☐ Copie Juré
☐ Copie Conseiller
☐ Copie Infothèque

h e g

Haute école de gestion
Genève

Proptech : la Data Science appliquée à l'immobilier

Travail de Bachelor réalisé en vue de l'obtention du Bachelor HES

par :

Yaffet TAMENU

Conseiller au travail de Bachelor :

Julien RIBON, enseignant vacataire

Genève, le 15.01.2021

Haute École de Gestion de Genève (HEG-GE)

Filière économie d'entreprise, orientation banque et finance

Hes·so GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

Déclaration

Ce travail de Bachelor est réalisé dans le cadre de l'examen final de la Haute école de gestion de Genève, en vue de l'obtention du titre de Bachelor of Science en économie d'entreprise, orientation banque et finance.

L'étudiant a envoyé ce document par email à l'adresse d'analyse remise par son conseiller au travail de Bachelor pour analyse par le logiciel de détection de plagiat URKUND. http://www.urkund.com/fr/student/392-urkund-faq

L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le travail de Bachelor, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du conseiller au travail de Bachelor, du juré et de la HEG.

« J'atteste avoir réalisé seul le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le 15.01.2021

Yaffet Tamenu

Remerciements

Je souhaiterais adresser mes remerciements à M. Ribon, mon conseiller pour ce travail, pour sa confiance et pour les échanges qui m'ont permis de concevoir cette étude.

Je tiens également à remercier mes proches pour m'avoir soutenu durant mon cursus académique. À mes parents, je suis reconnaissant des valeurs qu'ils m'ont transmises, de leur soutien indéfectible et de leurs encouragements qui m'ont maintenu droit quand je doutais.

Je remercie mes amis pour les échanges passionnés qui nourrissent ma créativité et mon ambition, nécessaires à tout projet de vie. Je remercie particulièrement Arnaud von Gross et Fanny Destenay pour leur art de vivre que j'aime partager en leur compagnie et pour leur relecture de ce travail qui les aura passionnés, à n'en point douter.

Pour finir, je remercie Chloé, sans qui je ne serai pas tout à fait la personne que je suis.

Résumé

La Proptech, ou Property Technology, regroupe un vaste ensemble de technologies qui ont pour principe sous-jacent la transformation digitale du secteur immobilier sous toutes ses formes et à chaque étape du cycle de vie de cet actif très particulier.

Ces technologies exploitent et transforment des données afin qu'elles procurent aux acteurs du marché immobilier une information exploitable et qu'elles garantissent une décision la plus éclairée que possible. Du Big Data à la *data science* en passant par l'informatique décisionnelle, les données sont au cœur de cette transformation digitale et nous parcourrons dans la première partie de ce travail l'évolution de ces technologies et leur impact sur notre manière d'appréhender l'environnement ainsi que le pouvoir prédictif qu'elles recèlent, sous la forme de modèles de régression, qui dévoilent les relations qu'elles peuvent avoir entre elles, sous-tendues par des concepts statistiques, mathématiques et algorithmiques.

L'environnement de la proptech suisse sera décrit de manière plus spécifique, afin de pouvoir se représenter les technologies qui sont actuellement utilisées pour répondre à nos besoins de mieux concevoir nos bâtiments, de mieux appréhender notre utilisation de l'espace et de mieux interagir avec ces lieux de vie.

Une partie de ce travail présentera brièvement le langage de programmation le plus utilisé par les *data scientists*, Python, dont les possibilités d'utilisation seront testées dans le cadre d'un cas pratique, dans le contexte du marché immobilier locatif genevois, élaboré selon deux axes de recherche. L'un de ceux-ci consistera à déterminer la disponibilité de l'information, sa fiabilité et sa complétude, pour un particulier souhaitant anticiper l'évolution du marché immobilier locatif local, en se défaisant du manque de transparence qui caractérise ce secteur. Le second reposera sur la capacité de cet outil à représenter l'information sous un format plus clair que les rapports et autres *factsheets* traditionnels.

Les limites de cette utilisation reposent sur l'incomplétude des données et leur fiabilité peu satisfaisante, qui réduisent la capacité des modèles de *data science* à délivrer des résultats plus complets et plus techniques. Des standards de récolte et d'échange de ces données devraient pouvoir aider les acteurs de l'immobilier à mieux exploiter ces technologies et permettraient de catalyser les interactions entre le secteur traditionnel et les start-ups de la proptech. Malgré ces frictions, la proptech et la *data science* donneront l'impulsion d'un profond remaniement du secteur immobilier.

Table des matières

Liste des tableaux

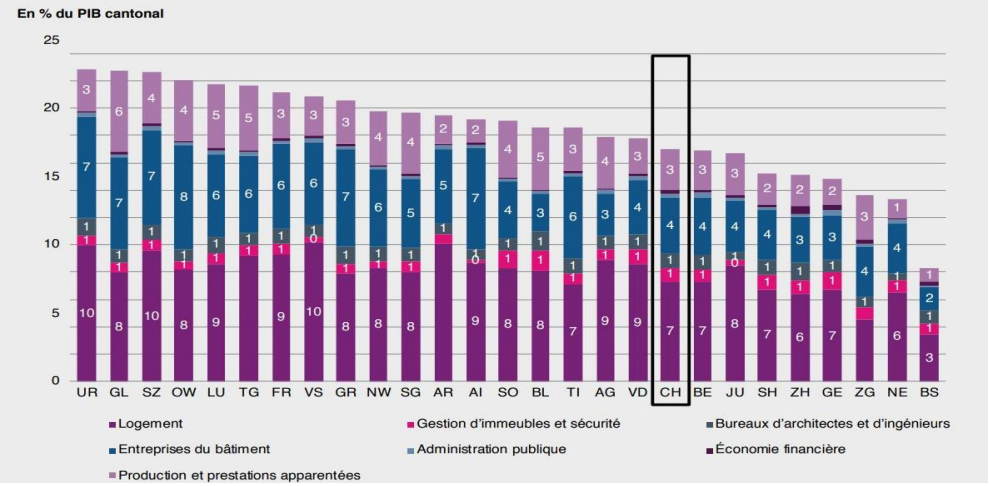
Liste des figures

1. Introduction

L'étude entreprise dans le cadre de ce travail ne saurait éluder la question de la crise sanitaire actuelle et de son impact sur l'économie mondiale. Cette crise, qui a frappé de plein fouet nos économies, aura un impact négatif profond sur l'activité économique mondiale. Selon les perspectives de l'économie mondiale publiées par le Fonds Monétaire International, le recul du PIB réel prévisionnel pour 2020 des suites de la pandémie de COVID-19 se précisait au mois d'octobre à -4,4% (FMI, 2020). La crise économique, que d'aucuns appellent aujourd'hui le « Grand Confinement », aura bouleversé tous les secteurs économiques et l'immobilier n'est pas en reste. Son impact sur ce secteur d'activité particulier se dessine sous divers aspects, allant de la réduction de la mobilité à l'augmentation du chômage et au recul de l'emploi dans un contexte où l'incertitude atteint un niveau paroxystique (GOODMAN, MCTAVISH, KLEIN, GOPI, 2020). Les chocs économiques de la crise mondiale sur les niveaux de production et de consommation auront des conséquences indéniables et complexes sur le marché immobilier suisse, et ce travail s'intéressera à étudier de manière plus précise comment suivre cet impact au niveau du canton de Genève, au travers de l'étude d'un cas pratique issu de l'analyse des données du marché immobilier locatif et de son environnement ainsi que de son contexte particulier.

Au niveau national, la part de contribution du secteur immobilier au PIB s'élevait en 2017 à 17% et ce secteur à lui seul contribuait à 14% de l'emploi national (HEV SCHWEIZ 2020). Malgré l'importante des disparités de la part de la valeur ajoutée brute imputable à l'immobilier dans le PIB des cantons, observables dans la figure 1 pour l'année 2017, la tendance à une dynamisation de l'emploi au sein du secteur

Figure 1-1 : Part de la valeur ajoutée brute imputable à l'immobilier dans le PIB des cantons 2017



(HEV SCHWEIZ, RÜTTER SOCECO AG, 2020, p. 23)

immobilier s'observe à Genève, qui a connu, entre 2011 et 2017, une augmentation de 10% du nombre d'emplois principalement imputables au secteur de la gestion d'immeubles et de la sécurité ainsi qu'au secteur de la construction (HEV SCHWEIZ, 2020, p. 24). Au 1er juin 2020, le nombre de logements vacants en Suisse atteignait un record de quelques 78'800 logements (OFL, 2020), tandis que le canton de Genève n'enregistrait que 1'169 logements vacants, soit un taux de vacance de 0.49%. Ces intrications macroéconomiques sur le logement au niveau cantonal permettront ainsi de maintenir tout au long de ce travail une compréhension de la pertinence des données utilisées afin de les analyser et de les utiliser de la manière la plus adéquate. Elles seront mises en exergue lors du développement des résultats obtenus dans le cas pratique et sous-tendront les prédictions que ce projet d'étude laisse présager.

Dans ce tableau général, qui dessine les pourtours d'un contexte plus spécifique, nous définirons et étudierons les différentes technologies que sont le Big Data, la Data Science et la Business Intelligence. Il sera constamment fait référence tout au long de ce travail à la relation entre les technologies évoquées et l'intérêt de leur utilisation dans le cadre du marché immobilier. Après une première explication relative aux moyens d'extraction des données d'intérêt, la deuxième partie de ce travail présentera diverses analyses des techniques et outils mathématiques et statistiques qui permettront d'exploiter ces données sous la forme de résultats visuels, graphiques et cartographiques. Un cas concret d'utilisation et d'exploitation de ces données achèvera d'expliquer et de prouver la puissance de ces technologies lorsqu'elles sont mises au service du marché de l'immobilier genevois, en démontrant quelles utilisations peuvent être faites de ces données et quelles perspectives sont dévoilées par l'utilisation de ces technologies dans ce secteur économique. Les limites de cette étude clôtureront la partie pratique de ce travail, qui s'inscrivent dans le cadre plus large de la data governance et de son importance pour les entreprises, ouvrant des perspectives sur la diffusion de ces technologies au sein de ce secteur économique.

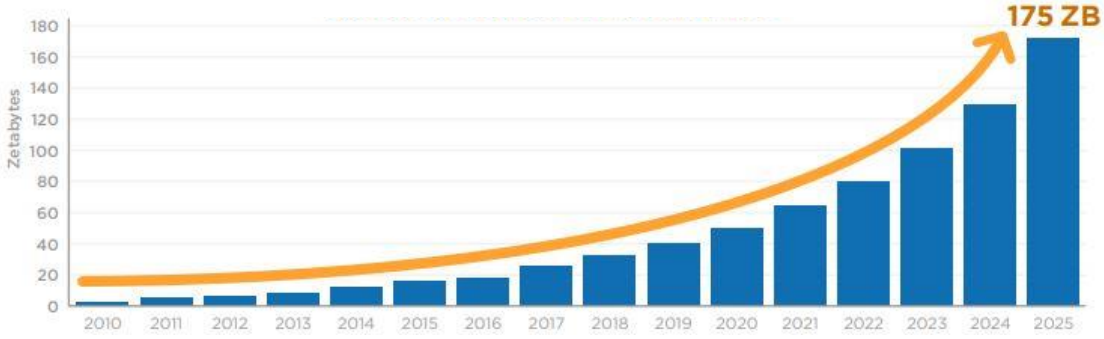
Le marché immobilier suisse, dans sa configuration actuelle, permet-il aux acteurs économiques d'exploiter toutes les ressources offertes par ces nouvelles technologies ? Sommes-nous à l'aube d'une nouvelle manière de concevoir les métiers de l'immobilier, de construire nos habitations et d'utiliser les espaces disponibles ? Quel contrôle supplémentaire permettent ces données et quelle valeur ajoutée ont-elles, d'aucuns se plaisant déjà à les appeler l'or noir du 21ème siècle ? Ce travail saura apporter à ces questions quelques éléments de réponse et aider le lecteur à mieux appréhender la puissance de la Data Science et de ces nouvelles technologies.

2. Technologies de la Proptech

En préambule aux définitions des termes et des technologies présentées ci-après, une base importante qui fondera la démarche de cette étude doit être évoquée. L'accès à l'information dans le marché immobilier n'est pas à la portée du public, si bien que le poids des acteurs dans ce marché biaise l'établissement d'un équilibre tel que l'on pourrait le concevoir dans un marché de concurrence pure et parfaite. Si le marché immobilier est un terme fréquemment utilisé, c'est qu'il est préférable d'éluder la complexité de ce qu'est l'immobilier, à dessein… Nombre d'acteurs tirent profit de cette asymétrie de l'information, bénéficiant des caractéristiques imparfaites des marchés immobiliers en question. L'immobilier est par définition peu liquide et les biens échangés sont hétérogènes. Il n'existe pas un marché immobilier mais une pléthore de micromarchés dont les caractéristiques varient fortement géographiquement. Ce sont ces constats simples qui permettent d'expliquer le besoin d'intermédiation important qui existe lorsque l'on s'intéresse à acquérir, vendre, estimer ou louer un bien immobilier. La principale raison qui fonde ce besoin repose sur une notion simple : sans données, il n'y a pas d'informations, donc pas de décision.

Le Forum Économique Mondial publiait le 26.04.2019 un article dans lequel était annoncé le chiffre vertigineux de 33 zettaoctets (10^{21}) de données numériques produites durant l'année 2018 (FEM, Statista France, 2019).

Figure 2-1 : Annual Size of the Global Datasphere



(IDC Global DataSphere, 2018, p. 6)

La quantité de données produites annuellement ne cesse de croître et cette tendance semble s'inscrire dans un horizon de long terme. Pourtant, un problème persiste. Dans un article publié le 13 janvier 2018, le magazine Forbes titrait que la donnée brute semblait perdre de l'intérêt, mais que c'est son utilisation qui devenait l'enjeu (FORBES, 2020). Quel intérêt peut avoir une donnée si elle n'est pas exploitée ? Cette partie, après un tour d'horizon des technologies actuelles exploitables dans le contexte du marché immobilier, se concentrera sur les moyens à mettre en œuvre afin d'exploiter les données efficacement dans le processus de prise de décision.

2.1 Le Big Data

Avant de s'intéresser au Big Data dans sa singularité, il semble opportun de contextualiser l'émergence de cette technologie dans le plus vaste ensemble qu'est la « révolution numérique » et des changements profonds qu'elle a provoqués au sein de la société. Les technologies de l'information ont bouleversé notre conception du monde, au-delà de toute autre technologie précédente. Ces changements s'appuient d'une part sur une évolution technologique dont seront décrits en premier lieu les origines et les aspects distinctifs. Après avoir présenté quels outils et quelles technologies ont sous-tendu notre capacité à échanger et à analyser les données, le paragraphe suivant s'intéressera à la naissance du Big Data et à ses premières utilisations comme indicateur précoce de tendance ou de prévisions. Ce n'est qu'à ce moment qu'une pleine compréhension de la rupture technologique qu'est le Big Data sera possible. Ce travail veillera notamment à présenter quelles utilisations en sont faites dans le cadre du marché immobilier en présentant les entreprises qui participent à ce progrès technologique.

Au cœur de toutes ces technologies, la donnée représente l'élément fondamental qui a suscité l'intérêt des scientifiques. Afin de la mesurer, de la stocker, de l'analyser et de l'exploiter, l'Homme n'a eu de cesse au cours des derniers siècles de chercher quels supports et quels outils seraient les plus rapides et plus efficaces à la réalisation de ces tâches. Cette rétrospective historique se focalisera essentiellement sur les technologies qui ont permis la dématérialisation des données et leur échange quasi-instantané. La société, telle que nous la connaissons aujourd'hui, ne saurait exister sans la contribution de ceux qui ont construit les fondations de l'ère numérique et ce travail se doit d'en faire la mention.

2.1.1 La « Révolution numérique »

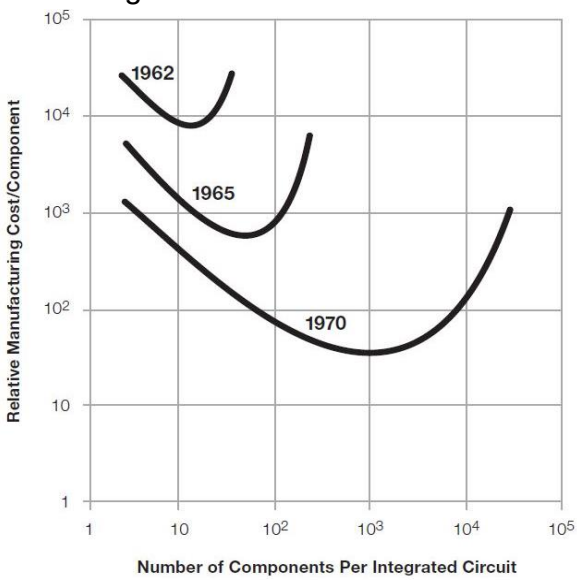
Les grandes révolutions, qui ont accompagné les progrès de l'humanité, ont permis à l'être humain de mieux comprendre et utiliser son environnement. De l'apparition de l'écriture à l'imprimerie, qui permettent d'échanger de l'information d'un individu à un autre sur un support matériel transmissible, de la création des outils aux machines, qui dispensent les hommes de fournir l'entièreté du travail de production, les technologies, de la plus rudimentaire à la plus élaborée, ont accompagné l'émancipation humaine. Les limites spatiales et temporelles, contraintes de notre environnement, se modifient, et avec elles, notre capacité à appréhender le monde, à concevoir l'avenir, nous donnant la possibilité de nous consacrer à des tâches sans cesse plus complexes.

Ce besoin d'émancipation est la genèse de l'automatisation et des technologies qui permettent l'implémentation de ces processus. Avec la découverte des nouvelles sources d'énergie que sont l'électricité et le pétrole, la capacité humaine à produire et à déplacer sa production prend une dimension mondiale. Étayant ces avancées technologiques, la nécessité d'un réseau de communication à longue portée performant et efficient devient une priorité. Ce dernier s'appuie sur deux composantes principales : les composants matériels, au travers desquels est émise, transportée et reçue l'information, ainsi que la capacité à dématérialiser les données, à les mesurer, les transformer et les stocker.

Alors que l'année 1858 voit le déploiement du premier câble transatlantique reliant les systèmes de communication télégraphique entre l'Europe et les États-Unis, il faut attendre le milieu du 20ème siècle pour voir naître les premiers ordinateurs entièrement électroniques. Les dimensions de ces machines, utilisées principalement dans le cadre militaire, impressionnaient, affichant, pour l'ENIAC par exemple, un poids de plus de 30 tonnes pour une surface occupée de 167 mètres carrés (WIKIPEDIA, Histoire des ordinateurs, 2020). Ces caractéristiques physiques semblent ne pas avoir de rapport direct avec la chronologie de la science des données, mais elles permettent néanmoins d'introduire quelques lois qui ont façonné l'industrie des systèmes électroniques et de ce fait, modifier en profondeur notre capacité à utiliser ces systèmes pour traiter les données mesurées.

En 1965, Gordon E. Moore, cofondateur d'Intel, énonce dans le magazine Electronics, la loi suivante qui portera son nom. Il constata que la complexité des semi-conducteurs depuis 1959 doublait tous les ans à coût constant, ce qui lui laissa supposer que cette

Figure 2-2 : La Loi de Moore



(ELECTRONICS VOL.38 N°8, GORDON E. MOORE, 1965)

dynamique se poursuivrait vraisemblablement dans le court terme (GORDON E. MOORE, 1965). La **loi de Moore** et son adaptation de 1975, dans laquelle le délai sera revu à 18 mois, naissent de cette observation empirique qui aura une étonnante pertinence et sera par la suite largement utilisée dans l'industrie, sous la forme d'une feuille de route publiée par l'Institut des ingénieurs électriciens et électroniciens, plus connu sous le sigle IEEE en anglais. Bien que la course à la miniaturisation semble se confronter à un obstacle majeur, l'atome et le nœud de 3 nanomètres (SCIENCES ET VIE, 2019), il n'en demeure pas moins que la puissance de calcul des processeurs a connu un développement exponentiel. C'est ce progrès technologique qui a permis l'essor fulgurant de l'informatique et la diffusion de son utilisation dans tous les pans de la société. Dès lors, la transformation des données calculées par ces processeurs ouvre la voie à une utilisation presque illimitée des résultats de ces opérations. Alors que les scientifiques plongent leur regard dans l'univers de l'atome, nos écrans atteignent un degré de réalisme ressemblant à s'y méprendre avec une fenêtre ouverte sur le monde.

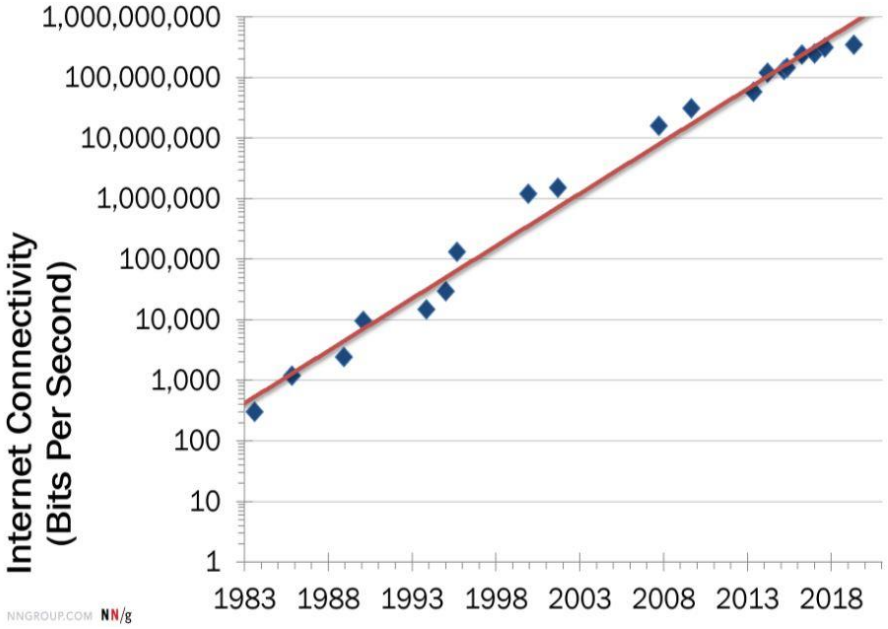
Parallèlement à cette expansion massive de la puissance de calcul, d'autres prouesses technologiques sont réalisées en matière de capacité de stockage. Intimement interdépendants, le calcul et le stockage ne sauraient avoir évolué l'un sans l'autre. Sur une période de cinquante ans, la densité d'informations enregistrées par un disque dur a été multipliée par 50 millions, passant de quelques 2000 bits à 100 milliards de bits, soit 100 gigabits. L'ordre de grandeur décrits par ces préfixes décimaux intègre notre quotidien, si bien que le 1er août 2005, dans un article publié dans la revue scientifique Scientific American, intitulé « **Loi de Kryder** », le journaliste William J. Walter Jr. décrit les observations de Mark Kryder, CTO de Seagate Corp., donnant naissance au principe qui servira de référence pour synthétiser la manière dont évolue la capacité de stockage :

> *Inside of a decade and a half, hard disks had increased their capacity 1,000-fold, a rate that Intel founder Gordon Moore himself has called "flabbergasting." (LOI DE KRYDER, MARK KRYDER, William J. Walter Jr. 2005)*

La capacité de stockage double tous les 13 mois alors que le coût est divisé par deux mais ces données n'auraient pas une utilité aussi importante sans la possibilité de les échanger, de les transmettre.

À cette fonction de transmission des données, un modèle mathématique existe, là encore. Postulée le 4 avril 1998 par Jakob Nielsen, expert mondialement reconnu dans le domaine de l'utilisabilité des sites web, la **Loi de Nielsen**, parue dans un article intitulé « Nielsen's Law of Internet Bandwith » prédit que la vitesse de croissance des réseaux publics est de 50% par an, soit un doublement tous les 21 mois (NIELSEN, 1998). Après avoir relevé ses propres données de connexion depuis 1984, son article, mis à jour en 2019, démontre l'étonnante précision de ce postulat comme le prouve le graphe suivant :

Figure 2-3 : La Loi de Nielsen



(NIELSEN 1998-2019)

De ces lois, qui sous-tendent l'accroissement de notre capacité à mieux gérer le traitement, le stockage et la transmission des données, il nous est permis de discerner les conditions de l'émergence du Big Data.

2.1.2 Caractéristiques du Big Data

Le terme Big Data ne trouve pas dans l'histoire une date de référence quant à la démocratisation de son utilisation. Dans un article du magazine Forbes publié en mai 2013, le journaliste Gil Press date sa première utilisation d'octobre 1997 dans la librairie digitale de l'Association for Computing Machinery, au sens communément employé de nos jours. L'article, publié par Michael Cox et David Ellsworth, présente les défis de la visualisation de données dont le volume excède les ressources de l'ordinateur, ce qu'ils nomment le problème du « Big Data » (PRESS Forbes 2013). De nombreuses définitions sont utilisées pour en expliquer la spécificité et ce travail s'en

tiendra à la définition proposée ci-dessous, qui se base sur les caractéristiques particulières liées au traitement et à l'utilisation des mégadonnées :

> *"Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value." (DE MAURO, GRECO, GRIMALDI 2016)*

Cette définition présente un élément fondamental de la problématique du Big Data, ou mégadonnées, en tant que leur traitement, leur analyse et leur stockage nécessitent des ressources spécifiques.

