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.

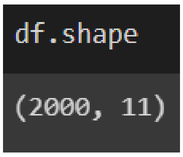
In this blog post, we delve into the exploration of a meticulously crafted dataset, encompassing variables such as age, gender, BMI, blood pressure, glucose levels, cholesterol levels, physical activity, diet type, insulin usage, and HbA1c levels. Through this analysis, we aim to unravel patterns, correlations, 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:**

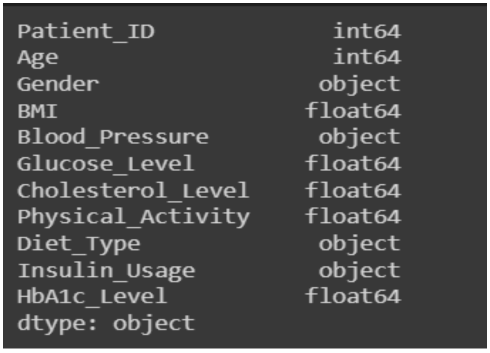
- Captures whether patients use insulin for diabetes management (Yes/No).
- Provides insights into the prevalence of insulin use within the dataset.

11. **HbA1c_Level:**

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

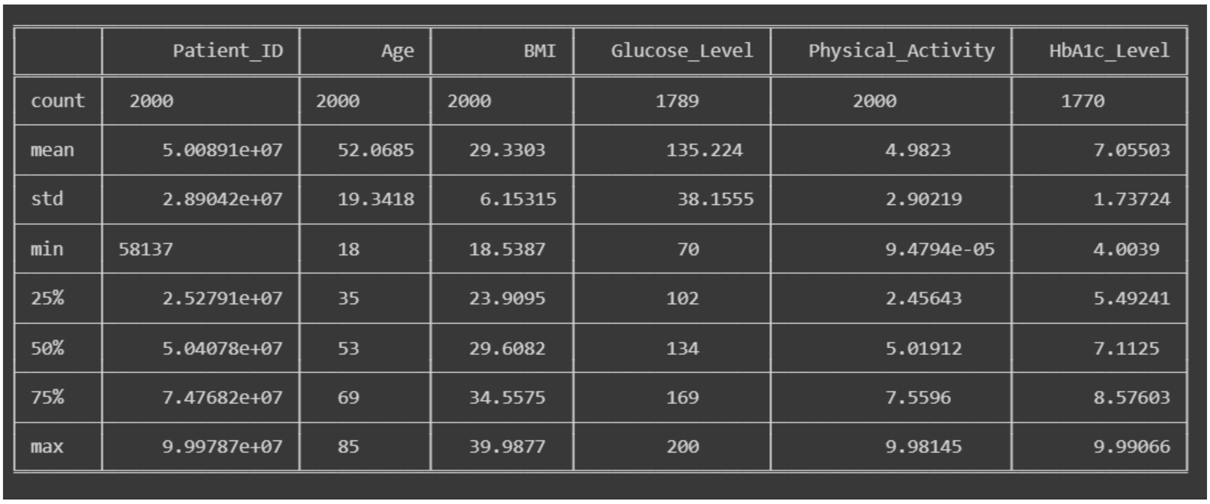
Exploratory Data Analysis (EDA)

