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:

Serial No.	Title	Page No.
1	Project Proposal	4
1.1	Overview of the project topic	4
1.2	Objectives and expected outcomes	5
1.3	Proposed methodology	5
2	Data Preprocessing	6
2.1	Data acquisition and sourcing	6
2.2	Steps taken for cleaning and preparing the data.	9
2.3	Challenges faced and how they were addressed	14
3	Initial Exploratory Data Analysis (EDA)	16
3.1	Initial findings and patterns observed.	16
3.2	Visualization techniques used for initial insights.	16
3.3	Interpretation of the initial analysis	17
4	Final Exploratory Data Analysis (EDA)	31
4.1	Advanced analysis and deeper insights	31
4.2	Application of Complex and Simple Visualization Techniques	42
4.3	Key findings	46
	Conclusion	48
	Limitations	49
	Future Work	50

1. Project Proposal

1.1. Project Overview:

In today's rapidly advancing society, we can forecast the future and develop AI technology, yet the ultimate challenge remains: minimizing fatalities caused by heart diseases. This category of illnesses, known as cardiovascular diseases (CVDs), directly impacts the heart and surrounding blood vessels and can greatly compromise one's health and overall quality of life. It is vital to understand key facts and factors related to CVDs and their potential consequences for individuals.

High Mortality Rates: It's a grim reality that cardiovascular diseases continue to reign as the leading cause of death worldwide. In the US alone, the statistics are even more alarming - with one CVD-related death occurring every 30 seconds, as reported by the 2021 survey conducted by the Centers for Disease Control and Prevention (CDC).

Risk Factors: A lack of physical activity, elevated blood pressure, smoking, elevated cholesterol levels, and diabetes are among the many factors contributing to the onset of cardiovascular diseases.

Age Factor: As individuals age, they become more vulnerable to a potentially serious cardiovascular condition that poses a risk to their lives.

Financial Burden & Background: Managing cardiovascular illnesses can incur significant expenses for both patients and healthcare systems. The costs of medications, medical procedures, and hospital stays can create a substantial financial burden.

Accessibility: Unfortunately, a significant portion of the population remains unable to access essential healthcare services due to a lack of facilities equipped with the necessary resources to effectively address their needs.

Psychological Effects: The detrimental impact of stress and depression on one's mental health can also have serious consequences for cardiovascular health. A multitude of reasons can contribute to the development of mental health issues, such as lifestyle modifications, uncertainty about the situation, and concerns about the future.

Complications: A rich medical history including heart attacks, heart failure, and strokes are just some of the consequences that can stem from cardiovascular disorders. This puts individuals with these conditions at a higher risk of experiencing exacerbating health challenges. Extensive research by the World Health Organization has shown that once someone falls victim to these diseases, their chances of survival are bleak, as they trigger a domino effect of ailments that rapidly deteriorate overall health. As a result, our paramount objective is to pinpoint and tackle the major contributing factors that heighten the risk of developing CVDs. We aim to offer practical solutions that can effectively reduce these risks and provide appropriate treatments to prevent untimely deaths.

1.2. Objectives and Expected Outcomes:

The primary goal of this research is to gain insights into the prevalence of cardiovascular diseases (CVDs) and their effects on individuals, as well as potential implications for healthcare systems. We aim to evaluate the prevalence of CVD within various categories, recognize key risk factors and trends, and develop prevention strategies. Despite advancements in the medical field, CVDs continue to pose a considerable global health challenge, resulting in significant mortality rates and financial burdens.

Key Findings:

- Complexity of CVDs
- Major and Minor factors impacting CVD.

Implications:

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

1.3. Project Methodology:

The CVD risk analysis project adheres to standard and proven project management practices, and these include the following:

1. **Data reading and Understanding**: During the preliminary stage, the data is meticulously examined and comprehended from its original source. From there, we proceed to establish the objectives or ultimate findings of the study.
2. **Data Cleaning/Pre-processing:** During this stage, the dataset undergoes thorough cleaning and preparation for subsequent processing and visualization. Below, you will find a comprehensive description of the processing methods used.
3. **Approach and Methods**: In this phase, we focus on fine-tuning the techniques that best suit the specific characteristics and nature of the data. This crucial step ensures that our methods are accurately tailored to optimize the outcome.
4. **Data Visualization & Analysis**: The analysis effectively employs a variety of charts and visualization techniques that align with the specific characteristics of the data.
5. **Conclusion, Improvements & Lessons learned**: In this report, we provide a concise overview of the primary discoveries, contemplate the overall involvement in the project, and highlight potential avenues for enhancing future projects.