Dans le but de permettre la conceptualisation de critères discriminants à la catégorie de datasets considérés comme des mégadonnées, l'introduction de dimensions spécifiques à ces jeux de données a été proposée. En 2001, Doug Laney analysait la manière dont le commerce en ligne se trouvait confronté à un défi majeur dans le management des données collectées sur les utilisateurs et leurs transactions. Il décomposait alors en 3 dimensions ce problème : Volume, Vitesse, Variété (LANEY, 2001). Les 3 V du Big Data justifient et rendent nécessaire une nouvelle approche technologique pour valoriser ces jeux de données. La dimension du **volume** fait référence à l'explosion massive des quantités de données disponibles, rendant difficiles la sélection de données pertinentes et exploitables pour l'entreprise. La **vitesse** à laquelle fait référence Laney, soulève la problématique du traitement des données en temps réel. Cette dimension représente un intérêt tout particulier par les entreprises, permettant la proposition de contenu publicitaire personnalisé, ou la détection instantanée de fraudes… La **variété**, décrite par Laney comme la plus grande barrière à un management des données efficient, distingue les données générées par le Big Data par leur nature et leur provenance multiples. L'enjeu principal du traitement dimensionnel de cette variété de données réside dans l'exploitation de données de type non-structuré. On entend par le terme « non-structuré » l'impossibilité de représenter certaines données numériques dans un format prédéfini. Les informations qui proviennent de données non-structurées sont « toujours destinées à des humains » (MONINO, SEDKAOUI 2016, p. 13), en ce sens qu'elles ne sont par essence pas facilement adaptées au traitement par une machine. Qu'il s'agisse de texte, de vidéos, d'images ou de sons, la multiplicité de données de ce type est le cheval de bataille des entreprises, du fait que ces données représentent près de 80% des données générées par celles-ci et que leur volume tend à doubler chaque année (DELL 2020).

Ces dimensions semblent pourtant ne pas avoir le caractère discriminant souhaitable à la définition de ce que sont les datasets du Big Data. C'est à cette fin que Rob Kitchin

et Gavin McArdle ont proposé dans un article datant de janvier 2016 une approche systématique pour définir ce qui constitue de manière discriminante les mégadonnées. Cette « taxinomie » basée sur l'analyse de 26 datasets communément acceptées comme faisant partie du Big Data, les explore sous le prisme de sept axes formulés par Kitchin, que l'on retrouve dans le tableau ci-dessous. Le premier résultat important de leur analyse soutient qu'il n'existe pas une seule entité de datasets que l'on nommerait Big Data, mais qu'il en existe de multiples. Le deuxième élément, qui introduit le caractère discriminant du Big Data, s'appuie sur deux principales caractéristiques déterminantes : **la vitesse** et **l'exhaustivité** (KITCHIN et MCARDLE 2016). Alors que certains critères comme le volume ne semblent pas être retenus par Kitchin et McArdle, ces derniers avancent que le Big Data est caractérisé par des données générées de manière continue. D'autre part, les données provenant de petits jeu de données semblent selon leur taxonomie ne pas englober l'ensemble des observations de la population, mais n'en captent qu'un échantillon. Le Big Data, quant à lui, tend à capter toutes les observations disponibles, et de ce fait, toute la population (n=all).

Table 1 : Comparaison de Kitchin – Small vs. Big Data

	Small data	**Big Data**
Volume	Limited to large	Very large
Velocity	Slow, freeze-framed/ bundled	Fast, continuous
Variety	Limited to wide	Wide
Exhaustivity	Samples	Entire populations
Resolution and indexicality	Course and weak to tight and strong	Tight and strong
Relationality	Weak to strong	Strong
Exstensionality and scalability	Low to middling	High

(KITCHIN et MCARDLE 2016)

Ces éléments nous conduisent à penser que l'instantanéité de ces données et leur caractère exhaustif permettent l'élaboration de modèles prédictifs à court terme qui surpassent les modèles antérieurs, en réduisant notamment le temps de latence à l'agrégation et à l'analyse de ces données. Ce postulat a été soumis à des tests en temps réel, empiriquement modélisés pour répondre en particulier à cette latence de traitement des données. Ces tests ont eu de nombreux échos dans le monde scientifique, où la capacité prédictive de ces modèles s'est avérée d'une étonnante pertinence.

2.1.3 Prévoir le présent

Prévoir le présent, ou **nowcasting** en anglais, représente un défi majeur pour les entreprises et pour la société, qui pourraient s'appuyer sur ces prédictions pour ajuster leurs décisions en temps réel. Ce travail mentionnera l'exemple de **Google Flu Trends**, dont l'objectif est d'aider à prédire l'activité de la grippe saisonnière à un stade précoce en se basant sur les requêtes effectuées sur le moteur de recherche. Publiée dans la revue scientifique Nature le 19 février 2009, les résultats de ce modèle laissent entrevoir l'impact majeur de l'application globalisée du Big Data à d'autres domaines. L'écho que ce modèle peut avoir dans le contexte actuel de pandémie mondiale prouve l'utilité de la démarche, et justifie naturellement que l'on s'intéresse à la question de la rapidité de détection des cas de grippes. Cette démarche pourrait être entreprise d'une façon similaire pour modéliser le comportement des intervenants sur le marché immobilier. Ce travail y reviendra plus tard dans le cadre d'une preuve de concept spécifiquement conçue pour le marché immobilier genevois.

Les épidémies de grippe suscitent historiquement un intérêt médical particulier. Les virus de la grippe ont pour particularité de muter de manière plus fréquente que d'autres types de virus, étant donné le fait qu'ils consistent essentiellement en des virus à ARN. La spécificité de ces virus rend difficile l'élaboration d'un vaccin ou d'un traitement garantissant le contrôle épidémiologique et la propagation du virus au sein de la population. L'enjeu principal auquel sont confrontés les scientifiques réside principalement dans la détection de la recrudescence de l'épidémie et la mise en place de dispositifs sanitaires adéquats pour limiter sa propagation (Delort 2018, p.32). Outre la création du vaccin, dont le temps de conception et de fabrication demeure, dans l'état actuel de la microbiologie, incompressible au-delà d'une certaine limite, il s'agit donc de mesurer, dans un délai aussi court que possible, le taux de contamination observé dans la population à un instant donné. L'idée présentée dans l'article intitulé « *Detecting influenza epidemics using search engine query data* » se fonde sur l'observation suivante : la fréquence relative de certaines requêtes, effectuées sur le moteur de recherche Google, démontrent un niveau élevé de corrélation avec le pourcentage de visites médicales pour lesquelles un patient présente des symptômes associés à la grippe (GINSBERG, MOHEBBI, PATEL, BRAMMER, Nature, 2009, p.1012). L'article mentionne en préambule que les systèmes traditionnels de surveillance, le CDC (Center for Disease Control and Prevention) aux États-Unis et l'EISS (European Influenza Surveillance Scheme) en Europe, reposent principalement sur des données virologiques et cliniques dont la publication comporte un temps de latence d'une à deux semaines. Le modèle, développé quelques paragraphes plus loin

dans l'article, suggère d'estimer la probabilité d'une visite médicale dans une région spécifique pour des symptômes assimilables à la grippe. L'unique variable explicative qu'ils proposent d'utiliser pour ce faire se base sur la probabilité qu'une requête aléatoire soumise depuis la même région soit associée à des symptômes assimilables à la grippe. La capacité d'automatiser la méthode pour déterminer cette variable par régression permet de comprendre la portée du Big Data. Ce sont en effet 5 années de requêtes sur Google, entre 2003 et 2008, soit plusieurs centaines de milliards de requêtes individuelles, qui ont été agrégées dans le modèle américain, anonymisées et isolées pour chaque état.

Sans décrire en détail la méthodologie utilisée afin de modéliser ces données par régression linéaire, modélisation qui fera l'objet d'une partie dédiée à la data science appliquée dans le cadre du marché immobilier genevois, il est intéressant de se concentrer sur l'approche générale qu'ont utilisés les équipes de Google et du CDC (Delort 2018, p. 33-34). Contrairement aux modèles analytiques traditionnellement utilisés, le choix d'utiliser les données comme l'origine du raisonnement, et non comme un argument d'une variable introduite dans un modèle prédéfini, est l'un des aspects qui dévoila la puissance du Big Data. C'est cette démarche, que l'on nomme raisonnement inductif, qui a été privilégiée dans le modèle utilisée pour Google Flu Trends. En les comparant avec les données historiques provenant du CDC, le modèle récompense les requêtes agrégées par région qui démontrent la meilleure corrélation avec les variations régionales provenant de l'historique du CDC : Afin de déterminer le nombre de requêtes inclues dans la modélisation de la variable explicative de la régression, les modèles de *n* requêtes ont été comparés par leur corrélation moyenne aux données historiques. Le résultat délivré par cette évaluation identifie que le modèle basé sur les 45 premières requêtes obtient la corrélation maximale, comme le montre la figure 2-4 ci-après (GINSBERG, MOHEBBI, PATEL, BRAMMER, Nature, 2009, p.1013). Ces 45 termes de recherche, bien que sélectionnées de manière automatique, ont un lien évident avec les symptômes de la grippe. Cela permet notamment de valider le postulat selon lequel les termes de recherche sur Google constituent, au moins partiellement, un indicateur précoce de la recrudescence de l'épidémie de grippe.

Figure 2-4 : Google Flu Trends : Corrélation moyenne par nombre de requêtes

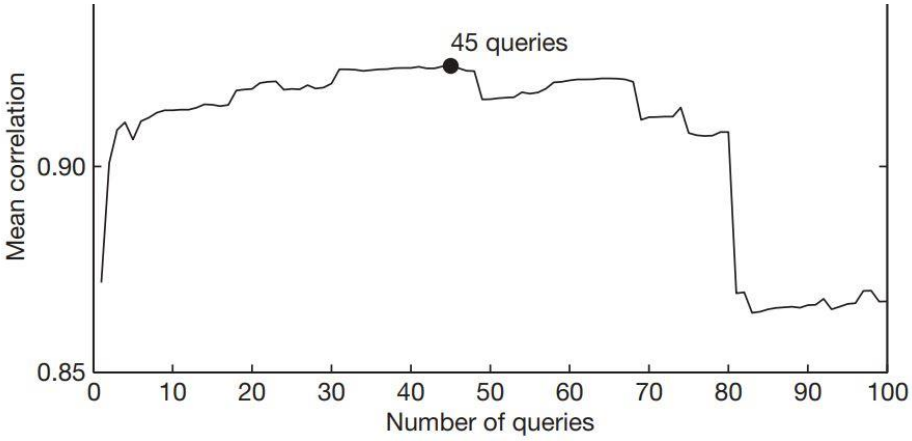


Figure 1 | An evaluation of how many top-scoring queries to include in the ILI-related query fraction. Maximal performance at estimating out-of-sample points during cross-validation was obtained by summing the top 45 search queries. A steep drop in model performance occurs after adding query 81, which is 'oscar nominations'.

(GINSBERG, MOHEBBI, PATEL, BRAMMER, Nature, 2009, p.1013)

En observant le graphique, on constate qu'après l'ajout de la 81ème variable, « oscar nomination », le modèle perd de sa concordance, ce qui permet d'anticiper l'une des problématique principale du Big Data, à savoir que les modèles obtenus par inférence comportent intrinsèquement une part d'erreur qu'il faut savoir apprécier. Dans le cas présent, la saisonnalité de certaines requêtes semble coïncider avec l'épidémie de grippe. Ces modèles, dépendants de la source des données qu'ils exploitent, sont négativement influencés par les corrections apportées à cette source, notamment par l'autocomplétion des recherches sur Google introduite après la modélisation de Google Flu Trends (Delort 2018, p. 38-39). Ce travail abordera plus loin de manière plus étendue les limites de l'utilisation des mégadonnées, mais le résultat suivant permet toutefois de constater ce que permet le Big Data.

Après avoir validé le modèle sur la base des données historiques disponibles mises à disposition par le CDC, le modèle Google Flu Trends a été mis à l'épreuve durant la période 2007-2008 pour évaluer sa précision prédictive à court terme. La prévision du présent, ou nowcasting, permet ainsi de réduire d'une à deux semaines le délai d'agrégation des données réelles, et donne les résultats suivants, dont on notera l'étonnante pertinence, laissant présager des applications que cette technologie permettrait dans d'autres domaines.

Figure 2-5 : Google Flu Trends - Prévisions du modèle

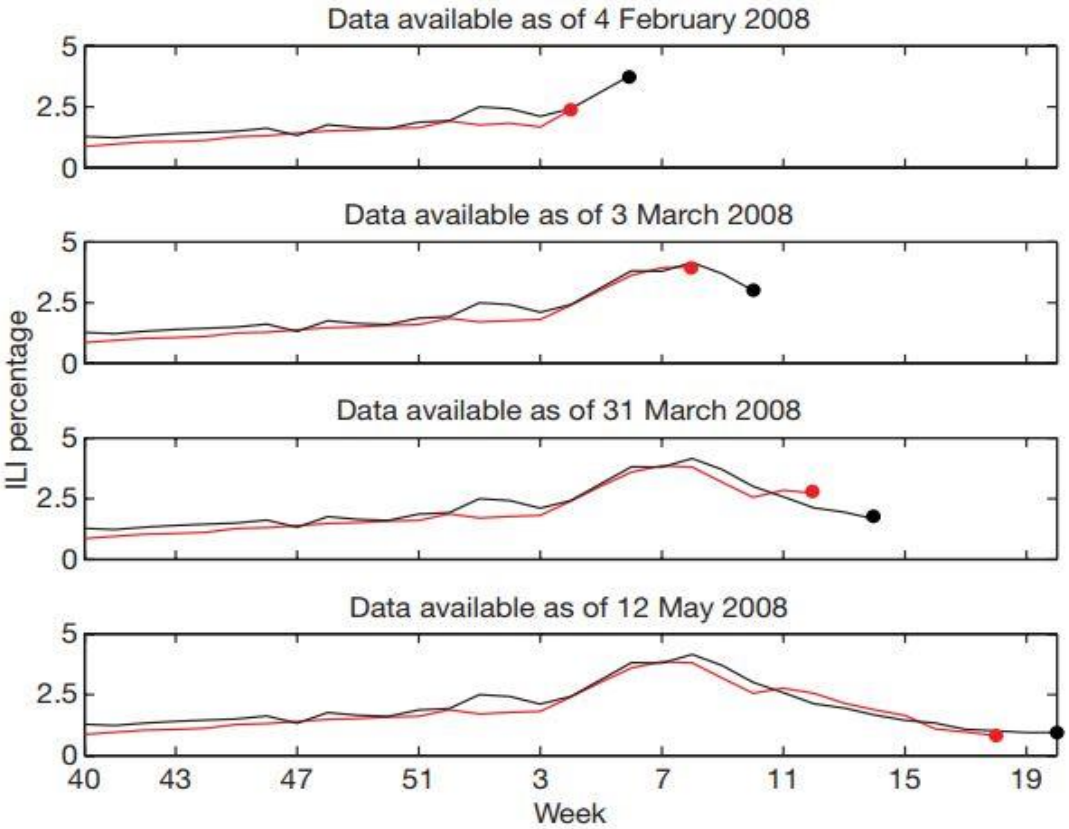


Figure 3 | ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5 we detected a sharply increasing ILI percentage in the mid-Atlantic region; similarly, on 3 March our model indicated that the peak ILI percentage had been reached during week 8, with sharp declines in weeks 9 and 10. Both results were later confirmed by CDC ILI data.

(GINSBERG, MOHEBBI, PATEL, BRAMMER, Nature, 2009, p.1013)

2.2 Immobilier : un marché en révolution

Doté d'une réputation traditionnaliste qui semble intrinsèque à ce secteur d'activité, l'immobilier représente le principal actif mondial et semble ne pas encore avoir complètement absorbé les bénéfices de la transformation digitale. Qu'en est-il réellement ? La partie qui suit aura pour but de renseigner le lecteur sur les bouleversements que connaît ce secteur fondamental à nos économies.

2.2.1 Tour d'horizon du secteur immobilier

Selon une étude de la société Savills datée de 2016, la valeur de l'immobilier mondial se montait à quelques 217'000 milliards de dollars (BARNES, TOTSEVIN, Savills World Research, 2016, p.4), en incluant l'immobilier résidentiel et commercial, mais aussi l'immobilier industriel, l'hôtellerie ou encore les terrains agricoles. Plus encore, l'étude révèle que cette classe d'actif représente 60% des actifs mondiaux et pèse près de 3 fois le PIB mondial. L'étude propose également une comparaison du secteur avec d'autres classes d'actifs, le résultat étant la figure suivante qui impressionne au vu de l'importante contribution relative de l'immobilier dans la richesse mondiale.

Figure 2-6 : Savills : Global asset universe 2016

Asset*	Investable (trillions)	Non-investable (trillions)	All (trillions)
ALL REAL ESTATE	**$81**	**$136**	**$217**
RESIDENTIAL	$54	$108	$162
HIGH QUALITY, GLOBAL, COMMERCIAL	$19	$10	$29
AGRICULTURAL LAND	$8	$18	$26
OTHER INVESTMENTS	**-**	**-**	**$155**
EQUITIES	$55	-	$55
OUTSTANDING SECURITISED DEBT	$94	-	$94
ALL GOLD EVER MINED	-	-	$6
GLOBAL MAINSTREAM ASSET UNIVERSE	**-**	**-**	**$372**

***(values in US$ trillions – rounded)** Sources: Savills Research, Bank for International Settlements, Dow Jones Total Stock Market Index, Oxford Economics

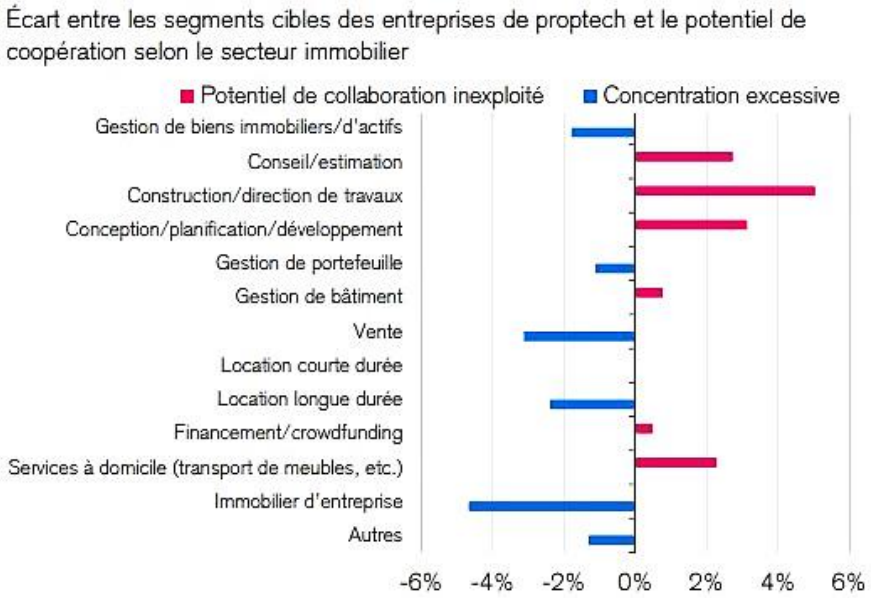
(BARNES, TOTSEVIN, Savills World Research, 2016, p.4-5)

L'étude mentionne par ailleurs que la majeure composante de la valeur immobilière mondiale consiste en des propriétés détenues à titre privé et que les propriétaires y résident dans une vaste proportion, ce qui permet de développer les deux déductions suivantes, à savoir que l'immobilier est le secteur le moins concentré en termes de propriété et le plus corrélé à la fortune propre des ménages (BARNES, TOTSEVIN,

Savills World Research, 2016, p.5). Dès lors, on est en droit de s'intéresser de manière plus spécifique à cette proportion non négligeable du marché immobilier qu'est le résidentiel et qui représente quelques 2,5 milliards de propriétés au niveau mondial, selon la même étude. Une large partie de ces propriétés, soit 75% environ selon la figure 2-6, ne figurent pas dans la catégorie des investissements réalisables, catégorie qui est évaluée quant à elle à 54'000 milliards de dollars et représente les propriétés qui sont activement intégrées au marché immobilier dans un objectif transactionnel (BARNES, TOTSEVIN, Savills World Research, 2016, p.5). D'autre part, le secteur immobilier dans une large mesure semble avoir pris conscience de la nécessité d'intégrer le digital dans la gestion de leur portefeuille. Les sociétés immobilières semblent pour une majeure partie, avoir mis en place une stratégie digitale (WEIR, PYLE, GRUNEWALD, KPMG International, 2019, p. 6), avec une volonté affirmée de créer une stratégie de gestion de leurs données (WEIR, PYLE, GRUNEWALD, KPMG International, 2019, p. 7). En Suisse, le principal moteur de ce changement de paradigme, souhaité et accueilli positivement par les acteurs traditionnels, met en évidence un besoin d'améliorer l'efficience opérationnelle dans la gestion courante des affaires, de réduire les coûts et de faciliter le processus décisionnel (WEIR, PYLE, GRUNEWALD, KPMG International, 2019, p. 8). Alors que les entreprises du secteur de la proptech en Suisse semblent surtout être freinées par des moyens financiers limités, empêchant un marketing efficace, les sociétés immobilières helvétiques connaissent quant à elles des lacunes imputables à des processus décisionnels complexes et longs. Le terme proptech sera défini ci-après, étant donné qu'il nécessite une plus profonde compréhension du contexte dans lequel il est utilisé et de la disruption que ce terme expose. Dès lors, il s'agit avant tout pour les proptechs de faire connaître leur produit et de se profiler comme le partenaire de confiance pour un domaine d'expertise spécifique (HASENMAILE, LOHSE, NÄF, RIEDER, WALTERT, Credit Suisse AG, Investment Solutions & Products ,2019, p. 9). La devise suivante, qui semble animer la concurrence intense qui règne dans l'environnement des start-ups de la proptech : « The winner takes it all » (HASENMAILE et al., 2019, p.9), explique la rapidité avec laquelle ces entreprises doivent s'établir sur le marché. Alors que le taux de pénétration des services de la proptech semble important, avec un taux considérable de 86% des sondés déclarant collaborer avec les start-ups, la grande majorité de ces relations sont récentes et près du tiers n'ont pas répondu aux attentes des sociétés immobilières qui initient souvent ces relations d'affaire (HASENMAILE et al., 2019, p.11). La coopération entre ces deux secteurs pourrait voir naître des opérations d'absorption de la part des sociétés immobilières qui détiennent pour 26% d'entre elles, une participation financière dans une ou plusieurs firmes (HASENMAILE

et al., 2019, p.11). La limite semble donc floue entre la collaboration et la phagocytose. Un élément reste toutefois intéressant. Les marchés ciblés par les entreprises de la proptech ne semblent pas toujours être en adéquation avec les attentes du secteur immobilier, et des divergences sont constatées quant au niveau de concurrence qui existe en fonction des divers marchés. Alors que certains d'entre eux semblent surexploités et verront vraisemblablement certaines start-ups souffrir d'une concurrence trop importante, d'autres dévoilent un écart à combler. Il serait intéressant pour les nouveaux produits et services proposés par les acteurs de la proptech de se concentrer sur ces activités, mentionnées dans la figure ci-dessous. On y découvre notamment que ce sont principalement les secteurs du conseil et de l'estimation, de la construction et de la direction de travaux ainsi que celui de la conception, de la planification et du développement qui pourrait bénéficier d'une collaboration intéressante dans les prochaines années. C'est le secteur du conseil et de l'estimation qui sera retenu pour le développement du cas pratique qui appuiera les réflexions de cette étude.

