Risk Analysis of Cardiovascular Disease in the USA

Milestone: Project Report

Group 4

Group Details:

Anjali Haryani
+1 716 277 6868
haryani.a@northeastern.edu
Effort Percentage: 20%
Signature: Anjali Haryani

Jaamie Maarsh Joy Martin
+1 206.229.9258
joymartin.j@northeastern.edu
Effort Percentage: 20%
Signature: Jaamie Maarsh Joy Martin

Kalyan Kumar Chenchu Malakondaiah
+1 206.679.0953
chenchumalakondaia.k@northeastern.edu
Effort Percentage: 20%
Signature: Kalyan Kumar Chenchu Malakondaiah

Payal Chavan
+1 206.290.1057
chavan.pay@northeastern.edu
Effort Percentage: 20%
Signature of Student: Payal Chavan

Shubham Gaur
+1 206.673.0365
gaur.sh@northeastern.edu
Effort Percentage: 20%
Signature of Student: Shubham Gaur

Submission Date: 12/10/2023

Table of Contents

Project Overview:

We are in a day and age of high technological advancement, from predicting the futureto creating AI bots to replace humans but the task has still been to reduce the number of deaths from heart diseases. Cardiovascular diseases (CVDs) are a group of illness that affects the heart and the blood vessels. These disorders can have serious consequences for people's health and overall well-being. Here are some important facts/factors concerning cardiovascular disorders and their consequences to people.

High Mortality Rates: Across the globe, cardiovascular illnesses are the primary causeof mortality. Especially in the US, approximately one individual dies for every 30 seconds of CVD, according to the Disease Control and Prevention (CDC) survey in 2021.

Risk Factors: A sedentary lifestyle, high blood pressure, smoking, high cholesterol, diabetes, and other risk factors all play a part in the development of cardiovascular illnesses.

Age Factor: Elderly people are more susceptible to a very intense cardio condition,which could be life-threatening.

Financial Burden & Background: It can be very expensive to manage cardiovascularillnesses. Medication expenses, medical procedure costs, and hospital stays can put aheavy financial strain on patients and healthcare systems.

Accessibility: There is a large population of people who are still not accessible tohealth care facilities having the adequate and necessary equipment to address theproblem.

Psychological Effects: Stress and depression are some of the psychological effects that can have an adverse effect and can lead to cardiovascular illness. Mental health issues can arise from a variety of factors, including changes in lifestyle, ambiguity aboutthe condition, and future worries.

Complications: A history of medical conditions such as Heart attacks, heart failure, and strokes are just a few of the issues that can result from/lead to cardiovascular disorders. People with cardiovascular diseases may experience worsening health issues because of these consequences. According to research conducted by the World Health Organization (WHO), there is minimal possibility of saving a person's life once they have been impacted by the disease, as they cause a chain reaction of illnesses that depletes the health. Therefore, our primary objective is to identify and answer the major inculcating factors that are high risk in causing CVDs and thereby providing possible solutions to reducing them and ensuring appropriate treatment that can prevent premature deaths.

Objective:

This study aims to obtain the status of cardiovascular diseases (CVDs), their impact on individuals, and potential consequences on health care systems. The main objectives are to determine the prevalence of CVD across different categories, identify risk factors and patterns, and prevent them. Statistics show that, despite technological advances in medicine, CVD remains a major health problem worldwide, causing significant mortality and economic costs.

Key Findings:

1. Complexity of CVDs
2. Major and Minor factors impacting CVD.

Implications:

1. Focusing on early detection
2. Addressing specific problems for the underlying factors.

Project Methodology:

The CVD risk analysis project, like most others, follows an established and effective project management practices, which are as follows:

1. **Data reading and Understanding**: It is the initial or the introductory phase where in the data has been looked at carefully and understood from the data source. Following which, we tend to fix/come up with the objectives/outcome of the study.
2. **Data Cleaning/Pre-processing:** In this stage, the dataset is cleaned and made ready for further processing and visualizations. A detailed explanation on the processing is mentioned below.
3. **Approach and Methods**: This stage involves the finalizing the techniques appropriate to the characteristics and nature of the data.
4. **Data Visualization & Analysis**: The analysis uses various charts and visualization techniques appropriate to the characteristics of the data.
5. **Conclusion, Improvements & Lessons learnt**: We summarize the key findings, reflect on the overall project experience, and outline opportunities for improvement in future projects.

Data Sources:

The dataset under study is being taken from Kaggle named as cardiovascular disease dataset and the link for the same is given below:

CVD dataset

CVD dataset notebook

Data Overview:

The specific problem under study is to identify the type of individuals who are more susceptible in getting CVDs based on the existing data points like physical exercise dietary patterns, usage of psychoactive drugs etc., are some which are under consideration. All data points correspond to the following below.

Attributes	Description
HeartDiseaseorAttack	Indicates Respondents that have ever reported having coronary heart disease
HighBP	Indicates whether an individual has high blood pressure
HighChol	Indicates whether an individual has high cholesterol levels
CholCheck	Indicates whether an individual has undergone cholesterol level checks
BMI	Body Mass Index (BMI) is a measure of body fat based on height and weight. This attribute likely contains the BMI values for individuals.
Smoker	Indicates whether an individual is a smoker
Stroke	Indicates whether an individual has experienced a stroke
Diabetics	Indicates whether an individual has diabetes
PhysActivity	Indicates the level of physical activity or exercise engagement of the individuals
Fruits	Frequency or quantity of fruit consumption
Veggies	Frequency or quantity of vegetable consumption

HvyAlcoholConsumption	Person having heavy Alcohol condition (Adult Men - 14 drinks/week & Adult Women - 7 drinks/week)
AnyHealthcare	Indicates whether the individual has any form of healthcare
GenHlth	General health status or self-reported health assessment of the individuals
PhysHlth	Physical health status or self-reported physical health assessment of the individuals
DiffWalk	Likely indicates any difficulty with walking
Sex	Gender of the individuals (e.g., male or female)
Age	The age of the individual in years
Education	Highest grade/degree completed
Income	Person's annual household income

Data Description:

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey conducted by the Centers for Disease Control and Prevention (CDC). The study collects data from over 400,000 Americans each year on health-related risk behaviors, chronic health issues, and usage of preventive treatments. It has taken place every year since 1984. The dataset contains 22 elements and 253,680 cleaned BRFSS 2015 survey responses that will be used largely for the binary categorization of heart disease. The dataset mostly contains Boolean values, and the detailed description are as follows.

Response Variable:

- HeartDiseaseorAttack - Contains the values (0-Not reported, 1- previouslyreported)

Independent Variables:

- HighBP - Contains the values (0- No presence, 1- Presence of High BP)
- HighChol - Contains the values (0- No presence, 1- Presence of High Chol)
- CholCheck - Contains the values (0- Not gone checked, 1- underwent check)
- BMI - Contains the following values: (Below 18.5: Underweight, 18.5 – 24.9:Normal or Healthy, 25.0 – 29.9: Overweigh, 30.0 and Above: Obese)
- Smoker - Contains the values (0- Non-Smoker, 1- Smoke.

Other Chronic Health Conditions:

- Stroke - Contains the values (0- Not experienced 1- experienced)
- PhysActivity - Contains the values (0- No Activity, 1- Very Active)
- Fruits - Contains the values (0- Less consumption, 1- frequent consumption)
- Veggies - Contains the values (0- Less consumption, 1- frequent consumption)
- HvyAlcoholConsumption - Contains the values (0- Less consumption, 1- frequent consumption)
- AnyHealthcare - Contains the values (0- No healthcare 1- Has Health care)

Individual's health attributes:

- MentHlth - Contains the values (0- No Mental health condition 1- Has Mentalhealth condition)
- NoDocbcCost - Contains the values (0- No cost 1- Facing cost)
- GenHlth - Contains the values (0- Good Health 1- Poor Health)
- PhysHlth - Contains the values (0- Good Physical health 1- Poor Physical health)
- DiffWalk - Contains the values (0- less likely 1- More likely)

Demographics:

- Sex - Contains the values (0- Female 1- Male)
- Age - Contains the following values: ('Young', 'Adult', 'Elderly')
- Education - Contains the values from 1-6, where 1 is the low education level and6 is the highest education level)
- Income - Contains the values from 1-8, where 1 is the low-income level and 8 isthe highest income level)

Challenges Faced:

Here is the set of challenges that were encountered during the initial phase of data analysis:

- Adapting to a column with age values ranging from 1 to 10 presented a challenge, which was addressed by categorizing it into three groups: "young," "adult," and "elderly" using feature engineering approach. This transformation aimed to enhance interpretability and facilitate analysis, offering a more insightful perspective on age-related patterns within the dataset.
- The BMI column, initially numerical, transformed categorical values: 'Underweight,' 'Normal Weight,' 'Overweight,' and 'Obese.' Hence, feature engineering was applied to this column for clearer representation of weight status within distinct categories. This conversion enhances the categorical understanding of BMI for analytical purposes.
- The "Sex" column, initially represented by 0 and 1 values, was redefined to enhance clarity, with 0 now denoting female and 1 indicating male. This modification simplifies the interpretation of gender within the dataset, providing a more intuitive representation for analytical purposes. The revised coding facilitates a straightforward understanding of gender distinctions in the data.
- The "Smoker" column, initially represented by 0 and 1 values, was redefined to enhance clarity, with 0 now denoting “non-smoker” and 1 indicating “smoker”. This modification simplifies the interpretation of smoking status within the dataset, providing a clearer representation for analytical purposes.
- The “Education” and “Income” columns contained uncertain data. Hence, one-hot encoding technique was used on these columns to have consistent data and to maintain data accuracy for our analysis. Later, it was iterated back because it caused increased dimensionality to the dataset.
- Several numerical columns like “High BP”, “High Cholesterol”, “Physical Activity”, “Smoker”, “Diabetes”, “Stroke”, “Chol Check” had inconsistent data. Hence, “Min-Max Scaling” was applied on these columns to ensure data consistency and accuracy.

Data Preprocessing:

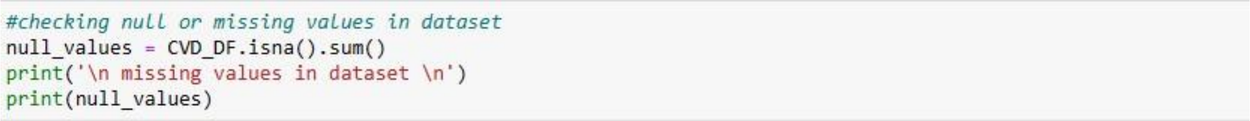
Data preprocessing is an important phase in the data mining process that consists of cleaning, transforming, and merging raw data into a structured format appropriate for further analysis, modeling, and machine learning. Data preparation aims to improve data quality and make it more appropriate for analysis.

Cleaning, instance selection, normalization, one-hot encoding, data transformation, feature extraction, and feature selection are all examples of data preparation procedures.

Given below are the steps undertaken for data cleaning and preprocessing:

Step 1: Checking null/missing values:

To ensure that the data is clean and ready for analysis, it is important to check for null or missing values in the dataset. However, there were no missing values found in our dataset.

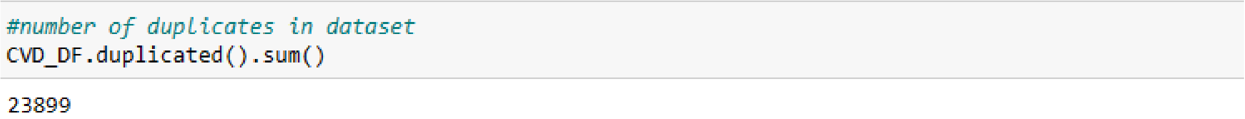


```
#checking null or missing values in dataset
null_values = CVD_DF.isna().sum()
print('\n missing values in dataset \n')
print(null_values)
```

Fig 1.1- Checking null/missing values in the dataset

Step 2: Checking duplicate values:

While cleaning data, it’s necessary to ensure that there are no duplicate values present in the dataset. They can lead to incorrect analysis and biased results. In our dataset, we found out that there were 23899 duplicate records. Hence, we eliminated them.

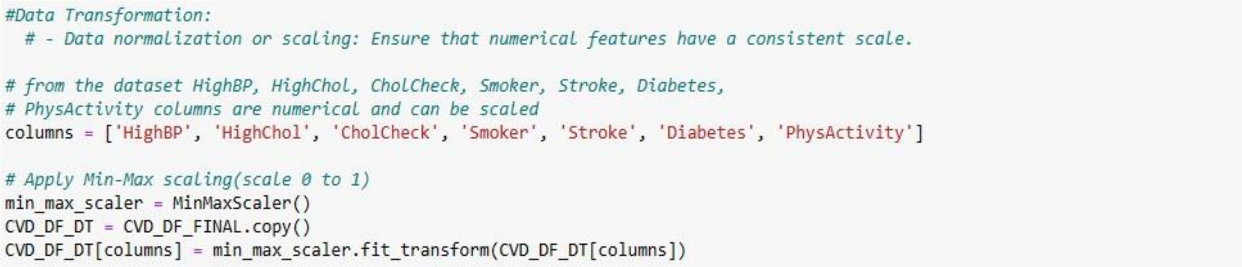


```
#number of duplicates in dataset
CVD_DF.duplicated().sum()

23899
```

Fig 2.1 Checking for duplicate records

Step 3: Data Transformation:

Data transformation is the process of converting data into a more usable format for analysis. After the data has been cleansed, data transformation techniques such as "data normalization" and "scaling" can be used. Data normalization is the process of scaling (Min-Max scaling) data to a given range to prevent bias caused by differing units of measurement. To maintain data consistency, several numerical features were scaled between 0 and 1.

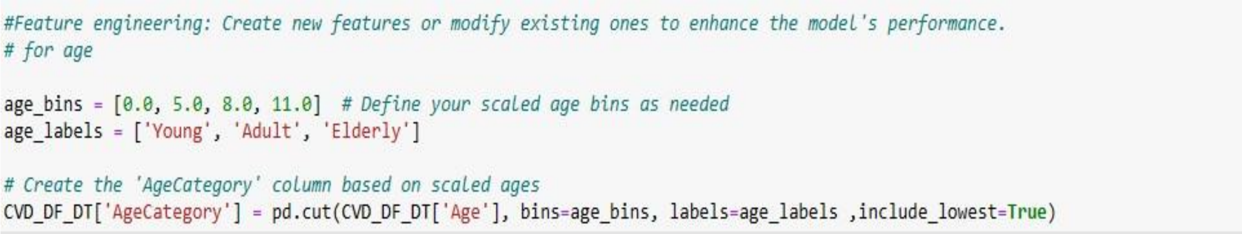


```
#Data Transformation:
  # - Data normalization or scaling: Ensure that numerical features have a consistent scale.

# from the dataset HighBP, HighChol, CholCheck, Smoker, Stroke, Diabetes,
# PhysActivity columns are numerical and can be scaled
columns = ['HighBP', 'HighChol', 'CholCheck', 'Smoker', 'Stroke', 'Diabetes', 'PhysActivity']

# Apply Min-Max scaling(scale 0 to 1)
min_max_scaler = MinMaxScaler()
CVD_DF_DT = CVD_DF_FINAL.copy()
CVD_DF_DT[columns] = min_max_scaler.fit_transform(CVD_DF_DT[columns])
```

Fig 3.1- Scaling of features

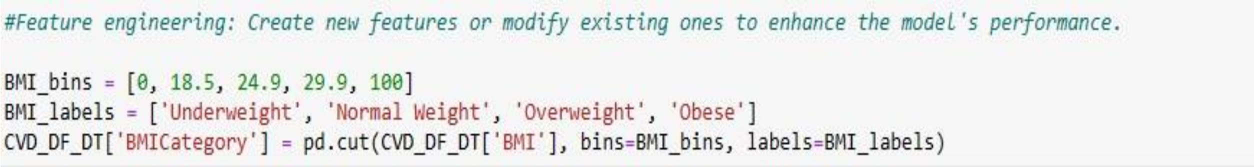
Step 4: Feature Engineering:

Feature engineering is the process of selecting, transforming, extracting, combining, andmanipulating raw data to generate the desired variables for analysis. With the use of feature engineering, we were able to extract valuable information from the already existing features, such as determining a person’s “age” based on only three categories or groups rather than using various distinct ages. Similarly, we performed feature engineering on “BMI” column and created 4 different BMI categories. These new featuresincrease or enhance the model to attain greater accuracy.



```
#Feature engineering: Create new features or modify existing ones to enhance the model's performance.
# for age

age_bins = [0.0, 5.0, 8.0, 11.0]  # Define your scaled age bins as needed
age_labels = ['Young', 'Adult', 'Elderly']

# Create the 'AgeCategory' column based on scaled ages
CVD_DF_DT['AgeCategory'] = pd.cut(CVD_DF_DT['Age'], bins=age_bins, labels=age_labels ,include_lowest=True)
```

Fig 4.1 (a)- Feature Engineering for “age” column



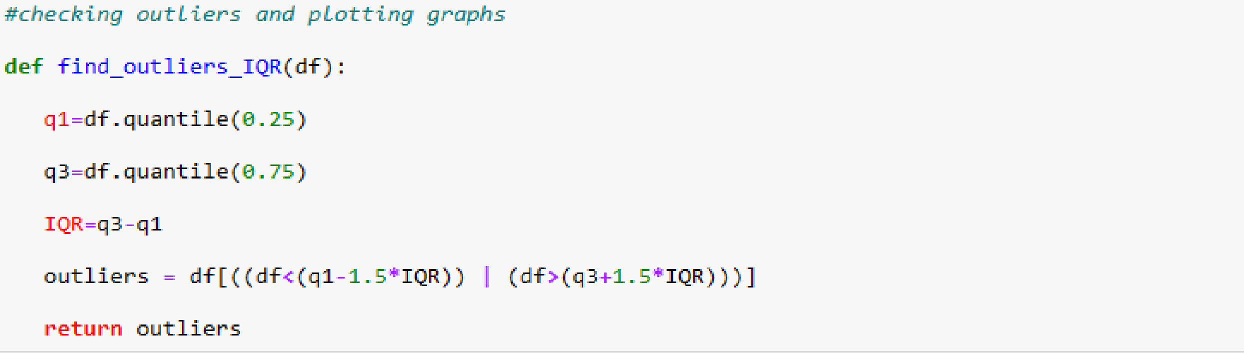
```
#Feature engineering: Create new features or modify existing ones to enhance the model's performance.

BMI_bins = [0, 18.5, 24.9, 29.9, 100]
BMI_labels = ['Underweight', 'Normal Weight', 'Overweight', 'Obese']
CVD_DF_DT['BMICategory'] = pd.cut(CVD_DF_DT['BMI'], bins=BMI_bins, labels=BMI_labels)
```

Fig 4.1 (b)- Feature Engineering for “BMI” column

Step 5: Checking for outliers:

Outliers are extreme values that differ from most other data points in a dataset . They canhave a big impact on our statistical analyses and skew the results. It's important to carefully identify potential outliers in our dataset and deal with them in an appropriate manner for accurate results.

