**DIABETES DATA ANALYSIS USING PYTHON**

**Members**

1. -Mahmoud Adel 233561
2. -Khaled Ahmed 238277

**Table of Contents**

**Part 1: Data Analysis and Insights**

1. **Correlation Analysis**
   * Correlation Matrix
   * Key Insights:
     + Age and Pregnancies
     + Glucose and Diabetes Outcome
     + BMI and Diabetes Outcome
2. **Data Visualization**
   1. Scatter Plots: BMI vs. Blood Pressure
   2. Distribution Plots:
      1. Pregnancies
      2. SkinThickness
      3. Age
   3. Boxplots:
      1. SkinThickness
      2. Blood Pressure
      3. Diabetes Pedigree Function
   4. Pie Chart: Diabetes Outcome Distribution

**Part 2: Statistical Analysis and Findings**

1. **Statistical Analysis**
   * Central Tendency (Mean, Median, Mode)
   * Dispersion (Variance, Standard Deviation)
   * Min and Max Values
   * Outlier Analysis

**Part 3: Conclusion**

* Summary of Findings
* Implications for Diabetes Prediction

**PART 1: Data Analysis and Insights**

* 1. **Correlation Matrix**

A screenshot of a computer screen

Description automatically generated

1. **What Do the Colors Mean?**
   * **Red areas** mean a strong positive relationship (both values increase together).
   * **Blue areas** mean a negative or weak relationship (one increases while the other decreases, or no connection at all).
   * The closer the numbers are to 1 or -1, the stronger the relationship.
2. **Key Insights:**
   * **Age and Pregnancies:** Strong positive link (0.54). Older people tend to have had more pregnancies.
   * **Glucose and Outcome:** Moderate positive link (0.47). Higher glucose levels are connected to a higher chance of having diabetes.
   * **BMI and Outcome:** Moderate link (0.29). Higher BMI also slightly increases the chance of diabetes.
3. **Weaker Relationships:**
   * Variables like SkinThickness, Insulin, and BloodPressure have weak links to diabetes (Outcome). They may not predict diabetes as clearly on their own.
4. **What It Means for Diabetes:**
   * Important factors for predicting diabetes include **Glucose, BMI, Age, and Pregnancies**.
   * Other factors like Insulin or SkinThickness might be less important on their own but could still contribute when combined with others.
5. **Scatter plot**A diagram of a scatter plot

   Description automatically generated

**Overall Relationship:**

* Higher BMI generally leads to higher Blood Pressure (red line shows this trend).

**Spread of Data:**

* Most of the data points are clustered between 20 and 40 for BMI and between 50 and 100 for Blood Pressure.
* There are some outliers, particularly:
  + Very low BMI values with varying blood pressure.
  + A few high BMI values that deviate from the trend.

**Outliers:**

* There are points at the extremes, such as very low Blood Pressure with high BMI or vice versa

**Density:**

* The central cluster suggests that most individuals in this dataset fall into a moderate range for both Blood Pressure and BMI.

1. **Distribution plot**
   1. (Histogram) A graph with a red line

      Description automatically generated

**Most Common Values:**

* Most of the data is concentrated around 0-2 pregnancies, indicating that most individuals in the dataset had few or no pregnancies.

**Declining Frequency:**

* The frequency decreases as the number of pregnancies increases, showing that higher numbers of pregnancies are less common.

**Skewed Distribution:**

* The data is positively skewed, with a long tail extending toward higher pregnancy counts.

**Outliers:**

* There are few individuals above 15, shown by the sparse bins at the far right.

**Rare Cases:**

* Few individuals have more than 10 pregnancies, as shown by the low frequency in the higher bins.

b) A graph of a graph

Description automatically generated

**Most Common Values:**

* SkinThickness values around **0 to 20** are the most frequent, with a peak at or near zero.

**Decline in Frequency:**

* The frequency decreases as SkinThickness increases beyond 20, indicating fewer individuals with higher SkinThickness values.