Figure 2-7 : Segments au potentiel inexploité ou surexploité



(HASENMAILE et al., 2019, p.11)

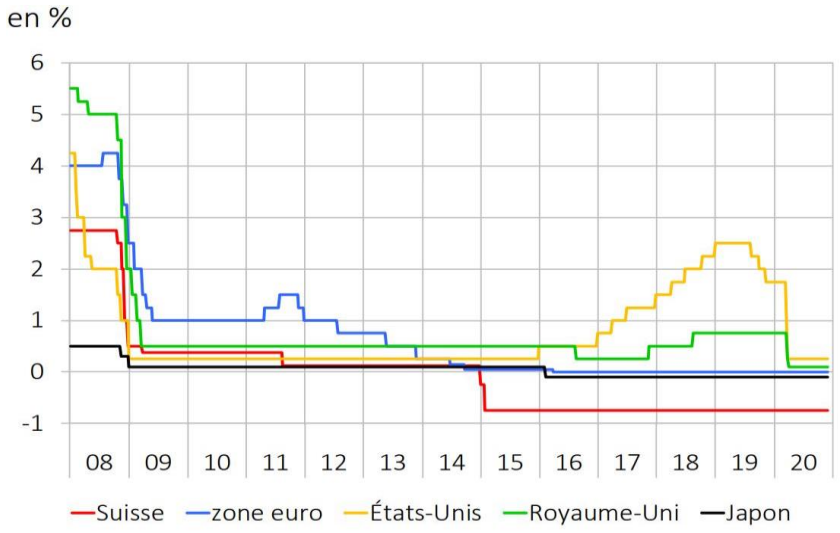
Avec des investissements en constante augmentation, totalisant depuis 2012 plus de 43 milliards de fonds levés (GIBOR, HAREL, TRAJTENBERG, Deloitte Netherlands, 2020), les start-ups de l'immobilier et de la construction se voient confronter aux frictions avec les acteurs du secteur dans un contexte post-pandémique qui rebat les cartes de l'allocation des espaces de vie et de leur fonction (DELOITTE CENTER FOR FINANCIAL SERVICES, 2020). Pour un bon nombre d'entre elles, la pandémie leur aura imposé des mesures de réduction des coûts, principalement pour les entreprises

spécialisées dans le cosharing et les espaces de coworking. Toutefois, de nombreux autres secteurs verront d'intéressantes opportunités se dévoiler dans des horizons de temps variés. De quelles manières les technologies du Big Data sont-elles en mesure de résoudre ces problématiques idiosyncratiques inhérentes au marché immobilier dans le contexte post-pandémique qui se profile ? D'autre part, si le marché immobilier est très fortement corrélé aux performances des économies locales, quels besoins ces technologies sont-elles susceptibles de résoudre pour les acteurs du secteur à l'ère du quantitative easing et des taux bas ? Ce travail s'intéressera, dans cette partie, au marché suisse de la proptech, d'une façon plus précise et spécifique, afin de mettre en évidence les problématiques locales intrinsèques à ce secteur. Ces éléments de réponse seront mis en perspective de manière plus globale avec les besoins du secteur au niveau mondial dans le contexte particulier de la crise sanitaire.

2.2.2 La Proptech et la data

En préambule, le contexte des taux d'intérêt négatif a modifié de manière profonde la structure des investissements à l'échelle mondiale. Le secteur immobilier est devenu ces dernières années, sous l'impulsion de la politique expansionniste des banques centrales des pays développés, une opportunité de diversification des placements et un potentiel de rendements intéressants pour les investisseurs.

Figure 2-8 : BNS - Taux d'intérêt de référence



(SECO, Tendances Conjoncturelles, 2020)

Ces mesures ont eu pour principal objectif de soutenir l'économie en luttant contre les impacts économiques de la crise sanitaire. Elles profitent toutefois à certains acteurs économiques, notamment aux entreprises, aux états ainsi qu'aux propriétaires de biens immobiliers ; quant aux banques, aux épargnants et aux caisses de pension, les mesures précitées tendent à affaiblir leurs marges pour les premières, répercutant ces

coûts supplémentaires sur les détenteurs d'épargne, respectivement sur la fortune des épargnants et les avoirs des caisses de pension. Dans le but de soutenir leurs finances, ces derniers cherchent à tirer profit de la conjoncture en cherchant notamment des placements sûrs, en première ligne desquels figurent les biens immobiliers. C'est dans ce contexte que le marché immobilier attire l'intérêt grandissant des investisseurs, en quête d'une stratégie de placement sûre intégrée à une gestion efficiente de leurs portefeuille d'actifs. À ces fins, des données disponibles, utilisables et immédiates soutiennent indéniablement le processus de décision, et cette tendance vient nourrir dans le secteur immobilier une forte demande de conseils qui se heurte à son immobilisme et à son inertie. Les entreprises technologiques répondent en cela à cette demande, avec agilité, ainsi qu'à ces nouveaux besoins, leur offrant une opportunité de se distinguer auprès de ces divers acteurs en leur proposant une valeur ajoutée, des données intelligemment captées, reproductibles et évolutives, intégrées à une stratégie adaptée aux clients.

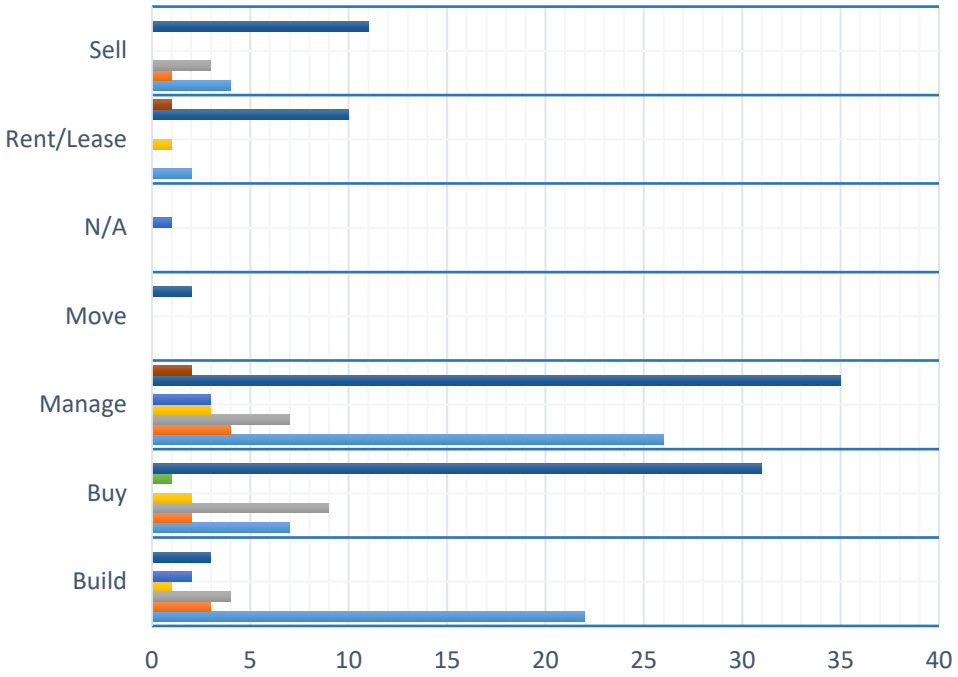
Diminutif de « Property Technology », la proptech est définie comme suit par James Dearsley, cofondateur d'Unissu, société leader de l'information sur les entreprises de la proptech, et Andrew Baum, *Professor of Practice* à *la Saïd Business School, University of Oxford*, et lauréat du *UK PropTech Association Special Achievement Award* en 2019 :

> *"PropTech is one small part of the wider digital transformation of the property industry. It describes a movement driving a mentality change with the real estate industry and its consumers regarding technology-driven innovation in the data assembly, transaction, and design of buildings and cities." (BAUM, DEARSLEY, Unissu, 2019)*

Unissu, cofondée en 2018 à Londres par James Dearsley, présenté précédemment et Eddie Holmes, tous deux experts confirmés en proptech, est une plateforme et un annuaire dont l'objectif principal est de permettre aux professionnels de l'immobilier de comprendre l'écosystème de la proptech et de sélectionner le produit ou le service qui correspond à leurs besoins. La plateforme Unissu recense près de 8'489 sociétés actives à travers le monde dans le domaine de la proptech à la fin de l'année 2020, dont 2'085 sociétés basées aux États-Unis, 957 basées au Royaume-Uni, 531 sociétés françaises, 408 sociétés espagnoles, 341 sociétés allemandes et 203 sociétés helvétiques. Une analyse des résultats pour les proptechs suisses semblait nécessaire à la compréhension de l'activité de ces sociétés en suisse et des technologies qu'elles utilisent afin de se différencier de l'offre de services du secteur. Ce travail se propose d'en dessiner les pourtours et de présenter l'une d'entre elles dans la partie suivante, dont l'ingéniosité et l'innovativité méritait d'être relevée.

Le graphique ci-dessous propose un aperçu des secteurs d'activité sur lesquels opèrent les proptechs suisses et le cycle de vie dans lequel elles interviennent. On constate que pour 65 % d'entre elles, le secteur résidentiel représente un marché dans lequel elles interviennent, avec un focus sur les phases d'achat et de management. La part d'entre elles qui opèrent dans le *retail*, soit le commerce de détail, est faible et ne constitue pas un marché encore développé pour la branche proptech en Suisse.

Figure 2-9 : Proptech Suisse - Secteurs et Cycles de vie



	Build	Buy	Manage	Move	N/A	Rent/Lease	Sell
Retail			2			1	
Residential	3	31	35	2		10	11
Financial Service (non renseigné)		1					
Commercial + Retail	2		3		1		
Commercial + Residential + Retail	1	2	3			1	
Commercial + Residential	4	9	7				3
Commercial + Residential	3	2	4				1
Commercial	22	7	26			2	4

(Unissu, 2020)

Tout secteur confondu, ce sont les phases de management et d'achat qui semblent avoir permis au plus grand nombre de ces start-ups de trouver des applications dans le secteur immobilier. On constate cependant qu'elles sont encore très peu représentées pour la phase du déménagement, avec seulement deux start-ups qui opèrent sur ce cycle de vie du bien, ou de la phase de location.

La topologie de la proptech est encore relativement complexe compte tenu de l'étendue des applications de ces technologies dans le secteur immobilier. Les principales classification apparaissent dans le tableau ci-dessus et concernent le secteur d'activité et la phase du cycle de vie de l'entreprise. Parmi les autres classifications qui apparaissent sur le site Unissu, on trouve le type de technologie utilisée, le business modèle, la sous-catégorie industrielle à laquelle s'apparentent la start-up et le service qu'elle propose.

Afin que le lecteur puisse se faire une idée globale du type de technologie qui est actuellement mis en œuvre dans la branche de la proptech suisse, voici une liste non exhaustive, néanmoins relativement précise et générale, de technologies qu'il paraît important de mentionner. Ce sont ces technologies qui assurent au secteur de la proptech sa valeur ajoutée et qui permettront, par les données qu'elles génèrent, ainsi que par la facilitation des transactions exécutées sur l'ensemble de la chaîne de valeur immobilière, de véritablement changer la manière dont est pensé et commercialisé l'immobilier.

Table 2 : Technologies utilisées par la proptech en Suisse

3D Modelling	**Crowdfunding**	**Parking**
3D Printing	**Data Analytics**	**Peer to Peer**
Artificial Intelligence	**Design**	**Predictive analytics**
Augmented Reality	**Digital Twin**	**Robotics**
Automated Valuation	**Drones**	**Sensors**
Big Data	**Geolocation**	**Smart Building**
BIM	**Geospatial**	**Smart City**
Blockchain	**Internet of Things**	**Smart Home**
CAD	**Location analytics**	**VR/AR**
CRM	**Mapping**	**Work Flow Management**

(Unissu, 2020)

Dans ce cadre, le principal intérêt de la technologie Big Data consiste en la captation de données à faible densité d'informations pour trouver par induction et de manière exploratoire, du fait de leur importante volumétrie, des modèles à capacité prédictive (Delort 2018, p. 41). Il se trouve que les données générées par le secteur immobilier ne sont pour l'heure que peu accessibles et restent pour l'essentiel récoltées et traitées de manière informelle et peu, voire pas standardisée. La *Royal Institution of Chartered Surveyors*, organisme professionnel de renommée mondiale ayant pour vocation la formation des professionnels du secteur immobilier, la promotion de la profession et des normes déontologiques et éthiques ainsi que l'élaboration des solutions pour l'avenir du secteur, s'est exprimé par la voix d'Andrew Knight, directeur international

des standards de données, sur l'avenir de la proptech. Il décrit que les acteurs du secteur agissent, sous l'impulsion de facteurs et comportements spécifiques au marché immobilier, de manière à ne pas permettre l'échange de données (KNIGHT, RICS, 2020). Si la source des données doit être qualitativement irréprochable, ce dernier semble encourager une plus grande transparence du marché, sous réserve de la confidentialité qu'impose la sensibilité de certaines données. La confidentialité de certaines données immobilières, tout comme l'avantage concurrentiel qu'elles permettent, est sans doute l'une des principales raisons de la complexité d'implémenter de manière globale et directe les solutions technologiques à ce marché.

La bataille du contrôle des données reste une question à laquelle le secteur immobilier devra trouver des réponses, bon gré mal gré. Cette question n'empêche pourtant pas les acteurs d'avancer dans cette nouvelle ère qu'augurent ces nouvelles technologies. L'émergence de sociétés spécialisées dans la collecte des données produites durant tout le cycle de vie des biens immobiliers est proche. Avec l'entrée des géants de la technologie dans le marché de l'Internet des Objets, notamment Alphabet avec Nest, Samsung avec SmartThings et Amazon avec Alexa, qui ont commencé leur installation progressive au sein des ménages, l'intérêt de ces sociétés pour le marché immobilier commercial semble logique (VON DITFURTH, AHOLT, Deloitte France, 2018, p.4). Leur avantage concurrentiel est immense et il se peut que ces acteurs deviennent les outsiders d'une véritable transformation du marché immobilier, dans une course qu'ils semblent bien placés pour remporter.

L'ère de la robotique, de la réalité virtuelle ou de la réalité augmentée, des capteurs, de l'Internet des Objets et du Building Information Modeling, présage de véritablement remodeler le modèle traditionnel du fonctionnement du secteur immobilier, dont le nerf de la guerre devient la donnée. Ce travail se propose, dans la partie qui suit, de comprendre comment la valoriser et l'exploiter de telle sorte qu'elle devienne un support décisionnel. Par ailleurs, et avant de clore cette partie, il faut garder à l'esprit que le besoin technologique suscité par la pandémie, véritable catalyseur de ces transformations, n'a pas encore dévoilé son ampleur réelle. Alors que le vaccin contre ce nouveau virus est accueilli avec l'espoir d'un retour à la normale, il se peut qu'une nouvelle norme s'établisse et qu'elle redéfinisse complètement nos espaces intérieurs de leur fonction à leur utilisation (SUN, The Real Deal New York, 2020).

3. La Data Science et la Business Intelligence

Au-delà des données, il y a leur utilisation et leur pouvoir indicatif, sur l'état d'un objet ou d'un système à un instant donné, mais aussi leur pouvoir prédictif, soit leur projection dans le temps et l'extrapolation de leur évolution à partir des observations, ainsi que la détermination du biais de cette estimation. Le principal intérêt de la récolte des données, qu'elles soient issues ou non du Big Data, consiste en la possibilité d'en extraire une information utile et de modéliser leur comportement de façon à permettre une prise de décision. Ce chapitre s'intéressera en particulier aux outils mathématiques, statistiques et techniques ainsi qu'aux différents logiciels nécessaires à la fouille des données présentes sur le web, aussi appelé *data mining*, à leur exploration et leur classification, ainsi qu'à l'exploitation et à la visualisation des données à l'ère du Big Data.

Le langage de programmation Python sera étudié dans cette partie de manière plus approfondie, tant son utilisation s'est imposée auprès des *data scientists* du monde entier comme un outil de premier choix. L'ensemble du cas pratique sera développé dans ce langage pour les raisons évoquées précédemment, tout autant que pour sa popularité, aidant tout programmateur débutant à trouver de très nombreuses sources de formation et d'apprentissage en ligne, ou pour sa simplicité et sa flexibilité, que ce travail abordera dans la partie qui sera dédiée à ce langage.

D'autre part, ce travail proposera, pour clore la thématique de l'exploitation des données, de parcourir l'état actuel de l'informatique décisionnelle, ou Business Intelligence et des logiciels qui s'en servent. La Business Intelligence, dont l'acronyme communément utilisé BI sera utilisé ci-après, concentre l'attention de l'informatique autour de la prise de décision et du pilotage. Son utilisation à l'ère de la veille stratégique en fait un atout primordial et sa mention dans le cadre de ce travail devait être faite, tant elle s'est imposée dans l'univers professionnel comme un moyen rapide et efficace d'analyser, de représenter et d'interpréter les données produites en entreprise.

Les thèmes qui vont suivre serviront avant tout à expliquer quelles technologies se cachent derrière une prise de décision orientée données, et le bénéfice de cette pratique pour toute entreprise ou organisation qui l'utilise. Avant de recenser ces outils et ces technologies, l'intérêt de la prise de décision orientée données, ou *data-driven decision-making, DDD*, doit être clairement établi. Il ne s'agit pas de comparer une entreprise particulière à une autre d'un point de vue concurrentiel, mais d'identifier comment la DDD accroît la performance relative d'une entreprise par rapport à

l'ensemble de ses concurrents en fonction de son degré d'utilisation. La généralisation des avantages de l'utilisation de la DDD a été étudiée par l'économiste Erik Brynjolfsson, Lorin M. Hitt et Heekung Hellen Kim dans un article paru le 22 avril 2011, intitulé « Strength in Numbers : How Does Data-Driven Decisionmaking Affect Firm Performance ? ». L'analyse menée dans ce rapport se base sur l'étude de 179 sociétés cotées en bourse aux États-Unis et détermine une méthode de classification de celles-ci en fonction de leur degré d'utilisation de la DDD. Les résultats de cette étude ne laissent subsister de doutes quant à l'importance de la DDD pour les entreprises. Elle démontre non seulement qu'il existe une relation linéaire positive entre l'utilisation de la DDD et la productivité, mais aussi que les sociétés qui utilisent la DDD augmentent leur productivité d'un ordre de grandeur de 5 à 6 pourcents et que cette relation s'étend à de nombreuses autres mesures de performance, dont le retour sur investissements, l'utilisation des actifs ou la rentabilité des capitaux propres, semblant suggérer qu'un lien de causalité avéré existe (BRYNJOLFSSON, HITT et KIM 2011, p.1).

Ces résultats permettent la conclusion suivante. Il faut admettre que tout raisonnement orienté données déploie des avantages concurrentiels stratégiques supérieurs à un raisonnement intuitif. L'intérêt de cette pratique suscite dès lors des questions quant aux concepts qu'elle utilise, aux outils et aux technologies qui la rendent possible. Afin d'en saisir toute la portée et les nuances, il paraît utile de comprendre comment fonctionnent ces technologies. Si le but n'est pas de les utiliser ou de les mettre en pratique, tout décideur doit savoir identifier les avantages et inconvénients d'un projet de data mining et définir, dans sa phase conceptuelle, toute erreur flagrante ou hypothèse erronée (PROVOST et FAWCETT 2018, p. 11).

3.1.1 Statistiques, data mining et data science

Les modèles statistiques classiques constituent la base de la data science actuelle et permettent de synthétiser un jeu de données, ou datasets, de manière à rendre son exploration plus aisée. Quelques éléments sont cependant à mettre en évidence, qui ont trait aux possibilités d'automatisation des calculs permises par l'informatique. On notera notamment que la statistique classique, avant l'ère informatique, se basait sur des calculs manuels limités. Cette seule contrainte limite les volumes de données qui sont incorporées dans les études et oblige souvent les statisticiens à travailler selon la méthode dite hypothético-déductive (TUFFÉRY et SAPORTA 2017, p. 3), avec de fortes hypothèses préalables, parmi lesquelles on citera les hypothèses sous-tendant la régression linéaire notamment, à savoir la linéarité, la normalité et l'homoscédasticité (TUFFÉRY et SAPORTA 2017, p. 3), ainsi que les principes de l'inférence statistique.

Afin de comprendre les principales thématiques qu'aborde la data science, les bases statistiques classiques sont un prérequis important qui serviront non seulement à mieux appréhender le problème à résoudre mais aussi à comprendre les données et leurs relations. Ceci servira notamment à éviter la redondance de certaines informations dans les modèles développés, ou à choisir l'approche la moins coûteuse et la plus efficiente. Cette partie propose donc un bref rappel des outils statistiques d'intérêt dans le cadre d'une étude de data science.

On fera ici une distinction importante entre les statistiques univariées, dont les mesures de tendance centrale, de dispersion et de forme renseignent sur les caractéristiques d'une variable unique et les statistiques multivariées, qui mettent en évidence des liens entre les variables explicatives, dont la corrélation pour les variables quantitatives ou l'indépendance pour les variables qualitatives. Si les statistiques univariées permettent d'extrapoler des caractéristiques observées dans un échantillon à la population dans son ensemble, ces prédictions sont limitées à un critère en entrée et manqueront de satisfaire les exigences de modèles à plusieurs facteurs. Les statistiques multivariées comblent les lacunes du modèle univarié en proposant un ensemble de techniques qui permettront notamment de déceler des similarités entre les données, ou de classifier les données en fonction des paramètres pris par les facteurs en entrée. Les statistiques sont un domaine d'étude vaste et complexe et ce travail n'ambitionne pas d'en faire un exposé exhaustif. Il sera fait mention des principaux outils statistiques employés pour résoudre des problèmes usuels et d'apporter des précisions sur les conditions de l'utilisation des modèles statistiques courants. Les mesures de tendance centrale, de dispersion et de forme ne seront pas développées dans ce chapitre, afin de maintenir un niveau de développement plus détaillé pour les thématiques principales utilisées dans la pratique.

3.1.1.1 Statistiques probabilistes et inférentielles

Ce domaine d'étude des statistiques regroupe les notions d'échantillon, de paramètre et de population. L'intérêt spécifique de ces études consiste en l'extrapolation de caractéristiques observées sur un échantillon à l'ensemble de la population et en la mesure des erreurs de cette extrapolation.

Quelques termes sont importants à définir avant de développer cette partie. Un **estimateur** est défini, de manière succincte, comme une statistique observable sur un échantillon de taille n, comme la moyenne, la proportion ou la variance, que l'on cherchera à généraliser à un paramètre inconnu d'une population de taille N, avec n étant inférieur à N. L'estimation, quant à elle, consiste en la valeur prise par l'estimateur

pour un échantillon de taille n donné. Les estimateurs doivent posséder les deux propriétés suivantes, à savoir l'**absence de biais** et la **convergence**. Le biais décrit l'erreur intrinsèque produite par l'estimateur. On cherchera à utiliser des estimateurs qui ne génèrent pas de telles erreurs. La convergence exprime le fait que lorsque l'échantillon de taille n est suffisamment grand, alors l'estimation doit converger vers le paramètre réel de la population, avec un intervalle de confiance qui tend vers zéro à mesure que n tend vers N. Un estimateur est d'autant plus précis que l'erreur entre les observations et la prévision qu'il génère est faible. La somme des carrés des résidus, ou SCR, mesure l'écart entre les valeurs prises par une variable expliquée et les valeurs estimées. Les résidus sont donc les écarts que le modèle ne peut expliquer. La SCR se résume par la formule suivante, avec y_i la valeur prise par la variable expliquée et $\hat{y}_i$ la valeur estimée par le modèle :

Formule 1 $$SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

(BAUDOT 2020)

La moyenne arithmétique de la somme des carrés des résidus, aussi appelée moyenne des carrés des erreurs, MCE, est une notion importante puisque c'est celle qui doit être minimisée dans les calculs de régression. Il faut comprendre que, bien qu'elle dépende de la SCR, la MCE intègre la notion de degrés de liberté. On divise la SCR par $\mathrm{n} - \mathrm{k} - 1$, avec k correspondant au nombre de variables explicatives, sauf dans le cadre des séries chronologiques, où le dénominateur est égal à n. La MCE est aussi connue sous le nom d'erreur quadratique moyenne et sa définition formelle est donnée ci-dessous :

Formule 2 $$MCE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)]$$

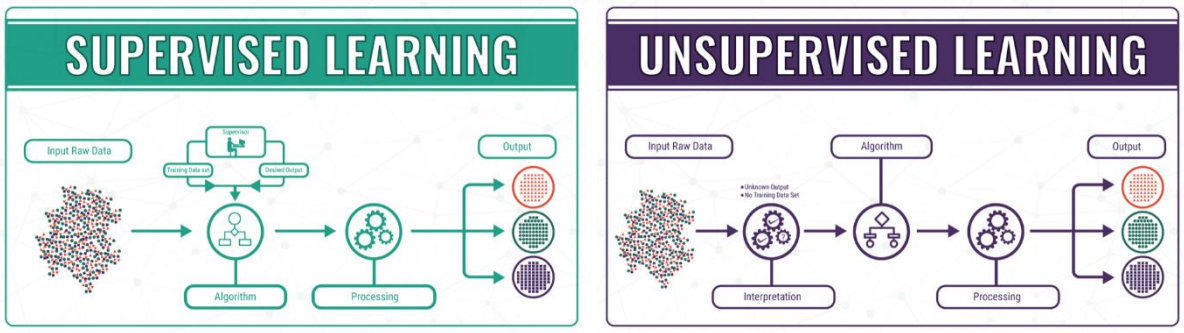
(Wikipédia 2020)

Les outils statistiques prédictifs décrits dans la partie qui suit se basent sur ces notions élémentaires de la régression et ce travail citera les principales techniques de régression à disposition des *data scientists*.

