Projet de M8 : Traitement de données réelles.

DUPONT Nathan, TONDENIER Hugo, HAMDANI Ali

Résumé

L'objectif de ce projet est de mettre en œuvre les différentes notions étudiées lors des cours de M8, grâce à un jeu de donnés réel.

Table des matières

Table des figures

Liste des tableaux

1 Introduction

Nous voulions choisir un jeu de données dont l'étude nous paraissait utile et concrète. Le thème de l'économie nous a d'abord intéressé, mais après avoir rapidement étudié plusieurs jeux de données financiers, notamment sur des cours en bourse, nous avons préféré chercher un jeu de données dont les variables ne dépendent pas du temps et sont plus facilement analysables.

Finalement nous avons choisi un sujet assez proche, qui reste dans le domaine économique : la location d'appartements à l'échelle des États-Unis. Cette étude pourrait permettre d'identifier les profils de location de logements les plus fréquents, ce qui peut être important pour les investisseurs et promoteurs immobiliers et les gouvernements locaux. En effet, nous trouvons très intéressant d'étudier les caractéristiques qui définissent le prix de location d'un bien immobilier aux États-Unis. De plus, le jeu de données contient la latitude et la longitude de chaque appartement, ce que nous avons trouvé très intéressant pour effectuer une analyse géographique et cartographier nos données. Le but de ce projet est de répondre aux questions suivantes :

- Est - il possible de prédire avec fiabilité le prix d'un appartement aux États-unis ?
- Quelles sont les caractéristiques qui influent le plus sur le prix d'un appartement ?
- Quel est l'impact des facteurs géographiques sur la location d'appartements ?

2 Description des données

2.1 Récupération et tri du jeu de données

Nous avons trouvé ce jeu de données sur le site de l'université d'Irvine (`https://archive-beta.ics.uci.edu/dataset/555/apartment+for+rent+classified`). Le jeu de données comporte 100 000 observations d'appartement et 22 variables. Nous avons dû effectuer un tri dans lequel nous avons conservé les 13 variables qui nous semblaient pertinentes ainsi que les 99 820 observations utilisables.

En effet, certaines observations étaient incomplètes et donc inexploitables. Les variables que nous avons utilisées sont :

Variable	Type
Équipements	Qualitative
Animaux autorisés	Qualitative
Ville	Qualitative
État	Qualitative
Source	Qualitative
Frais	Qualitative(binaire)
Photo	Qualitative(binaire)
Salles de bain	Quantitative
Chambres	Quantitative
Prix	Quantitative
Surface	Quantitative
Latitude	Quantitative
Longitude	Quantitative

TABLE 1 – Type des variables utilisées

2.2 Description des données

Dans ce projet, la variable qui nous paraît la plus importante et sur laquelle nous allons nous concentrer est le prix des appartements. Notre étude va donc se faire en fonction de cette variable. Nous avons étudié les différentes variables selon leur impact sur le prix de location des appartements. Nous allons donc tenter d'établir des relations entre les variables et la variable prix. Dans un premier temps, décrivons les différentes variables de notre jeu de données.

2.2.1 Première visualisation des données.

Pour s'approprier les données que nous possédions et mieux comprendre chaque variable, nous avons commencé par observer et analyser les moyennes, médianes et variances des variables quantitatives.

Variables	Moyenne	Médiane	Variance
Salles de Bain	1.445	1.0	0.299
Chambres	1.728	2.0	0.560
Prix	1 527.4	1 350	808 346
Surface	956.2	900.0	133 362

TABLE 2 – Descripteurs Monovariables

Les données concernant les salles de bain et les chambres nous indiquent que les biens sont surement plus des appartements que des maisons. En effet, le

nombre moyen de chambres est de 1.73, et le nombre moyen de salles de bain est de 1.445. Cependant, au vu de la surface moyenne, la taille du terrain est sûrement prise en compte (car la surface est très grande par rapport au reste des informations). De plus, nous remarquons une variance très élevée pour les variables prix et surface. Nos données sont donc très dispersées par rapport à la moyenne. Cela est plutôt cohérent au vu du nombre d'appartements du jeu de données et de la zone géographique étudiées.

Nous avons ensuite voulu connaître le nombre d'appartements selon chaque ville :

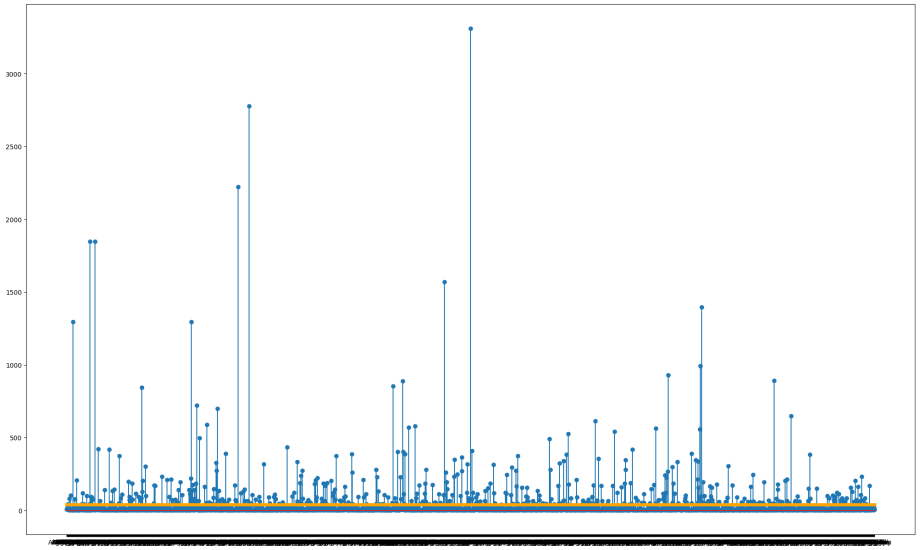

FIGURE 1 – Nombre appartements par ville (moyenne en orange)

Le jeu de données contient 2557 villes, et le nombre moyen d'appartements en location par ville est de 39. Nous voyons donc que certaines grandes villes possèdent beaucoup plus d'appartements en location que la moyenne. Il faudra alors prendre en compte le fait que certaines villes auront une plus grande influence dans notre étude. Le nombre maximum d'appartements dans une ville est alors de 3311, à Los Angeles.

Ensuite, nous avons tracé le prix moyen de location par ville :

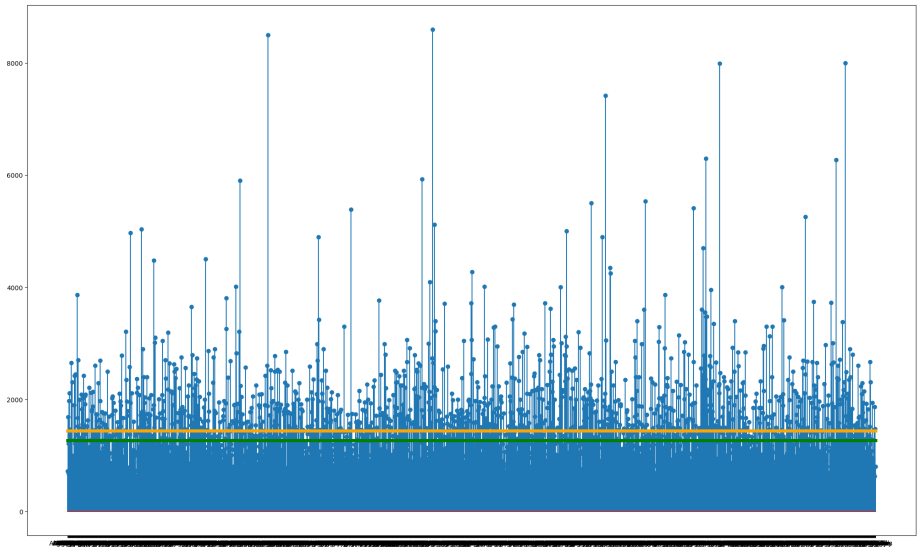

FIGURE 2 – Prix moyen par ville (moyenne en orange, médiane en vert)

Voici un graphique plus lisible, sur lequel on a affiché uniquement les villes possédant plus de 100 appartements :

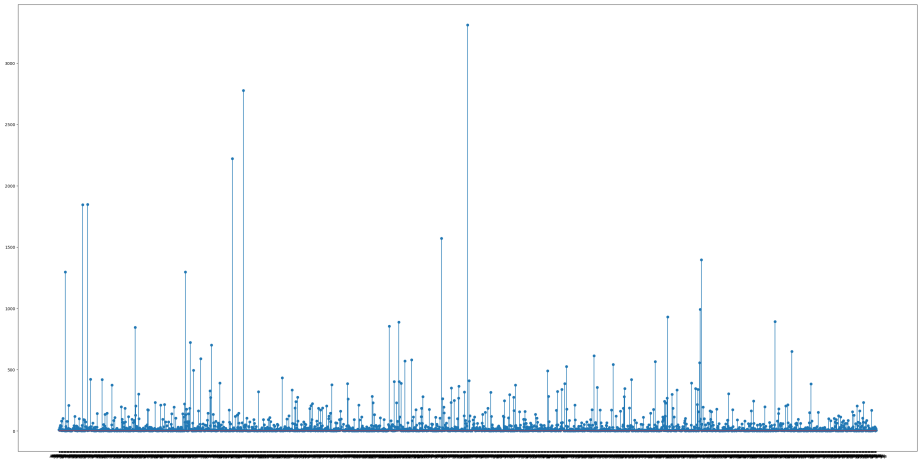

FIGURE 3 – Prix moyen des villes avec plus de 100 appartements

Ces deux graphiques rendent bien compte de la variance très élevée de la variable prix. En effet, de très nombreuses villes affichent un prix moyen beaucoup plus élevé que la moyenne. Beaucoup de villes ont aussi un prix moyen de location très faible (la moitié des villes sont sous la médiane, tracée en vert sur le graphe).

Nous voulions ensuite voir si le prix et la surface étaient liés. Nous avons donc tracé, pour chaque État, dans un premier temps le prix moyen des appartements, et, dans un second temps, le prix par mètre carré (prix / surface) :

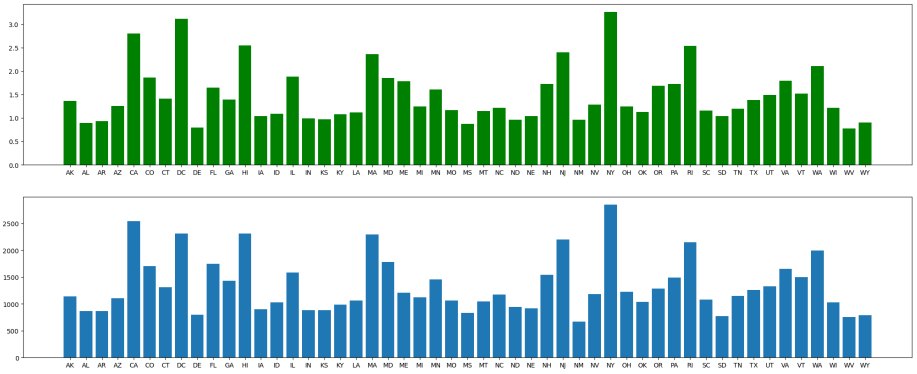

FIGURE 4 – Prix moyen par État (bleu) et prix par mètre carré moyen (vert)

Nous voyons alors que la tendance semble être la même pour les deux graphiques. Les États les plus chers sont New-York, District de Columbia, et la Californie dans les deux cas, et les moins chers Delaware et le Nouveau - Mexique.

