Project Title: Data Visualization and Statistics Using Python

Student 1: Ziad Mohamed Ibrahim
Student 1 **ID**: 247425

Student 2: Mohamed Hesham Mohamed

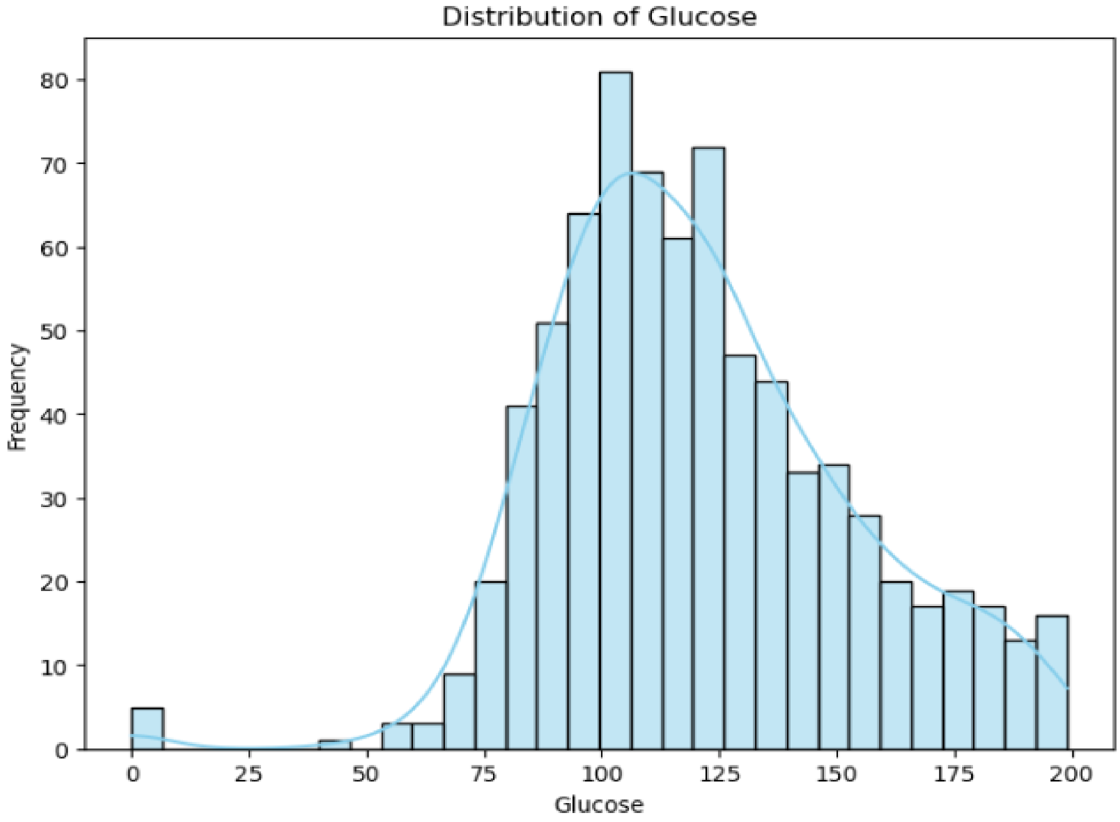
Student 2 **ID**: 244327

Introduction

This project aims to analyze a diabetes dataset using Python to uncover patterns and insights through visualizations analysis. The dataset includes features such as glucose levels, BMI, age, and diabetes outcomes. By visualizing and summarizing the data, we aim to better understand the trends and relationships within the dataset.

Visualizations and Analysis

1. Distribution of Glucose Levels

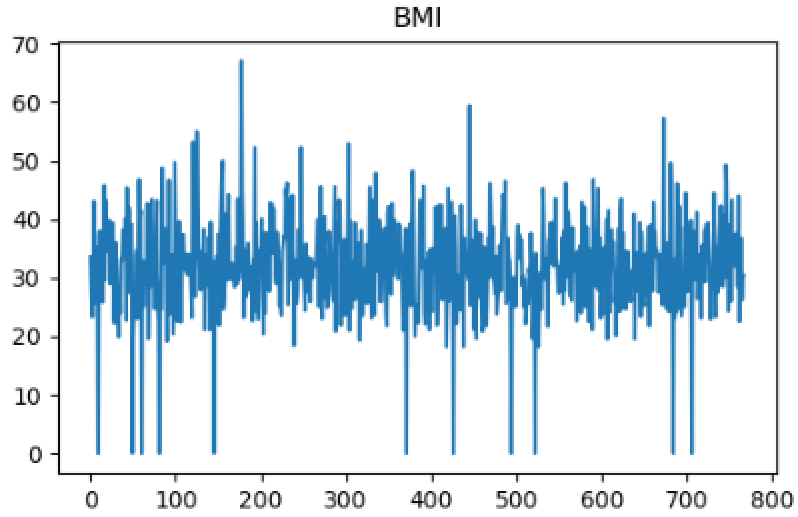


Type of Visualization: Histogram

Description:

The histogram shows the frequency distribution of glucose levels in the dataset. The x-axis represents the glucose levels, and the y-axis represents the number of individuals with those levels.