In our CVD dataset, we have identified outliers by using the "Interquartile range method". The interquartile range method involves calculating the interquartile range (IQR) and identifying the values that are more than 1.5 times the IQR away from the firstor third quartile. Once we have identified the outliers, we can deal with them by either removing them from the dataset, replacing them with a more appropriate value, or leavingthem in the dataset if they represent true values from natural variation in the population.



```
#checking outliers and plotting graphs

def find_outliers_IQR(df):

   q1=df.quantile(0.25)

   q3=df.quantile(0.75)

   IQR=q3-q1

   outliers = df[((df<(q1-1.5*IQR)) | (df>(q3+1.5*IQR)))]

   return outliers
```

Fig 5.1 Function to find outliers in dataset & check distribution of data

In our data analysis, we've utilized the interquartile range (IQR) and box plot techniques to assess potential outliers. Now, we shall check the outliers for the features, which showed a greater number of outliers in our dataset.

1. **BMI**

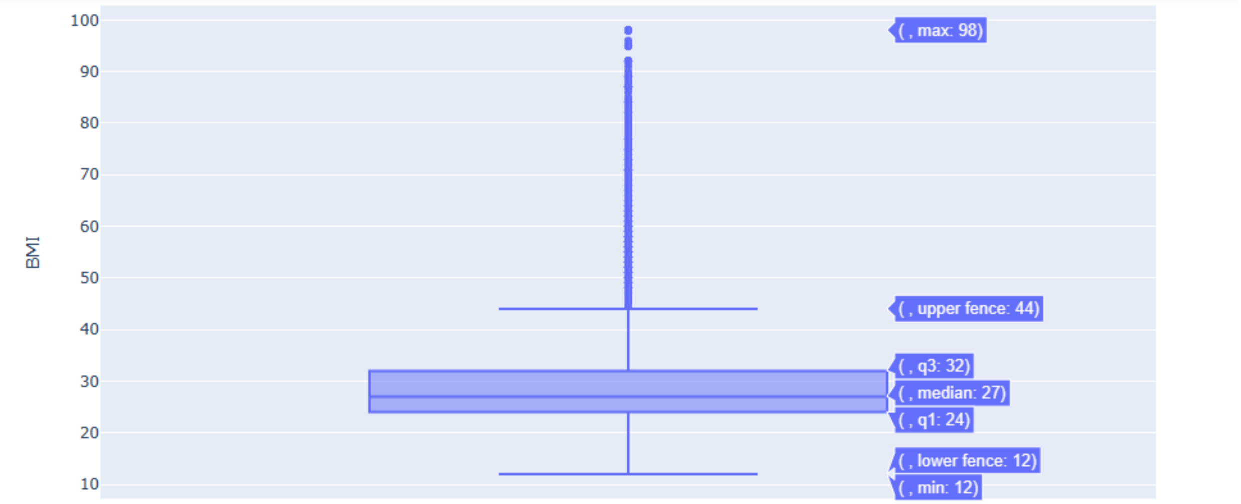


Fig 5.1 (b)- Box plot of "BMI" column

From the given box plot, we can see that the BMI values range from 12 to 98. The medianBMI value is 27, which is the value that separates the upper half of the data from the lowerhalf. Since the median is closer to the lower quartile of 24, we can say that the distribution is positively skewed. The interquartile range (IQR) is 8, which is the difference between the third quartile (Q3) and the first quartile (Q1). Since the distribution is not symmetric or normal, we cannot assume that the mean and the median are equal. As the BMI represents the true value for the natural variation in the population, we've made a conscious decision not to label them as outliers and keep them as it is for our future analysis.

2. MentHlth

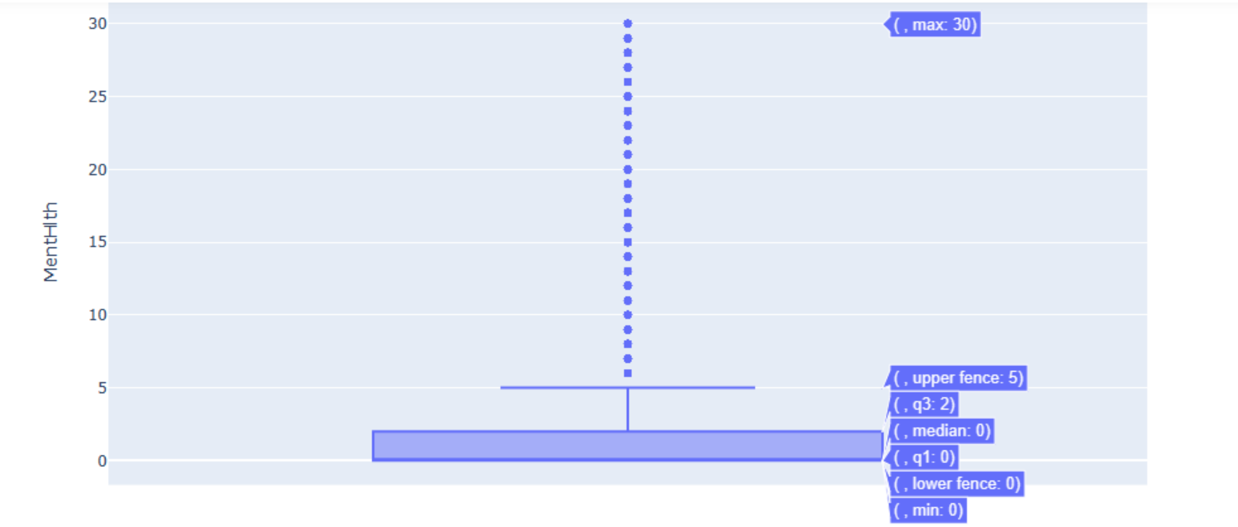


Fig 5.1 (c)- Box plot of “MentHlthcolumn

From the above box plot, we can interpret that the range of values lie between 0 to 30. The median value is 0, suggesting that there are 50% of values that are positive,and 50% of values that are negative. The distribution of the data is positively skewedas the median is closer to the lower quartile of 0. The IQR(Q3-Q1) is 2. Since the distribution is not symmetric or normal, we cannot assume that the mean and the median are equal. If we remove the outliers, there will be very few values left in this column for our analysis, hence we have chosen to keep the outliers in our dataset.

3. **GenHlth**

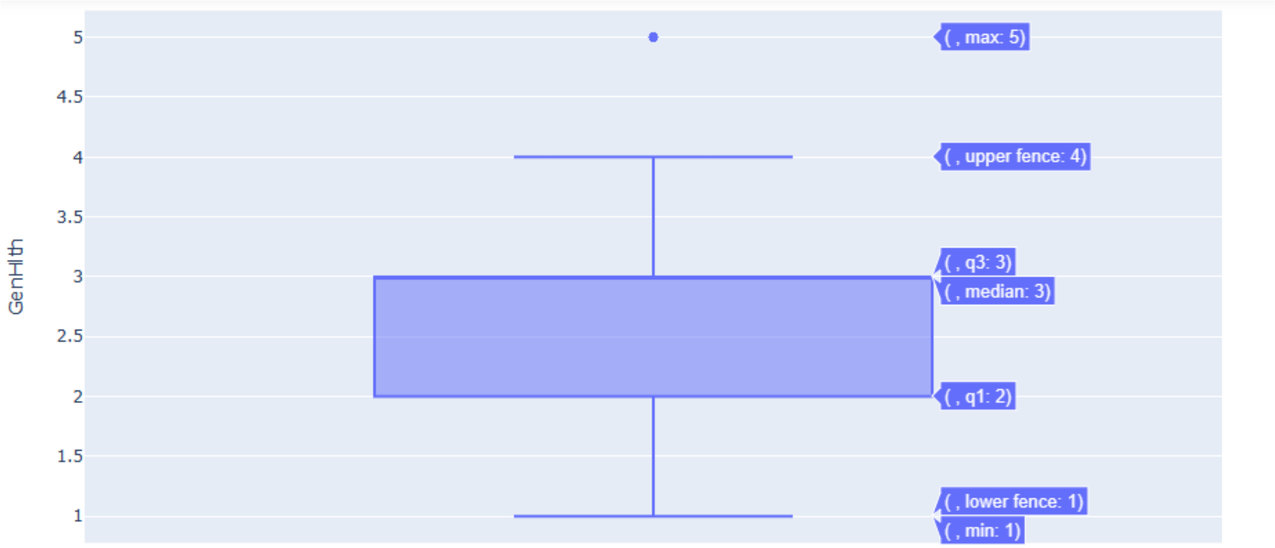


Fig 5.1 (d)- Box plot of “GenHlth” column

The distribution of the above box plot appears to be positively skewed, meaning that the data is clustered towards the lower end of the graph, with a few outliers towards the higher end. The median value is 3, the lower quartile is 2, and the upper quartile is 3. The minimum value is 1 and the maximum value is 5. The IQR(Q3-Q1) is 1. The “GenHlth” column has values ranging from 1 to 5, which can be categorized as a range from very weak to very healthy. Since all these categories are important for our analysis, we have decided to keep the outliers in our dataset.

EDA Approach & Methods:

During this analysis, we applied Min-Max scaling to normalize the data within a range of 0 to 1. Following this scaling process, we selectively considered columns with values falling within this normalized range for further investigation. Moreover, we utilized feature engineering techniques to derive meaningful insights from the existing features. A noteworthy outcome of this approach was the creation of a new feature that categorizes age into three groups. This innovative categorization not only simplifies the representation of age but also contributes to enhancing the model's accuracy by retaining crucial information.

In our analysis, we employ various visualization techniques tailored to the characteristics of the data. Bar charts are chosen for their effectiveness in handling categorical data, particularly when dealing with non-string components and a limited number of distinct values. We opt for pie charts due to the dataset's small number of categories, offering a clear representation of elevated cholesterol proportions in different age groups. Line plots aid in illustrating trends and variations over a continuous scale, facilitating the interpretation of correlations between BMI and high blood pressure.

Histograms are employed to identify patterns in data distribution, focusing on physical health and BMI. Additionally, outlier detection using the Interquartile Range (IQR) informs decisions on column retention or discarding. Furthermore, scatterplots are utilized to assess heart disease prevalence across age categories, exploring relationships between variables like age and heart disease presence for potential correlations. The heatmap is a valuable tool for visualizing connections between different variables and heart disease, offering a comprehensive perspective on data relationships. Overall, these visualizations serve as valuable tools in interpreting connections within the dataset, particularly concerning heart disease.

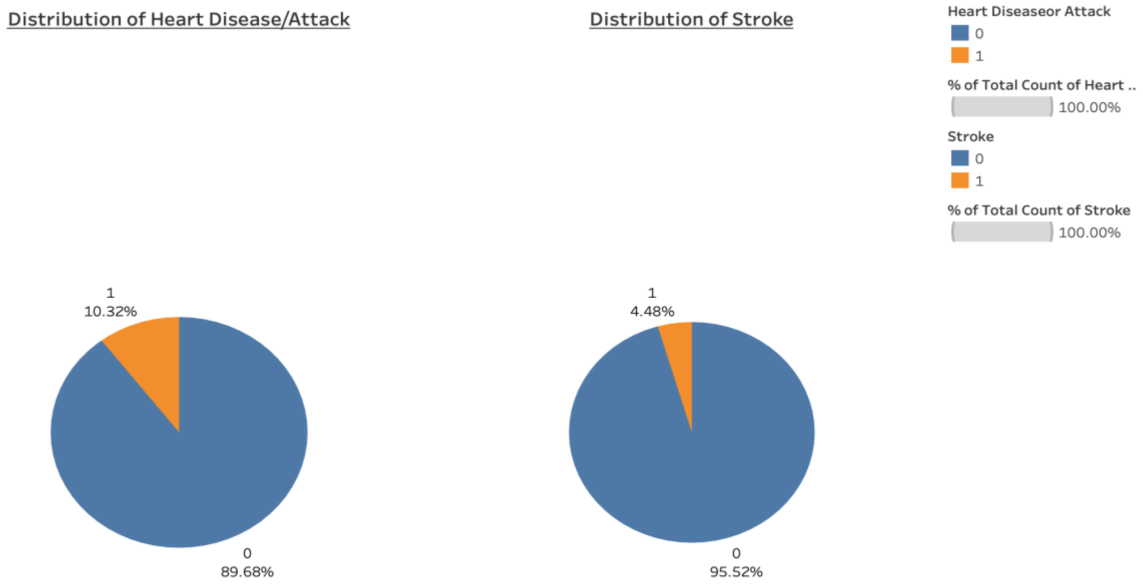
Feedback & Iteration:

Initially, one-hot encoding was applied to the education column, resulting in the creation of multiple columns based on distinct values. However, upon evaluation, these changes were reversed due to the absence of meaningful insights derived from the individual columns.

In our revised approach, our strategy has transitioned from examining relationships across all columns to a more targeted focus on cardiovascular disease (CVD). The current investigation involves plotting categories to discern proportional trends related to heart disease. In the bivariate analysis, each category is compared to assess the proportionality of heart disease within it. Furthermore, in the multivariate analysis, we explore relationships with other variables to uncover potential correlations and gain a deeper understanding of the data.

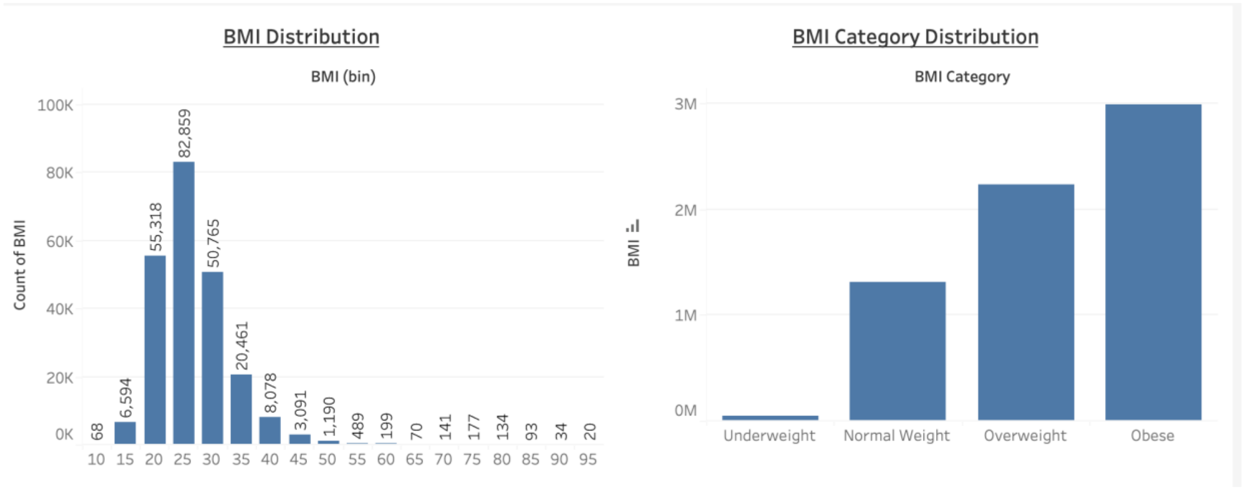
Initial EDA Analysis:
First, we started with checking the distribution of data for different factors.

Univariate Plots:



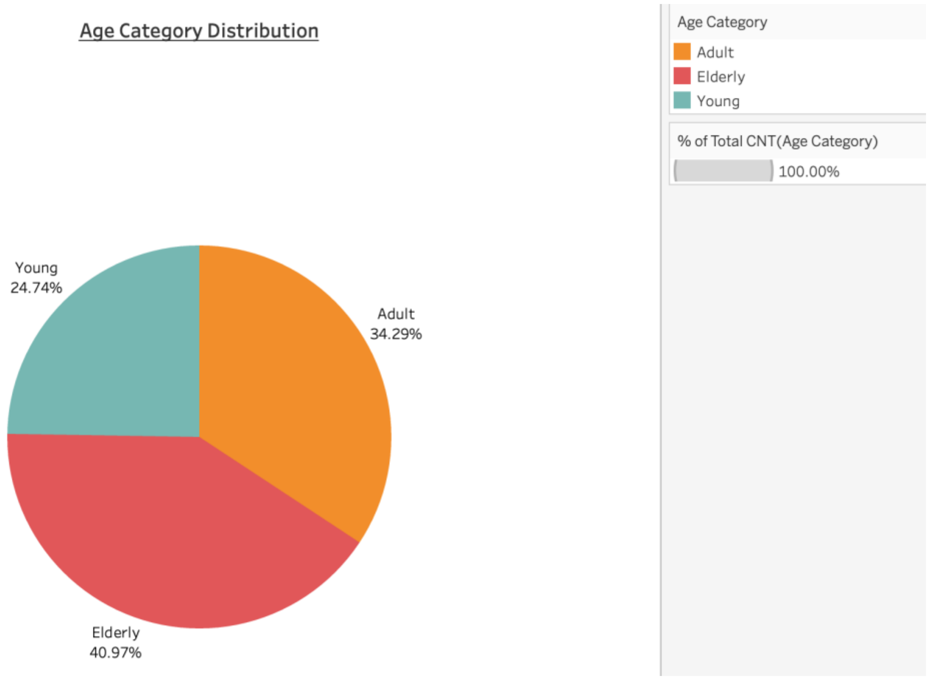
From the analysis of the pie charts, it is evident that within the dataset, approximately 10.3% of individuals had a history of heart disease or a heart attack, while the majority did not exhibit such conditions. Additionally, only 4.8% of the population had a history of stroke, with the remainder showing no signs of stroke.

Furthermore, the data suggests that the prevalence of heart disease or heart attack is higher compared to stroke within the studied population. It also highlights the importance of understanding these health conditions for targeted healthcare interventions. Additionally, the relatively low percentages indicate that the dataset may consist largely of individuals without a documented history of these cardiovascular issues.



