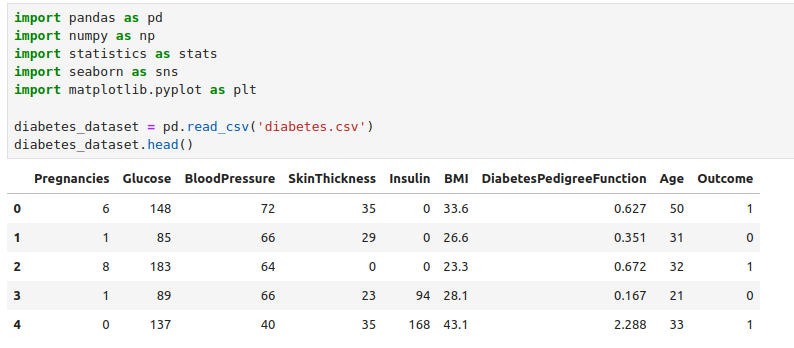
A1. Statistics

## The chosen dataset

Argumentative points of why I have choose the diabetes dataset;



1. **Relevance and Importance**:

* ***Health Impact***: Diabetes is a major public health concern affecting millions of people worldwide. Understanding the factors that contribute to diabetes can help in early detection, prevention, and effective management.
* ***Predictive Modeling***: This dataset provides an excellent opportunity to develop predictive models that can help identify individuals at high risk of developing diabetes. Such models can be instrumental in medical research and clinical decision-making.

2. **Rich Feature Set**:

* ***Diverse Attributes***: The dataset includes a variety of features that are clinically relevant to diabetes, such as glucose levels, blood pressure, BMI, insulin levels, and others. This diversity allows for comprehensive analysis and the exploration of different relationships and interactions among features.
* ***Demographic Data***: Features like age and pregnancies provide additional context that can enhance the understanding of diabetes risk factors.

3. **Accessibility and Size**:

* ***Publicly Available***: The diabetes dataset is readily available from sources like the UCI Machine Learning Repository, making it accessible for research and educational purposes.
* ***Manageable Size***: With 768 observations and 9 features, the dataset is large enough to provide meaningful insights but small enough to be processed and analyzed without requiring extensive computational resources.

4. **Statistical Characteristics**:

* ***Variability***: The dataset exhibits a wide range of values for each feature, which is beneficial for training robust machine learning models. The presence of both high and low values allows for a thorough analysis of the factors influencing diabetes.
* ***Balance***: While the dataset has a class imbalance (around 35% positive cases for diabetes), this is typical in medical datasets and provides an opportunity to explore techniques for handling imbalanced data.

5. **Educational Value**:

* ***Learning Opportunity***: This dataset is widely used in educational settings to teach data preprocessing, feature engineering, and model evaluation. It provides a practical context for applying machine learning algorithms and statistical analysis techniques.
* ***Benchmarking***: The dataset serves as a standard benchmark for comparing the performance of different algorithms and methodologies, making it a valuable tool for both beginners and experienced practitioners.

6. **Real-world Application**:

* ***Clinical Relevance***: Insights derived from this dataset can have real-world applications in improving patient outcomes, developing targeted intervention strategies, and informing public health policies.
* ***Interdisciplinary Research***: The dataset supports interdisciplinary research, bridging the gap between data science and healthcare, and fostering collaboration among data scientists, clinicians, and public health experts.

7. **Potential for Improvement**:

* ***Data Quality Challenges***: The presence of missing or zero values in some features (e.g., insulin and skin thickness) provides an opportunity to apply data imputation and cleaning techniques, which are critical skills in data science.
* ***Feature Engineering***: The dataset allows for extensive feature engineering, such as creating new features from existing ones or transforming features to improve model performance.

## What data do we have?

**Pregnancies**: This feature represents the number of times the individual has been pregnant. It's a discrete numerical feature. **Scale of Measurement**: Ratio.

