Final Project

Aya Ibrahim

ERM 412: Data Visualization

Data Analysis, Visualization, & Storytelling

Managing Heart Health

The Behavioral Risk Factor Surveillance System (BRFSS): Cardiovascular Diseases Risk Prediction Dataset

&

University of California at Irvine (UCI) Cleveland's Database: Heart Disease Dataset

Table of Contents

Executive Summary

Heart disease is a critical global death threat. This paper's motivation lies at the intersection of data science and healthcare, aiming to utilize data analysis to observe heart disease risk factors and enhance public health.

The primary objectives include identifying patterns and associations among lifestyle factors and health metrics to understand their role in heart disease risk. The paper focuses on the analysis of real-world heart disease datasets from (1) *The Behavioral Risk Factor Surveillance System (BRFSS),* and (2) *The University of California at Irvine (UCI) Machine Learning Repository Cleveland's Database*, to further explore risk factors, facilitating early intervention and educating the general public on how to manage their heart health proactively.

This project aims to fulfill its purpose with the application of techniques such as univariate, bivariate, and multivariate analysis through exploratory data analysis to reveal insightful patterns and trends related to the occurrence of diseases.

It is important to note that exploratory analysis does not fully establish a causal relationship between the variables and occurrences due to the fact that correlation does not imply causation. However, it tells a lot about how certain factors interact with one another.

I. Introduction and Background

From its ancient origins in Egypt to the groundbreaking innovations of the 20th and 21st centuries, heart disease remains one of the most critical causes of death globally, with coronary heart disease being one of the most common types, killing at least 375 thousand people every year (Centers for Disease Control and Prevention, 2023). Heart disease, also known as cardiovascular disease (CVD), can refer to various different abnormalities within an individual's heart. These can include heart attack, stroke, heart failure, arrhythmia, heart valve problems, etc. (American Heart Association, 2017). Generally speaking and for the context of this paper, most heart diseases involve a process of blood-flow blocking known as atherosclerosis which happens as a result of arterial plaque that may eventually lead to blood clotting and thus, a heart attack (American Heart Association, 2017). The very first case of atherosclerosis dates back to 1580 and 1550 BC which suggests its prevalence is more than previously believed (Baystate Health, 2022).

A stroke can happen in a similar way to a heart attack but with a blood clot blocking a blood vessel transmitting blood to our brain instead (American Heart Association, 2017). Heart diseases are fatal and dangerous because most heart attacks are expected to lead to heart failure as the heart gets disrupted with it not pumping blood as it normally would.

It is indeed surprising that despite its prevalence, not until the 20th century did heart disease treatment see great innovations (Baystate Health, 2022). President Dwight D. Eisenhower, the 34th president of the United States, had a heart attack that captured the nation's attention and opened eyes to how limited medication options were there for treating such diseases that needed specialized care

(Baystate Health, 2022). Mortality rates were especially reduced after the opening of the first coronary care unit at Bethany Hospital in Kansas to one of the most significant breakthroughs in the 1960s and 1970s, the ability to perform coronary revascularization procedures, including bypass surgery and angioplasty, and even alternatives to the open-heart surgery (Baystate Health, 2022).

Given the historical prevalence and impact of heart diseases, one might wonder about the underlying factors and lifestyle habits contributing to their development. My motivation for selecting this topic lies at the intersection of data science and healthcare. Data science is a powerful tool, and utilizing data analysis and machine learning to predict heart disease risk is not only fascinating but also holds significant potential for improving public health. This journey begins with curiosity and exploratory analysis. Despite our awareness of common risk factors like family history, high blood pressure, chronic stress, diabetes, high cholesterol, and obesity, understanding the pivotal contributors remains a challenge.

Hence, my primary objectives for this project include the identification and analysis of patterns and correlations among lifestyle factors, demographics, and health metrics. This will offer a deeper understanding of their influence on heart disease risk and may provide insights and recommendations for individuals to proactively manage their cardiovascular health. The core mission of this paper is to examine high-risk individuals based on real-world data, enabling early intervention, potentially saving lives, and alleviating the burden on healthcare systems. This paper aims to explore the dynamic relationship between data analysis, visualization, and storytelling while addressing the pressing issue of heart disease, which affects countless lives worldwide.

