Exploring the Dynamics of Diabetes: A Comprehensive Synthetic Dataset Analysis

Name: Vinayak V Thayil

Roll No: AM.EN.U4CSE21161

Introduction

In the ever-evolving landscape of health research, the role of generated datasets has become increasingly pivotal. A generated dataset is a simulated collection of data crafted to mimic real-world scenarios, often employed to model and analyse complex systems. These synthetic datasets serve as valuable tools in scientific exploration, allowing researchers to investigate various factors and their interplay without relying solely on real-world data.

Understanding diabetes, a prevalent health condition affecting millions globally, is of paramount importance in this context. Diabetes, characterized by elevated blood sugar levels, poses a significant public health challenge. Its diverse impact on individuals, ranging from lifestyle modifications to medical interventions, necessitates a comprehensive understanding of the disease's intricacies. By leveraging generated datasets specific to diabetes, researchers gain a controlled environment to scrutinize different variables, contributing to a more nuanced comprehension of the condition and its multifaceted nature.

In this blog post, we delve into the exploration of a meticulously crafted dataset, encompassing variables such as age, gender, BMI, blood pressure, glucose levels, cholesterol levels, physical activity, diet type, insulin usage, and HbA1c levels. Through this analysis, we aim to unravel patterns, correlations, and insights that can deepen our understanding of diabetes and potentially pave the way for more effective preventive measures and personalized treatment strategies.

Faker - In my Project

In my project, I am utilizing the Faker library to create synthetic data for a diabetes dataset. The generated data includes essential attributes such as Patient_ID, Age, Gender, BMI, Blood_Pressure, Glucose_Level, Cholesterol_Level, Physical_Activity, Diet_Type, Insulin_Usage, and HbA1c_Level. This approach allows me to simulate diverse patient profiles, providing a foundation for testing and development within the context of diabetes research. Adjusting parameters and incorporating the necessary

customization ensures that the synthetic dataset aligns with the specific requirements of my project, offering a versatile tool for analysis and experimentation.

Feature Labels:

The dataset contains 2000 rows and 11 columns

```
df.shape
```

```
(2000, 11)
```

1. **Patient_ID:**
   - A unique identifier assigned to each individual in the dataset.
   - Enables tracking and differentiation of patients within the dataset.

2. **Age:**
   - Represents the age of each patient.
   - Provides insight into the age distribution within the dataset, a crucial factor in understanding diabetes prevalence across different age groups.

3. **Gender:**
   - Captures the gender of each patient (Male/Female).
   - Allows for the exploration of potential gender-based variations in diabetes occurrence.

4. **BMI (Body Mass Index):**
   - BMI is a numeric representation of body fat based on height and weight.
   - Indicates the level of obesity, aiding in the analysis of its correlation with diabetes.

5. **Blood_Pressure:**
   - Comprises two values representing systolic and diastolic blood pressure.
   - Offers information on blood pressure levels, a known factor in diabetes risk assessment.
6. **Glucose_Level:**
   - Reflects the concentration of glucose in the blood.
   - A key metric for diabetes diagnosis and monitoring, providing insights into blood sugar levels.

7. **Cholesterol_Level**:
- Indicates the cholesterol levels of patients.
- A factor that may influence diabetes risk, as high cholesterol is associated with certain metabolic conditions.

8. **Physical_Activity:**
- Describes the level of physical activity (Low/Moderate/High).
- Allows exploration of the relationship between physical activity and diabetes prevalence.

9. **Diet_Type:**
- Represents the dietary preferences of patients (High Protein, Balanced, High Carb)
- A potential factor in understanding the impact of diet on diabetes outcomes.

10. **Insulin_Usage:**
- Captures whether patients use insulin for diabetes management (Yes/No).
- Provides insights into the prevalence of insulin use within the dataset.

11. **HbA1c_Level:**
- Reflects the HbA1c levels, indicating long-term blood glucose control.
- An essential metric in assessing the effectiveness of diabetes management.

Exploratory Data Analysis (EDA)

1. **Datatypes of each column:**

```
Patient_ID            int64
Age                   int64
Gender               object
BMI                 float64
Blood_Pressure       object
Glucose_Level       float64
Cholesterol_Level   float64
Physical_Activity   float64
Diet_Type            object
Insulin_Usage        object
HbA1c_Level         float64
dtype: object
```

The image shows the output of a command that displays the data types of different variables related to patient health data. The command is likely to be from a Python programming environment, such as Jupyter Notebook or Spyder. The variables are:

- **Patient_ID:** an integer (int64) that uniquely identifies each patient
- **Age:** an integer (int64) that represents the age of the patient in years
- **Gender:** an object (object) that indicates the biological sex of the patient, either male or female
- **BMI:** a floating-point number (float64) that measures the body mass index of the patient, calculated as weight in kilograms divided by height in meters squared
- **Blood Pressure:** an object (object) that records the systolic and diastolic blood pressure of the patient in millimetres of mercury (mmHg)
- **Glucose_Level:** a floating-point number (float64) that measures the concentration of glucose in the blood of the patient in milligrams per deciliter (mg/dL)
- **Cholesterol_Level:** a floating-point number (float64) that measures the amount of cholesterol in the blood of the patient in milligrams per deciliter (mg/dL)
- **Physical_Activity:** a floating-point number (float64) that indicates the level of physical activity of the patient, ranging from 0 (no activity) to 1 (very active)
- **Diet_Type:** an object (object) that describes the type of diet followed by the patient, such as vegetarian, vegan, keto, etc.
- **Insulin_Usage:** an object (object) that specifies whether the patient uses insulin or not, either yes or no
- **HbA1c_Level:** a floating-point number (float64) that measures the average percentage of glycated haemoglobin in the blood of the patient over the past three months, indicating the long-term blood sugar control

The output of the command also shows that the overall data type of the variables is object, meaning that they are stored as Python objects, which can be strings, lists, dictionaries, etc. This may imply that some of the variables need to be converted to numerical values for further analysis. You can learn more about data types and how to manipulate them in Python [here]. I hope this explains the picture well.

