Analysis of Diabetes Indicators among Pima Indians
By Kaavya Majumder

Introduction:

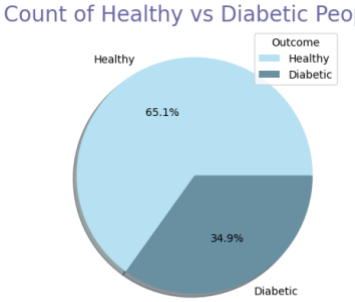
The Pima Indian Diabetes dataset consists of several medical predictor variables and one target variable, `Outcome`, indicating whether a patient has diabetes. The primary aim of this analysis is to determine whether pathological factors (such as glucose and insulin levels) or family history (represented by the diabetes pedigree function) serve as better indicators of diabetes and how they affect each other.

Data Loading and Preprocessing:

- 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)
```

- I then made a pie chart comparing the healthy and diabetic patient percentages.

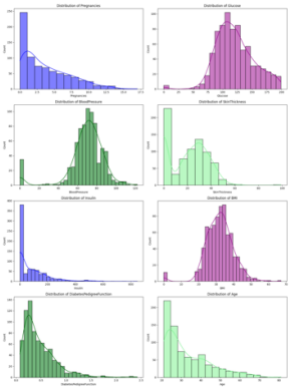


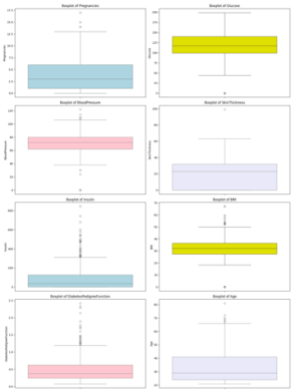
Exploratory Data Analysis (EDA):

Univariate Analysis:

Histograms and Boxplots were used to explore the distribution of individual features.

- **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.

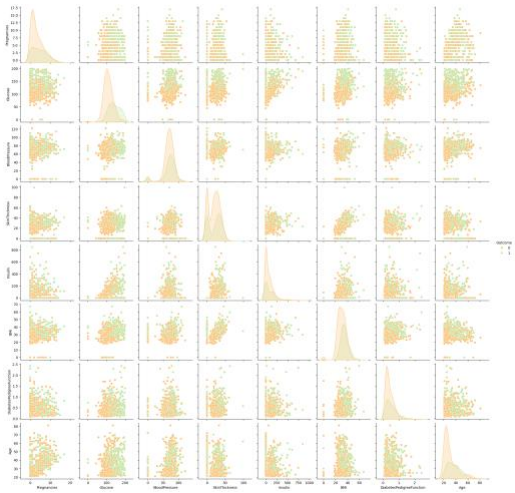


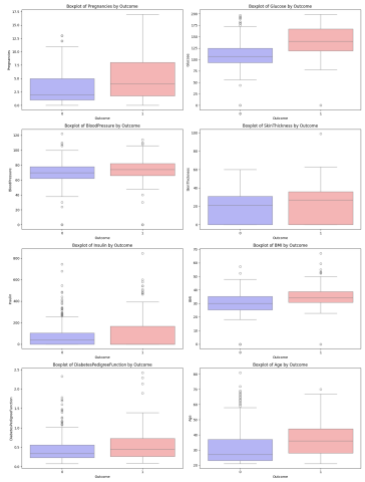


Bivariate Analysis:

Pair plots and bivariate Boxplots were used to explore relationships between features.

- **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.





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:

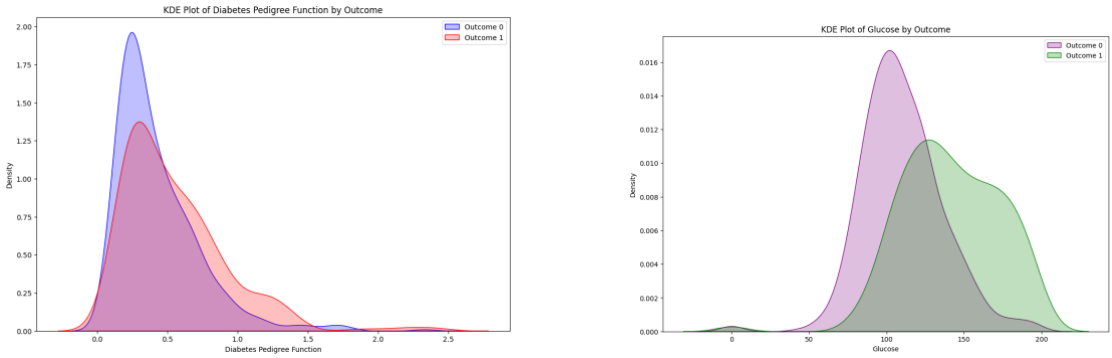
What is a better indicator of diabetes among the Pima Indians: pathological factors (such as glucose and insulin levels) or family history (pedigree function)?

Advanced Visualization and Comparison:

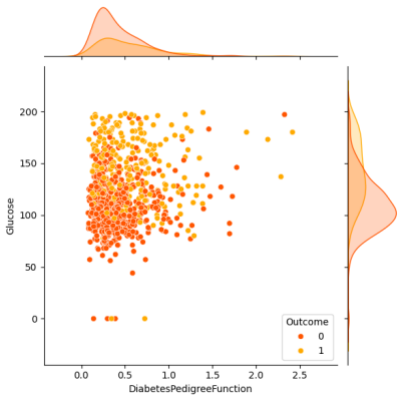
- **Bubble Plots:** Bubble plots were used to visualize the relationship between `DiabetesPedigreeFunction`, `Glucose`and `Outcome`. Larger bubbles representing higher ` DiabetesPedigreeFunction ` often align with higher `Glucose` levels and positive diabetes outcomes, emphasizing the interplay between these factors.

- **Violin Plots:** 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.

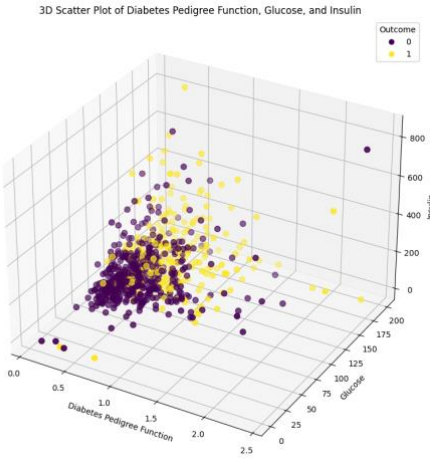


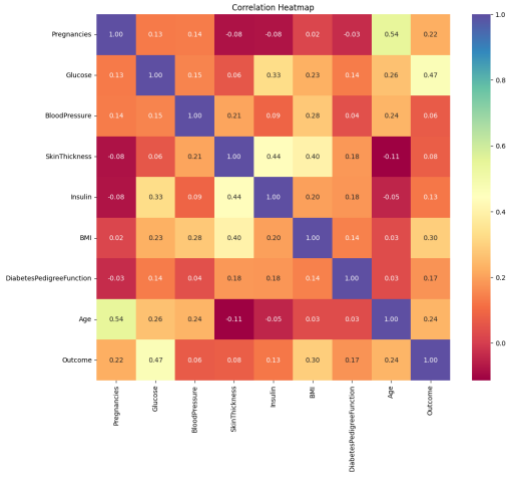
- **Joint Plots:** Joint Plots for `DiabetesPedigreeFunction` and `Glucose` with marginal distributions illustrated a strong relationship between high glucose levels and diabetes. The plots highlighted that patients with high `DiabetesPedigreeFunction` and `Glucose` levels are more likely to be diabetic.



Correlation Mapping:

- **3D Scatter Plots (Multivariate):** Provided a more nuanced view of the interactions between `DiabetesPedigreeFunction`, `Glucose`, and `Insulin`, highlighting that while high `DiabetesPedigreeFunction` often co-occurs with high `Glucose` in diabetic patients, there are exceptions.





- **Heatmaps:** Correlation heatmaps showed that `Glucose` has the highest positive correlation with `Outcome`, followed by `BMI` and `DiabetesPedigreeFunction`. This indicates that while family history is significant, pathological factors like glucose levels are more strongly associated with diabetes.

Modelling and Prediction:

- **Pathological Factors Model:** Using features like `Glucose`, `Insulin`, and `BMI` yielded the highest predictive accuracy, indicating these are strong indicators of diabetes.
- **Family History Model:** Using only `DiabetesPedigreeFunction` and `Age` resulted in lower accuracy, suggesting these are less definitive predictors on their own.
- **Combined Model:** Integrating both pathological factors and family history provided the best predictive performance, underscoring that while pathological factors are more immediate indicators, family history still plays a significant role.

Conclusion

The analysis shows that pathological factors, particularly `Glucose`, are better immediate indicators of diabetes among the Pima Indians. However, family history, as represented by the `DiabetesPedigreeFunction`, also contributes valuable information and should not be overlooked. Combining both pathological and hereditary factors offers the most comprehensive prediction model, highlighting the multifaceted nature of diabetes risk.

Creating a Conda Environment:

- I first created a new conda environment on terminal using the following command: conda create --name data-science-final python=3.9 and made sure the python version was up to date
- Then I activated it using the command conda activate data-science-final
- Then I forged base using conda install conda-forge::r-base
- I then exported my environment as a yml file using the command conda export --from-history>data-science-final.yml