1. **Datatypes of each column:**



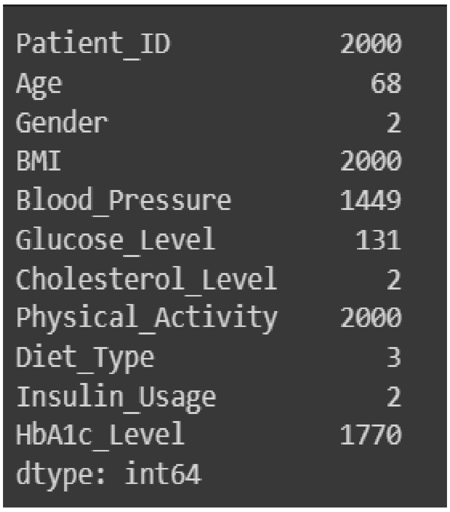
```
Patient_ID              int64
Age                     int64
Gender                 object
BMI                   float64
Blood_Pressure         object
Glucose_Level         float64
Cholesterol_Level     float64
Physical_Activity     float64
Diet_Type              object
Insulin_Usage          object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

2. Statistics Description of the dataset:



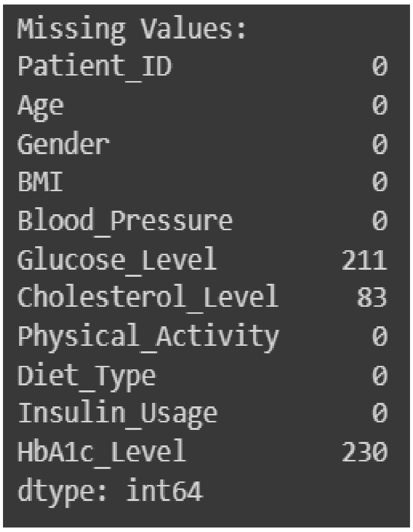
	Patient_ID	Age	BMI	Glucose_Level	Physical_Activity	HbA1c_Level
count	2000	2000	2000	1789	2000	1770
mean	5.00891e+07	52.0685	29.3303	135.224	4.9823	7.05503
std	2.89042e+07	19.3418	6.15315	38.1555	2.90219	1.73724
min	58137	18	18.5387	70	9.4794e-05	4.0039
25%	2.52791e+07	35	23.9095	102	2.45643	5.49241
