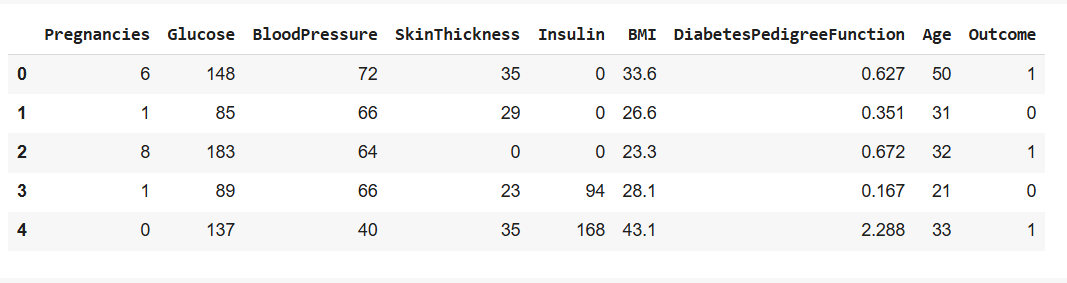
Data Visualization and Statistics

NAME: Yasmin hamada ID: 247057

NAME: Tasneem Khaled ID: 245737

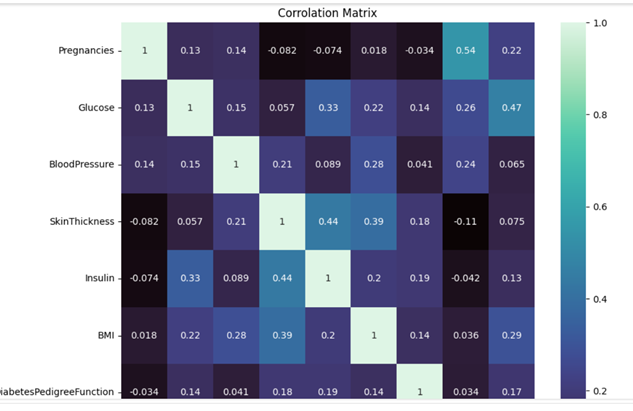
Visualizations:



The table contains data related to pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, age, and diabetes outcome for 768 individuals. Each row represents an individual, and each column represents a different attribute.

**Correlation Matrix**: A table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. The values range from 0 to 1.

This graph is a **heatmap** visualization of a correlation matrix



**Strongest Positive Correlation**:

* **Pregnancies and Age (0.54)**: This suggests older participants generally have more pregnancies, as expected.

**Moderate Positive Correlations**:

* **Glucose and Outcome (0.47)**: Indicates that higher glucose levels are linked to a higher likelihood of the outcome variable (likely diabetes in this case).
* **SkinThickness and Insulin (0.44)**: Suggests a moderate relationship between skin thickness and insulin levels.

**Weak Correlations**:

* Many variables, such as **BloodPressure and Outcome (0.06)**, have very weak correlations, suggesting they may not significantly impact the outcome.

**Potential Feature Relationships**:

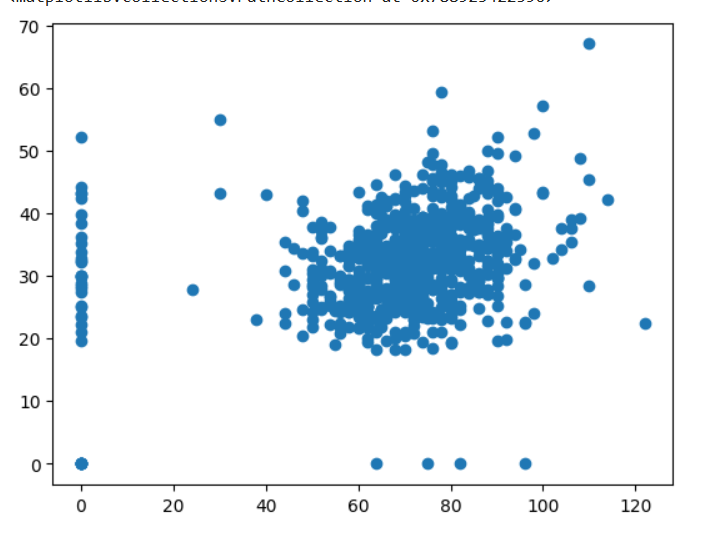
* Variables with moderate correlations, such as **BMI and Outcome (0.29)**, could play a role in predicting the outcome and may be worth investigating further.