**Skewed Distribution:**

* The data is positively skewed, with a longer tail toward higher SkinThickness values.

**Outliers:**

* There are rare cases where SkinThickness values exceed 60, as indicated by the sparse bins at the right end.

c) A graph of a graph with a red line

Description automatically generated

**Most Common Values:**

* Individuals aged 20-30 are the most frequent, with a clear peak in this age group.

**Decline in Frequency:**

* The frequency decreases steadily beyond 30, indicating fewer older individuals in the dataset.

**Skewed Distribution:**

* The data is positively skewed, with a longer tail toward higher ages, meaning younger individuals are more common.

**Outliers:**

* There are few individuals above 70, shown by the sparse bins at the far right.

**4)Pie chart**

A blue and orange pie chart

Description automatically generated

**Proportions**:

* 65.1% of the population does not have diabetes.
* 34.9% of the population has diabetes.

**Imbalanced Distribution**:

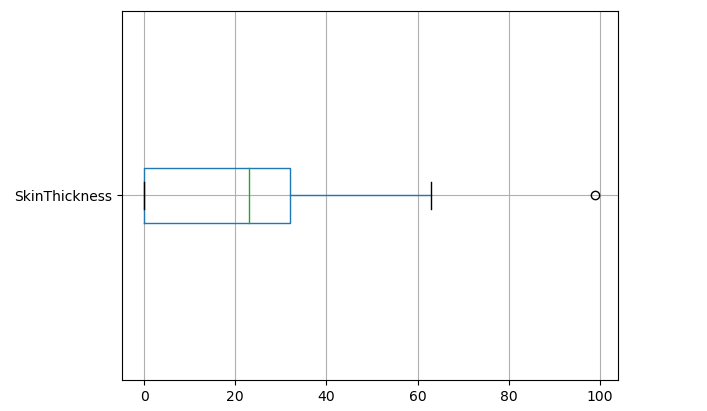
* The dataset appears to be imbalanced, with a higher proportion of "No Diabetes" outcomes compared to "Diabetes" outcomes. This imbalance could influence any predictive modeling or statistical analyses, as models may be biased towards the majority class (No Diabetes).

**Dataset Implications**:

• This dataset's 34.9% diabetes prevalence may offer information about the community under study, possibly indicating greater rates than local or worldwide averages (depending on context).

**5)Boxplot**

* 1. Skinthickness



**Central Tendency**:

* The **median** SkinThickness value is around **20 mm**, indicating that half of the individuals have SkinThickness values below this level.

**Interquartile Range (IQR)**:

* The IQR, represented by the box, ranges roughly from **15 mm to 30 mm**, showing where the middle 50% of the values lie. This suggests most SkinThickness measurements are within this range.

**Spread and Variability**:

* The whiskers extend to approximately **0 mm** and **60 mm**, representing the range of most of the data. Values beyond this range are less common.

**Outliers**:

* There is a noticeable **outlier** at **100 mm**, indicating an unusually high SkinThickness value compared to the rest of the dataset.
  1. BloodPressure

A graph with a blue and green square

Description automatically generated

**Central Tendency:**

* The median blood pressure (green line inside the box) is approximately 80, meaning half of the individuals have blood pressure values below this level.

**Interquartile Range (IQR):**

* The box spans roughly 70 to 85, indicating that the middle 50% of individuals have blood pressure values within this range. This suggests a relatively tight distribution of typical blood pressure readings.

**Spread and Variability:**

* The whiskers extend from about 40 to 100, indicating that most blood pressure values lie within this range. Values outside this range are considered less frequent.

