Huong Phan
CS 382: Data Cleaning Assignment

Question 1

- Replace all the missing values of continuous data (normalized-losses, bore, stroke, horsepower, peak-rpm, price) by the mean value of data for each variable, because this will not change the mean values of continuous data. However, this will affect any correlation between the variables that are replaced, for example, there would be a lot of points with the same horizontal or vertical coordinates in a scatterplot of two variables, which will decrease the reliability of the regression plot. So when correlation between variables is considered, it's better to ignore the missing continuous data.
- Replace missing values of categorical data (num-of-doors) by the most frequently appeared value, because the missing categorical data is more likely to be the most frequently appeared value in the data than other less frequently appeared values i.e the probability of the missing categorical data to be the most frequently value is the highest.

Question 2

- Convert categorical data into numerical data: Categorical data in column 1, column 6 and column 16 can be converted into numerical data, because categories in those columns are numbers such as "-3", "0", "four", "two", "eight",…

Question 3

- Exploratory analysis:

+ Scatterplots and linear regression plots for:

1. Correlation of horsepower against curb weight: strong positive correlation, r-value = 0.7885
2. Correlation of price against curb weight: strong positive correlation, r-value = 0.8938
3. Correlation of peak-rpm against curb weight: weak negative correlation, r-value = -0.2620

4. Correlation of peak-rpm against wheel base: weak negative correlation, r-value = -0.2925
5. Correlation of horsepower against wheel base: weak positive correlation, r-value = 0.5145
6. Correlation of price against wheel base: strong positive correlation, r-value = 0.7348

+ Distribution plots:

1. We can see that distribution plots are all skewed right for curb weight, length, width, wheel base, car price, meaning the median < mean for these variables. The distribution plots for height and symboling is almost symmetric, meaning median is almost equal to the mean.
2. Test the above interpretations using the describe function. The above is more or less correct, except that the distribution of symboling data actually skews left because the mean < median for symboling data.

+ Bar plots for different categories:

1.1 Bar plot of body style against mean city-mpg and highway-mpg
Hatchback has the highest mean city-mpg and highway-mpg, while convertible has the lowest mean city-mpg and highway-mpg
1.2 Bar plots of body style against mean bore and stroke.
Hardtop has the highest mean value for stroke, but convertible has the highest mean value for bore; wagon has the lowest mean value for stroke, but hatchback has the lowest mean value for bore.
1.3 Bar plots of body style against mean bore/stroke ratio
The mean bore/stroke ratio is highest in wagon, lowest in hatchback. All body types are generally short-stroke (bore/stroke ratio > 1), except for hatchback.
1.4 Bar plots of body style against mean price
Convertible is generally the most expensive, while hatchback is generally the cheapest.

2.1 Bar plots of fuel type against mean city-mpg and highway-mpg

Engines with diesel have higher mean city-mpg and highway-mpg than those with gas.

2.2 Bar plots of fuel type against mean bore and stroke

Engines with diesel have higher mean bore and stroke than those with gas.

2.3 Bar plots of fuel type against mean bore/stroke ratio

Engines with gas are short-stroke, while the opposite happens in engines with diesel.

2.4 Bar plots of body style against mean price

Engines with gas are generally more expensive than those with gas.

3.1 Bar plots of engine location against mean city-mpg and highway-mpg

The highest mean city-mpg and highway-mpg are the highest in cars with engine location ohc, lowest in that cars with engine location ohcv.

3.2 Bar plots of engine location against mean bore and stroke

Cars with engine location ohcf have the highest mean value for bore, but cars with engine location ohc have lowest mean value for bore. Cars with engine location dohc have the highest mean value for stroke, but cars with engine location ohcf have lowest mean value for stroke.

3.3 Bar plots of engine location against mean bore/stroke ratio

The mean bore/stroke ratio is highest in cars with engine location ohcf, lowest in cars with engine location ohc. Cars with all engine locations are generally short-stroke (bore/stroke ratio > 1), except for cars with engine location ohc.

3.4 Bar plots of engine location against mean price

Cars with engine location ohcv are generally the most expensive, while cars with engine location ohcf are generally the cheapest.

4.1 Bar plots of fuel system against mean city-mpg and highway-mpg

The highest mean city-mpg and highway-mpg are the highest in cars with fuel system lbbl, lowest in that cars with fuel system mfi.

4.2 Bar plots of fuel system against mean bore and stroke

Cars with fuel system mfi have the highest mean value for bore and stroke, while cars with fuel system 2bbl have lowest mean value for bore and stroke.

4.3 Bar plots of fuel system against mean bore/stroke ratio
The mean bore/stroke ratio is highest in cars with fuel system mpfi, lowest in cars with fuel system lbbl. Cars with fuel systems mpfi and 2bbl are generally short-stroke (bore/stroke ratio > 1), while cars with other fuel systems are long-stroke.
4.4 Bar plots of fuel system against mean price
Cars with fuel system idi are generally the most expensive, while cars with fuel system 2bbl are generally the cheapest.

5.1 Bar plots of num of cylinders against mean city-mpg and highway-mpg
Both the mean city-mpg and highway-mpg increases as the number of cylinders increases.
5.2 Bar plots of num of cylinders against mean bore and stroke
Cars with eight cylinders have the highest mean value for bore, but cars with five cylinders have the highest mean value for stroke. Cars with three cylinders have the lowest mean value for both bore and stroke.
5.3 Bar plots of num of cylinders against mean bore/stroke ratio
The mean bore/stroke ratio is highest in cars with eight cylinders, lowest in cars with five cylinders. Only cars with five cylinders are long-stroke, the rest are short-stroke.
5.4 Bar plots of num of cylinders against mean price
The price of cars generally increases as the number of cylinders increases; yet, the price of cars with five cylinders are generally higher than those with six cylinders.

+ Box plot of curb weight by symboling: we can see that cars with symboling 0 have the widest range of curb weight.

Thoughts on the assignment: The assignment was fun to do but also time-consuming. I think it could be improved in a lot of ways, and I hope to learn more techniques of exploratory analysis in the following class, especially when we get to use scikit-learn.