Decoding Health Patterns: Unveiling Predictive Insights through EDA

Introduction

In the changing world of healthcare, using data analysis to understand patient well-being is crucial for shaping personalized wellness. The initial phase of data analysis is critical because it not only uncovers hidden insights but also lays the groundwork for subsequent analyses that guide future analyses. Exploratory Data Analysis (EDA) is a method that aims to reveal important patterns in a dataset that might be easily overlooked.

This blog, "Decoding Health Patterns: Revealing Predictive Insights with EDA," explores the concept of Exploratory Data Analysis. It shows how this approach can uncover important patterns and predictive signs in healthcare datasets.

Dataset Generation

In crafting a comprehensive understanding of healthcare dynamics through Exploratory Data Analysis (EDA), the foundation lies in the dataset generation process. The dataset for our analysis has been meticulously generated using the Faker library, ensuring authenticity and diversity in its composition. This synthetic dataset encompasses 1500 rows, each meticulously designed to simulate a diverse set of healthcare scenarios. It includes a combination of categorical and numerical attributes, intentionally selected to mirror the diverse facets inherent in patient profiles.

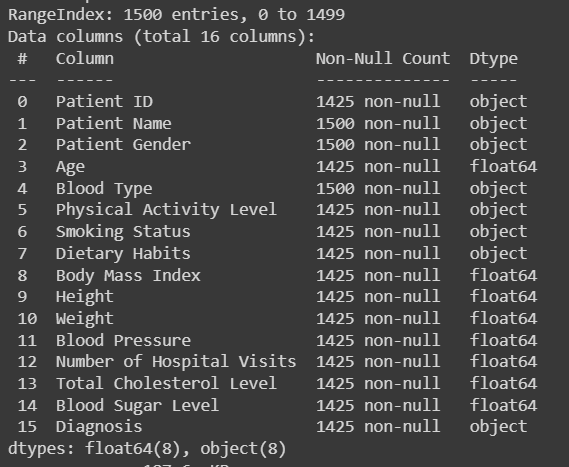
Attributes:

* Patient ID (unique identifier)
* Patient name
* Patient Gender (Male, Female, Other)
* Blood Type (A, B, AB, O)
* Physical Activity Level (Sedentary, Low Activity, Moderate Activity, Active, Very Active)
* Smoking Status (Smoker, Non-Smoker, Former Smoker)
* Dietary Habits (Vegetarian, Mediterranean, Balanced, High Protein, Low Carb)
* Age (in years)
* Body Mass Index (BMI)
* Height (in cm)
* Weight (in kilograms)
* Blood Pressure (mmHg)
* Number of Hospital Visits (integer)
* Total Cholesterol Level (mg/dL)
* Blood Sugar Level (mg/dL)
* Diagnosis (e.g., Healthy, Chronic Disease, Acute Condition)

The addition of categorical and numerical attributes, such as Patient Gender, Blood Type, and Body Mass Index, enriches our dataset, providing a multifaceted perspective on the complexities of healthcare dynamics.

Missing Value Analysis

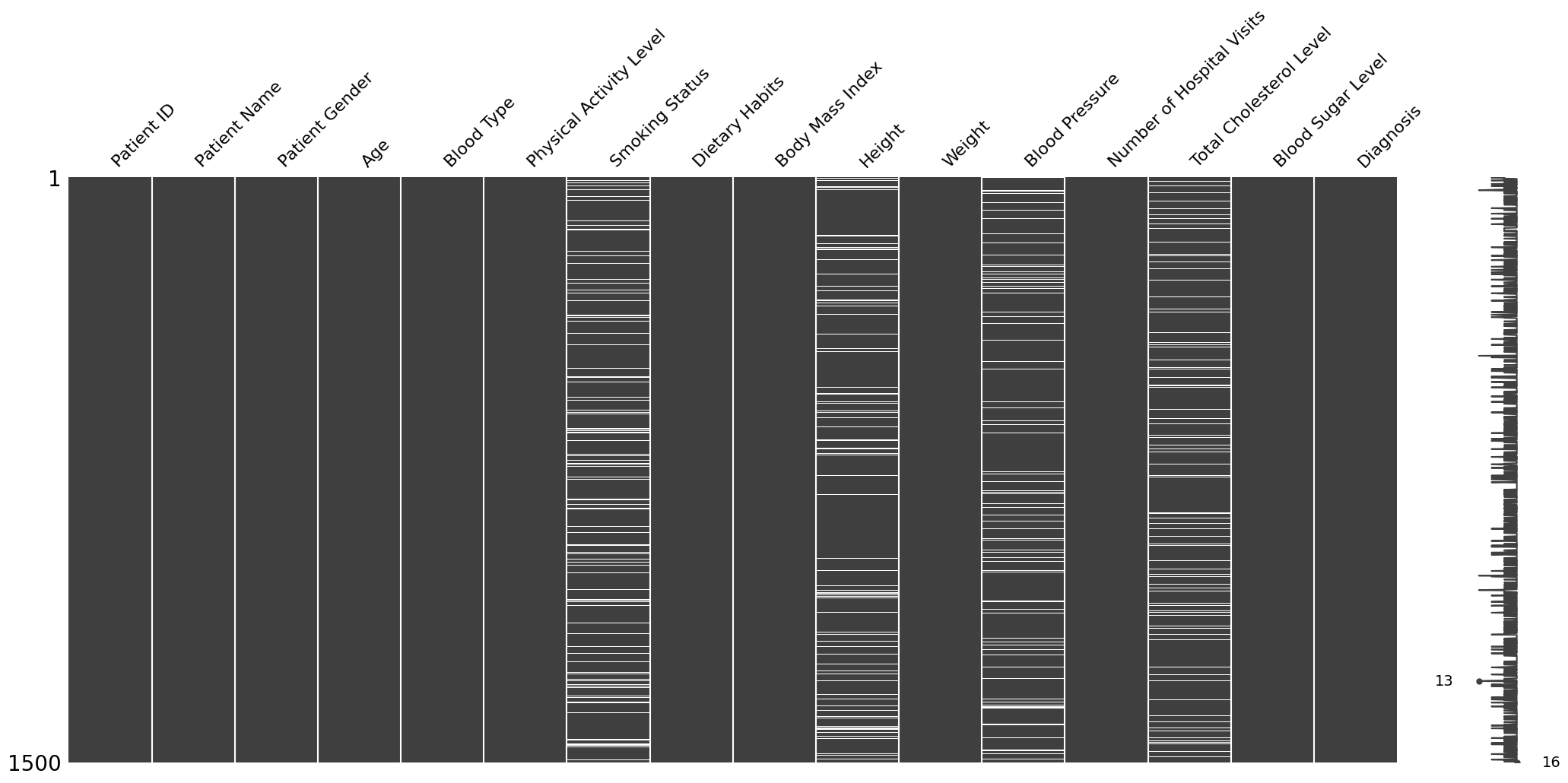
The presence of missing values can significantly impact the reliability and accuracy of analyses. Detecting and addressing these missing entries is a crucial step in ensuring the integrity of our exploratory data analysis (EDA).