50%	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

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
```

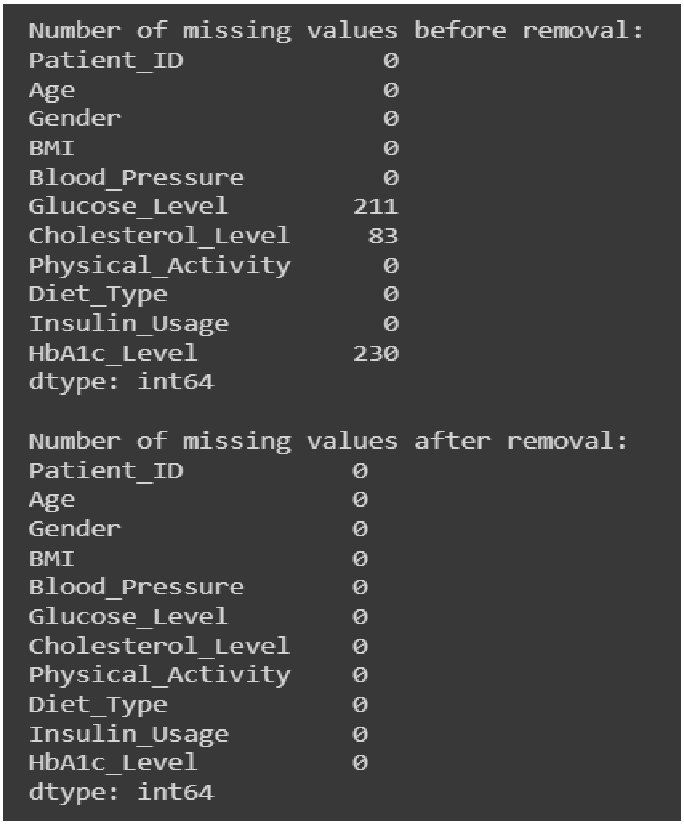
4. Check the Missing Values



```
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
```

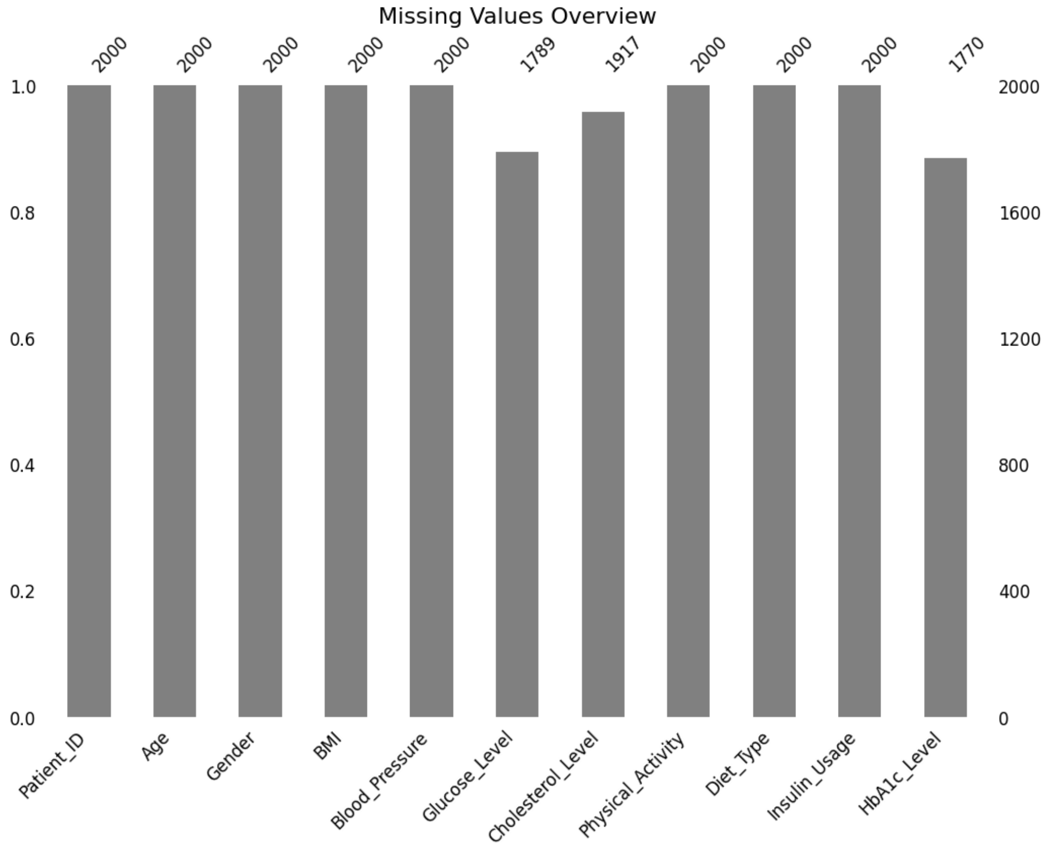
Missing Values are detected in the dataset

Removing the Missing Values



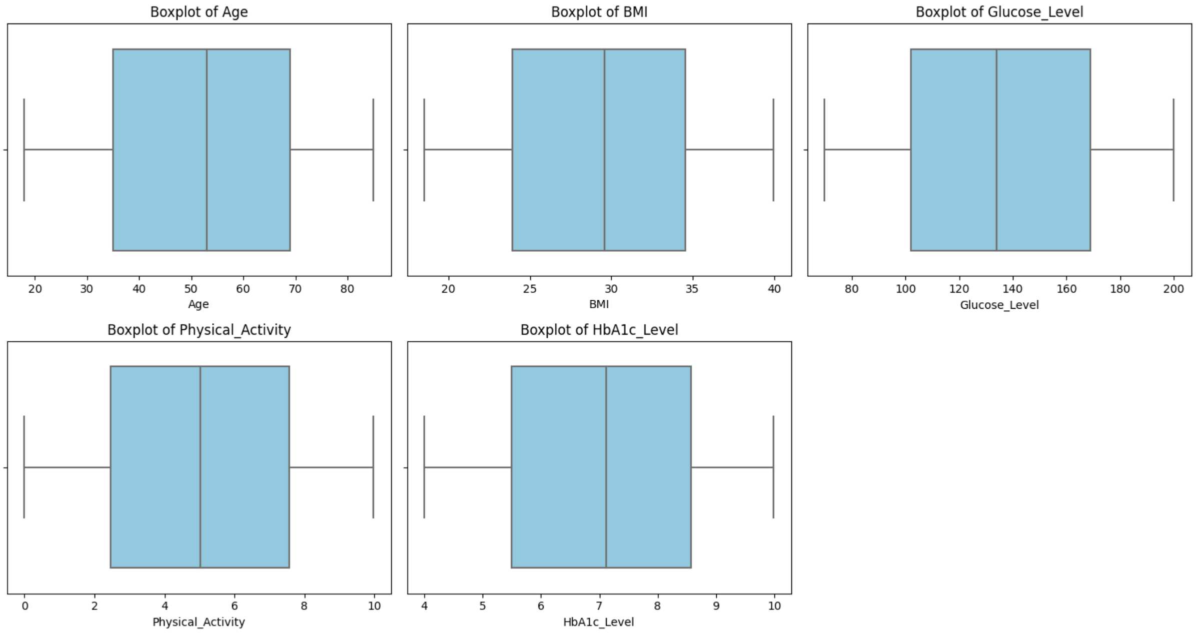
```
Number of missing values before removal:
Patient_ID              0
Age                     0
Gender                  0
BMI                     0
Blood_Pressure          0
Glucose_Level         211
Cholesterol_Level      83
Physical_Activity       0
Diet_Type               0
Insulin_Usage        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

5. Missing Values Graph



