Assignment 2

Data link Q1: 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)

I first checked the dataset for missing values. The column **New_Price** had more than **85% missing values**, which makes imputation unreliable and would introduce noise into the dataset.
Therefore, I dropped the New_Price column.

The **Seats** column had only a few missing values. Since it is a categorical-type numeric feature (2, 4, 5, 7 seats),

I imputed missing values in Seats using the mode, because the most frequently occurring seat count is appropriate.

For **Mileage, Engine, and Power**, the columns originally contained units (kmpl, CC, bhp). After converting them to numeric, each had a small number of missing values.
I imputed these missing values using the median, because these variables may contain outliers and the median is more robust than the mean.

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)

I removed the textual units from the following columns and kept only the numeric values:

- **Mileage:** removed "kmpl" / "km/kg"
- **Engine:** removed "CC"
- **Power:** removed "bhp"
- **New_Price:** removed "Lakh" (numeric part extracted before column was dropped in part a)

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

The categorical variables:

- **Fuel_Type** (Petrol, Diesel, CNG, etc.)
- **Transmission** (Manual / Automatic)

were converted into numerical format using **one-hot encoding**.
This created new binary columns such as:

- Fuel_Type_Diesel, Fuel_Type_Petrol, etc.
- Transmission_Manual or Transmission_Automatic

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)

I created a new feature called **Car_Age**, representing how old each car is:

$$\text{Car\_Age} = 2025 - \text{Year}$$

This feature captures vehicle age, which is important because older vehicles typically have lower prices.
I also created an additional feature:

$$\text{Price\_per\_KM} = \frac{\text{Price}}{\text{Kilometers\_Driven} + 1}$$

This gives an idea of the cost-efficiency of a car.

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 a subset of important columns such as:
Name, Year, Price, Car_Age, Kilometers_Driven

Filter:

- Cars manufactured **after 2015**
- Cars priced **between 5 and 15 lakhs**

Rename:

- Renamed Name → Car_Model
- Renamed Kilometers_Driven → KM_Driven

Mutate:
Created a categorical feature:

- **Age_Category** (New, Medium, Old) based on Car_Age

Arrange:
Sorted cars by **Price (descending)** to find the most expensive cars.

Summarize + Group By:

Grouped by **Owner_Type** and calculated:

- Count
- Mean Price
- Min/Max Price
- Median
- Standard Deviation

This helped understand how resale prices vary with number of previous owners.

Data link Q2: https://app.box.com/s/7qv44umhw0vnzgmoe9krfkfkv5kf2atv

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:

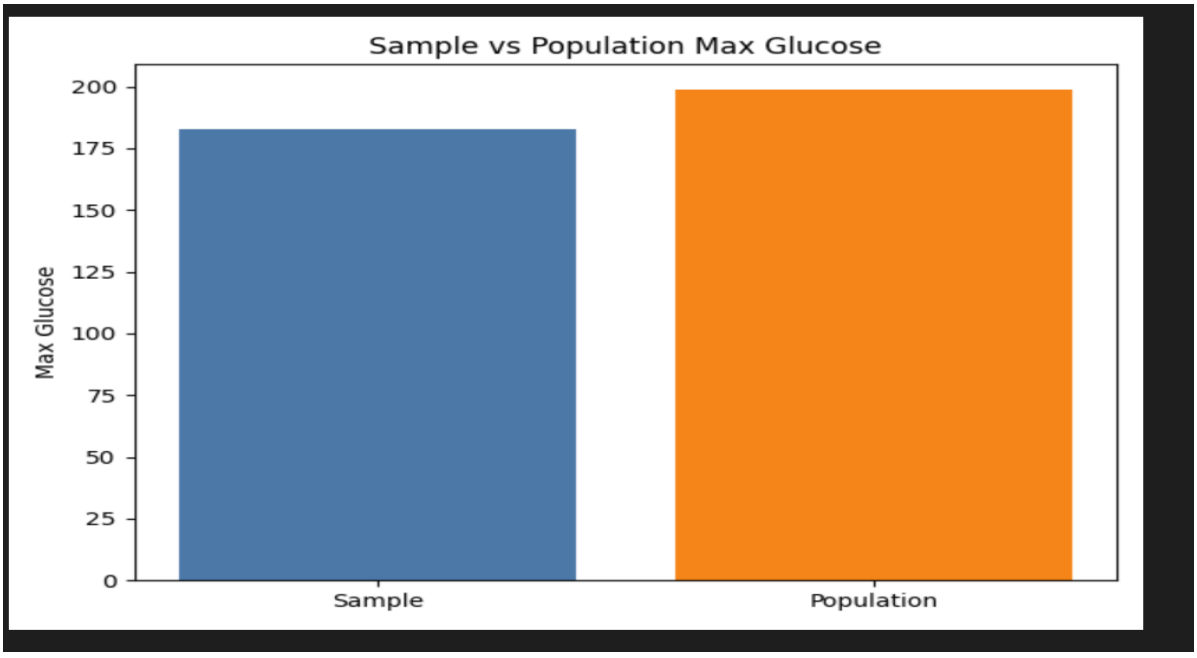
a) Set a seed (to ensure work reproducibility) and take a random sample of 25 observations and find the mean Glucose and highest Glucose values of this sample and compare these statistics with the population statistics of the same variable. You should use charts for this comparison. (5 points)

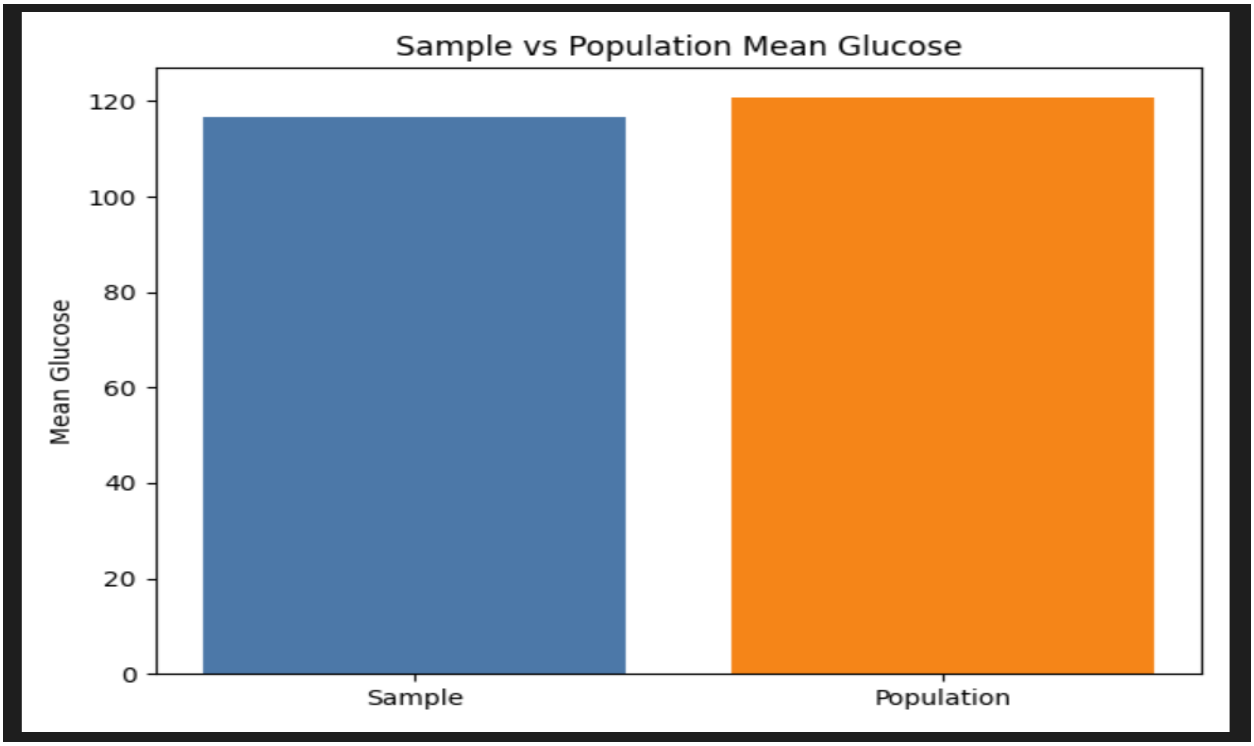
I set a random seed for reproducibility and took a **sample of 25 observations**.

