Project Report

Data Analysis on General Health Data using Statistical Methods

Bharath C Beeravelly

bbeerave@stevens.edu

MSc. in Machine Learning

Stevens Institute of Technology

1 Introduction

Embarking on an exploration of health metrics, this project delves into a dataset sourced from a comprehensive survey involving over 300,000 participants. The dataset encapsulates various health indicators, and our study is devoted to unraveling their significance. Employing advanced statistical methods and modeling techniques, our project endeavors to provide a nuanced understanding of the intricate relationships and dynamics among these health metrics.

2 Data Exploration

2.1 Meet the Data

Within the realm of our health dataset lies a wealth of comprehensive information pertaining to individuals' well-being. As previously highlighted, this extensive data is derived from a survey involving the collection of health-related information from over 300,000 patients. Encompassing 18 distinct metrics, it is worth noting that the analysis in subsequent sections places particular emphasis on HeartDisease as a pivotal aspect of our data exploration.

In the array of 18 columns constituting our dataset, we encounter a mix of nominal, ordinal, and continuous variables. Table 1 below provides a comprehensive overview of the data types associated with each column. This classification is crucial for our analytical approach, as it allows us to tailor specific statistical methods and techniques to the nature of each variable.

Variable	**Data Type**
HeartDisease	Nominal
BMI	Continuous
Smoking	Nominal
AlcoholDrinking	Nominal
Stroke	Nominal
PhysicalHealth	Ordinal

MentalHealth	Ordinal
DiffWalking	Nominal
Sex	Nominal
AgeCategory	Nominal
Race	Nominal
Diabetic	Nominal
PhysicalActivity	Nominal
GenHealth	Ordinal
SleepTime	Continuous
Asthma	Nominal
KidneyDisease	Nominal
SkinCancer	Nominal

Table 1: Information about Quantitative and Qualitative

2.2 Data Visualization

2.2.1 Quantitative columns

In our pursuit of a deeper understanding of the dataset, we initiated the exploration through data visualization of all variables. The initial focus was on qualitative columns, and histograms were employed to illuminate the distribution patterns of the data. Notably, these histograms, as seen in figure 1, did not exhibit the characteristic shape of a normal distribution, which is often considered ideal. The assessment of normality for these variables will be addressed in subsequent sections.

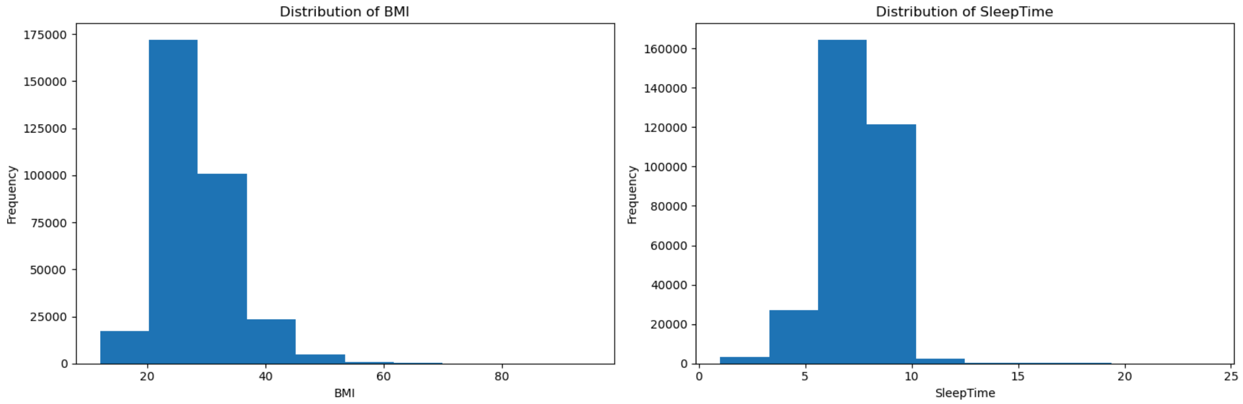


Figure 1: Histogram charts of columns BMI and SleepTime

Following the histogram analysis, we delved into box plot visualization to identify outliers within the dataset. As seen in figure 2, for the BMI column, approximately 10,000 outliers were identified. In which most of the outliers lie outside 1.5 times IQR, suggesting that the outliers are of data of the extreme obese people. And for the SleepTime column, the count stood at around 4,500 outliers. It's worth highlighting that these outliers represent a minor fraction, constituting roughly 5% of the entire dataset. Given the scope of this project, outlier management strategies will be thoroughly discussed in a later section.