6. Check for Outliers

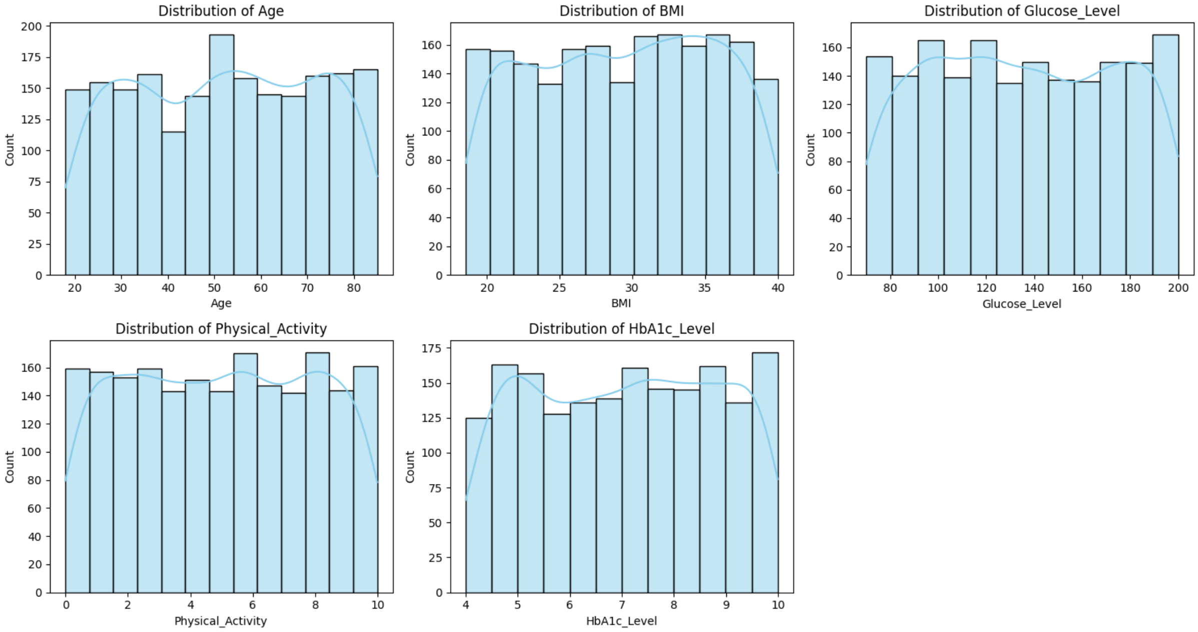
As missing are detected in the dataset, we can check for the outliers by using box plot graph.



No outliers are detected in the dataset.

7. Histogram Plot – Numerical Values

Numerical Values distributed in the dataset

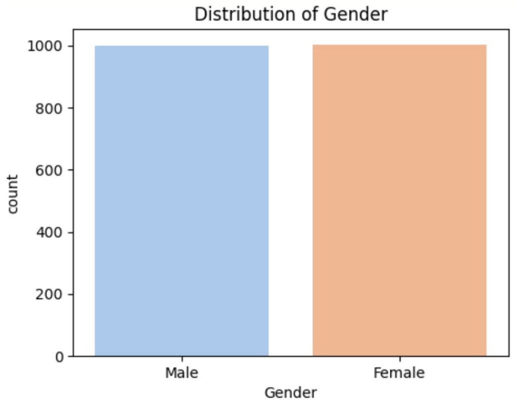


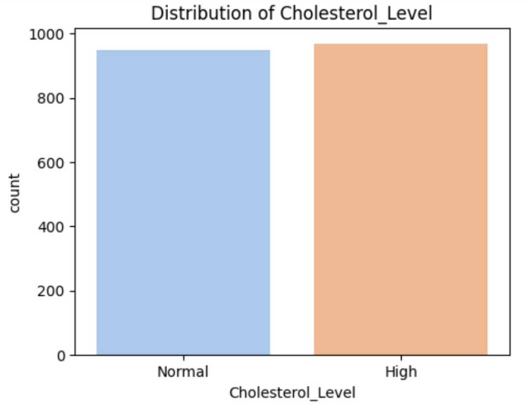
Insights from the graph:

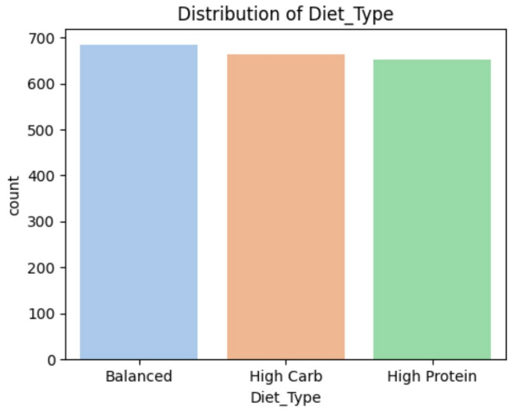
Each numerical columns (Age, BMI, Glucose Level, Physical activity and HbA1c Level) are almost equally distributed for each interval.

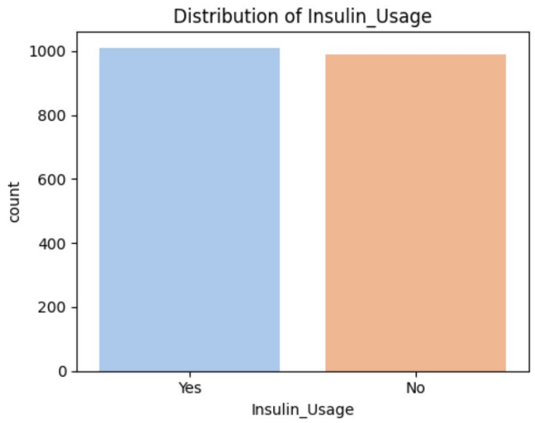
8. Categorical Values

Categorical Values distributed in the dataset









- **Gender:** There is no significant difference between the genders in terms of diabetes risk or management.