**Negative Correlations**: Some variables, such as **Age and SkinThickness (-0.11)**, show weak negative correlations, indicating a slight inverse relationship.

# scatterplot

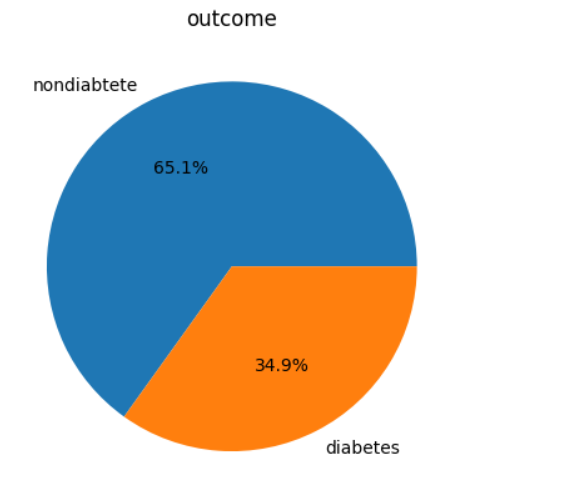
**scatter plot:** a type of data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis. Each dot on the scatter plot represents an individual data point.

The x-axis is labeled "Blood Pressure," and the y-axis is labeled "BMI." The plot consists of numerous red dots representing data points, showing how blood pressure and BMI values are related across different individuals.



**pie chart:** circular statistical graphic that divides a whole into slices to illustrate numerical proportions. Each slice of the pie represents a category of data, and the size of the slice is proportional to the quantity or percentage it represents.

The majority of individuals in this dataset (65.1%) are healthy, while 34.9% have diabetes. This visual representation helps quickly grasp the overall health distribution and can be useful in analyzing and communicating health trends or outcomes.



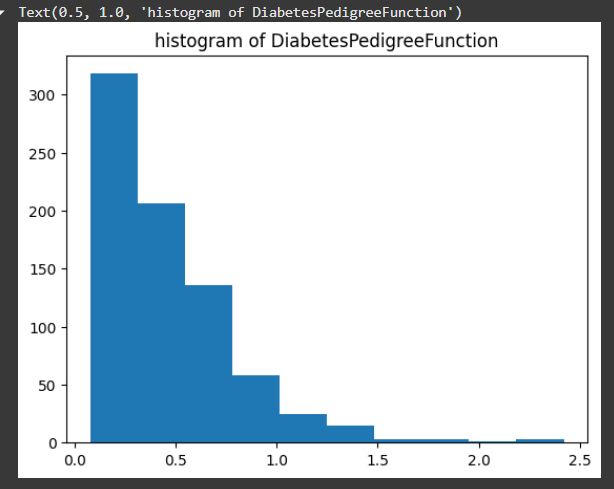
# Histograms

A graph of blood pressure

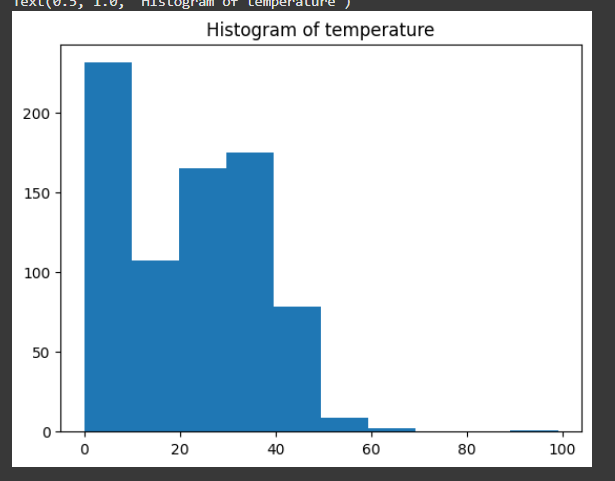
Description automatically generatedt

**Histogram:** A histogram visualizes the frequency distribution of certain dataset

This histogram means that age that range around 60-80 have high blood pressure in the dataset



The histogram shows that most individuals in the dataset have a low DPF , concentrated around 0.



This histogram means that have higher temperatures have ages that range from 0-10.