2. Data Preprocessing

2.1. Data Sources

The dataset being analyzed is sourced from Kaggle and is titled the cardiovascular disease dataset. The corresponding link is provided below:

CVD dataset

CVD dataset notebook

Data Overview:

The particular issue being investigated involves identifying individuals at a higher risk of developing cardiovascular diseases (CVDs) by analyzing existing data points such as physical exercise, dietary patterns, and the use of psychoactive drugs. The relevant data points are outlined 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 Centers for Disease Control and Prevention (CDC) conducts the Behavioral Risk Factor Surveillance System (BRFSS), an annual telephone survey. This survey, initiated in 1984 and carried out consistently each year, gathers information from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and preventive treatment utilization. The dataset utilized for this study comprises 22 elements, including 253,680 cleaned responses from the BRFSS 2015 survey. This data is predominantly employed for binary categorization related to heart disease and is primarily composed of Boolean values. A detailed description of the dataset is provided below.

Response Variable:

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

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 Mental health 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 and 6 is the highest education level)
- Income - Contains the values from 1-8, where 1 is the low-income level and 8 is the highest income level)

2.2. Data Preprocessing:

Data preprocessing is a crucial step within the data mining process, involving tasks such as cleaning, transforming, and merging raw data to create a structured format suitable for subsequent analysis, modeling, and machine learning. The objective of data preparation is to enhance data quality and ensure its suitability for thorough analysis.

Data preparation procedures encompass a range of tasks, including cleaning, instance selection, normalization, one-hot encoding, data transformation, feature extraction, and feature selection.

Listed below are the procedures carried out for data cleaning and preprocessing:

Step 1: Checking null/missing values:

In order to guarantee the cleanliness and readiness of the data for analysis, it is essential to examine the dataset for any null or missing values. Nonetheless, no missing values were detected 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 2.2.1- Checking null/missing values in the dataset

Step 2: Checking duplicate values:

During the data cleaning process, it is crucial to verify the absence of duplicate values in the dataset, as they can result in inaccurate analyses and biased outcomes. In our dataset, we identified 23,899 duplicate records and subsequently removed them.

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

```
23899
```

Fig 2.2.2 Checking for duplicate records

Step 3: Data Transformation:

Data transformation involves converting data into a format that is more suitable for analysis. Once the data has undergone cleansing, techniques like "data normalization" and "scaling" can be applied. Data normalization, specifically Min-Max scaling, adjusts the data to a predefined range, mitigating bias arising from varied units of measurement. In order to ensure data consistency, various numerical features were scaled to fall within the range of 0 to 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 encompasses the activities of choosing, transforming, extracting, combining, and manipulating raw data to create the necessary variables for analysis. Through the application of feature engineering, we successfully extracted valuable information from existing features. For instance, we simplified the representation of a person's "age" into three categories instead of using numerous individual ages. Similarly, we engaged in feature engineering for the "BMI" column, creating four distinct BMI categories. These newly engineered features contribute to the enhancement of the model, leading to improved

```
#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)
```

accuracy.

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

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

Step 5: Checking for outliers:

Outliers refer to exceptionally high or low values that deviate significantly from the majority of data points within a dataset. These outliers can exert a substantial influence on our statistical analyses, potentially distorting the results. It is crucial to meticulously identify potential outliers in our dataset and address them appropriately to ensure accurate and reliable results.

Within our cardiovascular disease (CVD) dataset, outliers have been detected through the application of the "Interquartile range method." This method entails computing the interquartile range (IQR) and pinpointing values that deviate more than 1.5 times the IQR from either the first or third quartile. After identifying these outliers, we can address them by either excluding them from the dataset, substituting them with a more fitting value, or retaining them in the dataset if they genuinely reflect 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 and check distribution of data.

In our data analysis, we employed both the interquartile range (IQR) and box plot methods to evaluate potential outliers. Next, we will examine the outliers specifically for the features that exhibited a higher prevalence in our dataset.

1. BMI

Fig 5.1 (b)- Box plot of “BMI” column

Examining the provided box plot, it is evident that BMI values span from 12 to 98. The median BMI, denoted by the value 27, serves as the point of demarcation between the upper and lower halves of the data. Given that the median is in closer proximity to the lower quartile of 24, it indicates a positive skewness in the distribution. The interquartile range (IQR), representing the difference between the third quartile (Q3) and the first quartile (Q1), is 8. Due to the distribution's lack of symmetry or normality, we cannot assume that the mean and the median are equivalent.

