**SQL Hands-on-3**

**Problem Statement:**

Amir is working as a Data analyst in an XYZ automobile company. The company is in talks to launch some old models with some changes and renovations. To do this he was asked to work on the automobile dataset of the year 1985. Help Amir sort, clean, and process the data to analyze the facts that are required for the launch of Upcoming models for the company.

The dataset provided to the data analyst is [here](https://www.kaggle.com/datasets/toramky/automobile-dataset).

**Dataset Description:**

This data set consists of the following columns:

1. **Symboling**: It is the insurance risk level of the car. (3 is riskier where as -3 is pretty safe).
2. **Normalized losses**: It is the relative average loss payment per insured vehicle year and represents the average loss per car per year.
3. **Make**: It is the brand of the vehicle.
4. **Fuel\_type**: The type of fuel used by the vehicle.
5. **Aspiration**: The type of engine aspiration used by vehicles.
6. **Num-of-doors**: The number of doors the vehicle is having.
7. **Body-style**: The body style of the vehicle.
8. **Drive\_wheels**: The type of wheels the vehicle is using.
9. **Engine\_location**: The location of the engine in the vehicle.
10. **Wheel\_base**: The base of the wheel.
11. **Length, width, height, and curb\_weight** are the quantitative attributes of the vehicle.
12. **Engine\_type**: The type of engine used by the vehicle.
13. **Num-of-cylinders**: The number of cylinders used by the vehicle.
14. **Engine\_size**: The size of the engine used by the vehicle.
15. **Fuel\_system**: The type of fuel used by the vehicle.
16. **Bore:** The inner diameter of the cylinder used in the engine of the vehicle.
17. **Stroke**: The type of power cycle used by a piston engine by the vehicle.
18. **Compression ratio**: The compression ratio of the vehicle.
19. **Horsepower**: The power of the vehicle.
20. **Peak\_rpm**: How fast any vehicle can operate at any given time.
21. **City\_mpg:** The score a car will get on average in city conditions, with stopping and starting at lower speeds.
22. **Highway\_mpg:** the average a car will get while driving on an open stretch of road without stopping or starting, typically at a higher speed
23. **Price**: The price of the vehicle.(Note: the data is from 1975 so the prices may be low).

**Tasks:**

Amir has divided the task into two categories. The first one is Data cleaning where you will be performing necessary queries in order to make the data-efficient for accurate analysis.

Help him to clean the data by performing the following set of tasks.

**Sub-Tasks-1: Data Cleaning.**

1. Amir is discussing with Pranav how to start the tasks and Pranav suggested that they need to create and insert values manually by writing the script. But Amir thought it would not be an inefficient way. Suggest and implement an efficient way to create and insert values of the automobile data in the SQL Server Management Studio.
2. Pranav is worried about the amount of data he is gonna work with. Help him write a query to find out the size of the data he is dealing with.
3. While checking the attribute values in the database, Amir has found out that there are some missing values in the following columns.
4. normalized-losses
5. bore
6. stroke
7. horsepower
8. peak-rpm
9. price

Since the data is large he is having a hard time figuring out how many missing values are there in each of these columns.

Write a set of queries in order to display a number of missing values in each column.

1. As a next step in cleaning the data, Arun suggests Pranav replace missing values in each of these columns with the null values.

Use an efficient way to replace the missing values in each of the columns mentioned above with the NULL value.

1. Amir has a really good eye for analysis. He states that NULL values wouldn’t contribute to the analysis so he asks Arun to replace all of the null values with the mean of the column if it is a numeric value or with the mode, if it is a categorical column. Help arun complete the task.

Now, that Amir has made sure the data is cleaned, he wants to perform the following tasks and help the company decide on some important factors related to future releases.

**Sub-Tasks: 2. Data Analysis**

1. What are the brands that were present in 1985
2. How many cars are from the brand 'Nissan'?
3. Which brand is having the highest number of vehicles to sell.
4. How many vehicles are with least insurance risk
5. Which brand vehicle is of least insurance risk?
6. What are the different fuel types involved with the vehicles?
7. How many diesel vehicles do we have in the given dataset.
8. Display the fuel\_type and the number of vehicles with each fuel type.
9. How many sedan type of cars is up for sale in the Nissan brand.
10. What is the highest price of the car?
11. Which one is the costliest brand among all the cars in the given data?
12. What is the highest price of a sedan car?
13. Which one is the costliest hatchback car?
14. Display the engine\_size whose compression ratio is least.
15. What is the highest horsepower of the cars available in the given dataset?
16. Which brand is offering the highest power among all the vehicles.