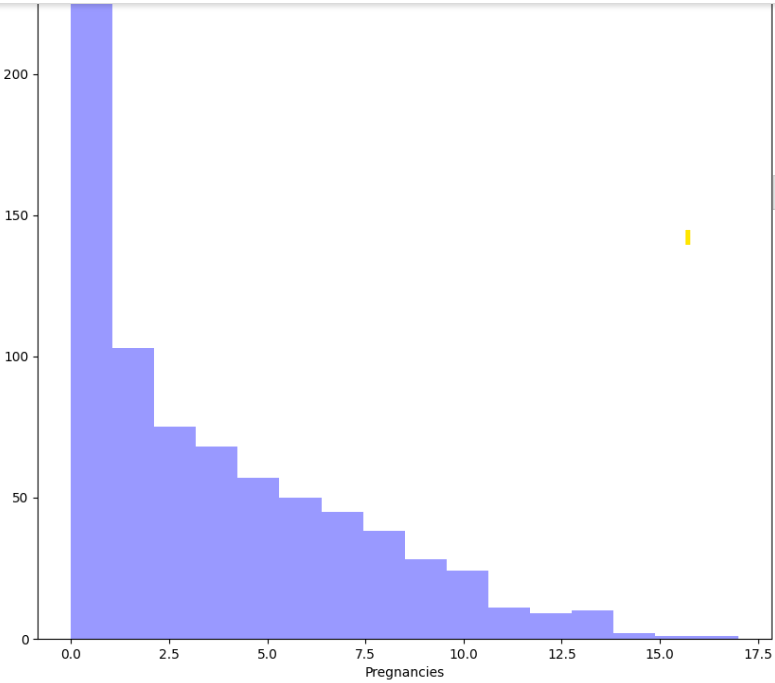
**Distribution plot**: a graphical representation that shows how data values are distributed across a range, providing insights into the shape, spread, and central tendency of the data. These plots are used to understand the overall pattern and variability of the dataset

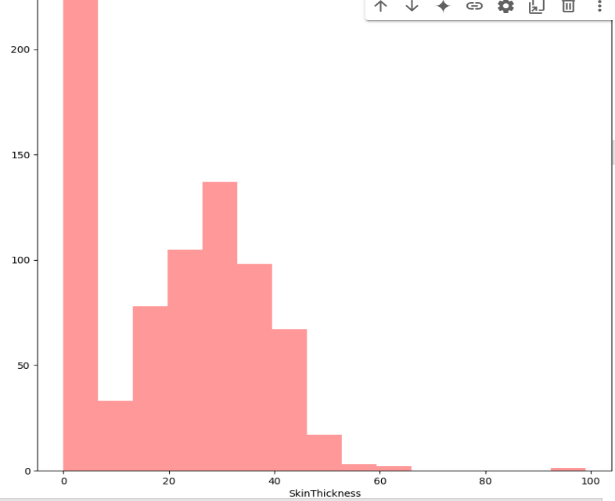
The histogram reveals that most individuals have low Skin Thickness values, with a significant drop-off as the values increase. This skewness can provide insights into the population's characteristics or potentially guide further investigation into what factors contribute to these values.

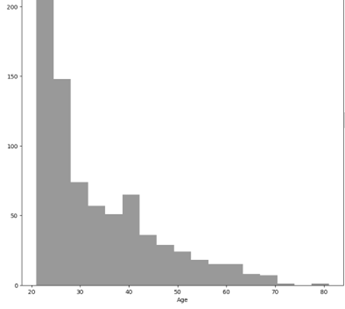
•This data is right-skewed.

This distribution plot provides a clear visual representation of how common different numbers of pregnancies are within the dataset. Most individuals have had zero pregnancies, and the frequency decreases as the number of pregnancies increases.

•This data is right-skewed.



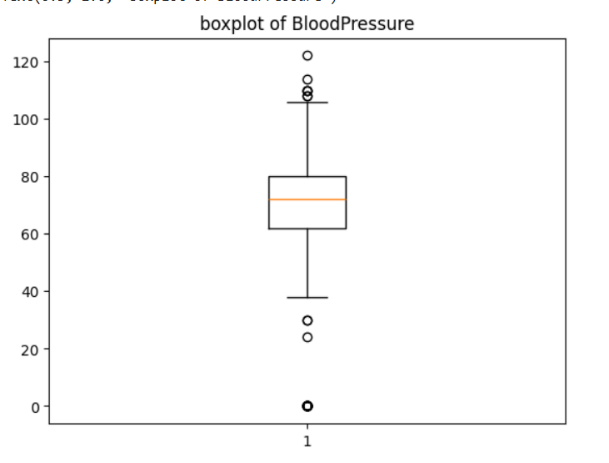




The distribution meaning younger ages are more frequent. This pattern suggests that the dataset has a larger number of younger individuals, with the frequency dropping as the age increases, especially after the age of 40.