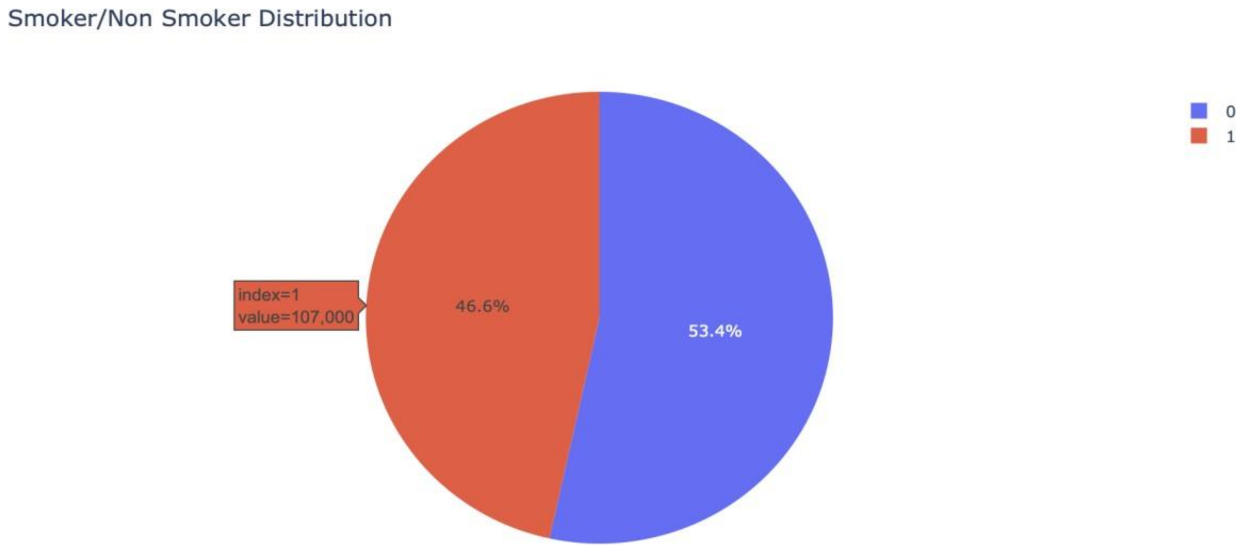
The distribution of BMI in the dataset follows a normal distribution, indicating a balanced spread across various weight categories. Notably, a significant portion of individuals falls within the BMI category of 30, suggesting a concentration around this weight status in the sampled population.

Analyzing the bar chart for BMI categories reveals that 'Obese' and 'Overweight' are the predominant categories among CVD patients, encompassing approximately 9000 individuals. Conversely, 'Underweight' stands as the least prevalent category, with fewer than 500 patients, while around 6000 patients fall within the 'Normal Weight' category.



The dataset is predominantly composed of individuals in the "elderly" age group, comprising

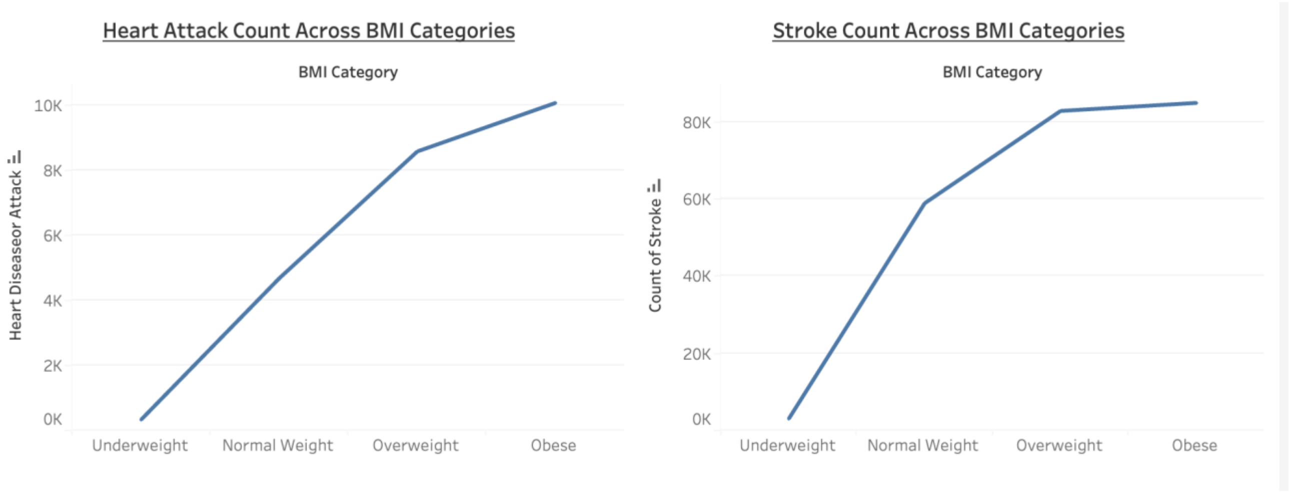
41% of the total. Conversely, the "young" age group is less common, representing only 24.7%. Understanding this age distribution is vital for contextualizing health patterns within specific demographic segment.



The pie chart illustrates the frequency of smokers and non-smokers, with a higher count of non-smokers compared to smokers. This information is pivotal for understanding the distribution of smoking habits within the analyzed population. The visual representation highlights the prevalence of non-smoking individuals in the dataset.

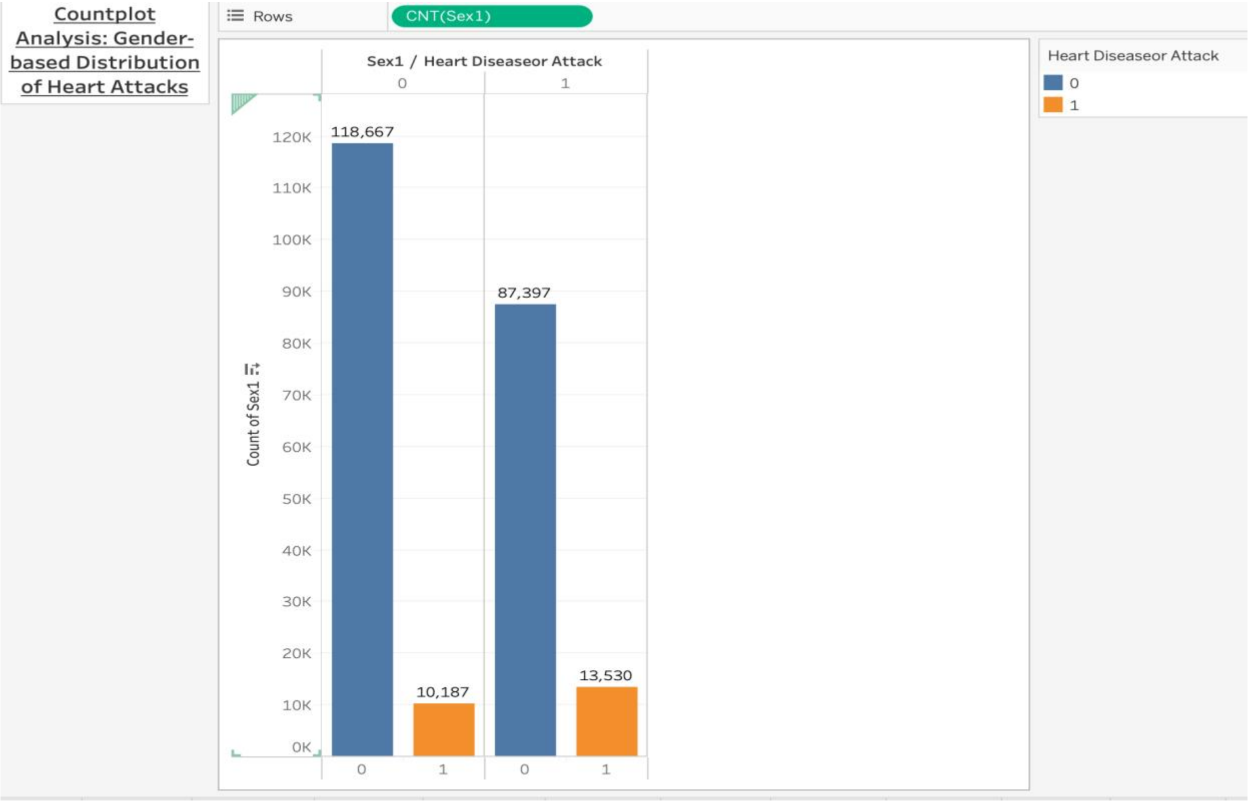
Now, we have visualized the relationship between the heart attack column and various factors in the dataset.

Bivariate Plots:

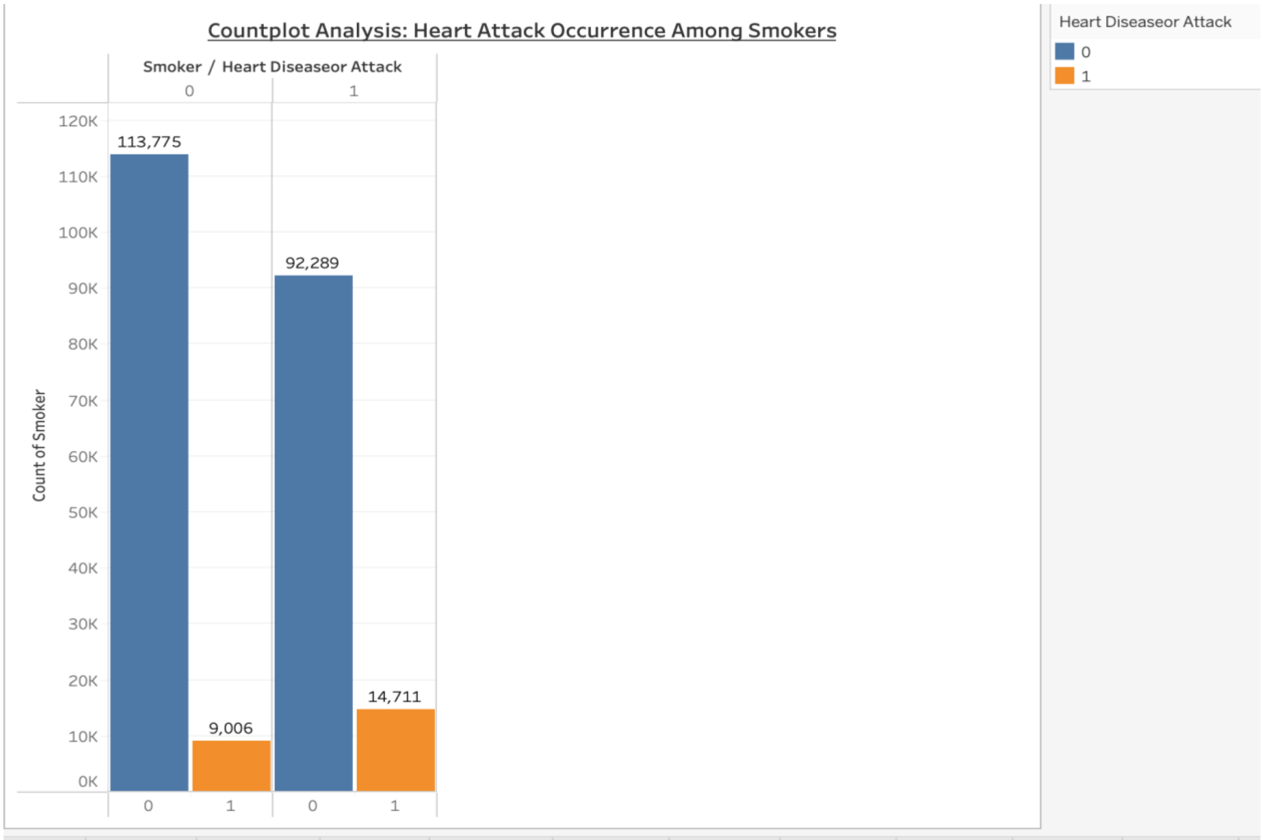


Through the utilization of line plots in both scenarios, a discernible trend emerges: an increase in an individual's BMI correlates with an elevated likelihood of experiencing heart disease and stroke. This suggests a direct proportionality between BMI and the occurrence of heart-related issues.

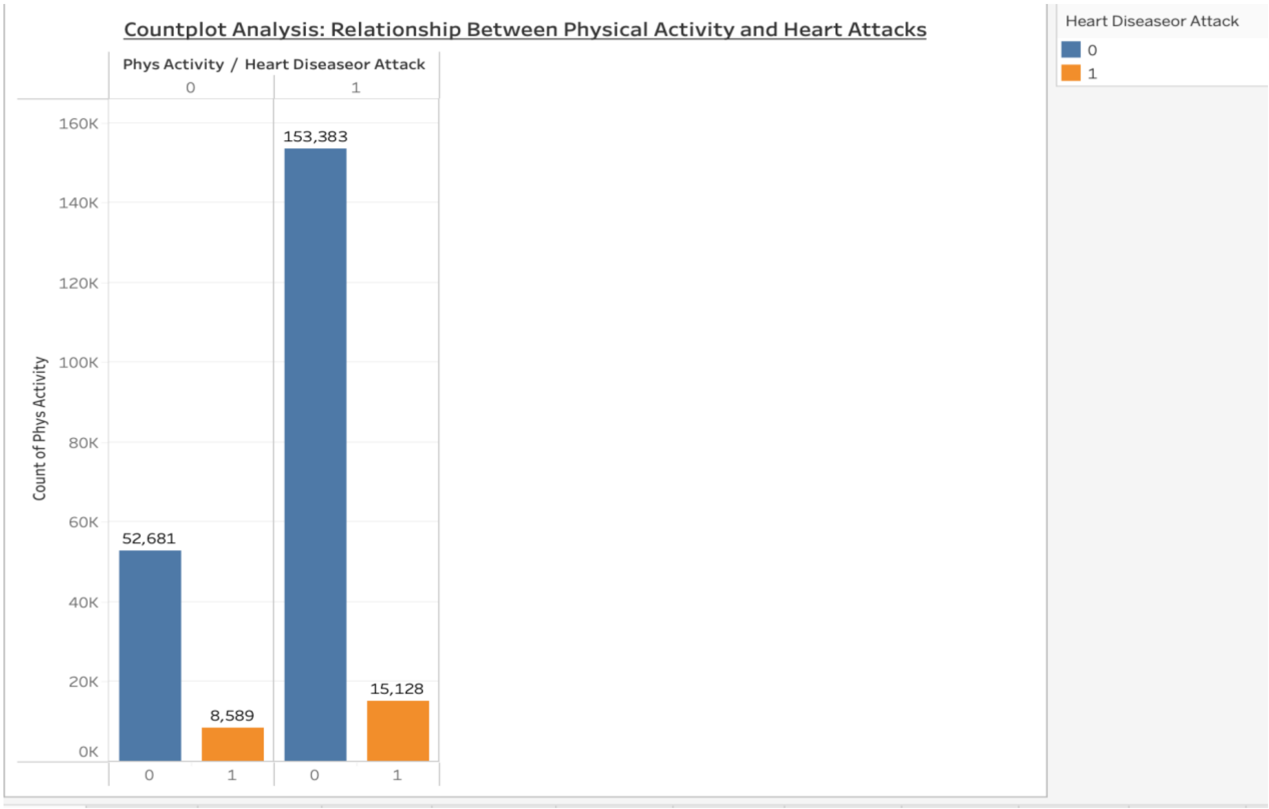
It is noteworthy to consider that higher BMI values may indicate an increased risk of cardiac events. Individuals with elevated BMIs might benefit from closer monitoring of cardiovascular health and adopting lifestyle changes to mitigate the associated risks.



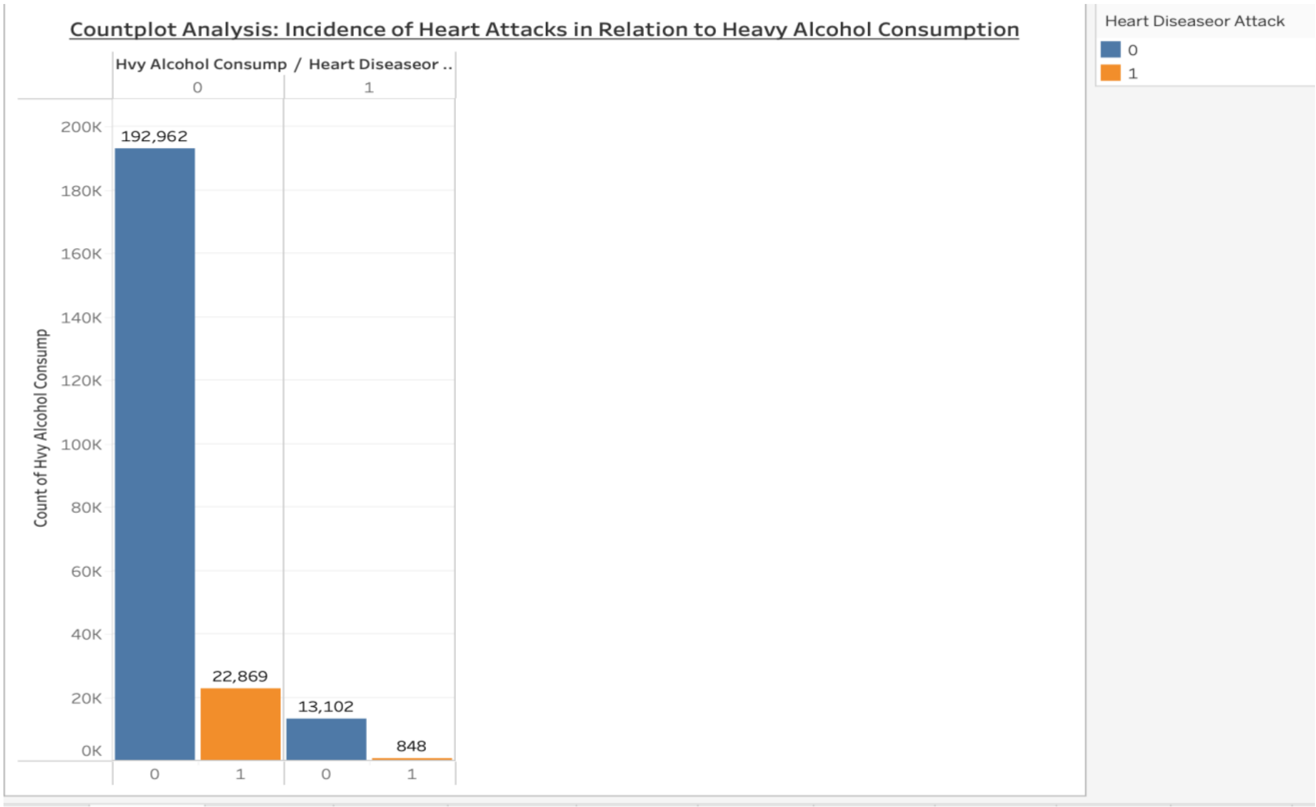
Analyzing the dataset with 0 representing females and 1 representing males, there is a notable predominance of females. The focus of the visualization is on heart attack occurrences relative to gender. The data unveils a significant trend—males demonstrate a heightened susceptibility to heart attacks compared to females. This gender-specific insight accentuates the need to address gender-based disparities in cardiovascular health. The larger representation of females prompts further exploration into heart related issues across genders. This underscores the importance of gender-specific healthcare strategies to address distinct cardiovascular challenges faced by males and females, highlighting the need for balanced gender representation in healthcare research.



