Final Project Write-Up: Pima Indians Diabetes Database

**Introduction**

The goal of this project is to explore the Pima Indians Diabetes Database to predict whether or not a patient has diabetes based on diagnostic measurements. The dataset is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases and consists of various medical predictor variables.

First, I loaded the dataset and displayed the first few rows to understand its structure.

Then, examined the data types and structure of the dataset.

**Question**:

* How does BMI affect the likelihood of diabetes?

I chose this question because BMI (Body-mass index) is a crucial indicator of health, and understanding its relationship with diabetes can provide valuable insights.

**Distribution of BMI**

I explored the distribution of BMI values in the dataset.

**BMI vs Outcome**

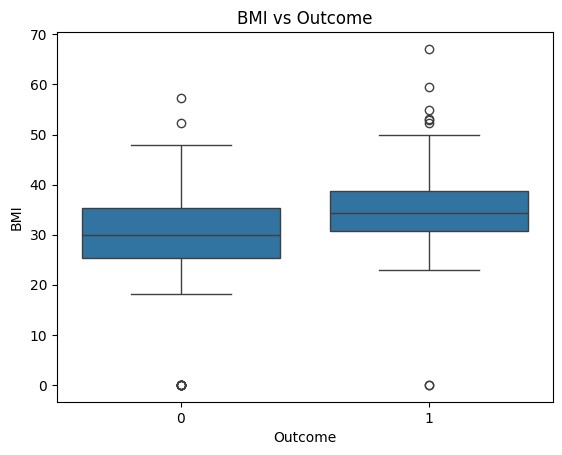
I analyzed the relationship between BMI and diabetes outcome.

* Is there a correlation between the number of pregnancies and diabetes?

**Correlation Heatmap**

I created a heatmap to visualize correlations between different variables.

**Conclusion**



The box plot displays the relationship between BMI (Body Mass Index) and Outcome (no diabetes vs. diabetic).

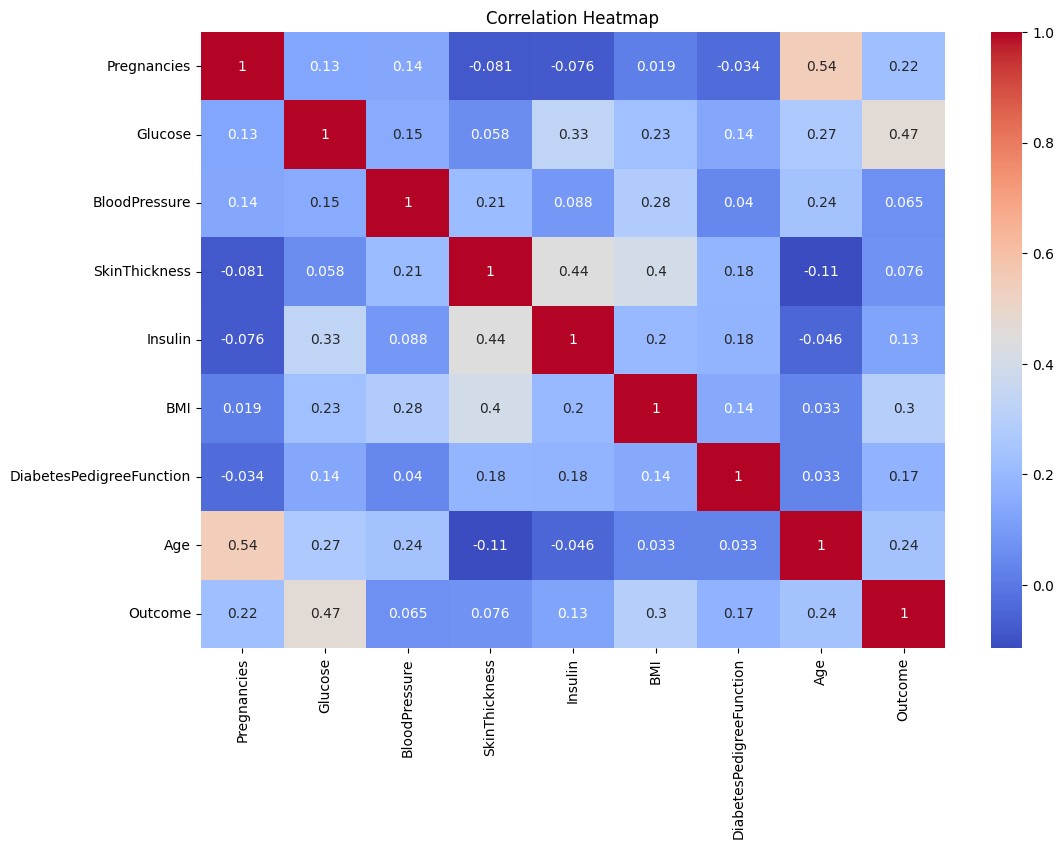
Median BMI: The median BMI is similar for both groups (Outcome 0 and Outcome 1). The thick line inside each box represents the median, and it appears to be close for both groups.