II. Datasets Description

1) *The 2021 Behavioral Risk Factor Surveillance System (BRFSS), Center for Disease Control:* Cardiovascular Diseases Risk Prediction Dataset

Data Collection & Methodology:

The first dataset to be explored is one I found on Kaggle, a platform for data science competitions. However, the original source of this dataset is extracted from a study of 438,693 records from the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is the "nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services" further operated by the World Health Organization (WHO) and the Center for Disease Control (CDC). The BRFSS conducts annual surveys and questionnaires with adults from all 50 states forming data on a representative sample of the US population.

Although the original dataset consists of 438,693 records and 304 distinct variables, Kaggle bases its dataset on a research article published by the European Journal of Computer Science and Information Technology that uses only 308,854 records and 19 distinct variables. This research article (Lupague et al., 2023) subsets the data to utilize machine learning algorithms in the early detection and prevention of cardiovascular disease. The study concluded that the Logistic Regression model, developed using personal

attributes from the dataset, holds promise for use in the medical field. Further enhancement could be achieved by incorporating additional medical attributes into the dataset, potentially improving the accuracy and utility of CVD risk prediction (Lupague et al., 2023, p. 44). This is why exploring an additional dataset will be very useful for the project's overall purpose.

Data Intake Report & Variable Description:

Total number of observations/instances (rows)	308,854
Total number of features/attributes/variables (columns)	19
Total number of files	1
Base Format of the file	Microsoft Excel Comma Separated Values File (.csv)
Size of the data	30.9 MB (32,453,765 bytes)

```
> glimpse(cvd_raw)
Rows: 308,854
Columns: 19
$ General_Health              <chr> "Poor", "Very Good", "Very Good", "Poor",…
$ Checkup                     <chr> "Within the past 2 years", "Within the pa…
$ Exercise                    <chr> "No", "No", "Yes", "Yes", "No", "No", "Ye…
$ Heart_Disease               <chr> "No", "Yes", "No", "Yes", "No", "No", "Ye…
$ Skin_Cancer                 <chr> "No", "No", "No", "No", "No", "No", "No",…
$ Other_Cancer                <chr> "No", "No", "No", "No", "No", "No", "No",…
$ Depression                  <chr> "No", "No", "No", "No", "No", "Yes", "No"…
$ Diabetes                    <chr> "No", "Yes", "Yes", "Yes", "No", "No", "N…
$ Arthritis                   <chr> "Yes", "No", "No", "No", "No", "Yes", "Ye…
$ Sex                         <chr> "Female", "Female", "Female", "Male", "Ma…
$ Age_Category                <chr> "70-74", "70-74", "60-64", "75-79", "80+"…
$ Height_cm                   <dbl> 150, 165, 163, 180, 191, 183, 175, 165, 1…
$ Weight_kg                   <dbl> 32.66, 77.11, 88.45, 93.44, 88.45, 154.22…
$ BMI                         <dbl> 14.54, 28.29, 33.47, 28.73, 24.37, 46.11,…
$ Smoking_History             <chr> "Yes", "No", "No", "No", "Yes", "No", "Ye…
$ Alcohol_Consumption         <dbl> 0, 0, 4, 0, 0, 0, 0, 3, 0, 0, 0, 0, 8, 4,…
$ Fruit_Consumption           <dbl> 30, 30, 12, 30, 8, 12, 16, 30, 12, 12, 30…
$ Green_Vegetables_Consumption <dbl> 16, 0, 3, 30, 4, 12, 8, 8, 12, 12, 20, 8,…
$ FriedPotato_Consumption     <dbl> 12, 4, 16, 8, 0, 12, 0, 8, 4, 1, 2, 30, 2…
```

As shown in the overview above, the dataset includes a lot of variables that seem to play a critical role in influencing the development of CVD. These include various factors such as general health, checkup frequency, BMI, physical activity, presence of other disease or mental health disorders, smoking history, alcohol consumption, and other dietary variables, including even demographic variables such as age, height, and gender.

As per the analysis I conducted above using R, the dataset contains 12 categorical variables, 8 of which are binary, and 7 other numerical variables for a total of 19 attributes. The author to publish the dataset on Kaggle claims that he/she cleaned and preprocessed the dataset which can indicate that it has no missing or duplicate values. However, I will be reconfirming that through analysis.

