ACTIVITY 2

Activity Reference:
https://www.kaggle.com/code/sachinrajput17/exploratory-data-analysis-for-automobile-data/notebook

Download the auto-mpg dataset and do the following:

1. Drop column Origin from the dataset.
2. Find out all the missing values in the dataset and replace it with its most appropriate replacement.
   Finding out missing values: We can use .sum() method after applying .isnull() this will return sum of missing values within each columns in the dataframe.
   ```
   import pandas as pd
    df = pd.read_csv('IMDB-Movie-Data.csv')
   df.isnull().sum()
   ```

Replacement Decision: How to decide? May be take help from below provided link:
Link 1:

Example: df['salary'] = df['salary'].fillna(df['salary'].mode())

3. Find and remove duplicate entries for the column ' CAR Number'
4. Get the basic insights:
   - ✔ Display the first five and last five rows
   - ✔ Display all the column names in the dataset
   - ✔ Display the concise summary of your dataset
   - ✔ Display the name of the car with maximum number of horsepower.

5. In our dataset, the fuel consumption column is "mpg" and is represented by mpg (miles per gallon) unit. Assume we are developing an application in a country that accepts fuel consumption with the L/100km standard. **change the name of the column to "L/100km".**
6. Normalize the column "Weight" so that the values range from 0 to 1
7. Normalization is the process of transforming values of several variables into a similar range. Typical normalizations include scaling the variable so the variable average is 0, scaling the variable so the variance is 1, or scaling variable so the variable values range from 0 to 1
8. **Approach:** replace original value by (original value)/(maximum value)
9. In our dataset, "horsepower" is a real valued variable ranging from 48 to 288, it has 57 unique values. What if we only care about the price difference between

cars with high horsepower, medium horsepower, and little horsepower (3 types)? Can we rearrange them into three ‘bins' to simplify analysis?

10. We will use the Pandas method 'cut' to segment the 'horsepower' column into 3 bins
11. Use a histogram to visualize the distribution of bins we created above.
12. Detect outliers using Z-score and remove them