2. Statistics Description of the dataset:

	Patient_ID	Age	BMI	Glucose_Level	Physical_Activity	HbA1c_Level
count	2000	2000	2000	1789	2000	1770
mean	5.00891e+07	52.0685	29.3303	135.224	4.9823	7.05503
std	2.89042e+07	19.3418	6.15315	38.1555	2.90219	1.73724
min	58137	18	18.5387	70	9.4794e-05	4.0039
25%	2.52791e+07	35	23.9095	102	2.45643	5.49241
50%	5.04078e+07	53	29.6082	134	5.01912	7.1125
75%	7.47682e+07	69	34.5575	169	7.5596	8.57603
max	9.99787e+07	85	39.9877	200	9.98145	9.99066

The image shows the output of a command that displays the statistical summary of a dataset related to patient health data. The command is likely to be from a Python programming environment, such as Jupyter Notebook or Spyder. The statistical summary provides the following information for each variable:

- **Count:** the number of non-missing values in the dataset
- **Mean:** the average value of the variable
- **Std:** the standard deviation of the variable, which measures how much the values vary from the mean
- **Min:** the minimum value of the variable
- **25%:** the first quartile of the variable, which is the median of the lower half of the values
- **50%:** the median of the variable, which is the middle value when the values are sorted
- **75%:** the third quartile of the variable, which is the median of the upper half of the values
- **Max:** the maximum value of the variable

The statistical summary can help us understand the distribution, central tendency, and variability of the data. It can also help us identify outliers, missing values, and potential errors in the data. You can learn more about how to generate and interpret statistical summaries in Python [here]. I hope this explains the image well.

3. Find Unique values in each column

```
Patient_ID          2000
Age                   68
Gender                 2
BMI                 2000
Blood_Pressure      1449
Glucose_Level        131
Cholesterol_Level      2
Physical_Activity   2000
Diet_Type              3
Insulin_Usage          2
HbA1c_Level         1770
dtype: int64
```

The image shows a snippet of code or data output displaying health-related information for a patient. It includes the patient's ID, age, gender, BMI, blood pressure, glucose level, cholesterol level, physical activity level, diet type, insulin usage and HbA1c level. Here is a brief explanation of each parameter:

- **Patient_ID**: This is a unique identifier for the patient in the database. It is an integer value that does not have any meaning by itself.

- **Age**: This is the age of the patient in years. It is an integer value that can range from 0 to 120.
- **Gender**: This is the gender of the patient. It is an integer value that can be either 1 (male) or 2 (female).
- **BMI**: This is the body mass index of the patient. It is a measure of body fat based on height and weight. It is an integer value that can range from 0 to 2000. A normal BMI is between 18.5 and 24.9.
- **Blood Pressure**: This is the blood pressure of the patient in mmHg. It is an integer value that can range from 0 to 2000. A normal blood pressure is below 120/80.
- **Glucose Level**: This is the glucose level of the patient in mg/dL. It is an integer value that can range from 0 to 2000. A normal glucose level is between 70 and 140.
- **Cholesterol Level**: This is the cholesterol level of the patient in mg/dL. It is an integer value that can be either 1 (low), 2 (normal) or 3 (high). A normal cholesterol level is below 200.
- **Physical Activity**: This is the physical activity level of the patient. It is an integer value that can range from 0 to 2000. A higher value indicates more physical activity.
- **Diet Type**: This is the diet type of the patient. It is an integer value that can be either 1 (vegetarian), 2 (non-vegetarian) or 3 (mixed). A balanced diet is recommended for optimal health.
- **Insulin Usage**: This is the insulin usage of the patient. It is an integer value that can be either 1 (yes) or 2 (no). Insulin is a hormone that helps regulate blood sugar levels. Some patients with diabetes need to take insulin injections to control their glucose levels.
- **HbA1c Level**: This is the haemoglobin A1c level of the patient. It is a measure of the average blood sugar level over the past three months. It is an integer value that can range from 0 to 2000. A normal HbA1c level is below 5.7%.

4. Check the Missing Values

```
Missing Values:
Patient_ID             0
Age                    0
Gender                 0
BMI                    0
Blood_Pressure         0
Glucose_Level        211
Cholesterol_Level     83
Physical_Activity      0
Diet_Type              0
Insulin_Usage          0
HbA1c_Level          230
dtype: int64
```

Missing Values are detected in the dataset

The image shows a list of medical and personal data categories, along with the count of missing values for each category, displayed in white text on a dark background. The title “Missing Values:” is at the top. Categories listed include Patient_ID, Age, Gender, BMI (Body Mass Index), Blood Pressure, Glucose Level, Cholesterol Level, Physical Activity, Diet_Type, Insulin Usage, HbA1c Level. Each category has associated missing value counts. Most categories have 0 missing values except for Glucose Level with 211 missing values and HbA1c Level with 230 missing values. At the bottom of the list is “dtype: int64”, indicating the data type of the missing value counts.