- **Sample mean Glucose** was compared to **Population mean Glucose (all 768 patients)**.
- **Sample maximum Glucose** was compared to **Population maximum Glucose**.

I created two bar charts:

1. Sample vs Population **Mean Glucose**

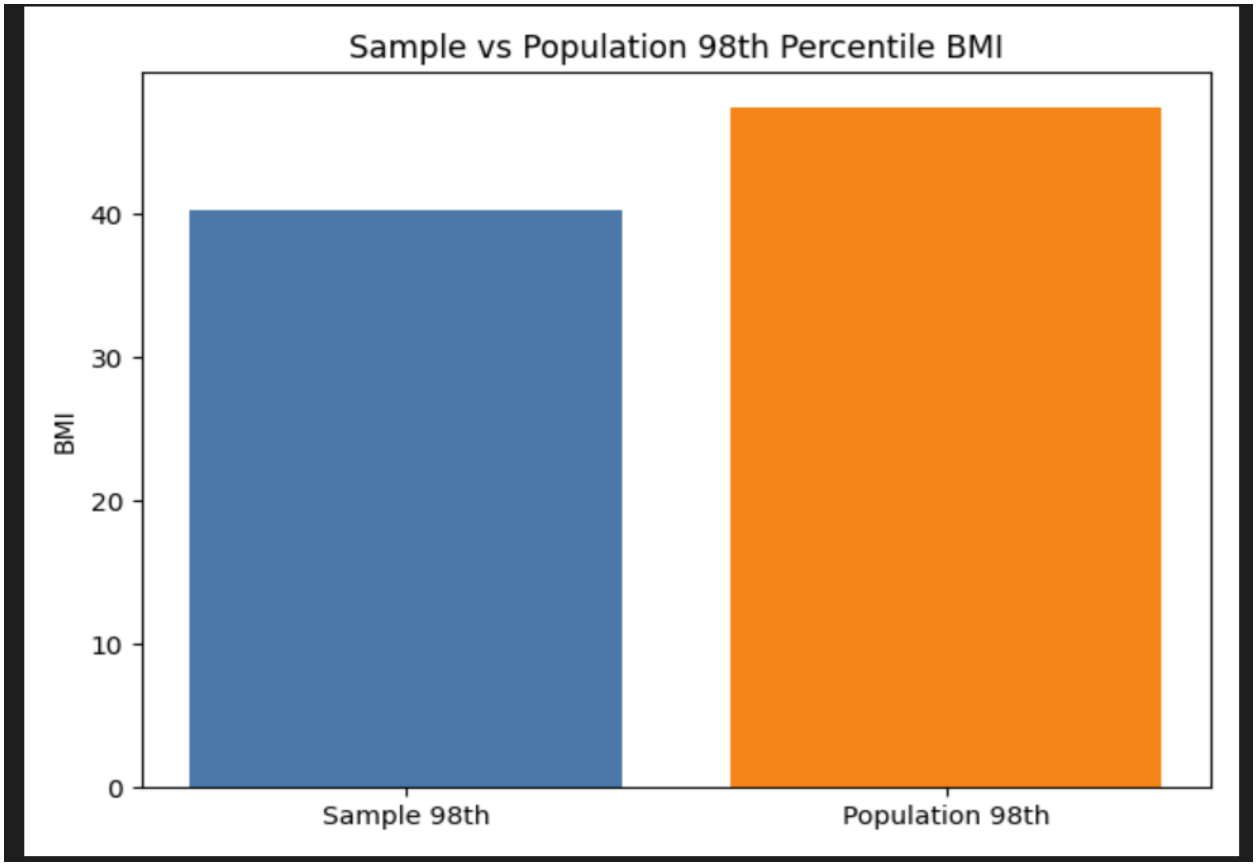




Findings:
Because the sample size is small (n = 25), the sample mean and max differed from the population, showing sampling variability. The population statistics are more stable due to larger size.

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

I computed the **98th percentile of BMI** for:



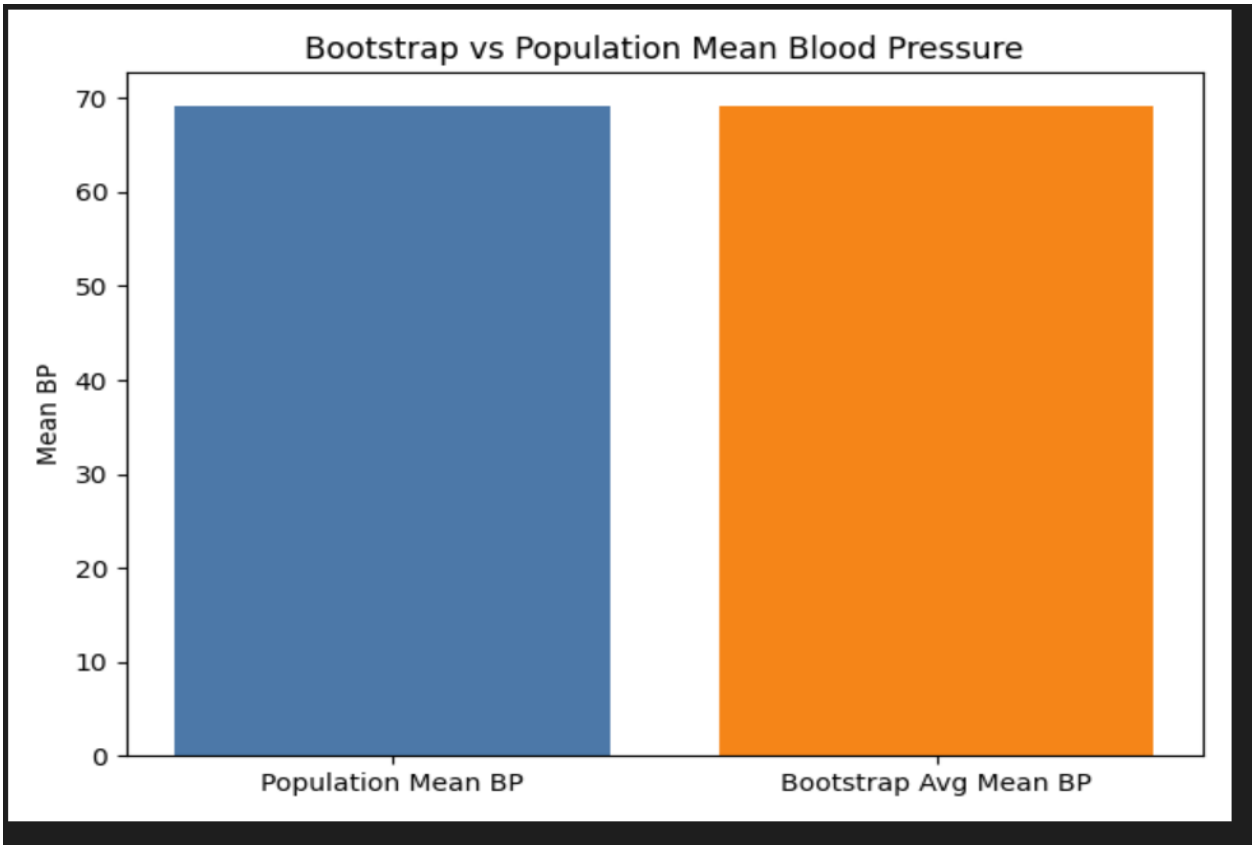
- The sample of 25
- The entire population of 768