3.1.1.2 Algorithmes de la data science

Afin de distinguer les algorithmes, on se réfère souvent à leurs usages destinés à résoudre des problématiques courantes de traitement des données (BIERNAT et LUTZ 2016, p. 15). Les principales distinctions sont faites quant au mode d'apprentissage de l'algorithme, supervisé ou non supervisé, ainsi qu'au type de traitement que l'algorithme applique, à savoir la régression et la classification. Les algorithmes supervisés sont basés sur la notion de couples entrée-sortie, connus préalablement au traitement. On cherchera dans le cas des algorithmes supervisés à prédire les valeurs de sortie de nouvelles données grâce à une fonction de prédiction que l'algorithme aura déterminé sur la base de données d'entraînement. Dans le cas des algorithmes non supervisés, l'algorithme ne sera alimenté que par les données d'entrée et devra déterminer seul la meilleure répartition sous-jacente des données qu'il traite.

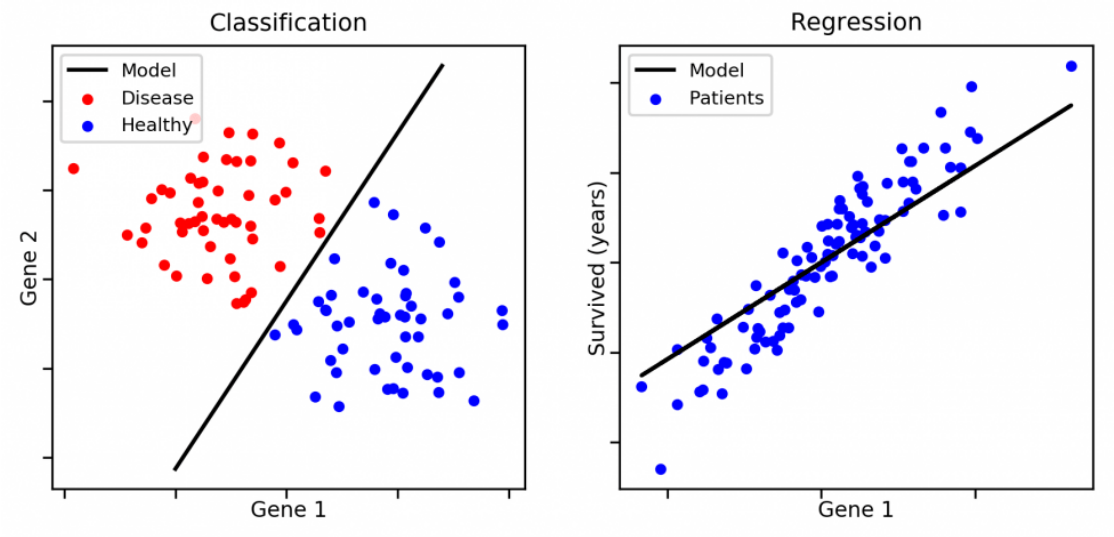
Figure 3-1 : Algorithmes supervisés et non supervisés



(THAKUR 2018)

Concernant les distinctions de traitement, dans le cadre des algorithmes supervisés, ce sont les valeurs de sortie qui diffèrent. La régression permet à l'algorithme d'assigner aux données une valeur de sortie dans l'ensemble continu des réels ($\mathbb{R}$), tandis que la classification a pour but d'assigner ces données à des classes prédéfinies dans le modèle.

Figure 3-2 : Algorithmes - Classification et Régression



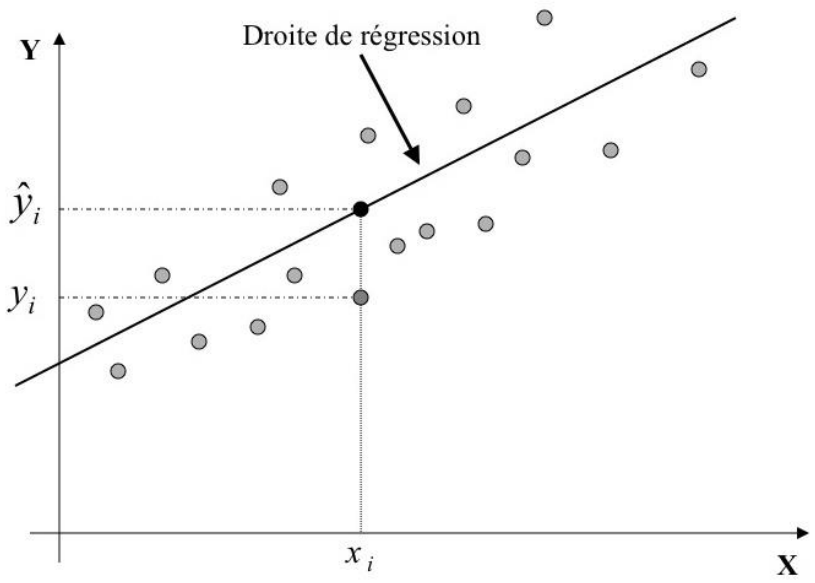
(Le DataScientist 2019)

Ce travail s'intéressera particulièrement aux algorithmes de régression, dont le pouvoir prédictif recèle un intérêt particulier pour l'étude du marché immobilier, notamment en ce qui concerne l'estimation et l'évolution des prix des biens immobiliers.

L'algorithme le plus simple pour appréhender les outils à disposition des data scientists consiste en un modèle de **régression linéaire univariée** héritée des statistiques classiques. Ce modèle vise à trouver dans un jeu de données la droite de régression qui évalue au mieux les données de sortie, en ne se basant que sur une seule et unique variable explicative, ou degré de liberté. Nommons X cette variable explicative, soit la donnée de notre problème. Elle pourrait correspondre au nombre de mètre-carré d'un appartement par exemple dans le contexte du marché immobilier. La valeur de sortie, qui est recherchée, appelons-la Y, correspondrait par exemple au loyer locatif dudit appartement. On définit une fonction hypothèse qui vise à approximer le plus exactement possible cette valeur Y se basant sur X. On modélise cette fonction hypothèse par $h(X) = \theta_0 + \theta_1 X$ (BIERNAT et LUTZ 2016, p. 36). Notre base de données contient des couples de données (x, y) qui serviront à déterminer le couple (θ_0, θ_1) tel que $h(X)$ approxime au mieux notre jeu de données. On devine que l'approximation comporte intrinsèquement une erreur, calculée par la somme des carrés des résidus SCR décrite précédemment. Pour chaque couple (x, y), autrement dit pour chaque appartement de l'échantillon de test, la distance entre ce point et la droite au point x_i représente l'erreur d'appréciation du modèle. La SCR calcule la somme des carrés de cette distance pour chaque point.

La figure ci-dessous permettra au lecteur de mieux se représenter graphique les concepts développés plus haut, avec $(y_i - \hat{y}_i)$ les résidus, ou erreurs d'appréciation du modèle de la formule 1 de la SCR en page 25. Par la suite, on prendra $\hat{y}_i = h(x_i)$.

Figure 3-3 : Régression linéaire univariée



(OpenClassrooms 2020)

La fonction de coût, qui représente cette SCR dans les modèles de régression linéaire algorithmique, est définie par :

Formule 3 $$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h(x_i) - y_i)^2$$

En remplaçant $h(x_i) = \theta_0 + \theta_1 x_i$, on obtient la formule suivante :

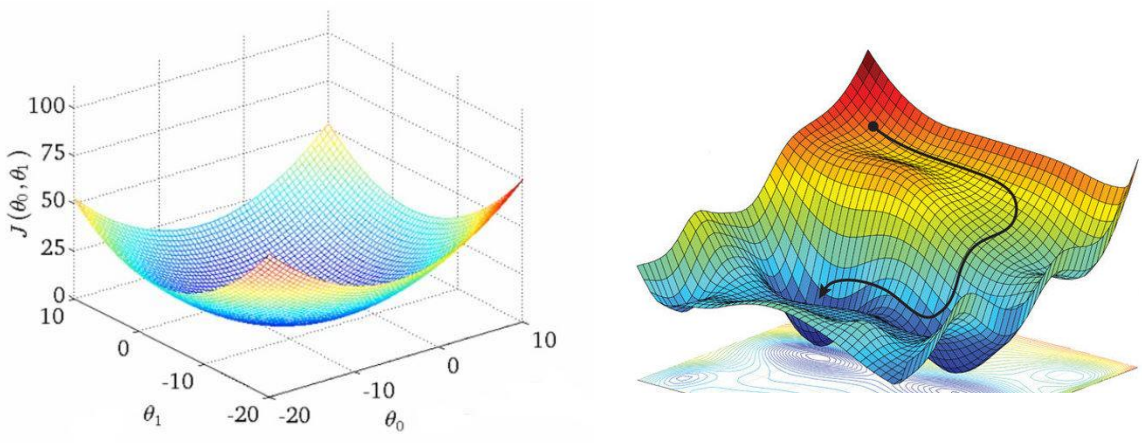
Formule 4 $$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(\theta_0 + \theta_1 x_i - y_i)^2$$

(BIERNAT et LUTZ 2016, p. 37)

avec m le nombre de couples (x, y) observés, ou d'appartements dont on connaît la surface et le prix locatif. La multiplication par 2 du dénominateur n'a pour objectif que de simplifier le calcul de la dérivée de l'étape suivante. Il faut remarquer que cette fonction J ne dépend que de $(\theta_{0,}\theta_1)$. Ces paramètres définissent respectivement l'ordonnée à l'origine et la pente de la droite de régression. On cherchera donc à minimiser la fonction de coût J afin d'obtenir la droite qui minimise la distance entre nos observations y_i et nos prévisions $h(x_i)$. Pour ce faire, il faudra déterminer pour quel couple $(\theta_{0,}\theta_1)$ la fonction de coût atteint son minimum.

La minimisation de la fonction de coût J s'obtient par la méthode de la descente de gradient. Afin de visualiser cette méthode itérative, la figure suivante modélise la fonction J dans l'espace $(\theta_{0,}\theta_1)$ pour mettre en évidence une caractéristique particulière de celle-ci, à savoir qu'elle est convexe.

Figure 3-4 : Descente de gradient



(BENZAKI 2017 à gauche et AMINI 2018 à droite)

Cette méthode est intuitive et relativement aisée à se représenter après l'observation du graphique ci-dessus. Si une balle était lâchée du haut de l'arête d'un bol, quelle trajectoire adopterait-elle ? Elle suivrait la pente la plus forte jusqu'à s'immobiliser à l'endroit le plus bas du bol (BIERNAT et LUTZ 2016, p. 39). On choisit pour cela un couple aléatoire de valeurs $(\theta_{0,}\theta_1)$ qui initialiseront la descente de gradient. L'algorithme suivant, qui devra être itéré jusqu'à convergence, permettra de calculer le couple $(\theta_{0,}\theta_1)$ tel que la fonction de coût atteint son minimum. On rappellera que pour une fonction convexe le minimum local est un minimum global et que la convergence devrait être observée lorsque les dérivées partielles $\frac{\delta}{\delta\theta_j}J$, avec $j\epsilon(0,1)$ se rapprochent de zéro.

Descente de gradient $$\theta_j \coloneqq \theta_j - \alpha\frac{\delta}{\delta\theta_j}J(\theta_0,\theta_1)$$

(BIERNAT et LUTZ 2016, p. 39)

La vitesse de convergence dépendra notamment du choix d'initialisation qui pourra réduire le nombre d'itérations nécessaires avant d'atteindre la convergence, mais aussi du paramètre α. Ce paramètre, appelé *learning rate*, représente le pas entre deux itérations. Il faudra le déterminer de telle sorte qu'il ne soit pas trop important, afin de ne pas risquer de dépasser le minimum de la fonction de coût, voire de diverger (BIERNAT et LUTZ 2016, p. 39), ou trop petit, au risque d'allonger la durée de

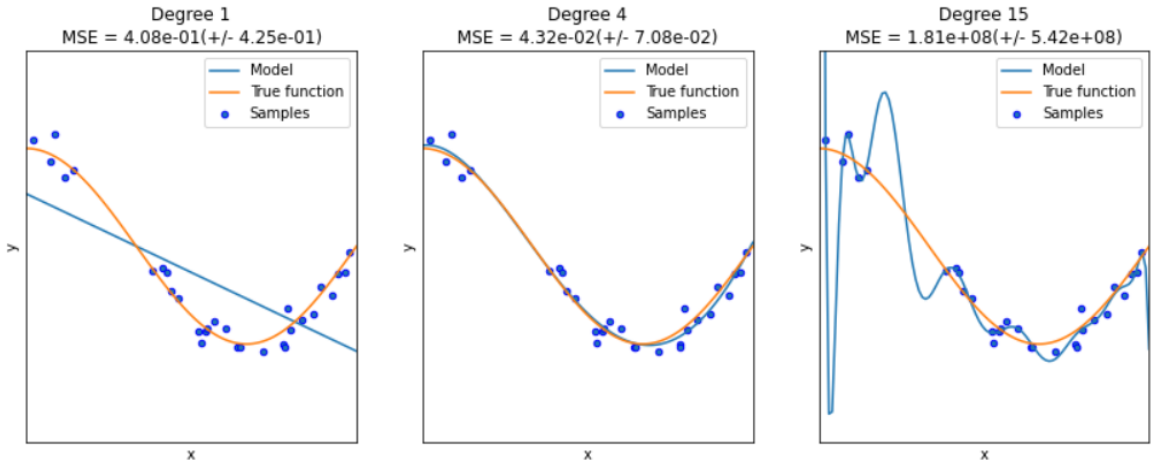
convergence de l'algorithme. Cette méthode est utilisée pour d'autres types de modèles, dont la régression polynomiale, logistique, ou le *Support Vector Machine*, ou SVM, en faisant un incontournable du *machine learning*.

Après avoir vu le fonctionnement d'un modèle de régression linéaire univariée, les modèles suivants, sans les décrire en détail, seront brièvement cités pour évoquer leurs particularités. La régression linéaire multivariée se base sur des données d'entrée multiples, contrairement au modèle univarié où une seule donnée d'entrée était utilisée. Le nombre de degrés de liberté de la fonction hypothèse augmente, permettant une meilleure approximation des données d'entrée, avec un modèle plus fiable. Le principal problème qui pourrait se poser ici pour la descente de gradient consiste en la différence d'échelle existante entre les données d'entrée. Si cette différence est très importante, la convergence de l'algorithme du gradient sera freinée. Pour contourner le problème, il suffit de normaliser les données d'entrée, en d'autres termes, de les mettre à la même échelle, par centrage et réduction par exemple comme c'est le cas en statistiques classiques (BIERNAT et LUTZ 2016, p. 41-46).

Tout comme la régression linéaire multivariée, la régression polynomiale se base sur de multiples données d'entrée. La principale distinction qu'elle possède avec la précédente consiste en sa propriété d'introduire de la non-linéarité dans le modèle de régression. Un polynôme est défini par l'expression suivante : $\sum_{k=0}^{n} a^k$ Le polynôme de degré k égal à 1 correspond au modèle linéaire. La présentation de ce modèle, bien que similaire aux précédents, permet d'introduire deux notions fondamentales des algorithmes de data science, le surapprentissage et le compromis biais-variance. Le surapprentissage définit l'erreur que commet un modèle en s'adaptant de manière trop spécifique aux données d'entraînement. La généralisation d'un tel modèle, du fait que sa stabilité et son évaluation sont mauvaises, n'est pas recommandée (BIERNAT et LUTZ 2016, p. 55-57). Tous les modèles sont concernés par cette erreur et il est nécessaire de s'assurer qu'un modèle ne s'ajuste pas trop aux données d'entrainement. C'est le dilemme du compromis biais-variance. Les procédures de validation croisée permettent de régler les paramètres du modèle de telle sorte que le bon compromis soit trouvé entre la complexité du modèle, ou du degré du polynôme, et sa capacité prédictive (BIERNAT et LUTZ 2016, p. 58-59).

La figure suivante illustre ce problème en fonction du degré de complexité du polynôme, avec un modèle sous-ajusté à gauche à biais élevé et variance faible, un modèle correct au centre et un modèle sur-ajusté à droite à biais faible mais à variance élevée.

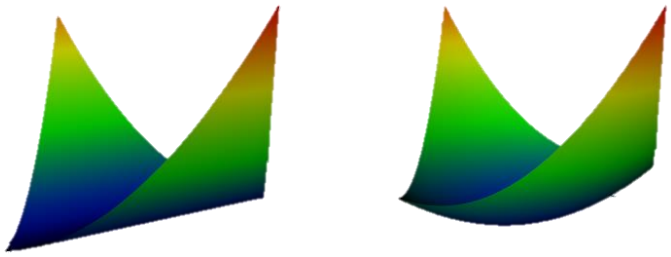
Figure 3-5 : Surapprentissage et compromis biais-variance



(TRIPATHI, Mayank, 2020)

L'un des inconvénients de la régression linéaire est son absence de contrainte sur les paramètres du modèle. Dès lors, la topologie de l'espace de solution de la fonction de coût J peut être soumis à la formation de vallées ou d'arêtes, comme ci-dessous dans la figure de gauche, qui rendent difficiles la détermination du minimum et peuvent aboutir à des modèles très différents (BIERNAT et LUTZ 2016, p. 61-62). La régression régularisée pallie ce problème, tel qu'on peut le constater dans la figure de droite ci-après.

Figure 3-6 : Régression linéaire et régression ridge



(GLEN_B, StackExchange, 2015)

La régression régularisée a pour but d'empêcher cette instabilité. Elle permet de modifier l'espace de solution en ajoutant à la fonction de coût J une fonction de pénalité $P(\lambda, \Theta)$, avec λ un paramètre positif déterminé de manière empirique, dont voici la formule générale :

Formule 5 $$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h(x_i) - y_i)^2 + P(\lambda, \Theta)$$

(BIERNAT et LUTZ 2016, p. 61)

Les méthodes de régression *ridge*, *LASSO* et *ElasticNet* sont construites sur ce modèle. Pour la régression *ridge*, la fonction de pénalité $P(\lambda, \Theta)$ se base sur la notion de distance euclidienne, que l'on élève au carré, donnant pour le vecteur des paramètres la pénalité suivante :

Formule 6 $$P(\lambda, \Theta)_{ridge} = \lambda \sum_{j=1}^{n} \Theta_j^2$$

(BIERNAT et LUTZ 2016, p. 63)

La minimisation de la fonction de coût $J(\theta)_{ridge}$ est bien pénalisée par la formule 6, étant donné que la SCR est augmentée par cette fonction de pénalité, avec λ positif et n le nombre de paramètres du modèle, qui correspond au nombre de variables explicatives.

Toutefois, la régression *ridge* comporte deux inconvénients majeurs. La présence de forte corrélation entre les variables du modèle induira des coefficients θ_j très proches les uns des autres, péjorant la sélection des variables appropriée. De plus, aucune pénalité n'est appliquée aux variables qui nuisent à la capacité prédictive du modèle (KHAROUBI 2016, p. 17).

La régression *LASSO*, acronyme de Least Absolute Shrinkage and Selection Operator, se base quant à elle sur la distance de Manhattan. Conceptuellement, la formulation de la distance pour un espace de dimension n est donnée par $(|x_1|^p + \cdots + |x_n|^p)^{1/p}, \forall\, p$, avec le cas $p = 2$, qui correspond à la distance euclidienne vue précédemment. Pour le *LASSO*, on a $p = 1$, donnant l'expression suivante pour la fonction de pénalité :

Formule 7 $$P(\lambda, \Theta)_{LASSO} = \lambda \sum_{j=1}^{n} |\Theta_j|$$

(BIERNAT et LUTZ 2016, p. 64)

Le principal avantage de la méthode *LASSO* consiste à réduire à zéro les variables qui nuisent au modèle, conservant ainsi celles qui y contribuent le plus fortement. Elle a cependant l'inconvénient de ne pas pouvoir conserver plus de prédicteurs qu'il y a d'observations dans le modèle et de pénaliser les prédicteurs trop fortement corrélés entre eux, en n'en retenant qu'un seul (KHAROUBI 2016, p. 21).

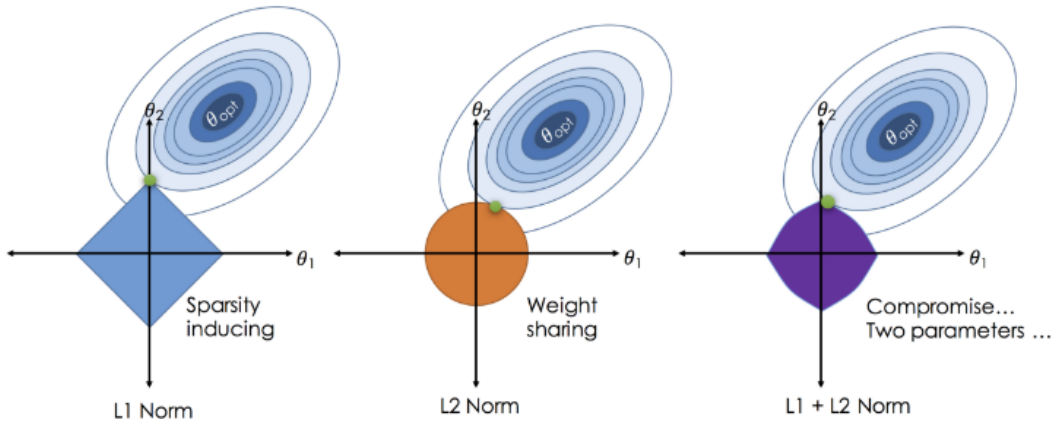
La méthode *ElasticNet* comporte les avantages des deux méthodes précédentes. Bien que la régression *ridge* ne permette pas la réduction du nombre de variables, elle demeure plus efficace dans le cas de variables corrélées entre elles. La régularisation suivante permet de solutionner cette problématique, en fusionnant le *ridge* et le *LASSO*. Voici ce que devient la fonction de pénalité, dont le théorème a été strictement démontré par Zou et Hastie, dans leur article intitulé "Regularization and variable selection via the elastic net" (ZOU et HASTIE 2005, p. 307-309):

Formule 8 $$P(\lambda,\Theta)_{ElasticNet} = \lambda \sum_{j=1}^{n} [\frac{1}{2}(1-\alpha)\Theta_j^2 + \alpha\,|\Theta_j|]$$

(BIERNAT et LUTZ 2016, p. 65)

Le paramètre α , borné par 0 et 1, a pour fonction de définir l'équilibre entre le *ridge*, obtenu pour une valeur de $\alpha = 0$ et le *LASSO*, obtenue pour $\alpha = 1$. La figure suivante permet de visualiser la manière dont est modifiée la topologie de l'espace de solutions de la fonction de coût en fonction des pénalités appliquées par les modèles de régularisation vu ci-dessus, avec le *LASSO* à gauche avec une forme caractéristique de losange, le *ridge* au centre en forme de rond, et l'*ElasticNet* à droite, qui combine les deux formes précédentes.

Figure 3-7 : Lasso, Ridge, ElasticNet



(SNEIDERMAN, Robby, 2020)

Contrairement à l'idée reçue qui veut que plus un modèle est complexe, plus il permet d'obtenir de bonnes prédictions, les méthodes ensemblistes sont fondées sur l'idée que c'est le nombre d'estimateurs, même s'ils sont de moindre qualité, qui, assemblés, vont permettre d'obtenir une bonne estimation. L'algorithme *random forest* repose sur le concept des arbres décision, qui sont les outils algorithmiques parmi les plus utilisés en *data mining*. D'un point de vue conceptuel, un arbre de décision cherche à scinder la population en classes prédéfinies, chaque étape de séparation, appelée *nœud*, étant réalisée en choisissant la variable qui sépare le mieux les individus. À la dernière étape de l'arbre, appelée *feuille*, la proportion des individus appartenant spécifiquement à l'une des classes prédéfinies est par construction plus importante, l'individu devant respecter l'ensemble des règles pour y parvenir (TUFFÉRY et SAPORTA 2017, p. 649). Le *random forest* s'appuie sur quelques particularités des arbres de décision, à savoir que leur performance dépend fortement de l'échantillon de départ et que l'ajout de nouvelles observations peut grandement modifier le modèle (BIERNAT et LUTZ 2016, p. 102). La force du *random forest* est d'utiliser cette propriété des arbres en les démultipliant, créant ainsi une forêt d'arbres décisionnels, chacun basés sur un tirage aléatoire des observations, dont l'assemblage, que l'on nomme *tree bagging*, permet d'échantillonner une première fois le problème. Un deuxième échantillonnage, le *feature sampling*, est réalisé sur les variables du problème. La variance d'un ensemble de B variables indépendantes et identiquement distribuées, soit nos arbres de décision, de variance σ^2, est donnée par :

Formule 9 $$V_{forest} = \rho\sigma^2 + \frac{1-\rho}{\mathrm{B}}\sigma^2$$

(BIERNAT et LUTZ 2016, p. 104)

On constate que pour réduire la variance de l'ensemble, le nombre d'arbres B permet d'affaiblir considérablement le deuxième terme, par *bagging*, sachant que pour le premier terme, c'est le *feature sampling* qui permettra de réduire la corrélation entre les arbres et donc de réduire la variance globale.

Sans aller plus loin dans le fonctionnement de l'algorithme *random forest*, dont on comprend le fonctionnement global, on mentionnera cependant que de nombreux paramètres sont à définir pour le paramétrer correctement, allant du nombre d'arbres, au critère de séparation, jusqu'au nombre de cœurs de CPU (BIERNAT et LUTZ 2016, p. 109). Les résultats du *random forest* dépendent du traitement et peuvent être une moyenne s'il s'agit de régression ou un vote s'il s'agit de classification.