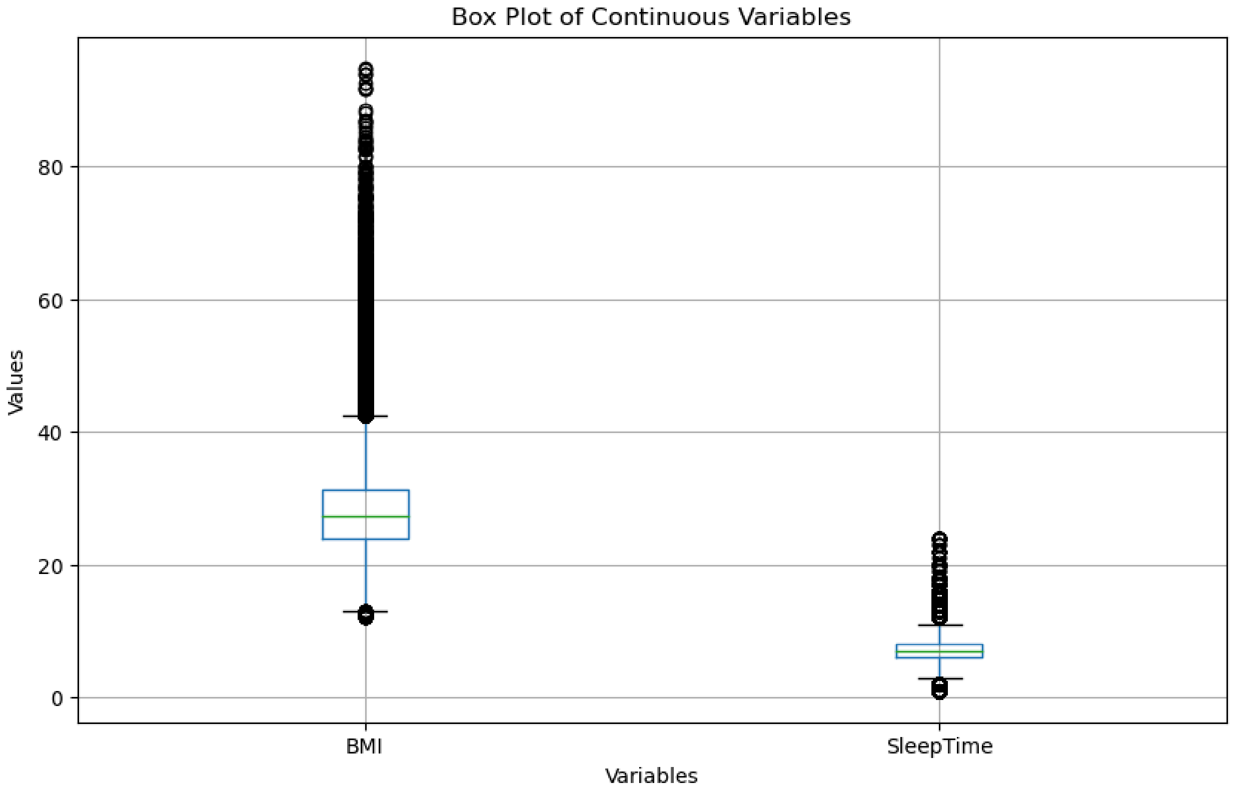


Figure 2: Boxplot of columns BMI and SleepTime

2.2.2 Qualitative columns

Progressing in our analysis, we've employed barcharts to visualize the occurrences of categorical variables. Our initial focus has been on plotting bar charts for nominal variables, revealing a notable observation. Figure 3 illustrates the imbalanced nature of these variables, with the frequency of each occurrence depicted for enhanced clarity. This insight into the distribution of nominal variables sets the stage for a more in-depth examination of their impact on the overall dataset.