**Outliers:**

* Several outliers are present:
  + On the lower end, there are a few values below 40, which may represent either extremely low blood pressure readings or potential data errors.
  + On the upper end, outliers exceed 100, with some going beyond 120, representing cases of hypertension.
  1. DiabetesPedigreeFunction

A graph with a blue square and black line

Description automatically generated

**Central Tendency:**

**•** The median DPF value is around 0.3, indicating that half of the individuals have a DPF value below this level.

**Interquartile Range (IQR):**

• The middle 50% of the values lie between approximately 0.2 and 0.6, suggesting most individuals have moderate DPF values.

**Spread and Variability:**

• The whiskers extend from around 0.1 to 1.0, showing where the majority of the data points fall. Values outside this range are less common.

**Outliers:**

• Numerous outliers exist beyond 1.0, with some exceeding 2.0, indicating that a significant number of individuals have unusually high DPF values

**Part 2 : Statistical Analysis and Findings**

* 1. **Central Tendency**

A screenshot of a computer screen

Description automatically generated

1. Pregnancies

* Mean: 3.84, Median: 3, Mode: 1
* Insight: The average number of pregnancies is around 4, with most individuals having only 1 pregnancy. The skew suggests that while most individuals have had fewer pregnancies, some outliers have significantly more.

2. Glucose

* Mean: 120.89, Median: 117, Mode: 99
* Insight: Glucose levels are an important predictor of diabetes. The average glucose level is slightly above the normal fasting range (70–99 mg/dL), suggesting a predisposition to higher glucose levels in this population.

3. Blood Pressure

* Mean: 69.15, Median: 72, Mode: 70
* Insight: The blood pressure (likely diastolic) is slightly below the normal range (80-89 mmHg). It could indicate a healthier blood pressure trend overall, but a few outliers might pull the mean lower.

4. Skin Thickness

* Mean: 20.54, Median: 23, Mode: 0
* Insight: Skin thickness, often used to estimate body fat, shows many zero values (mode = 0), indicating missing or zero-filled data. This could skew analysis and may need imputation or further investigation.

5. Insulin

* Mean: 79.80, Median: 30.5, Mode: 0
* Insight: The insulin variable shows a significant discrepancy between the mean and median, with a mode of 0, indicating many missing or zero-filled values. This suggests a need for imputation or a careful handling strategy.

6. BMI (Body Mass Index)

* Mean: 31.99, Median: 32, Mode: 32
* Insight: The average BMI is near the obesity threshold (≥30), indicating that obesity is prevalent in the dataset, a major risk factor for diabetes.

7. Diabetes Pedigree Function

* Mean: 0.47, Median: 0.372, Mode: 0.254
* Insight: This variable quantifies diabetes likelihood based on family history. Most values cluster around 0.254–0.372, indicating a moderate familial risk across the dataset.

8. Age

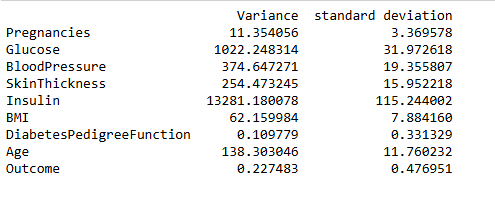
* Mean: 33.24, Median: 29, Mode: 22
* Insight: The average age of participants is 33, with a majority being younger (mode = 22). This relatively young cohort may include individuals at early risk for diabetes.

9. Outcome

* Mean: 0.35, Median: 0, Mode: 0
* Insight: The outcome variable likely represents whether someone has diabetes (1) or not (0). With a mean of 0.35, only about 35% of individuals have diabetes in the dataset, indicating an imbalance (potential class imbalance in the dataset for modeling).

Significance of These Statistics

* Data Quality Issues: Variables like SkinThickness and Insulin show potential issues with missing data (mode = 0). This may require cleaning or imputation before modeling.
* Diabetes Risk Factors: Higher mean glucose and BMI indicate a predisposition to diabetes. Coupled with familial risk (Diabetes Pedigree Function), these variables could be significant predictors.
* Age and Pregnancies: Age distribution and pregnancies (specific to women) indicate a younger cohort, which can affect predictive modeling for diabetes onset.
  1. **Dispersion**



1. Pregnancies

* Variance: 11.35
* Standard Deviation: 3.37  
  This indicates that the number of pregnancies among individuals varies moderately, with a spread of about 3.37 from the mean. Pregnancies are likely categorical or discrete values, so the range is not very high.