From the above picture we can identify that some attributes have missing values. To illuminate the path forward, we employ the missingno library, a valuable tool for visualizing and comprehending the distribution of missing values within our synthetic healthcare dataset.

Visualizing Missing Values

The matrix elegantly represents missing values as white lines, allowing us to discern patterns and concentrations of missingness across various attributes. Each row corresponds to an entry in the dataset, while columns depict different attributes.

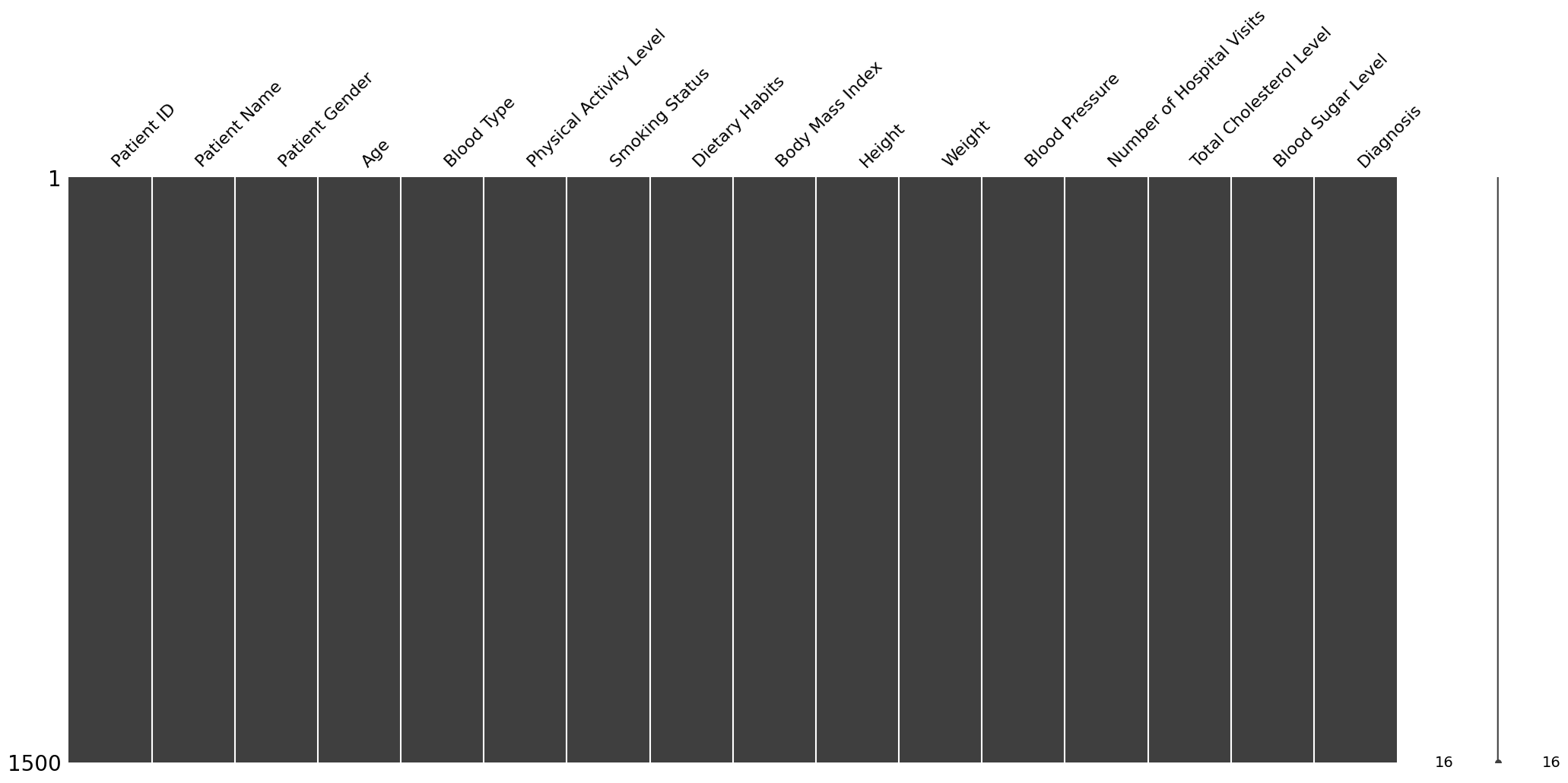


Observing the matrix, we notice that missing values are not uniformly distributed across all attributes. Certain columns, such as 'Height,' 'Blood Pressure,' 'Smoking Status,' and 'Total Cholesterol Level,' exhibit a higher rate of absences. The reasons for missing data can be as diverse as the dataset itself. Individuals might be unwilling to disclose sensitive information such as 'Height,' 'Smoking Status,' or other health-related attributes.

