RAPPORT DU PROJET

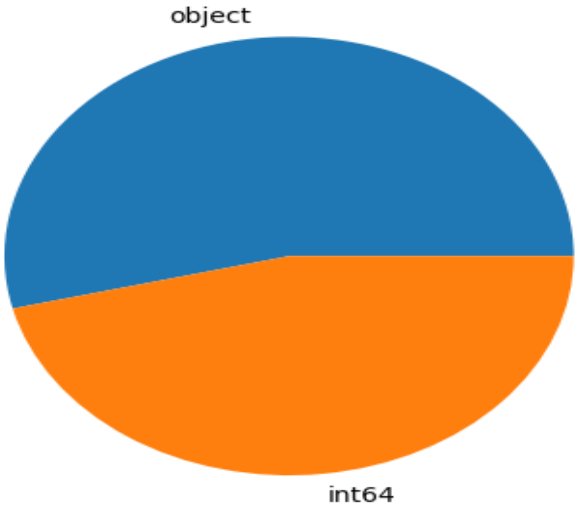
TACHE 2 : Tracé de Graphiques et Analyse de Données

Faire une analyse de données avancée en Python ou R implique plusieurs étapes, allant de la compréhension des données brutes à la communication ou interprétation des résultats issue des différents graphes affichés. Voici les principales étapes détaillées :

1. COMPREHENSION ET PREPARATION DES DONNEES (Prétraitement)

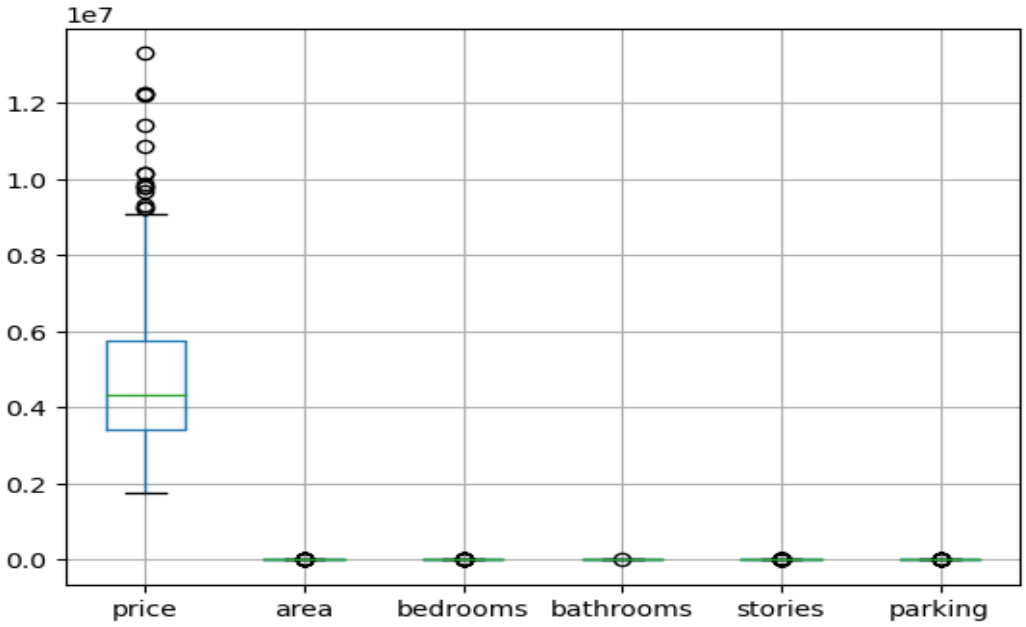
- Collecte des données
  Elle consiste généralement d'aller à la rencontre de nos données à traiter tout en créant un sondage pour pouvoir récupérer ses données dans une base de données ensuite l'importer dans un environnement de travail pour son analyse.
  Alors nous allons utiliser un fichier '**Housing.csv**' télécharger sur **Kaggle** sur lequel on est amené à faire une analyse en python et en R.
  Pour importation de cette base de donnée (dataset), on utilise souvent la bibliothèque pandas en python et en R lire notre base à l'aide de 'read.csv' pour les fichiers de type CSV.

