

Algo 2 matching, graph, clustering

Guillaume Houbion, Pierre Delsirie

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.

Nature du DataSet : 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 cylindres, le volume généré par les cylindres, la puissance, le poids, l'accélération, l'année, et enfin l'origine.

Pour la création des graphs nous avons choisis de faire comme dans le cours avec des matrices de dissimilarités en faisant la distance euclidienne entre nos données.

Nous avons choisi de faire du clustering et pour cela nous avons utilisé un algorithme Kmeans qui est une méthode de partitionnement de données se basant sur la division des différents points en k groupes (les clusters), ainsi qu'une Elbow méthode en calculant le WCSS (Within Cluster Sum of Square).

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 modèles 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 soit pas les mêmes.