Handling Missing Data

Various techniques come into play, each offering a unique solution to address the gaps in our dataset. Imputation techniques, such as mean or median substitution, provide avenues to fill gaps without compromising dataset integrity. Alternatively, excluding entries with missing values offers a straightforward approach.

Using the scikit-learn library's SimpleImputer, missing values were effectively handled for both numerical and categorical attributes. For numerical columns such as 'Height,' 'Blood Pressure,' and others, the missing values were imputed with the mean, ensuring a representative replacement. Concurrently, categorical attributes like 'Smoking Status' underwent imputation with the most frequent values, preserving the inherent characteristics of the dataset.



Our healthcare dataset now has no blank spaces, representing a thorough treatment of missing values across both numerical and categorical attributes.

Duplicate Data

Duplicate data can adversely affect model accuracy by introducing bias. However, in our Faker-generated healthcare dataset, each entry is associated with a unique set of attributes related to patients' health profiles. No duplications exist in this dataset, reinforcing its reliability. To check for duplicates, the duplicated() method in pandas can be employed.

Outliers

Before diving into the analysis of this synthetically created dataset, it's essential to ensure the absence of outliers. While this dataset may appear to be free of outliers, it is crucial to acknowledge that in many real-world scenarios, handling outliers becomes essential as they can significantly influence the integrity of the data analysis. Outlier handling techniques include 1) Z-score method, 2) Interquartile Range (IQR) method.

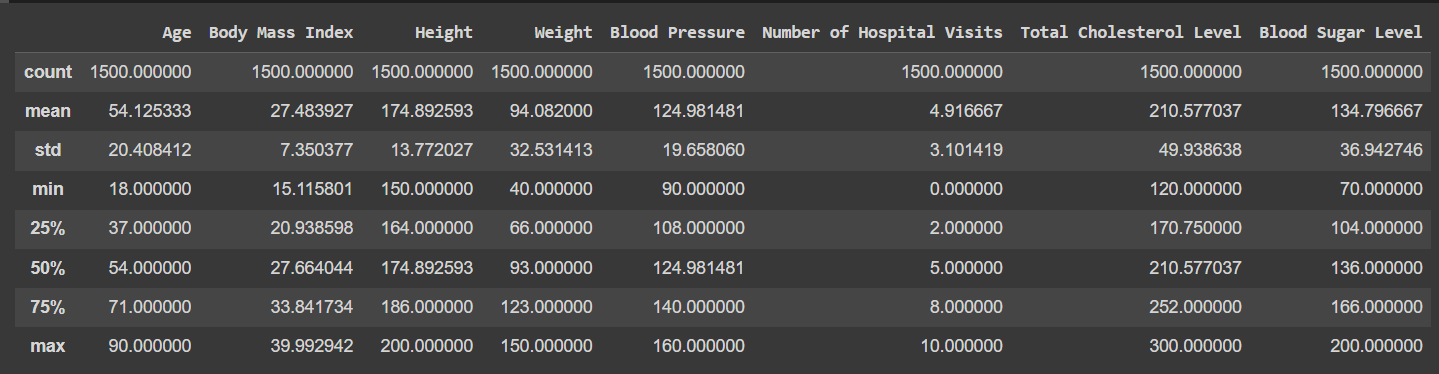
Noisy Data

Unwanted noise can interfere with a clear signal, and noisy data can interfere with the clarity of health data. Each entry, a unique combination of health attributes, adds to the dataset's diversity. Techniques such as sampling and binning can be used to remove noisy data.

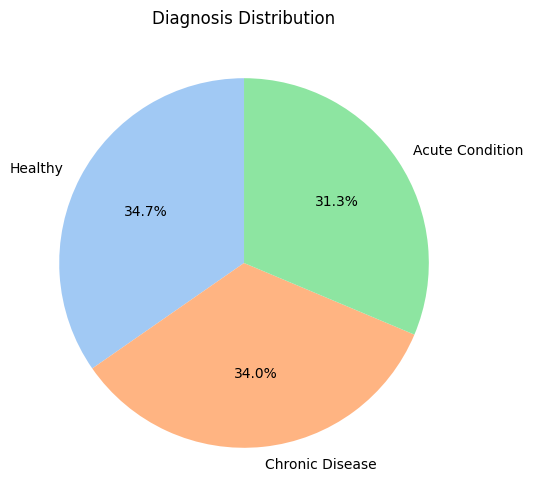
Visualizations

After dealing with missing values, duplicates, and noise, our dataset is ready for visualization. Let's embark on a visual journey through our healthcare dataset, uncover meaningful insights and patterns that lie within the data.

To embark on our exploratory data analysis journey using Python, we first import essential libraries to unveil correlations and patterns within the dataset. Let's initiate our exploration by gaining insights into the dataset. By executing df.describe().