NB : Nous pouvons noter que le prix par mètre carré est assez faible (entre 0.5 et 3$/m^2$). Notre explication est la suivante : tout d'abord, la location étant mensuelle, il est plutôt logique que le prix soit relativement faible, puisqu'il est payé tous les mois. De plus, une supposition que nous avons faite (même si ce n'est pas précisé dans la description du jeu de donnée), est que la surface de l'appartement comporte également le terrain de celui-ci, car cette surface est pour la plupart très grande pour un appartement. Rappelons que la moyenne de la surface est de $956m^2$ et la médiane $900m^2$.

2.2.2 Analyse des coefficients de corrélation

Tout d'abord, nous avons décidé d'analyser les potentielles corrélations et liens entre nos variables (par couple de variables). Pour ce faire, nous avons analysé la matrice de corrélation de nos variables. Cette étude n'est pertinente et utilisable que pour les variables quantitatives de notre jeu de données. Nous avons donc calculé les coefficient de corrélations entre nos variables pour déterminer si certaines variables pourraient possèder un lien entre elles, de manière assez importante.

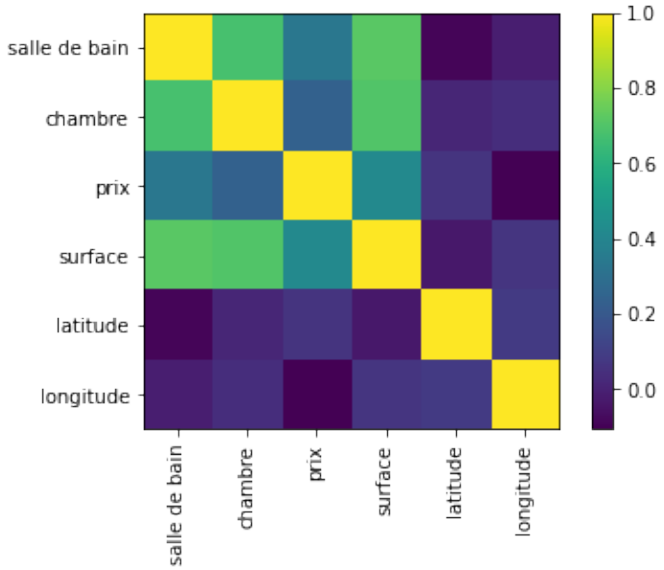

FIGURE 5 – Matrice de corrélation des variables quantitatives.

Nous pouvons noter que les variables les plus corrélés entre elles sont le nombre de salles de bain avec le nombre de chambres, et avec la surface, ce qui semble relativement évident. Cependant, cela ne nous apporte pas d'informations utiles à notre étude. Mais en réalité, la variable pour laquelle les corrélations nous intéresse le plus est le prix, voici ses coefficients de corrélation avec les autres variables :

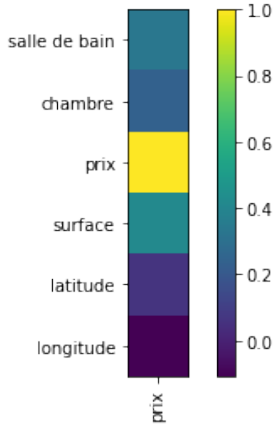

FIGURE 6 – Coefficients de corrélation des variables avec le prix

Comme nous pouvons le voir, la variable surface est celle qui possède le plus

de lien avec le prix parmi toutes les variables quantitatives. Malheureusement son coefficient de corrélation avec le prix est de 0.412, ce qui est trop faible et n'est pas suffisant pour en tirer de réelles conclusions.

Complétons notre analyse afin de trouver de potentiels liens entre nos variables, notamment avec le prix.

2.2.3 Analyse de la distance du χ_2

Avant de commencer cette analyse, notons que celle-ci n'est pas forcément la plus pertinente, car le prix est une variable quantitative et non qualitative. Cependant, comme nous possédons un grand nombre de variables qualitatives, nous avons décidé de quand même analyser cette distance avec le prix, dans le but de voir les potentiels liens entre le prix et nos variables, tout en sachant que la fiabilité de ces résultats n'est pas garantie et que l'on devra s'en méfier.

Afin de voir quelles variables sont les plus liées au prix des appartements, nous avons calculé la distance du χ_2 entre le prix et chacune des autres variables. Pour ce faire, nous avons créé un tableau de contingence entre chaque variable et le prix. Nous avons ensuite calculer la distance du χ_2 pour toutes ces variables, et nous avons obtenu les résultats suivants :

Variable	Salle de Bains	Chambres	Frais	Photos	Animaux
D_{χ_2} avec le prix	777 953	917 957	8 447	13 630	20 467
Surface	Ville	Etat	Latitude	Longitude	Source
1 901 193	12 606 887	393 413	30 487 403	30 633 654	121 531

TABLE 3 – Distance du χ_2 entre le prix et les autres variables

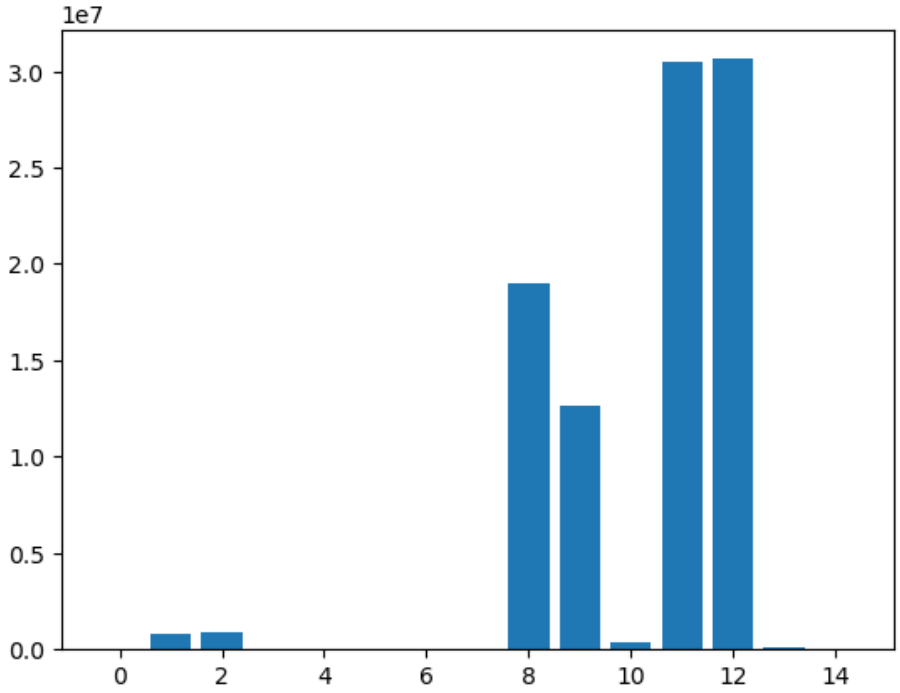

FIGURE 7 – Distance du χ_2 entre le prix et les autres variables

Notons que l'on n'a pas utilisé la variable équipements pour calculer la distance du χ_2. En effet, cette variable étant représentée sous forme d'une chaîne de caractères regroupant tous les équipements pour chaque observation, le nombre de modalités pour la variable équipement aurait été trop élevé, et la distance du χ_2 calculée non représentative de l'indépendance ou le lien des variables : les équipements auraient été très peu liés au prix, sans refléter forcément la réalité.

NB : Nous notons que les scores de χ_2 sont plutôt importants pour toutes les variables (de l'ordre de 10^3 à 10^7). Ceci peut s'expliquer par le fait que, d'une part, la variable prix possède beaucoup de modalités, et, comme la distance du χ_2 fait une somme pour chaque modalité, cela revient à sommer beaucoup de termes, et augmenter la valeur finale. D'autre part, on multiplie cette somme par le nombre d'observations n, et, n étant à 99 820, cela augmente considérablement la valeur de la distance du χ_2.

Nous remarquons que la distance du χ_2 la plus grande entre le prix et les autres variables est la longitude. Cela signifie donc que le prix et la longitude sont les variables les moins indépendantes entre elles, et donc les plus liées. Avoir un appartement sur les côtes américaines influence donc le prix plus fortement que les autres variables. Cependant, nous voyons aussi que le prix de location à également l'air dépendant de la surface. Nous allons donc, dans un premier temps, pour la régression, essayer de faire une régression linéaire simple avec la

surface, avec le prix comme variable à expliquer.

2.2.4 Analyse en Composante Principale

Nous avons également fait une ACP, afin de savoir quelle variable de notre jeu de données résume, influence et à le plus d'impact sur celui-ci. Pour faire cette ACP, nous devons créer la matrice de variance/covariance (à un facteur n près), et il était de ce fait indispensable d'avoir des variables quantitatives. Nous avons donc fait l'ACP avec toutes les variables quantitatives : le nombre de chambres, salles de bain, le prix, la surface, la latitude et la longitude. Nous avons également transformer nos variables qualitatives en variables quantitatives en les codant avec des chiffres. Malgré tous les inconvénients que ce choix implique (détails un petit peu plus bas, section 3.1), nous trouvions quand même plus intéressant d'avoir le plus de variables possibles dans notre ACP, pour que l'on sache parmi toutes laquelle résume le mieux les informations du jeu de données.

Finalement, nous obtenons les résultats suivants :

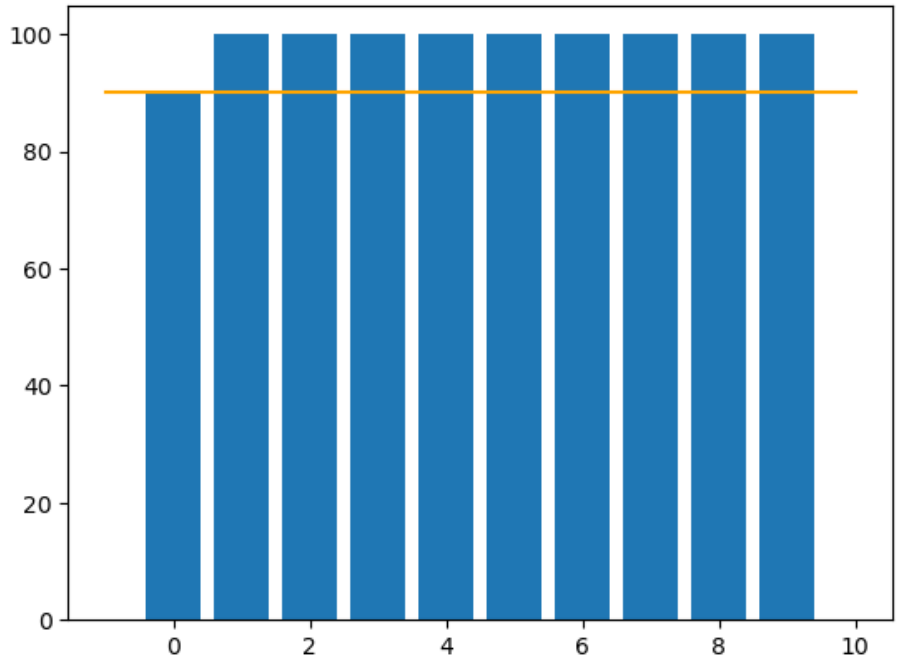

FIGURE 8 – Pourcentage cumulé d'information expliquée par l'ACP (droite orange : 90% de l'information expliquée)

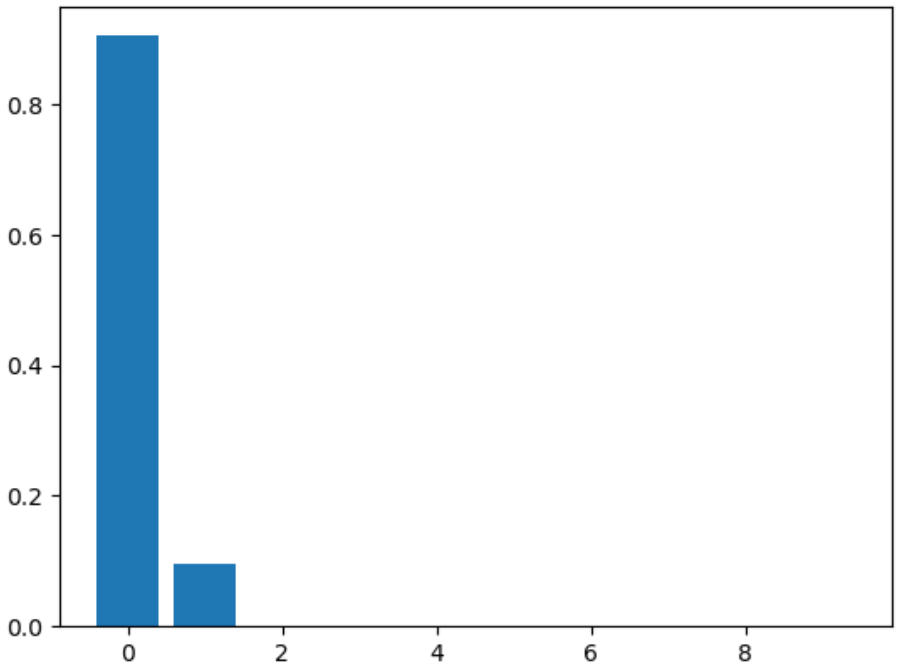

FIGURE 9 – Proportion d'information expliquée par chaque vecteur propre

Nous voyons qu'un seul vecteur propre suffit à expliquer plus de 90% de l'information. Nous dépassons les 99% avec seulement les 2 plus grandes valeurs propres.

