

**School of Computer Science**

**Masters in applied computing (M.A.C)**

**Advanced Database Topics- COMP 8157**

**Professor Dr. Shafaq Khan**

**Lab Assignment**

**by**

**Rohit Kumar (110088741)**

Table of Contents

[**Part 1: Data Exploration** 3](#_Toc135946388)

[1. Import and Summarized the Vehicle dataset: 3](#_Toc135946389)

[2. Structure and Dimension of Vehicle dataset: 4](#_Toc135946390)

[3. Display the rows: 4](#_Toc135946391)

[4. Average km driven for each car type: 5](#_Toc135946392)

[5. Average selling price of the cars per year: 6](#_Toc135946393)

[6. Display the unique combination: 6](#_Toc135946394)

[7. Display the different combination and sort it based on count: 7](#_Toc135946395)

[**Part 2: Data Pre-Processing:** 9](#_Toc135946396)

[8. Display missing values in dataset: 9](#_Toc135946397)

[9. Display columns with missing values: 9](#_Toc135946398)

[10. Replaced missing values with most repeated: 10](#_Toc135946399)

[11. Find and remove duplicate rows if exit. 10](#_Toc135946400)

[12. Replace the attribute values: 11](#_Toc135946401)

[13. Adding a new field: 12](#_Toc135946402)

[14. Create a new dataset: 13](#_Toc135946403)

[15. Shuffling the rows randomly: 13](#_Toc135946404)

[**Part 3: Data Visualization:** 14](#_Toc135946405)

[16. Create a scatter plot: 14](#_Toc135946406)

[17. Creating a box plot 16](#_Toc135946407)

[18. Scatter plot using k-means: 16](#_Toc135946408)

[19. Scatter plot using hierarchical clustering: 17](#_Toc135946409)

[20. Creating a bar plot: 18](#_Toc135946410)

[21. Creating a correlation plot: 18](#_Toc135946411)

[22. Scatter plot using DBSCAN clustering: 19](#_Toc135946412)

# **Part 1: Data Exploration**

## Import and Summarized the Vehicle dataset:

Q. Import the Vehicle dataset, summarize it and explain the output.

A picture containing text, font, white, screenshot

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Summary function, summarizes all the information and provide information about all the columns such as min. value of each column, max. value of each column, mean value of each column, median value of each column, etc.



A picture containing text, font, screenshot, number

Description automatically generated

## Structure and Dimension of Vehicle dataset:

Q. Show the structure and dimension of the dataset and explain it.

Str function in R, explains the structure of vehicle dataset, the number of columns in dataset with data types and the possible values.

A picture containing text, font, line, white

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Dim function provides the dimension of vehicle dataset number of rows and columns.

A picture containing text, screenshot, font, line

Description automatically generated

A picture containing text, screenshot, font, line

Description automatically generated

## Display the rows:

Q. Show the column names of the Vehicle dataset and the first 3 rows and the last 6 rows of it.

A picture containing text, font, screenshot, white

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, font, screenshot, white

Description automatically generated

A screenshot of a computer screen

Description automatically generated with low confidence

## Average km driven for each car type:

Q. Show the average Kms\_Driven for each type of car (Car\_Name) in the dataset.

A picture containing text, font, line, screenshot

Description automatically generated

A screenshot of a computer

Description automatically generated

## Average selling price of the cars per year:

Q. What is the average Selling\_Price of the cars in each year?

A picture containing text, screenshot, font, line

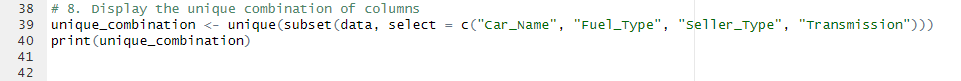
Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

## Display the unique combination:

Q. Show the unique combinations of Car\_Name, Fuel\_Type, Seller\_Type, and Transmission in the Vehicle dataset.



A screenshot of a computer

Description automatically generated with medium confidence

## Display the different combination and sort it based on count:

Q. What are the different combinations of Car\_Name, Fuel\_Type, Seller\_Type, and Transmission in the Vehicledataset, and how many times does it occur? (Display all such in both ascending and descending orders)

A picture containing text, screenshot, font

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated with medium confidence

# **Part 2: Data Pre-Processing:**

## Display missing values in dataset:

Q. Find if there are any missing values in the Vehicle dataset.

A picture containing text, font, line, white

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

## Display columns with missing values:

Q. Find which columns contain missing values in the vehicle’s dataset. What are the total missing values for each column?

A picture containing text, font, line, screenshot

Description automatically generated

A picture containing text, font, line, screenshot

Description automatically generated

## Replaced missing values with most repeated:

Q. Replace the missing values in the dataset with the most repeated value of that field. Check if the missing values were replaced successfully.

A picture containing text, screenshot, font

Description automatically generated

A computer code on a white background

Description automatically generated with low confidence

## Find and remove duplicate rows if exit.

Q. Find if the dataset has duplicate rows. Remove them, if exist.

A screen shot of a computer code

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated with low confidence

## Replace the attribute values:

Q. Replace the values of the following attributes:

a Fuel\_Type: “Petrol”: 0, “Diesel”: 1, “CNG”: 2

b Seller\_Type: “Dealer”: 0, “Individual”: 1

c Transmission: “Manual”: 0, “Automatic”: 1

Show the conversion output of the specific attribute.