**Glucose**: This feature represents the plasma glucose concentration after a 2-hour oral glucose tolerance test. It's a continuous numerical feature. **Scale of Measurement**: Interval.

**Blood Pressure**: This feature represents the diastolic blood pressure (mm Hg) of the individual. It's a continuous numerical feature. **Scale of Measurement:** Ratio

**Skin Thickness**: This feature represents the thickness of the skin (triceps skin fold thickness) measured in mm. It's a continuous numerical feature. Scale of Measurement: Ratio.

**Insulin**: This feature represents the 2-hour serum insulin (mu U/ml) of the individual. It's a continuous numerical feature. Scale of Measurement: Ratio.

**BMI**: This feature represents the Body Mass Index (weight in kg / (height in m)^2) of the individual. It's a continuous numerical feature. Scale of Measurement: Ratio.

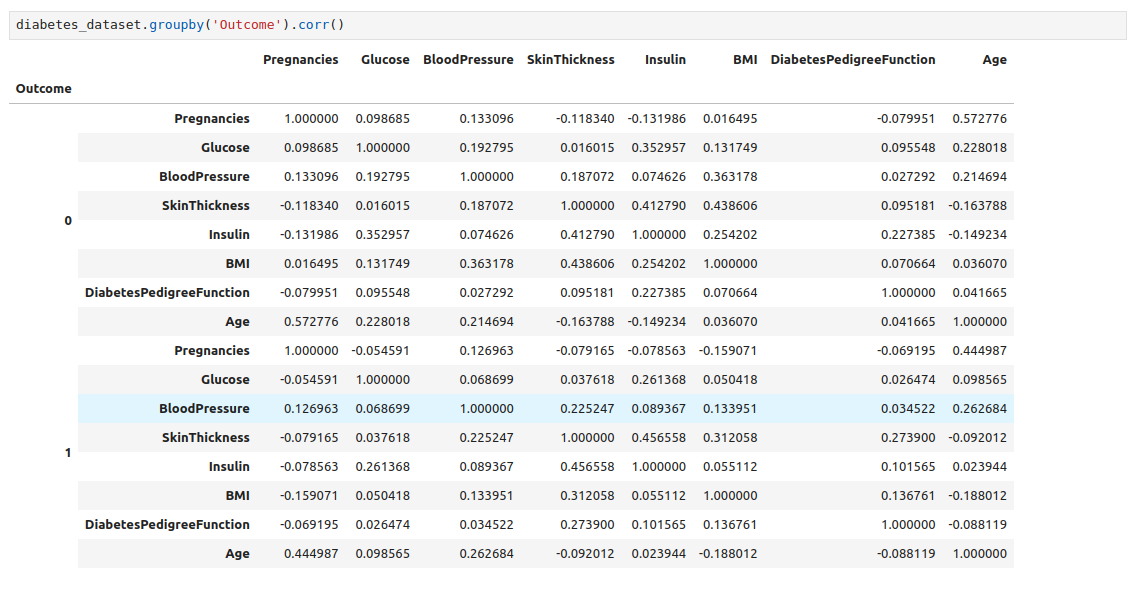
**Diabetes Pedigree Function**: This feature represents a function that scores the likelihood of diabetes based on family history. It's a continuous numerical feature. Scale of Measurement: Ratio

**Age**: This feature represents the age of the individual in years. It's a discrete numerical feature. Scale of Measurement: Ratio.

**Outcome**: This feature represents whether the individual has diabetes or not, with "1" indicating the presence of diabetes and "0" indicating the absence. It's a categorical feature, specifically binary. Scale of Measurement: Nominal.

