**LAB 1**

Author:

|  |  |
| --- | --- |
| **Abhirup Ranjan** | **110091866** |

Submitted To:

|  |
| --- |
| **Professor** |
| **Dr. Shafaq Khan** |

**Table of Contents**

[1. Part 1: Data Exploration 2](#_Toc135936914)

[2. Part 2: Data Pre-Processing 8](#_Toc135936915)

[3. Part 3: Data Visualization 12](#_Toc135936916)

# **Part 1: Data Exploration**

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

head keyword is used to display the top 6 lines from the csv dataset imported whereas the summary keyword give the entire overview detail of the dataset.

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with low confidence

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

Keyword ‘str’ is used for provided the dataset structure of every field.

A screenshot of a computer

Description automatically generated with medium confidence

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

A screenshot of a computer

Description automatically generated with medium confidence

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

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

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

A screenshot of a computer

Description automatically generated with medium confidence

1. 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

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

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

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

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

A screenshot of a computer

Description automatically generated with low confidence

1. Find which columns contain missing values in the vehicles dataset. What are the total missing values for each column

A screenshot of a computer

Description automatically generated with low confidence

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

A screenshot of a computer

Description automatically generated with medium confidence

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

A screenshot of a computer

Description automatically generated

1. Replace the values of the following attributes:
   1. Fuel\_Type: “Petrol”: 0, “Diesel”: 1, “CNG”: 2

A screenshot of a computer

Description automatically generated with medium confidence

* 1. Seller\_Type: “Dealer”: 0, “Individual”: 1

A screenshot of a computer

Description automatically generated

* 1. Transmission: “Manual”: 0, “Automatic”: 1

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated with medium confidence

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

A screenshot of a computer

Description automatically generated

1. 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

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

A screenshot of a computer

Description automatically generated with low confidence

# **Part 3: Data Visualization**

1. Import the Vehicle dataset. Create a scatter plot of the Selling\_Price Vs Present\_Price. Colour code the points based on the Transmission
   1. 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, screenshot, font, design

Description automatically generated

A graph with red and blue dots

Description automatically generated with low confidence

* 1. Add open triangles to the plot.

A screen shot of a computer

Description automatically generated with low confidence

A picture containing text, screenshot, line, plot

Description automatically generated

* 1. What do you understand from the output.

It is evident from the above diagram that manual vehicles are more than the automatic vehicles. 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.

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

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing screenshot, text, diagram, plot

Description automatically generated

1. 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 picture containing text, screenshot, font

Description automatically generated

A picture containing text, line, plot, diagram

Description automatically generated

1. 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 screenshot of a computer code

Description automatically generated with medium confidence

A picture containing text, plot, line, font

Description automatically generated

1. Add a new field called ‘Age’, and calculate it using the field ‘Year’. Create a barplot for the following fields of the dataset:
   1. ‘Age’, ‘Year’, ‘Transmission’, ‘Seller\_Type’, ‘Fuel\_Type’ and ‘Owner’

A screenshot of a computer

Description automatically generated

* 1. Add labels, titles, and colours to the plot.

A picture containing text, screenshot, font, line

Description automatically generated

A picture containing screenshot, line, plot, diagram

Description automatically generated

1. 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

By examining the correlation plot, we may identify the relationships between the dataset's variables and discover more about how they relate to one another.  It helps to understand the connections and correlations between the different automobile attributes in the dataset.

In this case, the correlation between "Selling\_Price" and "Present\_Price" is positive, indicating a strong positive link between the selling price and present price of the cars in the dataset. This implies that the selling price tends to climb as the current price does.

There is little to no correlation between the transmission type of the cars and the fuel type because the correlation between "Transmission" and "Fuel\_Type" is almost negligible. This indicates that a car's gearbox type is not much influenced by the type of fuel it uses.

A screenshot of a computer program

Description automatically generated with medium confidence

A picture containing colorfulness, screenshot, design

Description automatically generated

1. 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 screenshot of a computer code

Description automatically generated with low confidence

A picture containing text, line, plot, screenshot

Description automatically generated