# Rapport d’analyse des données

Introduction :

Dans ce rapport, un ensemble de données sur l’immobilier sera expliquer et analyser afin d’avoir de déduire des informations qui pourraient être significatives et indispensable pour un modèle d’apprentissage.

L’objectif est de prédire des données en utilisant un modèle d’apprentissage en passant par data cleaning, data analysis et la data visualisation.

La dataset contient 25 caractèristiques (features) et 196 lignes.

**Présentation des données :**

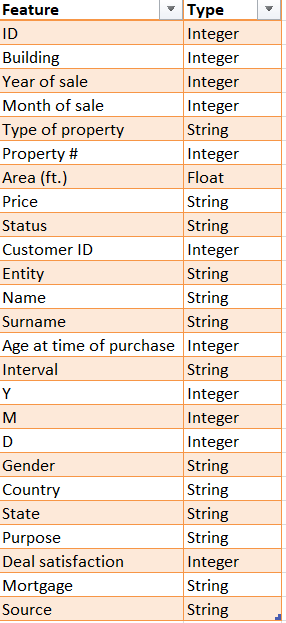
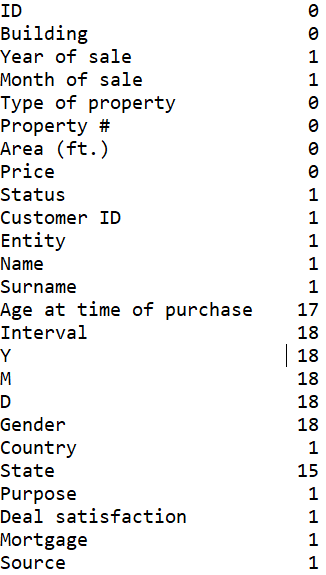


Figure 1 - Présentation des données de la data set

### Data cleaning

Dans un premier temps, après avoir balayé la data set, on remarque des valeurs manquantes dans certaines colonnes, l’image ci-dessous représente quelle colonnes contient des valeurs manquantes et combien.



Donc l’objectif est de rendre toutes ces valeurs à zéro après le data cleaning.

La première chose à faire dans cette étape est de supprimer les lignes en doubles, ensuite arranger tous les types des valeurs par exemple la colonne Price, lui enlever tout caractère spéciaux afin qu’elle soit de type **float**. Et enfin, traiter ces valeurs manquantes colonne par colonne :

**Colonne** **Age at time of purchase**:

On remarque plus on descend dans la data set plus la valeur de cet attribut augmente, donc remplir le manquement il est plus judicieux d’incrémenter la dernière valeur +1 et le faire avec le reste des valeurs manquantes.

**Colonne** **State**:

Afin de ne pas avoir des données incohérentes, j’ai créé un mapping qui lie chaque ville manquante à un pays existant et donc remplacer ces valeurs par des villes correspondantes.

**Colonne** **Gender**:

Etant donné que cette valeur est catégorielle, l’utilisation de la fréquence était plus adéquate. Donc prendre la valeur qui revient le plus dans la colonne dans ce cas c’était M.

**Colonnes** **Y**:

On remarque plus on descend dans la data set plus la valeur de cet attribut diminue, donc remplir le manquement il est plus judicieux d’incrémenter la dernière valeur -1 et le faire avec le reste des valeurs manquantes.

**Colonnes** **M et D**:

Puisque ces valeurs sont valeurs entières, la médianne était préférable pour avoir des données cohérentes avec le type.

**Colonnes** **Interval**:

Supprimer les données nulles et remplacer avec la dernière valeur renseignée pour tout le reste des valeurs.

Afin de ne pas voir d’autres valeurs manquantes, un seuil de colonnes vides a été fixé sinon on supprime la ligne.

