Assignment 2

1) The provided data 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)

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)

C) Change the categorical variables ("Fuel_Type" and "Transmission") into numerical one hot encoded value. (4 points).

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)

2) The data file diabetes.csv contains data of 768 patients. In this data there are 8 attributes (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age) and 1 response variable (Outcome). The response variable, Outcome, has binary value (1 indicating the outcome is diabetes and 0 means no diabetes). For this assignment purposes we will consider this data as a population. Use this data to perform the following:

a) Set a seed (to ensure work reproducibility) and take a random sample of 25 observations and find the mean Glucose and highest Glucose values of this sample and compare these statistics with the population statistics of the same variable. You should use charts for this comparison. (5 points)

b) Find the 98th percentile of BMI of your sample and the population and compare the results using charts. (5 points)

c) Using bootstrap (replace= True), create 500 samples (of 150 observation each) from the population and find the average mean, standard deviation and percentile for BloodPressure and compare this with these statistics from the population for the same variable. Again, you should create charts for this comparison. Report on your findings. (10 points)

Submission:
Create a public GitHub repo and upload the folders for the assignment on the GitHub and submit the link to Canvas. **Please note that the work needs to be organized as folders and subfolders and only notebooks are acceptable source files.**