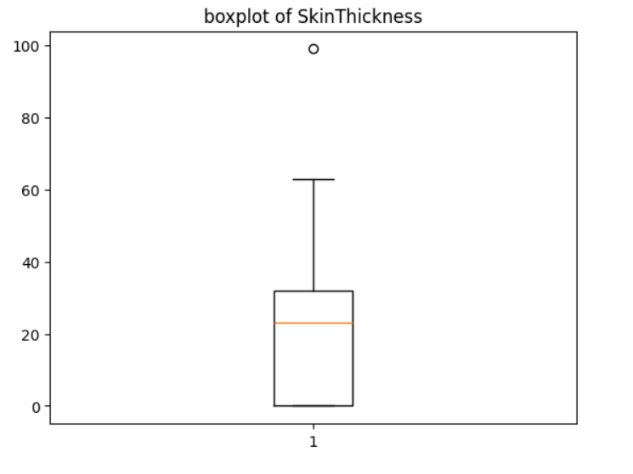
•This data is right-skewed.

**Boxplot** (also known as a box-and-whisker plot): graphical representation of the distribution of a dataset based on a five-number (min, max, median, Q1, Q3). It provides insights into the dataset’s central tendency, variability, and presence of outliers.



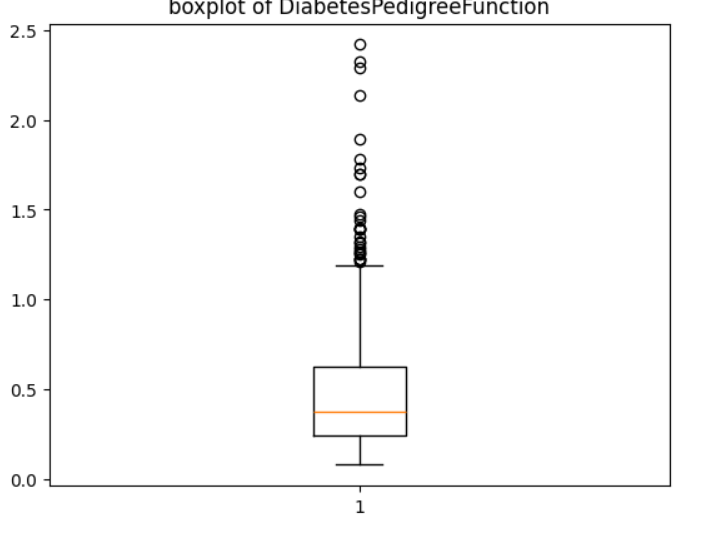
The central box represents the interquartile range (IQR), which includes the middle 50% of the data. In this box plot, the lower quartile (Q1) is 0, the median is 23, and the upper quartile (Q3) is 32, and the minimum is 0, the maximum is around 60.

There are data points marked outside the whiskers, indicating outliers. In this plot, an outlier is marked with a circle at the value of 100, suggesting that this value is significantly higher than the rest of the data.



The central box represents the interquartile range (IQR), which includes the middle 50% of the data. In this box plot, the lower quartile (Q1) is 0.24, the median is 0.37, and the upper quartile (Q3) is 0.63, and the minimum is around 0.1, the maximum is around 1.2.

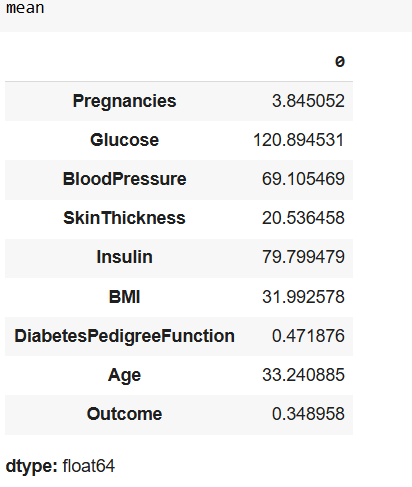
These outliers are higher than the upper whisker, indicating they are significantly different from the rest of the data.