### Data Analysis and Visualisation

Pour une observation de la dataset, on peut trouver des relations entre certaines colonnes et en déduire des informations cruciales pour le modèle d’apprentissage, mais ces relations peuvent vérifier qu’avec une visualisation de ces données. On peut retrouver plusieurs types :

Visualisation des prix par rapport à la superficie :

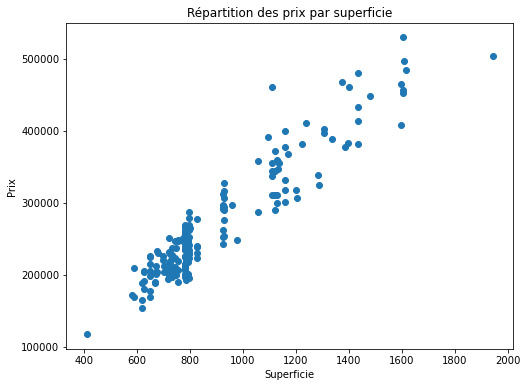
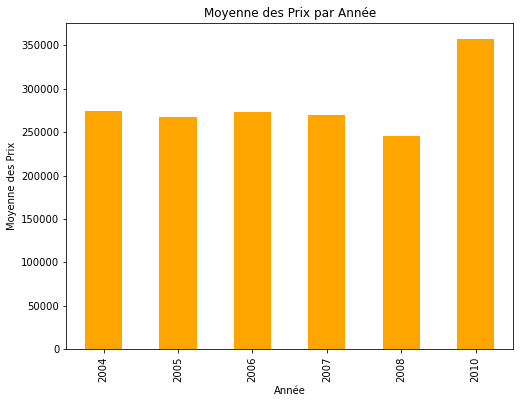


Figure 2 - Prix par rapport à la superficie

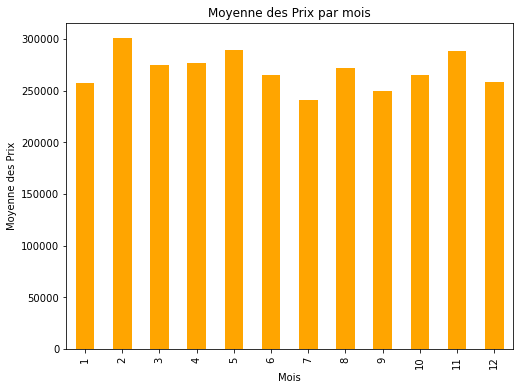
On peut voir que plus la superficie augmente, plus le prix augmente, cela représente relation qui pourrait être pertinente pour nos tests.

Visualisation des prix au courant des années:



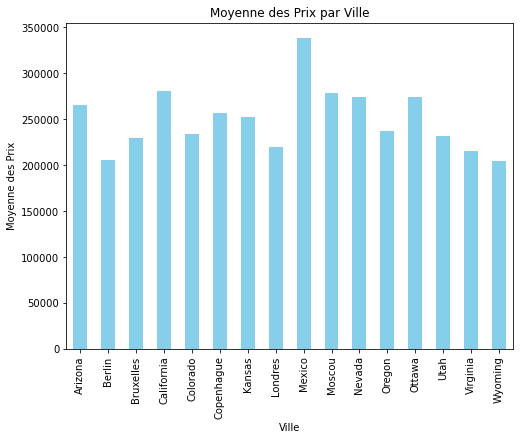
On constate qu’il y’a plus ou moins une égalité entre les années, à part en 2010 où l’augmentation des prix est bien présente, cela nous pourrait donc pas être très pertinent pour nos tests.

Visualisation des prix par mois:



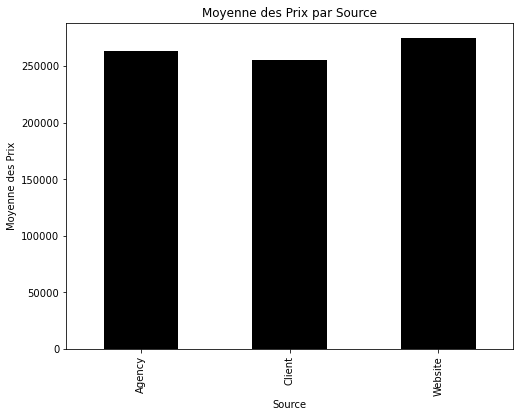
Même chose pour les mois, ils sont plus ou moins constant.