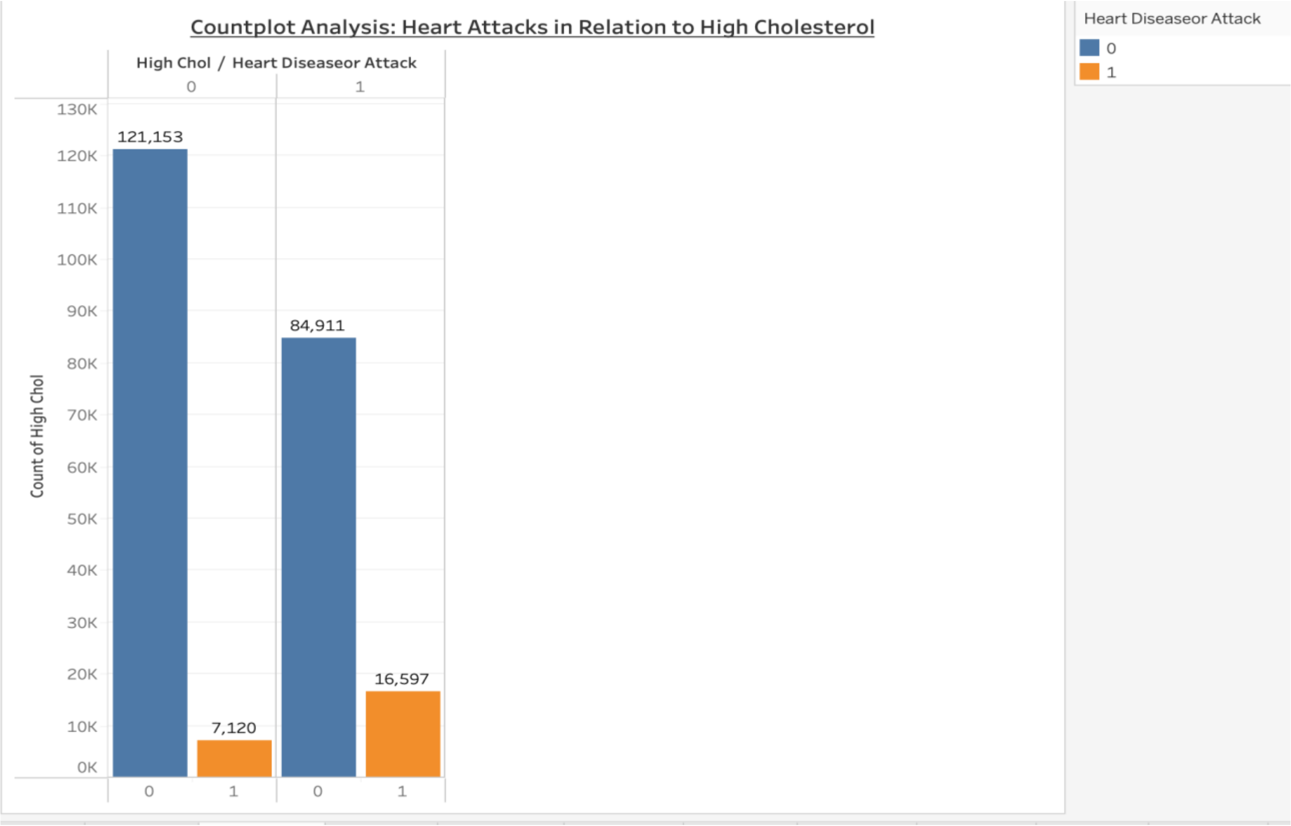
In the dataset, non-smokers outnumber smokers. Analyzing heart disease occurrences reveals a slight trend: smokers show a slightly higher likelihood of experiencing heart disease. This underscore smoking as a risk factor for heart issues, suggesting a correlation. The prevalence of heart disease among smokers emphasizes the need for targeted preventive measures and public health campaigns. The dataset advocates for comprehensive cardiovascular interventions considering lifestyle factors like smoking. The observed contrast signals potential impactful interventions, especially in populationswith a higher prevalence of smoking.



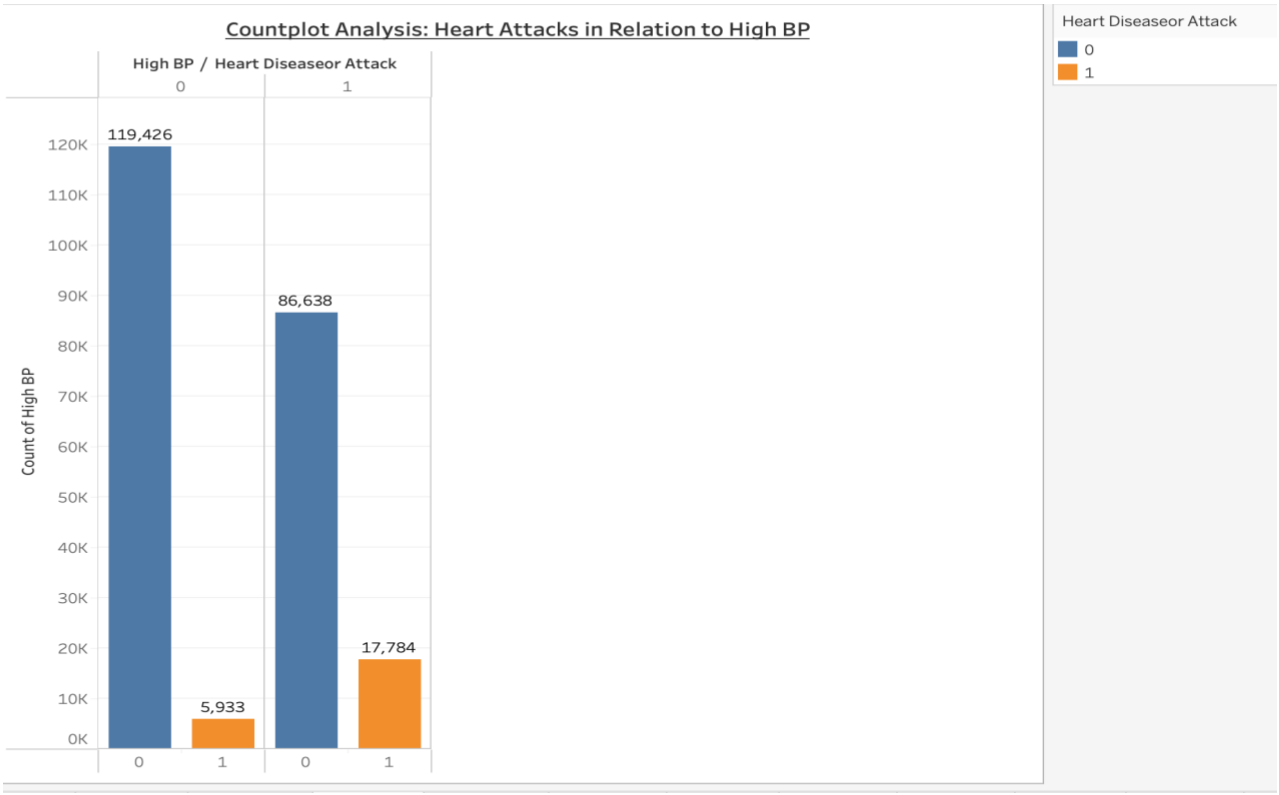
The dataset reveals a substantial portion engaged in regular physical activity. Interestingly, individuals with an active lifestyle show a slightly higher susceptibility to heart disease, possibly linked to activities like running. This nuanced observation emphasizes the need for a detailed understanding of the physical activity-heart healthy relationship. Despite the acknowledged benefits of exercise, the dataset suggests complexity in this association, highlighting the importance of exploring specific types and durations of physical activities. Additionally, personalized health recommendations considering activity profiles are crucial. The dataset provides valuable insights into the intricate interplay between physical activity and heart health, urging further exploration for nuanced public health strategies.



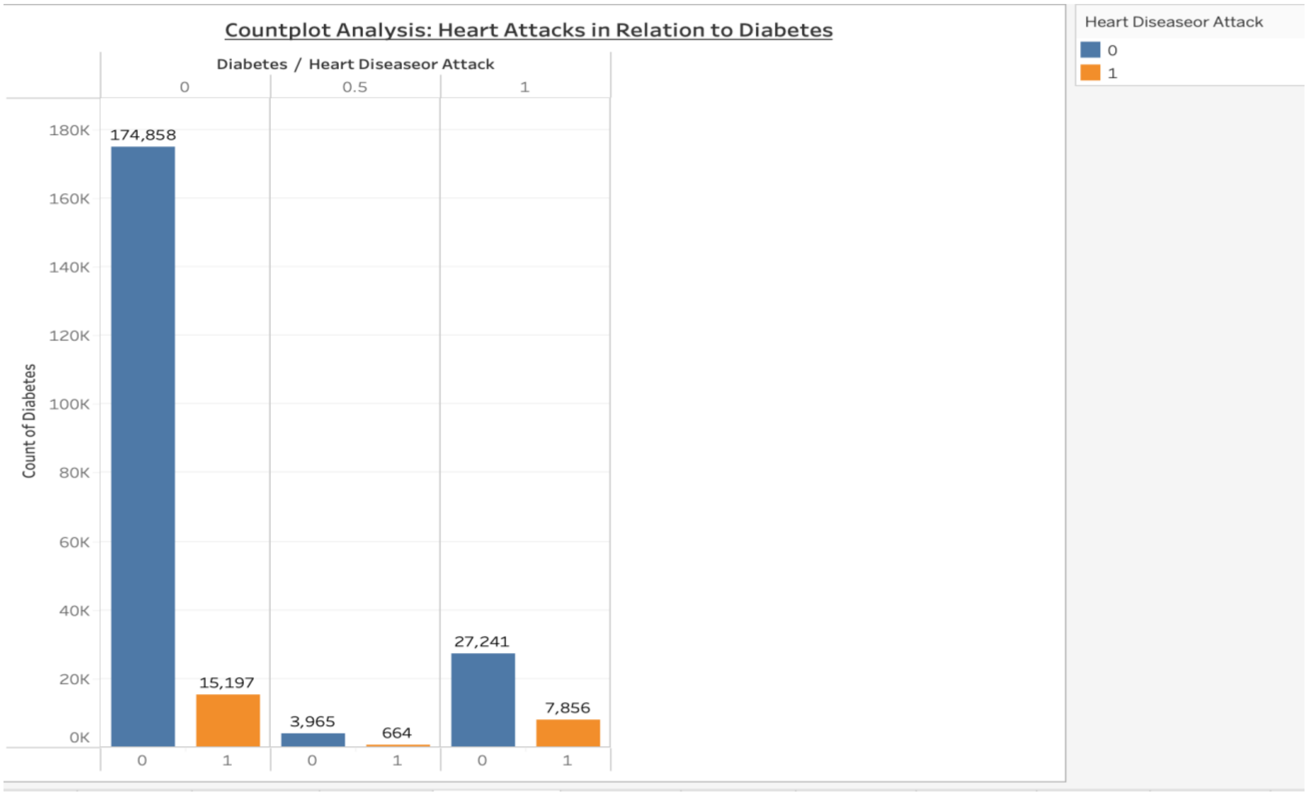
In the dataset, those with heavy alcohol consumption are fewer than those with moderate intake. Surprisingly, heavy alcohol consumption is associated with a lower prevalence of heart disease. The dataset suggests a potential protective effect linked to heavy alcohol consumption. However, the lower count of individuals with both heavy alcohol consumption and no heart disease implies a complex relationship. Further investigation is crucial to understand the nuanced interplay between alcohol consumption patterns and cardiovascular outcomes. These insights emphasize the need for comprehensive analyses considering various factors and provide a foundation for nuanced discussions on alcohol's intricate relationship with cardiovascular health.



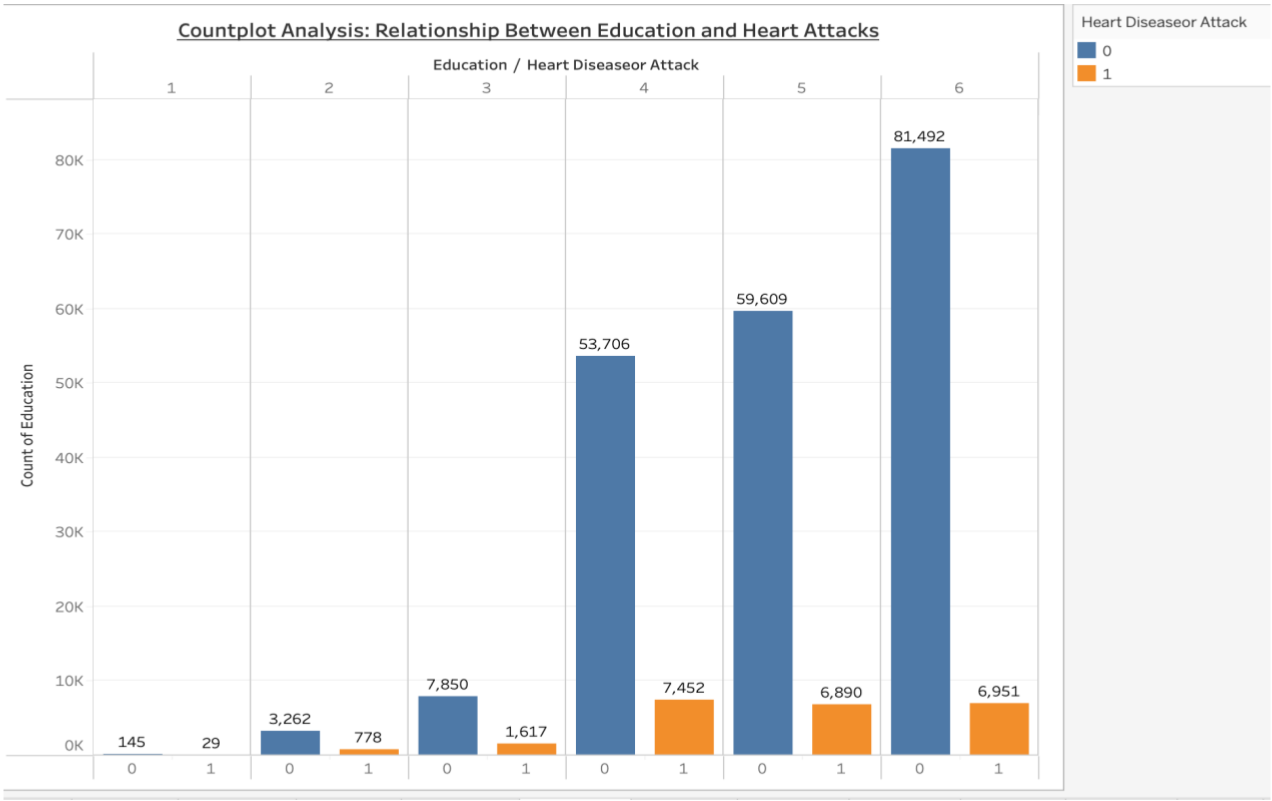
In the dataset, fewer individuals exhibit high cholesterol levels, but there's a striking association with an increased prevalence of heart disease. This underscores the potential significance of cholesterol levels in cardiovascular health outcomes, suggesting higher cholesterol may elevate susceptibility to heart-related issues. Monitoring and managing cholesterol levels emerge as crucial preventive measures against heart disease. The dataset encourages targeted interventions for those with elevated cholesterol, emphasizing lifestyle modifications and medical interventions. These insights pave the way for focused investigations and interventions to mitigate the impact of elevated cholesterol on cardiovascular outcomes.



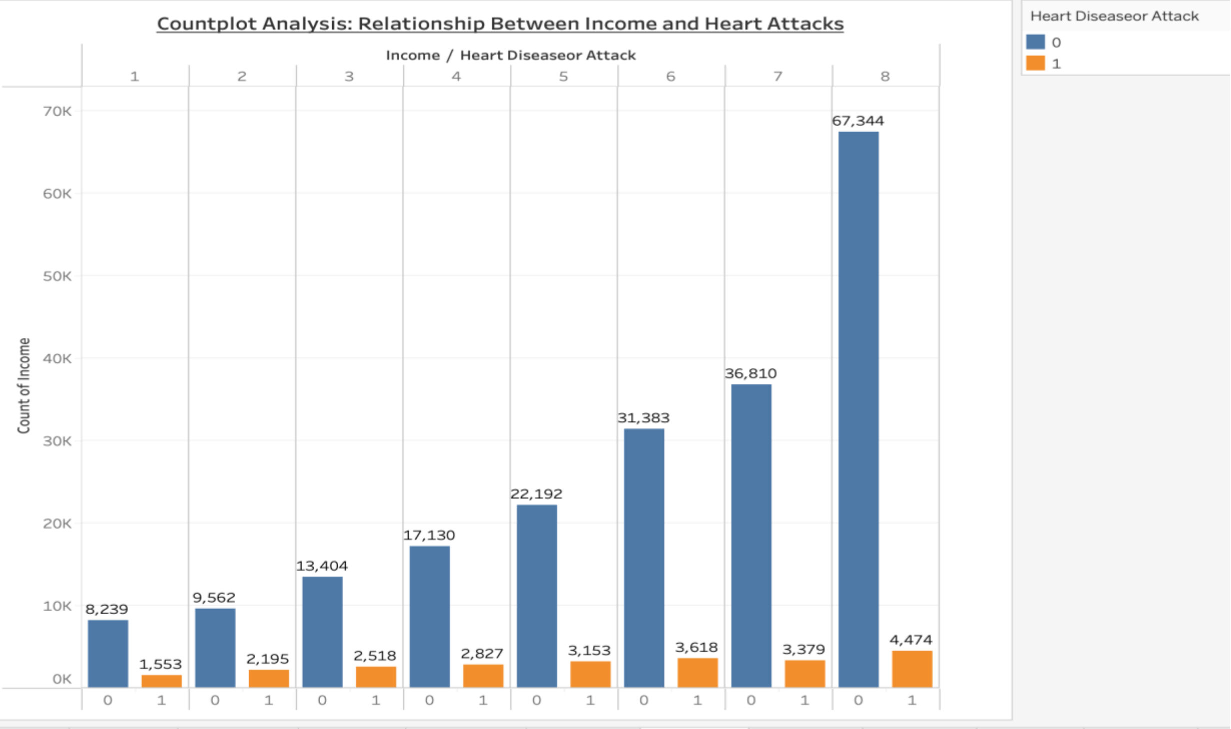
In the dataset, fewer individuals exhibit high blood pressure, but a distinct trend emerges—those with high blood pressure have a larger proportion with a history of heart disease. This underscores high blood pressure's potential significance in cardiovascular health outcomes, indicating an increased susceptibility to heart-related issues. Monitoring and managing blood pressure become crucial preventive measures against heart disease. The dataset encourages targeted interventions, focusing on lifestyle modifications and medical interventions tailored to manage blood pressure. These insights pave the way for focused investigations and interventions to mitigate the impact of elevated blood pressure on cardiovascular outcomes.