Finalement, nous avons calculé la contribution de chaque variable au premier axe vectoriel :

Variable	Salle de bain	Chambres	Prix	Surface	Latitude
Contribution au 1er axe vectoriel (valeur absolue)	2.09×10^{-4}	2.03×10^{-4}	9.8×10^{-1}	1.7×10^{-1}	1.6×10^{-3}
Variable	Longitude	Frais	Photos (recodée)	Etat (recodée)	Source (recodée)
Contribution au 1er axe vectoriel (valeur absolue)	1.5×10^{-3}	1.8×10^{-6}	3.8×10^{-6}	3.0×10^{-3}	4.8×10^{-6}

TABLE 4 – Contribution de chaque variable au 1er axe vectoriel de l'ACP

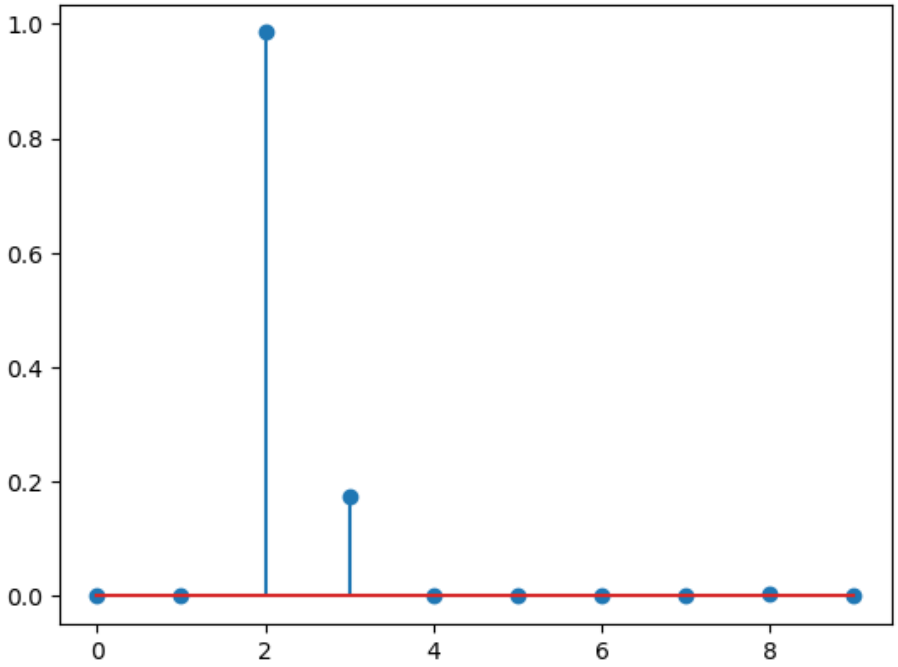

FIGURE 10 – Contribution de chaque variable au 1er axe vectoriel de l'ACP

Nous remarquons que c'est le prix qui influence largement (plus de 5 fois plus que la deuxième) le plus le premier axe vectoriel.

Nous avons donc en quelque sorte justifié le fait d'axer notre étude sur le prix

des appartements, car c'est bien cette variable qui contribue le plus à donner des informations sur les observations de notre jeu de données.

2.2.5 Affichage de la carte des États unis selon le prix.

Nous allons maintenant utiliser la latitude et la longitude des observations, afin de savoir si la localisation d'un appartement a une influence sur le prix de celui-ci (ce dont on peut se douter à priori, d'autant plus avec les valeurs de χ_2). Nous allons donc afficher les localisations des appartements , grâce à leurs coordonnées. Les points devront être de différentes couleurs selon le prix des appartements. Nous nous sommes alors posés la question suivante : comment séparer les appartements en différentes gammes de prix. En effet, la variable prix possédant beaucoup de modalités (3477 prix différents), afficher une couleur différente pour chaque prix n'aurait pas été assez visible et analysable sur la carte. Nous avons donc fait le choix de séparer les prix d'appartements en quatre groupes différents, selon les quatre quartiles de la variable prix. Ce choix nous permettait notamment d'avoir autant de points de chaque couleur, pour un meilleur rendu, plus lisible.

Prix	1014	1350	1795
Quartile	Q1	Q2 = M	Q3

TABLE 5 – Valeur des quartiles, variable « prix »

Pour afficher les appartements et bien rendre compte de la géographie des États-unis (même si le grand nombre d'observations permettait déjà de bien reconnaître les frontières du pays), nous avons utiliser le module Cartopy (Voir annexe). Un avantage d'utiliser ce module est notamment qu'il est adapté à Matplotlib, ce qui nous a permis de pouvoir superposer nos points et la carte de manière lisible et facile. La première carte que nous avons obtenue est la suivante :

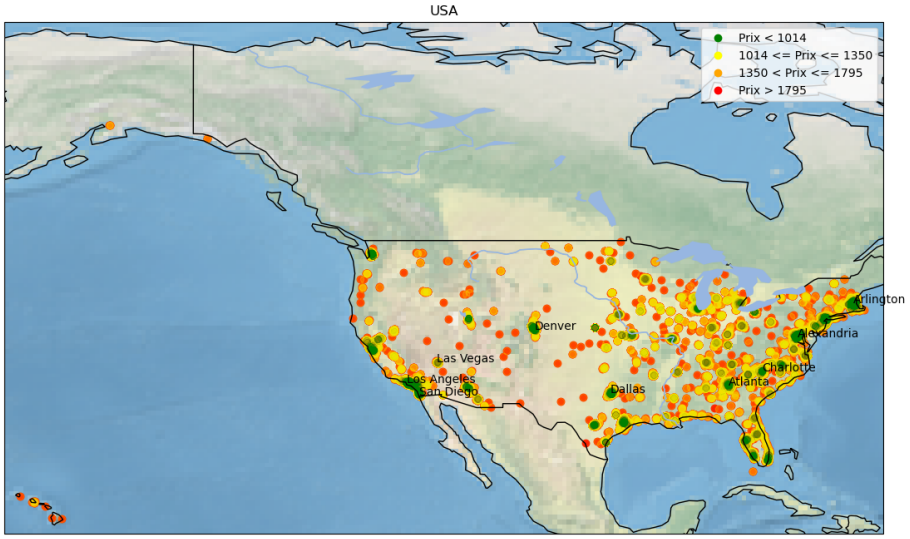

FIGURE 11 – Prix des appartements pour toutes les observations

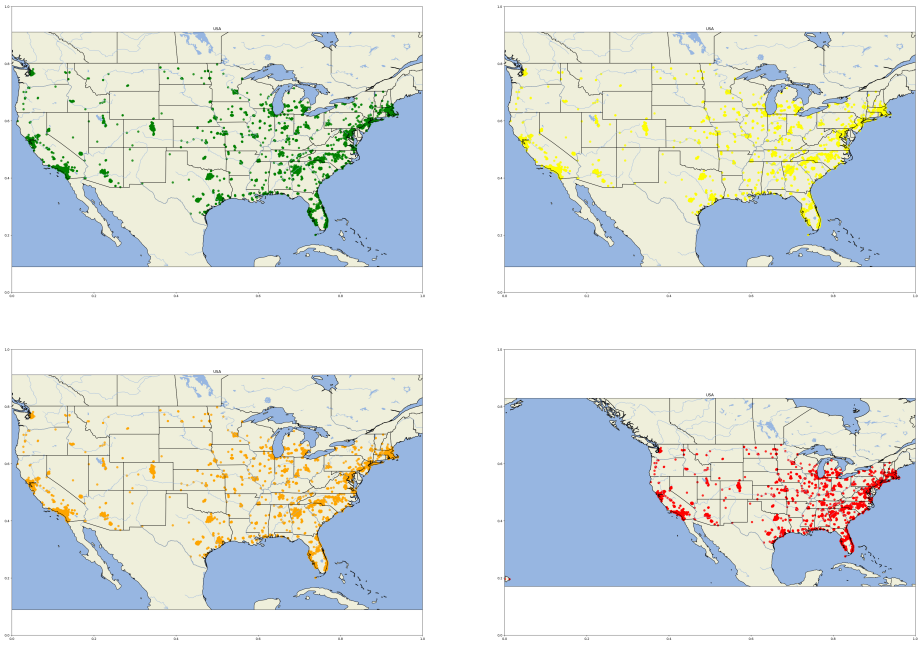

FIGURE 12 – Prix des appartements par quartile

Cependant, nous remarquons que, même en essayant de jouer avec la transparence des points, ceux-ci se superposent beaucoup (ce qui est normal en voulant

afficher 100 000 sur une relativement “petite” zone), rendant une analyse finale assez compliquée. Nous avons alors remarqué que la superposition est dûe au fait que, pour une même ville, tous les appartements possèdent la même latitude et longitude (elles ne sont pas assez précises). Donc si un point dans Q1 est affiché en premier pour une certaine ville, et qu'un autre point de Q2 est affiché pour la même ville, alors le deuxième recouvrera le premier, ce qui explique notre rendu. Pour pallier à ce problème, nous avons choisi de calculer la moyenne des prix pour chaque ville, puis d'afficher un seul point pour chaque ville, la couleur dépendant du prix moyen. Nous avons également affiché le nom des 10 villes possédant le plus de locations d'appartements.