Pour finir sur la partie des algorithmes, ce travail évoquera le *gradient boosting*, méthode ensembliste non linéaire dont le fonctionnement s'inspire du *boosting*.

Le concept utilisé par le boosting est de transmettre l'information de l'arbre de niveau k à son successeur de niveau k+1. Ainsi, pour *random forest*, qui construisait une forêt d'arbres en parallèle, le *boosting* crée ces arbres en série, l'assemblage des arbres se fait sur la base d'une somme pondérée. Le coefficient de pondération récompense lors de l'agrégation les arbres dont l'erreur était la plus faible.

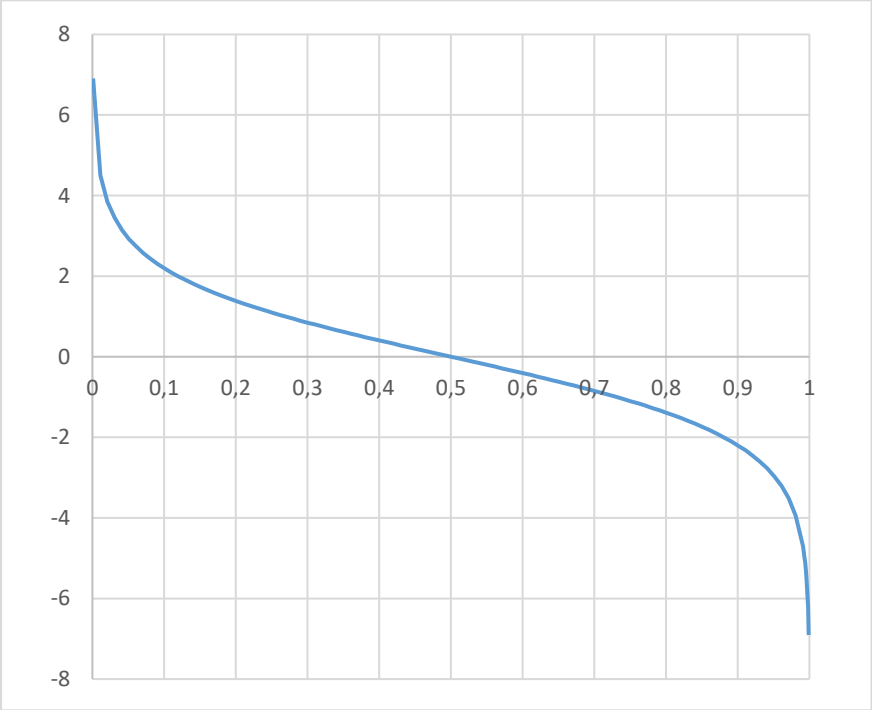
Ainsi, nous avons les formules suivantes qui serviront à comprendre comment fonctionne le *boosting*. Sachant que X est une matrice $m * n$ d'observations et H_i correspond aux arbres, avec $i \epsilon [1, B]$, chacun procédant à un traitement noté $h_i(x)$, l'algorithme aura pour résultat :

Formule 10 $$\mathrm{H(X)} = \mathrm{signe}\left(\sum_{i=1}^{B} \alpha_i \, h_i(X)\right)$$

(BIERNAT et LUTZ 2016, p. 104)

avec $\alpha_i = \frac{1}{2} \ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$ les coefficients de pondération, dont on voit ci-dessous l'évolution en fonction de la valeur prise par les erreurs ε_i :

Figure 3-8 : Fonction des coefficients de pondération du *boosting*



On constate que la forme exponentielle prise par la fonction lorsque les erreurs tendent vers 0 et 1 pondère, de manière très marquée, la contribution du traitement $h_i(x)$ à

l'ensemble. Plus l'erreur est faible, plus la contribution de $h_i(x)$ au modèle final sera importante, et inversement.

L'algorithme *gradient boosting* utilise les principes du *boosting* en les combinant avec la descente de gradient dont le principe a été décrit précédemment. Dans le cas du *gradient boosting*, on cherchera à approximer les valeurs de sortie par un algorithme final H, construit itérativement sur la base de B fonctions hypothèse h. Le concept de la descente de gradient consiste ici pour chaque itération de l'algorithme à minimiser l'erreur produite par l'itération précédente. Si $h_1$va réaliser un premier apprentissage des données, au terme de la première itération, on obtient $\mathrm{H} = h_1$. La seconde itération h_2 cherchera à améliorer H, tel que : $\forall\, i\epsilon\{1, \dots, m\}$, avec m le nombre d'observations, on a $h_2(x_i) \approx y_i - H(x_i)$, dont le terme de droite correspond aux résidus de l'étape précédente. Le *gradient boosting* cherchera ici non plus à surpondérer ces résidus, mais à minimiser la fonction de coût $J = \sum_{i=1}^{m} j(y_i, H(x_i))$, avec $j(y_i, H(x_i)) = \frac{(y_i - H(x_i))^2}{2m}$ le gradient négatif utilisé pour l'algorithme de descente, obtenu ici par la méthode des moindres carrés (BIERNAT et LUTZ 2016, p. 122-123).

Ainsi, nous obtenons la définition suivante pour l'algorithme final :

Formule 11 $$H(x_i) \coloneqq H(x_i) - \frac{\delta J}{\delta H(x_i)}, \forall\, i\epsilon\{1, \dots, m\}$$

(BIERNAT et LUTZ 2016, p. 123)

L'avantage consiste à pouvoir utiliser diverses fonctions de coût, parmi lesquelles ont été vues précédemment la distance euclidienne, la distance de Manhattan, ou encore la fonction Huber, combinaison des deux précédentes, ainsi que la quantile. Ce sont principalement les points atypiques de notre jeu de données initial, ou *outliers*, qui conditionneront le choix de la fonction de coût, qui doit être choisie « en fonction du critère sur lequel sera évalué en production votre algorithme » (BIERNAT et LUTZ 2016, p. 123).

Ceci conclut la partie concernant les algorithmes de la *data science* que ce travail n'a pas cherché à décrire exhaustivement, mais à permettre de comprendre l'intrication importante qui existent entre les méthodes statistiques classiques et les méthodes algorithmiques utilisées pour résoudre des problématiques courantes des *data scientits*. D'autres algorithmes tels que le *Naive Bayes*, la régression logistique ou le *Support Vector Machine* sont des outils nécessaires à la maîtrise des analyses de données et le lecteur est invité à se référer aux ouvrages mentionnés dans la bibliographie en cas d'intérêt pour ces méthodes.

3.1.2 Python et la data science

Python, créé par le Néerlandais Guido van Rossum à ses heures perdues, est un langage de programmation devenu un incontournable de la *data science* grâce à ses nombreuses librairies de fonctions et ses quatre styles de programmation. S'inspirant du langage ABC ; Guido van Rossum écrit en décembre 1989 sa première version du langage Python, dont la première publication date du 20 février 1991 (Rossum 2009).

Parmi les caractéristiques qui rendent son étude intéressante pour des projets de data science, on citera notamment sa philosophie open source, son processus d'interprétation qui implique qu'il ne doit pas être compilé pour être exécuté, sa syntaxe lisible et claire, avec l'indentation comme identification des blocs, ou encore sa popularité, qui le rendent compatible avec de nombreuses plateformes. D'autre part, étant un langage de haut niveau, Python permet de concentrer l'attention du programmeur sur les résultats plutôt que sur l'aspect technique du programme. Il est le troisième langage le plus utilisé selon l'index TIOBE qui est un indicateur de popularité des langages de programmation. Il est même élu pour la quatrième fois le langage de l'année par le site, ce qui est un record.

Figure 3-9 : Index TIOBE

Jan 2021	Jan 2020	Change	Programming Language	Ratings	Change
1	2	^	C	17.38%	+1.61%
2	1	v	Java	11.96%	-4.93%
3	3		Python	11.72%	+2.01%
4	4		C++	7.56%	+1.99%
5	5		C#	3.95%	-1.40%
6	6		Visual Basic	3.84%	-1.44%
7	7		JavaScript	2.20%	-0.25%
8	8		PHP	1.99%	-0.41%
9	18	^^	R	1.90%	+1.10%
10	23	^^	Groovy	1.84%	+1.23%
11	15	^^	Assembly language	1.64%	+0.76%
12	10	v	SQL	1.61%	+0.10%
13	9	vv	Swift	1.43%	-0.36%
14	14		Go	1.41%	+0.51%
15	11	vv	Ruby	1.30%	+0.24%
16	20	^^	MATLAB	1.15%	+0.41%
17	19	^	Perl	1.02%	+0.27%
18	13	vv	Objective-C	1.00%	+0.07%
19	12	vv	Delphi/Object Pascal	0.79%	-0.20%
20	16	vv	Classic Visual Basic	0.79%	-0.04%

(index | TIOBE 2021)

Dans cette partie, ce travail s'intéressera en particulier à décrire les principales fonctionnalités du langage et d'expliquer les raisons pour lesquelles Python se révèle être un langage de premier choix pour débuter avec la *data science*.

3.1.2.1 Les styles de programmation en Python

Le langage Python permet d'adopter des styles de programmation différents en fonction du but recherché par le programmeur.

Voici les 4 styles de programmation que permet d'adopter Python, dont on développera les spécificités :

- la programmation fonctionnelle
- la programmation impérative
- la programmation orientée objet
- la programmation procédurale

La programmation fonctionnelle désigne un paradigme de programmation déclaratif selon lequel toute affectation aux valeurs des données est interdit. Ces effets de bord, comme ils sont appelés, ou effets secondaires, sont évités parce que chaque instruction du code est une fonction qui peut posséder plusieurs paramètres en entrée mais une seule valeur de sortie (MUELLER, MASSARON et ENGLER 2019, p. 12).

La programmation impérative désigne le paradigme selon lequel les instructions exécutées modifient l'état du programme. C'est ce style de programmation qui est le plus répandu et se différencie par définition du paradigme précédent.

La programmation orientée objet désigne le paradigme qui consiste à conceptualiser les données comme des objets, ou plus exactement, à affecter ces données à des objets dont on aura précédemment défini la structure, le comportement et les relations avec les autres objets. Ne permettant pas l'encapsulation des données, soit le processus qui masque les attributs et les méthodes de ces objets, Python ne supporte pas complètement cette approche (MUELLER, MASSARON et ENGLER 2019, p. 12).

Pour finir, la programmation procédurale décrit le paradigme selon lequel les instructions du code sont incorporées dans des fonctions, réutilisables à différents endroits du code.

Compte tenu de cette polyvalence, Python est un langage d'une grande souplesse qui permet de réaliser des tâches variées et complexes.

3.1.2.2 Les librairies en Python

Cette partie est consacrée aux diverses librairies fréquemment utilisées dans le langage Python. Sans être exhaustive, cette liste permet néanmoins au lecteur de se faire une bonne idée des possibilités offertes par Python dans le cadre de la *data science*.

Une librairie, ou bibliothèque, comme son nom le laisse deviner, est une collection de fonctions qui sont utilisables en l'état, ce qui signifie que le programmeur pourra les utiliser sans avoir à les réécrire. Elles sont importées dans Python dans la première partie du script afin de pouvoir être utilisées.

Parmi les librairies utilisées pour la *data science*, on citera en premier lieu le recueil de librairies SciPy, parmi lesquelles on trouvera NumPy, une librairie qui permet de manipuler des tableaux et d'effectuer des calculs d'algèbre linéaire. Cette librairie sert donc à effectuer des calculs fondamentaux sur les données du programme et de ce fait, son utilisation est largement répandue dans le domaine (MUELLER, MASSARON et ENGLER 2019, p. 33). Pour la visualisation des données, la librairie matplotlib s'inspire de l'outil MATLAB afin de produire un rendu graphique facilitant la représentation des données, lui donnant une place de premier plan dans le bagage des *data scientists*.

La librairie pandas sert quant à elle à l'analyse des données dans Python, avec des possibilités d'analyse similaires à celles offertes par des langages spécifiquement conçus comme R (MUELLER et al. 2019, p. 33).

Pour tous les algorithmes vus dans la partie précédente, utilisera la librairie Scikit-learn, dont les propriétés pour le minage et l'analyse de données en font un outil très puissant pour le *machine learning* (MUELLER et al. 2019, p. 33).

Pour conclure sur cette partie dédiée aux librairies communément utilisées, ce travail s'intéressera au scraping de données sur le web. Cette technique consiste à récolter sur des pages web des informations que l'on souhaite utiliser dans un programme sans devoir effectuer cette prise d'informations de manière manuelle. La librairie dédiée Beautiful Soup permet l'analyse de données HTML ou XML afin d'utiliser le code source des pages web dans le programme et d'extraire les informations voulues grâce à différentes méthodes. Son intérêt dans le cadre de ce travail est d'une grande importance, étant donné que les données du marché immobilier ne sont pas accessibles de manière simple pour le grand public, même si l'offre est largement

publiée sur des sites internet tels que le site comparis en Suisse, qui fera l'objet du cas pratique dans le chapitre suivant.

3.1.2.3 PriceHubble : la data science en immobilier

La société PriceHubble, fondée en janvier 2016, dont l'objectif est de radicalement améliorer la compréhension et la transparence des marchés immobiliers sur la base d'un raisonnement orienté données. Société suisse, dont le siège est à Zürich, elle utilise les données du Big Data ainsi que toutes les techniques les plus poussées de *machine learning* et d'intelligence artificielle pour modéliser la structure, le fonctionnement et les particularités des marchés immobiliers et de leur environnement, leur faisant gagner en transparence (LinkedIn 2021). De ce fait, elle est devenue en peu de temps un acteur incontournable pour les investisseurs institutionnels, comme Axa, BNP Paribas, SwissLife etc… Solution B2B par essence, la société est présente dans 5 pays à travers le monde, de Paris à Tokyo en passant par Vienne et Berlin. Fondée par le Dr. Stefan Heitmann, connu pour avoir également fondé MoneyPark AG, ancien de McKinsey, et Markus Stadler, mathématicien passé par Bain & Company, la société emploie de nombreux *data scientists* dont les parcours mentionnent, pour leur CTO Mario Ubeda Garcia par exemple, de très nombreuses recommandations pour le langage Python. De nombreux autres profils semblent également avoir comme compétence clé l'utilisation de Python, laissant suggérer que ce programme est utilisé chez PriceHubble pour une partie du processus de *data mining*.

Cette présentation se conclut par la présentation de leur outil de datavisualisation dont on ne peut qu'admirer l'élégance et le pouvoir de représentation.

Figure 3-10 : PriceHubble: datavisualisation

(Actu IA 2019)

3.1.3 Business Intelligence

La Business Intelligence, ou informatique décisionnelle, décrit l'ensemble des outils informatiques et logiciels, permettant d'extraire, de représenter et de comparer l'évolution des données et des indicateurs calculés à partir de celles-ci. La BI, sous son acronyme communément usité, est avant tout dédiée à la prise d'une information éclairée par les décideurs de l'entreprise, avec des programmes tels que Tableau ou QlikView ou PowerBI.

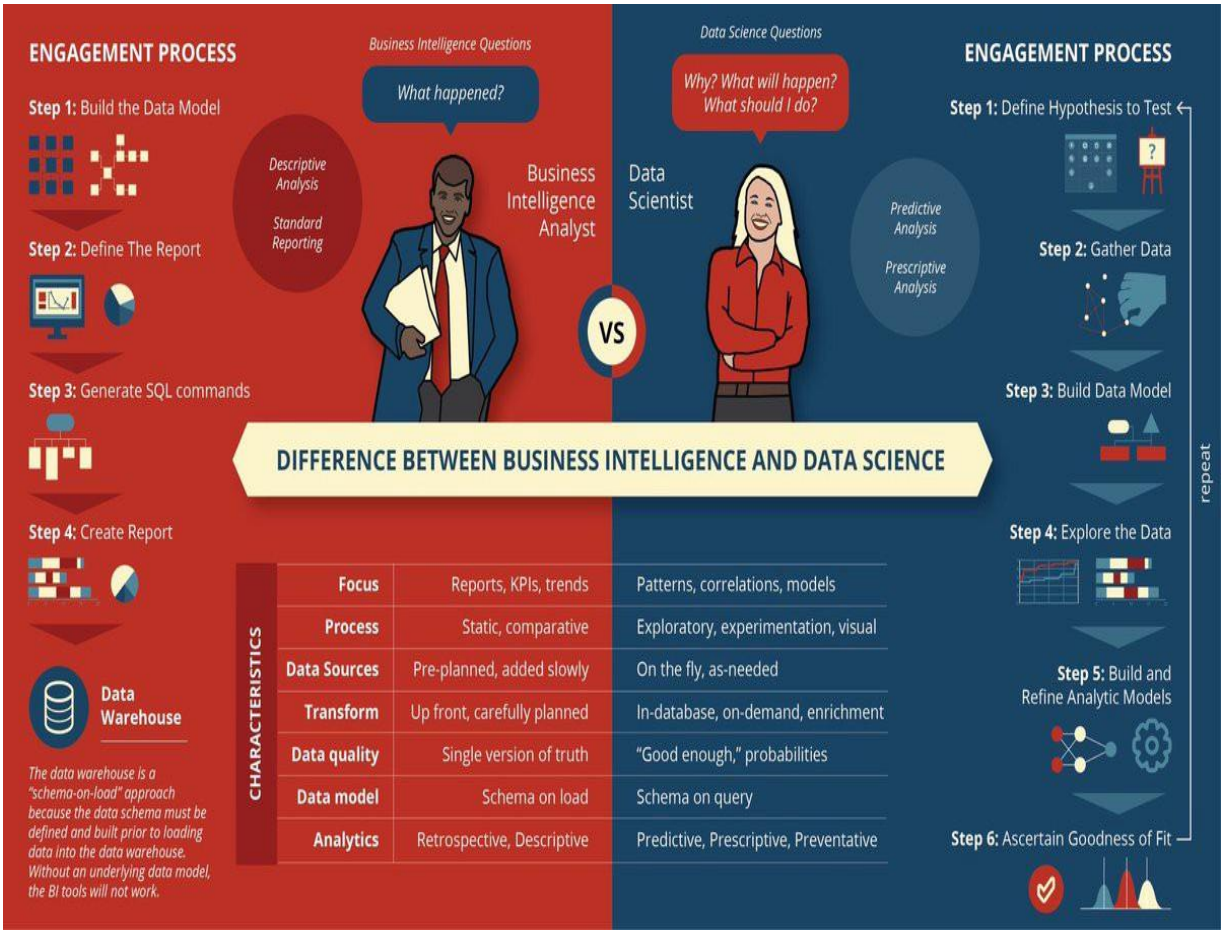
Devenue incontournable avec la diffusion de la veille stratégique, la BI repose sur des données avant tout, mais leur utilisation diffère en cela de la *data science* qu'elles sont utilisées en tant que tel, dans un raisonnement de type déductif. A contrario, en *data science*, c'est le raisonnement inductif qui est employé (Delort 2018, p. 42).

Dès lors, les objectifs poursuivis n'étant pas les mêmes, on ne pourra pas travailler sur les données de façon similaire si l'on cherche à les suivre selon les principes de la BI ou si l'on cherche à les utiliser dans le cadre d'un projet de *data science*.

Autre élément qui nécessite d'être mentionné, la BI produit des tableaux de bord qui s'alimentent sur des entrepôts de données, ou Data Warehouse, des rapports et autres outils d'aide à la gestion, à la prise de décision, à la veille concurrentielle, mais ne permet pas en tant que tel l'analyse du comportement d'achat d'un client sur le web ou la modélisation de l'évaluation du prix d'un bien immobilier sur des critères variables.

La figure ci-dessous résume bien les principales nuances qu'il faut comprendre entre ces deux manière de travailler avec les données.

Figure 3-11 : Business Intelligence et Data Science



(MORENO 2019)

4. Le Data mining

La partie qui suit a pour objectif de démontrer les possibilités qu'offre la *data science* dans le cadre du marché immobilier. L'idée générale qui sous-tend cet objectif se fonde sur le constat suivant : les marchés immobiliers, contrairement à d'autres marchés d'actifs de taille comparable, demeurent relativement peu transparents. Plus encore, le besoin d'intermédiation y est si important que de nombreux courtiers semblent avoir adopté un comportement qui peut nuire à la recherche et à l'échange d'informations. Cette rétention leur permet en effet de facturer des services, dans le cas des « chasseurs » d'appartements, pour assurer à un locataire la conclusion d'un contrat de bail par exemple. Fort de ce constat, ce travail se place dans le paradigme d'un particulier à la recherche d'informations sur l'état du marché locatif immobilier genevois et cherche à répondre à la question encadrant l'étude selon deux axes. Quelles possibilités offrent les technologies de la *data science* en termes d'obtention d'informations et à quel niveau de détail ces informations sont-elles disponibles pour le grand public ? D'autre part, ce travail consistant également à analyser de manière plus technique l'état du marché immobilier, dans un horizon de temps plus court que celui mis à disposition du grand public, le second axe déterminera s'il est possible d'obtenir une source d'informations anticipée d'une qualité suffisante avec quelques outils qu'utilisent les *data scientists*.

Avant de développer les résultats obtenus, la première partie se propose de schématiser le déroulement d'une étude de *data science*, d'en expliquer les points principaux et de contextualiser précisément la manière dont seront extraits, concaténés, et expliqués les résultats de l'étude

4.1 Déroulement d'une étude de data science

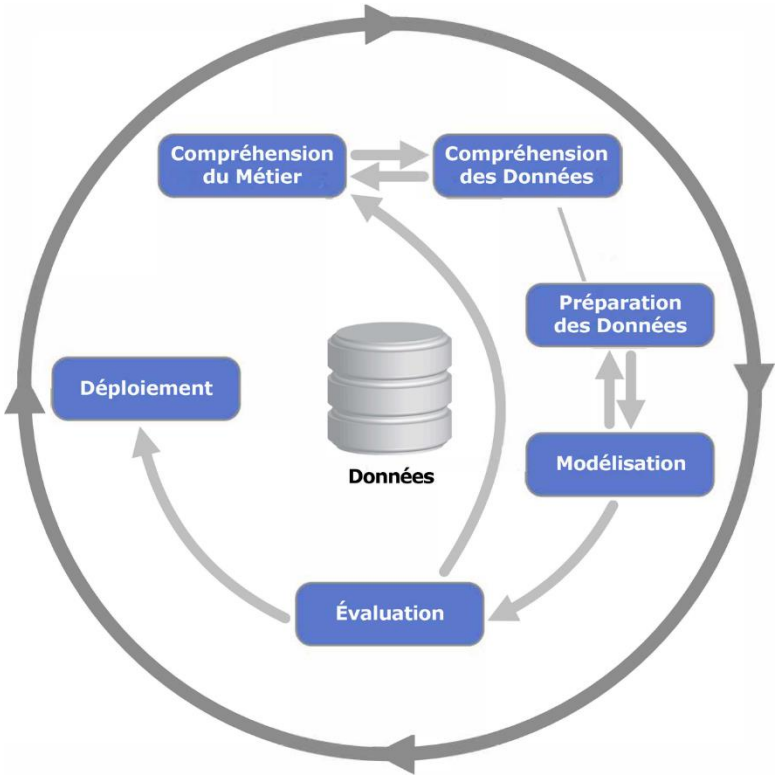
Le processus selon lequel se déroule une étude de *data mining*, ou forage de données, débute non pas avec leur obtention, mais avec la compréhension du problème à résoudre. Avant de commencer tout projet de *data science*, il est primordial de savoir quelle problématique devra être résolue et quels sont les outils les moins coûteux pour la résoudre.

Le *Cross Industry Standard Process for Data* Mining, ou CRISP-DM, propose un processus définit selon 6 phases principales le processus d'une étude. L'étude qui suit se base sur ce modèle et cherche avant tout à poser les bases solides d'une modélisation utilisables au-delà du cadre strict de cette étude. Sachant que l'on ne peut pas tout analyser, il sera nécessaire de restreindre le champ de l'étude aux

questions fondamentales et conserver pour des phases ultérieures toute information susceptible de résoudre des questions subsidiaires du problème.

Afin de visualiser le déroulement du CRISP-DM, la figure suivante illustre sous la forme d'un diagramme récursif les liens entre les différentes phases du processus.