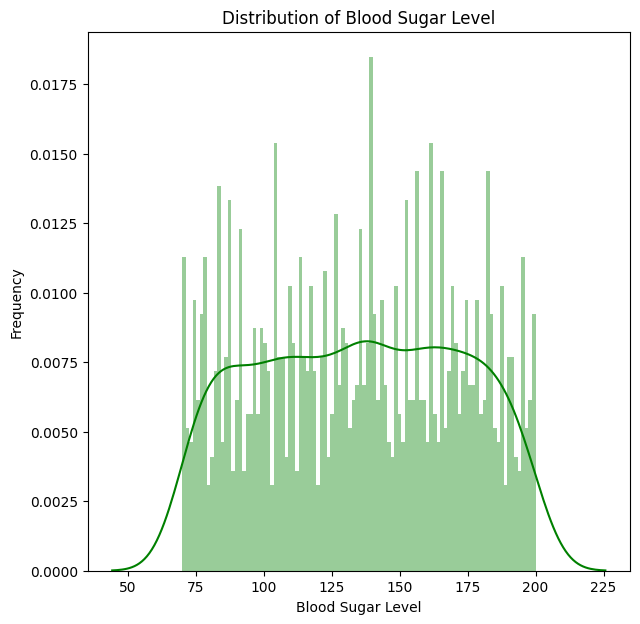


This command provides key statistical metrics such as minimum, maximum, and other details, offering a comprehensive overview and scaling of the dataset. Let's look at a visual representation of our healthcare dataset now.



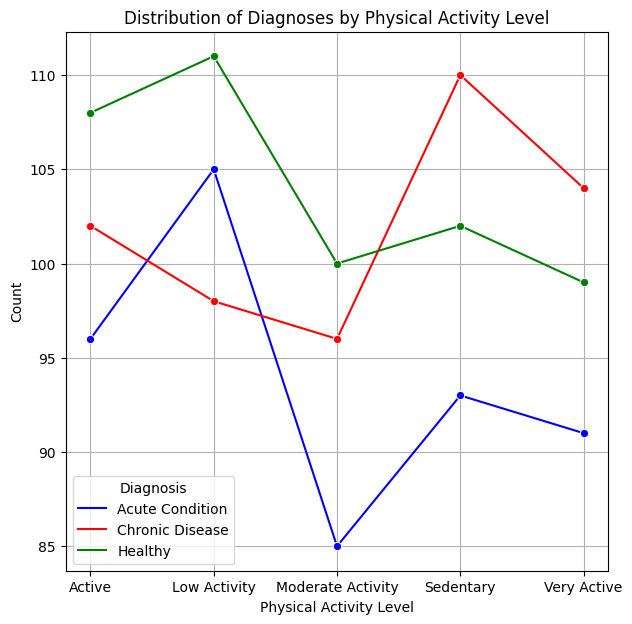
The "Diagnosis Distribution" pie chart provides a comprehensive overview of the distribution of health diagnoses within our synthetic healthcare dataset. Each slice of the pie represents a different diagnosis category. The pie chart reveals a balanced distribution of health conditions in our synthetic dataset, with 34% chronic, 34.7% healthy, and the remaining percentage representing acute conditions. Dataset covers all the possible classes equally.

Blood sugar also have high impact on patients’ health condition. Now let’s look at how Blood sugar is distributed.



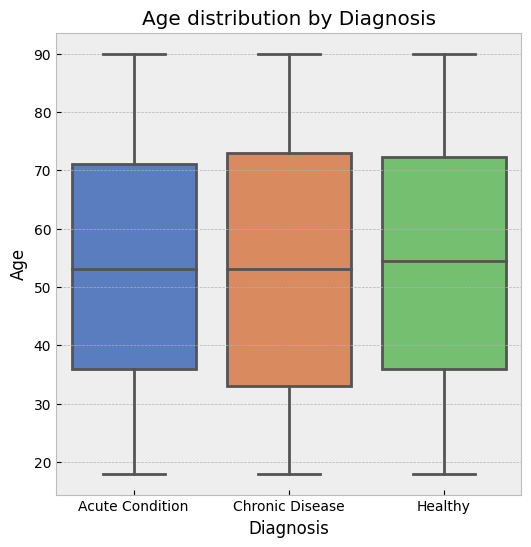
The "Blood Sugar Level Distribution" histogram provides valuable insights into the distribution of blood sugar levels in our synthetic healthcare dataset. The histogram illustrates the frequency of different blood sugar levels. The median blood sugar level (136) is higher than the mean (134.79), which is a hallmark of left-skewed distributions. The majority of the 1500 blood sugar measurements in this dataset fall above the average value, indicating a tendency towards higher levels.

Let's look at the "Distribution of Diagnoses by Physical Activity Level" line plot to see if there are any patterns and whether different diagnoses have relationships with different activity levels.