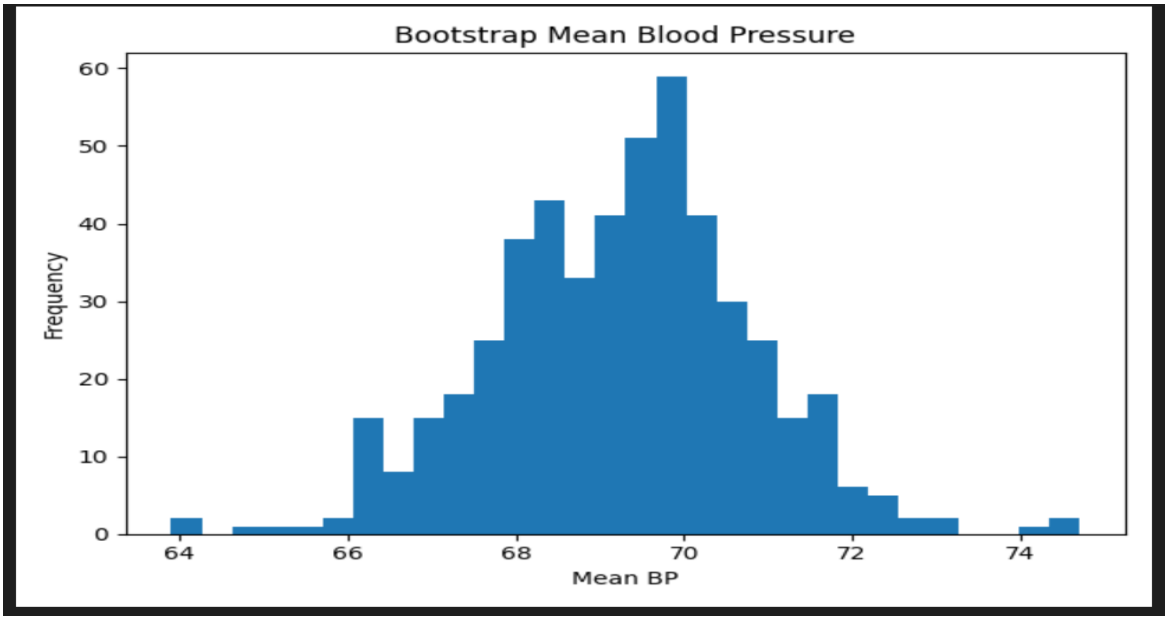
A bar chart was used to compare both values.

Findings:
The sample 98th percentile differed noticeably from the population 98th percentile, again because extreme percentiles are highly sensitive to sample size. Smaller samples tend to produce more variable high-percentile estimates.

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)

Using bootstrap (sampling **with replacement**):





- I created **500 bootstrap samples**, each of size **150**.
- For each sample, I computed:
  - Mean BloodPressure
  - Standard deviation
  - A chosen percentile (98th)

I then compared the **average bootstrap statistics** to the **population statistics** for BloodPressure:

- Population mean
- Population standard deviation
- Population percentile

I plotted histograms and bar charts showing the differences.

Findings:

- The **average bootstrap mean** was very close to the **population mean**, showing bootstrap is an unbiased estimator.
- The **average bootstrap standard deviation** slightly differed from the population SD, which is normal because sample SD fluctuates more.
- The **bootstrap percentile** also approximated the population percentile but had more variability.