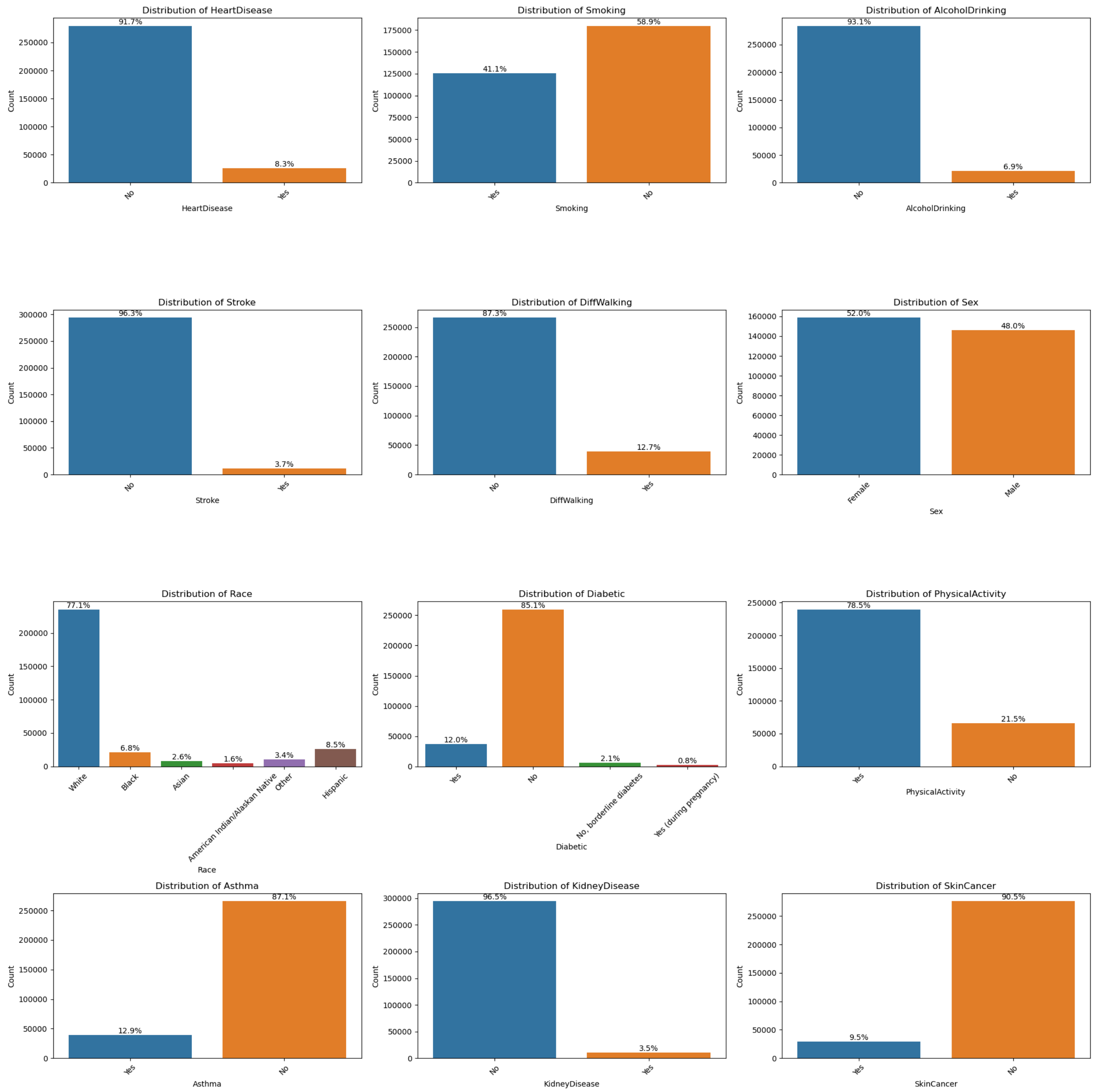


Figure 3: Bar chart of nominal variables

Subsequent to visualizing the occurrences of nominal variables through barcharts, our attention shifted to ordinal variables, employing bar charts for illustration. Figure 4 provides a visual representation, unveiling distinctive patterns within the dataset. Notably, variables such as 'GenHealth' and 'AgeCategory' exhibit a fair distribution of occurrences. Contrastingly, 'PhysicalHealth' and 'MentalHealth', both comprising 30 classes, showcase a concentration of over 60% of occurrences in a single class. This observation prompts a closer examination of the distribution dynamics within these specific ordinal variables, offering valuable insights into potential patterns or disparities.

2.2.3 Normality Tests

The normality test is a crucial step in statistical analysis that assesses whether a given dataset follows a normal distribution. Normality is a fundamental assumption in many statistical methods, and confirming or rejecting this assumption guides the appropriate selection of statistical tests. Various techniques, such as visual methods like histograms and quantitative methods like statistical tests, can be employed to evaluate normality.

As we advance in our analysis, a pivotal focus will be directed towards assessing the normality of key variables, specifically BMI and SleepTime. Given their substantial impact on health metrics, comprehending the distribution of these variables is important for ensuring the precision of our statistical inferences. To scrutinize the normality, Quantile-Quantile (QQ) plots will be employed. In Figure 4, the divergence between the straight line and the plot becomes evident, signifying a noticeable departure from a typical normal distribution. Consequently, we can assert that both the BMI and SleepTime distributions deviate significantly from the normal distribution assumption. This observation holds paramount importance in shaping subsequent steps of our analytical journey.