Given that BMI reflects the natural variation within the population, a deliberate decision has been made not to categorize these values as outliers. They will be retained as-is for our subsequent analyses.

2. MentHlth

Fig 5.1 (c)- Box plot of “MentHlthcolumn

Based on the provided box plot, it is evident that the values range from 0 to 30. The median, with a value of 0, indicates an equal distribution with 50% of values being positive and 50% negative. The data's distribution is positively skewed, given that the median is closer to the lower quartile of 0. The interquartile range (IQR), calculated as the difference between the third quartile (Q3) and the first quartile (Q1), is 2. Due to the asymmetry and non-normality of the distribution, it is not reasonable to assume that the mean and the median are equal. Opting to remove outliers would significantly reduce the number of values in this column for analysis. Consequently, a decision has been made to retain the outliers in our dataset.

3. GenHlth

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

The box plot above illustrates a positively skewed distribution, indicating that the data is concentrated towards the lower end of the graph, with a few outliers toward the higher end. The median value is 3, with the lower quartile at 2 and the upper quartile at 3. The range spans from a minimum value of 1 to a maximum value of 5. The interquartile range (IQR), calculated as the difference between the third quartile (Q3) and the first quartile (Q1), is 1. The values in the "GenHlth" column range from 1 to 5, representing a spectrum from very weak to very healthy. Given the significance of all these categories for our analysis, a decision has been made to retain the outliers in our dataset.

2.3. 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 a 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.

3. Initial Exploratory Data Analysis:

3.1. Initial findings and patterns observed

In the course of our analysis, we implemented Min-Max scaling to standardize the data, constraining it to a range of 0 to 1. Subsequent to this scaling procedure, we meticulously examined columns containing values within this normalized range for in-depth exploration. Additionally, we employed feature engineering methods to extract valuable insights from the current set of features. A significant result of this strategy was the development of a novel feature classifying age into three distinct groups. This inventive categorization not only streamlines the depiction of age but also plays a pivotal role in improving the model's accuracy by preserving essential information.

3.2. Visualization techniques used for initial insights

In our analysis, we utilize a diverse set of visualization techniques tailored to the unique characteristics of the data. Bar charts are selected for their efficacy in handling categorical data, especially when dealing with non-string components and a limited number of distinct values. The choice of pie charts is driven by the dataset's small number of categories, providing a clear representation of elevated cholesterol proportions within different age groups. Line plots prove instrumental in illustrating trends and variations over a continuous scale, facilitating the interpretation of correlations between BMI and high blood pressure.

Histograms play a crucial role in identifying patterns in data distribution, with a specific focus on physical health and BMI. Moreover, the application of outlier detection using the Interquartile Range (IQR) informs decisions regarding column retention or discarding. Additionally, scatterplots are employed to assess heart disease prevalence across age categories, delving into relationships between variables such as age and the presence of heart disease to identify potential correlations. The heatmap emerges as a valuable tool for visualizing connections between different variables and heart disease, providing a comprehensive perspective on data relationships. Collectively, these visualizations serve as invaluable tools for interpreting connections within the dataset, particularly concerning heart disease.

Feedback & Iteration:

Initially, we employed one-hot encoding on the education column, leading to the generation of multiple columns based on distinct values. However, upon assessment, these modifications were reverted due to the absence of meaningful insights gleaned from the

individual columns. In our revised approach, there has been a shift from scrutinizing relationships across all columns to a more focused examination centered around cardiovascular disease (CVD). The ongoing investigation involves the plotting of categories to discern proportional trends pertaining to heart disease. In the bivariate analysis, each category is systematically compared to assess the prevalence of heart disease within it. Additionally, in the multivariate analysis, we delved into relationships with other variables to uncover potential correlations, aiming for a more comprehensive understanding of the data.

3.3. Interpretation of the initial analysis

First, we started with checking the distribution of data for different factors.

3.3.1. Univariate Plots:

The examination of the pie charts reveals that within the dataset, around 10.3% of individuals have a history of heart disease or a heart attack, while the majority do not exhibit such conditions. Additionally, only 4.8% of the population has a history of stroke, with the remaining portion showing no signs of stroke.

Moreover, the data indicates a higher prevalence of heart disease or heart attack compared to stroke within the studied population. This underscores the importance of comprehending these health conditions for the implementation of targeted healthcare interventions. Additionally, the relatively low percentages suggest that the dataset may predominantly consist of individuals without a documented history of these cardiovascular issues.

