Visualisation 2

Guillaume Houbion, Pierre Delsirie

Dataset:

Nous avons choisi ce dataset en nous posant la question suivante peut-on trouver des différences notoires entre les différents modèles en fonction des années et des origines.

Le dataset que nous avons choisi contient des informations sur des modèles de voiture venant des US d'Europe et enfin du Japon. Les différentes informations sont la consommation de carburant (MPG), le nombre de cylindré, le volume générer par les cylindrés, la puissance, le poids, l'accélération, l'année, et enfin l'origine.

Visualization 1 :

Pour le fichier visualization_1 nous avons utilisé la méthode parallele coordinate avec le module plotly pour afficher les données. Il est possible de choisir un intervalle de points à afficher (par défaut tous les points sont affichés). Cette méthode permet de très bien voir les tendances dans le dataset et l'impact de chaque donnée sur la consommation de carburant.

Visualization 2 :

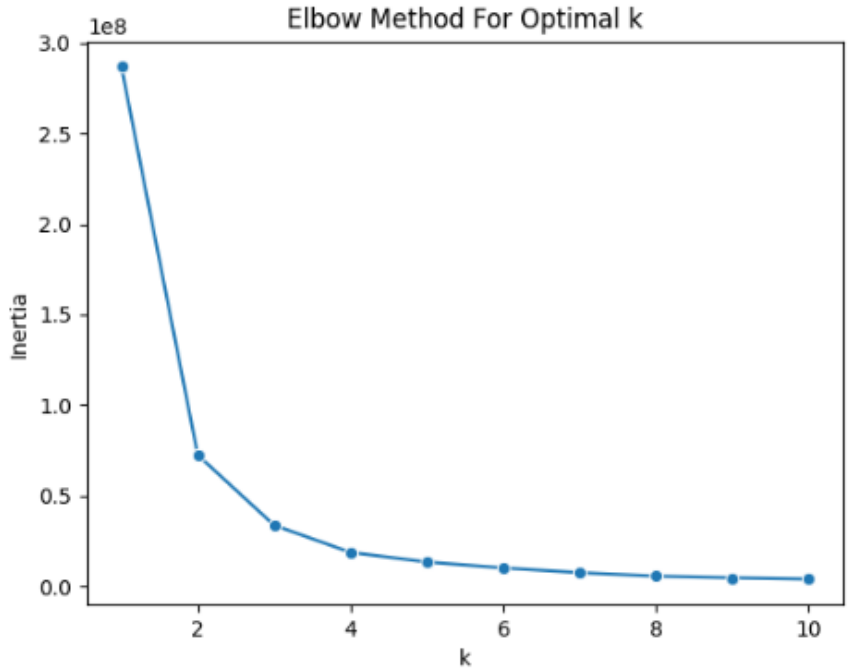
Pour le fichier visualization_2 nous avons utilisé la méthode scatterplot avec le module seaborn pour afficher les données. Il est possible de choisir un intervalle de point ainsi que les colonnes à afficher (par défaut tous les points sont affichés et les colonnes sont "Acceleration" et "Weight"). Cette méthode permet de visualiser seulement certains éléments mais ne permet pas d'avoir une vue d'ensemble de toutes les données.

Supervised learning :

Pour supervised_learning nous avons utilisé la méthode de régression linéaire à l'aide de scikit learn. Le but est de prédire la cosommoation (MPG) pour un poid choisi. Le modèle a un coefficient de détermination de 0.69, il est donc moyennement précis.

Unsupervised learning :

Pour unsupervised_learning nous avons utilisé la méthode Kmeans à l'aide de scikit learn. Nous avons utilisé l'inertie pour déterminer le nombre optimal de cluster pour le dataset.



Résultats :

Les résultats nous permettent pour commencer de montrer des choses simples comme le fait que plus une voiture est lourde plus sa consommation d'essence est importante.

On peut aussi remarquer que les US produisent beaucoup plus de modèle de voiture que les autres régions du monde. Et que leur moteur sont souvent plus puissants que ceux des voitures Européenne ou Japonaise.

On peut aussi remarquer de manière plus générale que la consommation des voitures augmente au fil des années surtout si on compare la moyenne de la consommation entre les voitures produites en 82 et celle produite en 70.

En ce qui concerne la distribution des données nous pouvons parler simplement du fait que sur notre dataset les voitures américaines sont surreprésenter en comparaison des voitures Européenne et Japonaise, ce qui peut nous amener à nous demander si toutes les conclusions que nous pouvons tirer des données sont forcements justes. Il est tout à fait possible qu'avec un dataset différents certaines conclusions ne soient pas les mêmes.

Répartitions du travail :

Côté répartition du travail Guillaume s'est occupé de faire Artificial Dataset, Visu 1 & 2 et Pierre de Analysis, Supervised learning & Unsupervised learning.