Visualization of Massive Data

Comments on the Dataset

made by mael.brocher@epitech.eu
& julien.fiant@epitech.eu

Nature of the dataset

We choose an open source dataset of cars with 407 rows and 9 columns.

Correlation between variables

We think this dataset has big correlation between Horsepower/Acceleration, Horsepower/Cylinders and Horsepower/Weight.

Explanation about the Visualization and Supervised learning

Visualisation 1

In this script we use scatter plots, to show the evolution of the correlation curve between 2 columns.
On default we set Horsepower/Acceleration, because for us it’s the best result for this visualization.

Visualisation 2

In this script we use box plots, to show a lot of statistics between 2 columns.
On default we set Horsepower/Origin, because for us it’s the best result for this visualization. Especially Origin which fit perfectly with box plots.

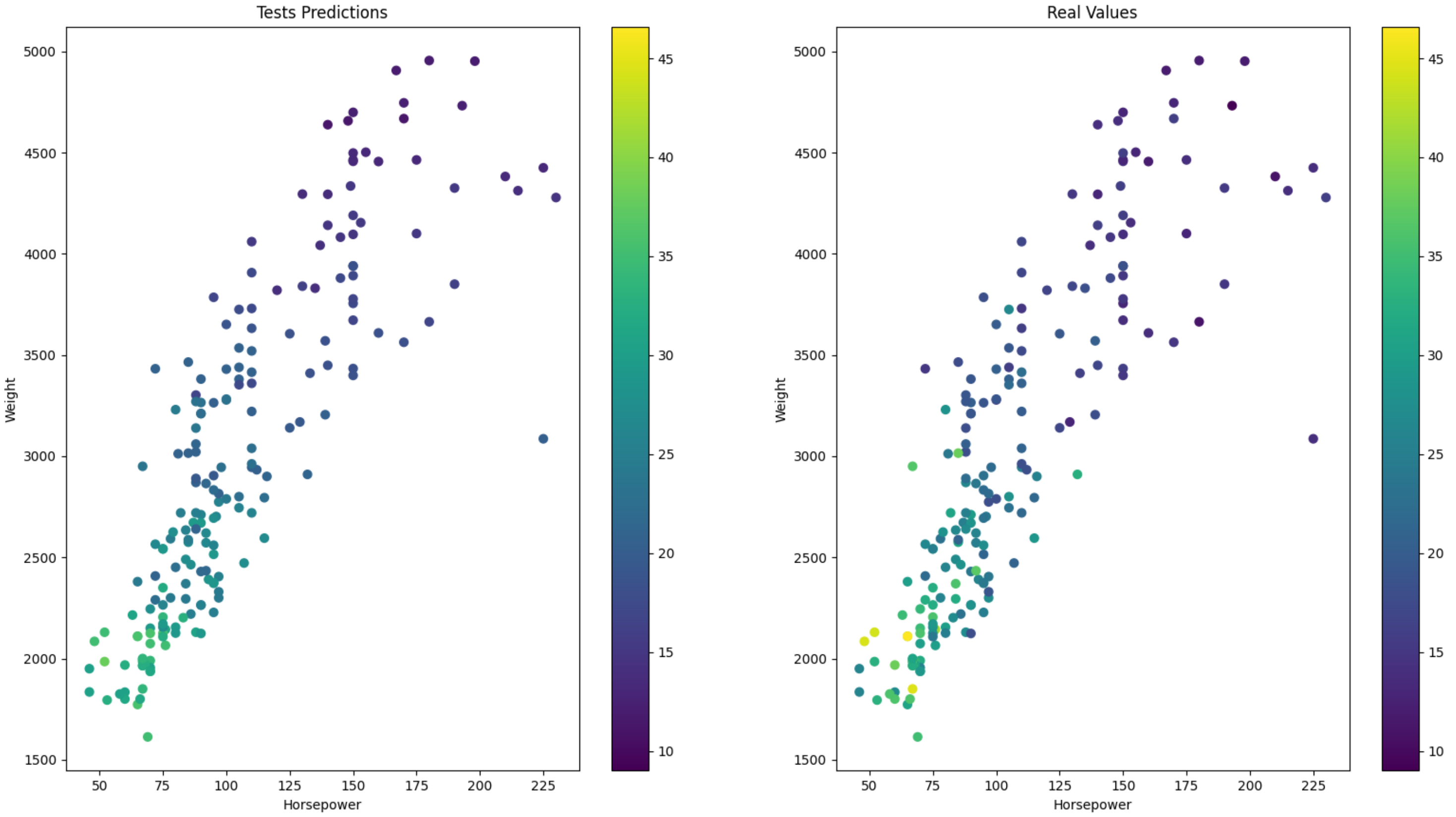
Supervised Learning

We used Knn (for k-nearest-neighbors), to predict as it's easy to implement with the python library sklearn.
Our goal is to predict the miles per gallon of different cars based on the weight and horsepower of the first half of the dataset.
The dataset might be too small to but the results obtained on the second half of the dataset were close to the real value.

Comments on the results obtained

As you can see here a screenshot of our result. The predictions seem to have worked and the root mean squared error is 4.502 which means that on the average there’s an error of 4.5 MPG in the consumption of gas.

On the left it’s the prediction of our knn model and on the right you can see our real value. the Y axis is the Weight, the X Axis is the Horsepower and the color bar is the miles per gallon



Tests Predictions

Real Values