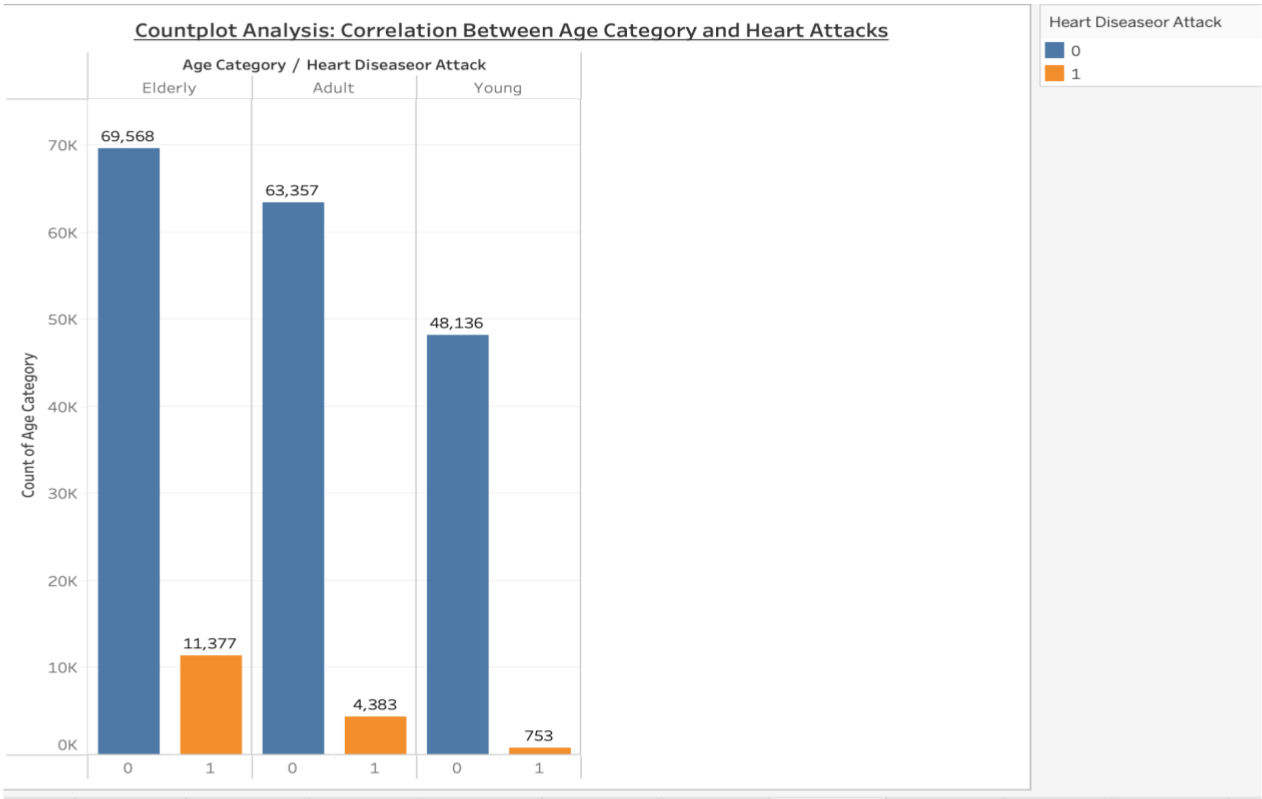
The dataset reveals a scarcity of individuals with diabetes, indicating a lower prevalence of this condition. Examining the relationship with heart disease unveils a larger proportion of individuals without diabetes having a history of heart disease. This emphasizes diabetes as a potential contributor to cardiovascular health outcomes. Individuals without diabetes seem to face a lower susceptibility to heart-related issues, making monitoring and managing diabetes crucial for heart disease prevention. The dataset calls for targeted interventions, including lifestyle modifications and glycemic control, to address the higher prevalence of heart disease observed in individuals with diabetes. These insights guide focused investigations and interventions to mitigate diabetes's impact on cardiovascular outcomes.



The dataset showcases a notable concentration of individuals in education level 6, indicating a higher prevalence of well-educated individuals. A distinct trend emerges, revealing an increase in heart disease frequency as education levels rise. This prompts considerations about the potential correlation between higher education and elevated heart disease risk. While causation isn't established, the association opens avenues for investigating socioeconomic and lifestyle factors contributing to this pattern. The dataset's six education levels enable a nuanced analysis, highlighting the importance of educational backgrounds in understanding health outcomes. Further exploration into characteristics associated with different education levels could unveil insights into the complex interplay between education, lifestyle, and cardiovascular health.



The dataset portrays a range of income levels, revealing a noticeable trend: as income increases, there is a corresponding rise in the likelihood of individuals having a history of heart disease. This prompts consideration of potential factors like stress and occupational demands influencing this pattern. The correlation suggests a complex interplay of socioeconomic factors impacting cardiovascular health, although causation isn't established. The dataset's multiple income levels allow for nuanced analysis, emphasizing the importance of socioeconomic factors in health outcomes. Deeper exploration into stressors and lifestyle attributes associated with varying income levels could provide valuable insights into the intricate relationship between income and cardiovascular health. In summary, the dataset suggests a potential association between income levels and heart disease, urging further investigation into underlying factors and providing context for understanding the complex dynamics involved.



The dataset prominently features individuals in the elderly category, indicating a significant presence of older individuals. An evident trend emerges, revealing a clear increase in the prevalence of heart disease with advancing age. This observation reaffirms the well-established correlation between age and elevated cardiovascular risk. As individuals enter older age brackets, the dataset suggests a corresponding rise in the likelihood of having a history of heart disease. Age, a recognized non-modifiable risk factor, is underscored in its importance for assessing and addressing cardiovascular risk. In summary, the dataset offers crucial insights into the relationship between age and heart disease likelihood, emphasizing the need for tailored interventions and healthcare strategies for different age groups.

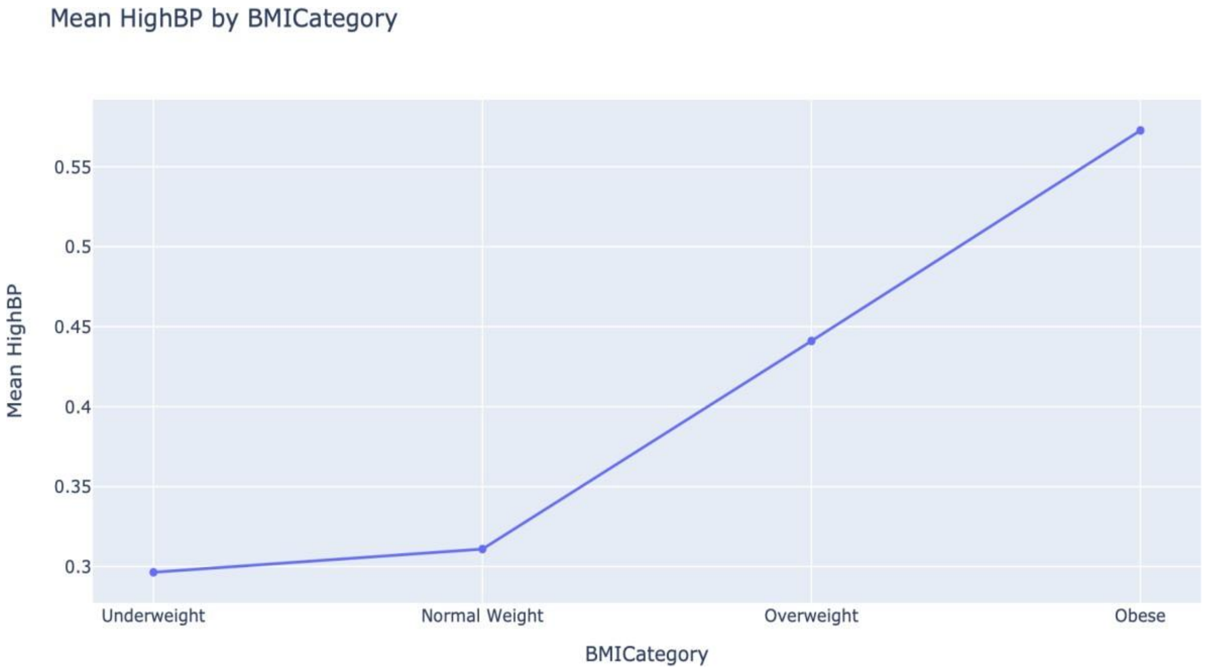
In the initial EDA we plotted:

1. Univariate plots: to check distribution of data.
2. Multivariate plots: to compare High BP, High Chol, Age, gender, smoker etc. columns with heart disease or attack column.

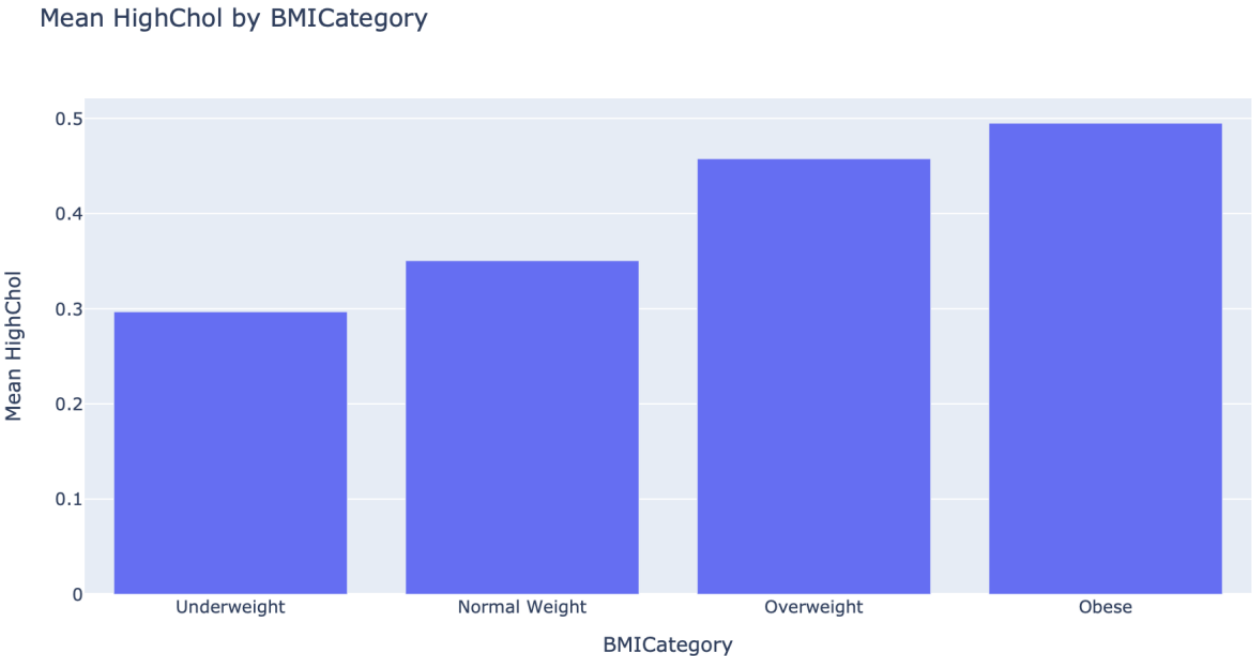
Final Exploratory Data Analysis:

In this EDA we will checking relationships between different factors and how they influence each other

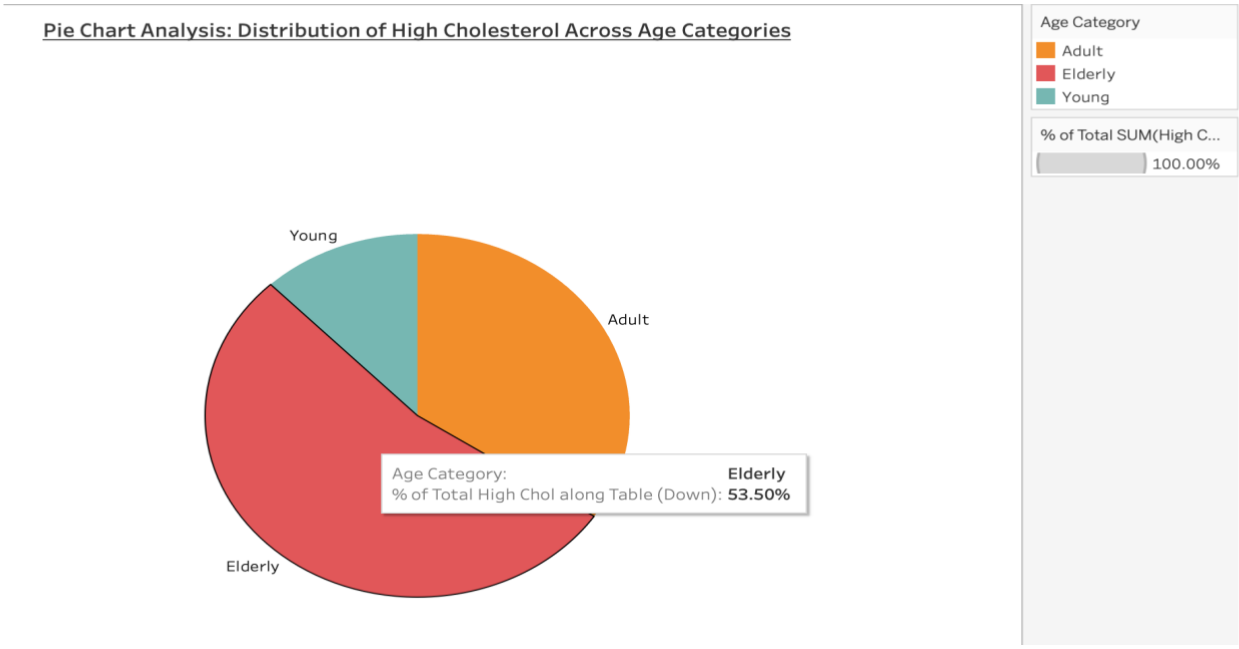
Bivariate Plots:



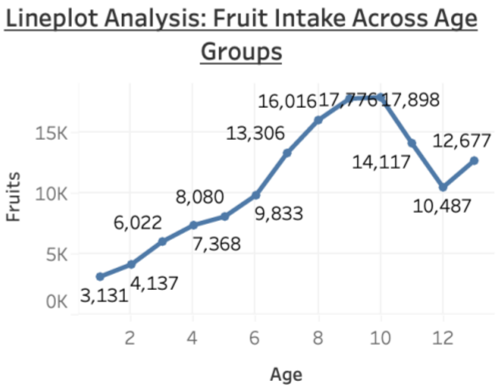
The bar chart effectively communicates the relationship between weight categories and high blood pressure levels, demonstrating a clear upward trend from underweight to obese individuals. This observation is significant as high blood pressure, a major cause of heart attacks, tends to escalate across varying weight categories, as depicted in the visual representation. The visual insight underscores the crucial connection between weight, high blood pressure, and the associated risk of heart attacks.

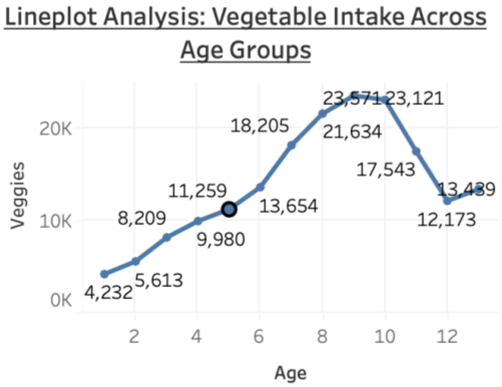


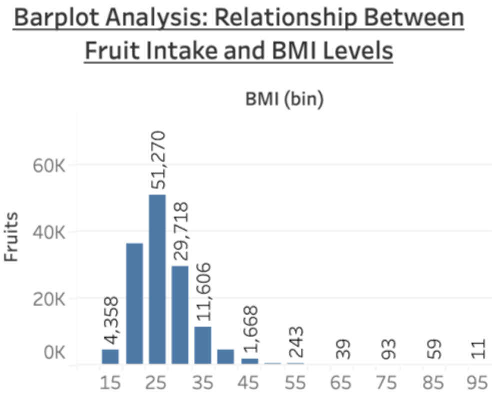
The straightforward bar chart visually conveys the association between weight categories and elevated cholesterol levels, highlighting that individuals ranging from underweight to obese tend to exhibit higher cholesterol levels. This observation gains significance, considering that higher cholesterol levels, a prevalent condition in these weight categories, represent a major contributor to heart attacks. The chart underscores the crucial link between weight, cholesterol levels, and the heightened risk of experiencing heart attacks.

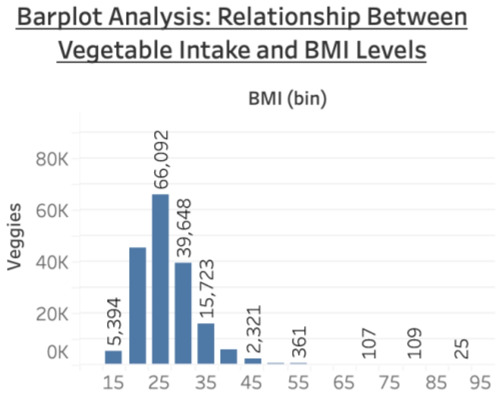


This pie chart vividly illustrates the distribution of elevated cholesterol levels across different age groups. Notably, the elderly population claims the largest share, emphasizing a higher prevalence of elevated cholesterol in this demographic. Following the trend, adults and young people exhibit progressively smaller portions, indicating a varying degree of cholesterol levels among age categories.