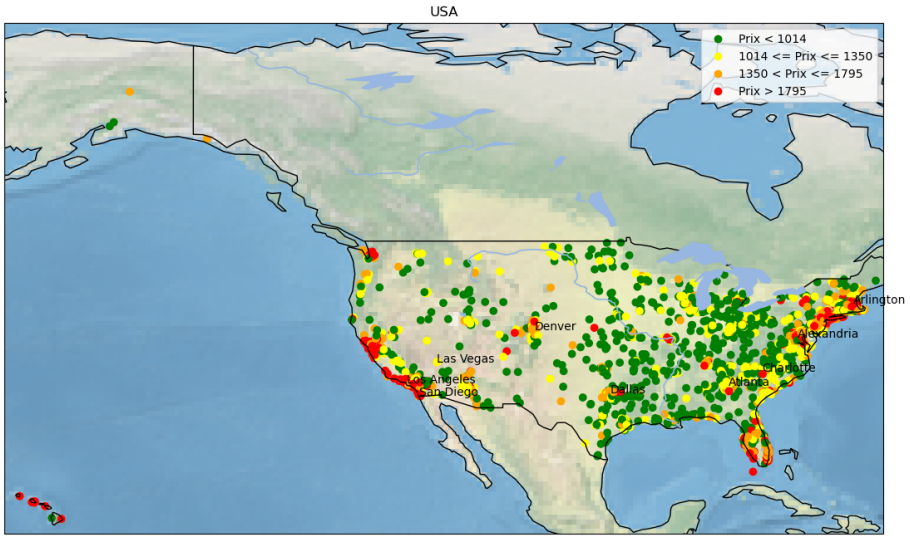

FIGURE 13 – Prix moyen pour chaque ville

Nous remarquons alors la tendance suivante : Cette carte nous montre bien que les locations d'appartements sont beaucoup plus chères sur les villes des côtes des États-Unis. En effet, ces zones sont les plus attractives des États-unis alors que les zones centrales sont plus abordables car ce sont des zones agricoles avec moins de grandes villes. Nous pouvons d'ailleurs voir sur l'arrière-plan de la carte que les zones désertiques (les plus claires) comportent moins d'appartements en locations et que ces derniers sont moins chers. Pour essayer de faire une analyse un petit peu plus fine, nous avons affiché également le prix moyen pour chaque État, suivant le même principe de code couleur. Nous remarquons alors que les États les plus chers sont : New-York, District de Columbia, et la Californie, et les moins chers : Delaware et le Nouveau - Mexique, ce qui confirme le graphique de la première partie, avec le prix moyen de chaque État.

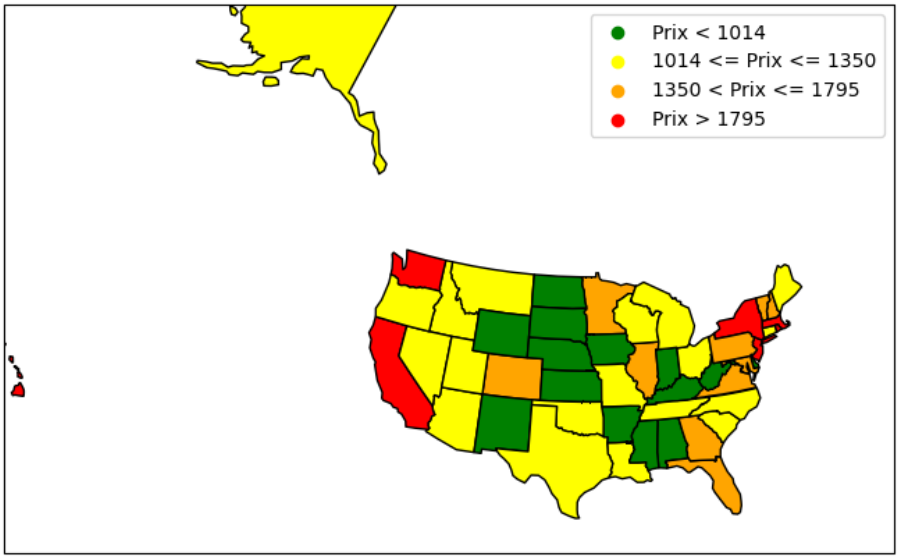

FIGURE 14 – Prix moyen par Etat

Finalement, pour regrouper et rendre plus lisible toutes les informations en lien avec la localisation des appartements par ville, nous avons également installé Folium (Voir annexe) , qui nous a permis de créer une carte dynamique sur laquelle on peut se déplacer et cliquer sur les villes. Nous avons donné, pour chaque ville, le prix moyen de la location, la surface moyenne, le nombre de chambres et salles de bain moyen, ainsi que le nombre d'appartement en location dans cette ville. Nous voulions afficher toutes les villes, mais le nombre étant trop conséquent et rendant le résultat trop peu fluide, nous avons dû réduire ce plot aux 996 villes avec le plus d'appartement en location. Nous avons fait le choix d'utiliser le module Folium pour rendre nos données plus lisibles et bien résumer les informations clés de chaque ville en palliant à la grande quantité de données à afficher. Nous avons pu, grâce à ce module, généré le code HTLM suivant permettant de visualiser la carte.

Le résultat est disponible via le fichier, nommé « Carte Projet M8 », disponible dans le ZIP avec ce rapport.

2.2.6 Analyse des équipements.

Enfin, nous avons décidé d'étudier la première variable de notre jeu de données : les équipements ; afin de savoir si ceux-ci étaient reliés à d'autres variables. Pour faire ceci, nous avons dû dans un premier temps créer une nouvelle matrice avec ces équipements. En effet, ceux-ci ont été importés comme une chaîne de caractère par observations, et il a alors fallu séparer chaque équipement au sein d'une même ligne.

Equipements
'Gym,TV'
'Club-house,Fireplace,Hot Tub,Patio/Deck,View'
'Cable or Satellite,Clubhouse,Dishwasher,Elevator,Garbage Disposal,Gated,Gym,Hot Tub,Internet Access,Patio/Deck,Pool,Refrigerator,Tennis'

TABLE 6 – Variable « équipements » avant la transformation (une colonne avec une chaîne de caractères par observation)

Observations	équipement 1	...	équipement i	...	...	...	équipement k
1	'Gym'	'TV'					
2	'Club-house'	'Fireplace'	'Hot Tub'	'Patio/ Deck'	'View'		
3	'Cable or Satellite'	'Club-house'	...	...	...	...	'Tennis"

TABLE 7 – Variable « équipements » après transformation

Une fois ce travail fait, nous avons déjà pu extraire les équipements disponibles dans les locations, ainsi que leur effectif dans ce jeu de données :

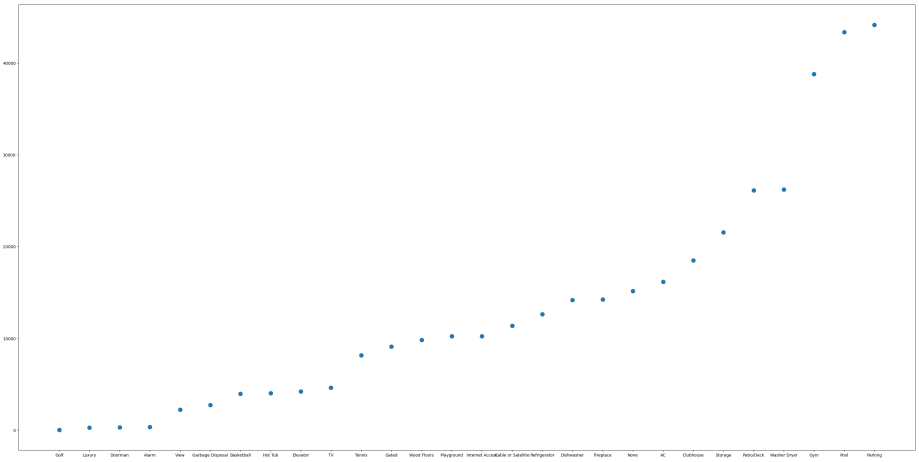

FIGURE 15 – Effectifs des équipements de notre jeu de données

Les équipements les plus présents sont alors : les parkings, les piscines et les salles de sport, avec respectivement 44 149, 43 374 et 38 795 occurrences. Nous avons ensuite affiché la localisation de différentes sortes d'équipements sur la carte, afin d'observer si certains étaient plus ou moins présents selon la géographie du pays :

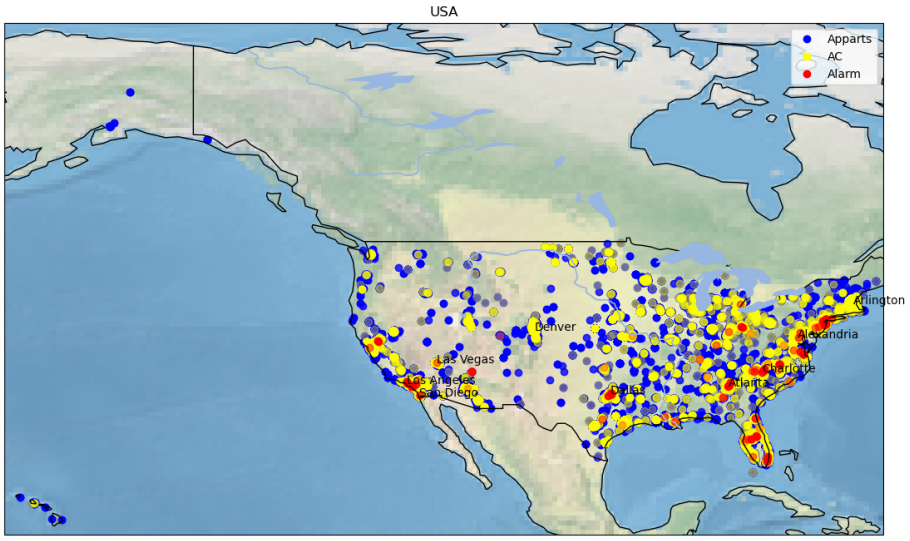

FIGURE 16 – Equipement en fonction de la géographie

Voici par exemple une carte où sont affichés tous les appartements (en bleu), les appartements possédant un système de climatisation (en jaune) et une alarme (en rouge). Nous pouvons déduire de cette observation que les alarmes sont regroupées en plusieurs tas, qui sont les plus grandes villes : il paraît cohérent que les appartements dans les grandes villes aient plus besoin d'une alarme que ceux à la campagne. Nous pouvons aussi voir que les climatisations sont présentes en majorité sur les côtes. Les villes proches des côtes, avec la plage et la chaleur, peuvent avoir plus besoin d'air conditionné que dans certaines autres régions.

3 Régression

Pour cette partie régression, notre objectif final est de créer un modèle de prédiction du prix d'un appartement selon certaines de nos variables, et cela adapté à chaque ville de notre jeu de données. Nous allons également faire une régression multiple sur nos données, ainsi qu'une Forward Selection afin de savoir quelle est la meilleure combinaison de variable alliant régression fidèle et efficacité de calcul.

3.1 Adaptation du jeu de données pour une régression multiple exploitable.

Tout d'abord, nous avons voulu effectuer une régression simple entre le prix et la surface, car, grâce aux résultats précédents, ces 2 variables nous semblaient corrélées.

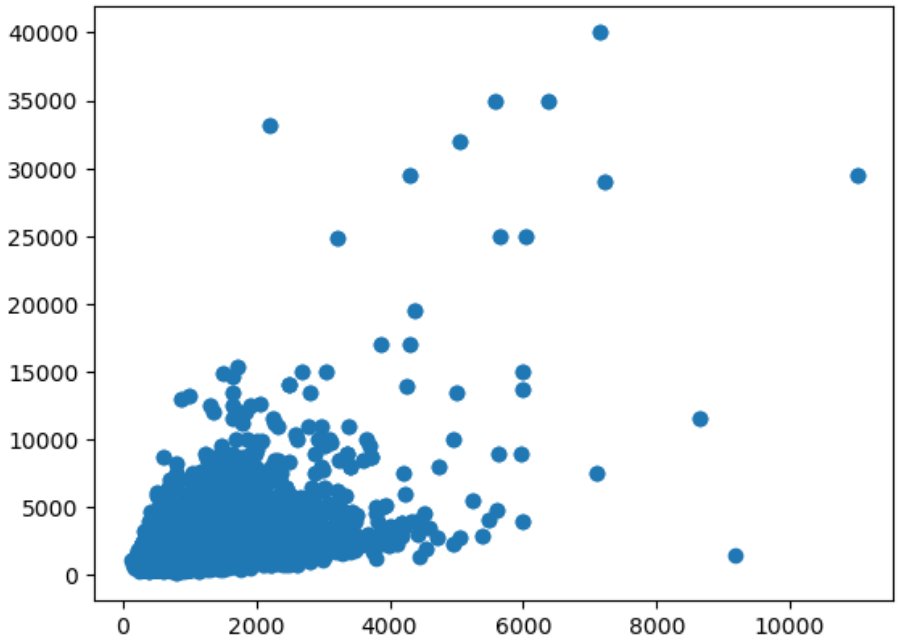

FIGURE 17 – Prix en fonction de la surface

On peut visuellement s'attendre à un mauvais R^2 car on est très éloignée d'une droite et, effectivement, le R^2 est de 0.191. Notre hypothèse n'étant pas assez bonne, nous avons décidé de faire une régression multi-variables avec uniquement les variables quantitatives, (car il y a des contraintes à utiliser des variables qualitatives pour faire une régression : encoder ces variables avec des chiffres n'est pas forcément représentatif de l'information contenue dans une variable, surtout quand il y a plus de 2 modalités) dans le but d'augmenter l'information pour avoir un meilleur R^2.

