MSA UNIVERSITY
جامعة أكتوبر للعلوم الحديثة والآداب

in partnership with
University of Bedfordshire

October University for Modern Sciences & Arts

Probability and Statistics Project

Diabetes Dataset Analysis

Name	ID
Erfan Nada	240097
Mohammed Ammar	246507

Dataset Description

Discrete Variables

Discrete variables are those that can take on a finite number of values, often counted in whole numbers. In this dataset, the following variables are discrete:

1. **Pregnancies**
2. **Outcome**

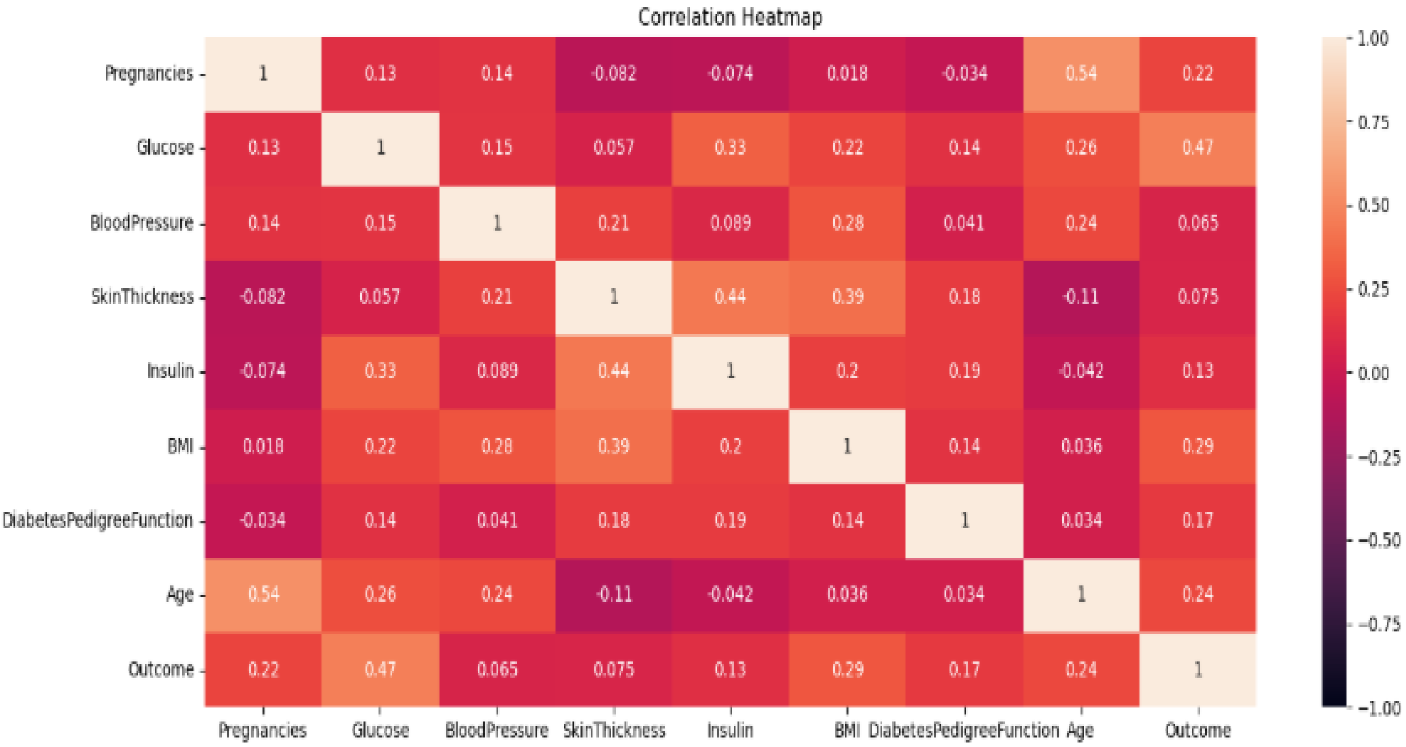
Continuous Variables

Continuous variables can take on an infinite number of values within a given range and are typically measured. In this dataset, the following variables are continuous:

1. **Glucose**
2. **Blood Pressure**
3. **Skin Thickness**
4. **Insulin**
5. **BMI (Body Mass Index)**
6. **Diabetes Pedigree Function**
7. **Age**

Visualizations

1. Correlation Heatmap



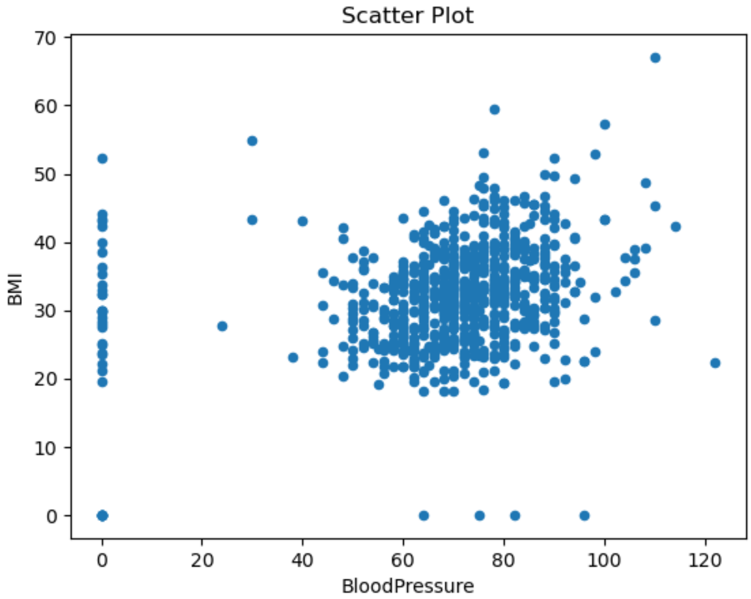
The correlation heatmap visually represents the relationships between different features in the diabetes dataset. Each cell in the heatmap displays the correlation coefficient between pairs of features, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation).

Insights:

- The heatmap reveals several key relationships:
  - **Glucose and Outcome:** The correlation coefficient is approximately 0.47, indicating a moderate positive correlation. This suggests that individuals with higher glucose levels are more likely to have a positive diabetes outcome.
  - **BMI and Outcome:** The correlation is also positive (0.29), suggesting that higher BMI may be associated with a higher risk of diabetes.

- **Age and Outcome:** The correlation coefficient is 0.238, indicating that older individuals tend to have a higher likelihood of diabetes.