## Statistics:

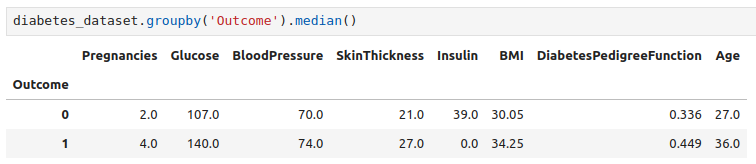
1. Correlation Matrix:
   * The correlation matrix provides the correlation coefficients between each pair of features for each outcome category (0: absence of diabetes, 1: presence of diabetes).
   * A positive correlation indicates that as one feature increases, the other also tends to increase, while a negative correlation indicates that as one feature increases, the other tends to decrease.
   * For example, in the "0" outcome category, there is a moderate positive correlation (0.572776) between "Pregnancies" and "Age", suggesting that as the number of pregnancies increases, so does the age.
   * Similarly, in the "1" outcome category, there is a moderate positive correlation (0.444987) between "Pregnancies" and "Age".



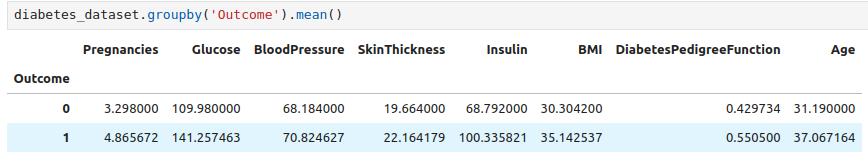
1. Mode, Median, and Mean:
   * Mode represents the most frequently occurring value in each feature.



* + Median represents the middle value of each feature when arranged in ascending order.



* + Mean represents the average value of each feature.

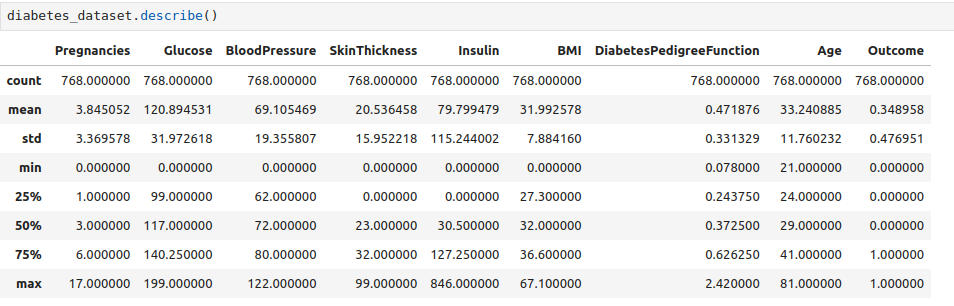


These measures provide different insights into the central tendency of the data distribution.

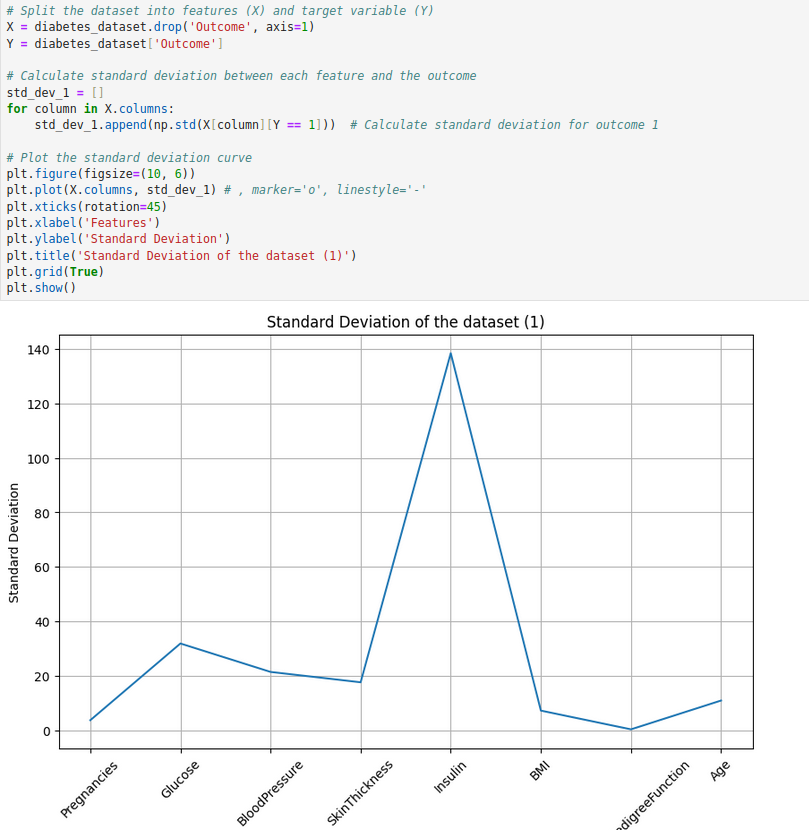
For example, the mode of "Pregnancies" is 1, indicating that the most frequent number of pregnancies is 1 in the dataset.