Cependant, le résultat n'était pas encore bon car nous avons obtenu un R^2 de 0.24. Selon nous, le problème venait du fait que l'information des variables qualitatives est beaucoup trop importante pour être négligée. En effet, la ville et l'État sont par exemple des facteurs nécessaires et déterminants pour trouver à quel prix louer un appartement. Nous avons donc décidé de faire du One Hot Encoding pour les villes et les États : nous avons transformé chaque ville et chaque État en variables binaires. Nous avions donc, dans les variables, une

colonne pour chaque ville et chaque État, avec un 1 si l'appartement fait partie de cette ville ou État et 0 sinon.

Cependant, le R^2 était encore trop peu satisfaisant pour pouvoir faire une prédiction intéressante par la suite. De plus, le One Hot Encoding a impliqué un problème supplémentaire lors du calcul de la régression : notre matrice contenant plus de 99 000 observations et 2600 variables, l'utilisation d'une fonction solve pour le calcul de H notamment prenait beaucoup trop de temps en calcul.

Nous nous sommes alors questionnés sur la raison de la faible corrélation linéaire de nos données, et comment faire en sorte de résoudre ce problème. Une réponse à ce problème a été pour nous le trop grand nombre d'observations dans notre jeu de données : le nombre et la proportion de données aberrantes étaient trop importantes avec ces 99 800 observations.

Pour résoudre ce problème, nous avons alors procéder de la manière suivante : Nous avons regardé si certaines villes posaient problèmes en séparant nos données pour chaque ville, et en effectuant une régression simple entre la surface et le prix pour chacune d'elles. Nous avions donc un coefficient R^2 pour chaque ville. Nous nous sommes alors rendus compte que, dans certaines villes, le R^2 était problématique car très proche de 0. Nous avons ensuite restreint notre étude aux villes possédant plus de 100 appartements pour que la régression soit un minimum représentative. Ensuite, nous n'avons gardé que les villes avec un R^2 supérieur à 0.60, pour que les données soient assez corrélées pour être interprétées.

A partir de ces informations, nous avons réduit notre jeu de données initial avec seulement ces observations. Nous avions donc une matrice de taille 6655×5, qui nous était certaine de posséder une certaine corrélation et donc allait être exploitable.

3.2 Régression multiple.

Nous avons donc fait une régression multiple entre le prix et les variables quantitatives : salle de bain, chambre, surface, latitude et longitude ; uniquement sur les villes préalablement sélectionnées. Grâce à cette technique, notre modèle de régression linéaire s'est amélioré et le R^2 a atteint 0.483 ce qui nous parait correct comparé au coefficient R^2 calculé sur l'ensemble du jeu de données. Pour illustrer l'état initial du modèle, voici un graphique du prix réel en fonction de nos prédictions de prix :

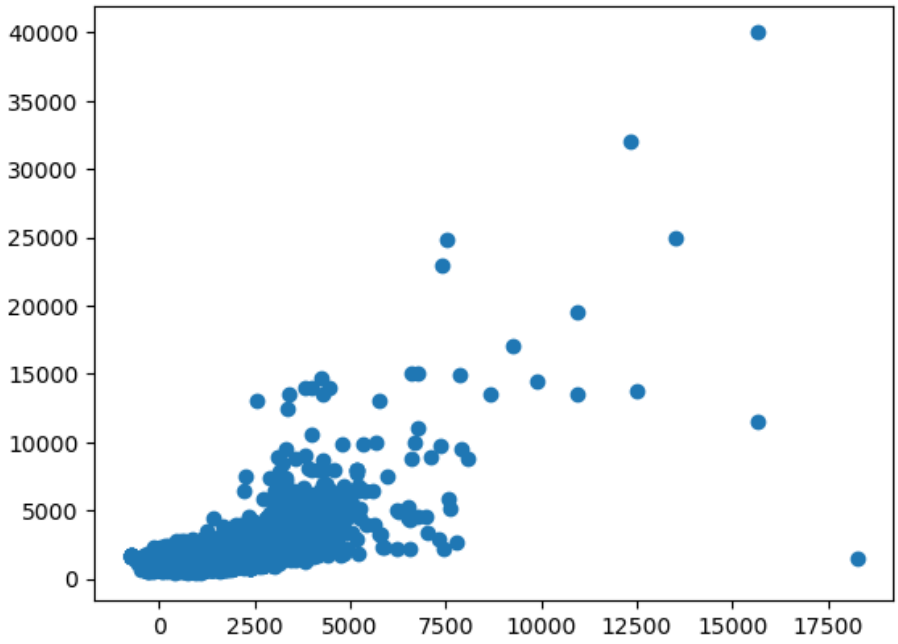

FIGURE 18 – Prix en fonction des prédictions de prix de notre modèle linéaire

Comme nous pouvons le voir sur ce graphique, nos prédictions sont surtout fausses dans les cas de prix élevés. Cependant, pour la majorité des points non aberrants, nous pouvons quand même noter une tendance linéaire claire.

Nous avons donc décidé de poursuivre avec cette régression et de faire du One Hot Encoding avec les villes pour améliorer le coefficient R^2 de notre modèle. Grâce à cela, nous avons obtenu un R^2 de 0.651.

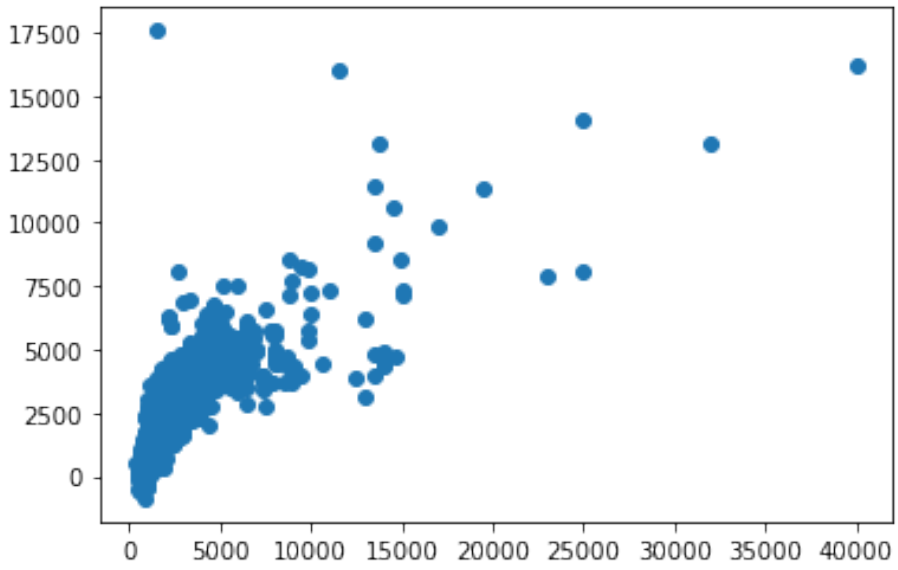

FIGURE 19 – Prix réel en fonction des prédictions de prix du modèle linéaire (avec One Hot Encoding).

Nous voyons que certains points posent encore problème : nous allons donc effectuer un diagnostic sur les observations et sur les variables afin de supprimer les points problématiques et d'améliorer notre modèle. Nous avons donc commencé par faire un diagnostic sur les observations. Pour ce faire, nous avons calculé les résidus, leviers et contributions pour chaque observation. Nous avons fait le choix de supprimer les observations avec une contribution supérieure à $\frac{4}{n}$. En effet, si une observation possède une trop forte contribution, elle impactera notre régression trop négativement.

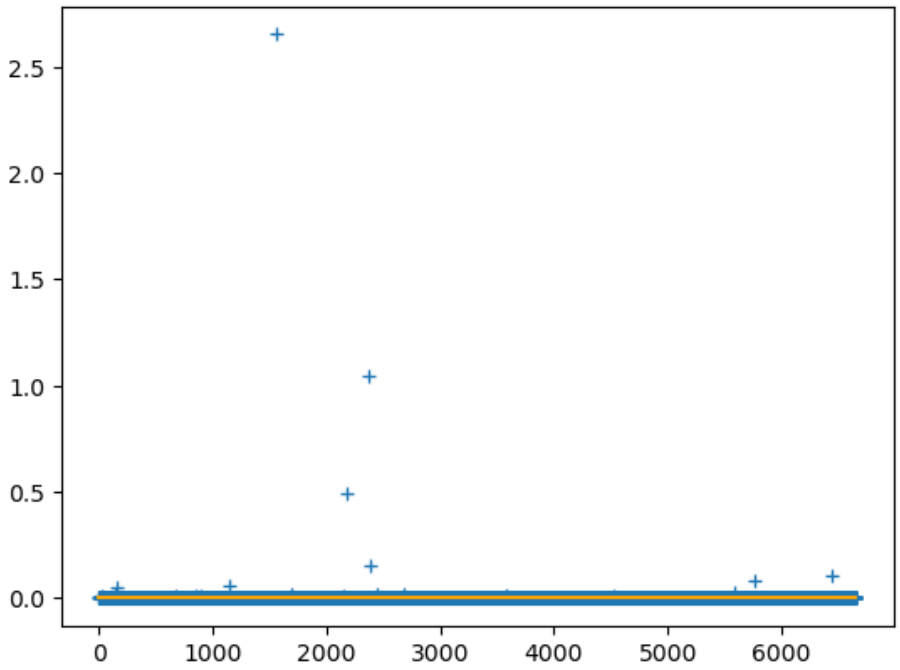

FIGURE 20 – Contributions des observations à la régression

Une fois ce tri effectué, il reste 6537 observations. Cette méthode a donc supprimé 118 observations qui avaient trop d'impact sur le modèle. Notre modèle de régression linéaire possède maintenant un coefficient $R^2 = 0.803$.

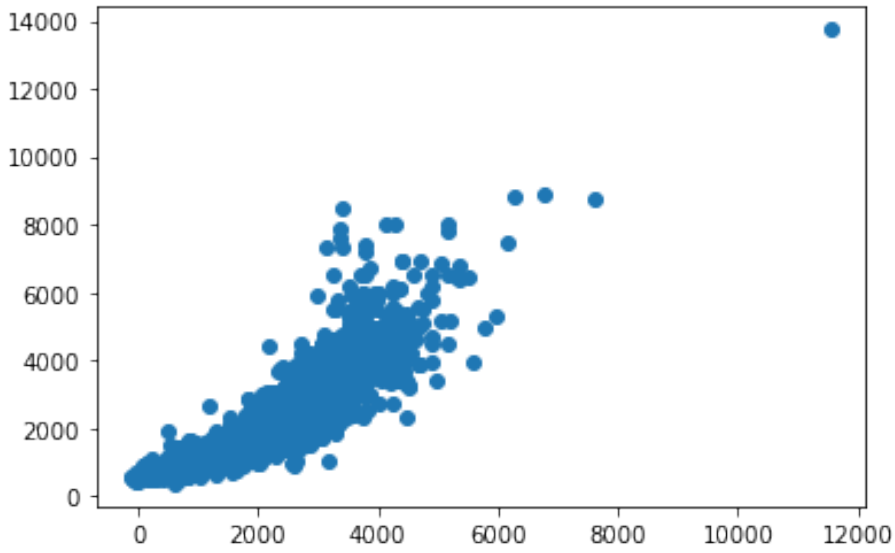

FIGURE 21 – Prédictions de prix du modèle de régression linéaire final en fonction du prix

Finalement, nous avons effectué une Backward Selection afin de trouver la meilleure combinaison permettant d'allier efficacité de calcul, en ne prenant pas toutes les variables, et précision dans la régression faite. Nous avons effectués une Backward Selection plutôt que de calculer tous les C_p de Mallows pour chaque combinaison de variables. En effet, avec l'utilisation du principe du One Hot Encoding, nous avons beaucoup de variables et le nombre de combinaisons explose.