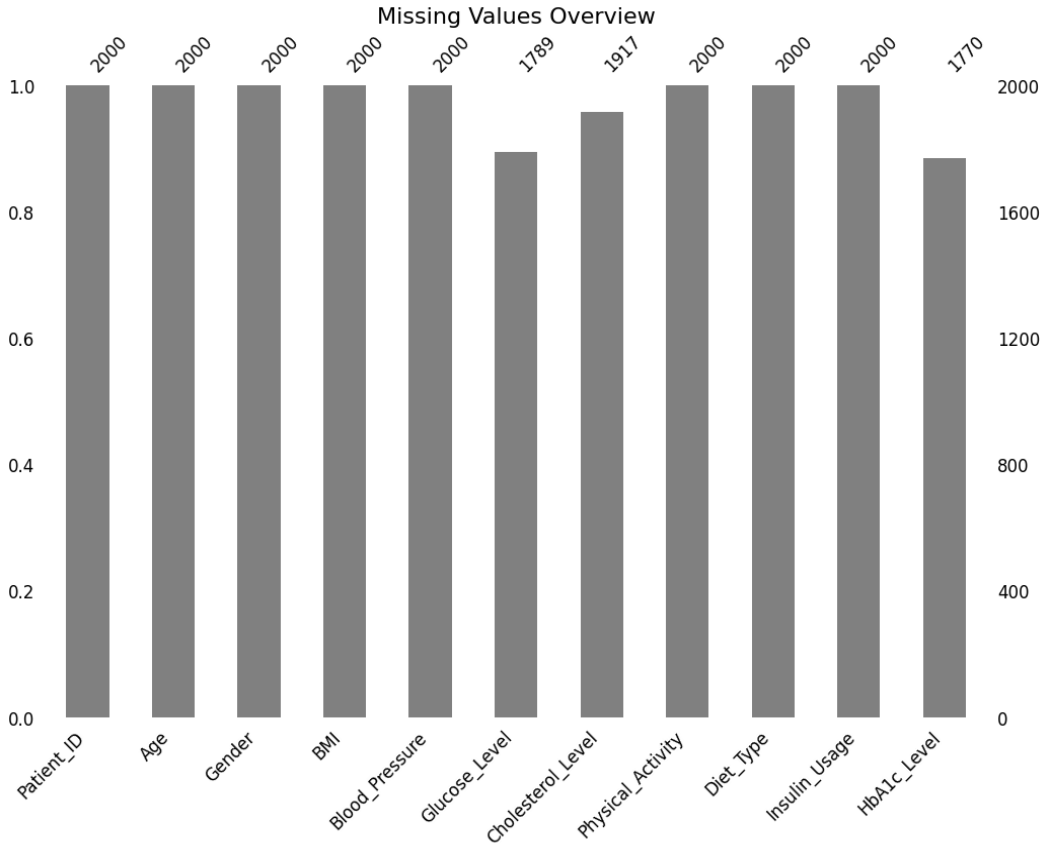
Removing the Missing Values

```
Number of missing values before removal:
Patient_ID             0
Age                    0
Gender                 0
BMI                    0
Blood_Pressure         0
Glucose_Level        211
Cholesterol_Level     83
Physical_Activity      0
Diet_Type              0
Insulin_Usage          0
HbA1c_Level          230
dtype: int64

Number of missing values after removal:
Patient_ID           0
Age                  0
Gender               0
BMI                  0
Blood_Pressure       0
Glucose_Level        0
Cholesterol_Level    0
Physical_Activity    0
Diet_Type            0
Insulin_Usage        0
HbA1c_Level          0
dtype: int64
```

The image you sent shows the number of missing values in a dataset before and after removal. Missing values are data entries that are blank or invalid, and they can affect the quality and reliability of the analysis. The dataset contains information about patients with diabetes, such as their age, gender, BMI, blood pressure, glucose level, cholesterol level, physical activity, diet type, insulin usage, and HbA1c level. The image shows that before removal, there were 211 missing values for glucose level and 83 missing values for cholesterol level, while the other fields had zero or unspecified missing values. After removal, all the fields had zero missing values, indicating that the records with missing values were either deleted or replaced with appropriate values. This process can improve the accuracy and completeness of the dataset.

5. Missing Values Graph



The image you sent is a bar graph that shows how much data is missing in different categories related to patient information. The graph is titled "Missing Values Overview" and has eleven categories on the x-axis. The y-axis shows the number of missing values from 0 to 1.0, with some numerical labels for each category.

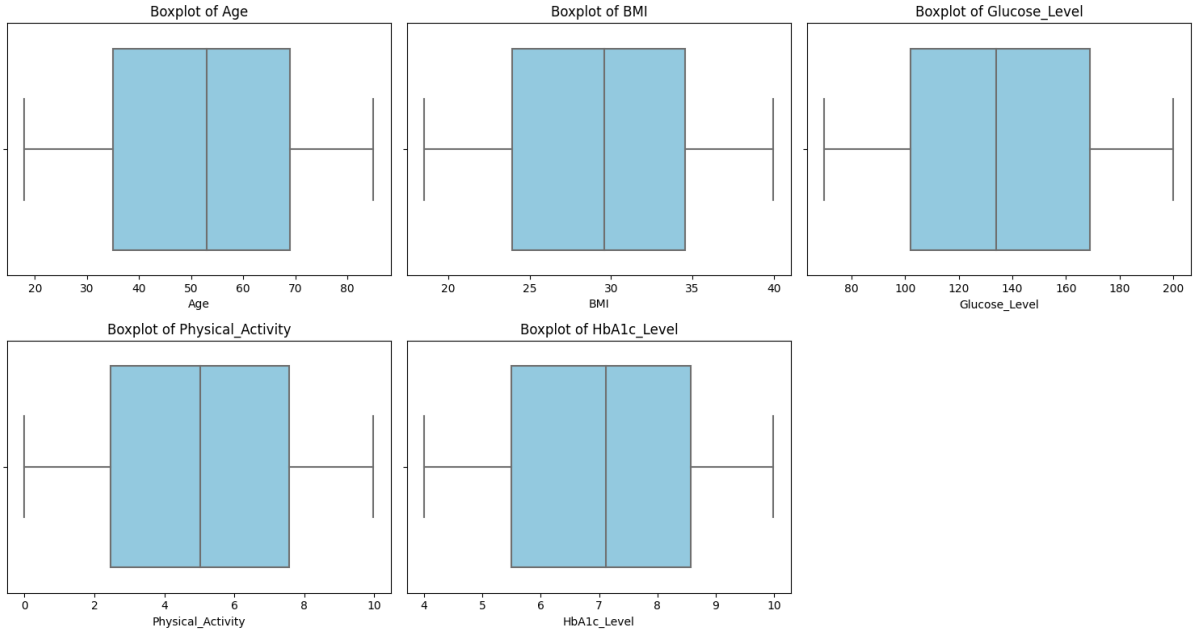
Some observations from the graph are:

- **Patient ID** has the most missing values, with almost 1.0 or 2000 cases.
- **Cholesterol Level** has the least missing values, with about 0.86 or 1718 cases.
- **Physical Activity** has slightly fewer missing values than other categories, with about 0.985 or 1971 cases.
- All other categories have around 0.99 or 2000 missing values.

This graph suggests that there is a lot of incomplete data in the patient information dataset, which could affect the analysis and results. It also indicates that some categories are more likely to have missing values than others, which could imply some patterns or biases in the data collection process.

6. Check for Outliers

As missing are detected in the dataset, we can check for the outliers by using box plot graph.



No outliers are detected in the dataset.