The BMI distribution in the dataset exhibits a normal distribution, reflecting a well-balanced spread across different weight categories. Particularly noteworthy is the concentration of individuals within the BMI category of 30, indicating a substantial presence around this weight status in the sampled population.

Upon examination of the bar chart illustrating BMI categories, it becomes evident that 'Obese' and 'Overweight' are the predominant categories among CVD patients, encompassing approximately 9000 individuals. In contrast, 'Underweight' emerges as the least prevalent category, with fewer than 500 patients, while approximately 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 segments.

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.

3.3.2. Bivariate Plots:

The employment of line plots in both instances reveals a clear trend: there is a noticeable correlation between an individual's BMI and a higher likelihood of experiencing heart disease and stroke. This implies a direct relationship between BMI and the occurrence of heart-related issues.

It's important to highlight that higher BMI values may serve as an indicator of an elevated risk of cardiac events. Individuals with elevated BMIs might find it beneficial to undergo closer monitoring of their cardiovascular health and consider adopting lifestyle changes to mitigate the associated risks.

Upon examining the dataset, where 0 denotes females and 1 represents males, a conspicuous prevalence of females is evident. The visualization primarily focuses on occurrences of heart attacks in relation to gender. The data reveals a noteworthy trend—males exhibit a heightened susceptibility to heart attacks compared to females. This gender-specific observation underscores the imperative to address gender-based disparities in cardiovascular health. The larger representation of females prompts a deeper exploration into heart-related issues across genders. This underscores the significance of implementing gender-specific healthcare strategies to tackle distinct cardiovascular challenges faced by both males and females, emphasizing the necessity for balanced gender representation in healthcare research.

The dataset unveils a significant segment involved in regular physical activity. Intriguingly, individuals with an active lifestyle exhibit a slightly higher susceptibility to heart disease, potentially associated with activities such as running. This nuanced observation underscores the necessity for a thorough comprehension of the relationship between physical activity and heart health. Despite the recognized benefits of exercise, the dataset suggests a complexity in this association, emphasizing the importance of exploring specific types and durations of physical activities. Additionally, personalized health recommendations considering activity profiles emerge as crucial. The dataset yields valuable insights into the intricate interplay between physical activity and heart health, underscoring the need for further exploration to inform nuanced public health strategies.

The dataset unveils a significant segment involved in regular physical activity. Intriguingly, individuals with an active lifestyle exhibit a slightly higher susceptibility to heart disease, potentially associated with activities such as running. This nuanced observation underscores the necessity for a thorough comprehension of the relationship between physical activity and heart health. Despite the recognized benefits of exercise, the dataset suggests a complexity in this association, emphasizing the importance of exploring specific types and durations of physical activities. Additionally, personalized health recommendations considering activity profiles emerge as crucial. The dataset yields valuable insights into the intricate interplay between physical activity and heart health, underscoring the need for further exploration to inform nuanced public health strategies.

Within the dataset, individuals with heavy alcohol consumption are outnumbered by those with moderate intake. Unexpectedly, heavy alcohol consumption is linked to a lower prevalence of heart disease. The dataset hints at a potential protective effect associated with heavy alcohol consumption. Nevertheless, the lower count of individuals with both heavy alcohol consumption and no heart disease implies a complex relationship. Further investigation is paramount to comprehend the nuanced interplay between alcohol consumption patterns and cardiovascular outcomes. These insights underscore the importance of comprehensive analyses considering various factors and lay the groundwork for nuanced discussions on the intricate relationship between alcohol and cardiovascular health.

Within the dataset, a smaller number of individuals display high cholesterol levels, yet there is a notable association with an increased prevalence of heart disease. This highlights the potential significance of cholesterol levels in cardiovascular health outcomes, suggesting that higher cholesterol may elevate susceptibility to heart-related issues. The importance of monitoring and managing cholesterol levels emerges as a crucial preventive measure against heart disease. The dataset underscores the need for targeted interventions for individuals with elevated cholesterol, emphasizing the importance of lifestyle modifications and medical interventions. These insights provide a foundation for focused investigations and interventions aimed at mitigating 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 indicates a scarcity of individuals with diabetes, suggesting a lower prevalence of this condition. Upon examining its relationship with heart disease, it becomes apparent that a larger proportion of individuals without diabetes have a history of heart disease. This underscores diabetes as a potential contributor to cardiovascular health outcomes. Individuals without diabetes appear to be less susceptible to heart-related issues, underscoring the importance of monitoring and managing diabetes for preventing heart disease. 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 provide guidance for focused investigations and interventions aimed at mitigating diabetes's impact on cardiovascular outcomes.