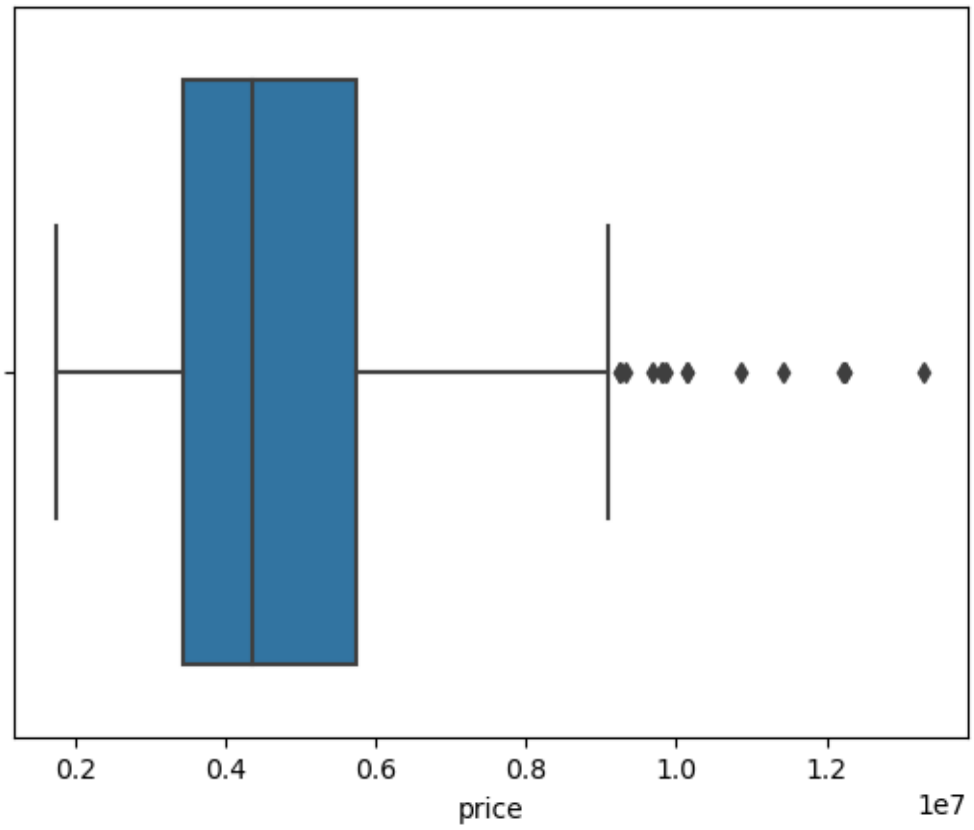
- Exploration initiale des données dans 'Housing.csv'
  Cette sous-étape a pour but de comprendre la structure des données encore appelé Métadonnées.
  - ✓ La taille du dataset est 543 lignes et 13 colonnes
  - ✓ Information sur les colonnes : `'price', 'area', 'bedrooms', 'bathrooms', 'stories', 'parking', 'mainroad', 'guestroom', 'basement', 'Hotwaterheating', 'airconditioning', 'prefarea', 'Furnishingstatus'.`



- ✓ Un résumé statistique sur le dataset ‘Housing.csv’

- ➢ Nettoyage des données
  - ✓ Il n’y a pas du tout de valeurs manquantes dans le dataset après vérification et visualisation.
  - ✓ Mais quant aux valeurs aberrantes, seule la colonne ‘price’ contient 15 valeurs aberrante avec un pourcentage de 2.752294%