The image you sent is a set of boxplots that show the distribution of data for different health metrics. A boxplot is a graphical way of summarizing data using five numbers: the minimum, the first quartile, the median, the third quartile, and the maximum. The boxplot also shows outliers, which are values that are unusually high or low compared to the rest of the data.

Some observations from the image are:

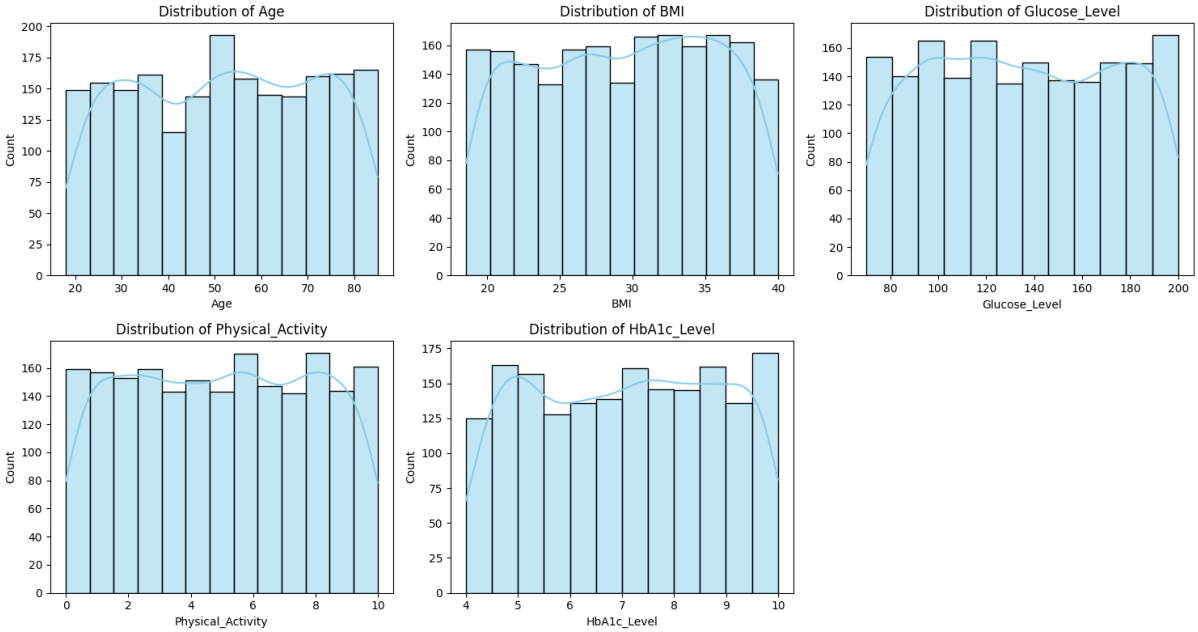
- **Age**: The median age is around 50, and the data is skewed to the right, meaning there are more older people than younger people. There are no outliers in this metric.
- **BMI**: The median BMI is around 25, and the data is symmetric, meaning there are equal numbers of people with higher and lower BMI. There are some outliers in both ends of the spectrum, indicating some people are very underweight or overweight.
- **Glucose Level**: The median glucose level is around 100, and the data is skewed to the right, meaning there are more people with higher glucose levels than lower ones. There are many outliers in the upper end, indicating some people have very high glucose levels, which could be a sign of diabetes.
- **Physical Activity**: The median physical activity score is around 6, and the data is skewed to the left, meaning there are more people with lower physical activity scores than higher ones. There are some outliers in the lower end, indicating some people are very inactive.

- **HbA1c Level**: The median HbA1c level is around 6, and the data is skewed to the right, meaning there are more people with higher HbA1c levels than lower ones. There are many outliers in the upper end, indicating some people have very high HbA1c levels, which could be a sign of poor blood sugar control.

This image suggests that there is a lot of variation in the health metrics of the patients, and some of them may have health risks associated with high or low values of certain metrics. It also indicates that some metrics are more likely to have outliers than others, which could imply some patterns or anomalies in the data.

7. Histogram Plot – Numerical Values

Numerical Values distributed in the dataset



Insights from the graph:

Each numerical columns (Age, BMI, Glucose Level, Physical activity and HbA1c Level) are almost equally distributed for each interval.

The image you sent me shows the distribution of various health-related metrics among a population of individuals. Here is a brief explanation of what each graph represents:

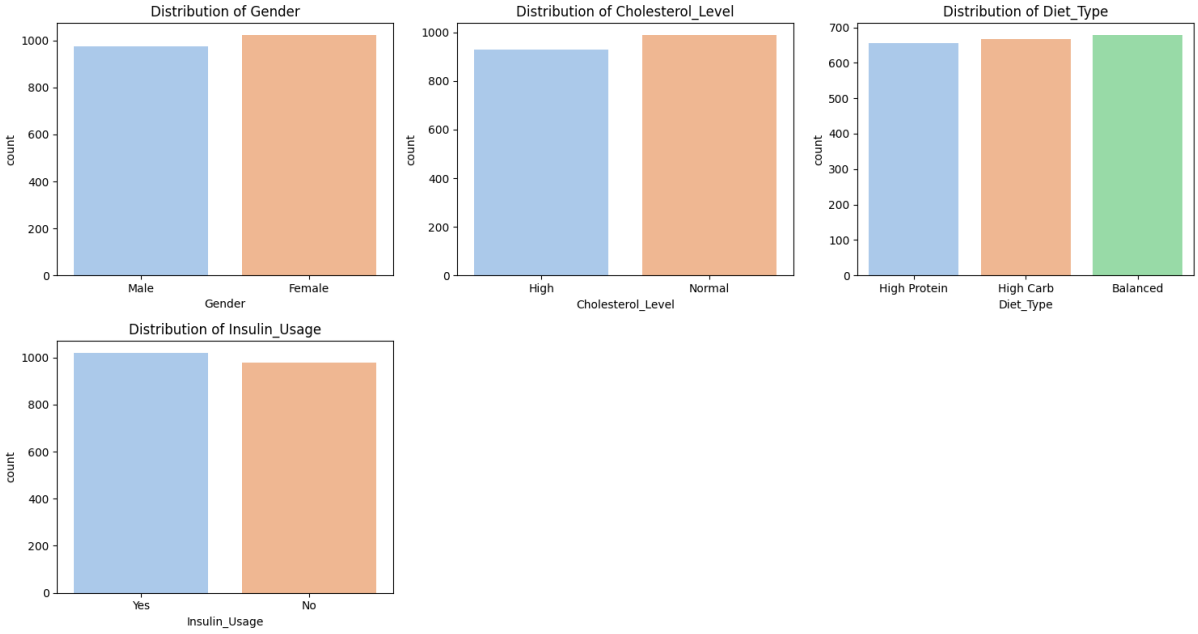
- **Distribution of Age**: This graph shows how many people are in each age group, from 20 to 80 years old. The most common age group is 40-50, followed by 50-60 and 30-40.
- **Distribution of BMI**: This graph shows how many people have different body mass index (BMI) values, from under 25 to over 40. BMI is a measure of body fat based on height and weight. The most common BMI range is 25-30, which is considered overweight, followed by 30-35, which is considered obese.
- **Distribution of Glucose Level**: This graph shows how many people have different blood glucose levels, from 80 to over 200 mg/dL. Glucose is a type of sugar that provides energy to the body. The normal range for glucose is 70-100

