**Principles of Data Science (5530-0001)-Assignment 2**

**Name: Tanuja Ramani Paricherla**

**ID: 16355123**

Given question:

The provided data (link above) contains various details and attributes associated with used cars. The

target variable, which is the central focus of analysis, is the price of the used cars, and it is measured in

lakhs. The data in this dataset is tabular, with rows and columns, where each row represents a specific

used car listing, and each column represents a particular attribute or feature of these cars. Features are

Make and model of the car, Location or city of sale, Year of manufacture, Mileage, Odometer

(kilometers driven), Fuel type (petrol or diesel), Transmission type (manual or automatic), Number of

owners, Engine displacement, Engine horsepower, Number of seats, and Price when the car was new.

Use this data to perform the following:

**a) Look for the missing values in all the columns and either impute them (replace with mean,**

**median, or mode) or drop them. Justify your action for this task.**

**b) Remove the units from some of the attributes and only keep the numerical values (for**

**example remove kmpl from “Mileage”, CC from “Engine”, bhp from “Power”, and lakh from**

**“New\_price”).**

**C) Change the categorical variables (“Fuel\_Type” and “Transmission”) into numerical one hot**

**encoded value.**

**d) Create one more feature and add this column to the dataset (you can use mutate function in**

**R for this). For example, you can calculate the current age of the car by subtracting “Year” value**

**from the current year.**

**e) Perform select, filter, rename, mutate, arrange and summarize with group by operations (or**

**their equivalent operations in python) on this dataset.**

Before Starting these tasks,

I performed data cleaning on the raw data which in internally comes under first question.

**a) Look for the missing values in all the columns and either impute them (replace with mean,**

**median, or mode) or drop them. Justify your action for this task.**I dropped the null values from the dataset and, I am dropping the first column i.e., unnamed column with the serial number. I am dropping this column because the serial numbers are not in order.

Coming to the question asked.

I am identifying the missing values in all columns. For the categorical columns, I am imputing the missing values with the mode because if we use the mode to impute missing values, you can preserve the data that is currently in that category field. By guaranteeing that the imputed values are representative of the complete collection of data, this reduces the likelihood of bias.

For the numerical columns, I am imputing the missing values with the mean because by imputing missing values using the mean, the distribution and central tendency of the available data in the numerical column are maintained. This technique ensures that the imputed values correspond to the distribution of the entire data set and reduces the likelihood of introducing discernible bias.

**A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated**

**b) Remove the units from some of the attributes and only keep the numerical values (for example remove kmpl from “Mileage”, CC from “Engine”, bhp from “Power”, and lakh from “New\_price”).**

I am removing the units from the attributes of Mileage, Engine, Power, New\_price columns and keeping only the numerical values.

A screenshot of a computer

Description automatically generated

**C) Change the categorical variables (“Fuel\_Type” and “Transmission”) into numerical one hot encoded value.**

In this question first I am encoding the categorical variables “Fuel\_Type” and “Transmission” into numerical values using dummies from pandas.

A screenshot of a computer

Description automatically generated

**d) Create one more feature and add this column to the dataset (you can use mutate function in R for this). For example, you can calculate the current age of the car by subtracting “Year” value from the current year**

Then I am adding a new column Current\_Age to the dataframe.

A screenshot of a computer

Description automatically generated

**e) Perform select, filter, rename, mutate, arrange and summarize with group by operations (or their equivalent operations in python) on this dataset.**

1. Select: I am selecting only 'Name', 'Location', 'Year', and 'Price' to show this operation.

2. Filter: I am filtering for cars from the year 2015 and 'First' owner type narrows down the dataset to only these matching rows.

3. Rename: I am renaming Kilometers\_Driven to KM\_Driven

4. Mutate: I have created a new column Price\_per\_KM which is calculated by dividing Price by Kilometers\_Driven, giving insight into the price per kilometer driven.

5. Arrange: I have done Sorting by 'Price' in descending order which shows most expensive cars at top

6. Summarize: Summarize mostly working with grouping aggregates data based on categories. I am grouping by Location and calculating the average price provides insights into how car prices vary by location.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated