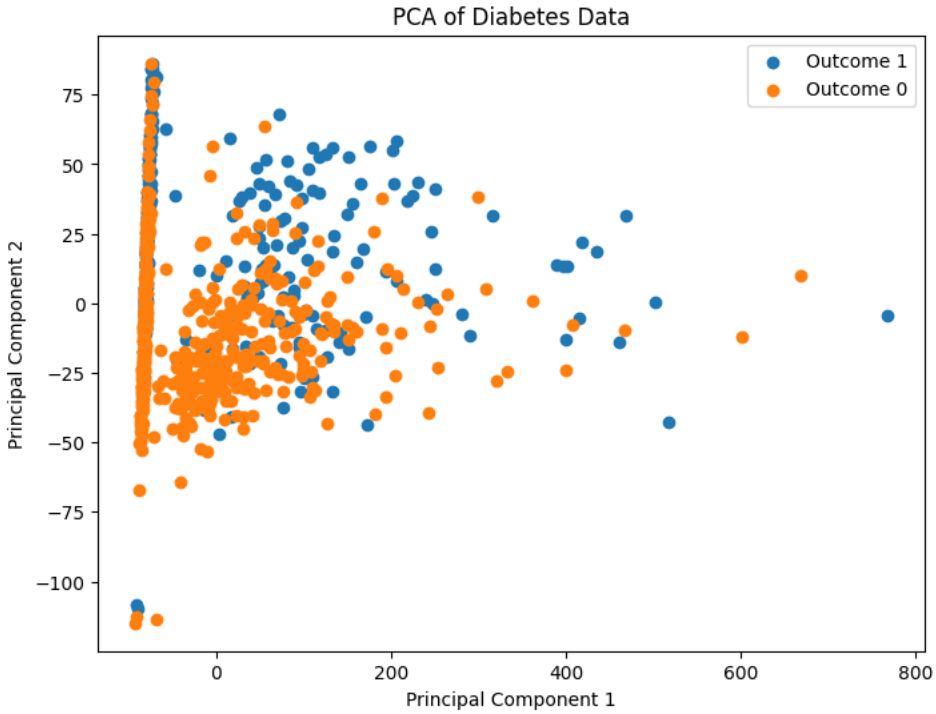
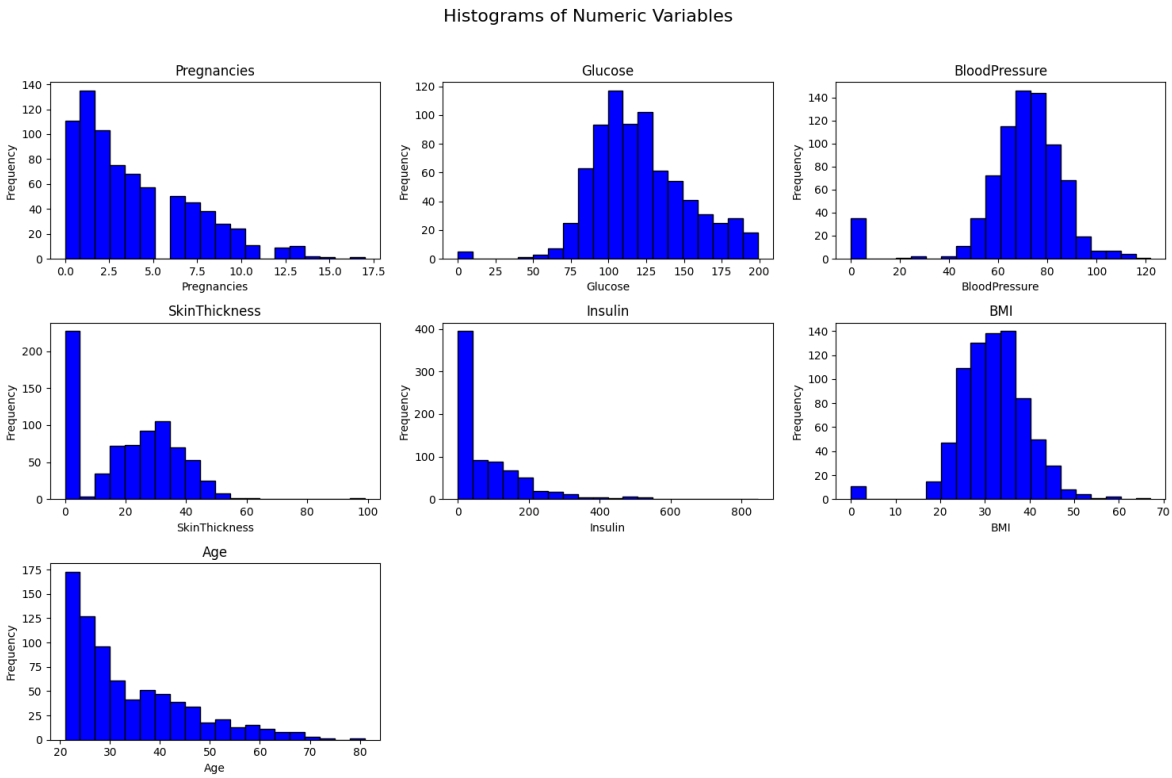
Daniela Alejandra Gonzalez
Assignment #7 DS5110.20722.202510 Section 4



In this plot, at first sight it is not a clear separation between the blue dots (Outcome 1, Diabetes presence) and orange dots (Outcome 0, Diabetes Absense) indicating that the first 2 components are not able to capture the differences between the Outcome class. Also, most of the dots are aligned with the PC1 and there are many points located far away to the PC1 (these points could be outliers). From this graph it can be seen that maybe changing the number of Principal Components can lead to a differentiation in the distribution of the dots.



These are histograms created for each feature. This plot is very useful to detect the distribution of the data in order to decide which statistical test to apply in case of test hypothesis before modelling the data.

Pregnancies: Most values range from 0 to 5 pregnancies, showing a clear decreasing trend as the number of pregnancies increases, indicating that having a high number of pregnancies is less common in the sample.

Glucose: Glucose levels are nearly normally distributed but slightly skewed to the left. The most common values fall between 80 and 150, reflecting normal blood glucose levels, though some values suggest possible hyperglycemia.

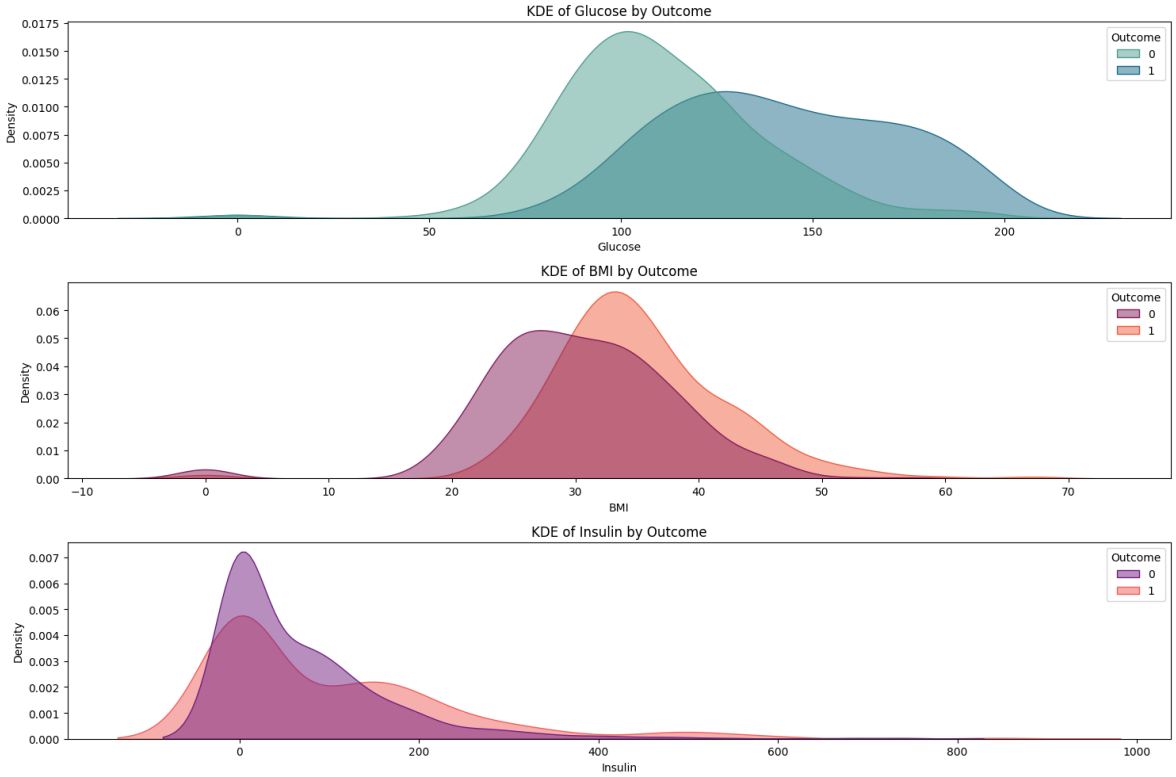
Blood Pressure: Blood pressure shows a symmetrical distribution, peaking around 70-80. However, a small number of values close to zero may indicate erroneous data or missing values.

Skin Thickness: Most skin thickness values are near zero, potentially indicating many missing or incorrectly recorded data points. Non-zero values appear to follow a normal distribution, peaking around 20-30.

Insulin: A significant number of insulin values are zero, suggesting missing data or individuals not using insulin. Non-zero values are right-skewed, with some extremely high values that could be outliers.

BMI: The BMI follows a normal distribution, peaking around 30, indicating that many individuals in this sample may be overweight or obese.

Age: Age distribution is right-skewed, with most individuals between 20 and 40 years old, suggesting a focus on younger adults in the sample.



This graph displays Kernel Density Estimations (KDE) for three variables (Glucose, BMI, and Insulin) against the outcome of diabetes (0 for no diabetes, 1 for diabetes). KDE plots are used to visualize the probability density of a continuous variable.

Glucose: Higher glucose levels are associated with a greater likelihood of diabetes.

BMI: A higher BMI indicates a greater probability of having diabetes.

Insulin: While less clear-cut than Glucose or BMI, higher insulin levels suggest a greater chance of diabetes. The overlapping colored areas represent the probability density of each variable for individuals with and without diabetes. The more significant the overlap between the curves for "0" and "1", the less that variable alone can definitively predict the presence of diabetes. Conversely, less overlap suggests a stronger association between that variable and diabetes. These plots show associations, not causations. While high glucose, BMI, and insulin levels are linked to diabetes, they don't necessarily cause it. Other factors may contribute to the development of diabetes.