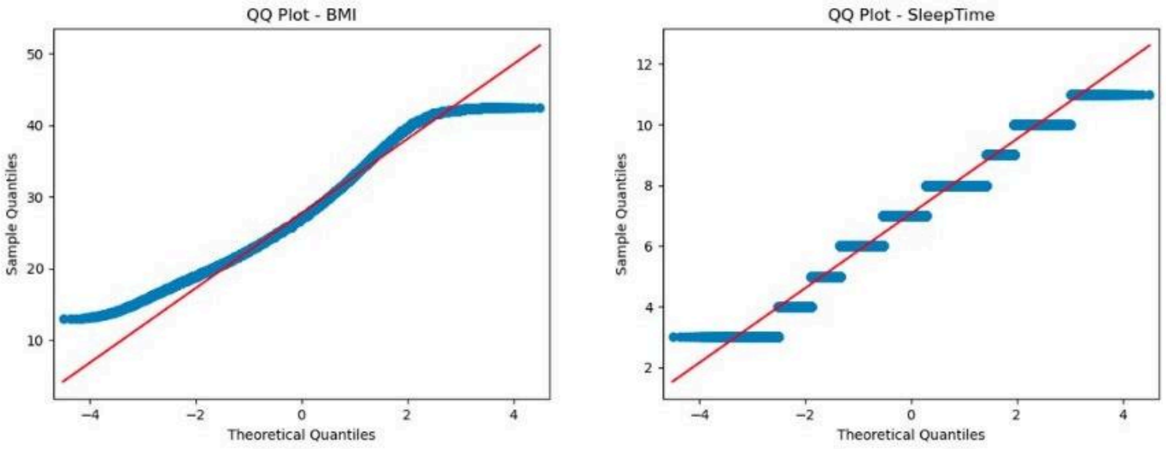
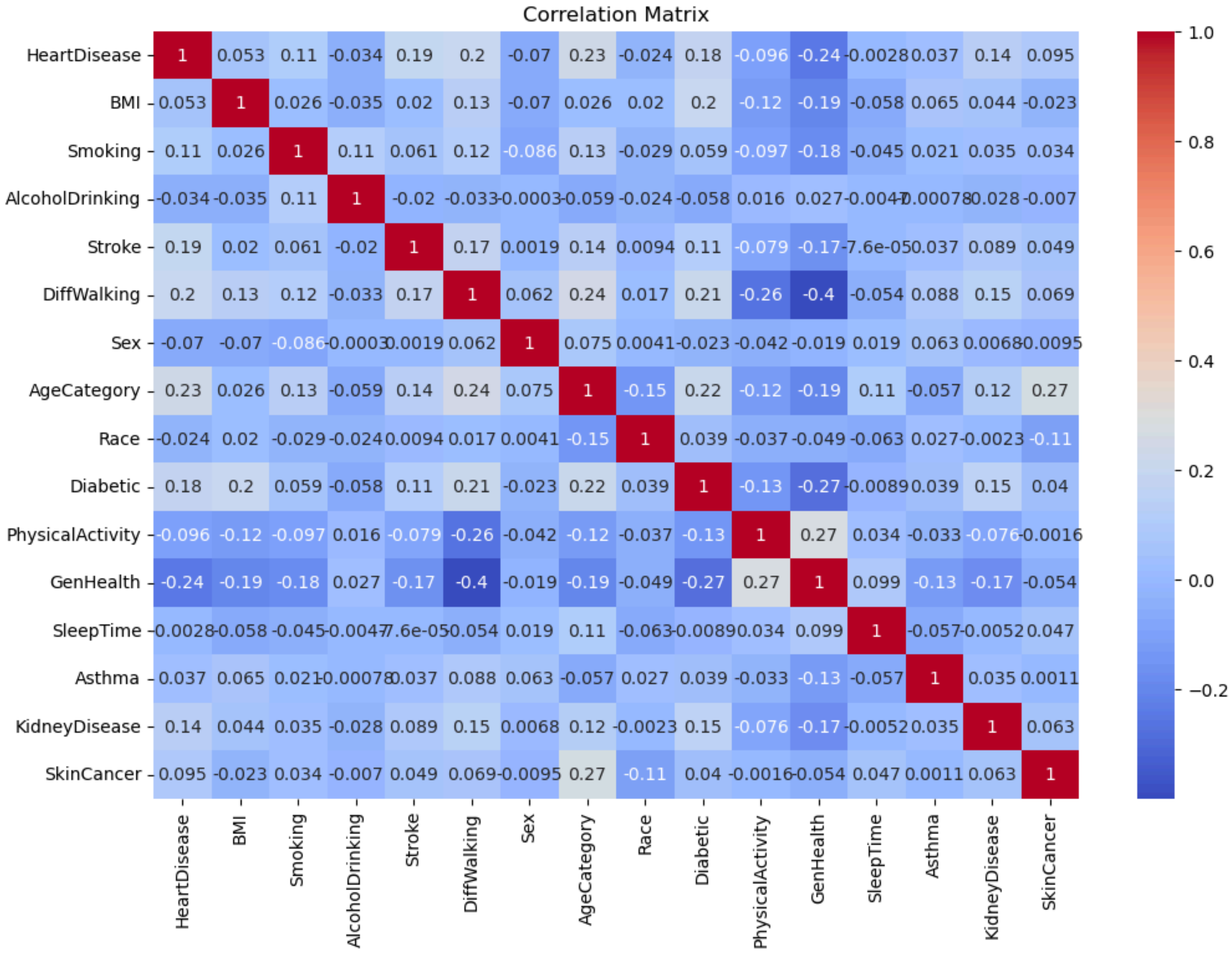


Figure 4: QQ Plots

2.4 Correlation Analysis

Correlation analysis is a fundamental statistical technique employed to assess the strength and direction of relationships between variables. A correlation matrix visually represents the pairwise correlations among variables, offering insights into potential associations within the dataset. Correlation values range from -1 to 1, where -1 indicates a perfect negative correlation, 1 denotes a perfect positive correlation, and 0 suggests no linear correlation.