Figure 4-1 : CRISP-DM



(Wikipédia, Cross Industry Standard Process for Data Mining, 2019)

Comme mentionné plus haut, tout problème de *data science* débute par la compréhension du métier, ou du problème à résoudre. L'idée de cette phase est avant tout de comprendre en détail le cas d'usage, ou *use case*, mais aussi de pouvoir revenir sur cette définition à mesure que le problème est redéfini par l'étude des données. En effet, il existe de manière constante dans les projets de *data mining* une assimilation des caractéristiques du problème et de ses données qui redéfinit le problème lui-même, ou du moins sa formulation et les réponses qui lui seront apportées. La phase de compréhension des données ne se résume pas à les trier ou les « nettoyer », il s'agit aussi d'évaluer le coût de leur obtention, la manière de les agréger, mais aussi les caractéristiques des sous-ensembles d'une population que l'on étudie. L'exemple de la détection de fraude parle de lui-même. Les fraudes visant les cartes de crédit sont identifiées par les banques, mais aussi par les détenteurs de carte et la fiabilité d'une telle détection est telle que l'identification par une variable d'une fraude est possible mais aussi concordante. Dans le cadre de la détection de fraude sur le

système d'assurance maladie américain Medicare, cette identification n'est pas aussi aisée. Les fraudes constituent un sous-ensemble de déclarations légitimes effectuées par les médecins ou les patients. Les techniques, bien que le problème semble similaire, ne pourront pas être identiques dans ces deux cas (PROVOST et FAWCETT 2018, p. 37).

Le problème ayant été défini, les données à extraire ayant été identifiées et récoltées, vient la phase de leur préparation qui est un domaine vaste et complexe dont la technicité relève à la fois des statistiques, de l'informatique matérielle et logicielle ainsi que des connaissances du domaine de l'étude. La manière dont sont préparées les données peut avoir un impact déterminant sur l'étude et sur les résultats qui en découlent. Certaines variables prédictives peuvent donner lieu au phénomène de fuite, comme dans le cadre de l'étude d'un comportement d'achat. Si l'on s'intéresse à prédire le panier moyen d'un consommateur, les catégories des articles achetés semblent un excellent indicateur qui a cependant un défaut majeur, appelé fuite, qui ne permet pas de déterminer le comportement d'achat, étant donné que cet indicateur est connu après la prise de décision (PROVOST et FAWCETT 2018, p. 39).

Les trois phases suivantes, la modélisation, l'évaluation et le déploiement, ont pour principal objectif de produire un modèle fiable et utilisable qui génère un retour sur investissement. Le déploiement peut d'ailleurs se révéler parfois beaucoup moins technique que les phases précédentes (PROVOST et FAWCETT 2018, p. 41).

Fort des éléments structurels de l'étude, la partie qui suit se focalisera sur le cas pratique en s'inspirant du modèle CRISP-DM. Les phases de modélisation, d'évaluation et de déploiement ne feront pas l'objet de cette étude, compte tenu du fait que notre problématique ne consiste qu'à déterminer quel niveau d'informations est accessible au grand public et si elles sont utilisables de manière anticipée.

4.2 Cas pratique : Immobilier locatif à Genève

La problématique de cette étude vise avant tout à déterminer à quel niveau d'informations un particulier peut avoir accès et si ces informations pourraient lui servir comme indicateur précoce de tendances sur le marché immobilier.

Dans ce cadre, un élément important est à mentionner. Les transactions effectives entre les acteurs du marché immobilier ne peuvent pas être à disposition du grand public, à moins d'avoir accès aux bases de données des régies immobilières, des banques et autres prestataires de services du secteur. Dès lors, le marché immobilier genevois, au niveau accessible par le grand public, se résume aux sources

d'information secondaires proposées par l'office cantonal de la statistique, aux études et rapports produits par les banques et cabinets de conseil comme Wüest Partner, JLL, ou CBRE, pour n'en citer que quelques-uns, ainsi qu'aux site d'annonces immobilières. Le principal inconvénient des sources secondaires dans le cadre de cette étude se résume à leur délai. Toutes les publications, rapports des administrations et autres analyses ne sont délivrés qu'une fois la période écoulée, dans un délai variant du mois au trimestre. Bien que l'immobilier ait pour caractéristique de ne pas évoluer de manière volatile, son illiquidité rend difficile toute transaction dans un contexte d'instabilité économique et financière. Preuve en est que la crise des *subprimes* aux États-Unis n'a pas été anticipée par les acteurs traditionnels du marché, démontrant que la complexité rend difficile l'anticipation de phénomènes extraordinaires.

Ainsi, seules les annonces immobilières répondent à notre horizon de temps que l'on souhaite le plus court possible. Cette source d'information, disponible en ligne, ne consiste pas en des transactions réelles, mais délivre une information précoce sur l'état de l'offre immobilière à un instant *t* donné. La demande ne saurait pas être modélisée de façon similaire, même s'il existe des études de *nowcasting* de la demande immobilière basées sur les techniques du Big Data, comme vu précédemment avec Google Flu Trends, notamment celles de la Banque de réserve de l'Inde (MITRA, SANYAL et CHOUDHURI 2017). Ce travail n'a cependant pas pour objectif de réaliser une telle étude.

D'autre part, la grande majorité des sites d'annonces en ligne visités semblent ne pas avoir intégré d'outils de *data science* à leurs pages. Il est fait mention ici de manière spécifique et non exhaustive aux données sur la démographie, la pollution, la criminalité, la taxation ou l'accessibilité, qui pourraient renseigner le visiteur dudit site sur l'environnement d'un bien immobilier de manière détaillée et le comparer de manière active en fonction de critères paramétrables. Le monde de l'annonce immobilière en ligne, pourtant déjà actif depuis de nombreuses années, n'a pour l'heure pas encore intégré l'ensemble des possibilités de la *data science*, au-delà de quelques représentations cartographiques basiques et *factsheets* disponibles en téléchargement. Ce travail ne se propose pas de gérer ces paramètres, mais il pourra démontrer que le niveau d'informations disponible en consultation sur ces sites est en-deçà des possibilités réelles sous-jacentes.

4.2.1 Data scraping

Après avoir identifié quelle source de données était la plus susceptible de remplir les critères de l'étude, il s'agit de récolter ces informations depuis une page web. Plusieurs sites ont été consultés à ce titre, afin de permettre d'identifier quel site semblait le plus à même de délivrer l'information la plus large possible. C'est le comparateur en ligne le plus connu de Suisse, Comparis, qui a été retenu à cet effet. Sans détailler la raison de ce choix, et pour n'en citer que la principale, il s'agissait avant tout de réduire l'étude des pages html à explorer afin de ne pas complexifier un travail par ailleurs conséquent. Comparis dispose d'un portail d'accès aux données présentes sur les sites d'annonces immobilières les plus courants. De ce fait, il en faisait un candidat idéal pour le *scraping* de données.

Un élément reste ici à mentionner qui a son importance. Le *scraping* de données sur le web doit être entrepris avec une attention particulière aux législations et aux fins pour lesquelles les données récoltées sont utilisées. En Suisse, plusieurs lois encadrent ces pratiques dont la loi sur la concurrence déloyale. La principale idée qu'il faut conserver à l'esprit se résume à ne pas exploiter le travail d'autrui par des méthodes déloyales et illicites. Un autre élément, qu'il est nécessaire de mentionner, s'articule autour d'éléments techniques du *scraping*. Sa mise en œuvre ne doit pas perturber le site internet qui est visé. On appelle attaques par déni de services des actions, volontaires ou non, résultat en la mise hors ligne d'un serveur. Les infrastructures importantes sauront se protéger contre un nombre élevé de requêtes, qui viendraient submerger les capacités de leurs serveurs. Toutefois, ces attaques peuvent être asymétriques, cette asymétrie étant définie par les ressources nécessaires à de telles attaques. Un site internet, même s'il est supporté par un réseau moderne et de taille suffisante, pourrait être attaqué par un ordinateur unique, le rendant indisponible. Les techniques de *scraping* doivent donc limiter ces intrusions à un niveau acceptable pour ne pas surcharger le trafic sur le site visé.

Dans un premier temps, il faut mentionner que toute récolte d'informations manuelles sur le site n'était pas réalisable. Il existe à chaque instant plusieurs milliers d'annonces uniques concernant des biens immobiliers sur Comparis. Dès lors, l'intérêt qu'un script réalise ces extractions était important.

La première partie du processus consistait à concentrer la zone de recherche des annonces en fonctions du lieu. On ne s'est intéressé qu'aux annonces immobilières répondant aux filtres « Louer » et dont le lieu était « Genève ». Ceci permettait de restreindre le nombre de pages à ce qui était nécessaire pour l'étude du marché

immobilier locatif genevois. Le script se sert de l'url de la page regroupant toutes les annonces afin d'extraire mille numéros d'annonce individuelle répondant aux critères ci-dessus. Les numéros sont stockés dans un fichier csv pour traitement ultérieur par un second script dont l'objectif est de consulter chaque annonce et d'extraire 35 caractéristiques, ou champs, contenus dans l'annonce.

Dans la figure ci-dessous, le script destiné à l'extraction des numéros d'annonce correspondant aux critères « Louer » et « Genève » est mis à disposition du lecteur pour consultation. Ce travail ne vise aucunement à inciter le lecteur à le reproduire.

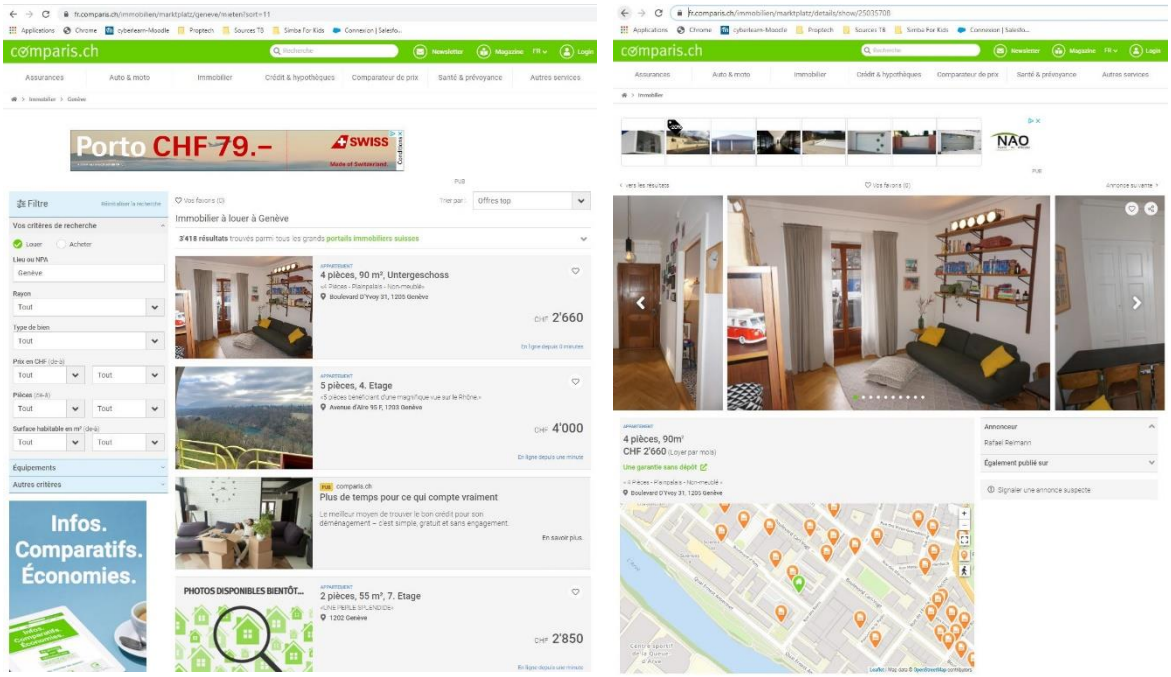
Figure 4-2 : Scraping des numéros d'annonce

```
File  Edit  View  Help   main.py - ...\PycharmProjects\ScrapData
main.py

from urllib.request import urlopen as uReq
from os import getcwd, chdir, mkdir
import urllib.request
import requests
import bs4
import pandas as pd
import re
import csv

"""ATTENTION : Changer la date afin de ne pas écraser les fichiers - Voir ligne 73-74"""


token = 'https://fr.comparis.ch/immobilien/result/list?requestobject=%7B%22DealType%22%3A10%2C%22SiteId%22%3A0%2C%22RootPropertyTypes%22%3A%5B%5D%2C%22PropertyTypes%22%3A%5B%5D%2C%22RoomsFrom%


def get_pages(token, nb):
    pages = []
    for i in range(1, nb + 1):
        j = token + str(i)
        pages.append(j)
    return pages

def get_annonces(token, num_annonces, index):
    pages = []
    for i in range(1, 1000):
        j = token + num_annonces[index]
        index += 1
        pages.append(j)
    return pages


pages = get_pages(token, 1)

""" 5 strategies to write unblock-able web scrapers in Python : https://towardsdatascience.com/5-strategies-to-write-unblock-able-web-scrapers-in-python-5e40c147bdaf """
headers = {
    'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/87.0.4280.88 Safari/537.36',
    }

for i in pages:
    response = requests.get(i, headers=headers)


soup = bs4.BeautifulSoup(response.text, 'html.parser')
for p in soup.find_all('script', {"id": "__NEXT_DATA__"}, {"type": "application/json"}):
    print(p)


results_texte = str(p)

chaine = "initialResultData"


def trouver(resultats, mot):
    index = 0
    while index < len(resultats):
        if mot == resultats[index:index+len(mot)]:
            return index + len(mot) + 12
        index = index + 1
    return -1


def extraction_resultats(index, mystring):
    begin = index
    end = index + 8999
    mystring = mystring[begin:end]
    mystring.replace(":","")
    return mystring.split(",")


indice_resultats = int(trouver(results_texte, chaine))

print(indice_resultats)

tableau_annonces = extraction_resultats(indice_resultats, results_texte)

print(tableau_annonces)

print(getcwd())
chdir('Annonces Geneve')

annonces_du_jour = pd.DataFrame(tableau_annonces, columns=['Numero annonce'])


"""Changer la date afin de ne pas écraser les fichiers"""

annonces_du_jour.to_csv('Annonces_GE_09.01.2021.csv', index=False)
```

Dans un deuxième temps, les fichiers journaliers d'annonce contenant les numéros individuels d'annonces répondant aux critères sont comparés hors script afin de ne conserver que les annonces qui ne figurent pas dans la base de données. Ceci a pour but de limiter le *scraping* à des pages qui n'ont pas été incorporées dans la base de données.

Le second script a été conçu pour récupérer les caractéristiques contenues dans les annonces et afin de bien comprendre ce que réalisent ces deux scripts, les figures ci-dessous serviront à illustrer ces différences. La figure de gauche contient l'ensemble des annonces répondant aux critères, la figure de droite présente une annonce individuelle répertoriée par un numéro unique d'annonce.

Figure 4-3 : Comparis



(Comparis.ch 2021)

Lors de cette étude, qui ne s'est limitée qu'à des annonces consultables via les critères prédéfinis, un constat surprenant sur l'étendue des possibilités offertes en termes de *scraping* a été réalisé. Le potentiel de *scraping* s'étend à la totalité des annonces parues, mais une telle démarche nécessiterait des moyens conséquents et comporterait des risques pour le site visé. Le travail en fera la mention dans le cadre de la gouvernance de données qui clôturera l'étude.

À titre consultatif, le script permettant l'extraction des caractéristiques des annonces et leur sauvegarde dans le format *.csv* est mis à la disposition du lecteur. Le recueil de ces informations n'a pas été une tâche aisée, dans la mesure où le site a évolué entre les premières tentatives de *scraping* et les suivantes. Ceci a rendu complexe la systématisation de la récolte de données, sans pour autant être un élément rédhibitoire. Ce commentaire s'inscrit dans le contexte du paradigme selon lequel l'information est disponible au grand public. Force est de constater qu'une certaine expérience est nécessaire, mais la popularité de Python et sa philosophie *open source* permettent de trouver sur des forums de nombreuses solutions aux erreurs les plus fréquentes.

Figure 4-4 : Scraping des données – 1/4

```
File  Edit  View  Help   Comparis_ScrapData.py - ...\PycharmProjects\ScrapData
Comparis_ScrapData.py
1  from urllib.request import urlopen as uReq
2  from os import getcwd, chdir, mkdir
3  import urllib.request
4  import requests
5  import pandas as pd
6  import re
7
8
9  token = 'https://fr.comparis.ch/immobilien/result/list?requestobject=%7B%22DealType%22%3A10%2C%22SiteId%22%3A0%2C%22RootPropertyTypes%22%3A
10
11
12 def get_pages(token, nb):
13     pages = []
14     for i in range(1, nb + 1):
15         j = token + str(i)
16         pages.append(j)
17     return pages
18
19
20 def get_annonces(token, num_annonces, index):
21     pages = []
22     for i in range(0, 354):
23         j = token + str(num_annonces[index])
24         index += 1
25         pages.append(j)
26     return pages
27
28
29 pages = get_pages(token, 1)
30 headers = {
31     'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/87.0.4280.88 Safari/537.36',
32 }
33 for i in pages:
34     response = requests.get(i, headers=headers)
35
36 import bs4
37
38 soup = bs4.BeautifulSoup(response.text, 'html.parser')
39 for p in soup.find_all('script', {"id": "__NEXT_DATA__"}, {"type": "application/json"}):
40     print(p)
41
42 results_texte = str(p)
43 chaine = "initialResultData"
44
45
46 def trouver(resultats, mot):
47     index = 0
48     while index < len(resultats):
49         if mot == resultats[index:index + len(mot)]:
50             return index + len(mot) + 12
51         index = index + 1
52     return -1
53
54
55 def extraction_resultats(index, mystring):
56     begin = index
57     end = index + 8999
58     mystring = mystring[begin:end]
59     mystring.replace("'", "")
60     return mystring.split(",")
61
```

Figure 4-5 : Scraping des données – 2/4

```python
indice_resultats = int(trouver(results_texte, chaine))

print(indice_resultats)

tableau_annonces = extraction_resultats(indice_resultats, results_texte)


print(getcwd())
chdir('Resultats')
print(getcwd())

""" Il faudrait sauvegarder la liste des annonces sous csv et les travailler hors programme pour supprimer les doublons"""
print(tableau_annonces)

df_annonce = pd.read_csv("Annonce_GE_DEF.csv", index_col=False)

print(df_annonce['number'].tolist())

annonce = "https://fr.comparis.ch/immobilien/marktplatz/details/show/"

"""Remplacer df_annonce['number'] par tableau_annonces pour extraire les nouvelles annonces"""

pages_annonces = get_annonces(annonce, df_annonce['number'], 0)
print("Pages annonces {}".format(pages_annonces))


def url_ok(url):
    """Function testing if URL can be accessed, parameter = URl string"""
    request = urllib.request.Request(url)
    request.add_header('user-agent', 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/87.0.4280.88
    request.get_method = lambda: 'HEAD'
    try:
        urllib.request.urlopen(request)
        return True
    except urllib.request.HTTPError:
        return False


def find_all_tables(url):
    """Function extracting all tables from a given webpage, parameter = URL string"""
    if url_ok(url):
        request = urllib.request.Request(url)
        request.add_header('user-agent', 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/87.0.4280
        uClient = urllib.request.urlopen(request)
        page_html = uClient.read()
        uClient.close()
        page_soup = bs4.BeautifulSoup(page_html, 'html.parser')
        print(page_soup)
        data = page_soup.find_all('script', {"id": "__NEXT_DATA__"}, {"type": "application/json"})
        return data[0].prettify('latin-1')



def extraction_data(pattern, caracteristiques_annonce):
    if re.search(pattern, caracteristiques_annonce) is None:
        return "N/A"
    else:
        substring = re.search(pattern, caracteristiques_annonce).group(1)
    return substring
```

Figure 4-6 : Scraping des données – 3/4+4/4

File Edit View Help Comparis_ScrapData.py - ...\PycharmProjects\ScrapData

Comparis_ScrapData.py

```python
prop_AdId = '"AdId":(.*?),'
prop_AdStatus = '"AdStatus":(.*?),'
prop_DealType = '"DealType":(.*?),'
prop_Site = '"Site":(.*?),'
prop_Title = '"Title":"(.*?)",'
prop_NumRooms = '"NumRooms":(.*?),'
prop_Address = '"Address":"(.*?)",'
prop_PropertyTypeId = '"PropertyTypeId":(.*?),'
prop_PropertyTypeText = '"PropertyTypeText":"(.*?)",'
prop_ComparisPoints = '"ComparisPoints":(.*?),'
prop_Price = '"Price":(.*?),'
prop_PriceText = '"PriceText":"(.*?)",'
prop_PriceType = '"PriceText":"(.*?)",'
prop_PriceTypeText = '"PriceTypeText":"(.*?)",'
prop_Area = '"Area":(.*?),'
prop_FoundForTheFirstTime = '"FoundForTheFirstTime":"(.*?)",'
prop_GeoPosLat = '"GeoPosLat":(.*?),'
prop_GeoPosLng = '"GeoPosLng":(.*?),'

prop_PropertyType = '"Key":"PropertyType","Value":"(.*?)",'
prop_NumRooms2 = '"Key":"NumRooms","Value":"(.*?)",'
prop_Floor = '"Key":"Floor","Value":"(.*?)",'
prop_LivingSpace = '"Key":"LivingSpace","Value":"(.*?)'
prop_YearOfConstruction = '"Key":"YearOfConstruction","Value":"(.*?)",'
prop_AvailableDate = '"Key":"AvailableDate","Value":"(.*?)",'
prop_HasBalconies = '"Key":"HasBalconies","Value":(.*?),'
prop_HasParkingIndoor = '"Key":"HasParkingIndoor","Value":(.*?),'
prop_HasFireplace = 'Key":"HasFireplace","Value":(.*?),'

prop_GrossRent = '"Key":"GrossRent","Value":"(.*?)",'
prop_NetRent = '"Key":"NetRent","Value":"(.*?)",'
prop_SideCost = '"Key":"SideCost","Value":"(.*?)",'

prop_AdvertiserInformation_Name = '"AdvertiserInformation":{"Name":(.*?),'
prop_VendorInformation_Name = '"VendorInformation":{"Name":"(.*?)",'
prop_VisitationContactInformation_Name = '"VisitationContactInformation":{"Name":"(.*?)",'
prop_PartnerName = '"PartnerName":"(.*?)",'

prop_Remarks = '"Remarks":(.*?),"MapData"'

Proprietes = [prop_AdId, prop_AdStatus, prop_DealType, prop_Site, prop_Title, prop_NumRooms, prop_Address,
              prop_PropertyTypeId, prop_PropertyTypeText, prop_ComparisPoints, prop_Price, prop_PriceText,
              prop_PriceType, prop_PriceTypeText, prop_Area, prop_FoundForTheFirstTime,
              prop_GeoPosLat, prop_GeoPosLng, prop_PropertyType, prop_NumRooms2, prop_Floor, prop_LivingSpace,
              prop_YearOfConstruction, prop_AvailableDate, prop_HasBalconies, prop_HasParkingIndoor, prop_HasFireplace,
              prop_GrossRent, prop_NetRent, prop_SideCost, prop_AdvertiserInformation_Name, prop_VendorInformation_Name,
              prop_VisitationContactInformation_Name, prop_PartnerName, prop_Remarks]

Annonce_data = []

"""POUR TESTER LES ERREURS D'EXTRACTION DES ANNONCES
for i in range(0, 34):
    print(extraction_data(Proprietes[i], results))
    Annonce_data.append(extraction_data(Proprietes[i], results))
```

File Edit View Help Comparis_ScrapData.py - ...\PycharmProjects\ScrapData

Comparis_ScrapData.py

```python
df = pd.DataFrame(columns=['prop_AdId', 'prop_AdStatus', 'prop_DealType', 'prop_Site', 'prop_Title',
                           'prop_NumRooms', 'prop_Address','prop_PropertyTypeId', 'prop_PropertyTypeText',
                           'prop_ComparisPoints','prop_Price', 'prop_PriceText', 'prop_PriceType',
                           'prop_PriceTypeText', 'prop_Area', 'prop_FoundForTheFirstTime',
                           'prop_GeoPosLat', 'prop_GeoPosLng', 'prop_PropertyType', 'prop_NumRooms2',
                           'prop_Floor', 'prop_LivingSpace', 'prop_YearOfConstruction',
                           'prop_AvailableDate', 'prop_HasBalconies', 'prop_HasParkingIndoor',
                           'prop_HasFireplace','prop_GrossRent', 'prop_NetRent', 'prop_SideCost',
                           'prop_AdvertiserInformation_Name', 'prop_VendorInformation_Name',
                           'prop_VisitationContactInformation_Name','prop_PartnerName','prop_Remarks'])

for j in range(0, 354):
    """print(find_all_tables(pages_annonces[j]))"""
    print(j)
    results_annonces = find_all_tables(pages_annonces[j])
    results = str(results_annonces)
    for i in range(0, 35):
        Annonce_data.append(extraction_data(Proprietes[i], results))
        """print(Annonce_data)
        print(len(df.columns))
        print(len(Annonce_data))"""
    df.loc[j] = Annonce_data
    Annonce_data.clear()


df.to_csv('Résultats 0 - 09.01.21.csv')
```

La première partie du code, jusqu'à la ligne 67, reprend le code du script précédent afin de conserver la possibilité d'extraire les annonces selon d'autres filtres en cas de nécessité. Elle n'est pas utilisée dans le cadre de ce script mais reste à disposition pour l'activer en cas de besoin. Il suffirait d'affecter au token de la ligne 9 l'url portant les filtres ciblés. Ce script récolte pour chaque annonce 35 caractéristiques, ou variables, qui sont intégrées à un dataframe de manière itérative pour chaque annonce contenue dans le fichier *Annonce_GE_DEF.csv*, dans lequel on aura préalablement inclus les numéros d'annonce ciblés. L'opération nécessite environ une dizaine de secondes par annonces, ce qui reste acceptable un terme de fréquence d'interrogation du serveur et ne risque pas de paralyser le site. La possibilité d'augmenter cette fréquence n'a pas fait l'objet de l'étude, les risques semblant trop importants par rapport aux besoins effectifs de l'étude.