Examining the dataset with all those different variables, everything seems to be pretty self-explanatory. I referred back to Kaggle for additional explanation and came up with the following comprehensive description of my data, including some questions and insights that will help fulfill the project's objective:

General_Health: A categorical variable representing the general health status of individuals. The unique values are: ['Poor', 'Very Good', 'Good', 'Fair', 'Excellent']. -- How do individuals' self-reported general health status relate to the risk of heart disease? Potentially identifying a correlation between health perception and actual health conditions.
Checkup: A categorical variable indicating when the last general health checkup was performed.

The unique values are: ['Within the past 2 years' 'Within the past year' '5 or more years ago' 'Within the past 5 years' 'Never'] -- How can regular checkups influence heart disease risk? Helps in exploring whether timely checkups are associated with better heart health.
Exercise: A binary variable indicating whether the individual engages in regular exercise ["Yes" or "No"]. -- What is the impact of regular exercise on heart disease risk? Helps answer whether exercise is a protective factor against heart disease.
Heart_Disease: A binary variable indicating whether the individual has been diagnosed with heart disease ["Yes" or "No"]. -- Target Variable
-- Does the history of these conditions correlate with an increased risk of heart disease? Help identify comorbidities that may contribute to heart disease.
Skin_Cancer: A binary variable indicating whether the individual has been diagnosed with skin cancer ["Yes" or "No"].
Other_Cancer: A binary variable indicating whether the individual has been diagnosed with any cancer other than skin cancer ["Yes" or "No"].
Depression: A binary variable indicating whether the individual has been diagnosed with depression ["Yes" or "No"].
Diabetes: A binary variable indicating whether the individual has been diagnosed with diabetes. The unique values are: ["Yes" or "No", "No, pre-diabetes or borderline diabetes", "Yes, but female told only during pregnancy"].
Arthritis: A binary variable indicating whether the individual has been diagnosed with arthritis ["Yes" or "No"].
Sex: A categorical variable representing the gender of individuals ["Female" or "Male"].

-- Are there gender-specific patterns in heart disease risk or gender-related disparities in heart disease prevalence?
Age_Category: A categorical variable categorizing individuals into age groups. The unique values are: ["70-74", "60-64", "75-79", "80+", "65-69", "50-54", "45-49", "18-24", "30-34", "55-59", "35-39", "40-44", "25-29"] -- At which life stages do individuals become more prone to heart disease? Does it hold true that as people age, they are more likely to develop heart disease due to changes in the heart and blood vessels, such as hardening of the arteries (atherosclerosis) and reduced elasticity of the heart walls (National Institute on Aging, 2018).
-- Is there a relation between body composition and heart disease risk? Analyzing BMI, in particular, can help identify obesity as a potential risk factor.
Height_(cm): Numeric variable representing the height of individuals in centimeters (cm).
Weight_(kg): Numeric variable representing the weight of individuals in kilograms (kg).
BMI: Numeric variable representing the Body Mass Index (BMI) of individuals, calculated from height and weight.
-- How can lifestyle factors and dietary choices influence heart disease risk?
Alcohol_Consumption: Numeric variable representing alcohol consumption habits. According to the BRFSS Questionnaire, the unit of measurement depends on whether the user inputs the intake to be per day, week, or month.
Smoking_History: A categorical variable indicating the smoking history of individuals ["Yes" or "No"].
Green_Vegetables_Consumption: Numeric variable representing the frequency or quantity of green vegetables consumption. According to the BRFSS Questionnaire, the unit of measurement depends on whether the user inputs the intake to be per day, week, or month.
Fruit_Consumption: Numeric variable representing the frequency or quantity of fruit consumption. According to the BRFSS Questionnaire, the unit of measurement depends on whether the user inputs the intake per to be day, week, or month.

FriedPotato_Consumption: Numeric variable representing the frequency or quantity of fried potato consumption. According to the BRFSS Questionnaire, the unit of measurement depends on whether the user inputs the intake to be per day, week, or month.