Upon conducting a correlation analysis on the dataset, as depicted in the figure 5, several noteworthy observations emerge. Notably, there is no strong correlation between the target variable and other variables, indicating a lack of a predominant linear relationship. However, a subtle positive correlation is identified between 'DiffWalking' and 'PhysicalHealth.' Additionally, negative correlations are observed in the pairs ('GenHealth', 'PhysicalHealth') and ('GenHealth', 'DiffWalking'). These nuanced correlation patterns provide a foundational understanding of the interplay between variables and guide further exploration in subsequent analytical phases.



	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
HeartDisease	1	0.053	0.11	-0.034	0.19	0.2	-0.07	0.23	-0.024	0.18	-0.096	-0.24	-0.0028	0.037	0.14	0.095
BMI	0.053	1	0.026	-0.035	0.02	0.13	-0.07	0.026	0.02	0.2	-0.12	-0.19	-0.058	0.065	0.044	-0.023
Smoking	0.11	0.026	1	0.11	0.061	0.12	-0.086	0.13	-0.029	0.059	-0.097	-0.18	-0.045	0.021	0.035	0.034
AlcoholDrinking	-0.034	-0.035	0.11	1	-0.02	-0.033	-0.0003	-0.059	-0.024	-0.058	0.016	0.027	-0.0047	-0.00078	-0.028	-0.007
Stroke	0.19	0.02	0.061	-0.02	1	0.17	0.0019	0.14	0.0094	0.11	-0.079	-0.17	-7.6e-05	0.037	0.089	0.049
DiffWalking	0.2	0.13	0.12	-0.033	0.17	1	0.062	0.24	0.017	0.21	-0.26	-0.4	-0.054	0.088	0.15	0.069
Sex	-0.07	-0.07	-0.086	-0.0003	0.0019	0.062	1	0.075	0.0041	-0.023	-0.042	-0.019	0.019	0.063	0.0068	-0.0095
AgeCategory	0.23	0.026	0.13	-0.059	0.14	0.24	0.075	1	-0.15	0.22	-0.12	-0.19	0.11	-0.057	0.12	0.27
Race	-0.024	0.02	-0.029	-0.024	0.0094	0.017	0.0041	-0.15	1	0.039	-0.037	-0.049	-0.063	0.027	-0.0023	-0.11
Diabetic	0.18	0.2	0.059	-0.058	0.11	0.21	-0.023	0.22	0.039	1	-0.13	-0.27	-0.0089	0.039	0.15	0.04
PhysicalActivity	-0.096	-0.12	-0.097	0.016	-0.079	-0.26	-0.042	-0.12	-0.037	-0.13	1	0.27	0.034	-0.033	-0.076	-0.0016
GenHealth	-0.24	-0.19	-0.18	0.027	-0.17	-0.4	-0.019	-0.19	-0.049	-0.27	0.27	1	0.099	-0.13	-0.17	-0.054
SleepTime	-0.0028	-0.058	-0.045	-0.0047	-7.6e-05	-0.054	0.019	0.11	-0.063	-0.0089	0.034	0.099	1	-0.057	-0.0052	0.047
Asthma	0.037	0.065	0.021	-0.00078	0.037	0.088	0.063	-0.057	0.027	0.039	-0.033	-0.13	-0.057	1	0.035	0.0011
KidneyDisease	0.14	0.044	0.035	-0.028	0.089	0.15	0.0068	0.12	-0.0023	0.15	-0.076	-0.17	-0.0052	0.035	1	0.063
SkinCancer	0.095	-0.023	0.034	-0.007	0.049	0.069	-0.0095	0.27	-0.11	0.04	-0.0016	-0.054	0.047	0.0011	0.063	1

Figure 5: Correlation Analysis

2.5 Data preprocessing

Addressing real-world data intricacies is a critical step to ensure the robustness of statistical analyses. Common challenges include missing values, non-uniform data types, and the presence of outliers. Fortunately, in our dataset, we are presented with a uniform and complete dataset, devoid of missing values. However, the existence of outliers, while constituting less than 5% of the data, demands attention.

Given the substantial size of our dataset, consisting of a large volume of entries, we have made the decision to address outliers in the 'BMI' and 'SleepTime' variables. Outliers in these variables are removed to enhance the generalizability of our statistical conclusions. This strategic approach aligns with the overarching goal of ensuring the reliability and validity of subsequent analyses conducted on this refined dataset.

3 Inferential Statistics

Inferential Statistics serves as a cornerstone in deriving meaningful insights and making informed conclusions about a population based on a sample of data. Unlike Descriptive Statistics, which provides a summary of the main features of a dataset, Inferential Statistics involves drawing inferences, predictions, and generalizations about a broader population from which the sample is drawn. Through the application of hypothesis testing, confidence intervals, and regression analysis, Inferential Statistics empowers researchers and analysts to extrapolate findings, assess relationships, and make predictions with a level of confidence, ultimately contributing to informed decision-making and a deeper understanding of the underlying phenomena.