Le principe de la Backward Selection est le suivant :

On commence l'algorithme avec toutes les variables. A partir du deuxième tour, à chaque tour de boucle, on calcule le C_p de Mallows de chaque sous ensemble $X^{(0)}$ privé d'une variable différente, on supprime la variable donnant le plus petit C_p de Mallows et on garde cette valeur de C_p, car le sous ensemble associé est le plus performant au sens des C_p de Mallows. A la fin, on compare les C_p de Mallows trouvés à chaque tour de boucle, et on sélectionne l'ensemble avec le plus petit C_p. En effet, cet ensemble sera celui alliant meilleure information transmise et nombre minimum de variables utilisées.

Dans notre cas, la plus faible valeur de C_p est celle trouvée au 2ème tour de boucle, avec un score de 26.1 :

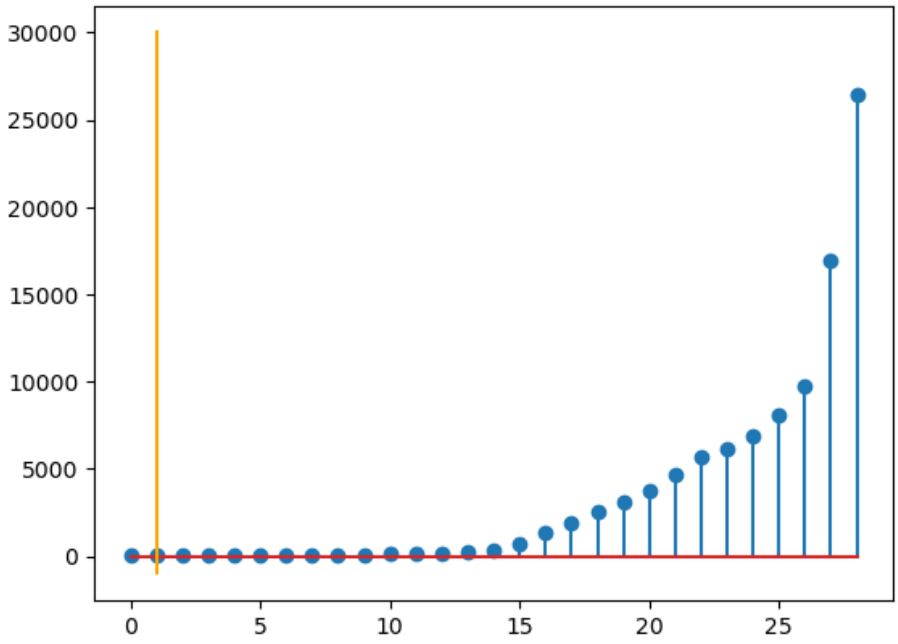

FIGURE 22 – Cp de Mallows calculés selon la Backward Selection

Cela signifie donc que le meilleur ensemble est celui comportant toutes les variables sauf les deux suivantes : Braintree et Knoxville. Sans ces deux variables (qui sont des colonnes de villes avec le One Hot Encoding), nous obtenons un score de R^2 de 0.803. Nous gardons donc la même précision dans la régression, avec deux variables en moins. La Backward Selection a donc été efficace.

Pour terminer, nous pouvons voir que le graphique de nos prédictions avec prix a une tendance assez claire en x^2, voire en x^4. On pourrait proposer un modèle de correction qui, à partir des prédictions de la première régression au carré et à la puissance 4, effectue une nouvelle régression pour se rapprocher encore plus du prix. Cela pourrait potentiellement améliorer les prédictions finales, mais la contrepartie est d'augmenter le sur-apprentissage de notre modèle.

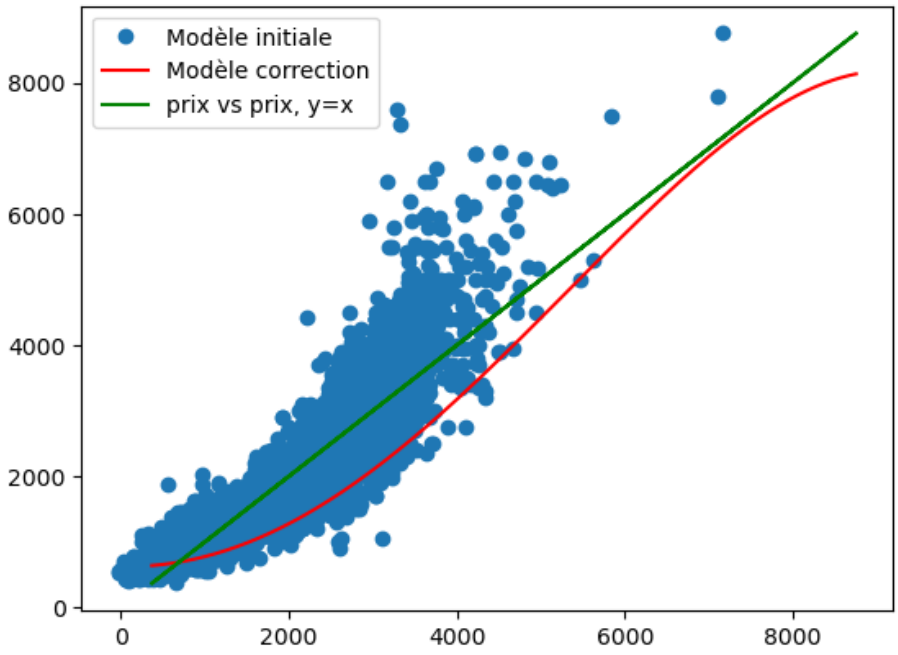

FIGURE 23 – Représentation du modèle de correction de degré 4.

Grâce à ce modèle de correction, nous avons réussi à atteindre une valeur maximale de R^2 de 0.854 après avoir appliqué successivement les deux modèles. Nous pouvons observer qu'il est possible et nécessaire de n'utiliser ce modèle de correction que sur certaines plages de prix. Nous avons par exemple un effet de sur-apprentissage du modèle de correction pour les prix en dessous de 100$ et les prix au-dessus de 7900$. Effectivement, dès que l'on sort de sa plage d'entraînement, le modèle de correction ne doit pas être utilisé car les prédictions ne sont plus du tout représentatives.

Heureusement, cette plage est assez large, et la correction est donc quand même utile.

3.3 Fonction prédiction.

Finalement, une fois cette régression multiple sur l'ensemble des villes gardées faites, nous nous sommes dit qu'il serait intéressant de pouvoir faire une prédiction du prix d'un appartement. Pour avoir la prédiction la plus précise possible, nous avons trouvé qu'une fonction qui adapte le modèle à la ville de l'appartement pouvait être plus pertinente et réaliste. Nous nous sommes alors placés du point de vue d'un particulier, qui,possédant un appartement et voulant faire un placement financier, aimerait un ordre d'idée du prix auquel il pourrait le louer. Nous avons donc créé une fonction permettant de faire cela. Cette fonction prend en entrée le nombre de chambres, le nombre

de salles de bain , la surface de l'appartement, et la ville dans lequel est situé l'appartement à louer. Nous restreignons nos observations à cette ville donnée, et calculons une régression multiple pour cette ville avec les variables qui sont demandées à l'utilisateur en entrée, avec le prix comme variable à expliquer. Ensuite, nous calculons une prédiction du prix avec les données de l'appartement en question. Nous demandons également avec quel sécurité (en pourcentage) l'utilisateur souhaite sa prédiction. Nous calculons ensuite, grâce au théorème du cours sur l'intervalle de prédiction, adapté à la représentation matricielle trouvé dans la littérature ($z + t_{\frac{\alpha}{2}} \sigma \sqrt{1 + X^*(X^T X)^{-1} X^{*^T}}$, avec $t_{\frac{\alpha}{2}}$ suivant une loi de Student à $n-2$ ddl, z la prédiction, σ l'écart type des erreurs, X la matrice des observations et X^* les données sur lesquelles nous effectuons la régression, cf : `https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-intRegmult.pdf`), un intervalle de confiance du prix prédit par notre fonction pour cet appartement. La signature de la fonction est donc :

```
1 def prediction_prix(NomVille : str,Bathroom,Bedroom,
    Square_feet,intervalle_conf)
```

Un exemple d'utilisation de cette fonction est :

```
1 estim = prediction_prix('Jacksonville',1,2,400,95)
```

Et la sortie de la fonction est :

```
1 'votre bien peut être loué entre 308.4$ et 361.1$ avec un
    intervalle de confiance à 95%'
```

La fonction peut être utilisée pour toutes les 2982 villes du jeu de données, bien que certaines petites villes sont peu représentées dans notre jeu de données, ce qui peut rendre l'estimation moins précise.

4 Test(s) statistique(s)

Nous avons effectué un test statistique pour répondre à la question suivante :

- Est-ce que le fait qu'un appartement soit sur les côtes américaines influence son prix ?

4.1 Test de Student et de Mann-Whitney-U

Au début, nous avons donc effectué un test de Student avec estimateurs de variance. Cependant, nous nous sommes vite rendu compte que nos données ne respectaient pas les hypothèses de test de Student. Effectivement, nos données ne sont pas distribuées normalement et la variance de prix des états côtiers est extrêmement différente de celle des états non côtiers.

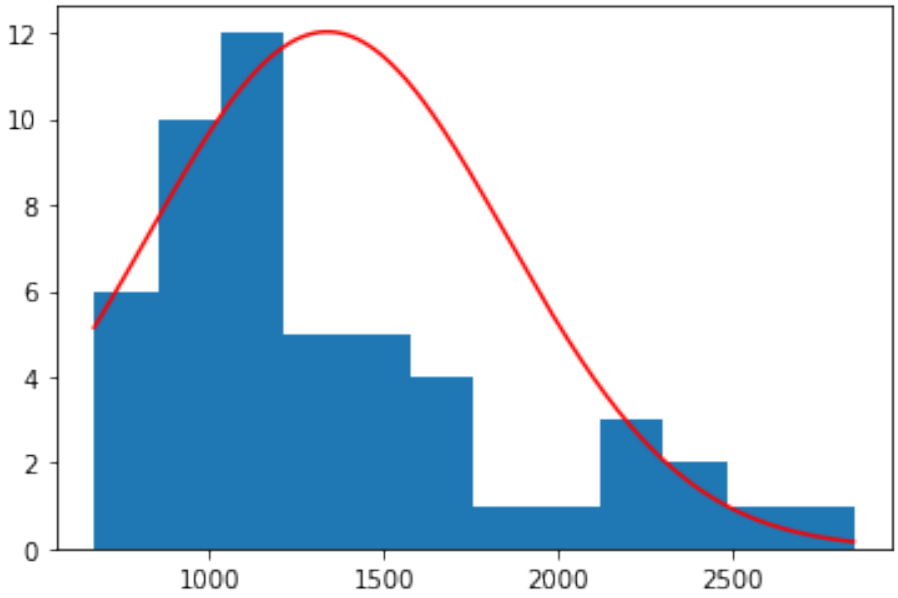

FIGURE 24 – Histogramme du prix moyen par État (bleu) et loi normale avec même moyenne et variance.