2. Glucose

* Variance: 1022.25
* Standard Deviation: 31.97  
  Glucose levels have significant variability, reflecting individual differences in blood sugar levels. This could be critical for understanding diabetes risk, as high glucose levels are a primary marker for the condition.

3. Blood Pressure

* Variance: 374.65
* Standard Deviation: 19.36  
  Blood pressure has a moderate variability, with a standard deviation of about 19.36. Elevated or low blood pressure levels are indicators of various health risks, including cardiovascular conditions.

4. Skin Thickness

* Variance: 254.47
* Standard Deviation: 15.95  
  Skin thickness variability is lower compared to glucose and blood pressure. This measure may be associated with body fat distribution, which could be relevant for assessing obesity or metabolic health.

5. Insulin

* Variance: 13281.18
* Standard Deviation: 115.24  
  Insulin levels have the highest variability, indicating a wide range of insulin concentrations in the population. This is significant, as abnormal insulin levels are closely linked to diabetes and insulin resistance.

6. BMI (Body Mass Index)

* Variance: 62.16
* Standard Deviation: 7.88  
  BMI shows moderate variability, reflecting differences in weight relative to height across individuals. High BMI values are commonly associated with obesity and related health issues.

7. Diabetes Pedigree Function

* Variance: 0.11
* Standard Deviation: 0.33  
  The diabetes pedigree function (a measure of family history of diabetes) has very low variability, suggesting most individuals have similar values, with less spread.

8. Age

* Variance: 138.30
* Standard Deviation: 11.76  
  Age shows moderate variability, which is expected in population datasets. Age is an essential factor influencing many health outcomes, including diabetes risk.

9. Outcome (Diabetes Presence)

* Variance: 0.23
* Standard Deviation: 0.48  
  The binary outcome (likely coded as 0 or 1 for the presence or absence of diabetes) has low variability. The standard deviation suggests a relatively balanced dataset where neither class dominates significantly.

Significance of These Statistics

1. Feature Importance: The variability in glucose and insulin levels suggests these are key features in predicting diabetes. High variability implies a strong ability to differentiate between individuals with and without diabetes.
2. Clinical Relevance: Blood pressure, BMI, and age also play a significant role in health outcomes, especially in conditions like diabetes and cardiovascular diseases.

**3)MIN & MAX**

A screenshot of a computer screen

Description automatically generated

1. Pregnancies

* Max: 17.00
* Min: 0.00  
  Pregnancies range from 0 to 17, indicating that some individuals have never been pregnant while others have had multiple pregnancies. This range aligns with expectations in datasets related to health and reproduction. High pregnancy counts could potentially correlate with health risks like gestational diabetes.

2. Glucose

* Max: 199.00
* Min: 0.00  
  The wide range in glucose levels is notable. A minimum value of 0 may indicate missing or erroneous data, as a glucose level of 0 is biologically implausible. This highlights the need for data cleaning or imputation. High glucose values (up to 199) suggest potential cases of diabetes or prediabetes.

3. Blood Pressure

* Max: 122.00
* Min: 0.00  
  A blood pressure value of 0 is also implausible, likely representing missing data. The upper limit of 122 is within normal to slightly elevated ranges, but the dataset might not include hypertensive individuals with extreme blood pressure values.

4. Skin Thickness

* Max: 99.00
* Min: 0.00  
  A minimum value of 0 likely represents missing data rather than actual measurements. A maximum value of 99 suggests significant variability in subcutaneous fat, which could be relevant for assessing obesity or metabolic health.