Lineplot Analysis: Fruit Intake Across Age Groups


Lineplot Analysis: Vegetable Intake Across Age Groups


Barplot Analysis: Relationship Between Fruit Intake and BMI Levels


Barplot Analysis: Relationship Between Vegetable Intake and BMI Levels

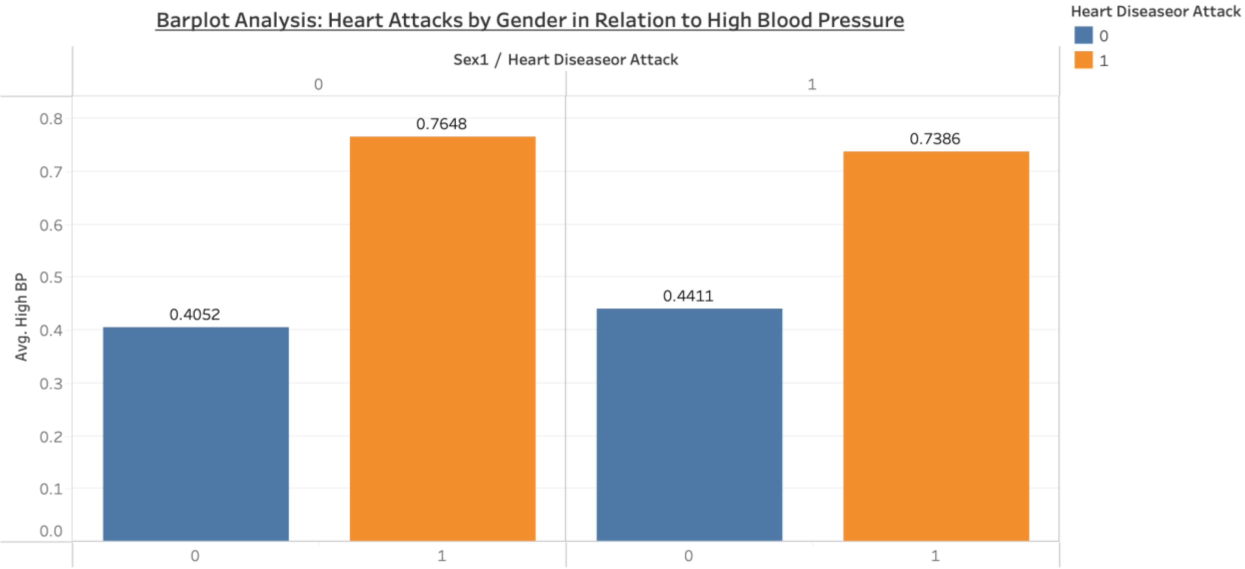
Fruits

A positive correlation with age is observed in fruit intake, showing an upward trajectory as individuals mature. Young children, particularly those under 2 years, have modest fruit intake. Adolescence sees a decline around ages 10-12, followed by a gradual resurgence from age 12 onward, highlighting age-specific variations in fruit consumption across different life stages.

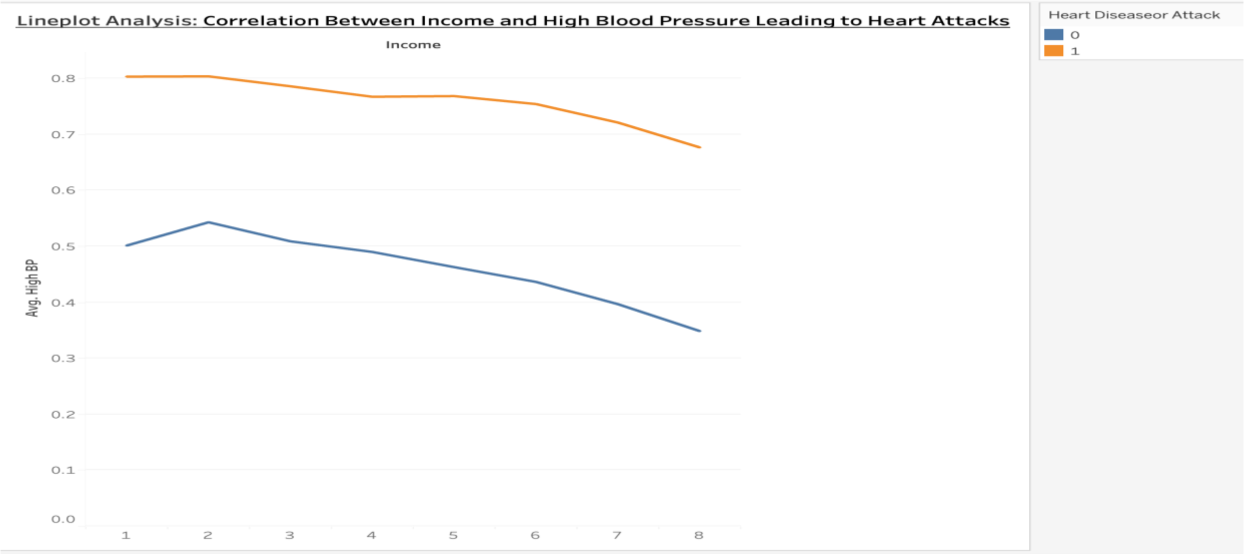
Veggies

A direct relationship with age is noted in vegetable intake, incrementally rising as individuals grow older. Young children, especially those under 2 years, have notably limited vegetable intake. Throughout adolescence, there's a decline around ages 10-12, followed by a gradual upturn from age 12 onward, emphasizing age-related fluctuations in vegetable consumption across distinct life phases.

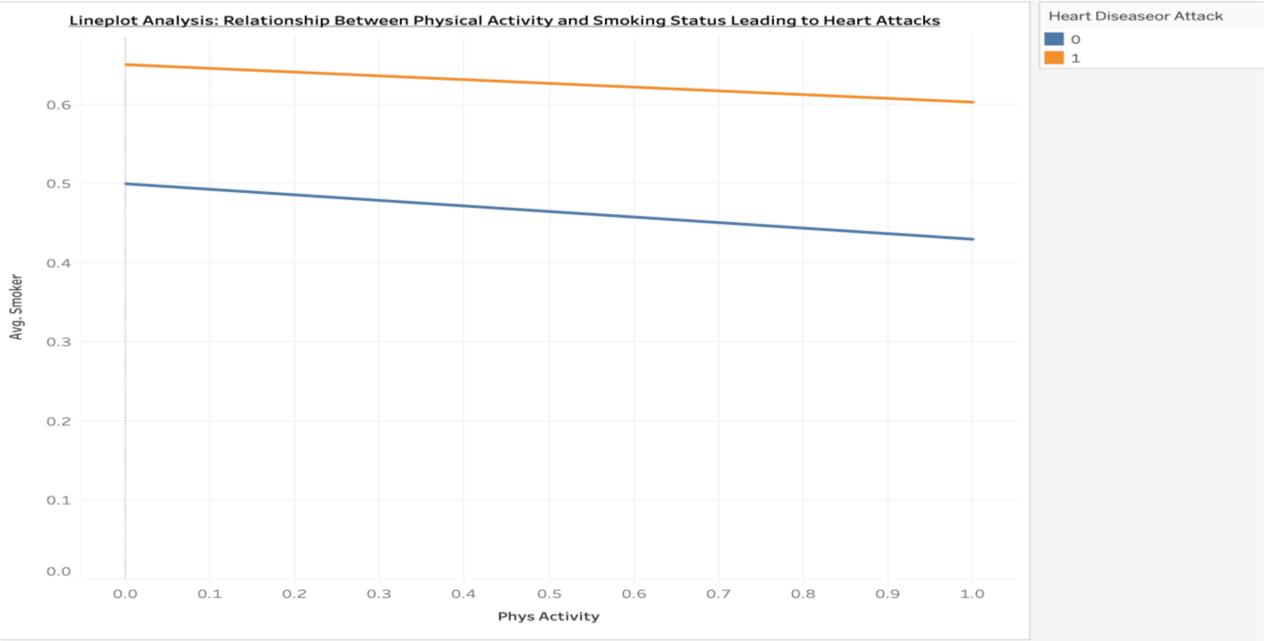
Multivariate Plots:



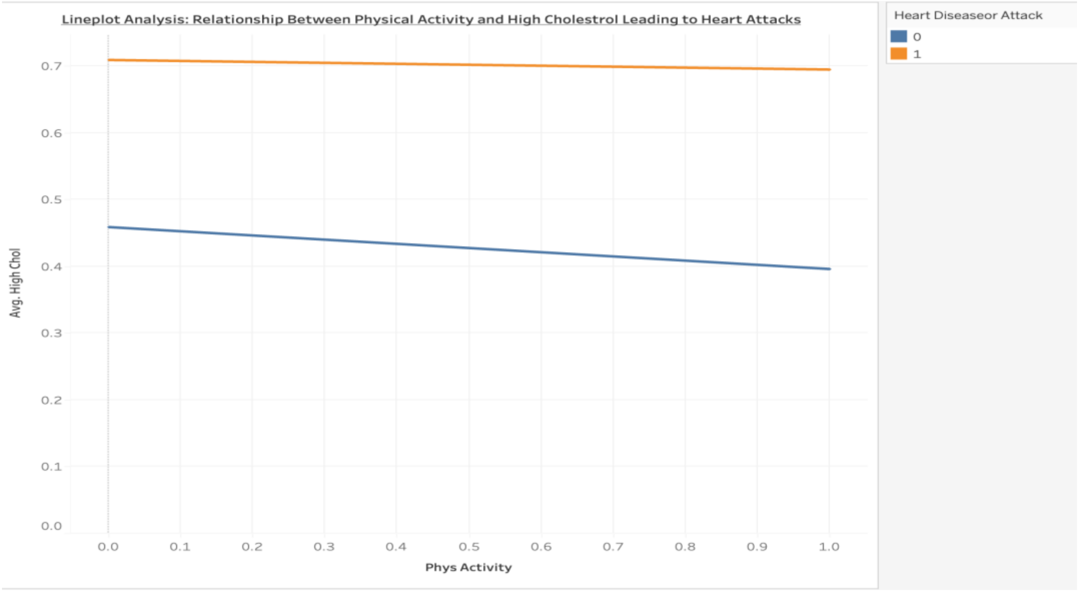
The analysis reveals that the risk of cardiovascular disease is directly linked to constant high blood pressure, irrespective of gender. However, women exhibit a higher susceptibility to the disease at the same blood pressure level compared to men. This observation emphasizes the nuanced gender-specific dynamics in cardiovascular risk, where high blood pressure plays a consistent role, but the impact differs between genders. Understanding these distinctions is crucial for tailored preventive strategies and healthcare interventions. The findings underscore the complexity of cardiovascular risk factors and advocate for gender-specific considerations in heart health assessments and intervention.



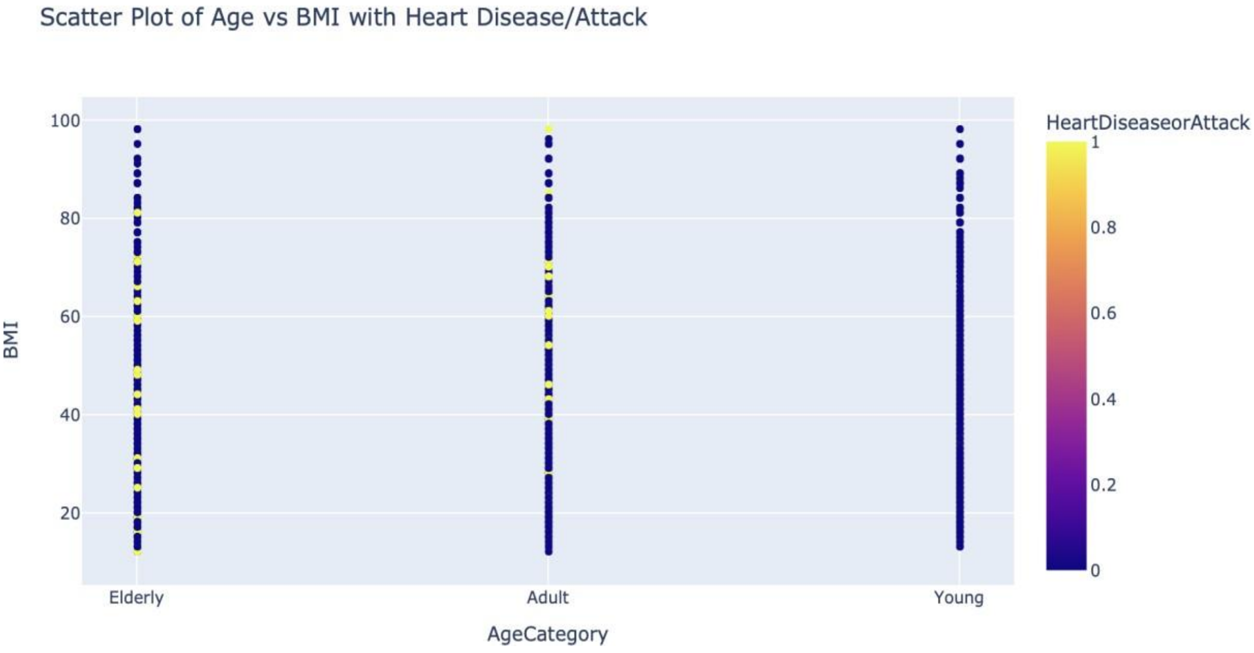
The graph depicts a downward trend, indicating that individuals with lower income who have high blood pressure tend to experience a higher prevalence of heart disease. This observation suggests a potential correlation between income levels, high blood pressure,and the risk of cardiovascular issues. The trend underscores the importance of socioeconomic factors in understanding heart health disparities, emphasizing the need for targeted interventions for individuals with lower income and elevated blood pressure.



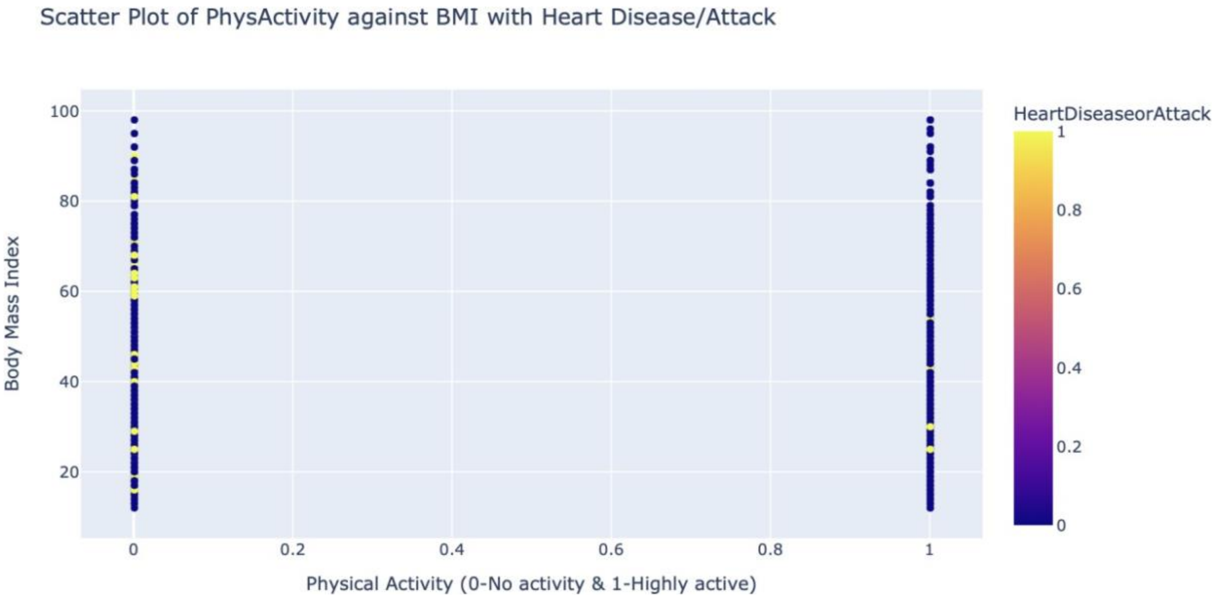
The line plot of Phys Activity against Smoker with Heart Disease/Attack reveals a distinct trend: as smoking increases, physical activity tends to decrease for individuals with and without heart disease. This observation underscores the potential interplay between smoking habits and physical activity in influencing cardiovascular health outcomes. The inverse relationship between smoking and physical activity levels suggests a complex interaction that may contribute to the risk of heart disease. These findings emphasize the importance of considering multiple lifestyle factors when assessing heart health, highlighting the need for targeted interventions that address both smoking and physical activity for comprehensive cardiovascular risk management.



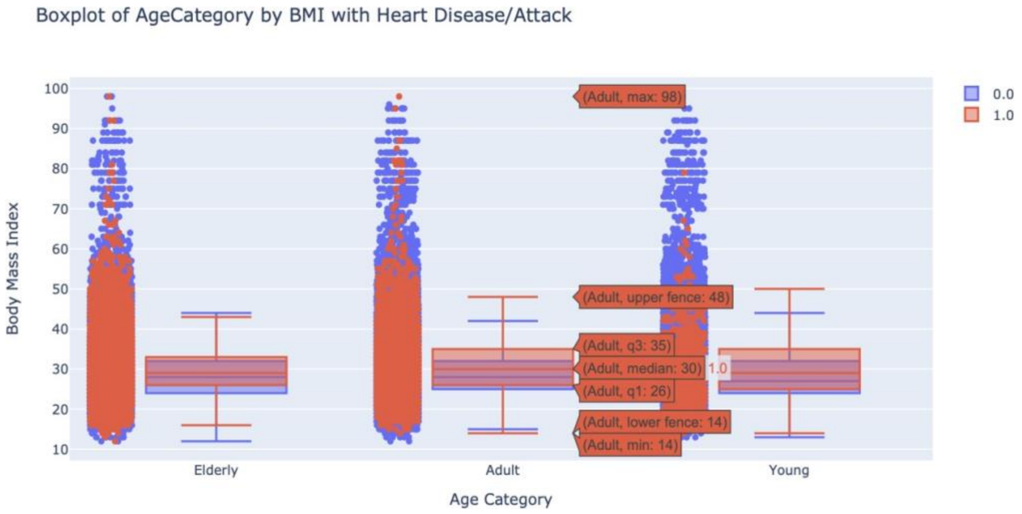
The line plot of PhysActivity against HighChol with Heart Disease/Attack illustrates a noteworthy trend: as physical activity increases, high cholesterol levels decrease for both individuals with and without heart disease. This observation suggests a potential protective effect of physical activity against elevated cholesterol, contributing to cardiovascular health. The inverse relationship emphasizes the role of regular exercise in managing cholesterol levels, a critical aspect in heart disease prevention. These findings underscore the importance of promoting physical activity as part of comprehensive cardiovascular risk reduction strategies. Tailored interventions focusing on increasing physical activity levels may offer significant benefits in mitigating the impact of high cholesterol on heart health.