0

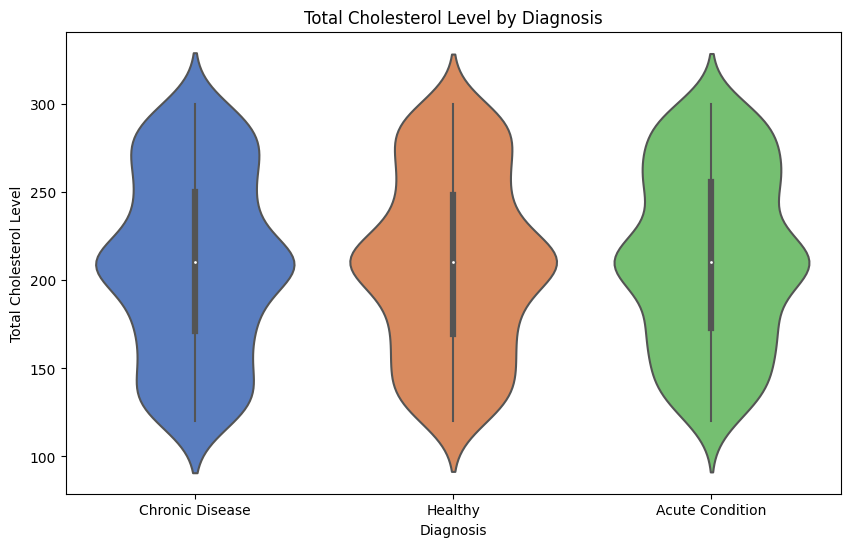
This visualization assists us in identifying any notable patterns or trends, allowing us to gain a more in-depth understanding of how physical activity may be associated with various health conditions in our synthetic healthcare dataset. From the plot we can see the count of healthy people with moderate physical activity that means people with moderate physical activity are less prone to diseases. people who are less active are more prone to chronic diseases.

Let’s look at how age and diagnosis are related

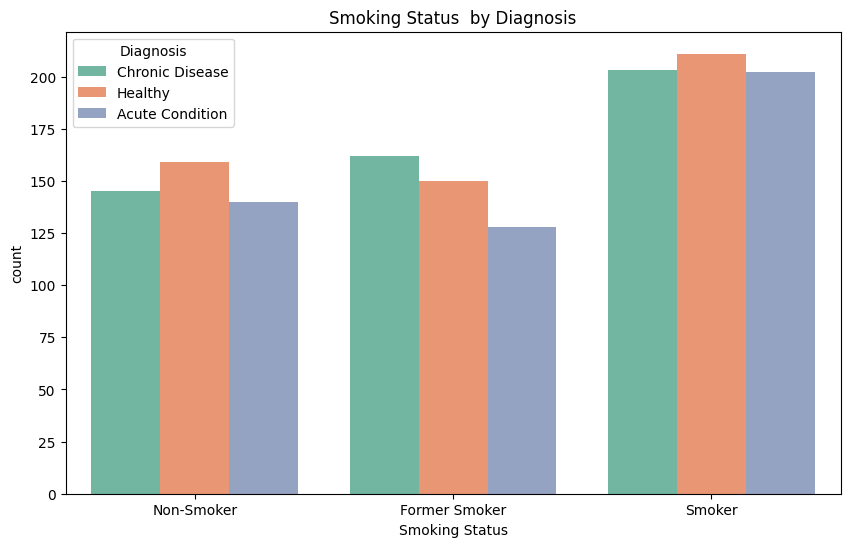


The plot consists of boxes representing the interquartile range (IQR) for each diagnosis category, with a line inside each box indicating the median age. people with high age are more prone to chronic diseases. Mostly healthy belong to the age group of 36 to 72.

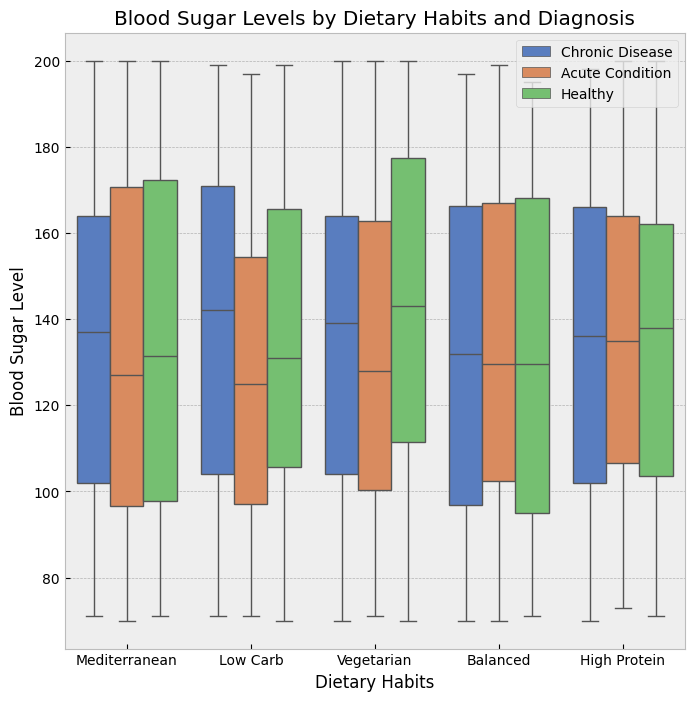
Let’s look at how total cholesterol level is related to diagnosis



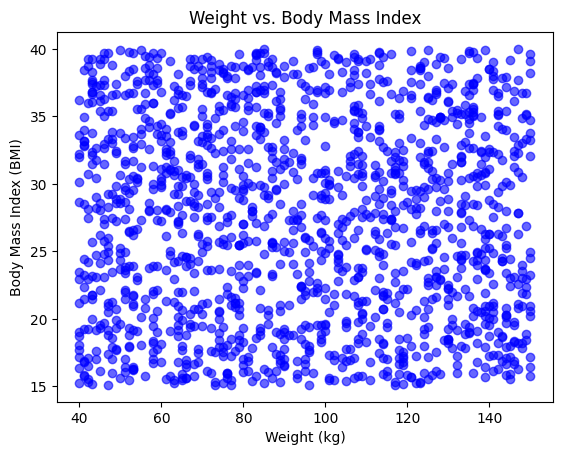
The "Total Cholesterol Level by Diagnosis" violin plot illustrates the distribution of total cholesterol levels across different health diagnoses.



