Dataset

Visualization of Massive Data

Yoann Kersulec

Benjamin Viguier

# L’ensemble de données

Les données sont issues du lien deux de l'énoncé, Il s'agit d’une liste de modèles de voitures et de leurs caractéristiques.

Le fichier csv contient 9 colonnes :

* Car : Modèle de voiture
* MPG : miles per galons ; consommation
* Cylinders : nombre de cylindres
* Displacement : Volumes compresse par l'ensemble des cylindres
* Horsepower : puissance
* Weight : poids
* Accélération : temps en seconde du 0 à 100 km/h
* Model : Annee de parution du véhicule
* Origin : Continent / Pays d’origine

## Pourquoi celui-là ?

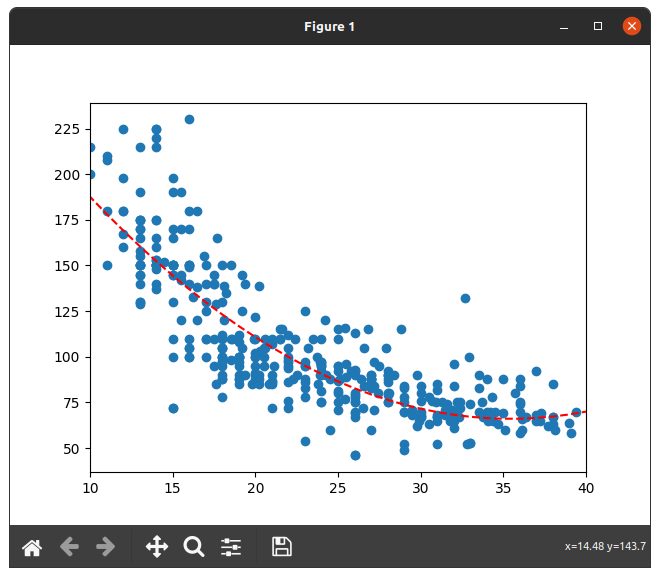
Nous avons choisi ce dataset car il y a des corrélation entre puissance et cylindrée, poids et puissance ou même puissance et origine de la voiture.

De plus ces données nous ont semblé compréhensibles et faciles à interpréter, et donc tout indiqué pour un projet pour des débutant.

## Analyse quantitative

Pour l’analyse, nous avons écrit un script pour effectuer un apprentissage supervisé.

Le script utilise deux colonnes (hardcodé, mais peut être modifier en changeant une ligne de code), les données de ces deux colonnes sont ensuite filtrées : on retire toute les paires incomplètes.



On utilise ensuite la fonction [curve-fit de scipy](#38pljkg72zvd) pour trouver une courbe en fonction des points

Ici on peut voir en rouge une courbe représentant la prédiction, Avec sur l’axe des abscisses la puissance et sur l’axe des ordonnées la consommation (plus le chiffre est grand moins on consomme)

# Librairies

## pyplot

On s’en sert pour générer un tableau en fonction d’un ensemble de points qu’on lui donne. C’est elle qui génère tous les visuels de nos trois exercices de ce projet.

[plt.plot](https://matplotlib.org/api/_as_gen/matplotlib.pyplot.plot.html#matplotlib.pyplot.plot) : Assigne les points, les paramètres sont un ensemble de tableaux de valeurs qui forment des points pour le scatter plot par exemple.

[plt.xticks](https://matplotlib.org/api/_as_gen/matplotlib.pyplot.plot.html#matplotlib.pyplot.plot) : Assigne les labels sur l’axe x pour le ‘parallel coordinate plot’.

[plt.xlim](https://matplotlib.org/api/_as_gen/matplotlib.pyplot.plot.html#matplotlib.pyplot.plot) : Limite les valeurs à un sous-ensemble de l’ensemble sur l’axe x. Dans le cas de l’analyse quantitative, le programme limite l’axe x à toutes les valeurs comprises entre 10 à 40.

[plt.show](https://matplotlib.org/api/_as_gen/matplotlib.pyplot.plot.html#matplotlib.pyplot.plot) : Affiche le graphique dans une fenêtre.

## scipy

[curve-fit](https://matplotlib.org/api/_as_gen/matplotlib.pyplot.plot.html#matplotlib.pyplot.plot) : Cette fonction permet de créer une courbe qui l’ont pourra tracer sur pyplot. Comme son nom l’indique, la courbe tracée va s’ajuster grâce à une fonction (objective() dans le cas du fichier analyse.py).