A picture containing text, font, screenshot, line

Description automatically generated

A screenshot of a computer code

Description automatically generated with low confidence

## Adding a new field:

Q. Add a new field called ‘Age’ and input the values by using the field Year. Show the output.

A picture containing text, screenshot, font, line

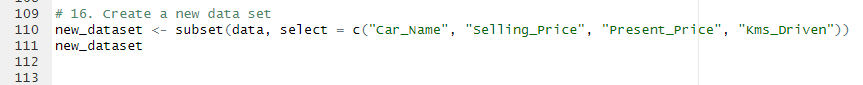
Description automatically generated

A screenshot of a computer screen

Description automatically generated with low confidence

## Create a new dataset:

Q. Create a new dataset by selecting only the columns “Car\_name”, “Selling\_Price”, Present\_Price”, and “Kms\_Drive”. Show the output of the new dataset.



A screenshot of a computer

Description automatically generated with medium confidence

## Shuffling the rows randomly:

Q. Shuffle the rows of the Vehicle dataset randomly and show the output.

A picture containing text, screenshot, font, line

Description automatically generated

A screenshot of a computer screen

Description automatically generated with medium confidence

# **Part 3: Data Visualization:**

## Create a scatter plot:

Q. Import the **Vehicle** dataset. Create a scatter plot of the Selling\_Price Vs Present\_Price. Colour code the points based on the Transmission.

a. Add labels, title, and colour to the plot. The colour should be red for Transmission type ‘0’ and blue for ‘1’.

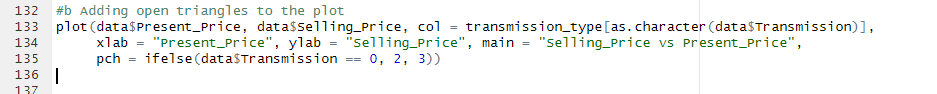
A picture containing text, font, line, screenshot

Description automatically generated

A graph with red and blue dots

Description automatically generated with low confidence

b. Add open triangles to the plot.



A picture containing text, screenshot, line, plot

Description automatically generated

c. What do you understand from the output.

As per above illustrations, it is clearly visible that manual transmission cars are higher than the automatic transmission. Additionally, selling price is less than the present price, moreover selling and present price of automatic transmission are higher in comparison to a manual transmission car type.

## Creating a box plot

Q. Create a box plot of the Selling\_Price Vs Transmission and Fuel\_Type.

A picture containing text, screenshot, font, line

Description automatically generated

A picture containing screenshot, text, diagram, plot

Description automatically generated

## Scatter plot using k-means:

Q. Create a scatter plot of the Selling\_Price Vs Kms\_Driven, and use k-means clustering to cluster the points into 4 clusters. Colour-code based on the cluster they belong to.

A screenshot of a computer code

Description automatically generated with low confidence

A picture containing text, line, plot, diagram

Description automatically generated

## Scatter plot using hierarchical clustering:

Q. Create a scatter plot of the Selling\_Price Vs Present\_Price, and use hierarchical clustering to cluster the points into 3 clusters? Colour-code the points based on the cluster they belong to.

A picture containing text, font, number, screenshot

Description automatically generated

A picture containing text, plot, line, font

Description automatically generated

## Creating a bar plot:

Q. Add a new field called ‘Age’, and calculate it using the field ‘Year’. Create a barplot for the following fields of the dataset:

a. ‘Age’, ‘Year’, ‘Transmission’, ‘Seller\_Type’, ‘Fuel\_Type’ and ‘Owner’

b. Add labels, titles, and colours to the plot.

A picture containing text, font, line, screenshot

Description automatically generated

A picture containing screenshot, line, plot, diagram

Description automatically generated

## Creating a correlation plot:

Q. Create a correlation plot of the whole dataset variables and explain the output. Do not forget to convert some of the variable’s datatype if required and possible.

A picture containing text, font, screenshot

Description automatically generated

A picture containing colorfulness, screenshot, design

Description automatically generated

With the help of correlation plot, we identify the connections between variables in the dataset and learn more about how they relate to one another by looking at the correlation plot. It aids in comprehending the relationships and linkages between various automotive attributes in the dataset.

The selling price and present price of the cars in the dataset have a high positive association because the correlation between "Selling\_Price" and "Present\_Price" is positive. This means that as the present price increases, the selling price also tends to increase.

The correlation between "Transmission" and "Fuel\_Type" is close to zero, which means there is little to no association between the cars' fuel type and transmission type. This means that the type of fuel used by a car does not have a significant influence on the type of transmission it has.

## Scatter plot using DBSCAN clustering:

Q. Create a scatter plot of the Selling\_Price Vs Kms\_Driven, and use DBSCAN clustering to cluster the points into 3 clusters. Color-code based on the cluster they belong to. Add a legend to the plot.

A picture containing text, font, screenshot

Description automatically generated

A picture containing text, line, plot, screenshot

Description automatically generated