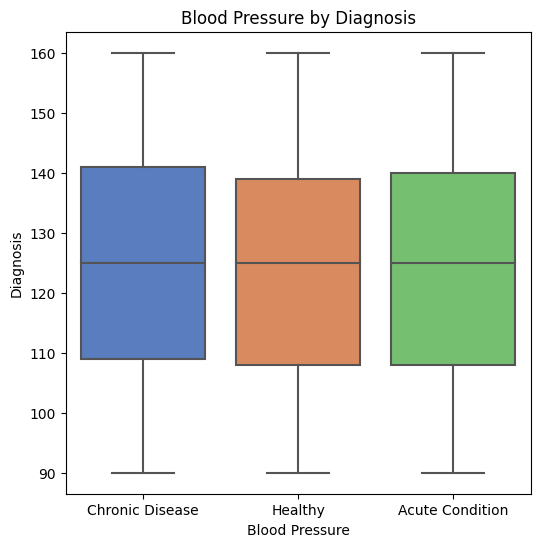
The "Smoking Status by Diagnosis" count plot shows the distribution of smoking status across different health diagnoses in our synthetic healthcare dataset, assisting in the investigation of potential links between smoking status and specific diagnoses. Each bar represents the number of people who fall into different smoking status categories, as indicated by the color-coded diagnoses. From the plot we can say that smokers have high tendency of getting a chronic disease. Non smokers are less prone to chronic diseases. Whereas former smoker getting a chronic disease is also not negligible.



From the figure, it is evident that the boxplot visualizes the distribution of blood sugar levels across various dietary habits. Notably there are no outliers in the dataset as it is synthetically generated using faker library. This dataset doesn’t have and outliers. Blood sugar levels tend to be lower for people on a Mediterranean or low-carb diet, compared to those on a high-protein diet. The median blood sugar level for people on a Mediterranean diet is around 100 mg/dL, while the median blood sugar level for people on a high-protein diet is around 120 mg/dL. The blood sugar levels for people with chronic diseases tend to be higher than the blood sugar levels for people with acute conditions or healthy people. The median blood sugar level for people with chronic diseases is around 140 mg/dL, while the median blood sugar level for people with acute conditions is around 110 mg/dL and the median blood sugar level for healthy people is around 100 mg/dL. This suggests that chronic diseases can have a significant impact on blood sugar levels.

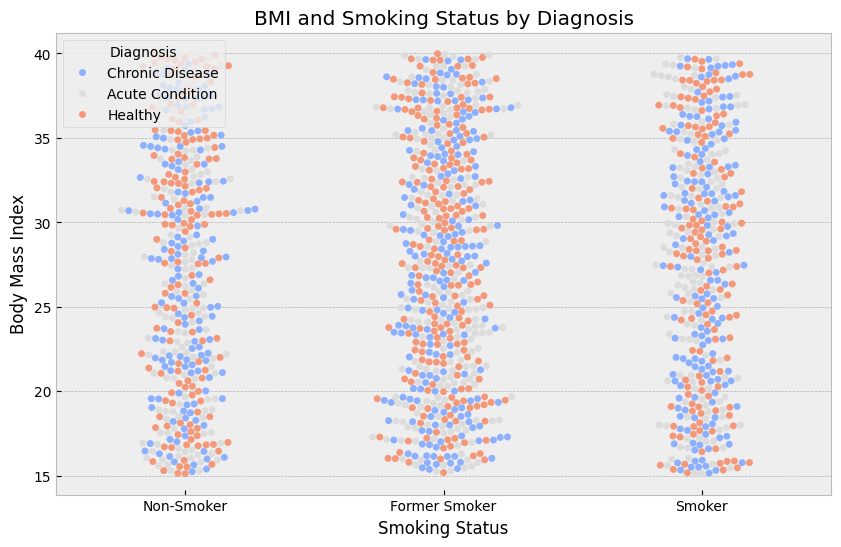


The scatter plot above represents the relationship between Weight and Body Mass Index (BMI) in our synthetic healthcare dataset. Each point on the plot corresponds to an individual entry in the dataset, with Weight (in kilograms) plotted on the x-axis and BMI on the y-axis. Observing the spread of points helps identify clusters or patterns within the dataset.