Data Limitations:

After examining this dataset, there might be limitations to be aware of:

- Self-Reporting: The data relies on individuals' self-reporting, which may introduce bias and inaccuracies, especially for variables like exercise, smoking, and dietary habits.
- Causation vs. Correlation: Like any dataset, correlations found will not always imply that one factor causes another.
- Data Quality: The accuracy of data might vary, and there may be errors in measurements, especially in dietary choice variables as I should aim throughout the analysis to keep the unit of measurement consistent for all records.
- External Factors: There may be external factors not covered in the dataset such as genetics that could influence heart disease risk.

2) ***The University of California at Irvine (UCI) Machine Learning Repository (Cleveland's Database):*** Heart Disease Dataset

Data Collection & Methodology:

The more factors that play a role in influencing the prevalence of heart disease are analyzed, the better the prediction will be and the more informed our society would be on public health. Therefore, another dataset to complement the BRFSS data is one I found from the University of California at Irvine's (UCI) Machine Learning Repository. This dataset dates back to 1988 when it was first donated to UCI, and is now published and stored in its repository. The dataset originally contained 76 variables and 303 instances. However, most published studies and research chose only 14 variables that are most relevant in predicting heart disease.

UCI's repository contains four databases for coronary artery heart disease: Cleveland, Hungary, Switzerland, and the VA Long Beach. Upon further research, the dataset I will be working with only includes the heart disease dataset from the Cleveland database. The UCI repository further reports that the Cleveland dataset was the only one to be processed and used by researchers up to date.

Data Intake Report & Variable Description:

Total number of observations/instances (rows)	303
Total number of features/attributes (columns)	14
Total number of files	1

Base Format of the file	Microsoft Excel Comma Separated Values File (.csv)
Size of the data	11.0 KB (11,319 bytes)

```
> glimpse(cleveland_raw)
Rows: 303
Columns: 14
$ age      <int> 63, 67, 67, 37, 41, 56, 62, 57, 63, 53, 57, 56, 56, 44, 52, 5…
$ sex      <int> 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1…
$ cp       <int> 1, 4, 4, 3, 2, 2, 4, 4, 4, 4, 4, 2, 3, 2, 3, 3, 2, 4, 3, 2, 1…
$ trestbps <int> 145, 160, 120, 130, 130, 120, 140, 120, 130, 140, 140, 140, 1…
$ chol     <int> 233, 286, 229, 250, 204, 236, 268, 354, 254, 203, 192, 294, 2…
$ fbs      <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0…
$ restecg  <int> 2, 2, 2, 0, 2, 0, 2, 0, 2, 2, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 2…
$ thalach  <int> 150, 108, 129, 187, 172, 178, 160, 163, 147, 155, 148, 153, 1…
$ exang    <int> 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1…
$ oldpeak  <dbl> 2.3, 1.5, 2.6, 3.5, 1.4, 0.8, 3.6, 0.6, 1.4, 3.1, 0.4, 1.3, 0…
$ slope    <int> 3, 2, 2, 3, 1, 1, 3, 1, 2, 3, 2, 2, 2, 1, 1, 1, 3, 1, 1, 1, 2…
$ ca       <int> 0, 3, 2, 0, 0, 0, 2, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0…
$ thal     <int> 6, 3, 7, 3, 3, 3, 3, 3, 7, 7, 6, 3, 6, 7, 7, 3, 7, 3, 3, 3, 3…
$ target   <int> 0, 2, 1, 0, 0, 0, 3, 0, 2, 1, 0, 0, 2, 0, 0, 0, 1, 0, 0, 0, 0…
```

```
In [7]: df.isnull().sum()

Out[7]: age         0
        sex         0
        cp          0
        trestbps    0
        chol        0
        fbs         0
        restecg     0
        thalach     0
        exang       0
        oldpeak     0
        slope       0
        ca          4
        thal        2
        target      0
        dtype: int64
```

All 14 variables are numerical. One thing that needs to be kept in mind while interpreting those values is another statement by UCI claiming that “Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).”

As per the analyses on the previous page, the right image shows that there are six missing values upon further analyses using Python. Also, unlike the BRFSS dataset, many of the variables from this one include medical terminology that the general public wouldn’t be familiar with. Data modification such as changing variable names to easier terms to grasp needs to be applied.