2. Scatter Plot of Blood Pressure vs. BMI

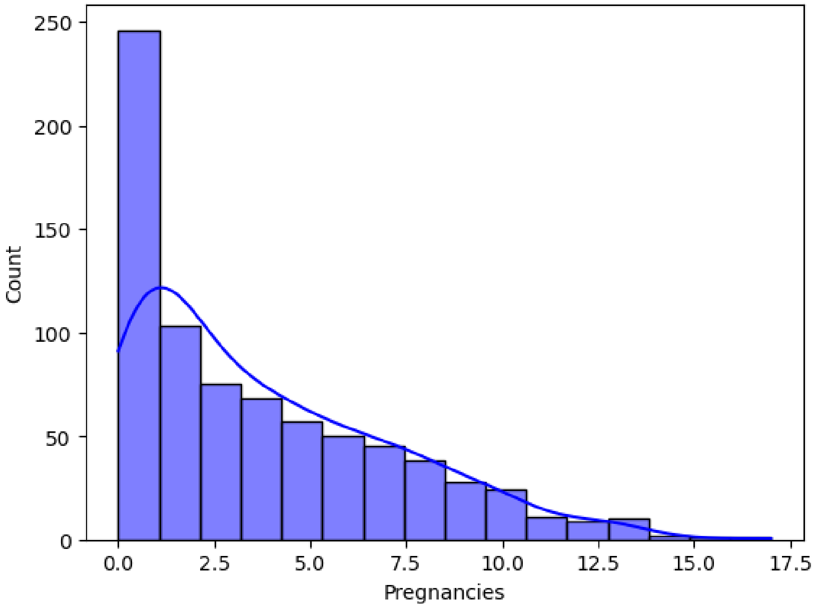


The scatter plot illustrates the relationship between Blood Pressure and BMI. Each point represents an individual in the dataset, with the x-axis representing Blood Pressure and the y-axis representing BMI.

Insights:

- The scatter plot shows a wide distribution of points, suggesting variability in BMI across different blood pressure levels.
- There is no linear appearance in the scatter plot; however, there appears to be clusters of data points that may indicate specific ranges of blood pressure and BMI that are more common.

3. Histogram of Pregnancies

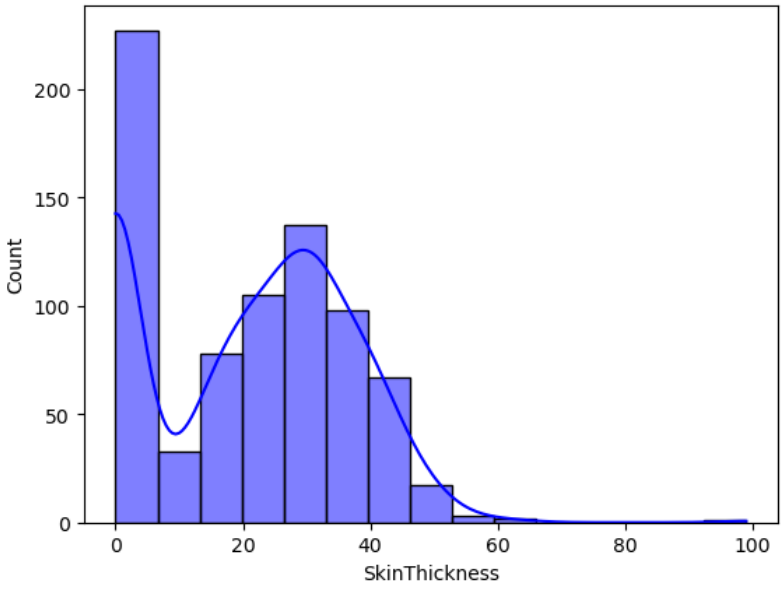


The histogram displays the distribution of the number of pregnancies among individuals in the dataset. The kernel density estimate (KDE) line provides a smoothed representation of the distribution.

Insights:

- The histogram shows that most individuals in the dataset have had between 0 to 5 pregnancies, with a noticeable drop-off for those with higher counts.
- The KDE overlay indicates that the distribution is somewhat right skewed towards lower numbers of pregnancies, suggesting that higher pregnancy counts are less common.

4. Histogram of Skin Thickness

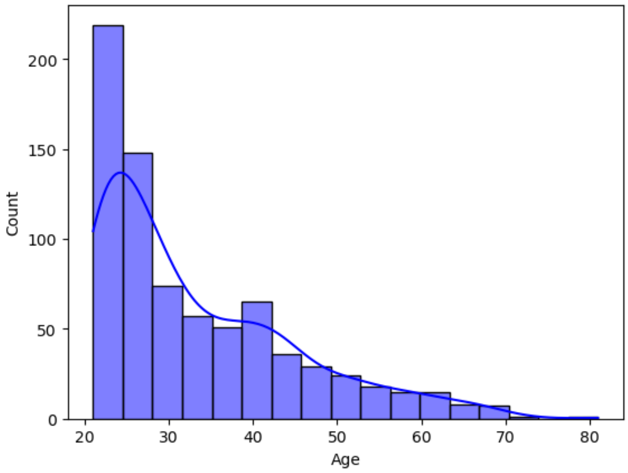


The histogram depicts the distribution of skin thickness measurements in the dataset, with the KDE line showing the overall shape of the distribution.