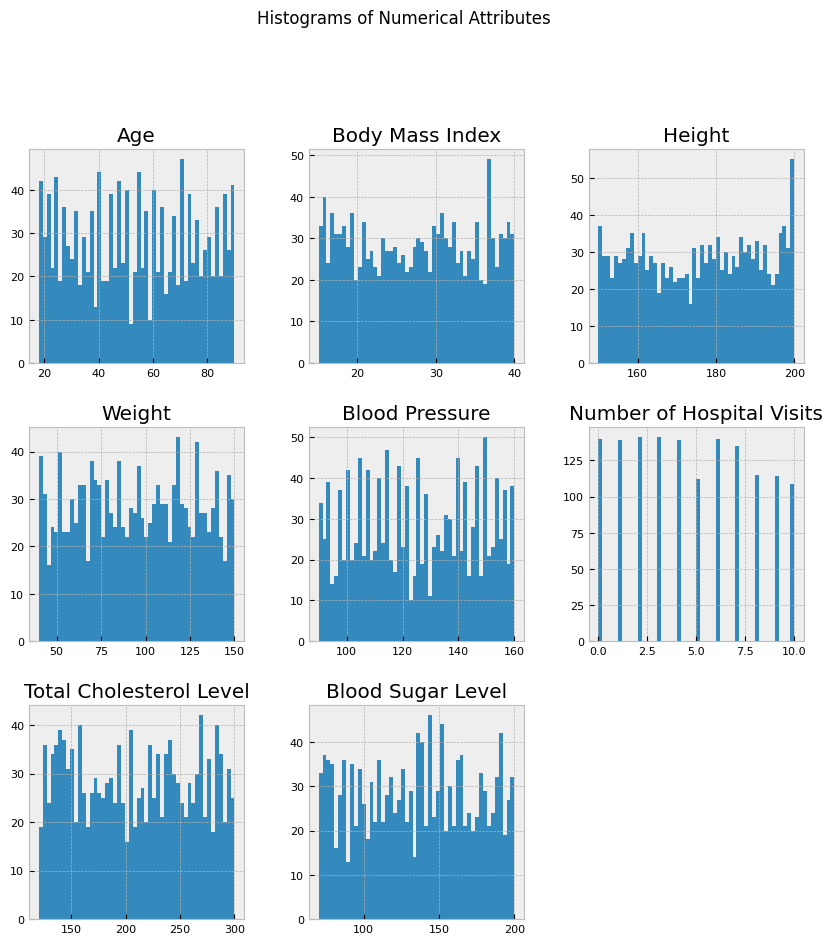
Now let’s find out the relation between Blood pressure and diagnosis with the help of boxplot which also gives info about outliers in the data. The box plot attached below visualizes the relationship between Blood Pressure (BP) levels and different health diagnoses in our synthetic healthcare dataset. The plot is categorized by Diagnosis on the x-axis, allowing for a quick comparison of the median, interquartile range (IQR), and potential outliers for each diagnosis group. From the box plot people with high blood pressure are more prone to chronic diseases where as people with moderate blood pressure are healthy. And people with low blood pressure are having acute conditions. We can say that blood pressure is highly related to diagnosis. There appears to be a difference in blood pressure between the three groups. The healthy blood pressure group has the lowest median blood pressure, followed by the acute condition group, and then the chronic disease group. This suggests that people with chronic diseases tend to have higher blood pressure than people with healthy or acute conditions.

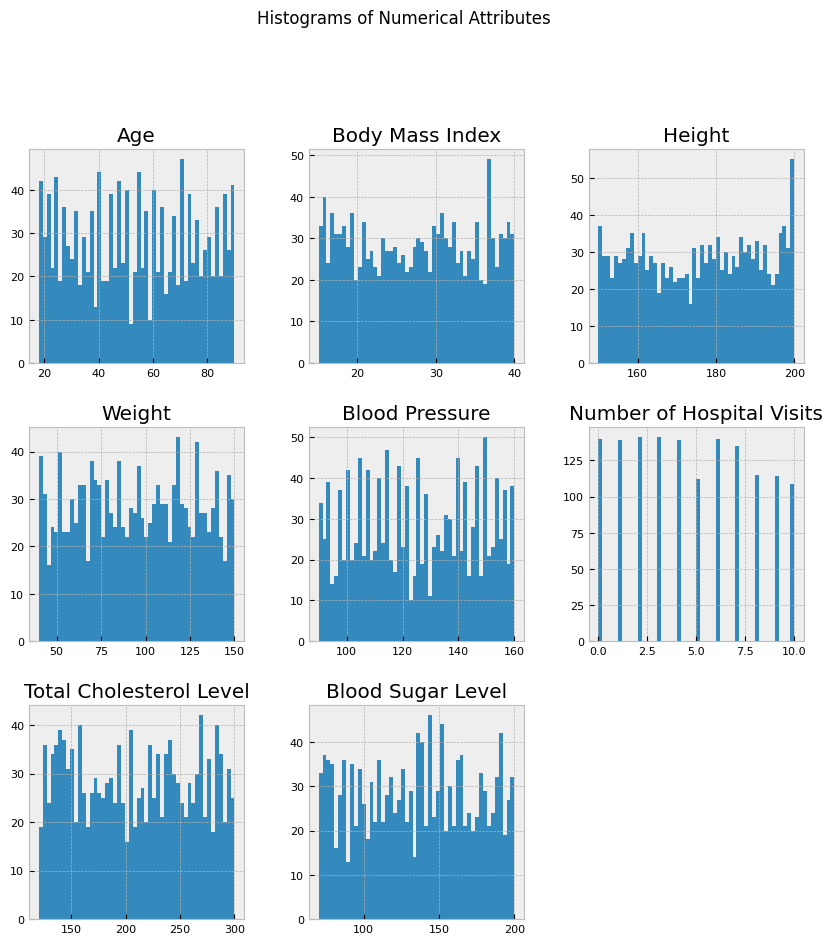
The blood pressure distribution in the chronic disease group is wider than the blood pressure distributions in the other two groups. This means that there is more variability in blood pressure among people with chronic diseases.



From the above Swarm plot, People with chronic diseases tend to have higher BMIs than people with healthy or acute conditions. This is especially true for people who are smokers. The data points for people with chronic diseases are spread out more horizontally than the data points for people with healthy or acute conditions, and there are more data points on the right side of the graph for chronic diseases. Smokers tend to have higher BMIs than non-smokers. This is true for all three diagnoses. The data points for smokers are spread out more horizontally than the data points for non-smokers, and there are more data points on the right side of the graph for smokers. There is some overlap in the BMI ranges between the different groups. For example, some people with healthy conditions have BMIs that are higher than the median BMI for people with chronic diseases.