Below is a comprehensive description of all 14 variables in this dataset, including some helpful questions and insights/notes.

age: A numerical variable indicating the age of the patient.
sex: A binary variable indicating the gender of the patient. The unique values are: [0 = "Female", 1 = "Male].
cp: A numerical variable indicating Chest Pain Type (aka Angina Pectoris). It has 4 unique values [1,2,3,4] each representing a different type: [1: Typical Anginal Pain, 2: Atypical Anginal Pain, 3: Non-anginal Pain, 4: Asymptomatic]. -- Atypical: Not typically associated with heart disease, but should not be ignored. Typical angina: Pain directly related to the heart, indicating potential coronary artery disease. Asymptomatic: No pain but potential silent heart attacks. Non-angina: Pain not related to the heart but may signal other conditions. The most dangerous in regard to CVD development in order are: Typical, Non-Anginal, Atypical, and Asymptomatic (Nakas et al., 2019).

trestbps: A numerical variable indicating the resting blood pressure on admission to the hospital. (Unit: mm Hg) -- It is commonly known that higher blood pressure leads to a higher risk of heart disease.
chol: A numerical variable indicating the concentration of serum cholesterol. (Unit: mg/dL) -- Higher cholesterol levels can lead to plaque buildup in arteries, a primary cause of atherosclerosis as previously mentioned. This buildup restricts blood flow to the heart, leading to heart diseases.
fbs: A binary variable indicating the value of fasting blood glucose/sugar. If fasting blood sugar > 120mg/dl then: [1 (True), 0 (False)]. -- High blood glucose levels (>120 mg/dL) can indicate diabetes, a significant risk factor for heart disease.
restecg: A numerical variable indicating the value of the resting electrocardiogram results with unique values [0: Normal, 1: Having ST-T wave abnormality l, 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria]. -- Abnormal ECG results can indicate heart problems.
thalach: A numerical variable indicating the value of maximum heart rate achieved. -- A low max heart rate during exercise might indicate that the heart is not receiving enough oxygen, which could be due to a blockage or other heart-related issues.
exang: A binary variable indicating the value of the presence of exercise-induced angina [1 = Yes, 0 = No]. -- The presence of angina during exercise is a strong indicator of some blockage or narrowing in the coronary arteries, which is a direct sign of heart disease.

oldpeak: A numerical variable indicating the value of ST segment depression induced by exercise relative to rest. -- An elevation or depression in the ST segment during or after exercise can indicate reduced blood flow to the heart, suggesting potential heart disease.
slope: A numerical variable indicating the value of the slope of the exercise ST section at the movement peak [1: upsloping; 2: flat; 3: downsloping]. -- The shape of the ST segment during exercise can provide insight into blood flow patterns in the heart. A downward (depressive) slope might indicate ischemia or heart disease.
ca: A numerical variable indicating the value of the number of major vessels colored by fluoroscopy with unique values: [0, 1, 2, 3]. -- The visibility of more vessels generally indicates a higher volume of blood flow, which could be interpreted as a healthier cardiovascular state. If no vessels are visible (0), it may suggest a higher risk of cardiovascular issues.
thal: A numerical variable indicating the value of the corresponding present Thalassemia type: [3 = Normal, 6 = Fixed Defect, 7 = Reversible Defect]. -- Thalassemia is a genetic blood disorder that affects the production of hemoglobin, a protein that carries oxygen in red blood cells. Normal indicates no abnormalities, Fixed indicates that it could be non-reversible or more chronic, and Reversible indicates that it is present but can be improved or managed.
target: A binary variable indicating the presence of heart disease [0: Absence, 1,2,3,4: Presence]. -- Target Variable

Data Limitations:

There aren't many limitations in this dataset but some to consider may be:

- External Factors: There are external factors not covered in the dataset such as genetics that could influence heart disease risk.
- Causation vs. Correlation: Like any dataset, correlations found will not always imply that one factor causes another.

III. Exploratory Data Analysis

1) *The 2021 Behavioral Risk Factor Surveillance System (BRFSS), Center for Disease Control:* Cardiovascular Diseases Risk Prediction Dataset

Data Cleaning & Pre-processing:

Upon further analysis in Python, the dataset, in fact, does not include any missing or duplicate values confirming what is claimed on Kaggle making it a good starting point for EDA. The data types seem appropriate for each column, with categorical variables stored as objects and numerical variables stored as integers or floats.