In this project, our focus extends beyond mere descriptive analyses, as we delve into the realm of Inferential Statistics to unravel the intricate relationships between the target variable, 'HeartDisease,' and other features within our dataset. The objective is to employ a diverse set of statistical tests, ranging from Chi-Square to regression models. However, it is important to note that, going forward, we are limiting our analysis to a sample size of 1000 since statistical tests tend to fail for larger datasets. By scrutinizing these statistical relationships, we aim to discern patterns, dependencies, and significant associations that can provide nuanced insights into the factors influencing heart disease. This methodical approach will not only contribute to a comprehensive understanding of the dataset but will also pave the way for informed decision-making and predictive modeling in the realm of health metrics.

3.1 Chi-Square Test of Independence

The Chi-Square test of independence stands as a statistical hypothesis test designed to ascertain the presence of a relationship between two categorical variables. Its primary goal is to assess whether the values of one categorical variable are contingent on the values of another and vice versa. In the context of this project, our focal point is the target variable, HeartDisease. To scrutinize the interplay between HeartDisease and various binary variables (Smoking, AlcoholDrinking, Stroke, DiffWalking, Sex, PhysicalActivity, Asthma, KidneyDisease, SkinCancer, Diabetic), we will employ the Chi-Square Test of Independence. The hypotheses for this test are formulated as follows:

H_0: The nominal variable and HeartDisease are independent of each other

H_1: The nominal variable and HeartDisease are not independent of each other

Upon applying the Chi-Square Test of Independence on all the above mentioned variables, we get the results as shown in Table 2. In every test, we reject the null hypothesis when we get p-value less than 0.05. This means that there is evidence to reject the null hypothesis

Variable	Conclusion
Smoking	HeartDisease is dependent on Smoking
AlcoholDrinking	HeartDisease is independent on AlcoholDrinking
Stroke	HeartDisease is dependent on Stroke
DiffWalking	HeartDisease is dependent on DiffWalking
Sex	HeartDisease is dependent on Sex
PhysicalActivity	HeartDisease is dependent on PhysicalActivity
Asthma	HeartDisease is independent on Asthma
KidneyDisease	HeartDisease is dependent on KidneyDisease
SkinCancer	HeartDisease is dependent on SkinCancer
Diabetic	HeartDisease is dependent on Diabetic

Table 2: Chi-Square Tests results

Proceeding with the analysis, we sought a visual comprehension of the results by creating bar charts depicting the observed versus expected outcomes. In Figure 6, a distinct pattern emerges: in instances where the null hypothesis was rejected, noticeable disparities are evident in the height of the respective bar charts. Conversely, when the null hypotheses were accepted, the bar charts exhibit nearly identical heights. This visual representation provides a clear and intuitive understanding of the Chi-Square Test of Independence outcomes, allowing for a nuanced interpretation of the significance of associations between HeartDisease and the examined binary variables.