The median and mean provide additional information about the central tendency of the data.

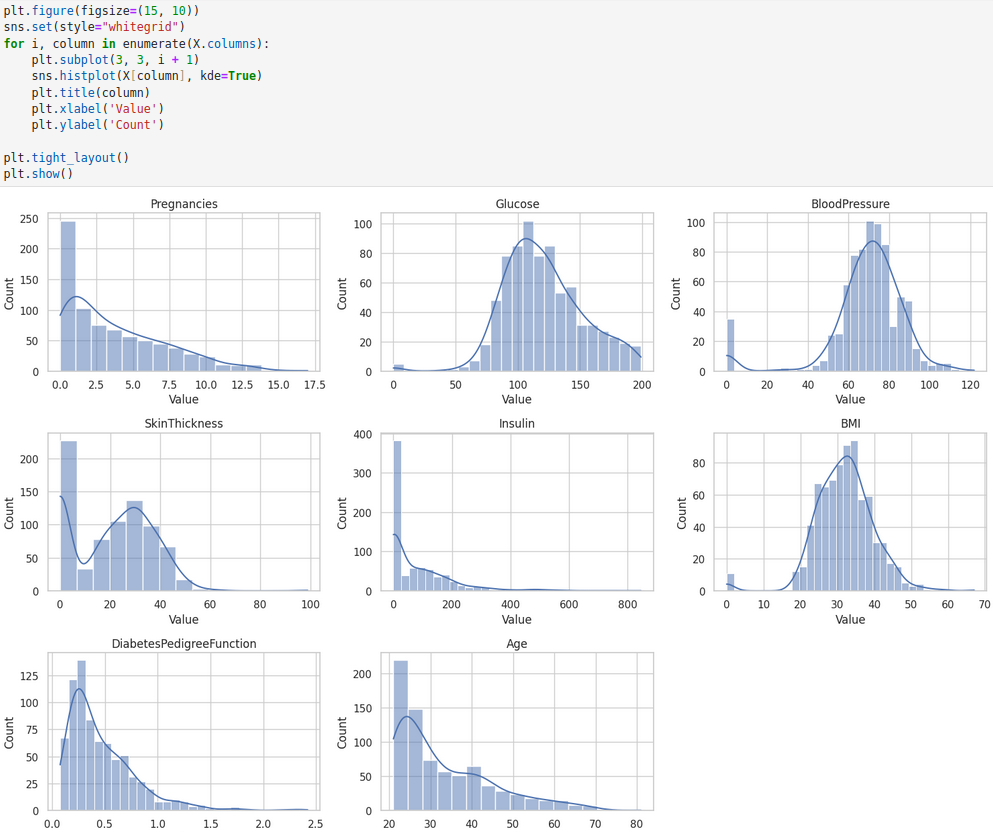
1. Descriptive Statistics:
   * Descriptive statistics (count, mean, std, min, 25%, 50%, 75%, max) provide a summary of the distribution of each feature.



* Standard Deviation Plot



* Data Distribution



They help understand the spread and variability of the data.

For example, the "Insulin" feature has a mean of approximately 79.80 and a standard deviation of approximately 115.24, indicating a relatively high variability in insulin levels.