Insights:

- The histogram reveals that skin thickness measurements are concentrated around lower values, with fewer individuals having very high skin thickness.
- The presence of a few outliers indicates that some individuals may have significantly higher skin thickness, which is relevant in diabetes assessments.

5. Histogram of Age

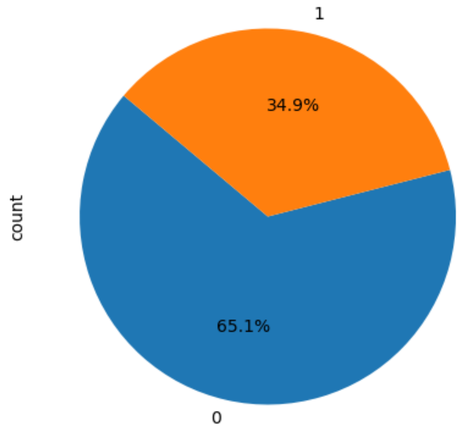


This histogram displays the distribution of ages among individuals in the dataset, with the KDE line providing a smoothed view of the age distribution.

Insights:

- The histogram shows that the majority of individuals are between the ages of 20 to 40, with fewer individuals in older age brackets.
- The age distribution is important for understanding the demographics of diabetes risk, as age is a significant factor in the likelihood of developing diabetes.
- The KDE line helps visualize the probability density, indicating that middle-aged individuals are more common in the dataset.

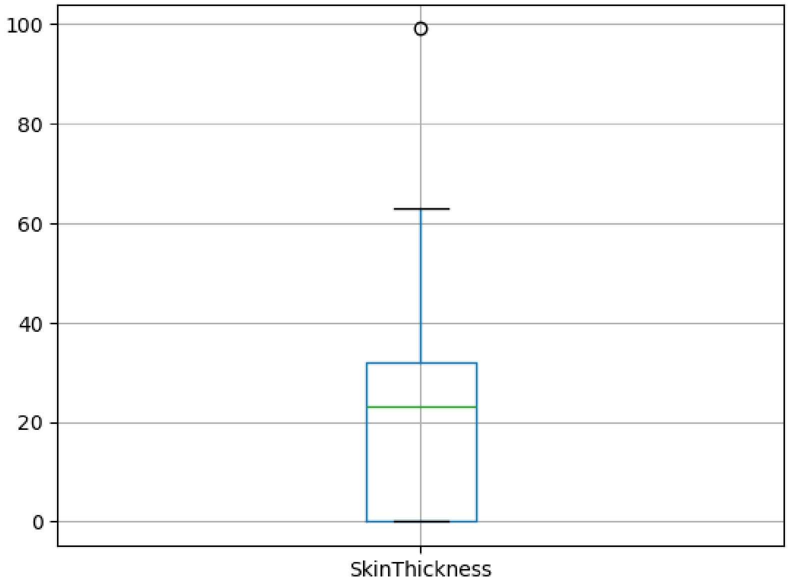
6. Pie Chart of Diabetes Outcomes



The pie chart visualizes the distribution of diabetes outcomes in the dataset. The 'Outcome' feature indicates whether an individual has diabetes (1) or does not have diabetes (0).

- The pie chart shows the proportion of individuals with diabetes compared to those without.
- The percentages displayed on the chart provide a clear visual representation of the prevalence of diabetes in the dataset.
- For example, if the chart indicates that approximately 34.9% of individuals have diabetes and 65.1% do not, this highlights the need for targeted health interventions for the affected population.
- Understanding the distribution of outcomes is crucial for evaluating the effectiveness of diabetes prevention and management strategies.