Malgré ces conditions, nous avons décidé d'explorer une approche particulière. Nous avons effectué le test de Student tout en sachant que son résultat n'avait rien de fiable. Après cela, nous allons effectuer un test de Mann-Whitney-U qui est un test relativement similaire au test de Student mais qui ne nécessite pas de distribution normale, ni de variance similaire entre les deux échantillons. Les conditions sur les données pour effectuer un test de Mann-Whitney-U sont les suivantes : il faut une variable quantitative et une qualitative, indépendantes et ordinales (qui possèdent un ordre). Nos données correspondent bien à ces conditions. Une fois les deux tests effectués, nous comparerons les deux résultats en accordant évidemment plus de crédibilité au test de Mann-Whitney-U, car nous respectons les hypothèses de ce test. Pour tous les test nous posons alors les hypothèses suivantes :

$$\begin{cases} H0 : \text{Être dans un État côtier n'a pas d'influence sur le prix d'un appartement.} \\ H1 : \text{Être dans un État côtier a une influence sur le prix d'un appartement.} \end{cases}$$

Dans le test de t-Student, nos hypothèses peuvent donc se traduire par :

$$\begin{cases} H0 : \bar{x}_{\text{côtier}} = \bar{x}_{\text{non-côtier}} \\ H1 : \bar{x}_{\text{côtier}} \neq \bar{x}_{\text{non-côtier}} \end{cases}$$

Le test de t-Student compare donc les moyennes des deux groupes étudiés.

Dans le cas du test de Mann-Whitney, on ne compare plus les moyennes des deux groupes, mais leur tendance centrale T_c. Pour calculer la tendance centrale d'un groupe, on procède de la manière suivante : Nous prenons le rang de chaque observation dans l'échantillon complet (1 pour l'État le plus cher, ..., 50 pour l'État le moins cher). Ensuite, nous calculons la moyenne des rangs de chaque sous-groupe (côtier et non-côtier). Ces moyennes sont la tendance centrale. Les hypothèses deviennent donc dans ce cas :

$$\begin{cases} H0 : T_{c_{\text{côtier}}} = T_{c_{\text{non-côtier}}} \\ H1 : T_{c_{\text{côtier}}} \neq T_{c_{\text{non-côtier}}} \end{cases}$$

Les tests sont donc bilatéraux, car nous n'imposons pas de relation ($<$ ou $>$). On s'autorise un risque de première espèce α fixé à 0.07. On s'autorise donc à se tromper dans 7% des cas en décidant que le fait d'appartenir à un État côtier a une influence sur le prix alors que ce n'est pas le cas. Dans un premier temps, nous réalisons le test de student et le test U de Mann-Whitney avec les données organisées de la manière suivante :

Etat	Côtier / Non-côtier	Prix moyen
Arizona	Non-côtier	865.8$
Californie	Côtier	2544$
Massuchussets	Côtier	2293$
...	...	...

TABLE 8 – Variables étudiée pour le test de Student

Il y a 16 États côtiers aux États-unis et 35 États non-côtiers.Nous possédons une variable quantitative, le prix et une variable qualitative, les États, que nous avons séparés en deux groupes : les États côtiers et non-côtiers.

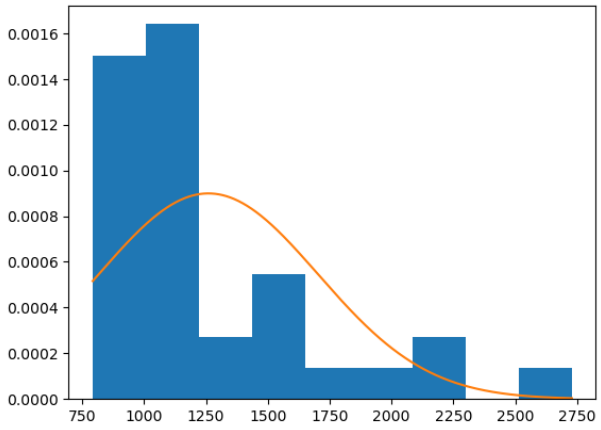

FIGURE 25 – Histogramme du groupe « Non-côtier », et loi normale de même moyenne et variance

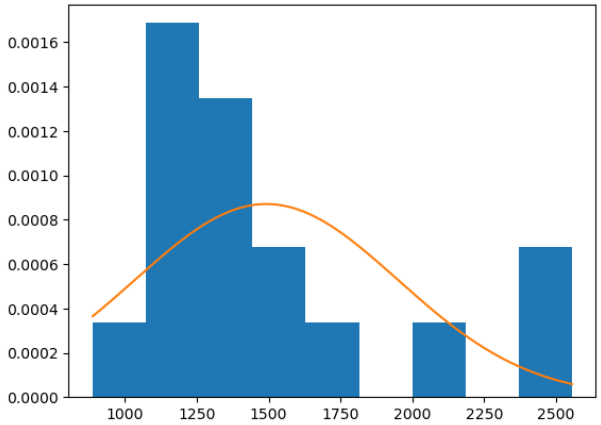

FIGURE 26 – Histogramme du groupe « Côtier », et loi normale de même moyenne et variance

Nous nous rendons bien compte ici que les observations des deux groupes ne suivent pas du tout une loi normale.

La moyenne du prix des États côtiers est de 1672$. Celle des États non-côtiers est de 1185$. Cette différence est-elle liée à un hasard raisonnable ?

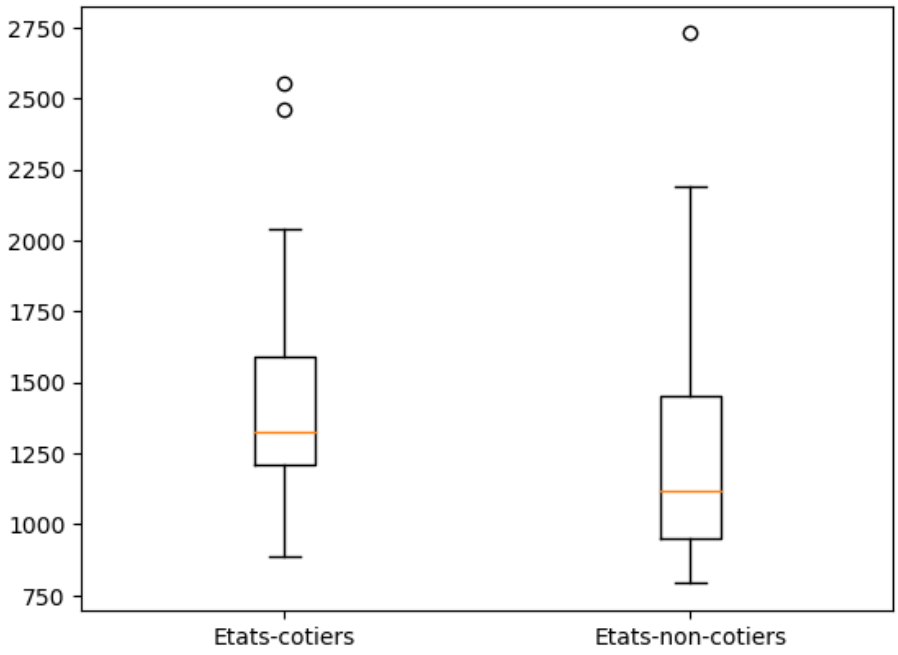

FIGURE 27 – Boîte à moustache : prix des États côtiers et non-côtiers

La médiane a pourtant l'air plutôt proche pour les deux groupes.

Le calcul de la p-valeur nous donne 0.00062 avec le test de Student et 0.02156 avec le test de Mann-Whitney, ce qui signifie que sous l'hypothèse d'indépendance H0, il y a respectivement pour les test de Student et de Mann-Whitney 0.062% et 2,15% de chance d'obtenir un tirage encore plus rare que le nôtre. Ces valeurs sont très inférieures au α de 7% que nous nous étions fixé au début du test. Notre tirage est donc beaucoup plus rare que ce que nous étions prêts à accepter au début du test. Nous rejetons donc H0 dans les deux cas.

On considérera alors que le fait d'être un État côtier influence effectivement le prix de location d'un appartement. La p-valeur de 2,15% reste évidemment la plus fiable dans le cadre de notre étude.

Comme nous savons qu'augmenter le nombre d'observations permet d'améliorer la fiabilité du test en réduisant l'influence des biais liés au manque de données, nous avons refait la même étude en organisant nos données différemment. Cette fois-ci, au lieu d'effectuer des moyennes de prix par État, nous avons classé tous les appartements comme appartenant à un État côtier ou à un État non-côtier pour les séparer en 2 groupes. Nous gardons donc les mêmes hypothèses pour effectuer les mêmes tests. Nous obtenons alors pour les deux tests une p-valeur de 0.0 ce qui signifie qu'elle est inférieure à 10^{-325}(valeur arrondie par Python).

Finalement, il y a donc peu de doute, on peut rejeter H0. Tous les tests

nous mènent alors à la même conclusion : le fait d'appartenir à un État côtier influence le prix d'un appartement.

5 Conclusion

Ce projet d'analyse d'un jeu de données d'appartement aux États-Unis a été une occasion pour nous d'appliquer les concepts et les connaissances acquises lors des cours de M8.

En étudiant ces données, nous avons identifié des tendances, confirmé des idée que nous avions ou parfois été surpris par les résultats que nous avons obtenus. Nous avons axé notre analyse autour du prix des appartements car c'est le premier facteur que l'on prend en compte lors de la recherche d'un appartement, que ce soit pour le locataire ou l'investisseur immobilier. L'analyse des prix des logements nous a permis d'identifier des tendances et des modèles, ainsi que les facteurs qui influencent le plus les prix des appartements. Par exemple, notre étude statistique nous a permis de déterminer que la longitude est l'information qui, pour notre jeu de données, influence le plus le prix d'un appartement aux Etats-Unis parmi toutes nos variables.

Nous avons aussi constaté que la localisation géographique était fortement liée au prix des appartements à différentes échelles. De plus, l'étude des caractéristiques des appartements, comme le nombre de chambres et de salles de bain, ainsi que les équipements fournis, nous a permis de mieux comprendre les préférences des locataires et des investisseurs, ainsi que les caractéristiques des appartements. Grâce à une régression, nous avons créé un modèle adapté à chaque ville permettant, pour un propriétaire, de prédire l'estimation du prix de location de son bien. Finalement, nous avons appuyé nos résultats à l'aide d'un test statistique.

Nous avons également rencontré beaucoup de problèmes différents qui nous ont permis d'apprendre de nombreuses choses, que ce soit au niveau de l'utilisation de Python ou des moyens de visualisation, de résolution de problèmes et d'analyse de nos données.

6 Annexes

6.1 Module Cartopy

Cartopy est une bibliothèque Python conçue pour la cartographie géospatiale. Elle facilite la création de cartes et la représentation de données géographiques. De plus, cette bibliothèque est compatible avec matplotlib, ce qui a été très utile et facile à prendre en main dans notre projet.