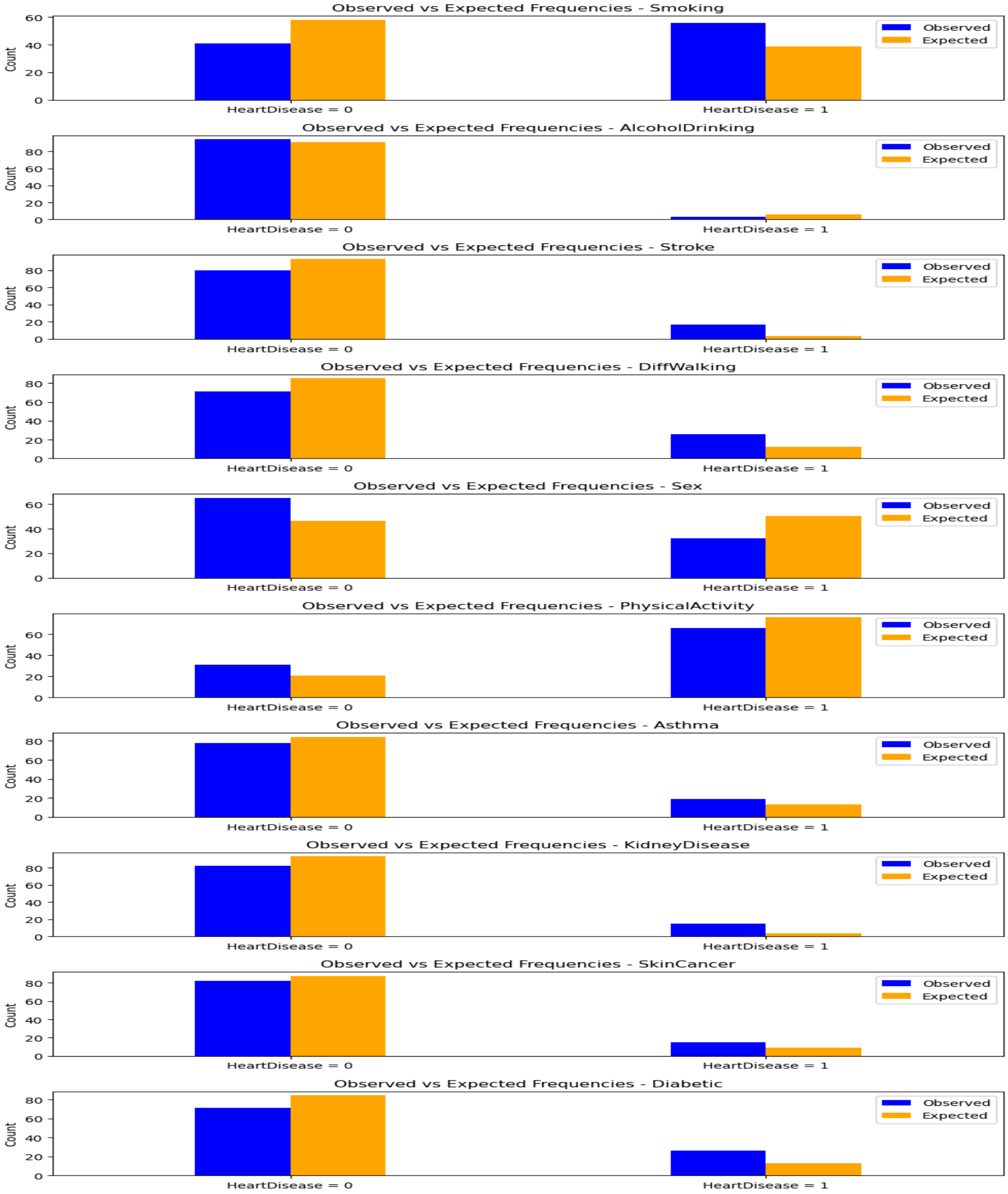


Figure 6: Observed v Expected frequencies

3.2 Kruskal Wallis Test

In the realm of statistical analysis for our project, we turn to the Kruskal-Wallis test to probe into the relationships between the target variable, HeartDisease, and multi-class variables such as Race, AgeCategory, and GenHealth. The Kruskal-Wallis test, a non-parametric method, becomes particularly relevant when dealing with multi-class categorical variables where the assumption of normality or equal variances may not hold. By exploring the medians across different groups within these multi-class variables, we aim to discern any statistically significant differences that could shed light on the impact of factors like Race, AgeCategory, and GenHealth on the likelihood of HeartDisease. Upon conducting the Kruskal Wallis Test, we find the conclusions as mentioned in table 3. These conclusions help us understand the interplay between the multi-class variables and HeartDisease

Variable	Conclusion
Race	HeartDisease is independent of Race
AgeCategory	HeartDisease is dependent on AgeCategory
GenHealth	HeartDisease is dependent on GenHealth

Table 3: Kruskal Wallis Tests

3.3 Mann Whitney U Test

In the analytical journey of our project, we employ the Mann-Whitney U test as a valuable tool to explore the relationships between the target variable, HeartDisease, and two continuous variables, SleepTime and BMI. The Mann-Whitney U test, also known as the Wilcoxon rank-sum test, is a non-parametric method designed to ascertain whether there are significant differences in the distributions of two independent samples. In the specific context of our project, we seek to investigate if the SleepTime and BMI values differ significantly between individuals with and without HeartDisease. This non-parametric approach proves particularly useful when dealing with variables that may not adhere to normal distribution assumptions or when the sample sizes are relatively small. By applying the Mann-Whitney U test, we aim to unravel potential distinctions in SleepTime and BMI that could offer valuable insights into their relationship with the occurrence of HeartDisease. Upon conducting the Mann Whitney U Test, we find the conclusions as mentioned in table 4.

To further elucidate the findings, we have plotted boxplots of each variable, with and without HeartDisease. From Figure 7, we can see that the range of occurrence where there is HeartDisease is quite similar to that of where there is no HeartDisease.

