Le besoin client :

Avoir une estimation des prix de vente d’un bien immobilier dans la France

Les données :

* Open Data sur le site data.gouv.fr (lien de la base de données : <https://www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncieres-geolocalisees/> )
* Scrapping du site seloger.com (idée arrêtée car les données sont en open data sur Data Gouv )

Création d’une base de données sur SQL Server:

Avec la base de données récupérée de grosses difficultés à l’intégrer.

Donc traitement des années 2017-2018-2019 : 7 453 214 entrées

Feature engineering : effectué sur Jupyter Notebook avec Python :

* + Réduction du nombre de colonnes de 40 à 15
  + Ajout d’une colonne id pour les biens
  + Suppression des doublons des id mutation et parcelle

Après traitement le dataset est à 2 224 142 entrées.

Data analysis and exploration

* + Séparation du dataframe unique en 3 :
    - Mutation
    - Cadastre
    - Bien
  + Export des différents dataframes

Création d’une base de données VentesImmobilieres

Intégration des données à partir des différents CSV pour la création des tables

Création d’un backup automatique

Visualisation avec PowerBI

Création d’une connexion entre SQL Server et PowerBi

Visualisation des données dvf SQL Server

Elle met en évidence des NaN

Il faudra donc revoir les données pour l’algorithme de ML

Nouveau traitement des données et mise en place d’un algo

* Nouveau traitement des NaN
* Visualisation des variations de prix
* Récupération des données de vente exclusivement
* Création d’une colonne Régions pour faciliter le travail des algorithmes