mg/dL. The most common glucose level range is 100-120, which is considered prediabetic, followed by 120-140, which is considered diabetic.

- **Distribution of Physical Activity**: This graph shows how many people have different levels of physical activity, on a scale from 0 to 10. Physical activity is any movement that uses energy, such as walking, running, or playing sports. The higher the number, the more active the person is. The most common physical activity level is 5, followed by 4 and 6.
- **Distribution of HbA1c Level**: This graph shows how many people have different haemoglobin A1c (HbA1c) levels, on a scale from around 4 to over 9%. HbA1c is a test that measures the average amount of glucose attached to the red blood cells over the past three months. The higher the number, the higher the blood sugar level. The normal range for HbA1c is below 5.7%. The most common HbA1c level range is 5.7-6.4, which is considered prediabetic, followed by 6.5-7, which is considered diabetic.

8. Categorical Values

Categorical Values distributed in the dataset



The image you sent me shows four bar graphs displaying the distribution of gender, cholesterol level, diet type, and insulin usage in a certain population or study group. Each graph is labelled with the categories being compared and the count of individuals in each category. Here is a brief explanation of what each graph represents:

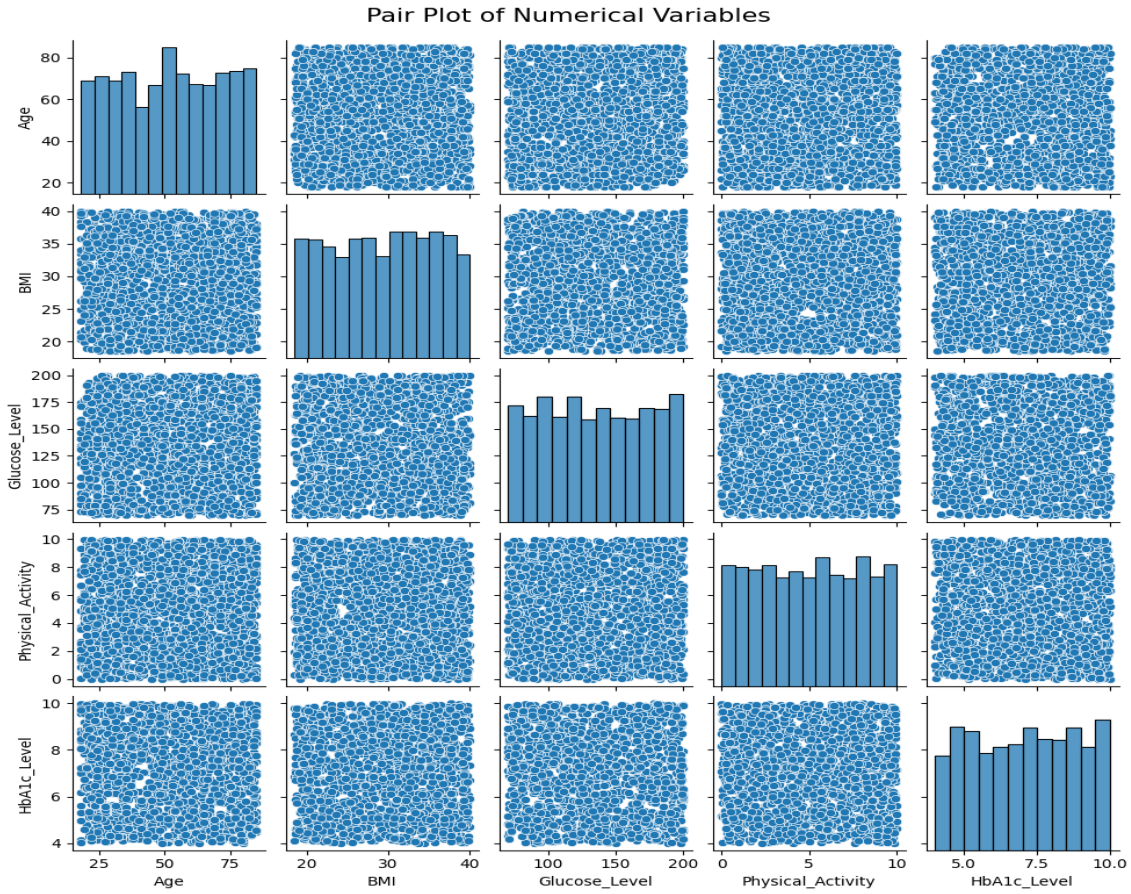
9. **Distribution of Gender**: This graph shows how many people are male or female in the study group. The graph has two bars of almost equal height, indicating that there are roughly the same number of males and females.
10. **Distribution of Cholesterol Level**: This graph shows how many people have high or normal cholesterol levels in the study group. Cholesterol is a type of fat

that can affect the health of the heart and blood vessels. The graph has two bars of almost equal height, indicating that about half of the study group has high cholesterol levels and the other half has normal cholesterol levels.