2. BMI Distribution by Diabetes Outcome

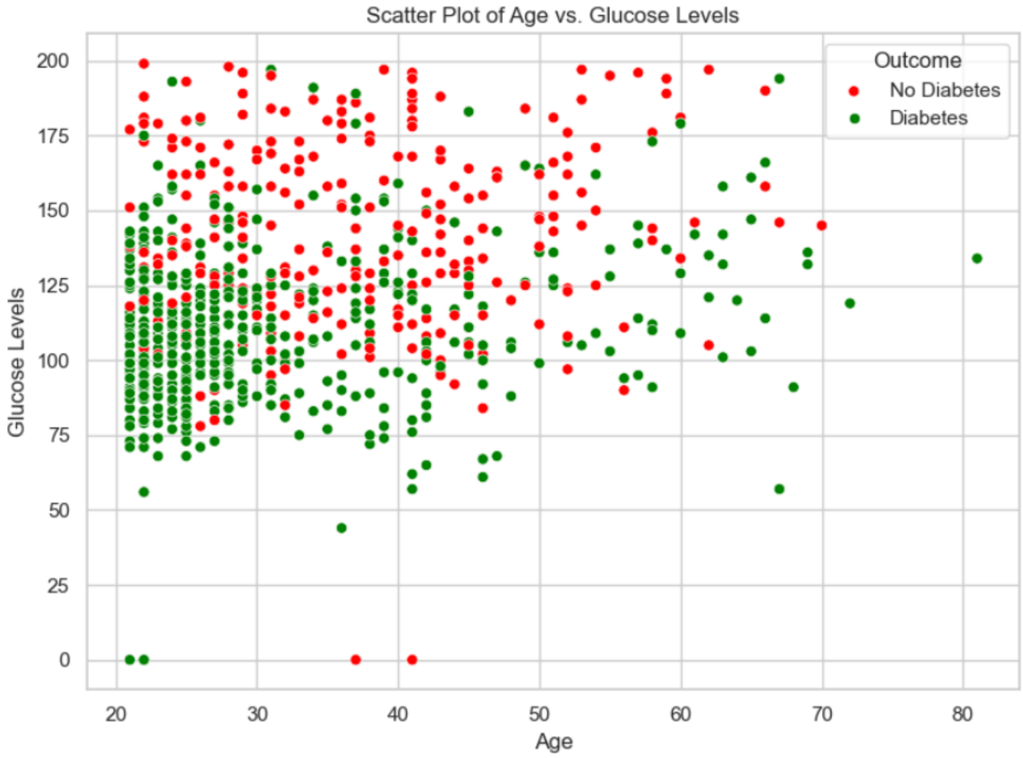


Type of Visualization: Box Plot

Description:

The box plot compares the BMI of individuals with and without diabetes The x-axis shows the diabetes outcome, and the y-axis shows BMI values.

3. Age vs. Glucose Levels

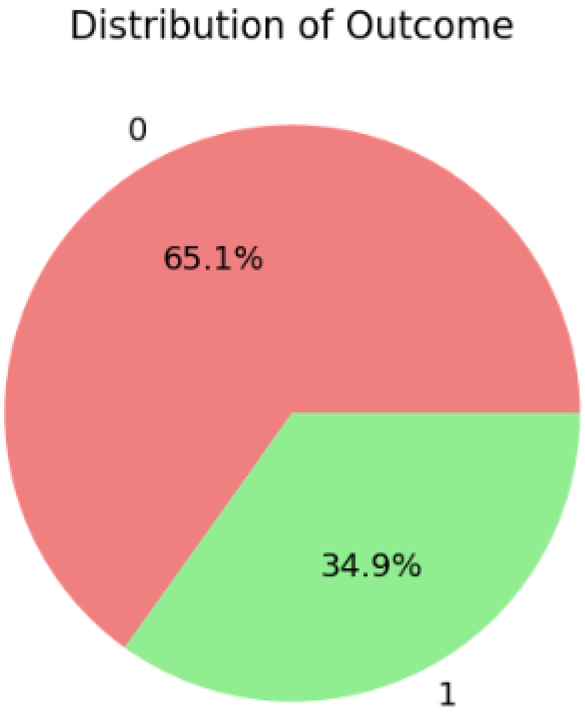


Type of Visualization: Scatter Plot

Description:

The scatter plot visualizes the relationship between age (x-axis) and glucose levels (y-axis), with different colors representing diabetes outcomes.

4.Distribution of Diabetes Outcome



Type of Visualization: Pie Chart

Description:

The pie chart shows the proportion of individuals with and without diabetes. The two categories are represented in red (diabetes) and green (no diabetes).