The dataset highlights a significant concentration of individuals with education level 6, indicating a higher prevalence of well-educated individuals. A distinct trend becomes evident, revealing an increase in the frequency of heart disease as education levels rise. This prompts considerations about the potential correlation between higher education and an elevated risk of heart disease. While causation is not established, the association opens avenues for investigating socioeconomic and lifestyle factors contributing to this pattern. The six education levels in the dataset enable a nuanced analysis, underscoring 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 illustrates a spectrum of income levels, revealing a discernible trend: as income increases, there is a corresponding elevation in the likelihood of individuals having a history of heart disease. These prompts considerations of potential factors such as stress and occupational demands influencing this observed pattern. While the correlation suggests a complex interplay of socioeconomic factors impacting cardiovascular health, causation is not firmly established. The dataset's inclusion of multiple income levels allows for a nuanced analysis, underscoring the importance of socioeconomic factors in shaping health outcomes. Delving deeper into stressors and lifestyle attributes associated with varying income levels could offer 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 showcases individuals in the elderly category, indicating a significant presence of older individuals. An unmistakable trend becomes apparent, 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 transition into 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 emphasized in its importance for assessing and addressing cardiovascular risk. In summary, the dataset provides crucial insights into the relationship between age and the likelihood of heart disease, underscoring the necessity for tailored interventions and healthcare strategies for different age groups.

In the initial EDA we plotted:

1. Employ univariate plots to assess the data distribution.
2. Utilize multivariate plots to compare columns such as High BP, High Chol, Age, gender, smoker, etc., with the heart disease or attack column.

4. Final Exploratory Data Analysis:

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

4.1. Bivariate Plots:

The bar chart clearly illustrates the relationship between weight categories and high blood pressure levels, with a noticeable upward trend from underweight to obese individuals. This observation is significant because high blood pressure, a major cause of heart attacks, tends to increase across varying weight categories, as shown in the visual representation. The visual insight emphasizes the crucial link between weight, high blood pressure, and the associated risk of heart attacks.

The bar chart clearly shows that individuals ranging from underweight to obese tend to exhibit higher cholesterol levels, with a noticeable upward trend. This observation is significant because higher cholesterol levels, a prevalent condition in these weight categories, represent a major contributor to heart attacks. The chart emphasizes the crucial link between weight, cholesterol levels, and the heightened risk of experiencing heart attacks.

The pie chart provides a clear representation of the distribution of elevated cholesterol levels across different age groups. The elderly population has the largest share, indicating a higher prevalence of elevated cholesterol in this demographic. The chart also shows that adults and young people have progressively smaller portions, indicating varying degrees of cholesterol levels among age categories.

Fruits

Fruit intake shows a positive correlation with age, with an upward trajectory as individuals mature. Young children, especially 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

The consumption of vegetables is directly related to age, with intake increasing as individuals grow older. Young children, particularly those under the age of 2, have limited vegetable intake. During adolescence, there is a decline in vegetable consumption around ages 10-12, followed by a gradual increase from age 12 onward. This highlights the age-related fluctuations in vegetable consumption across different life phases.

Multivariate Plots:

The risk of cardiovascular disease is directly linked to constant high blood pressure, regardless of gender. However, women are more susceptible to the disease than men at the same blood pressure level. This observation highlights the nuanced gender-specific dynamics in cardiovascular risk, where high blood pressure plays a consistent role, but the impact differs between genders. It is crucial to understand these distinctions 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 interventions.

The graph shows a downward trend, indicating that individuals with lower income and high blood pressure tend to have 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 shows that 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 scatter plot reveals a clear trend, indicating a higher susceptibility to heart disease among elderly individuals. The data suggests that individuals in the "Adult" category also face risks of heart disease, potentially influenced by factors such as physical activity and lifestyle. This observation underscores the necessity for additional visualizations to explore other contributing factors to heart disease risk in the "Adult" category, while also highlighting the distinct vulnerability of the elderly population. While the scatter plot provides a preliminary insight, it prompts the need for a more comprehensive analysis of various risk factors for a holistic understanding of the dynamics of heart disease.