Ces informations sont sauvegardées dans un fichier *.csv* qui servira à alimenter notre base de données d'annonces de la ville de Genève. La copie vers la base de données intégrale est faite de façon manuelle, mais une automatisation devrait être envisagée si le nombre de lignes devient trop important. Le format *.csv* est parfaitement adapté à l'utilisation qui est prévue dans ce travail et la mise en place d'une base de données, comme MongoDB, n'est pas indispensable.

4.2.2 Data cleaning

Avant de pouvoir utiliser les données récoltées, il faut s'assurer de leur complétude et de leur pertinence. Les sources d'erreurs pour les phases ultérieures du projet peuvent provenir de multiples facteurs. Les valeurs manquantes, aberrantes, ou de type incohérent avec les variables traitées doivent faire l'objet d'un retraitement avant de les utiliser. On appelle cette phase du projet le *data cleaning* ou nettoyage de données.

De nombreuses techniques existent pour cela, mais ce travail ne permet pas de s'y intéresser tant les exemples sont multiples et les cas variés. Le nettoyage de ces données n'a pas fait l'objet d'un script. Les données manquantes n'ont pas été inclues dans les phases ultérieures du projet si ces données étaient indispensables au traitement et calculs réalisé. Dans les cas où figuraient néanmoins des données aberrantes ou absentes qui ne portaient pas préjudice au travail, elles ont été identifiées par un traitement différencié, comme dans le cas de la cartographie, proposé ci-après.

4.2.3 Base de données : quelques chiffres

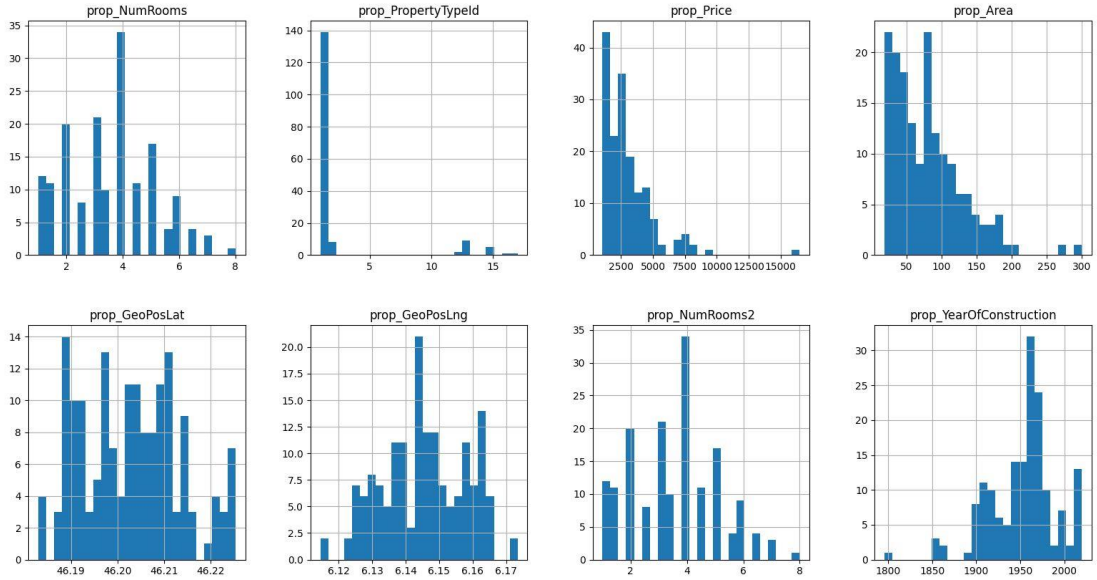
Après quelques semaines de récolte de données, notre base de données se compose de 2’762 annonces. De nombreuses annonces semblent ne pas posséder toutes les données d’intérêt. Comme mentionné précédemment, ces données seront ignorées dans le cas où elles sont nécessaires pour délivrer un résultat.

4.2.3.1 Visualisation via les graphes

Les histogrammes sont disponibles de manière relativement simple sur Python. Après avoir filtré le jeu de données pour retenir les valeurs qui pourront être utiles dans le script qui suit, cette partie propose quelques visualisations qui permettront au lecteur de comprendre les possibilités d’analyse que le jeu de données récoltées peu permettre.

La première figure qui a été obtenue consiste en un tableau d’histogrammes. Elle permet d’identifier rapidement la forme globale des données et d’identifier le cas échéant les valeurs aberrantes qui seraient présentes.

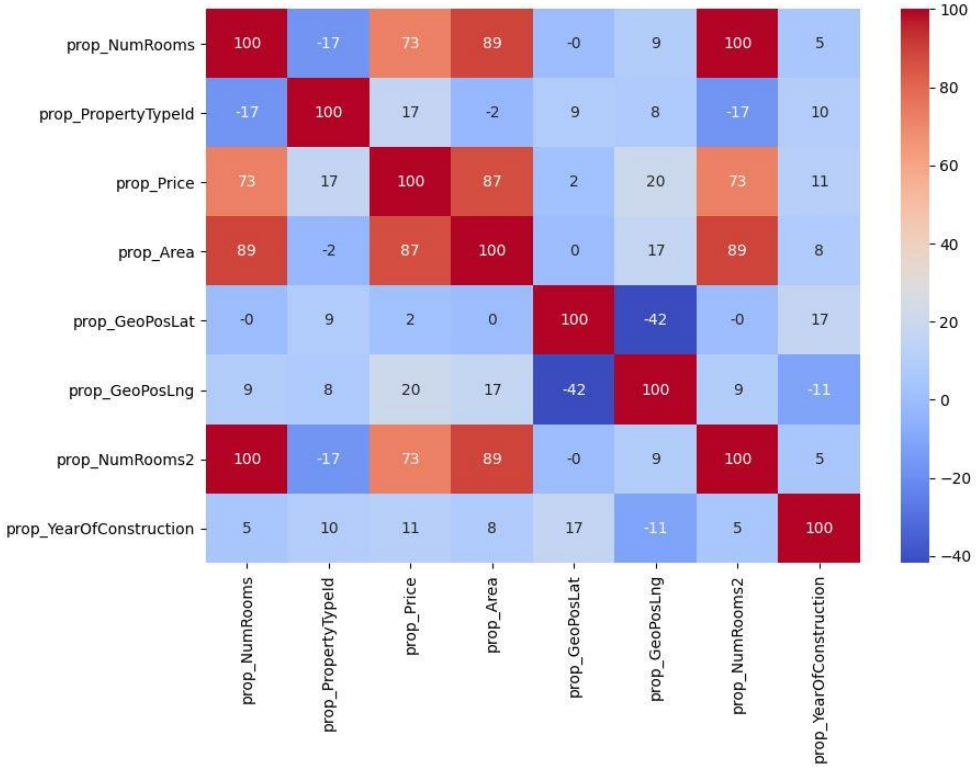
Figure 4-7 : Histogramme - Appartements



On constate rapidement la répartition des principales variables et les éventuelles formes caractéristiques des coefficients d’asymétrie, comme pour la variable *prop_Area* avec un *skweness* positif.

La deuxième figure présente une matrice de corrélation telle qu'on a pu l'obtenir selon le même jeu de données. Il est intéressant de constater que certaines données possèdent de fortes corrélations et permettent d'anticiper leur utilisation comme variables explicatives dans un modèle de régression.

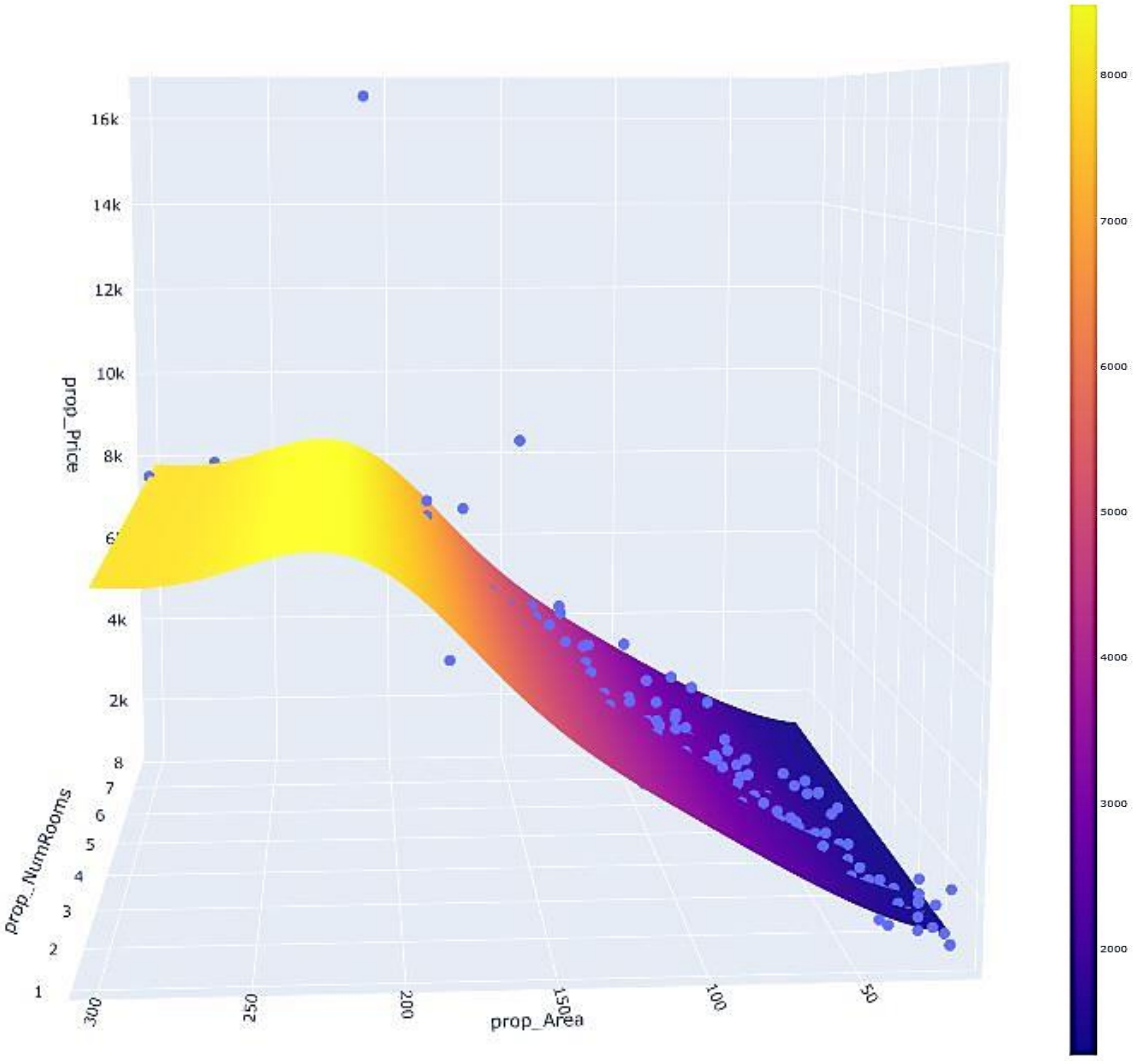
Figure 4-8 : Matrice de corrélations



On constate ici de manière visuelle et rapide que les champs les plus corrélés, outre les champs *prop_NumRooms* et *prop_NumRooms2* qui sont strictement identiques, que les champs les plus corrélés sont *prop_Price*, *prop_Area*, *prop_NumRooms*. Ces mêmes champs seront utilisés dans le modèle de régression qui suit.

La dernière figure démontre avec un résultat flagrant le type de prédiction que l'on peut obtenir avec les outils de *data science*. Une régression avec le Support Vector Machine a été utilisée dans cet exemple pour déterminer un modèle de prix, en fonction du nombre de pièces et de la surface habitable. Le modèle, correspondant à la surface colorée, ne vient pas s'adapter aux valeurs extrêmes mais suit globalement la forme du jeu de données, ce qui présume de l'absence de surajustement.

Figure 4-9 : Régression avec SVM



(Plotly 2021)

Le script relatif à cette partie est mis à disposition dans les annexes de ce travail.

4.2.3.2 Visualisation cartographique

Pour clore cette étude, ce travail propose de présenter une cartographie d'une partie des résultats obtenus. Pour rappel, la partie de *data cleaning* a été gérée hors du script et ce dernier ne se base donc pas sur l'ensemble du jeu de données disponible. Toutefois, l'étude permet de constater que la représentation cartographique laisse entrevoir une multitude de possibilités d'adjonction d'informations aux données extraites, sous la forme de cartes choroplèthes, de rasters et de masques, qui permettent d'identifier rapidement des données particulières, tout autant que la répartition géospatiale des annonces, le type d'objet proposé ou le prix annoncé.

Les cartes ci-dessous représentent les résultats obtenus en fonction du type d'objet, la taille des cercles servant à représenter les prix.

Figure 4-10 : Carte - Appartements

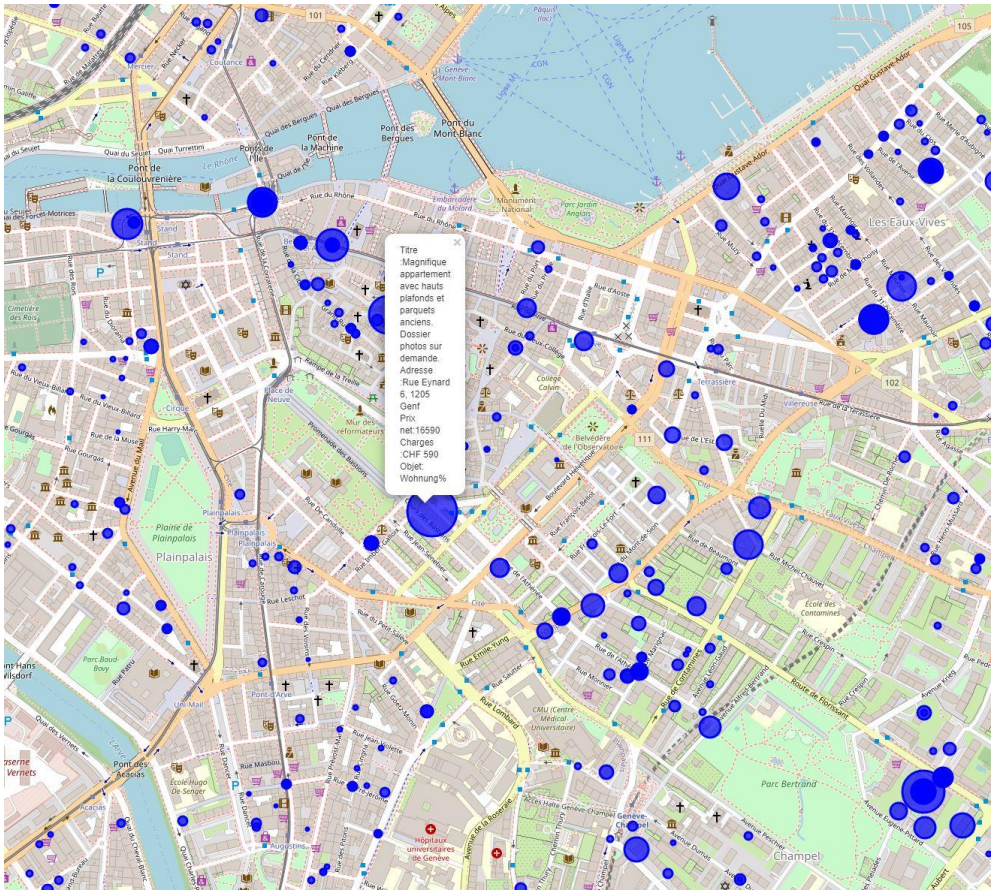
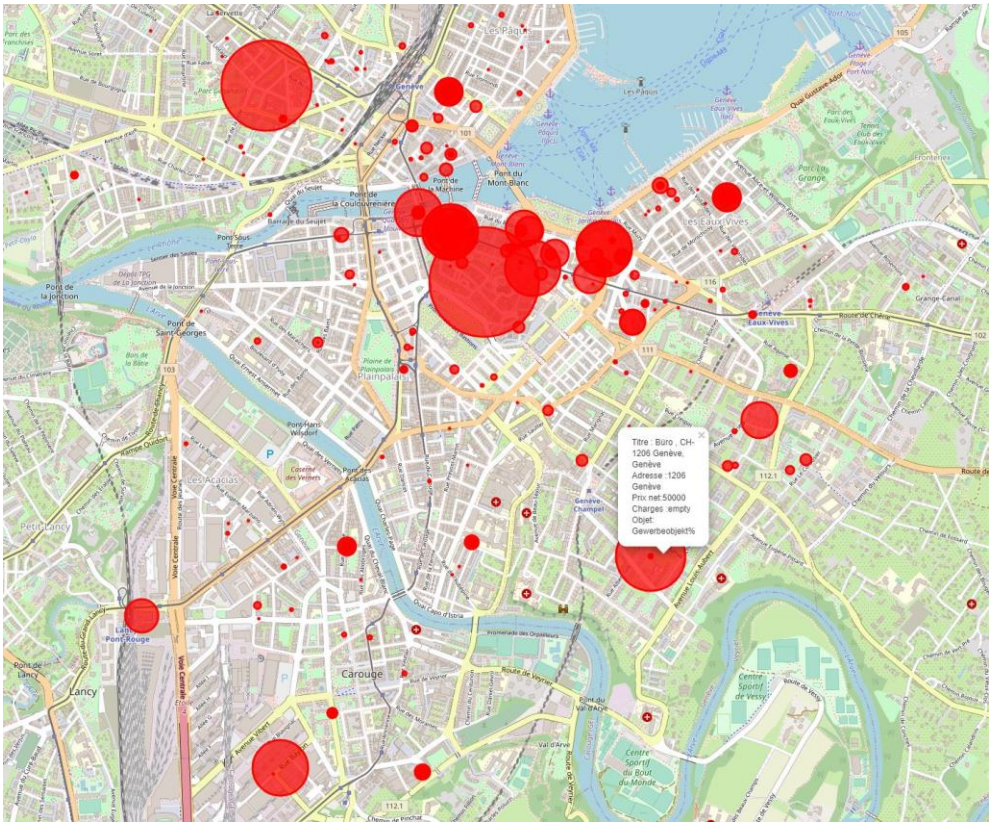


Figure 4-11 : Carte – Objets particuliers

Figure 4-12 : Carte - Locaux et Bureaux



4.3 Limites et Data Governance

Cette étude pratique nous aura permis de démontrer que les informations disponibles sur Internet permettent de mieux comprendre la topologie de l'offre immobilière à chaque instant. Prérogative des régies immobilières et autres acteurs traditionnels du marché, cette information semble pouvoir être utilisée de manière appliquée par un particulier sans que les ressources nécessaires soient trop importantes.

Les limites rencontrées dans cette étude, en dehors de l'aspect technique, se résument avant tout à la complétude des données. Aucune systématique clairement repérable n'a été identifiée dans les jeux de données récoltées. Les données manquantes ou les annonces sans lien avec les filtres, à l'instar d'un restaurant à remettre dans les filtres « louer » et « Genève », constituent autant de défi pour un traitement systématique des données. D'autre part, le risque juridique, bien que la loi n'empêche pas de manière stricte ces pratiques, reste bien présent et restreint les velléités de scalabilité de tels projets.

La problématique de cette étude visait à déterminer si l'obtention d'informations sur le marché immobilier locatif genevois était possible et quel était le niveau de détails de cette information, sa profondeur et sa fiabilité. Il semble clair que l'obtention d'une telle information est possible, avec un niveau de détail et une profondeur très importante. Parmi les critères que cette étude a su extraire, le nom de la personne de contact pour les visites était mentionné. L'extraction du numéro de téléphone de cette personne aurait été possible, mais les questions de protection des données semblaient trop importantes et les scripts de *scraping* n'ont pas visé ces données.

La data governance reste une thématique de premier plan dans ce contexte, pour tous les acteurs de l'annonce en ligne. Ce travail ne pourra pas en aborder toutes ces spécificités, même s'il permet de démontrer que les informations sur le web ne sont pas strictement privées et que leur utilisation peut aisément dépasser le strict cadre dans lequel elles ont été initialement fournies.

5. Conclusion

La conception de ce travail a débuté dans un contexte particulier, celui d'une crise sanitaire mondiale qui a rebattu les cartes de la définition et de l'utilisation de nos espaces de vie. Rythmée par des confinements successifs dont l'objectif est de contrôler la propagation du coronavirus SARS-CoV-2, l'année écoulée nous aura permis de découvrir les étonnantes capacités d'adaptation de nos économies, de notre propension à accepter le changement dans les situations exceptionnelles, laissant entrevoir le nouveau rôle central de ces technologies au quotidien.

Qu'il s'agisse du Big Data, de la *data science*, ou de la BI, l'information, émanant des données, représente une ressource qui doit être exploitée afin de permettre une prise de décision aussi éclairée que possible. Les technologies évoquées dans ce travail développent nos capacités à comprendre ces données, leurs interactions et les facteurs endogènes et exogènes qui les influencent. Elles améliorent ainsi notre faculté à conceptualiser notre environnement et à décider.

La proptech, qui représente un sous-ensemble de cette transformation digitale, semble pourtant ne pas répondre aux attentes du sein du secteur immobilier, bien que de nombreux acteurs traditionnels comprennent son intérêt et la nécessité de sa mise œuvre, initiant souvent eux-mêmes les relations d'affaires et participant au capital de certaines de ces start-ups. Deux pistes de réflexion expliquent ce paradoxe. D'une part, les paradigmes, dans lesquels évoluent ces deux parties prenantes, ne concordent pas sur de nombreux aspects, allant de la culture d'entreprise à la connaissance du métier. Coopération ou phagocytose sont les termes d'un accord encore flou, qui décidera de l'évolution de la proptech au sein du marché immobilier. D'autre part, l'écosystème de la proptech n'a pas encore atteint sa maturité. De nombreuses sociétés semblent en effet avoir focalisé leur attention sur des branches de l'activité immobilière dans lesquelles elles sont surreprésentés. La concurrence entre ces sociétés permettra d'asseoir à quels besoins réels elles devront répondre pour espérer survivre dans leurs jeunes années.

On aura vu que les données générées par le secteur immobilier, dépourvues de standard, rendent difficiles leur agrégation et leur exploitation. C'est l'une des principales raisons pour laquelle on ne voit pas encore de marché immobilier mondialement globalisé. Les régulations de chaque pays quant à cet actif, très corrélé à la fortune des ménages et aux richesses d'un état, varient de manière importante d'une juridiction à une autre. Considéré comme un bien stratégique, l'immobilier devra savoir s'imposer des standards communs de comparaison, sous l'impulsion d'une

économie mondiale à la recherche de croissance et de stabilité. La part de l'allocation immobilière dans les actifs des sociétés technologiques et les portefeuilles des fonds de pension augmente et avec elle, la demande de services spécialisés dans l'information immobilière. Ce travail a notamment présenté la société PriceHubble, spécialisée dans les solutions d'Automated Valuation Model, ou AVM, qui est un acteur innovant utilisant la *data science* pour améliorer la prédiction des modèles de valorisation des actifs immobiliers sur un horizon de court à moyen terme. Les réglementations semblent évoluer en faveur de telles solutions, avec l'attention qui est portée aujourd'hui sur le secteur immobilier dans l'évaluation du risque de crédit.

Afin de réellement comprendre quelle est la puissance de ces technologies, un cas pratique a été développé pour le marché immobilier locatif genevois. Celui-ci aura permis de comprendre que l'accès à l'information pour un particulier désireux d'anticiper l'évolution du marché immobilier local est possible, bien qu'il ait révélé que le niveau de complétude des données récoltées posait un problème important en termes d'extrapolation. Il a également été démontré que les sites d'annonce en ligne n'utilisaient encore que très peu les technologies de la *data science*, présageant du potentiel d'évolution de ces sites, indicateurs précoce de l'offre immobilière et par ricochet, indicateurs potentiels avancés de la demande immobilière en cas de récolte d'informations sur les utilisateurs. Certains acteurs semblent d'ailleurs vouloir exploiter cette opportunité, dont Properstar, qui semble vouloir bousculer le modèle de l'annonce en ligne, encore pensée comme un simple répertoire digital.

Ce travail cherchait ainsi à comprendre les tenants et aboutissants de la transformation digitale qui touche le secteur immobilier, ainsi que les ressorts sur lesquels s'appuie la proptech pour générer de la valeur dans ce marché. Force est de constater que ces technologies n'ont pas encore démontré toutes les facettes de leur pouvoir et qu'elles auront un rôle déterminant à jouer dans la conception des bâtiments, dans leur cycle de vie tout autant que dans notre manière d'interagir avec ces lieux de vie.