7. Boxplot of Skin Thickness

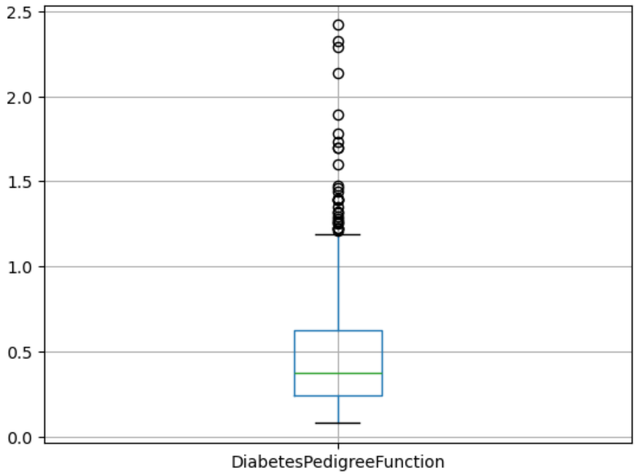


The boxplot provides a summary of the distribution of skin thickness measurements, highlighting the median, quartiles, and potential outliers.

Insights:

- The boxplot shows a median skin thickness of around 20 mm, with a significant range of measurements.
- Several outliers are visible, indicating that some individuals have much higher skin thickness measurements than the majority.
- This variability in skin thickness can be relevant in diabetes assessments, as it is one of the factors associated with insulin resistance.

8. Boxplot of Diabetes Pedigree Function

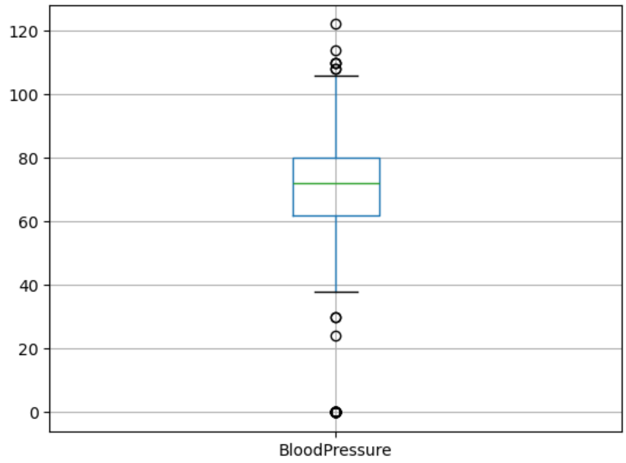


This boxplot summarizes the distribution of the diabetes pedigree function, which indicates the likelihood of diabetes based on family history.

Insights:

- The boxplot indicates a median value around 0.47, with a range that includes several outliers.
- The presence of outliers suggests some individuals have a significantly higher diabetes pedigree function, which may indicate a stronger genetic predisposition to diabetes.

9. Boxplot of Blood Pressure



The boxplot for blood pressure provides insights into the blood pressure distribution of individuals in the dataset.

Insights:

- The median blood pressure is around 72 mmHg, with a range that shows some individuals have significantly higher blood pressure levels.
- The presence of outliers indicates that some individuals may have hypertension, which is a known risk factor for diabetes.

Data Statistics

Mean Values:

Pregnancies	3.845052
Glucose	120.894531
BloodPressure	69.105469
SkinThickness	20.536458
Insulin	79.799479
BMI	31.992578
DiabetesPedigreeFunction	0.471876
Age	33.240885
Outcome	0.348958

The mean values indicate the average level of each feature in the dataset. For example, the mean glucose level of 120.89 mg/dL suggests that, on average, individuals in this dataset have elevated glucose levels, which is a risk factor for diabetes.

The mean BMI of 31.99 indicates a population that is, on average, classified as overweight, which is another risk factor for diabetes.

