BY KAAVYA MAJUMDER

Comparative Study of Diabetes Indicators: Pathological Factors vs. Family History

INSIGHTS FROM THE PIMA INDIANS DIABETES DATASET

09 JUNE, 2024

1. Methodology

DATA LOADING AND CLEANING

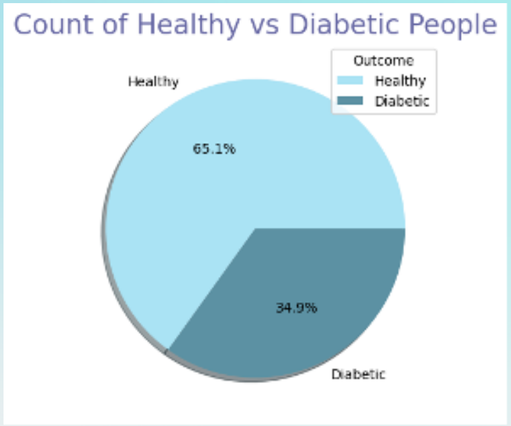
I began by **loading** the dataset and addressing any missing values. The dataset contains **768** samples and **9** features: `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`, `DiabetesPedigreeFunction`, `Age`, and `Outcome.

I then **cleaned** the data by replacing **NaN** values with the **mean** of the column.

```
#handle missing values (data cleansing)
df.fillna(df.mean(), inplace=True)
```

BASIC VISUALIZATION

I then made a pie chart showing the percentage of people who were healthy or diabetic from the outcome column using the matplotlib library.



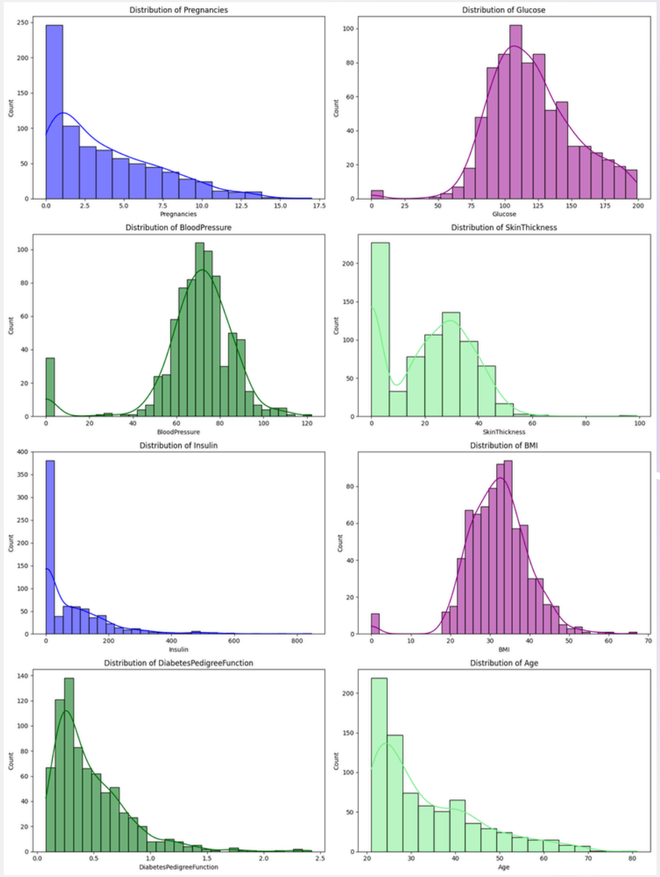
THE QUESTION

I wanted to find out if changes in pathological characteristics could result in diabetes in a person with low pedigree functions, and vice versa. I wanted to find out which had a greater effect on the outcome - physical condition or genetics? Thus I used `DiabetesPedigreeFunction`, `Glucose` and `Insulin` as my factors to conduct comparison to answer the following question:

What is a better indicator of diabetes among the Pima Indians: pathological factors (such as glucose and insulin levels) or family history (pedigree function)?

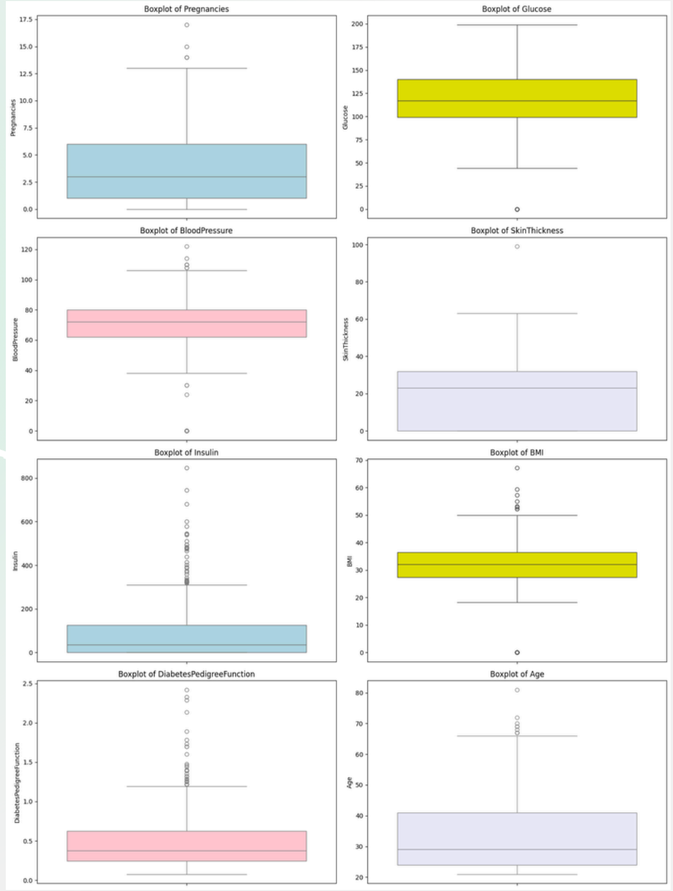
2. Descriptive Statistics

UNIVARIATE



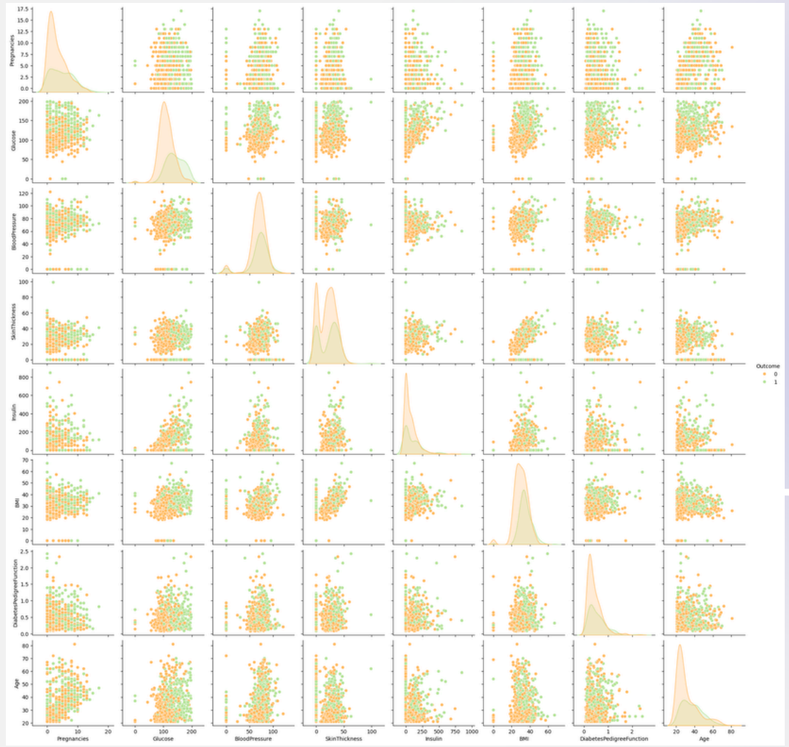
Histograms: These plots revealed that features such as `Pregnancies`, `Glucose`, and `DiabetesPedigreeFunction` are right-skewed, indicating that most patients have low to moderate values for these features.

Boxplots: Highlighted the presence of outliers in features like `Pregnancies` and `Insulin`, suggesting variability in these measures across the population.



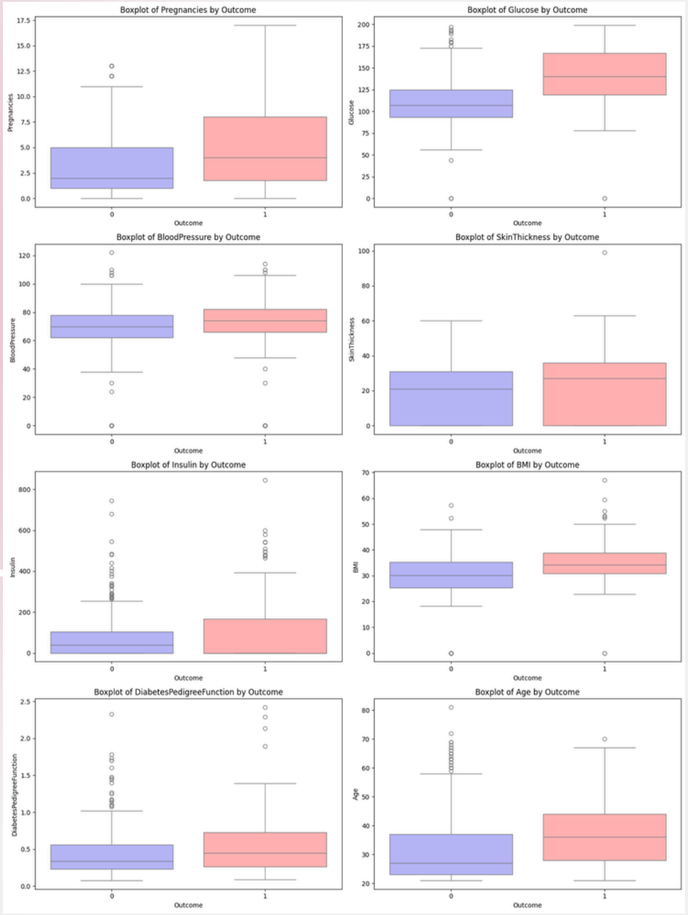
3. Descriptive Statistics

BIVARIATE

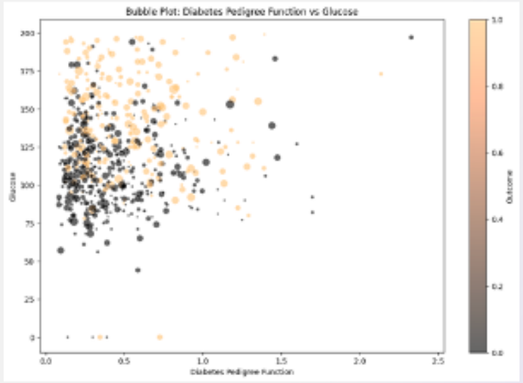


Pair plots: Plots of each column or feature were compared in scatterplots to show distribution of data across two axes, indicated that certain pairs of features, such as `Glucose` and `BMI`, show distinct separations between diabetic and non-diabetic patients.

Bivariate Boxplots: These show the distribution of data in each column across each outcome.



4. Advanced Visualization

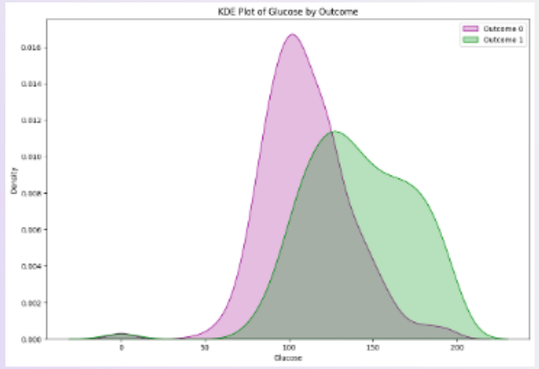


BUBBLE PLOTS

Used to visualize the relationship between `DiabetesPedigreeFunction`, `Glucose`and `Outcome`. Larger bubbles representing higher ` DiabetesPedigreeFunction ` often align with higher `Glucose` levels and positive diabetes outcomes, emphasizing the interplay between these factors.

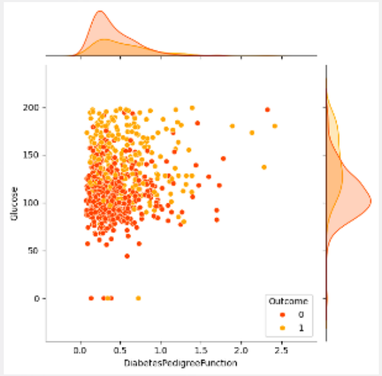
VIOLIN PLOTS

Compared the distribution of `DiabetesPedigreeFunction` for different outcomes. The plots showed that the median and distribution of `DiabetesPedigreeFunction` is higher in diabetic patients, suggesting it is a significant indicator, though there is considerable overlap. The higher values of glucose clearly result in more appearances of diabetes. Both the 0 and 1 outcomes are concentrated towards the lower values, thus not giving a clear conclusion for insulin.



