Assignments

1. Download the Automobile dataset from the LMS. The original dataset was available at https://archive.ics.uci.edu/dataset/10/automobile. Please use the dataset given in the LMS for the following exercises. You must develop the Python codes appropriately to answer the questions. You must submit your Jupyter notebooks with the answers.

a. Write a brief description about the dataset. Clearly mention the data type of each column.
b. Create a Pandas DataFrame (df) with the above dataset and add a header row with appropriate labels.
c. Find the number of missing values in each column.
d. Find the number of columns with duplicates and unique rows.
e. Print all the duplicated rows.
f. Drop all the duplicated rows and create a new DataFrame (df_new) with the unique rows.
g. Drop the rows with missing values from df_new and find the number of remaining rows. Do you think it is safe to remove the missing values?
h. Fill the missing values of df_new with the following (separately).
    I. Mean values of the numerical columns and most frequent non-null value for string columns (if exist).
    II. Median values of the numerical columns and most frequent non-null value for string columns (if exist).
i. Explain the importance of one-hot encoding.
j. Perform one-hot encoding for the categorical data in df_new. and print those columns after the transformation.

2. Find the following statistical parameters for each numerical column of the Automobile dataset in the DataFrame of **Part 1 h I**.

a. Mean
b. Median
c. Mode
d. Variance
e. Standard deviation
f. Minimum
g. Maximum