Variable	Conclusion
BMI	HeartDisease is independent of BMI
SleepTime	HeartDisease is independent of SleepTime

Table 4: Mann Whitney U Tests

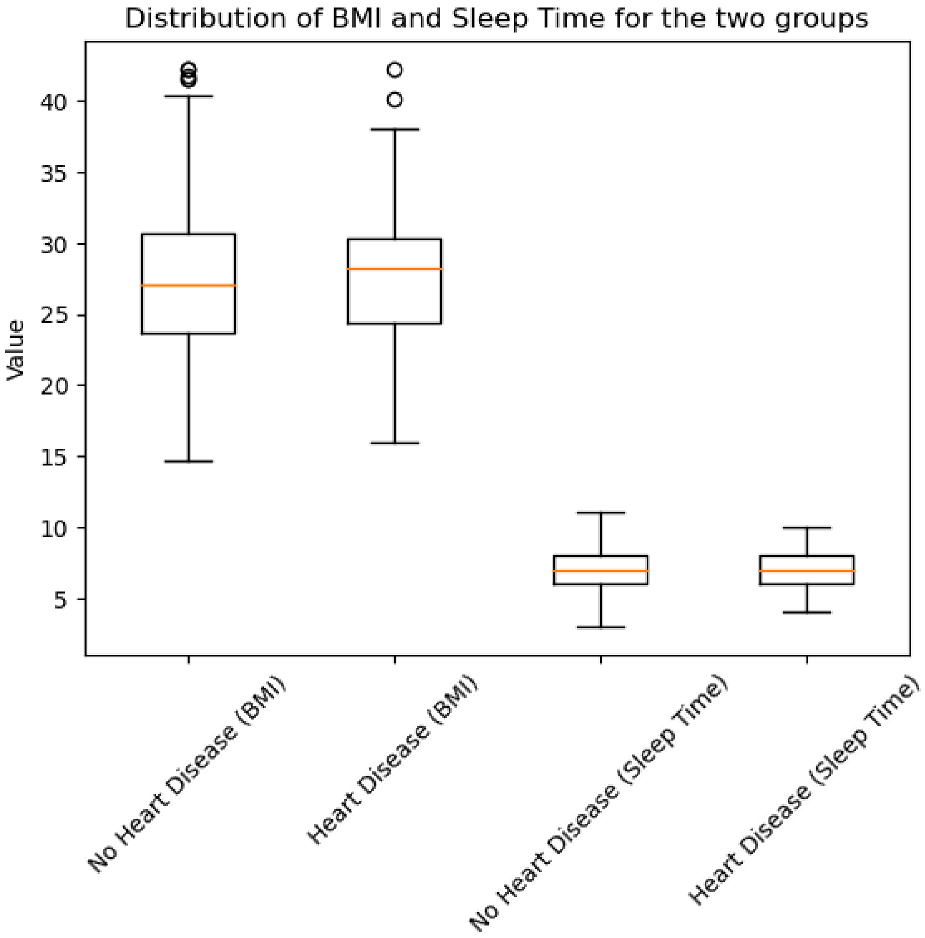


Figure 7: Boxplots of Continuous Variables

4 Regression Analysis

Regression analysis serves as a powerful statistical method employed to model and understand the relationship between a dependent variable and one or more independent variables. In our project, we embark on the application of regression analysis to unravel the intricate connections within our dataset. The central objective is to elucidate how variations in the independent variables contribute to the observed outcomes in the dependent variable. Whether exploring linear, multiple, or logistic regression, this analytical technique equips us with the tools to quantify and interpret the impact of different factors on the target variable. As we navigate through regression analysis, our aim is to derive meaningful insights, make predictions, and contribute to a comprehensive understanding of the underlying patterns and dynamics in our dataset.

For the logistic regression analysis in this project, given the binary nature of the target variable, HeartDisease, we embark on constructing various models to explore the impact of different variable combinations. Initially, individual logistic regression models are built, each featuring a single variable. Subsequently, models are developed encompassing all binary variables collectively, followed by models incorporating multi-class variables, all categorical variables, individual continuous variables, and ultimately, a model integrating all continuous variables. The purpose of this comprehensive approach is to discern the most influential factors and identify the optimal combination of variables for predicting HeartDisease.

In the process of model selection, we employ the R-squared (R2) value as a key metric to gauge the explanatory power of each model. The R2 value provides insights into the proportion of variability in the target variable explained by the model. By comparing R2 values across the diverse set of models, we aim to identify the model that best captures the variability in HeartDisease, offering a robust foundation for drawing meaningful conclusions and making accurate predictions within the scope of our analysis.

5 Conclusion

Throughout this project, our endeavor has been to conduct a comprehensive data analysis, unraveling the intricate relationships between the target variable, HeartDisease, and the diverse set of features. Leveraging statistical methods, we have not only gained valuable insights but also fortified our conclusions with robust evidence. The meticulous exploration of data has provided a nuanced understanding of the factors influencing heart disease, laying a solid foundation for informed decision-making.

Looking ahead, avenues for further exploration beckon. The project opens the door to the implementation of complex regression models, decision trees, and other advanced machine learning techniques. This step promises a deeper understanding and a potentially enhanced fit for our data. Additionally, the incorporation of methods such as Principal Component Analysis (PCA) and Factor Analysis is on the horizon. These techniques hold the potential to unravel latent patterns and dependencies within the data, offering a refined lens through which to make more precise conclusions and conduct insightful experiments. The journey from statistical analysis to advanced machine learning and exploratory techniques marks a progression toward a richer comprehension of our dataset and, ultimately, more informed insights into the dynamics of heart disease.