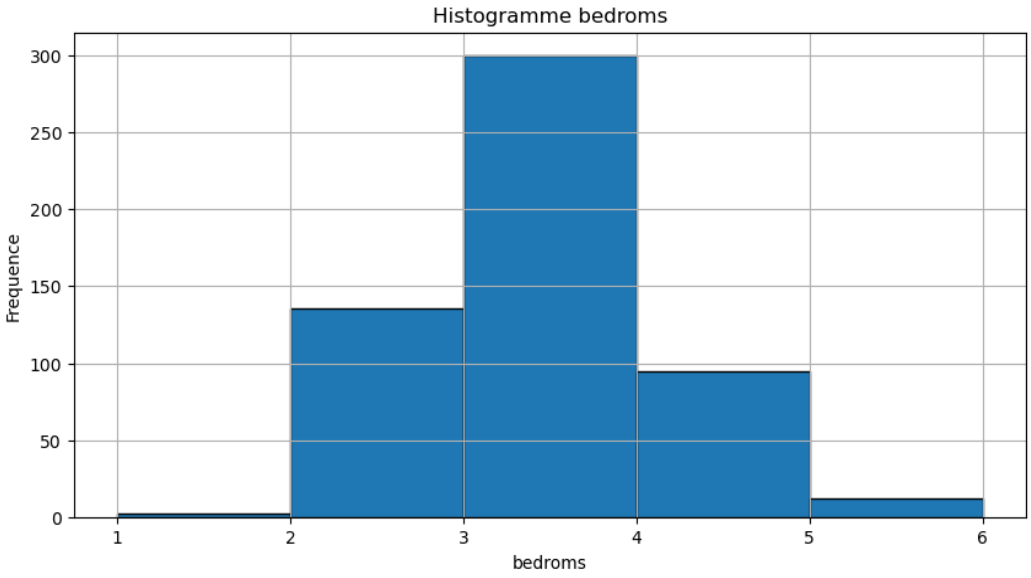




2. ANALYSE EXPLORATOIRE DES DONNEES (EDA)

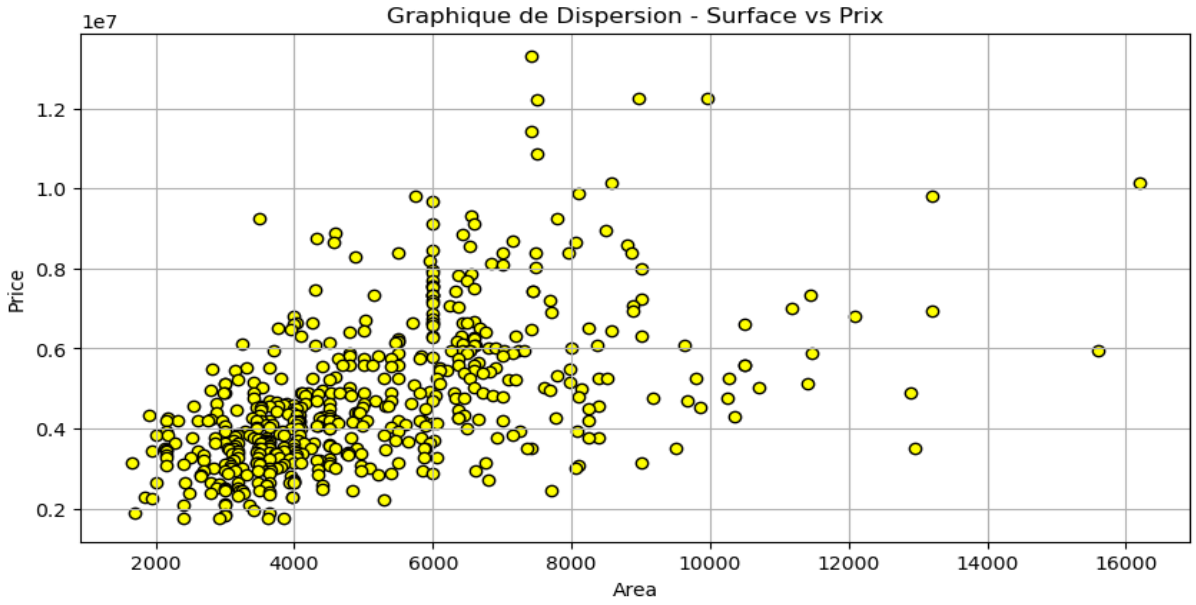
Pour faire les visualisations dans une analyse de données on utilise généralement la bibliothèque **matplotlib.pyplot** en python et **ggplot** en R

- Visualisation de la colonne '**bedrooms**'



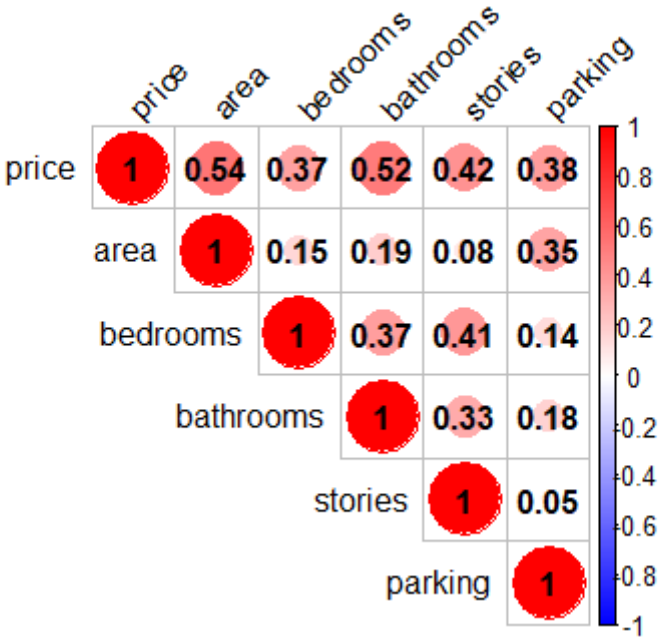
CONCLUSION : d'après la distribution de la colonne 'bedrooms', on dit dire qu'elle symétrique.

- Graphiques de dispersion de la colonne 'area' et 'price'



CONCLUSION : ce graphe permet une compréhension approfondie de la relation entre la surface et le prix des propriétés dans le jeu de données, ainsi que la détection des valeurs aberrantes.

- Matrix de corrélation du dataset avec '**ggplot**'



CONCLUSION : la matrix de corrélation est vraiment importation pour pouvoir faire du features selections, elle permet de voir la relation ou lien entre les colonnes de notre dataset en général. On en déduire ici qu'il n'y a pas une forte corrélation entre les colonnes de notre jeu de données.