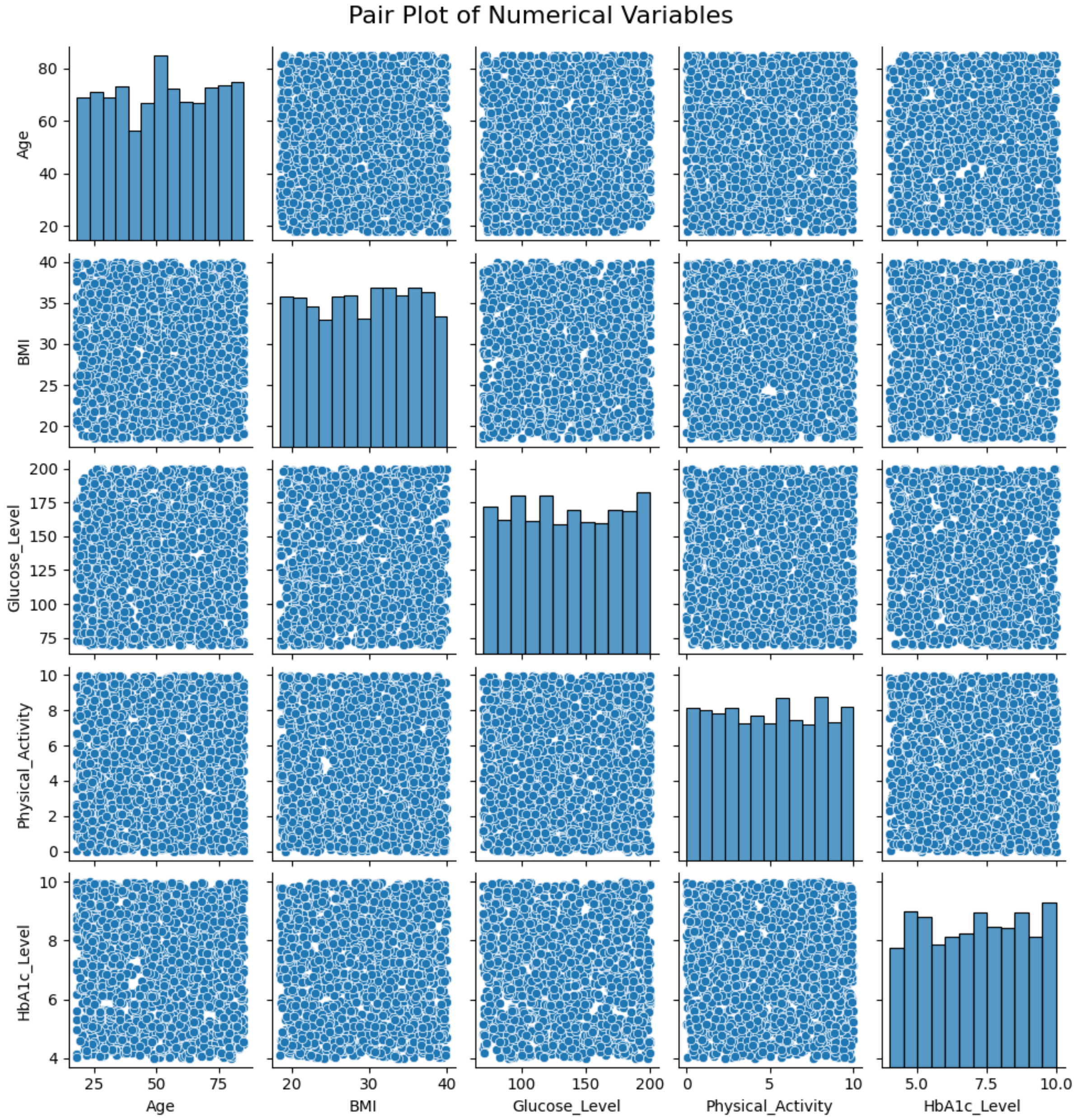
- **Cholesterol:** Cholesterol level is not a strong predictor of diabetes, as both normal and high levels are equally prevalent among the patients.

- **Diet:** Diet type may have some influence on diabetes, as high protein or high carb diets are more common than balanced diets. This could indicate that some patients are following specific dietary plans to control their blood sugar levels or prevent complications.