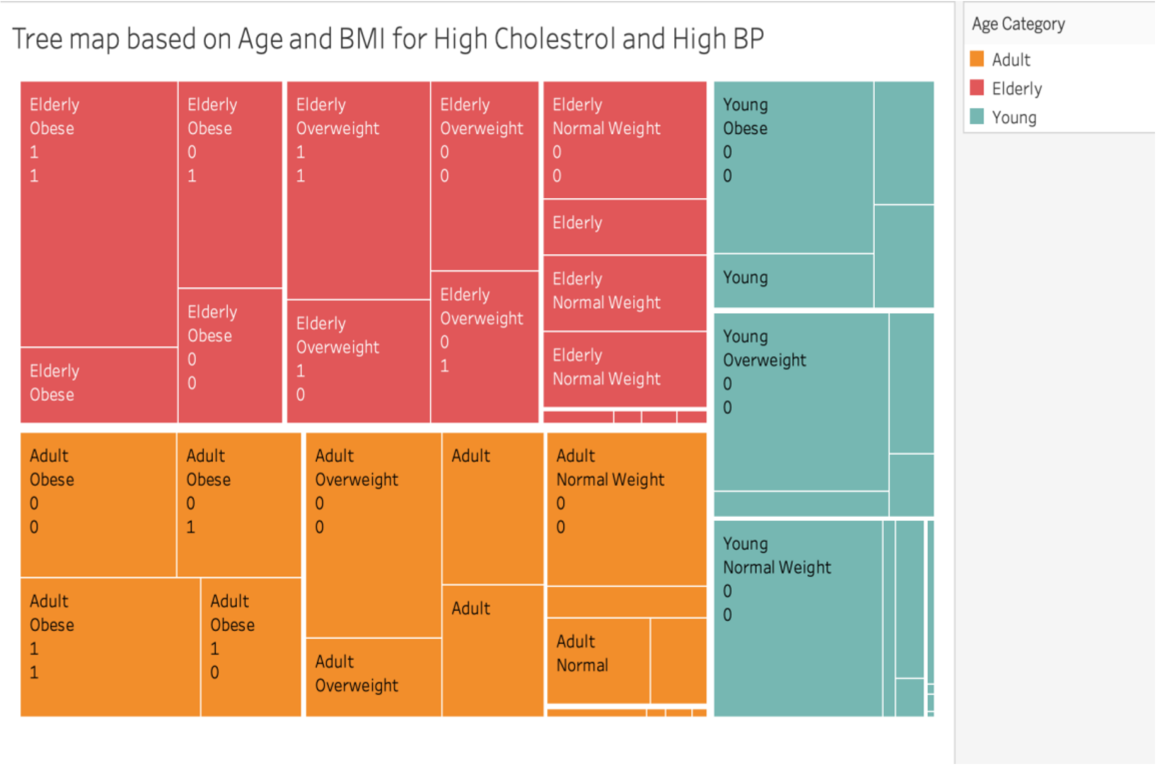
The scatterplot indicates a noticeable trend where elderly individuals show a higher susceptibility to heart disease. The data suggests that individuals in the "Adult" category also exhibit risks of heart disease, potentially influenced by factors like physical activity and lifestyle. This observation underscores the need for further visualizations to explore additional factors contributing to heart disease risk in the "Adult" category and emphasizes the distinct vulnerability of the elderly population. The scatterplot provides a preliminary insight, prompting a more comprehensive analysis of various risk factors for a holistic understanding of heart disease dynamics.



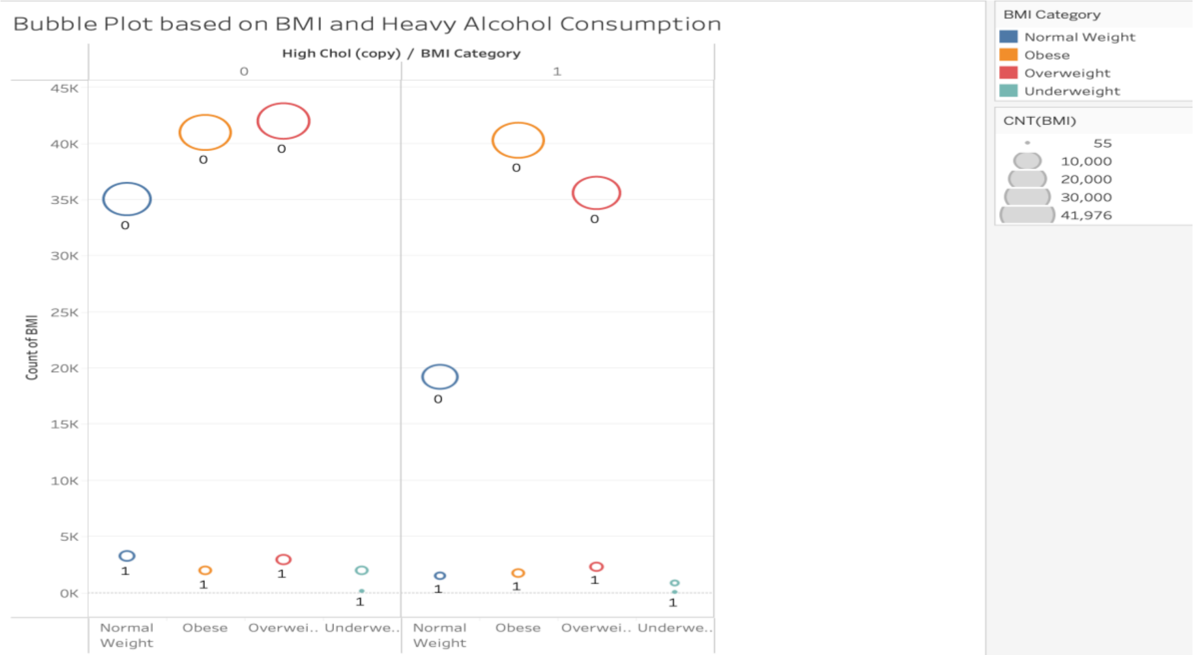
The scatterplot indicates a noticeable trend where elderly individuals show a higher susceptibility to heart disease. The data suggests that individuals in the "Adult" category also exhibit risks of heart disease, potentially influenced by factors like physical activity and lifestyle. This observation underscores the need for further visualizations to explore additional factors contributing to heart disease risk in the "Adult" category and emphasizes the distinct vulnerability of the elderly population. The scatterplot provides a preliminary insight, prompting a more comprehensive analysis of various risk factors for a holistic understanding of heart disease dynamics.



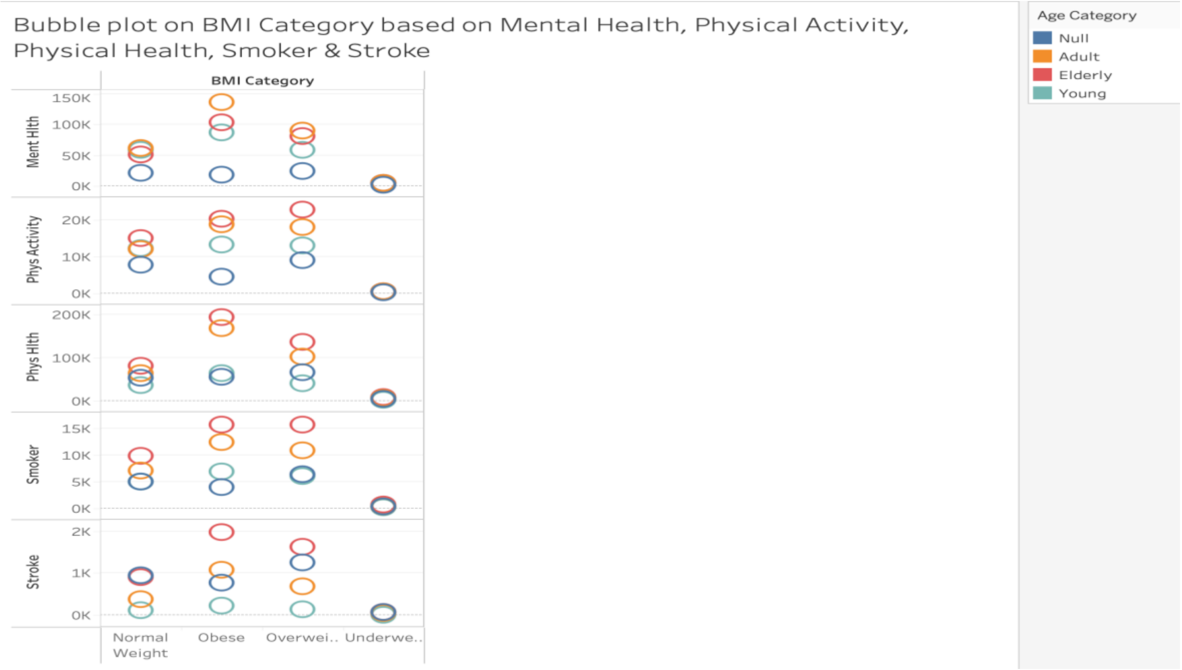
The lower median BMI observed in individuals with less or no risk of heart disease across age categories reinforces a direct correlation between BMI and cardiovascular risk. This underscores BMI's significance as an indicator, highlighting its role in assessing heart health across diverse age groups. The findings suggest that maintaining a lower BMI may be associated with a reduced risk of heart disease. This correlation emphasizes the importance of weight management in cardiovascular health considerations and supports the notion that BMI can serve as a valuable marker for assessing heart disease risk.



The tree map plots all the major factors against each other and re-affirming the fact that Age, BMI, Cholesterol and BP are directly proportional in causing the cardiovascular disease. Bigger the size of the box shows the higher number of people in that category. For instance, people in the elderly category who are obese, with high Cholesterol and BP (red category) tend to be more prone to disease rather than people in the yellow or the green category.

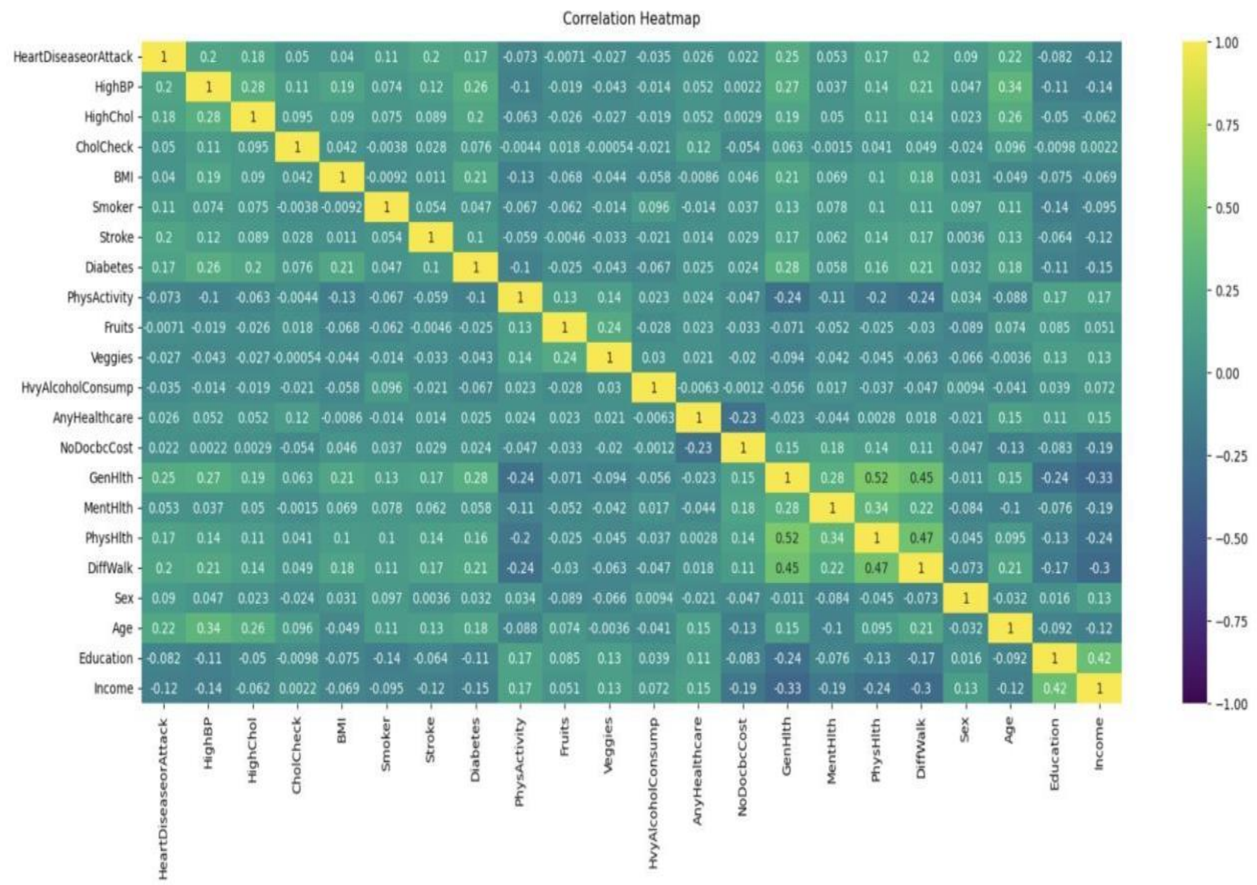


For the bubble chart above is plotted to find out how alcohol consumption with the couple of major factors react, taking into consideration BMI and Cholesterol. In people with the BMI category of "Normal weight", it is found that nature and complexity of data which is under consideration. It can be found that the alcohol consumption is inversely proportional from the other major factors.



In the advanced EDA we have plotted visuals between different factors which influence heart disease. We have also used some complex plots in our advanced analysis like bubble plot, tree map, heat map etc.

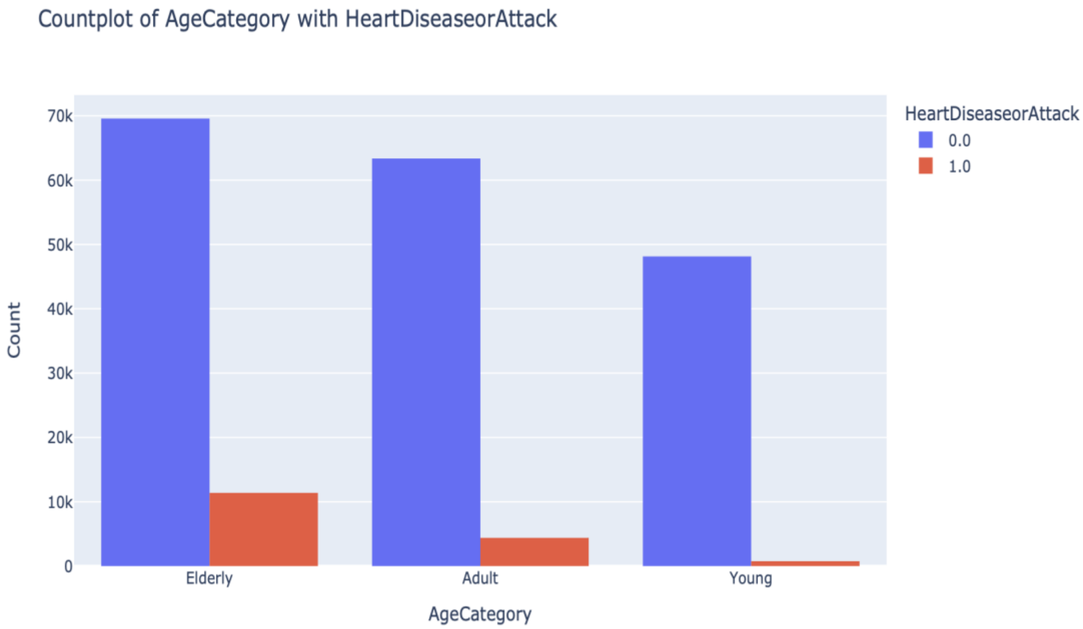
Correlation Matrix:



The likelihood of developing heart disease is intricately linked to various factors, among which a noteworthy contributor is BMI. Extensive analysis has revealed a robust positive correlation between BMI and the potential outcome of heart disease, emphasizing the significant impact of body mass index on overall cardiovascular health. Additionally, age emerges as a pivotal determinant, demonstrating a direct proportional relationship with the risk of heart disease. As individuals age, the likelihood of encountering this cardiovascular condition increases, marking age as a crucial factor in the health equation. Furthermore, the state of physical health plays a vital role, exhibiting a direct correlation with the probability of heart disease. These collective findings underscore the intricate interplay of BMI, age, and physical health in shaping the risk landscape for heart disease. Recognizing and understanding these correlations provides valuable insights for proactive health management, enabling informed preventive measures to mitigate the potential impact of cardiovascular issues.

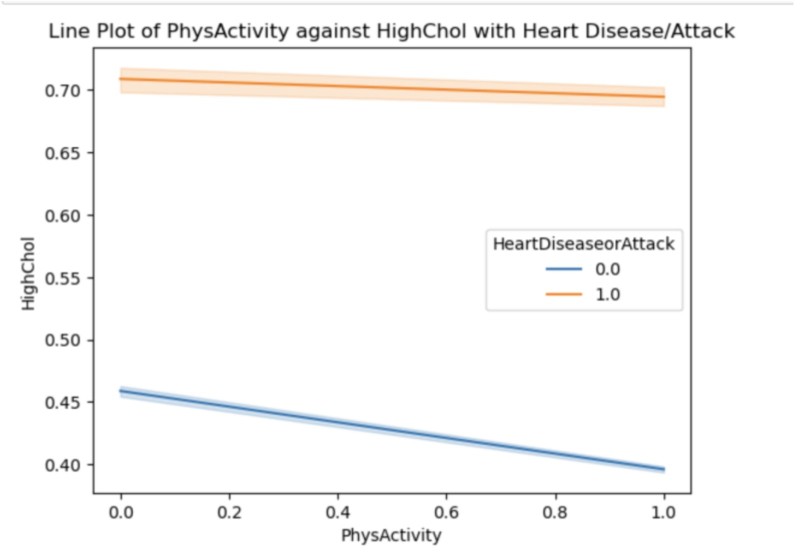
Visualization techniques used:

1. Bar Charts:



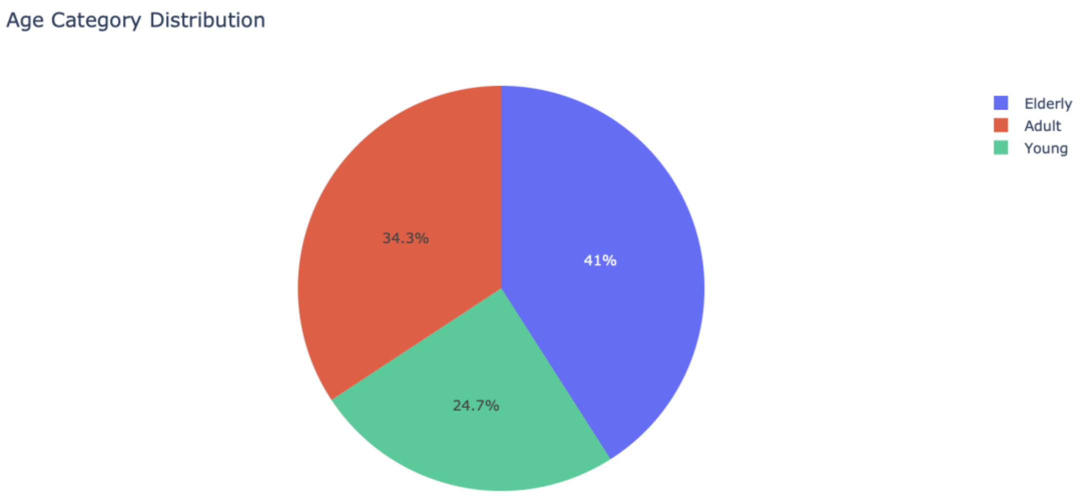
Bar Charts are an effective way to visualize and compare data across different categories. They are particularly useful for displaying qualitative data, such as the frequency of different types of risk factors or the distribution of risk factors among different populations. By presenting data in a clear and concise manner, bar charts can help identify patterns and trends that might not be immediately apparent from raw data. Count plots helps to determine the frequency of different risk factors such as High BP, High Cholesterol, BMI Category, Age Category, etc.