## Interpretation of the values:

1. Pregnancies:
   * Mean: The average number of pregnancies in the dataset is approximately 3.85.
   * Standard Deviation (Std): The standard deviation is approximately 3.37, indicating a moderate spread of data around the mean.
   * Mode: The most frequent number of pregnancies is 1.
   * Interpretation: Higher numbers of pregnancies might indicate a higher risk of developing diabetes, as pregnancy can affect insulin sensitivity and glucose metabolism. However, it's essential to consider other factors such as age and lifestyle.
2. Glucose:
   * Mean: The average glucose level is approximately 120.89 mg/dL.
   * Std: The standard deviation is approximately 31.97, indicating variability in glucose levels.
   * Mode: The most frequent glucose level is 100 mg/dL.
   * Interpretation: High glucose levels (hyperglycemia) are a key indicator of diabetes. Higher glucose levels are associated with an increased risk of diabetes, as insulin resistance or insufficient insulin production can lead to elevated blood sugar levels.
3. Blood Pressure:
   * Mean: The average blood pressure is approximately 69.11 mm Hg (diastolic).
   * Std: The standard deviation is approximately 19.36, indicating variability in blood pressure measurements.
   * Mode: The most frequent blood pressure measurement is 70 mm Hg.
   * Interpretation: Elevated blood pressure (hypertension) is a common comorbidity of diabetes. High blood pressure can contribute to the progression of diabetes complications, such as cardiovascular disease and kidney disease.
4. Skin Thickness:
   * Mean: The average skin thickness is approximately 20.54 mm.
   * Std: The standard deviation is approximately 15.95, indicating variability in skin thickness measurements.
   * Mode: The most frequent skin thickness is 0 mm.
   * Interpretation: Skin thickness may not directly influence diabetes outcome but could be related to obesity or insulin resistance. However, it's essential to interpret this feature in conjunction with other clinical parameters.
5. Insulin:
   * Mean: The average insulin level is approximately 79.80 μU/ml.
   * Std: The standard deviation is approximately 115.24, indicating considerable variability in insulin levels.
   * Mode: The most frequent insulin level is 0 μU/ml.
   * Interpretation: Insulin levels are directly related to diabetes, as insulin resistance or insufficient insulin production characterizes diabetes. Higher insulin levels may indicate insulin resistance, a precursor to type 2 diabetes.
6. BMI (Body Mass Index):
   * Mean: The average BMI is approximately 31.99 kg/m^2.
   * Std: The standard deviation is approximately 7.88, indicating variability in BMI measurements.
   * Mode: The most frequent BMI is 32.0 kg/m^2.
   * Interpretation: High BMI is strongly associated with an increased risk of type 2 diabetes. Obesity is a major risk factor for diabetes, as it contributes to insulin resistance and metabolic dysfunction.
7. Diabetes Pedigree Function:
   * Mean: The average diabetes pedigree function is approximately 0.47.
   * Std: The standard deviation is approximately 0.33, indicating variability in diabetes pedigree function values.
   * Mode: The most frequent diabetes pedigree function value is 0.254.
   * Interpretation: The diabetes pedigree function assesses the genetic predisposition to diabetes. Higher values indicate a stronger family history of diabetes, which can increase the risk of developing the condition.
8. Age:
   * Mean: The average age is approximately 33.24 years.
   * Std: The standard deviation is approximately 11.76, indicating variability in ages.
   * Mode: The most frequent age number is 22.
   * Interpretation: Age is a significant risk factor for diabetes, as the prevalence of diabetes increases with age. Older individuals are more likely to develop diabetes due to factors such as reduced insulin sensitivity and lifestyle changes.

## Jupyter Notebook source code:

<https://github.com/alexban14/DataMining_Diabetes_DS>

## Resources utilized:

https://medium.com/analytics-vidhya/statistics-mean-median-mode-variance-standard-deviation-47fab926465a

https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/

<https://www.geeksforgeeks.org/standard-deviation-plot/>

<https://www.youtube.com/watch?v=MXvd-bNU93E>