Le module Cartopy possède différentes projections de cartes. Dans notre cas, nous avons utilisé les projections PlateCarree et LambertConformal :

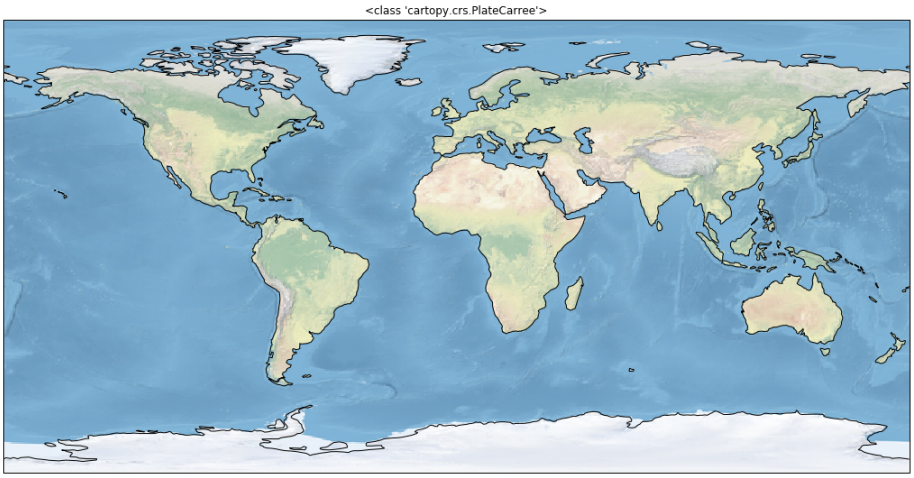

FIGURE 28 – Projection PlateCarree

FIGURE 29 – Projection LambertConformal

On commence par importer les bibliothèque cartopy.crs et cartopy.features :

```
1 import cartopy.crs as ccrs
2 import cartopy.feature as cfeature
```

A l'aide de Matplotlib, on crée une figure en choisissant une taille qui sera la taille de notre carte :

```
1 fig = plt.figure(figsize=(30, 20))
```

On ajoute un subplot en choisissant notre projection :

```
1 ax = fig.add_subplot(1, 2, 2, projection=ccrs.PlateCarree
    ())
```

Pour faire apparaitre à notre carte, on utilise la bibliothèque cartopy.feature puis on ajoute les éléments que l'on souhaite :

```
1 ax.add_feature(cfeature. « ElementAjouter »)
```

Les éléments à ajouter pouvant être : Borders (Frontières),Lakes(Lacs),Rivers(Rivières),etc... On obtient alors :

FIGURE 30 – Résultat de l'importation avec Cartopy

Pour notre projet, nous avions besoin d'une carte centrée sur les Etats-Unis. Nous avons alors utilisé alors la fonction

```
set_extent
```

qui permet de se focaliser sur une partie de notre projection, et nous y avons indiqué les coordonnées (latitude et longitude) des Etats-Unis :

```
1 ax.set_extent([-156, -66.5, 18, 65])
```

Nous obtenons alors la carte suivante :

FIGURE 31 – Résultat sans la fonction stock_img

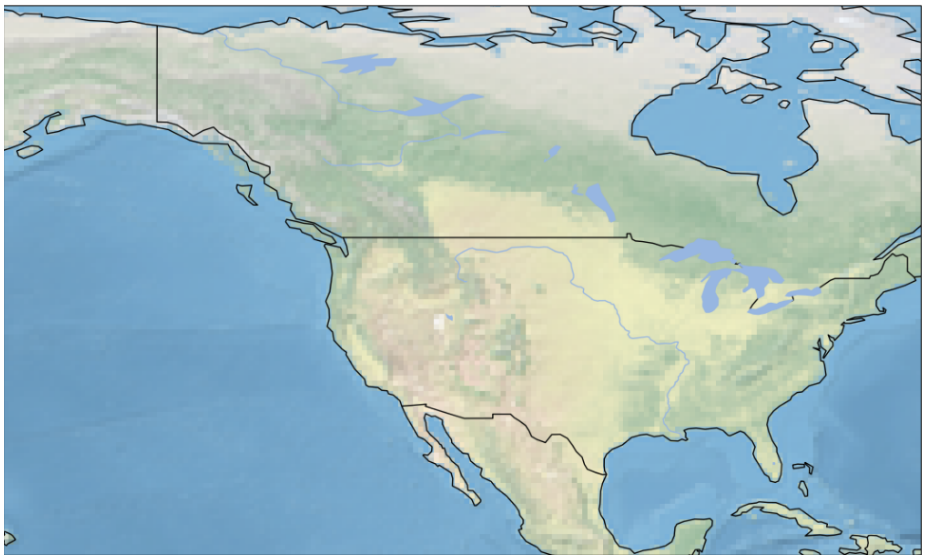

FIGURE 32 – Résultat avec la fonction stock_img

(la fonction stock_img incluse dans cartopy permet d'ajouter les reliefs)

Pour finir, en combinant ces éléments avec nos données de longitudes et latitudes ainsi que la fonction

```
plt.scatter()
```

de MatplotLib, on obtient une carte regroupant nos appartements sur une carte des États unis :

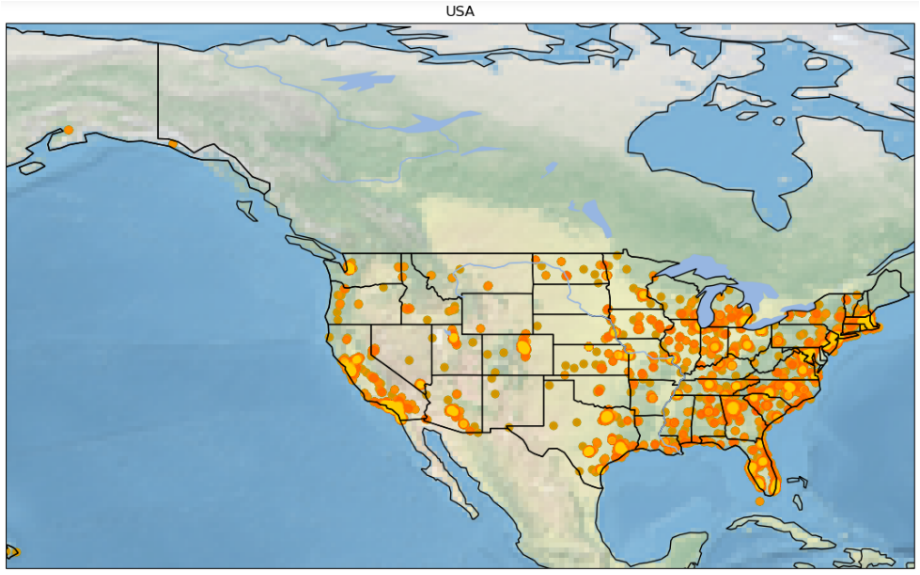

FIGURE 33 – Résultat final de notre utilisation du module Cartopy

6.2 Module Folium

Folium est une bibliothèque Python utilisée pour la visualisation interactive des données géospatiales sur des cartes. Folium est facile à utiliser et permet de créer rapidement des cartes interactives avec des fonctionnalités avancées et intéressantes.

Comme pour Cartopy, le module Folium possède plusieurs type de carte :

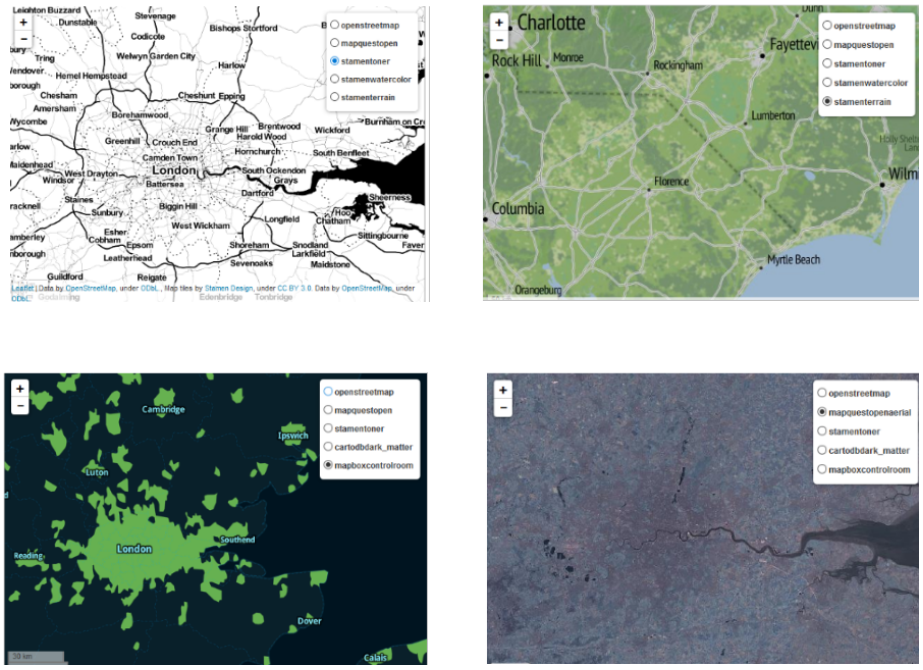

FIGURE 34 – Types de cartes de Folium (Stamen Toner / Stamen Terrain / Mapbox Control Room / MapQuest Open Aerial)

Il existe d'autres projections. Pour notre carte, nous allons utiliser la carte Stamen Terrain permettant d'afficher les reliefs.

On commence par importer la bibliothèque Folium :

```
1 import Folium
```

On crée notre carte avec la fonction

`folium.Map()`

et on la centre sur les États-unis avec des coordonnées et un niveau de zoom, puis on ajoute notre type de carte :

```
1 carte = folium.Map(location=[37.0902, -95.7129],
      zoom_start=4, tiles="Stamen Terrain")
```

On ajoute ensuite des points sur notre carte. Pour chaque point, on peut choisir la position sur la carte, la forme et couleur du point, le nom qui s'affiche lorsque l'on passe le curseur sur le point ainsi que le détail du point lorsque l'on appuie dessus.

On utilise la fonction :

```
1 folium.Marker().add_to(carte)
```

Dans notre cas :

```
1 folium.Marker([lon, lat],popup=name, tooltip=tool, icon
    =folium.Icon(color="red")).add_to(ca)
```

Avec :

- **lon,lat** sont les longitudes et latitudes de nos points
- **name** permet de coder en HTML le détail de notre point :
- on utilise les blocs :
- <**u**> : souligner
- <**strong**> : mettre en gras
- <**li**> : nouvelle ligne
- <**h1**> : mettre en titre
- <**ul**> : liste d'éléments
- **tooltip** définit le nom qui s'affiche lorsque l'on passe le curseur
- **icon** définit la couleur et la forme de notre point

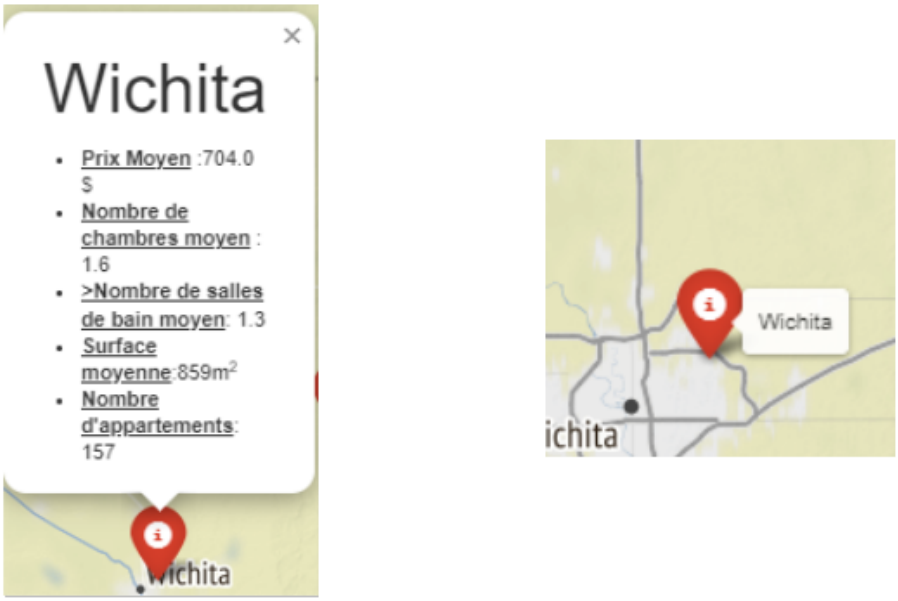

FIGURE 35 – Folium, rendu des arguments « name » et « tooltip »

FIGURE 36 – Folium, rendu de l'argument « icon » : marker par défaut / utilisé