5. Insulin

* Max: 846.00
* Min: 0.00  
  Insulin values range widely, with a maximum of 846 indicating hyperinsulinemia (excess insulin), commonly associated with diabetes or insulin resistance. A minimum value of 0 likely represents missing or unmeasured data.

6. BMI (Body Mass Index)

* Max: 67.10
* Min: 0.00  
  The maximum BMI indicates severe obesity, while a value of 0 is implausible and likely represents missing data. BMI is a critical feature for assessing obesity and its associated risks.

7. Diabetes Pedigree Function

* Max: 2.42
* Min: 0.078  
  The diabetes pedigree function measures the likelihood of diabetes based on family history. The range (0.078 to 2.42) suggests diverse genetic risks, with higher values indicating a strong hereditary predisposition.

8. Age

* Max: 81.00
* Min: 21.00  
  The age range indicates the dataset includes adults from young to elderly, which is important for studying diabetes risk as age is a significant factor.

9. Outcome (Diabetes Presence)

* Max: 1.00
* Min: 0.00  
  This binary variable indicates the presence (1) or absence (0) of diabetes. The range confirms it is a binary classification target.

**Significance and Actionable Insights**

1. Data Cleaning Required:
   * Variables with a minimum value of 0 (e.g., glucose, blood pressure, insulin, BMI, skin thickness) likely have missing or erroneous data. These should be handled through imputation or exclusion to improve data quality.
2. Feature Importance:
   * High ranges in glucose, insulin, and BMI reflect their critical roles in identifying diabetes risk and should be prioritized in predictive modeling.
3. Diversity in Data:
   * The wide range in variables like age and diabetes pedigree function indicates a diverse dataset, useful for generalizing results to different populations.
4. Potential Outliers:
   * Extremely high values (e.g., insulin = 846, BMI = 67.1) may represent valid data for outliers but should be verified to ensure they are not errors.

**Conclusion:**

**1. Relationships Between Variables:**

* **Age and Pregnancies:** Strong positive correlation (r=0.54r = 0.54); older individuals tend to have more pregnancies.
* **Glucose and Outcome:** Moderate positive correlation (r=0.47r = 0.47); higher glucose levels are associated with a greater likelihood of diabetes.
* **BMI and Outcome:** Moderate correlation (r=0.29r = 0.29); higher BMI increases the probability of diabetes.
* **Weak Relationships:** Variables such as SkinThickness, Insulin, and BloodPressure show limited predictive power for diabetes individually.

**2. Data Characteristics and Trends:**

* **Pregnancies:** Positively skewed with most individuals having 0–2 pregnancies. Few cases exceed 10 pregnancies.
* **Age:** Concentrated in the younger population (20–30 years); very few individuals are older than 70.
* **BMI and Blood Pressure:** BMI correlates positively with blood pressure. Most data cluster in the 20–40 BMI range and 50–100 for blood pressure, with a few outliers.

**3. Class Distribution (Outcome):**

* **Diabetes Prevalence:** 34.9% of the population has diabetes, while 65.1% does not. This imbalance may affect predictive modeling.

**4. Statistical Insights:**

* **Central Tendency and Spread:**
  + **Glucose:** High variability (mean = 120.89; standard deviation = 31.97), with values ranging from 0 (likely missing data) to 199.
  + **BMI:** Mean of 31.99, highlighting obesity prevalence.
  + **Insulin:** Extremely variable (standard deviation = 115.24), suggesting data issues or heterogeneity.
* **Outliers:** Notable for variables such as SkinThickness (max = 99 mm) and Insulin (max = 846).

**5. Data Quality Issues:**

* Variables like SkinThickness, Insulin, and BloodPressure have numerous zero values, likely representing missing data.
* Missing or erroneous data require imputation or exclusion to improve analysis reliability.

**6. Predictive Factors for Diabetes:**

* Key predictors identified include **Glucose**, **BMI**, **Age**, and **Pregnancies**.
* Familial history (as indicated by the Diabetes Pedigree Function) also plays a role but exhibits lower variability.