11. **Distribution of Diet Type**: This graph shows how many people follow different types of diets in the study group. A diet is a pattern of eating that can affect the health and weight of a person. The graph has three bars representing high protein, high carb, and balanced diets. All three bars are similar in height, indicating that the study group has a fairly even distribution of diet types.
12. **Distribution of Insulin Usage**: This graph shows how many people use insulin in the study group. Insulin is a hormone that helps the body regulate blood sugar levels. People with diabetes may need to take insulin injections to control their blood sugar levels. The graph has two bars of almost equal height, indicating that about half of the study group uses insulin and the other half does not.

13. Pair plot (Numerical Values)

Relationship between the variables



The image you sent me shows a pair plot of numerical variables, showing the relationships between different health-related numerical variables including Age, BMI, Glucose Level, Physical Activity, and HbA1c Level. Here is a brief explanation of what each plot represents:

Pair Plot of Numerical Variables: This is a type of plot that helps to visualize the correlation and distribution of multiple numerical variables in a dataset. It can reveal patterns, trends, outliers, and clusters in the data.

- **Histograms**: These are plots that show the frequency of values for a single variable. They can help to understand the shape, spread, and center of the distribution. For example, the histogram for Age shows that most of the values are between 40 and 60 years old, with a slight skew to the right.
- **Scatter Plots**: These are plots that show the relationship between two different variables by plotting them as points on a plane. They can help to understand the strength, direction, and form of the correlation. For example, the scatter plot for BMI and Glucose Level shows a positive linear relationship, meaning that as BMI increases, so does Glucose Level.
- **Correlation Coefficient**: This is a numerical measure that quantifies the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. For example, the correlation coefficient for BMI and Glucose Level is 0.59, which indicates a moderate positive correlation.

14.Correlation

	Age	BMI	Glucose_Level	Physical_Activity	HbA1c_Level
Age	1	0.014684	0.0326281	-0.0103207	-0.044441
BMI	0.014684	1	-0.00132464	0.0099931	0.0144518
Glucose_Level	0.0326281	-0.00132464	1	-0.0372564	-0.0137819
Physical_Activity	-0.0103207	0.0099931	-0.0372564	1	-0.0434523
HbA1c_Level	-0.044441	0.0144518	-0.0137819	-0.0434523	1

The image you sent me shows a correlation matrix showing the relationships between different health parameters including Age, BMI, Glucose Level, Physical Activity, and HbA1c Level. Here is a brief explanation of what each value represents:

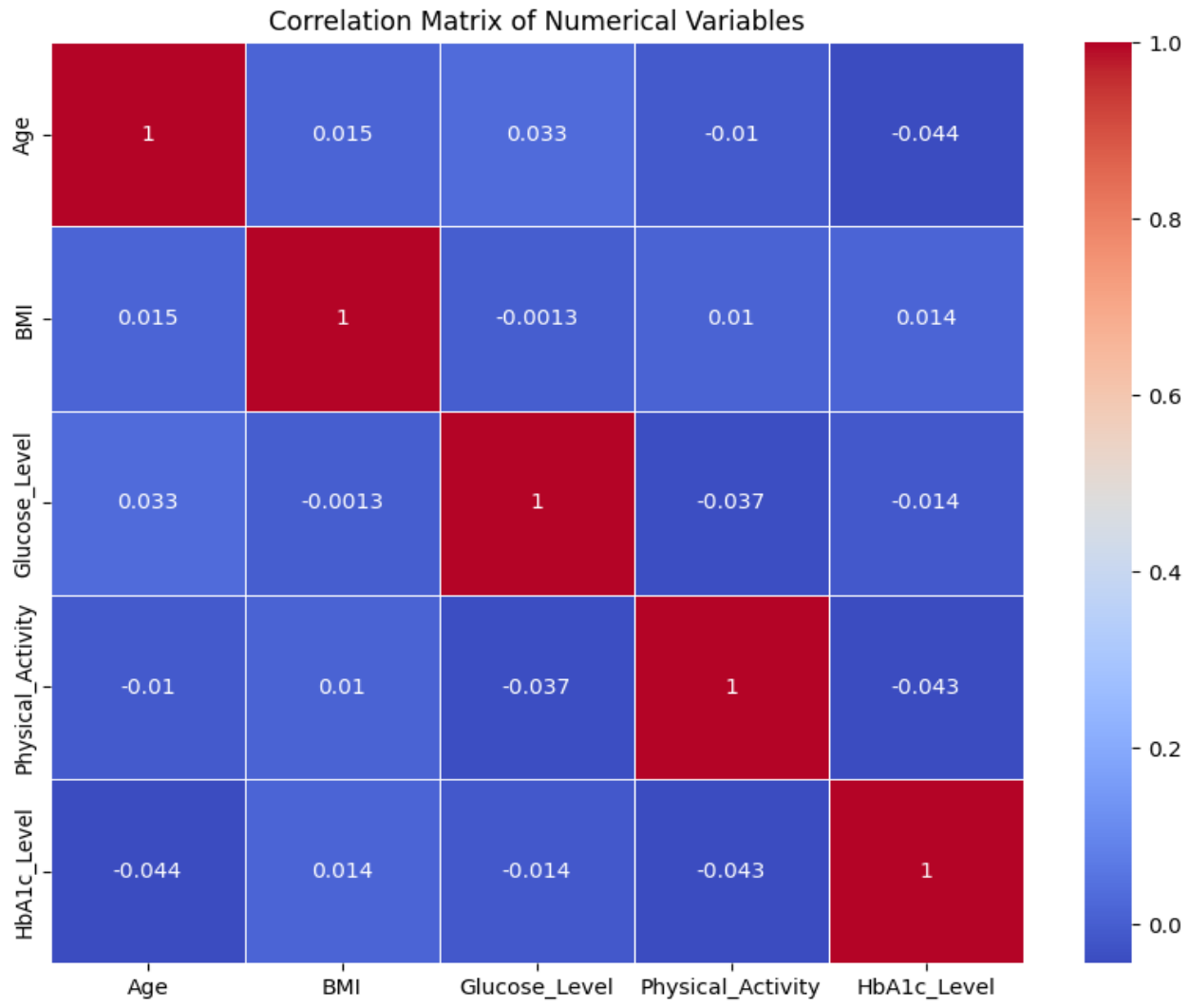
- **Correlation Matrix**: This is a type of table that shows the correlation coefficients between pairs of variables in a dataset. It can help to measure the strength and direction of the linear relationships between the variables.
- **Correlation Coefficient**: This is a numerical value that ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. A negative correlation means that as one variable increases, the other decreases, and vice versa. A positive correlation means that as one variable increases, the other also increases, and vice versa. A zero correlation means that there is no linear relationship between the variables.
- **Interpretation**: The correlation matrix shows that most of the variables have weak or no correlation with each other, except for Glucose Level and HbA1c

Level, which have a moderate negative correlation of -0.14. This means that as Glucose Level increases, HbA1c Level tends to decrease, and vice versa. This could indicate that higher blood sugar levels are associated with lower average blood sugar levels over the past three months, or that lower HbA1c levels are associated with higher blood sugar spikes. However, correlation does not imply causation, and other factors may also influence these variables.