Calculated Statistics:

Feature	Mean	Standard Deviation	Median	Mode	variance	Min	Max
Pregnancies	3.85	3.37	3.00	1	11.35	0	17
Glucose	120.89	31.97	117.00	99	1022.25	0	199
Blood Pressure	69.11	19.36	72.00	70	374.65	0	122
Skin Thickness	20.54	15.95	23.00	0	254.47	0	99
Insulin	79.80	115.24	30.50	0	13281.18	0	846
BMI	31.99	7.88	32.00	32	62.16	0	67.1
Diabetes Pedigree Function	0.47	0.33	0.37	0.254	0.11	0.078	2.42
Age	33.24	11.76	29.00	22	138.30	21	81
Outcome	0.35	0.48	0.00	0	0.23	0	1

Insights:

1. **Pregnancies**:
    - Average number of pregnancies is about (3.85). The median is (3). Most individuals had 3 pregnancies or fewer.
    - The standard deviation is (3.37) among the participants.
    - The mode is (1) , reflecting that many individuals in the dataset had only one pregnancy.
    - The variance indicate a spread of values but not extreme variability.
    - The min value is (0), which might indicate missing data or individuals with extreme physiological conditions, but max is (17) representing a participant with a very high number of pregnancies. .
2. **Glucose**:
    - The mean of glucose is (120.89). The median is (117). This is a slightly skewed distribution with some high glucose levels.
    - The standard deviation is (31.97). It can be noticeable variation in glucose levels between participants.
    - The mode is (99)mg/dL , occurs most frequently.
    - Shows moderate-to-high variance, reflecting a wide range of measurements in the population.

- The min value is (0), which might indicate missing data or individuals with extreme physiological conditions , but the max is (199) mg/dL.

3. **Blood Pressure**:
   - The blood pressure’s mean is (69.11)mm Hg. The median is (72)mm Hg. The values are so close and it suggest a symmetric distribution.
   - The standard deviation is (19.36) points to moderate variability.
   - The mode is (70) mmHg is the most frequent.
   - Shows moderate-to-high variance, reflecting a wide range of measurements in the population.
   - The min value is (0), which might indicate missing data or individuals with extreme physiological conditions , but the max is (122)mmHg which is on the higher side of normal.

4. **Skin Thickness**:
   - The mean of skin thickness is (20.54)mm. It’s median is (23)mm. It shows a skew towards lower values , and many participants may have missing or zero value.
   - High standard deviation (15.95) supports the wide spread of measurements.

- The mode is (0) , which might indicate missing or unmeasured values in the dataset for these features.
- Shows moderate-to-high variance, reflecting a wide range of measurements in the population as blood pressure and glucose.
- The min value is (0), which might indicate missing data or individuals with extreme physiological conditions , but the max value is (99) and it suggest some participants may have very high fat thickness.

5. **Insulin**:
   - The mean is (79.80) of insulin level , and the median is more lower than mean by (30.50). It shows a strong skew from some extreme high values.
   - The standard deviation is (115.24) supports this observation.
   - The mode is (0) , which might indicate missing or unmeasured values in the dataset for these features.
   - It has the highest variance by (13281.18) , indicating significant differences in insulin levels across individuals.
   - The min value is (0), which might indicate missing data or individuals with extreme physiological conditions , but the max is (846) and it indicates extremely high insulin levels.

6. **BMI**:
   - The mean and median BMI are almost identical (31.99 , 32), indicating a well-centered and likely symmetric distribution.
   - A standard deviation of (7.88) shows moderate variability in BMI.
   - The mode is (32) , shows a concentration around this level.
   - It has a moderate variance, suggesting that most BMI values are close to the average but with some outliers.
   - The min value is (0), which might indicate missing data or individuals with extreme physiological conditions , but the max BMI is (67.1) it shows obesity in some participants.

7. **Diabetes Pedigree Function**:
   - The mean of diabetes pedigree function is (0.47), the median is (0.37), and the standard deviation is (0.33). It is relatively tightly distributed around the mean.
   - The mode is (0.254) , indicating a common genetic predisposition value.
   - Has the lowest variance , means most values are tightly clustered near the mean.
   - The minimum value is **0.078**, representing a very low genetic predisposition to diabetes , but the max value is (2.42) it indicates a participants with very strong genetic.

8. **Age**:

   - The average participant age is (33.24) years, with a median age of (29) years. This indicates a younger unit.
   - A standard deviation of (11.76) shows a wide age range.
   - The mode is (22), reflecting a younger population in the dataset.
   - The variance indicating a spread of values but not extreme variability as the pregnancies.
   - The Min is (21) years, reflecting the youngest participants in the dataset , but the max is (81) years with the oldest individuals.

9. **Outcome**:

   - The average outcome values is (0.35), indicating approximately (35%) of participants are diabetic, this can tell that the dataset is being skewed towards non-diabetic individuals.
   - The mode is (0) , meaning the majority of individuals do not have diabetes.
   - The Min value is (0) indicates individuals without diabetes.