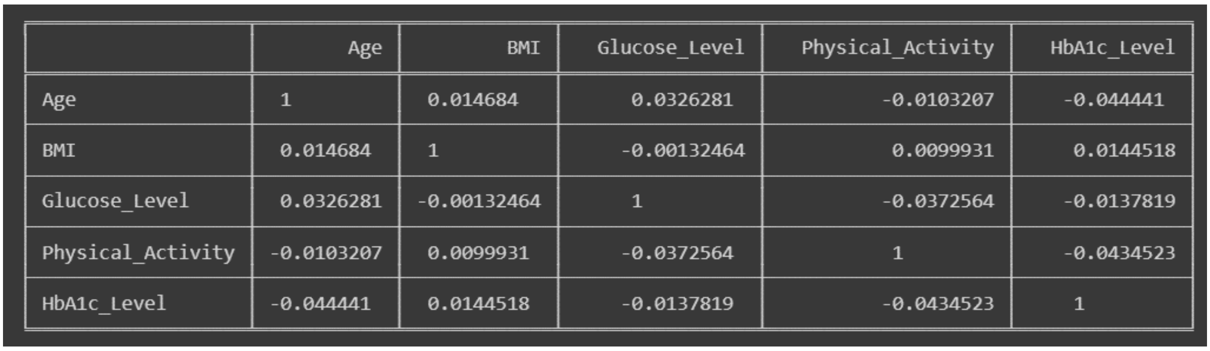
- **Insulin usage:** Insulin usage is not a clear indicator of diabetes severity, as both users and non-users are equally represented in the dataset. This could suggest that some patients can manage their diabetes with oral medications or lifestyle changes, while others require insulin injections.

9. Pair plot (Numerical Values)

Relationship between the variables



10. Correlation

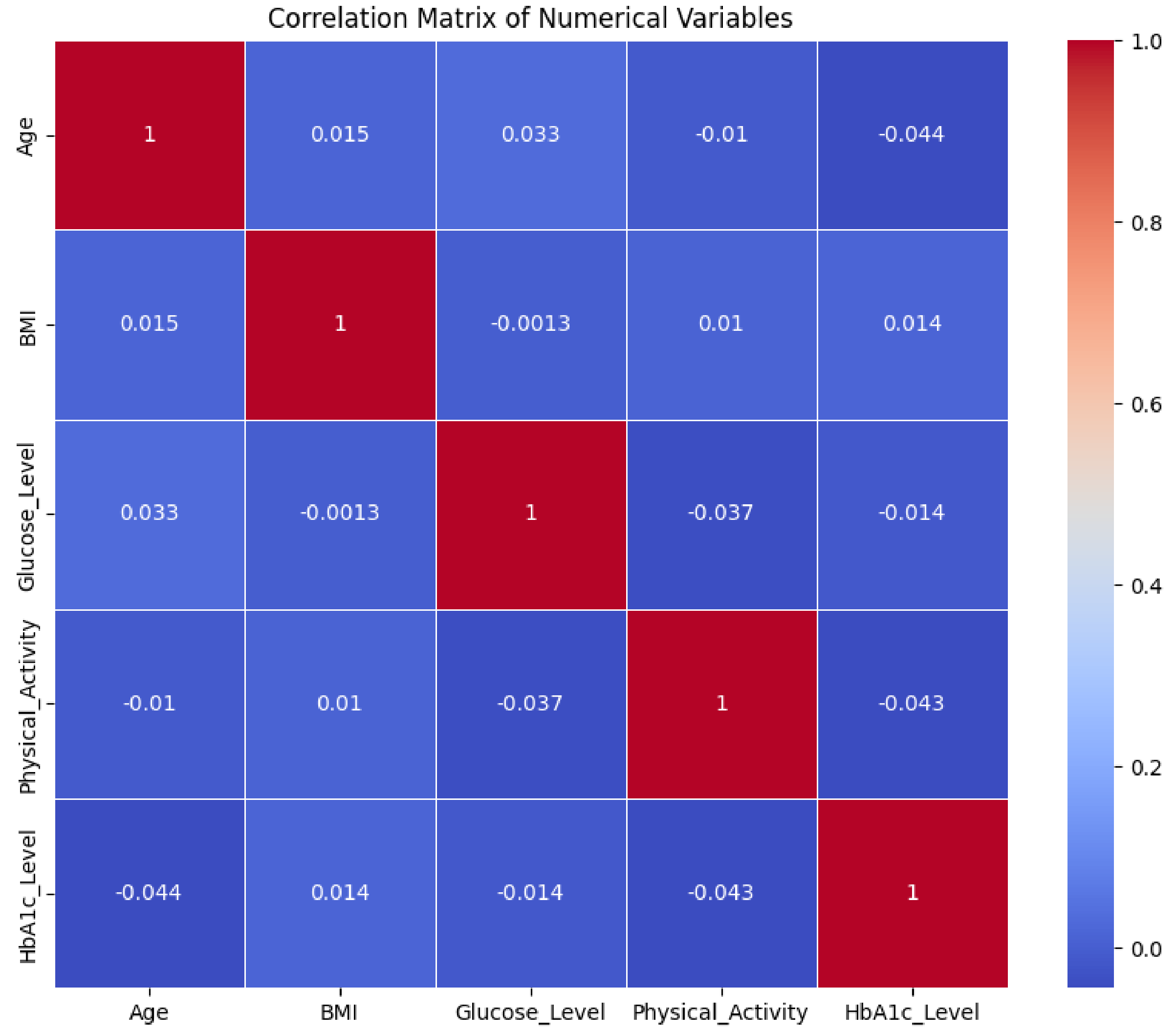


	Age	BMI	Glucose_Level	Physical_Activity	HbA1c_Level
Age	1	0.014684	0.0326281	-0.0103207	-0.044441
BMI	0.014684	1	-0.00132464	0.0099931	0.0144518
Glucose_Level	0.0326281	-0.00132464	1	-0.0372564	-0.0137819