Median Values:

Pregnancies	3.0000
Glucose	117.0000
BloodPressure	72.0000
SkinThickness	23.0000
Insulin	30.5000
BMI	32.0000
DiabetesPedigreeFunction	0.3725
Age	29.0000
Outcome	0.0000

The median values provide a measure of central tendency that is less affected by outliers. The median age of 29.00 suggests that the majority of individuals in the dataset are relatively young, which is important for understanding the demographic profile of the population at risk for diabetes.

Mode Values:

- **Pregnancies:** 1
- **Glucose:** 99
- **Blood Pressure:** 70
- **Skin Thickness:** 0
- **Insulin:** 0

- **BMI:** 30
- **Diabetes Pedigree Function:** 0.25
- **Age:** 22
- **Outcome:** 0

The mode indicates the most frequently occurring value for each feature. For example, a mode of 1 for pregnancies suggests that many individuals in the dataset have been pregnant only once.

The mode for glucose being 99 indicates that this is a common glucose level among individuals, potentially serving as a baseline for assessing elevated glucose levels.

Standard Deviation:

- **Pregnancies: 3.37**
- **Glucose: 31.97**
- **Blood Pressure: 19.36**
- **Skin Thickness: 15.95**
- **Insulin: 115.24**
- **BMI: 7.88**
- **Diabetes Pedigree Function: 0.33**
- **Age: 11.76**
- **Outcome: 0.48**

The standard deviation measures the variability of each feature. A standard deviation of 115.24 for insulin indicates considerable variability in insulin levels among individuals, suggesting that there are individuals with significantly higher or lower insulin levels, which is critical for diabetes risk assessment.

The relatively lower standard deviation for BMI (7.88) suggests that most individuals have BMI values that are closer to the mean, indicating a more consistent weight status within the population.

Variance:

- **Pregnancies: 11.35**
- **Glucose: 1022.25**
- **Blood Pressure: 374.65**
- **Skin Thickness: 254.47**
- **Insulin: 13281.18**
- **BMI: 62.16**
- **Diabetes Pedigree Function: 0.11**
- **Age: 138.30**
- **Outcome: 0.2**

Variance indicates the spread of the data points around the mean. High variance in insulin (13281.18) and glucose (1022.25) suggests that these features have a wide range of values, highlighting significant disparities in the population.

Maximum Values

Pregnancies	**17.00**
Glucose	**199.00**
BloodPressure	**122.00**
SkinThickness	**99.00**
Insulin	**846.00**
BMI	**67.10**
DiabetesPedigreeFunction	**2.42**
Age	**81.00**
Outcome	**1.00**

Minimum Values

Pregnancies	**0.000**
Glucose	**0.000**
BloodPressure	**0.000**
SkinThickness	**0.000**
Insulin	**0.000**
BMI	**0.000**
DiabetesPedigreeFunction	**0.078**
Age	**21.000**
Outcome	**0.000**

The maximum and minimum values highlight the variability in health indicators and the importance of addressing both extremes to effectively manage diabetes risk in the population.

The maximum and minimum values for glucose, insulin, BMI, and age provide critical insights into the health risks associated with diabetes. These values highlight the variability within the population and show the importance of targeted interventions for those at both ends. By focusing on these key indicators, healthcare providers can better identify individuals at risk and implement strategies to improve health outcomes.

1. **Mean vs. Median:**

    - The mean values for features like Glucose, BMI, and Age are higher than the median values, indicating a right-skewed distribution. This suggests the presence of outliers that influence the average.

- For instance, the mean glucose level of 120.89 mg/dL compared to a median of 117.00 mg/dL highlights that some individuals have significantly higher glucose levels, which is critical for diabetes risk assessment.

Conclusion

The findings of this project underscore the importance of monitoring key health indicators such as glucose levels, insulin sensitivity, BMI, and age in assessing diabetes risk. By identifying individuals at both ends of the health spectrum, healthcare providers can implement strategies to prevent diabetes onset and manage existing conditions. The insights gained from this analysis can inform public health initiatives aimed at reducing the prevalence of diabetes and improving overall health outcomes in the population.