2. Line Charts:



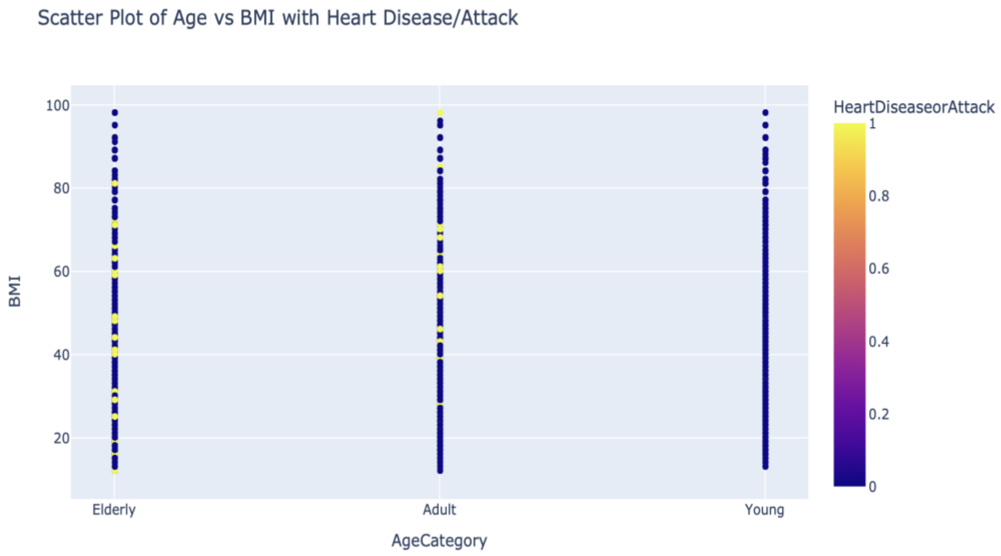
Line Charts are an effective way to visualize trends and changes over time. They are particularly useful for displaying quantitative data, such as trend of different risk factors over time. Line plots aid in illustrating trends and variations over a continuous scale, facilitating the interpretation of correlations between Physical Activity and High Cholesterol levels.

3. Pie Charts:



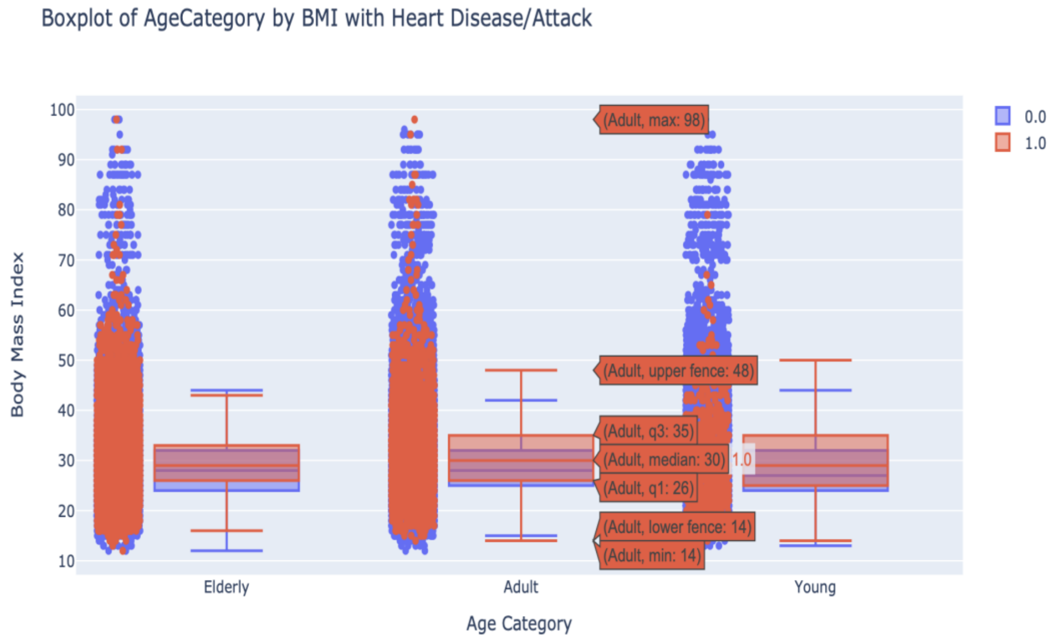
Pie charts are useful for visualizing data because they can help to show the relationship of different parts to the whole. We opt for pie charts due to the dataset's small number of categories, offering a clear representation of elevated cholesterol proportions in different age groups.

4. Scatter Plots:



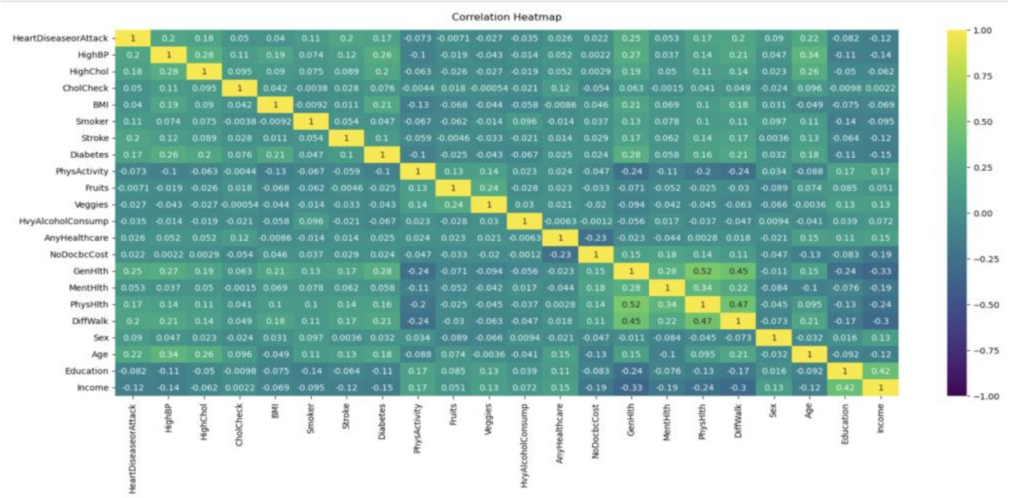
Scatter plots are the best way to observe the correlation between 2 different risk factors. We will use them to understand the relation between Age Category and BMI, Physical Activity and BMI., with respect to Heart Disease Attack. It provides more comprehensive analysis.

5. Box Plots:



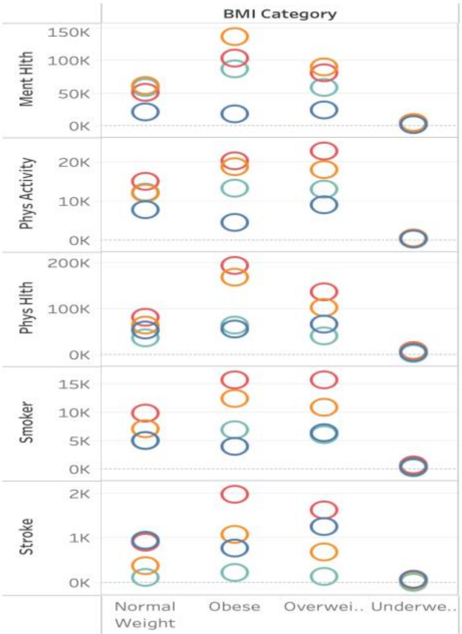
Box Plots are useful for visualizing the distribution of data. Additionally, it provides outlier detection using IQR. It helps in illustrating distribution of Age Category by BMI with respect to Heart Disease Attack. They are valuable for identifying patterns and correlations that may not be immediately apparent from the raw data.

6. Heatmaps:



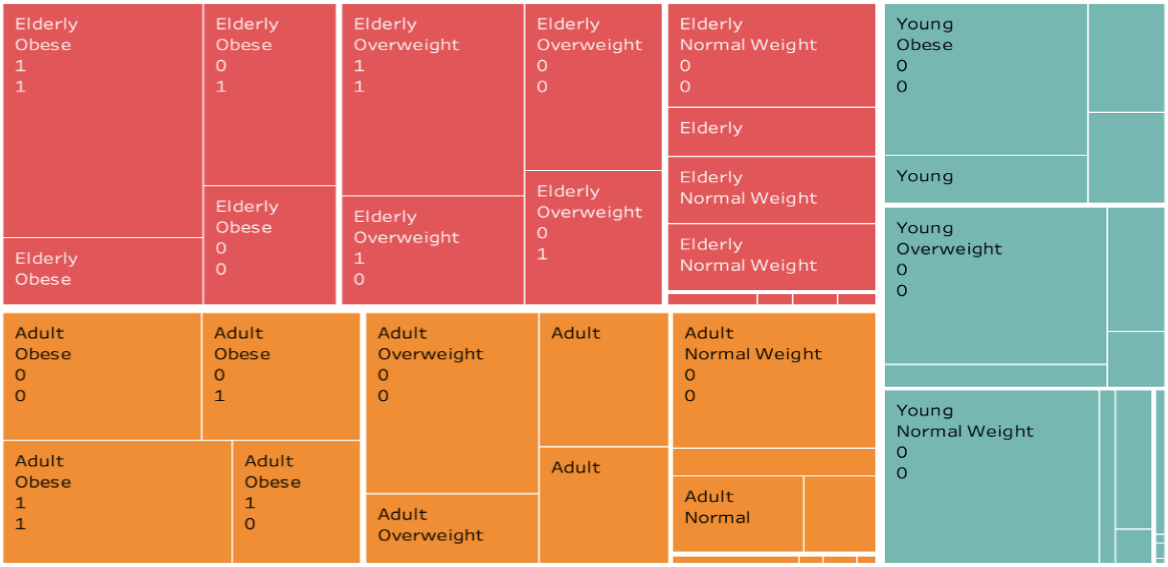
Heatmaps are useful because they provide visual appealing representation of the factors using color-coded intensity. They are an excellent way to show the correlation between different risk factors. Heatmaps are particularly useful for uncovering patterns, trends, and relationships within the data by highlighting areas of high or low values. They allow for quick and intuitive interpretation of the data, enabling to identify clusters, gradients, or anomalies. It helps us in visualizing connections between different variables and heart disease, offering a comprehensive perspective on data relationships.

7. Bubble Plots:



Bubble plots are employed in data visualization to effectively represent three variables simultaneously in a two-dimensional space. These plots integrate the features of a scatter plot with an additional dimension represented by the size of markers, or "bubbles." This visual representation is particularly useful when you want to convey the magnitude of a third variable, with larger bubbles indicating higher values. The clear and concise nature of bubble plots makes them suitable for comparing multiple data points at once, enabling the identification of patterns, trends, and potential outliers. Their enhanced readability in a 2D space, along with the option for interactive features, such as hovering and clicking on bubbles for additional information, makes bubble plots a versatile tool for conveying complex datasets and facilitating a deeper understanding of correlations and relationships within the data.

8. Tree Map



A tree map is a visualization that uses nested rectangles to represent hierarchical data. Each rectangle corresponds to a category, and its size reflects a quantitative value. This approach efficiently communicates the proportions and distribution of values within a hierarchy, making tree maps useful for visualizing structures like file directories or financial portfolios. They are widely applied in data analysis and business intelligence to highlight patterns in hierarchical relationships.

Key Findings:

In our analysis, it was revealed that around 10.3% of individuals had a documented history of heart disease or experienced a heart attack, while most of the population did not present such medical conditions. Furthermore, a mere 4.8% of the sampled population had a past occurrence of stroke, leaving the majority without any indications of stroke.

Major Factors Impacting Cardiovascular Health:

1. **Higher BMI:**
   Individuals with higher BMI, particularly in the obese category, show a significant association with cardiovascular issues, emphasizing the importance of weight management.

2. **Smoking:**
   Smoking is identified as a major factor with a slightly higher likelihood of

experiencing heart disease, underlining the necessity of smoking cessation programs for cardiovascular health.

3. **High Cholesterol:**
   A significant association exists between high cholesterol levels and an increased prevalence of heart disease, highlighting the importance of cholesterol monitoring and control in cardiovascular care.

4. **High Blood Pressure:**
   High blood pressure is a major factor, with individuals in this category exhibiting a larger proportion with a history of heart disease. Effective blood pressure management becomes crucial in preventing cardiovascular events.

5. **Age:**
   Age is a crucial factor, with a clear increase in the prevalence of heart disease observed with advancing age, particularly in the elderly category. Tailored healthcare strategies for older individuals may be essential.

Minor Factors Impacting Cardiovascular Health:

1. **Gender Disparities:**
   Males demonstrate a heightened susceptibility to heart attacks compared to females, indicating a gender-based trend. This insight emphasizes the need for gender-specific approaches in cardiovascular health research and interventions.

2. **Physical Activity:**
   Individuals engaged in regular physical activity show a slightly higher susceptibility to heart disease, presenting a noteworthy but minor finding. Balancing physical activity with other lifestyle factors is key for overall cardiovascular well-being.

3. **Education Level:**
   A correlation is observed between higher education levels and elevated heart disease risk, suggesting a potential minor influence. Further investigation into the link between education and cardiovascular health is warranted.

4. **Income Levels:**

As income increases, there is a corresponding rise in the likelihood of individuals having a history of heart disease, indicating a minor impact. Socioeconomic factors may contribute to variations in cardiovascular health outcomes.

5. **Nutrition:**
   Correlations are observed between age and fruit/vegetable consumption, suggesting increased awareness and healthier habits among older individuals, with a minor impact. Dietary patterns play a role in cardiovascular health but should be considered alongside other factors.

Factors Not Significantly Impacting Cardiovascular Health:

1. **Diabetes:**
   The dataset indicates a lower prevalence of diabetes, and it does not seem to have a significant impact on the history of heart disease. While diabetes is a concern, its role in cardiovascular outcomes in this dataset appears less pronounced.

2. **Heavy Alcohol Consumption:**
   Surprisingly, heavy alcohol consumption is associated with a lower prevalence of heart disease, contrasting with the expected impact on cardiovascular health. The complex relationship between alcohol and heart health requires further exploration and consideration of potential confounding factors.

Conclusion:

In conclusion, the analysis reveals a multifaceted landscape of factors influencing cardiovascular health, with some playing major roles, others contributing in minor ways, and a few showing unexpected patterns. Higher BMI, smoking, high cholesterol, high blood pressure, and age emerge as major players, emphasizing their significant impact on the prevalence of heart disease. These findings underscore the critical importance of weight management, smoking cessation, cholesterol monitoring, blood pressure control, and tailored healthcare for the elderly in cardiovascular care. Meanwhile, minor factors such as gender disparities, physical activity, education level, income, and nutrition provide additional nuances to the understanding of cardiovascular health, necessitating a balanced approach to lifestyle interventions. Surprisingly, the unexpected association between heavy alcohol consumption and a lower prevalence of heart disease calls for further exploration, highlighting the complexity of alcohol's role in cardiovascular health. Overall, this comprehensive analysis provides valuable insights for informed preventive measures and underscores the need for ongoing research to refine our understanding of these intricate relationships in cardiovascular health.

Limitations:

While the dataset offers valuable information, certain limitations present opportunities for further exploration and analysis.

Binary Data Representation:

Challenge: Many attributes are represented solely by binary values (0 and 1), limiting the clarity of data visualization, and hindering our ability to identify nuanced patterns and relationships.

Opportunity: Explore feature engineering techniques to expand the data representation and unlock its full potential. This could involve creating new features based on existing ones, utilizing techniques like binning or clustering to group binary values, or incorporating external data sources to enrich the representation.

Limited Population Representation:

Challenge: The dataset only includes a small percentage of individuals with heart conditions, potentially skewing the results and impacting the generalizability of our findings.

Opportunity: Leverage data augmentation techniques to artificially increase the representation of under-sampled groups. This could involve oversampling minority classes, utilizing synthetic data generation techniques, or employing transfer learning approaches trained on other datasets.

By addressing these limitations, we can unlock the full potential of the dataset and gain a deeper understanding of the factors driving heart conditions. This will ultimately enable us to develop more robust and generalizable models that can be applied effectively to a wider population.

Challenges:

While undertaking our analysis, we encountered several challenges that required innovative solutions to ensure accurate and insightful findings.

1. **Age Categorization:** The initial age range, spanning from 1 to 10, posed a challenge for interpretation and analysis. To address this, we implemented a three-category system: "young," "adult," and "elderly." This transformation transformed abstract numerical values into meaningful categories, facilitating the identification of age-related patterns within the data.

2. **BMI Categorization:** Initially presented as numerical values, the BMI column underwent a transformation into categorical values: "Underweight," "Normal Weight," "Overweight," and

"Obese." This categorization simplified interpretation and provided a clearer representation of weight status within distinct categories. This conversion enhanced the analytical value of BMI by offering a more readily understandable categorization.

3. **Gender Recoding:** The initial representation of "Sex" as binary values (0 and 1) presented a challenge to clarity. To address this, we implemented a more intuitive coding system, with 0 representing "female" and 1 representing "male." This revision facilitates a straightforward understanding of gender distinctions within the data, enhancing the analytical value of the "Sex" attribute.

By overcoming these challenges through thoughtful data transformations, we ensured the accuracy and accessibility of our analysis. These modifications provided a more insightful perspective on the data, enabling us to extract meaningful patterns and relationships that would have been obscured by the original format.

Future Work:

Alcohol & Diabetes: Future research endeavors should prioritize a detailed investigation into factors seemingly not significantly impacting heart health based on current findings, such as diabetes and the unexpected association of heavy alcohol consumption with lower heart disease prevalence. A comprehensive study, including longitudinal analyses and clinical trials, is essential to unveil the nuanced dynamics and potential underlying mechanisms. This focused research will contribute to a more precise understanding of these factors and aid in developing targeted strategies for cardiovascular risk prevention.

Health metrics categorized: The 'Physical Health' and 'Mental Health' columns, initially ranging from 0 to 30, are transformed into categorical data for improved interpretability. This categorization facilitates a more intuitive understanding of health metrics and simplifies the analysis of physical and mental well-being. The conversion provides a clearer representation of health statuses within distinct categories.