Physical_Activity	-0.0103207	0.0099931	-0.0372564	1	-0.0434523
HbA1c_Level	-0.044441	0.0144518	-0.0137819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

11.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.

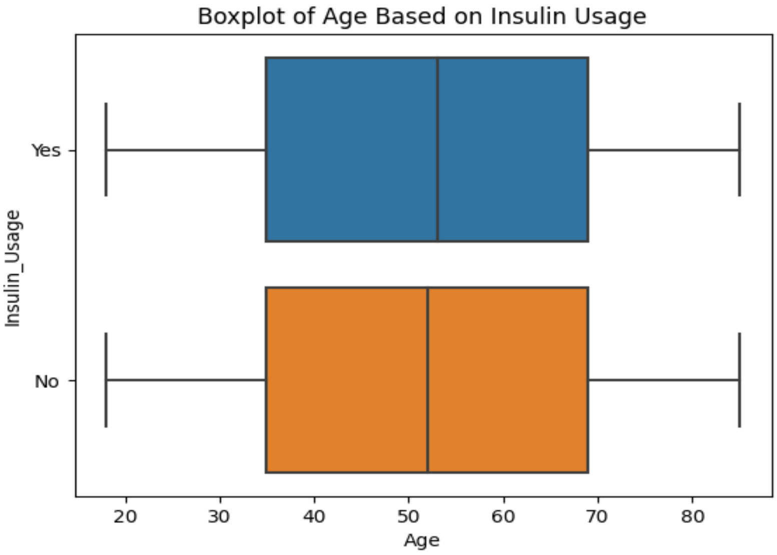
Another noteworthy observation pertains to the potential negative correlation between Insulin_Usage and Glucose_Level. This suggests that patients using insulin tend to exhibit lower glucose levels, emphasizing the crucial role of insulin in regulating blood sugar levels.

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.

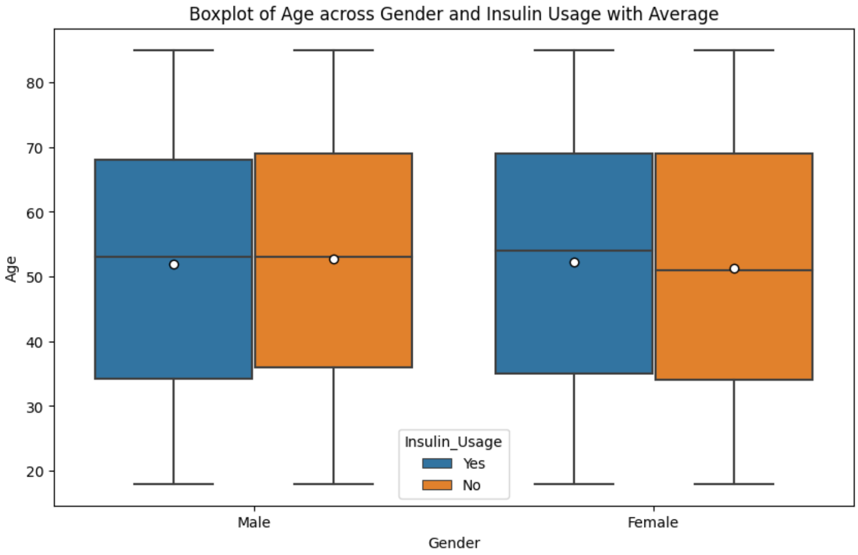
12.**Relationship between the numerical features**

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