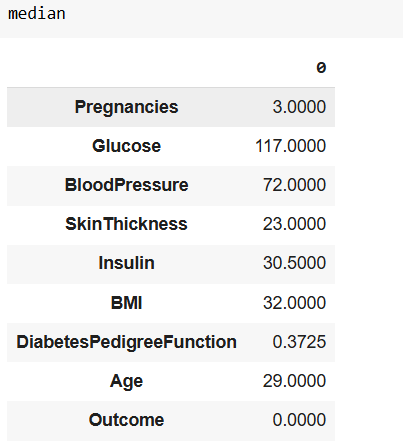
The central box represents the interquartile range (IQR), which includes the middle 50% of the data. In this box plot, the lower quartile (Q1) is 62.0, the median is 72.0, and the upper quartile (Q3) is 80.0, and the minimum is around 40, the maximum is around 105.

There are several circles outside the whiskers, representing outliers. These are individual data points that lie significantly outside the typical range.

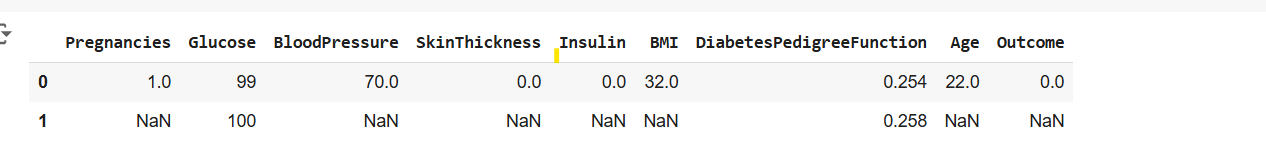
• The mean is the average of a set of numbers. You find it by adding up all the numbers and then dividing by how many numbers there are.



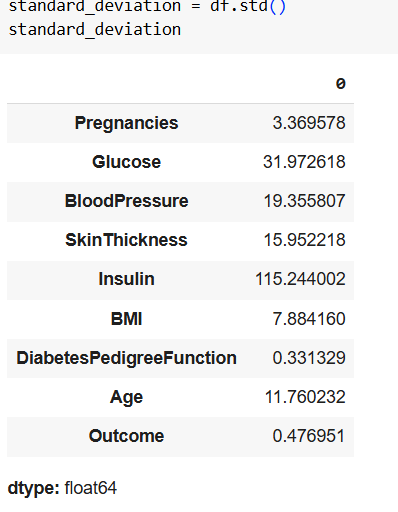
• The median is the middle value in a list of numbers. To find it, you arrange the numbers in order and pick the middle one. If there are two middle numbers, you take their average.



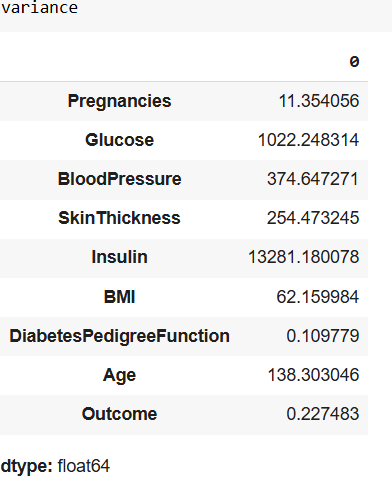
• The mode is the number that appears most frequently in a set of numbers. There can be more than one mode if multiple numbers appear with the same highest frequency, or no mode if all numbers are unique.



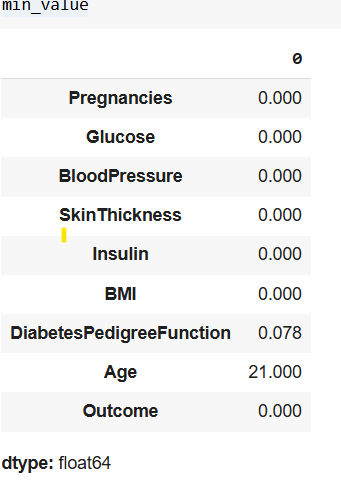
• Standard deviation is a measure of how spread out the numbers in a dataset are. It tells you how much the numbers typically differ from the average (mean) of the dataset.



• Variance is a measure of how much the numbers in a dataset differ from the average (mean). It is calculated as the average of the squared differences from the mean. Variance helps you understand the degree of spread in the data.



• The minimum is the smallest value in a set of numbers. It represents the lowest point in the data.



• The maximum is the largest value in a set of numbers. It represents the highest point in the data. 