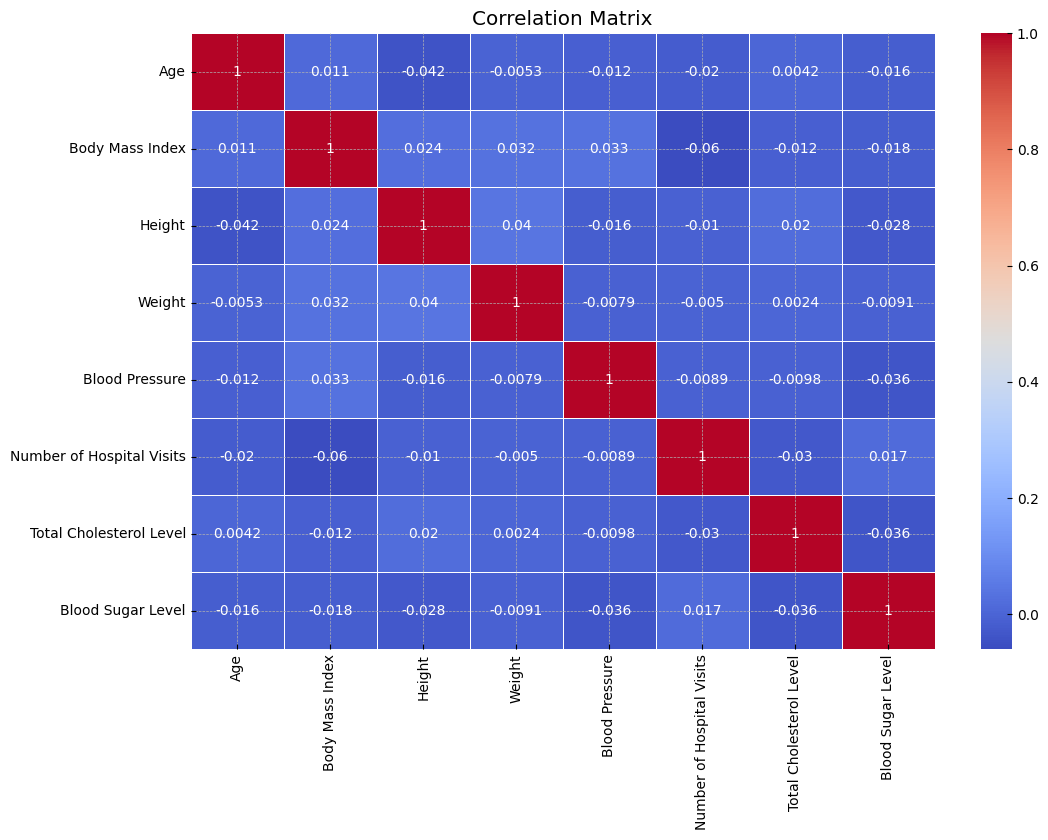
Now let’s look at how all the numerical attributes in our dataset are distributed using histograms.





Each subplot corresponds to a specific attribute, displaying the frequency of different values within that attribute. These histograms offer a quick overview of the data distribution and can aid in identifying patterns, central tendencies, and potential outliers in numerical features. From the histograms we can observe that blood pressure level and blood sugar level have some similar patterns.

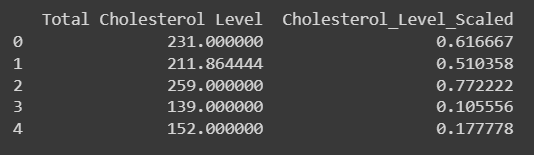
With a comprehensive exploration of our synthetic healthcare dataset through diverse visualizations. To further unravel the interdependencies among different attributes, we use to the correlation matrix. This matrix provides a quantitative measure of the strength and direction of relationships between numerical features. We can reveal key associations within the dataset by identifying strongly correlated features, providing a broader view of how different health-related factors interact.



In this matrix, values closer to 1 indicate a strong positive correlation, while values closer to -1 signify a strong negative correlation. Age and BMI, Height and Age, Blood pressure and BMI, Weight and Blood Sugar level are strongly corelated.

Data Transformation

Scaling is a crucial preprocessing step in data analysis, especially when dealing with features that have different units or magnitudes. It ensures that all attributes contribute equally to the analysis, preventing those with larger scales from dominating the results. Total Cholesterol Level, as a numerical attribute, can vary in magnitude significantly when compared to other features in a dataset. In this context, Min-Max Scaling is used to convert the 'Total Cholesterol Level' values to a common scale between 0 and 1. This not only improves understanding of the distribution of cholesterol levels, but it also improves the interpretability of machine learning models by preventing specific attributes from disproportionately influencing the outcomes. Below attached picture depicts how the values of total cholesterol values has changed after scaling.



Conclusion

In conclusion, our extensive Exploratory Data Analysis (EDA) of the healthcare dataset revealed valuable insights into a variety of critical aspects of the domain. We meticulously examined patterns and trends related to key health indicators and patient demographics using visualizations and statistical techniques.

We analysed the data using a variety of visualizations such as bar charts, scatter plots, and correlation matrices, revealing complex patterns and dependencies. Our analytical journey included attribute scaling and feature engineering. We gained a better understanding of how different variables relate and influence health outcomes by looking at correlations using heatmaps. We derived valuable insights from this synthetic healthcare dataset using a combination of analytical techniques and exploratory approaches, fostering a deeper understanding of the complex interconnections within individual attributes.