Data Cleaning - Report

For the **missing values** (?), I replaced the value with the mean of the column in:

- Normalised-Losses
- Price
- Horsepower

For the next 3:

- Bore
- Stroke
- Peak-rpm

I replaced it with null type as mean does not fit in those columns.

In ‘Number-of-doors’ column, I removed the rows which had ? under this category.
I chose removing or replacing based on the category of the division.

Below are some points from the analysis of the data set:

- Toyota is the make with most number of vehicles with 40% more than its follower Nissan
- Preferred fuel type is gas (175 vehicles) compared to diesel (25 vehicles)
- For drive wheels, fwd has most number of cars followed by rwd and 4wd. There are very less number of cars for 4wd.
- Curb weight of the vehicles is distributed between 1500 and 4000 (approx.)
- Symboling or the insurance risk rating have ratings between -3 and 3, but for our dataset it starts from -2. There are more vehicles in the range 0 to 1.
- Normalized-losses (average loss payment per insured vehicles) is higher for no. of vehicles between 65 and 70
- The most expensive car is manufacture by Mercedes benz and the least expensive is Chevrolet
- The premium cars costing more than 20000 are BMW, Jaquar, Mercedes benz and Porsche
- Less expensive cars costing less than 10000 are Chevrolet, Dodge, Honda, Mitsubishi, Plymoth and Subaru
- Rest of the cars are in the midrange between 10000 and 20000 which has the highest number of cars

Thoughts

The assignment was very helpful in learning how to analyse a data set by visualising it and changing values inside the data frame. Large quantity of data is simplified using pandas which helps in understanding the material easier. In a data set which has 204 rows and 26 columns, we were able to point specific points as mentioned above with the help of some functions.