**Assignment 2**

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)**

For task a – I found various missing values: mileage had 2 missing entries, engine had 36 missing entries, power had 36 missing entries, seats had 38 missing entries, and new price had 5032 missing entries.

A screenshot of a computer

Description automatically generated

Here is how I handled each of the missing entries: A screenshot of a computer program

Description automatically generated

I dropped mileage, engine, and power since they were small portions of the overall dataset. Additionally, I figured if I just put in an arbitrary value for each of these items it would compromise the dataset more than simple removal would.

For seats, I entered the mode since there is typically less variation in seat numbers for vehicles than there is in mileage or engine stats.

For new\_price, I cloned the original row into two separate rows and then modified the rows such that there is now only one price column and the new rows reflect what values a new car would have. Not only does this expand the data set, but now there is no confusion in the price metric if it were used as a target variable.

**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)**

This part was fairly straightforward. See my methodology below

A screenshot of a computer code

Description automatically generated

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

**encoded value. (4 points).**

Encoded True False Boolean values for the one-hot encode. And also created a new car age column for the next question.

A screenshot of a computer program

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)**

**e) Perform select, filter, rename, mutate, arrange and summarize with group by operations (or their equivalent operations in python) on this dataset. (4 points)**

   Some operations I did for this section

A screenshot of a computer code

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.