Principles of Data Science

Assignment 3

Bhavani Adula

16369897

Report Summary

In this assignment, I worked with the diabetes dataset to perform statistical analysis and visual comparisons between a random sample and the overall population. The goal was to assess the representativeness of sample statistics and explore the use of bootstrapping for deeper insights.

Part A: Data Sampling and Glucose Comparison

A simple random sample of 25 entries was taken from the full dataset. For both the sample and the population, I computed the **mean** and **maximum** glucose levels. The sample's mean glucose turned out to be close to the population average, but the **maximum value showed noticeable variance**, which is expected given the small sample size. I visualized these results using bar charts to highlight the differences, which showed that while averages can be reliable even in small samples, extreme values may vary significantly.

Part B: 98th Percentile of BMI

I calculated the **98th percentile of BMI** for both the sample and the population. The values differed slightly, again reinforcing those rare statistical measures (like percentiles near the tail) can be heavily influenced by the sample size and variability. The results were displayed using a bar plot, showing the percentile difference between groups.

Part C: Bootstrapping for Glucose Mean

To estimate the sampling distribution of the mean glucose level, I applied the **bootstrap sampling method**. I drew multiple resamples (with replacement) from the original sample and recalculated the mean glucose for each. This process allowed me to visualize the variability and expected range of the sample mean through a histogram. The distribution appeared approximately normal, supporting the idea that bootstrapping is an effective tool for estimating sampling uncertainty, even with a small initial sample.