Interquartile Range (IQR): The IQR, represented by the box itself, shows the range within which the middle 50% of the BMI values lie. Both groups have similar IQRs, indicating comparable variability in BMI within each group.

Range and Whiskers: The whiskers extend to show the range of the data, excluding outliers. Both groups have similar ranges, though there are some differences in the upper and lower extremes.

Outliers: Both groups have outliers, which are represented by the circles outside the whiskers. There are several high BMI outliers in both groups, and one very low BMI outlier in the Outcome 0 group.

**Overall, the box plot suggests that there is no significant difference in the distribution of BMI between the outcome groups. Both groups have similar medians, IQRs, and ranges, indicating that BMI might not be a distinguishing factor between the outcomes, or at least not strongly so based on this visual representation.**

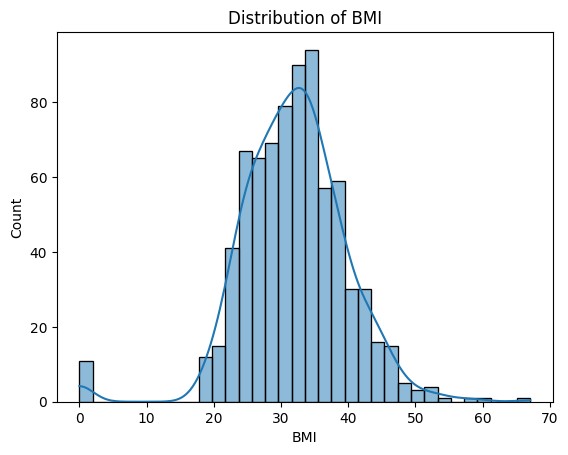


The correlation heatmap shows the relationship between different variables, including the number of pregnancies and the outcome (whether a person has diabetes or not).

From the heatmap:

The correlation coefficient between the number of pregnancies and the outcome **(diabetes) is 0.22.**

**A correlation coefficient of 0.22 indicates a positive but weak correlation between the number of pregnancies and the likelihood of diabetes. This means that as the number of pregnancies increases, there is a slight increase in the likelihood of having diabetes, but the relationship is not very strong.**



The histogram shows the distribution of BMI (Body Mass Index) in the dataset, with a smooth density curve overlay. Here are some key observations:

1. Distribution Shape: The distribution of BMI appears to be approximately normal (bell-shaped), with most values clustering around the mean.

2. Central Tendency: The peak of the distribution, indicating the mode, is around 30. This suggests that the most common BMI value in the dataset is around 30.

3. Spread: The BMI values range from below 10 to above 60, indicating a wide spread of BMI values within the dataset.

4. Skewness: The distribution is slightly right-skewed, with a longer tail on the right side. This suggests that there are some higher BMI values that are less common but extend the range of the data.

5. Outliers: There are a few outliers, particularly on the lower end (near 0) and higher end (above 50), but the majority of the data points lie within a more concentrated range around the mean.

Overall, the BMI values in this dataset are centered around 30, with a spread that indicates variation but mostly adheres to a normal distribution pattern with slight right skewness.