The scatter plot illustrates a clear pattern, indicating that elderly individuals have a heightened susceptibility to heart disease. The data suggests that individuals in the "Adult" category also encounter risks of heart disease, potentially influenced by factors like physical activity and lifestyle. This observation emphasizes the essential need for additional visualizations to explore additional factors contributing to heart disease risk in the "Adult" category, while also highlighting the distinct vulnerability of the elderly population. Serving as an initial insight, the scatter plot prompts the requirement for a more comprehensive analysis of various risk factors to gain a holistic understanding of the dynamics of heart disease.

The observed lower median BMI in individuals with minimal or no risk of heart disease across various age categories reaffirms a direct correlation between BMI and cardiovascular risk. This underscores the significance of BMI as an indicator, emphasizing its role in evaluating heart health across diverse age groups. The results imply that maintaining a lower BMI may be linked to a decreased risk of heart disease. This correlation underscores the importance of weight management in considerations of cardiovascular health and supports the idea that BMI can serve as a valuable marker for assessing the risk of heart disease.

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 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.

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 it 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 treemap is a visual representation that utilizes nested rectangles to depict hierarchical data. Each rectangle represents a category, and its size reflects a quantitative value. This method effectively conveys the proportions and distribution of values within a hierarchy, making treemaps valuable for visualizing structures such as file directories or financial portfolios. They find extensive applications in data analysis and business intelligence, aiding in the identification of patterns within hierarchical relationships.

Key Findings:

In our examination, it emerged that approximately 10.3% of individuals had a documented history of heart disease or had suffered a heart attack, with the majority of the population not showing such medical conditions. Additionally, only 4.8% of the sampled population had experienced a stroke in the past, leaving the majority without any history of stroke.

Major Factors Impacting Cardiovascular Health:

1. **Higher BMI:**
   A noteworthy correlation is observed between individuals with elevated BMI, particularly those in the obese category, and cardiovascular issues. This underscores the significance of weight management.

2. **Smoking:**
   Smoking is recognized as a prominent factor associated with a slightly increased likelihood of developing heart disease, underscoring the imperative for implementing smoking cessation programs to promote cardiovascular health.

3. **High Cholesterol:**
   Smoking is recognized as a prominent factor associated with a slightly increased likelihood of developing heart disease, underscoring the imperative for implementing smoking cessation programs to promote cardiovascular health.

4. **High Blood Pressure:**
   Elevated blood pressure emerges as a significant factor, with individuals in this category displaying a higher percentage of individuals with a history of heart disease. The effective management of blood pressure becomes imperative in preventing cardiovascular events.

5. **Age:**
   Age plays a pivotal role, with a distinct rise in the occurrence of heart disease noted as individuals age, especially within the elderly demographic. Tailored healthcare strategies for older individuals may be imperative.

Minor Factors Impacting Cardiovascular Health:

1. **Gender Disparities:**
   Men exhibit a greater vulnerability to heart attacks in comparison to women, highlighting a gender-related pattern. This finding underscores the importance of adopting gender-specific approaches in cardiovascular health research and interventions.

2. **Physical Activity:**
   People who consistently participate in physical activity demonstrate a slightly elevated susceptibility to heart disease, representing a notable yet minor discovery. Striking a balance between physical activity and other lifestyle factors is essential for comprehensive cardiovascular well-being.

3. **Education Level:**
   A correlation exists between higher levels of education and an increased risk of heart disease, indicating a potential but minor influence. Further exploration of the connection between education and cardiovascular health is warranted.

4. **Income Levels:**
   With an increase in income, there is a corresponding escalation in the probability of

individuals having a history of heart disease, signifying a minor impact. Socioeconomic factors may play a role in the variations observed in cardiovascular health outcomes.

5. **Nutrition:**
   Associations are noted between age and the consumption of fruits and vegetables, indicating heightened awareness and healthier habits among older individuals, albeit with a minor impact. While dietary patterns contribute to cardiovascular health, they should be evaluated in conjunction with other factors.

Factors Not Significantly Impacting Cardiovascular Health:

1. **Diabetes:**
   The dataset reveals a lower incidence of diabetes, and its influence on the history of heart disease does not appear to be substantial. While diabetes is a matter of concern, its role in cardiovascular outcomes within this dataset seems to be less prominent.

2. **Heavy Alcohol Consumption:**
   Unexpectedly, there is an association between heavy alcohol consumption and a reduced prevalence of heart disease, which contradicts the anticipated impact on cardiovascular health. The intricate relationship between alcohol and heart health warrants additional exploration, taking into account 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 Recording:** 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.