KDE PLOTS

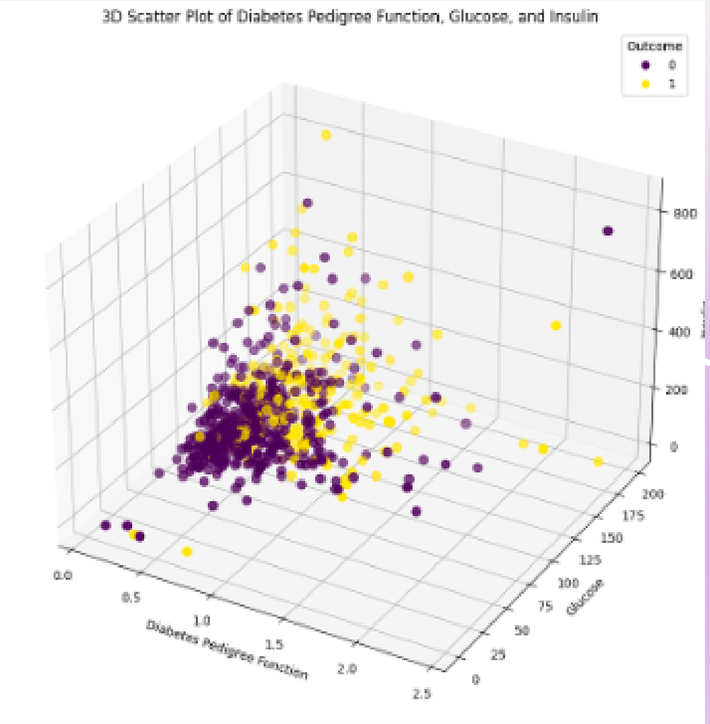
·KDE (Kernel Density Estimate) Plots for `DiabetesPedigreeFunction` showed that while the density is higher for diabetic patients, there is substantial overlap, indicating that `DiabetesPedigreeFunction` alone is not a definitive predictor. However, we clearly see that a higher glucose results in more outcomes of diabetes.



JOINT PLOTS

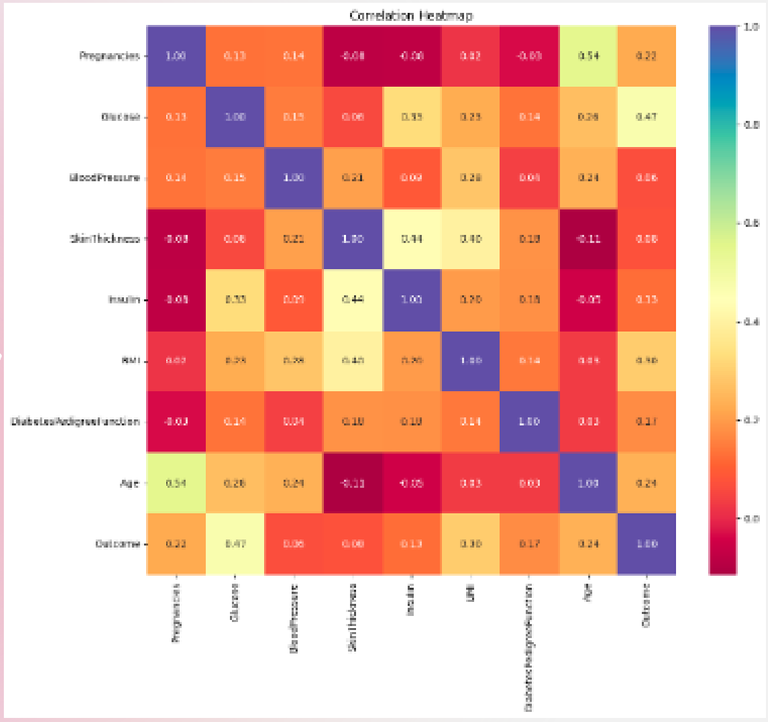
Joint Plots for `DiabetesPedigreeFunction` and `Glucose` with marginal distributions illustrated a strong relationship between high glucose levels and diabetes. The plots highlighted that patients with high `DiabetesPedigreeFunction` and `Glucose` levels are more likely to be diabetic.

5. Correlation Mapping and Conclusion



3D Scatter Plots (Multivariate): Provided a more nuanced view of the interactions between `DiabetesPedigreeFunction`, `Glucose`, and `Insulin`.

Heatmaps: Correlation heatmaps showed that `Glucose` has the highest positive correlation with `Outcome`, followed by `BMI` and `DiabetesPedigreeFunction`.



Conclusion: The analysis shows that pathological factors, particularly `Glucose`, Strong correlation with diabetes, higher predictive accuracy of diabetes among the Pima Indians. However, family history, as represented by the `DiabetesPedigreeFunction`, Moderate correlation, useful for identifying at-risk individuals with normal pathological markers. Combining both pathological and hereditary factors offers the most comprehensive, multi-facted prediction model.

Benefit for Society: Understanding the importance of pathological factors versus family history helps create personalized prevention strategies, such as targeting individuals with a strong family history for early interventions. This study can also develop cost-effective screening programs, targeted educational campaigns, guided research and drug development and effective public health campaigns.