Bibliographie

1965. MOORE, Gordon E, 1965. Cramming more components onto integrated circuits. Electronics. 1965. Vol. 38, n° 8, pp. 4.

1998. NIELSEN, Jakob, 1998-2019. Nielsen's Law of Internet Bandwidth. *Nielsen Norman Group* [en ligne]. [Consulté le 5 septembre 2020]. Disponible à l'adresse : https://www.nngroup.com/articles/law-of-bandwidth/

2001. LANEY, Doug, 2001. *3D Data Management: Controlling Data Volume, Velocity and Variety* [document PDF]. 6 février 2001. n° 949. META Group Inc.

2005. WALTER, William J., 2005. Kryder's Law - Scientific American. *Scientific American* [en ligne]. 1 août 2005. [Consulté le 27 juin 2020]. Disponible à l'adresse : https://www.scientificamerican.com/article/kryders-law/

2005. ZOU, Hui et HASTIE, Trevor, 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. avril 2005. Vol. 67, n° 2, pp. 301-320. DOI 10.1111/j.1467-9868.2005.00503.x.

2009. CHOI, Hyunyoung et VARIAN, Hal, 2009. Predicting the Present with Google Trends. [en ligne]. 10 avril 2009. [Consulté le 10 octobre 2020]. Disponible à l'adresse : https://poseidon01.ssrn.com/delivery.php?ID=226125090117005065064122021117121065052057047032095057123114009088087093088102118027096101058032022062055111095082001005116117114054094081027106018104016104122068061038050110028095096067123095007026010019068113094096126114097031022117066083126006007&EXT=pdf

2009. GINSBERG, Jeremy, MOHEBBI, Matthew H., PATEL, Rajan S., BRAMMER, Lynnette, SMOLINSKI, Mark S. et BRILLIANT, Larry, 2009. Detecting influenza epidemics using search engine query data. Nature. février 2009. Vol. 457, n° 7232, pp. 1012-1014. DOI 10.1038/nature07634.

2009. ROSSUM, Guido Van, 2009. The History of Python: A Brief Timeline of Python. *The History of Python* [en ligne]. 20 janvier 2009. [Consulté le 26 décembre 2020]. Disponible à l'adresse : https://python-history.blogspot.com/2009/01/brief-timeline-of-python.html

2011. BRYNJOLFSSON, Erik, HITT, Lorin M. et KIM, Heekyung Hellen, 2011. ID 1819486 : *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?* [en ligne]. SSRN Scholarly Paper. Rochester, NY : Social Science Research Network. [Consulté le 8 décembre 2020]. Disponible à l'adresse : https://papers.ssrn.com/abstract=1819486

2013. PRESS, Gil, 2013. A Very Short History Of Big Data. *Forbes* [en ligne]. 9 mai 2013. [Consulté le 6 septembre 2020]. Disponible à l'adresse : https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

2015. Pseudonyme: GLEN_B, 2015. StackExchange: Why is ridge regression called "ridge", why is it needed, and what happens when λ goes to infinity? *Cross Validated* [en ligne]. 8 mai 2015. [Consulté le 24 décembre 2020]. Disponible à l'adresse : https://stats.stackexchange.com/questions/151304/why-is-ridge-regression-called-ridge-why-is-it-needed-and-what-happens-when

2016. BIERNAT, Éric et LUTZ, Michel, 2016. Data science fondamentaux et études de cas: machine learning avec Python et R. Paris : Eyrolles. ISBN 978-2-212-14243-3.

2016. DE MAURO, Andrea, GRECO, Marco et GRIMALDI, Michele, 2016. A formal definition of Big Data based on its essential features. Library Review. 1 janvier 2016. Vol. 65, n° 3, pp. 122-135. DOI 10.1108/LR-06-2015-0061.

2016. KITCHIN, Rob et MCARDLE, Gavin, 2016. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*. 5 janvier 2016. Vol. 3, n° 1, pp. 205395171663113. DOI 10.1177/2053951716631130.

2016. MONINO, Jean-Louis et SEDKAOUI, Soraya, 2016. Big Data, Open Data et valorisation des données. London : ISTE Éditions. ISBN 978-1-78405-122-8.

2016. BARNES, Yolande, TOTSEVIN, Paul, 2016. Around The World In Dollars And Cents 2016. *Savills World Research* [en ligne]. 28 janvier 2016. [Consulté le 20 octobre 2020]. Disponible à l'adresse : https://pdf.euro.savills.co.uk/global-research/around-the-world-in-dollars-and-cents-2016.pdf

2016. KHAROUBI, Rachid, 2016. *Une nouvelle approche pour la sélection des variables dans le cas de modèles de discrimination en grandes dimensions* [en ligne]. Mémoire. UNIVERSITÉ DU QUÉBEC À MONTRÉAL. [Consulté le 24 décembre 2020]. Disponible à l'adresse : https://archipel.uqam.ca/8948/1/M14459.pdf

2017. TUFFÉRY, Stéphane et SAPORTA, Gilbert, 2017. Data mining et statistique décisionnelle : la science des données. ISBN 978-2-7108-1180-0.

2017. BENZAKI, Younes, 2017. Gradient Descent Algorithm : Explications et implémentation en Python. Mr. Mint : Apprendre le Machine Learning de A à Z [en ligne]. 15 mai 2017. [Consulté le 18 décembre 2020]. Disponible à l'adresse : https://mrmint.fr/gradient-descent-algorithm

2017. MITRA, Pratik, SANYAL, Anirban et CHOUDHURI, Sohini, 2017. Reserve Bank of India - Database. [en ligne]. 2017. [Consulté le 12 janvier 2021]. Disponible à l'adresse : https://m.rbi.org.in/Scripts/bs_viewcontent.aspx?Id=3516

2018. DELORT, Pierre, 2018. Le big data. 2ème édition. ISBN 978-2-13-079221-5.

2018. PROVOST, Foster et FAWCETT, Tom, 2018. Data science pour l'entreprise. ISBN 978-2-212-67570-2.

2018. AMINI, Alexander, 2018. Non-convex optimization. We utilize stochastic gradient descent to find a local optimum in our loss landscape. *ResearchGate* [en ligne]. mai 2018. [Consulté le 24 décembre 2020]. Disponible à l'adresse : https://www.researchgate.net/figure/Non-convex-optimization-We-utilize-stochastic-gradient-descent-to-find-a-local-optimum_fig1_325142728

2019. HASENMAILE, Frey, LOHSE, Alexander, NÄF, Philippe, RIEDER, Thomas et WALTERT, Fabian, 2019. Moniteur immobilier Suisse, Juin 2019 : *Les «jeunes fauves» sont devenus adultes* [en ligne]. Credit Suisse AG, Investment Solutions & Products. 22 mai 2019. [Consulté le 25 décembre 2020]. Disponible à l'adresse : https://www.credit-suisse.com/media/assets/private-banking/docs/ch/privatkunden/eigenheim-finanzieren/moniteurimmobilier-t2-2020.pdf

2019. GAUDIAUT, Tristan, 2019. Statista Digital Economy Compass 2019. *Forum Économique Mondial* [en ligne]. 26 avril 2019. [Consulté le 22 juillet 2020]. Disponible à l'adresse : https://fr.weforum.org/agenda/2019/04/la-totalite-des-donnees-creees-dans-le-monde-equivaut-a/

2018. THAKUR, Ankush, 2018. Comment démarrer avec l'apprentissage automatique. Geekflare [en ligne]. 10 octobre 2018. [Consulté le 18 décembre 2020]. Disponible à l'adresse : https://geekflare.com/fr/getting-started-with-machine-learning/

2019. Apprentissage Supervisé Vs. Non Supervisé, 2019. Le DataScientist [en ligne]. [Consulté le 18 décembre 2020]. Disponible à l'adresse : https://le-datascientist.fr/apprentissage-supervise-vs-non-supervise

2019. LEROUX, Hugo, 2019. 1965-2020 : La loi de Moore est morte - Science & Vie. *Science-et-vie.com* [en ligne]. 30 décembre 2019. [Consulté le 27 juin 2020].

Disponible à l'adresse : https://www.science-et-vie.com/science-et-culture/1965-2020-la-loi-de-moore-est-morte-53613

2019. UNISSU, 2019. What is PropTech? A definition, including a sector and region overview. [en ligne]. 5 mars 2019. [Consulté le 26 octobre 2020]. Disponible à l'adresse : https://www.unissu.com/proptech-resources/what-is-proptech

2019. MUELLER, John Paul, MASSARON, Luca et ENGLER, Olivier, 2019. *Data science avec Python pour les nuls*. ISBN 978-2-412-05072-9.

2019. Dataviz : PriceHubble diffuse une vidéo montrant l'évolution des prix de l'immobilier basée sur les données de la DGFiP, 2019. *Actu IA* [en ligne]. [Consulté le 10 janvier 2021]. Disponible à l'adresse : https://www.actuia.com/actualite/dataviz-pricehubble-diffuse-une-video-montrant-levolution-des-prix-de-limmobilier-basee-sur-les-donnees-de-la-dgfip/

2019. MORENO, Caio, 2019. Difference between BI (Business Intelligence) and Data Science. *Medium* [en ligne]. 25 février 2019. [Consulté le 10 janvier 2021]. Disponible à l'adresse : https://caiomsouza.medium.com/difference-between-bi-business-intelligence-and-data-science-1a9c7628bbdb

2020. DELOITTE CENTER FOR FINANCIAL SERVICES, 2020. Challenges in financial services post COVID-19 | Deloitte Insights. [en ligne]. 2020. [Consulté le 26 octobre 2020]. Disponible à l'adresse : https://www2.deloitte.com/us/en/insights/economy/covid-19/covid-19-financial-services-sector-challenges.html#impact-on-proptechs

2019. WEIR, Andrew, PYLE, Andy et GRUNEWALD, Sander, 2019. KPMG Global PropTech Survey 2019: Is your digital future in the right hands? *KPMG International* [en ligne]. octobre 2019. [Consulté le 27 décembre 2020]. Disponible à l'adresse : https://assets.kpmg/content/dam/kpmg/uk/pdf/2019/11/global-proptech-survey-2019.pdf

2019. Cross Industry Standard Process for Data Mining, 2019. *Wikipédia* [en ligne]. [Consulté le 12 janvier 2021]. Disponible à l'adresse : https://fr.wikipedia.org/w/index.php?title=Cross_Industry_Standard_Process_for_Data_Mining&oldid=164909522

2020. BAUDOT, Jean-Yves, [sans date]. Indicateurs d'écart : SCR, MSE, RMSE, MAE et MAPE. [en ligne]. [Consulté le 18 décembre 2020]. Disponible à l'adresse : http://www.jybaudot.fr/Stats/indicecarts.html

2020. VON DITFURTH, Jörg et AHOLT, Hendrik, 2018. Data is the new gold. *Deloitte France* [en ligne]. 2018. [Consulté le 28 décembre 2020]. Disponible à l'adresse : https://www2.deloitte.com/fr/fr/pages/immobilier/articles/data-is-the-new-gold.html

2020. GIBOR, Doron, HAREL, Amit et TRAJTENBERG, Maya, 2020. Proptech on the move. *Deloitte Netherlands* [en ligne]. 2020. [Consulté le 28 décembre 2020]. Disponible à l'adresse : https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/real-estate/ deloitte-nl-fsi-re-rep-2020-ch5.pdf

2020. Perspectives de l'économie mondiale, octobre 2020. *Fonds Monétaire International* [en ligne]. [Consulté le 20 novembre 2020]. Disponible à l'adresse : https://www.imf.org/fr/Publications/WEO/Issues/2020/09/30/world-economic-outlook-october-2020

2020. DELL INC., 2020. Données non structurées : stockage des données. [en ligne]. 2020. [Consulté le 6 septembre 2020]. Disponible à l'adresse : https://www.delltechnologies.com/fr-ch/learn/data-storage/unstructured-data.htm

2020. GOODMAN, Jonathan, MCTAVISH, Gavin, KLEIN, Florian et GOPI, Billa, 2020. COVID-19 : faire face à l'incertitude pendant et après la crise - L'analyse de scénarios :

un outil puissant pour améliorer la prise de décisions. [en ligne]. avril 2020. [Consulté le 20 novembre 2020]. Disponible à l'adresse : https://www2.deloitte.com/content/dam/Deloitte/ca/Documents/about-deloitte/ca-fr-about-covid-19-confronting-uncertainty-through-beyond-crisis-aoda.pdf

2020. POM+CONSULTING AG et RÜTTER SOCECO AG, 2020. L'importance de l'immobilier suisse pour l'économie nationale. *HEV Schweiz* [en ligne]. 23 septembre 2020. [Consulté le 20 novembre 2020]. Disponible à l'adresse : https://www.hev-schweiz.ch/fileadmin/sektionen/hev-schweiz/PDFs_Dateien/Atlas_der_Immobilienwirtschaft_Schweiz/Kurzbericht_Immobilienwirtschaft_FR.pdf

2020. OFFICE FÉDÉRAL DE LA STATISTIQUE, 2020. Logements vacants. [en ligne]. 5 octobre 2020. [Consulté le 21 octobre 2020]. Disponible à l'adresse : https://www.bfs.admin.ch/bfs/fr/home/statistiken/bau-wohnungswesen/wohnungen/leerwohnungen.html

2020. Histoire des ordinateurs. *Wikipédia* [en ligne]. [Consulté le 28 août 2020]. Disponible à l'adresse : https://fr.wikipedia.org/w/index.php?title=Histoire_des_ordinateurs&oldid=175457948

2020. SECO, Secrétariat d'Etat à l'économie, 2020. Tendances conjoncturelles Hiver 2020/2021. [en ligne]. 15 décembre 2020. [Consulté le 26 octobre 2020]. Disponible à l'adresse : https://www.seco.admin.ch/dam/seco/fr/dokumente/Wirtschaft/Wirtschaftslage/Konjunkturtendenzen/konjunkturtendenzen_winter_2020_2021.pdf.download.pdf/KT_2020_04.pdf

2020. KNIGHT, Andrew, [sans date]. Data Standards Will Encourage PropTech Revolution In Real Estate. *CIOApplicationseurope* [en ligne]. [Consulté le 26 octobre 2020]. Disponible à l'adresse : https://www.cioapplicationseurope.com/cxoinsights/data-standards-will-encourage-proptech-revolution-in-real-estate-nid-1448.html

2020. Global PropTech Directory - Over 8900 Companies Listed - Unissu. [en ligne]. [Consulté le 28 décembre 2020]. Disponible à l'adresse : https://www.unissu.com/proptech-companies?country=220&ordering=4

2020. SUN, Kevin, 2020. Proptech Pros Say Coronavirus Will Speed Up Tech Adoption. *The Real Deal New York* [en ligne]. 6 avril 2020. [Consulté le 28 décembre 2020]. Disponible à l'adresse : https://therealdeal.com/2020/04/06/proptech-and-the-pandemic-will-coronavirus-change-how-real-estate-works/

2020. Erreur quadratique moyenne, 2020. *Wikipédia* [en ligne]. [Consulté le 18 décembre 2020]. Disponible à l'adresse : https://fr.wikipedia.org/w/index.php?title=Erreur_quadratique_moyenne&oldid=168633554
Page Version ID: 168633554

2020. Analysez deux variables quantitatives par régression linéaire, 2020. OpenClassrooms [en ligne]. [Consulté le 18 décembre 2020]. Disponible à l'adresse : https://openclassrooms.com/fr/courses/4525266-decrivez-et-nettoyez-votre-jeu-de-donnees/4774671-analysez-deux-variables-quantitatives-par-regression-lineaire

2020. TRIPATHI, Mayank, 2020. Underfitting and Overfitting in Machine Learning. [en ligne]. 13 juin 2020. [Consulté le 18 décembre 2020]. Disponible à l'adresse : https://datascience.foundation/sciencewhitepaper/underfitting-and-overfitting-in-machine-learning

2020. SNEIDERMAN, Robby, 2020. From Linear Regression to Ridge Regression, the Lasso, and the Elastic Net. *Medium* [en ligne]. 6 novembre 2020. [Consulté le 24 décembre 2020]. Disponible à l'adresse : https://towardsdatascience.com/from-linear-regression-to-ridge-regression-the-lasso-and-the-elastic-net-4eaecaf5f7e6

2021. index | TIOBE - The Software Quality Company. [en ligne]. [Consulté le 2 janvier 2021]. Disponible à l'adresse : https://www.tiobe.com/tiobe-index/

2021. PriceHubble : À propos | LinkedIn. *Linkedin* [en ligne]. [Consulté le 10 janvier 2021]. Disponible à l'adresse : https://www.linkedin.com/company/pricehubble-ag/about/

2021. Mario Ubeda Garcia | LinkedIn. *Linkedin* [en ligne]. [Consulté le 10 janvier 2021]. Disponible à l'adresse : https://www.linkedin.com/in/marioug/

2021. Immobilier à louer à Genève - Comparer 3'421 annonces avec comparis.ch, [sans date]. *Comparis.ch* [en ligne]. [Consulté le 13 janvier 2021]. Disponible à l'adresse : https://fr.comparis.ch/immobilien/marktplatz/geneve/mieten?sort=11

2021. ML Regression. *Plotly* [en ligne]. [Consulté le 13 janvier 2021]. Disponible à l'adresse : https://plotly.com/python/ml-regression/

Annexe 1 : Script pour les graphes

```
from os import getcwd, chdir, mkdir
import geojson
import pandas as pd
import plotly.graph_objects as go
import plotly.express as px
import folium
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.svm import SVR
import numpy as np


df = pd.read_csv("Appartements_Data_masterfile.csv", sep=";",
low_memory = False)

print(df.head())

print(df.keys())

print(getcwd())

print(df.dtypes)

df_histogram = []

for feature in df.dtypes[df.dtypes == 'float64'].index:
    df_histogram.append(feature)
    print(df_histogram)

df_histogram.pop(0)
print(df_histogram)

colonnes = ['prop_AdId', 'prop_NumRooms', 'prop_PropertyTypeId',
'prop_Price', 'prop_Area', 'prop_GeoPosLat', 'prop_GeoPosLng',
'prop_NumRooms2', 'prop_YearOfConstruction']

df_donnees = df[colonnes]

print(df_donnees)

df_donnees_histogram = df_donnees.drop(columns=['prop_AdId'])

ax_list = df_donnees_histogram.hist(bins=25, layout=(2,4),
figsize=(15,15))

plt.show()

correlations = df_donnees_histogram.corr()

plt.figure(figsize=(4,4))
sns.heatmap(correlations*100, cmap='coolwarm', annot=True, fmt='.0f')
plt.show()


"""Cette partie reprend un exemple mis à disposition sur
https://plotly.com/python/ml-regression/"""
```

```
mesh_size = .1
margin = 0

df = px.data.iris()

X = df_donnees_histogram[['prop_NumRooms', 'prop_Area']]
y = df_donnees_histogram['prop_Price']

# Condition the model on sepal width and length, predict the petal
width
model = SVR(C=10000.)
model.fit(X, y)

# Create a mesh grid on which we will run our model
x_min, x_max = X.prop_NumRooms.min() - margin, X.prop_NumRooms.max()
+ margin
y_min, y_max = X.prop_Area.min() - margin, X.prop_Area.max() + margin
xrange = np.arange(x_min, x_max, mesh_size)
yrange = np.arange(y_min, y_max, mesh_size)
xx, yy = np.meshgrid(xrange, yrange)

# Run model
pred = model.predict(np.c_[xx.ravel(), yy.ravel()])
pred = pred.reshape(xx.shape)

# Generate the plot
fig = px.scatter_3d(df_donnees_histogram, x='prop_NumRooms',
y='prop_Area', z='prop_Price')
fig.update_traces(marker=dict(size=5))
fig.add_traces(go.Surface(x=xrange, y=yrange, z=pred,
name='PricePred'))
fig.show()
```

Annexe 2 : Script pour la cartographie

```
from os import getcwd, chdir, mkdir
import geojson
import pandas as pd
import plotly.graph_objects as go
import folium

df = pd.read_csv("MASTERFILE_Résultats Location Geneve_3.csv",
sep=";")

print(df.head())

print(df.keys())

print(getcwd())

carte = folium.Map(location=[46.204391, 6.143158], zoom_start=12,
tiles="OpenStreetMap")
carte1 = folium.Map(location=[46.204391, 6.143158], zoom_start=12,
tiles="OpenStreetMap")
carte2 = folium.Map(location=[46.204391, 6.143158], zoom_start=12,
tiles="OpenStreetMap")


radius_circle = .001

for i in range (0,len(df)):
    print(i)
    print(df['prop_Address'][i])
    if df['prop_Price'][i] > 100000:
        radius_circle = 0.0001
        color_circle = "yellow"
        folium.CircleMarker([df['prop_GeoPosLat'][i],
df['prop_GeoPosLng'][i]],
                            radius=radius_circle *
(df['prop_Price'][i]),
                            popup=('Titre    :' +
str(df['prop_Title'][i]) + '<br>'
                                   'Adresse :' +
str(df['prop_Address'][i]) + '<br>'
                                   'Prix net:' +
str(df['prop_Price'][i]) + '<br>'
                                   'Charges :' +
str(df['prop_SideCost'][i]) + '<br>'
                                   'Objet: ' +
str(df['prop_PropertyTypeText'][i]) + '%'
                                  ),
                            color=color_circle,
                            fill=True,
                            fill_opacity=0.7
                            ).add_to(carte)
    elif df['prop_Price'][i] == 0:
        radius_circle = 0.0001
        color_circle = "white"
        folium.CircleMarker([df['prop_GeoPosLat'][i],
df['prop_GeoPosLng'][i]],
                            radius=radius_circle *
(df['prop_Price'][i]),
                            popup=('Titre    :' +
str(df['prop_Title'][i]) + '<br>'
```

```
                                        'Adresse :' +
str(df['prop_Address'][i]) + '<br>'
                                        'Prix net:' +
str(df['prop_Price'][i]) + '<br>'
                                        'Charges :' +
str(df['prop_SideCost'][i]) + '<br>'
                                        'Objet: ' +
str(df['prop_PropertyTypeText'][i]) + '%'
                                       ),
                                color=color_circle,
                                fill=True,
                                fill_opacity=0.7
                                ).add_to(carte)
    elif df['prop_PropertyTypeText'][i] == "Wohnung":
        color_circle = "blue"
        radius_circle = .002
        folium.CircleMarker([df['prop_GeoPosLat'][i],
df['prop_GeoPosLng'][i]],
                                radius=radius_circle *
(df['prop_Price'][i]),
                                popup=('Titre    :' +
str(df['prop_Title'][i]) + '<br>'
                                        'Adresse :' +
str(df['prop_Address'][i]) + '<br>'
                                        'Prix net:' +
str(df['prop_Price'][i]) + '<br>'
                                        'Charges :' +
str(df['prop_SideCost'][i]) + '<br>'
                                        'Objet: ' +
str(df['prop_PropertyTypeText'][i]) + '%'
                                       ),
                                color=color_circle,
                                fill=True,
                                fill_opacity=0.7
                                ).add_to(carte1)

    elif df['prop_PropertyTypeText'][i] == "Gewerbeobjekt":
        color_circle = "red"
        radius_circle = .001
        folium.CircleMarker([df['prop_GeoPosLat'][i],
df['prop_GeoPosLng'][i]],
                                radius=radius_circle *
(df['prop_Price'][i]),
                                popup=('Titre    :' +
str(df['prop_Title'][i]) + '<br>'
                                        'Adresse :' +
str(df['prop_Address'][i]) + '<br>'
                                        'Prix net:' +
str(df['prop_Price'][i]) + '<br>'
                                        'Charges :' +
str(df['prop_SideCost'][i]) + '<br>'
                                        'Objet: ' +
str(df['prop_PropertyTypeText'][i]) + '%'
                                       ),
                                color=color_circle,
                                fill=True,
                                fill_opacity=0.7
                                ).add_to(carte2)
    else:
        color_circle = "black"
        folium.CircleMarker([df['prop_GeoPosLat'][i],
```

```
df['prop_GeoPosLng'][i]],
                              radius=radius_circle *
(df['prop_Price'][i]),
                              popup=('Titre    :' +
str(df['prop_Title'][i]) + '<br>'
                                     'Adresse :' +
str(df['prop_Address'][i]) + '<br>'
                                     'Prix net:' +
str(df['prop_Price'][i]) + '<br>'
                                     'Charges :' +
str(df['prop_SideCost'][i]) + '<br>'
                                     'Objet: ' +
str(df['prop_PropertyTypeText'][i]) + '%'
                                    ),
                              color=color_circle,
                              fill=True,
                              fill_opacity=0.7
                              ).add_to(carte)

"""chdir("data")

layer = folium.GeoJson(
    data=(open("swissBOUNDARIES3D_1_3_TLM_HOHEITSGEBIET.json",
"r").read()),
    name='borders'
).add_to(carte)

layer.add_to(carte)


chdir("..")
"""

carte.save('maCarte_Objets particuliers.html')
carte1.save('maCarte_Appartements.html')
carte2.save('maCarte_Locaux et Bureaux.html')
```