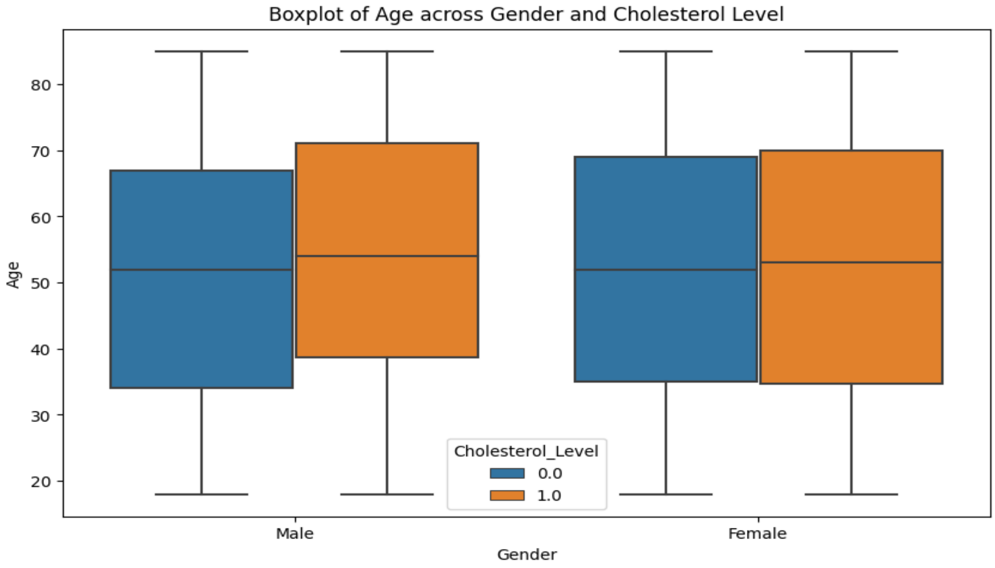
- Observe if the medians of the two groups (Insulin Usage and No Insulin Usage) differ significantly. A substantial difference might indicate an association between age and insulin usage.
- Check for differences in the spread of ages within each group. A wider spread in one group might suggest greater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

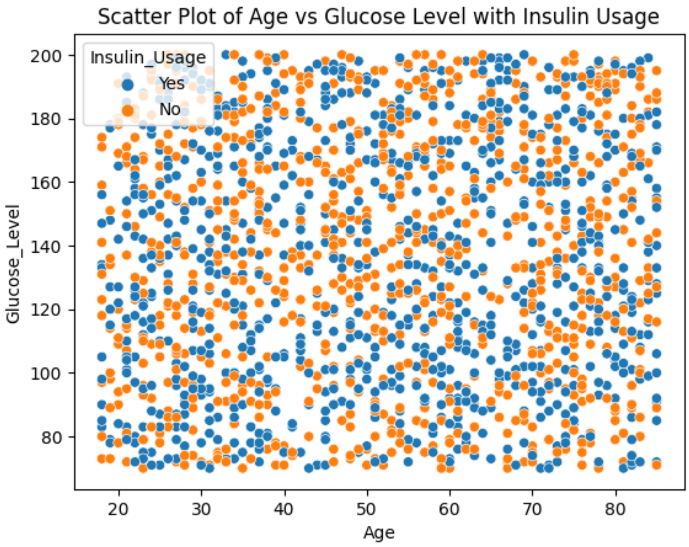
3. How does age vary across different genders concerning cholesterol levels?



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.

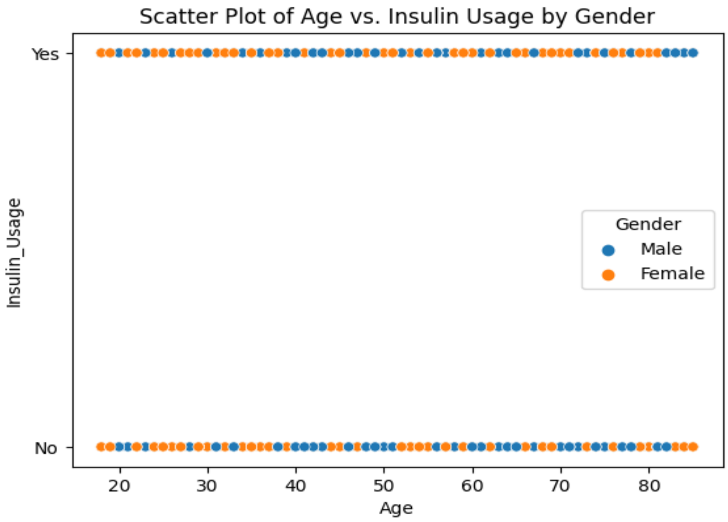
13. Scatterplot

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.