**Assignment 2**

**Mallikarjun Challa**

**16342908**

Data link:  <https://app.box.com/s/jm6pw202asu4xd3uypwtry2rqk691y1i>

1) 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.     (4 points)**

First, I checked for missing values in all columns of the dataset using the isnull().sum() function. This gave me a count of missing values in each column.

A screenshot of a computer

Description automatically generated

I chose to use the median of the corresponding columns to impute the missing values for the columns labeled "Mileage," "Engine," "Power," and "Seats." Since the median is less susceptible to outliers and extreme values—which are common in datasets about used cars—it was chosen over the mean. The mean could be skewed by the fact that certain used cars may have traveled a notably greater distance than others.

A 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”). (4 points)

**Handling Removal of Units from Attributes:**

To remove units from certain attributes while retaining only the numerical values, the following steps were taken:

1. **Mileage, Engine, and Power:**
   * Regular expressions were used to extract numerical values, discarding units.
   * Data types were converted accordingly (float for 'Mileage' and 'Power', numeric for 'Engine').
   * Missing values were filled with respective attribute medians.
2. **Price:**
   * 'Lakh' units were removed using string replacement, converting the column to float.
3. **New\_Price:**
   * A custom function removed units ('Lakh' and 'Cr'), converting 'Cr' values to lakhs.
   * Missing values were filled with the median of 'New\_Price'.
4. **Seats:**
   * Missing values were imputed with the median of 'Seats'.

C) Change the categorical variables (“Fuel\_Type” and “Transmission”) into numerical one hot encoded value. (4 points).

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.   (4 points)**

‘Fuel\_Efficiency\_Category’. This feature would categorize cars based on their mileage (kmpl or km/kg)

This function in this code divides mileage into three categories:

'Low' (less than 10 km/l)

'Medium' (10–20 kmpl)

'High' (more than 20 kmpl).

To create the new "Fuel\_Efficiency\_Category" feature, this function is then applied to the "Mileage" column.

I have also created a new feature ‘Kilometers\_Driven\_per\_Year’ by dividing the ‘Kilometers\_Driven’ by the age of the car.

df['Kilometers\_Driven'] / df['Car\_Age'] calculates the kilometers driven per year for each car. The result is stored in a new column ‘Kilometers\_Driven\_per\_Year’.

A white sheet with black text

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. (4 points)

  # Select

selected\_df = df[['Name', 'Price']]

print("Selected DataFrame:")

print(selected\_df.head())

print("\n")

# Filter

filtered\_df = df[df['Location'] == "Pune"]

print("Filtered DataFrame:")

print(filtered\_df.head())

print("\n")

# Rename

renamed\_df = df.rename(columns={'Year': 'Manufacture\_Year'})

print("Renamed DataFrame:")

print(renamed\_df.columns)

print("\n")

# Mutate

df['Car\_Age'] = datetime.datetime.now().year - df['Year']

print("DataFrame after Mutation:")

print(df[['Year', 'Car\_Age']].head())

print("\n")

# Arrange

sorted\_df = df.sort\_values('Price', ascending=False)

print("Sorted DataFrame:")

print(sorted\_df[['Name', 'Price']].head())

print("\n")

# Summarize with group by

summary\_df = df.groupby('Fuel\_Type\_Diesel')['Price'].mean()

print("Summary DataFrame:")

print(summary\_df)

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Submission:

Create a public GitHub repo and upload the folders for the assignment on the GitHub and submit the link to Canvas.