Visualisation des prix par Ville:



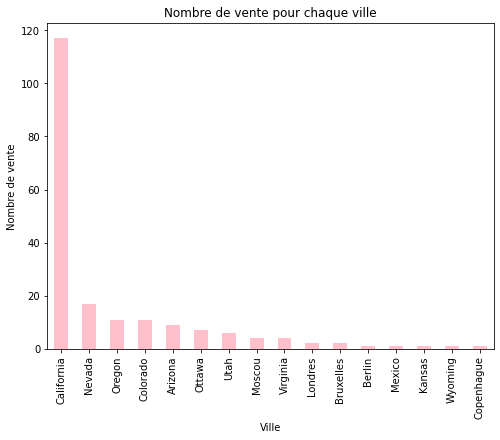
On constate qu’à Mexico que les prix sont nettement plus élevés qu’à Berlin, mais puisque il existe qu’une seule ligne qui contient Mexico, cela n’est pas suffisant pour le modèle d’apprendre.

Visualisation des prix par Source:



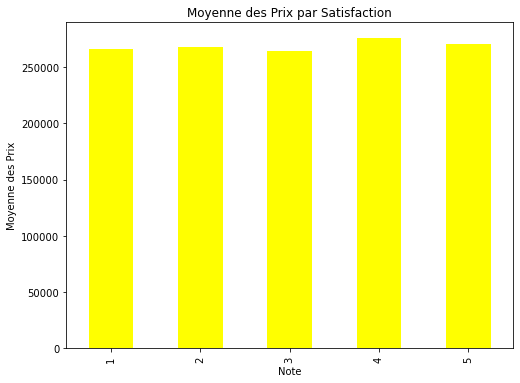
On constate qu’il y’a plus ou moins une égalité entre les sources, cela nous pourrait donc pas être très pertinent pour nos tests.

Visualisation du nombre de vente par ville:



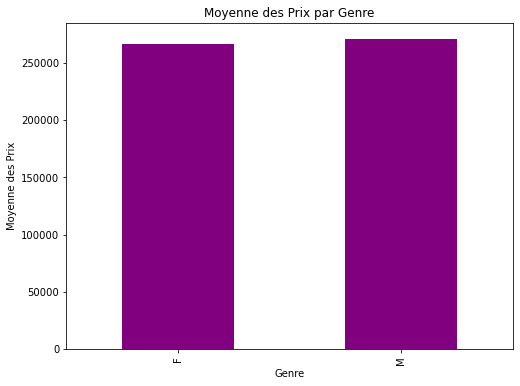
On voit que la ville de Californie est nombre nettement supérieur aux autres villes.

Visualisation des prix par rapport à la note de satisfaction:



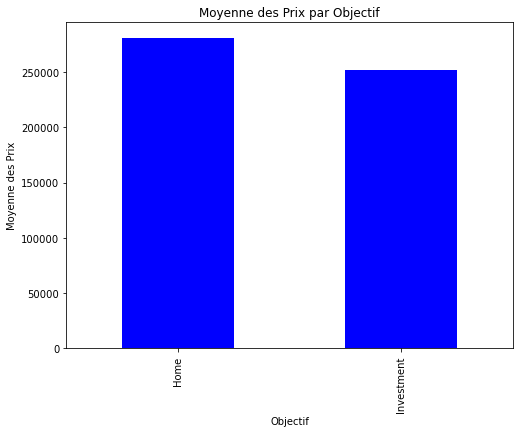
On remarque que la note de satisfaction n’a aucun rapport avec le prix.

Visualisation des prix par rapport au genre:



On remarque que le genre n’a aucun rapport avec le prix.

Visualisation des prix par rapport à l’objectif:



On constate que les clients qui veulent acquérir un bien en tant que maison, le prix est legerement plus élevé que pour un investissement.