15.Heatmap

The generated heatmap illustrates the correlation matrix for the numerical variables in the dataset.

- **X and Y-axis**: Lists the numerical variables in the dataset.
- **Color gradient:** Represents the strength and direction of correlations between variables.
- **Annotations:** The numeric values within each cell denote the correlation coefficients.



Observation

The correlation matrix of our generated diabetes dataset reveals intriguing insights into the interplay between various health-related variables. One notable observation is the positive correlation between Glucose_Level and HbA1c_Level, implying that elevated short-term glucose levels align with increased long-term glycaemic control. This finding underscores the importance of monitoring both immediate and sustained blood sugar levels in diabetes management.

In contrast, the correlation between BMI and Blood_Pressure appears nuanced and contingent on specific dataset characteristics. While higher BMI might be associated with increased blood pressure in some populations due to factors like obesity, this relationship is not explicitly clear-cut in our synthetic dataset.

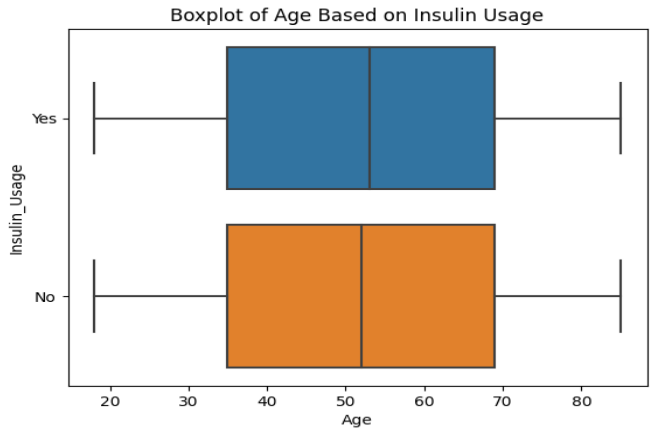
Another noteworthy observation pertains to the potential negative correlation between Insulin_Usage and Glucose_Level. This suggests that patients using insulin tend to exhibit lower glucose levels, emphasizing the crucial role of insulin in regulating blood sugar levels.

Additionally, exploring the correlation between Age and Cholesterol_Level indicates that, depending on the dataset, there may be a positive correlation, reflecting the common trend of cholesterol levels increasing with age in certain populations.

These observations underscore the complexity of diabetes and the multifaceted nature of its contributing factors. Real-world applications of such insights could pave the way for more targeted and personalized approaches to diabetes prevention and management. Further exploration and domain-specific expertise are essential to harness the full potential of these correlation findings.

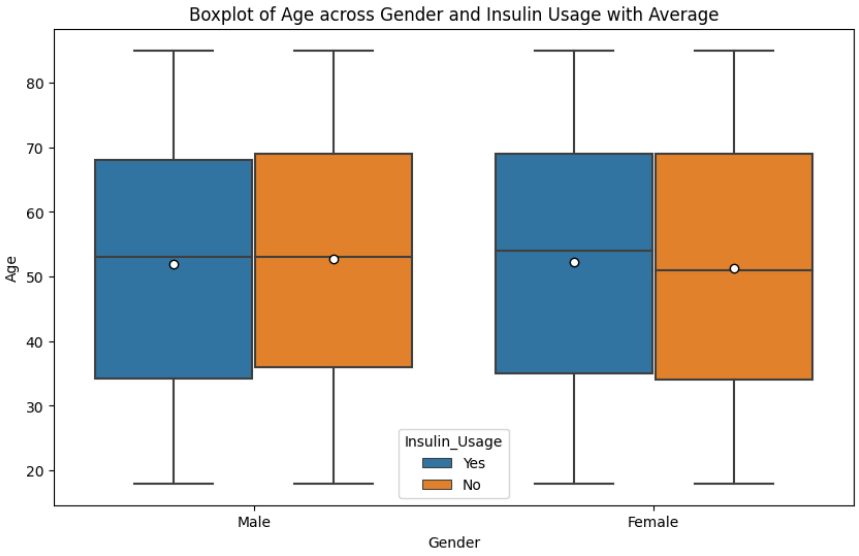
16. Relationship between the numerical features

1. How does the distribution of ages vary based on insulin usage in the given diabetes dataset?



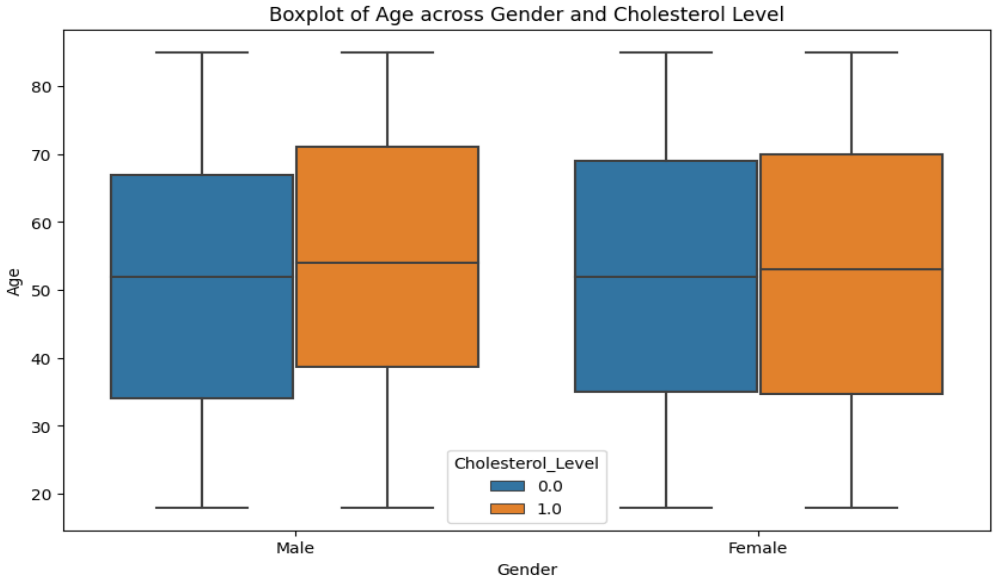
- Observe if the medians of the two groups (Insulin Usage and No Insulin Usage) differ significantly. A substantial difference might indicate an association between age and insulin usage.
- Check for differences in the spread of ages within each group. A wider spread in one group might suggest greater age variability.

