Dans un premier temps, nous nettoyons nos données, pour ce faire nous déterminons quelles sont les colonnes pertinentes et qui présentent des corrélations, pour ainsi nous permettre d'écarter les autres qui ne le sont pas.

Voici la liste des variables incluses dans notre dataset :

- APP_Libelle_etablissement
- SIRET
- Adresse_2_UA
- Code_postal
- Libelle_commune
- Numero_inspection
- Date_inspection
- APP_Libelle_activite_etablissement
- Synthese_eval_sanit
- Agrement
- geores
- filtre
- ods_type_activite

Nous souhaitons expliquer la variable « Synthese_eval_sanit ». Nous éliminons donc les variables non corrélées à cette dernière.

Nous avons décidé d'écarter les colonnes suivantes :

- Adresse_2_UA
- Libelle_commune
- geores
- Agrement
- ods_type_activite
- Filtre

La colonne Agreement est écarté car plus d'un tier des données sont vides donc les données sont considérées comme aberrantes.

La colonne filtre car elle est trop spécifique.

La colonne Adresse fait doublon avec la colonne Code Postal et est trop précise pour ce que nous voulons trouver de même pour la colonne libellée commune.

La colonne geores, car nous n'avons pas l'utilité de se baser sur des coordonnées qui sont encore moins lié qu'une adresse car l'échelle est complexe.

La colonne ods_type_activite comporte plus de ¾ de « Autres », donc pour éviter d'avoir une mauvaise interprétation nous l'avons évincé.

La colonne Libelle_commune qui pourrais servir si on cherche à expliquer notre valeur en fonction de chaque commune en segmentant par cette dernière mais ce n'est pas notre cas.

À la suite de notre trie de colonne, il nous a ensuite fallut affiner la colonne APP_Libelle_etablisseement, pour que cette colonne soit utilisable, il faut que les données soient plus simples car il y a des noms à rallonge. Dans un premier temps avec des regex on a trier les différents séparateurs pour pouvoir mieux visualiser les différents libellés. Ensuite, nous avons créé différents groupes dans lesquelles on a placer nos libellés.

La colonne APP_Libelle_etablissement, nous allons utiliser la colonne SIRET, cela nous enlèvera de la précision si une entreprise avec le même SIRET possède plusieurs établissements.

Nous avons commencé par tenter une régression linéaire, cependant ce ne fut pas concluant car nous devons prédire une variable qualitative et non une variable quantitative.

Par la suite pour modéliser et régler ce problème nous avons modifié la variable APP_Libelle_etablissement pour la classifier toutes les variations possibles en 12 catégories afin d'avoir des données qui sont plus facilement exploitable.

À la suite de la première génération du modèle, nous avons décidé de tenter de faire une matrice de corrélation.

Notre objectif est de répondre à des problématiques liés entre les différentes variables restantes et notre variable à expliquer « Synthese_eval_sanit ». Nous souhaitons prédire la satisfaction donner par les inspections en fonction du type d'activité, du libellée d'activité, de la date d'inspection et du code postal.

Pour notre régression nous constatons une précision qui n'est pas très élevé, nous avons donc deux possibilités soit les données ne sont pas suffisantes pour dégager un pattern, soit l'hypothèse de problèmatique n'est pas corecte.

La matrice de corrélation est une exploration de données et le résultat qu'elle nous donne est que les différentes variables utilisés dans sa construction ne sont visiblement pas corrélées. Sur cette problématique nous n'avons pas réussi à dégager un pattern de données lors de cette exploration. Nous pensons donc qu'elle ne s'impacte pas entre elles.