Visualisation de la heatmap:

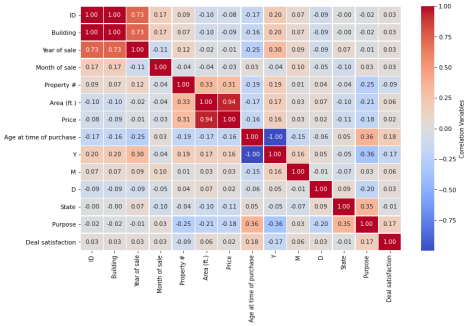
Ce graphe sert à voir les corrélations les plus fortes entre les colonnes :

Figure 3- HeatMap entre attributs

On constate qu’il y’a une corrélation très forte entre la **superficie** et le **prix** du bien (elle est proche à 1), c’est exactement ce qu’on avait déduit avec la figure-2, tandis que les autres attributs n’ont une corrélation remarquable (proche à -1) pour qu’on puisse l’exploiter dans modèle d’apprentissage.

N.B : le script DataViz.py contient plus de visualisations.

**Apprentissage automatique**

Après avoir analysé les données sur l’immobilier, la donnée la plus intéressante à prédire est celle du **Prix(Price).**

En se basant sur la heatMap (Figure – 2) on peut prendre la **superficie (Area)** comme donnée en input, étant donné qu’il existe une forte relation entre eux.

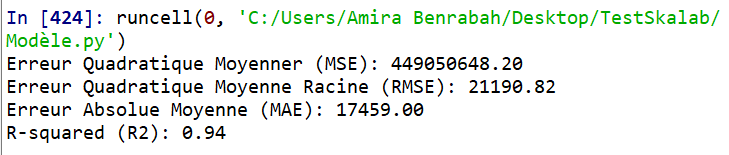
Puisque la donnée à prédire est une donnée réel(float) c’est-à-dire une valeur continue, le modèle d’apprentissage adéquat serait la régression linéaire simple.

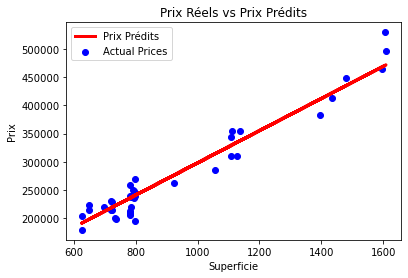
La dataset ne contient pas beaucoup de lignes, il est préférable de la diviser à 80 20, 80 % données d’entrainement, 20% données tests de manière aléatoire.

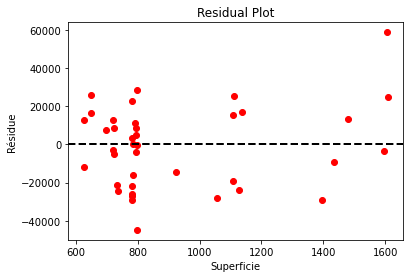
Les métriques qui évaluent la régression linéaire sont :

* Erreur Quadratique Moyenner (MSE)
* Erreur Absolue Moyenne (MAE)
* Erreur Quadratique Moyenne Racine (RMSE)
* R-squared (R2)

Après avoir diviser la donnée, lancer l’entrainement nous pouvons constaté les valeurs de ces métriques avec le graphe des valeurs prédites par rapport aux valeurs réelles ci-dessous :







Bien qu’il y est une forte corrélation entre ces deux attributs, ces résultats sont moyens et peuvent être amélioré, nous remarquons à travers ces graphes que certes que dans un bloc de superficie l’erreur est moindre mais plus celle-ci augmente (superficie) plus la prédiction est mauvaise, ce qu’on conclue qu’il existe des valeurs aberrantes.

Pour remédier à cela, il faut soit les supprimer soit en contraire rajouter des valeurs à la superficie, par exemple (Area², log(Area)…) afin de rajouter plus de données à l’ensemble de donnée, ou faire une normalisation des données.

On peut aussi rajouter des caractéristiques qui ont une forte corrélation avec le prix, ou modifier les existants, par exemple le prix des biens doit être réduit par rapport à la note de satisfaction.