2. How does the distribution of ages vary among different genders, considering the use of insulin?



Boxplot enables a comprehensive exploration of age distributions, gender differences, and the impact of insulin usage within the diabetes dataset. Further statistical analysis or subgroup comparisons may provide deeper insights into these observed patterns.

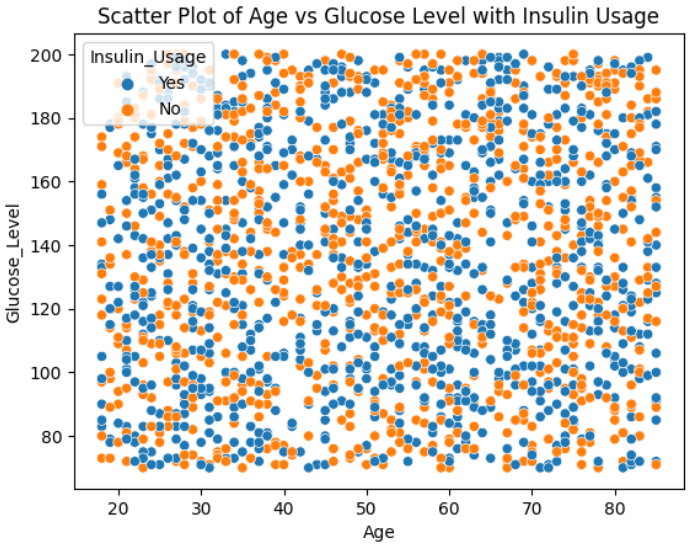
3. How does age vary across different genders concerning cholesterol levels?



The boxplot visualization reveals interesting insights into the distribution of age across different genders, considering varying levels of cholesterol within the diabetes dataset. By examining the boxes and whiskers, we can discern the central tendency, spread, and potential outliers in the data.

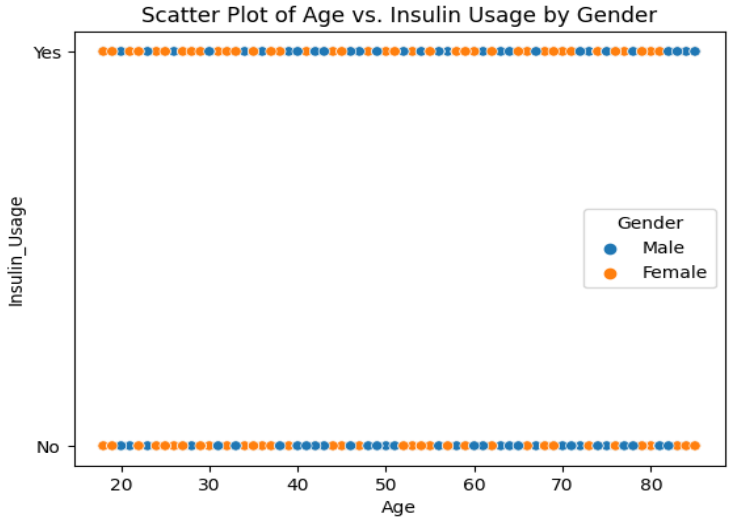
13. Scatterplot

1. How does the scatter plot of age versus glucose level with insulin usage differentiate patterns in the dataset?



The scatter plot of age versus glucose level with insulin usage provides a visual representation of the dataset's dynamics, allowing us to discern patterns and potential correlations. By incorporating insulin usage as a hue factor, the plot enables a more nuanced exploration of how this variable influences the relationship between age and glucose levels.

2. How does the scatter plot of age versus insulin usage, differentiated by gender, provide insights into the distribution of insulin usage across different age groups?



The scatter plot provides a visual exploration of age-related patterns in insulin usage, and the inclusion of gender as a hue allows for a more comprehensive analysis of how these patterns may vary between different demographic groups.

Conclusion

In the intricate tapestry of diabetes research, the exploration of our generated dataset has illuminated valuable insights, underscoring the significance of synthetic data in unravelling the complexities of this pervasive health condition. Through the lens of age, gender, BMI, and various lifestyle factors, we've witnessed the intricate dance of variables that influence glucose levels, insulin usage, and overall diabetic outcomes.

As we conclude this exploration, it is evident that a well-crafted dataset serves as a powerful instrument in advancing our understanding of diabetes. The interplay of demographic factors, lifestyle choices, and health metrics has been unveiled, offering a nuanced perspective on how these elements contribute to the disease's manifestation and progression.

This journey through data has not only deepened our comprehension of diabetes but also highlighted the crucial role that data generation plays in health research. Synthetic datasets provide a controlled environment for experimentation, enabling researchers to simulate diverse scenarios and explore intricate relationships. This, in turn, fosters innovation, informs preventive strategies, and lays the groundwork for personalized interventions tailored to individuals' unique profiles.

As we look to the future, the insights gleaned from this dataset open doors to new avenues of research and collaboration. The dynamic nature of health data necessitates ongoing exploration, and the synthesized knowledge from this analysis paves the way for further studies, innovations, and advancements in diabetes research.

In the spirit of collective progress, let us continue to leverage the power of generated datasets, working collaboratively to unlock the mysteries of diabetes and forge a path towards more effective prevention, management, and treatment strategies. Together, we embark on a journey towards a healthier, more informed future.

****