I have performed the modifications below using R on the dataset in ways I believe would be useful for future visualization and analyses:

- Recoded the values in the "*Diabetes*" column as it would be more useful if Diabetes only had two unique values "Yes" or "No" like the other disease variables.
- Added an index variable "*Patient_ID*" to each row to ensure we are working with the same individuals at various stages of the process.
- Recoded and added new columns to the dataset for all binary variables (Yes, No) to be numerical as (1, 0).
- Converted all consumption/intake values to be per month assuming that: Values greater than 7 are assumed to be in units of "per week" and are converted to "per month" by multiplying them by 4. Values greater than 1 but less than or equal to 7 are assumed to be in units of "per day" and are converted to "per month" by multiplying them by 30.
- Recoded certain category variables such as: *General_Health, Age_Category*, and *Checkup* to be as factors. This allows me to set the order and convert it to a numeric variable.
- Recoded and created a new column for BMI breaking them into levels: Underweight, Normal Weight, Overweight, and Obesity.
- Split data into separate tables to focus on one area at a time by building tables with variables that relate to each other combined in one:
    - Patient Demographics & Background Variables table

- Patient Dietary Choices table
- Patient Health Level & Other Factors table
- Diseases/Disorders table

```
Rows: 308,854
Columns: 34
$ Patient_ID                    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13…
$ General_Health                <chr> "Poor", "Very Good", "Very Good", "Poor",…
$ Checkup                       <chr> "Within the past 2 years", "Within the pa…
$ Exercise                      <chr> "No", "No", "Yes", "Yes", "No", "No", "Ye…
$ Heart_Disease                 <chr> "No", "Yes", "No", "Yes", "No", "No", "Ye…
$ Skin_Cancer                   <chr> "No", "No", "No", "No", "No", "No", "No",…
$ Other_Cancer                  <chr> "No", "No", "No", "No", "No", "No", "No",…
$ Depression                    <chr> "No", "No", "No", "No", "No", "Yes", "No"…
$ Diabetes                      <chr> "No", "Yes", "Yes", "Yes", "No", "No", "N…
$ Arthritis                     <chr> "Yes", "No", "No", "No", "No", "Yes", "Ye…
$ Sex                           <chr> "Female", "Female", "Female", "Male", "Ma…
$ Age_Category                  <chr> "70-74", "70-74", "60-64", "75-79", "80+"…
$ Height_cm                     <dbl> 150, 165, 163, 180, 191, 183, 175, 165, 1…
$ Weight_kg                     <dbl> 32.66, 77.11, 88.45, 93.44, 88.45, 154.22…
$ BMI                           <dbl> 14.54, 28.29, 33.47, 28.73, 24.37, 46.11,…
$ Smoking_History               <chr> "Yes", "No", "No", "No", "Yes", "No", "Ye…
$ Alcohol_Consumption           <dbl> 0, 0, 120, 0, 0, 0, 0, 90, 0, 0, 0, 0, 32…
$ Fruit_Consumption             <dbl> 120, 120, 48, 120, 32, 48, 64, 120, 48, 4…
$ Green_Vegetables_Consumption  <dbl> 64, 0, 90, 120, 120, 48, 32, 32, 48, 48, …
$ FriedPotato_Consumption       <dbl> 48, 120, 64, 32, 0, 48, 0, 32, 120, 1, 60…
$ Has_Exercise_Int              <int> 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0,…
$ Has_Heart_Disease_Int         <int> 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0,…
$ Has_Skin_Cancer_Int           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,…
$ Has_Other_Cancer_Int          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Has_Depression_Int            <int> 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0,…
$ Has_Diabetes_Int              <int> 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0,…
$ Has_Arthritis_Int             <int> 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0,…
$ Has_Smoking_History_Int       <int> 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1,…
$ General_Health_Factor         <fct> Poor, Very Good, Very Good, Poor, Good, G…
$ General_Health_Num            <dbl> 1, 4, 4, 1, 3, 3, 2, 3, 2, 2, 2, 2, 4, 2,…
$ Checkup_Factor                <fct> Within the past 2 years, Within the past …
$ Checkup_Num                   <dbl> 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5,…
$ Age_Factor                    <fct> 70-74, 70-74, 60-64, 75-79, 80+, 60-64, 6…
$ Age_Num                       <dbl> 11, 11, 9, 12, 13, 9, 9, 10, 10, 11, 12, …
```

Statistical Summary:

In this step, it would be useful to generate descriptive statistics to understand the distribution of the data and then proceed to univariate analysis, particularly focusing on the target 'Heart_Disease' variable to understand its distribution.

	count	unique	top	freq
General_Health	308854	5	Very Good	110395
Checkup	308854	5	Within the past year	239371
Exercise	308854	2	Yes	239381
Heart_Disease	308854	2	No	283883
Skin_Cancer	308854	2	No	278860
Other_Cancer	308854	2	No	278976
Depression	308854	2	No	246953
Diabetes	308854	2	No	266037
Arthritis	308854	2	No	207783
Sex	308854	2	Female	160196
Age_Category	308854	13	65-69	33434
Smoking_History	308854	2	No	183590
General_Health_Factor	308854	5	Very Good	110395
Checkup_Factor	308854	5	Within the past year	239371
Age_Factor	308854	13	65-69	33434
BMI_Category	308854	4	Obesity	108885

	count	mean	std	min	25%	50%	75%	max
Patient_ID	308854.0	154427.500000	89158.614358	1.00	77214.25	154427.50	231640.75	308854.00
Height_cm	308854.0	170.615249	10.658026	91.00	163.00	170.00	178.00	241.00
Weight_kg	308854.0	83.588655	21.343210	24.95	68.04	81.65	95.25	293.02
BMI	308854.0	28.626211	6.522323	12.02	24.21	27.44	31.85	99.33
Alcohol_Consumption	308854.0	41.576165	52.583808	0.00	0.00	1.00	80.00	210.00
Fruit_Consumption	308854.0	129.681778	93.918854	0.00	60.00	120.00	150.00	480.00
Green_Vegetables_Consumption	308854.0	81.270953	59.403767	0.00	48.00	64.00	120.00	512.00
FriedPotato_Consumption	308854.0	64.581595	53.427373	0.00	32.00	60.00	120.00	512.00
Has_Exercise_Int	308854.0	0.775062	0.417542	0.00	1.00	1.00	1.00	1.00
Has_Heart_Disease_Int	308854.0	0.080850	0.272606	0.00	0.00	0.00	0.00	1.00
Has_Skin_Cancer_Int	308854.0	0.097114	0.296113	0.00	0.00	0.00	0.00	1.00
Has_Other_Cancer_Int	308854.0	0.096738	0.295602	0.00	0.00	0.00	0.00	1.00
Has_Depression_Int	308854.0	0.200422	0.400316	0.00	0.00	0.00	0.00	1.00
Has_Diabetes_Int	308854.0	0.138632	0.345563	0.00	0.00	0.00	0.00	1.00
Has_Arthritis_Int	308854.0	0.327245	0.469208	0.00	0.00	0.00	1.00	1.00
Has_Smoking_History_Int	308854.0	0.405577	0.491004	0.00	0.00	0.00	1.00	1.00
General_Health_Num	308854.0	3.530448	1.031224	1.00	3.00	4.00	4.00	5.00
Checkup_Num	308854.0	4.617981	0.815120	1.00	5.00	5.00	5.00	5.00
Age_Num	308854.0	7.535888	3.523526	1.00	5.00	8.00	10.00	13.00

The descriptive statistics provide a summary of the following for the numerical variables:

- Heights range from 91 to 241 cm, weights from approximately 25 to 293 kg, and BMI from 12.02 to 99.33, indicating a wide range of body sizes, but the possibility of potential outliers. This also suggests that most individuals are overweight.
- Alcohol consumption, fruit consumption, and green vegetable consumption variables have a wide range of values, indicating varied dietary habits among the participants. However, the highest intake is found in fruits & green vegetables.
- The binary variables for exercise, heart disease, and other conditions are encoded as 0s and 1s, with the mean values indicating the proportion of positive responses. Individuals reporting "Yes" to any of the disorders/diseases (Arthritis, CVD, Diabetes, Depression, Skin Cancer, Other Cancer), and Smoking History are significantly less than those who report "No." It also showed that most individuals exercise.
- Regarding the target variable 'Heart_Disease', approximately 8.09% of the entries indicate the presence of heart disease, while the vast majority, 91.92%, do not have heart disease.

Categorical Variables statistics further indicated important insights and further confirmed what has been observed in the numerical statistics summary:

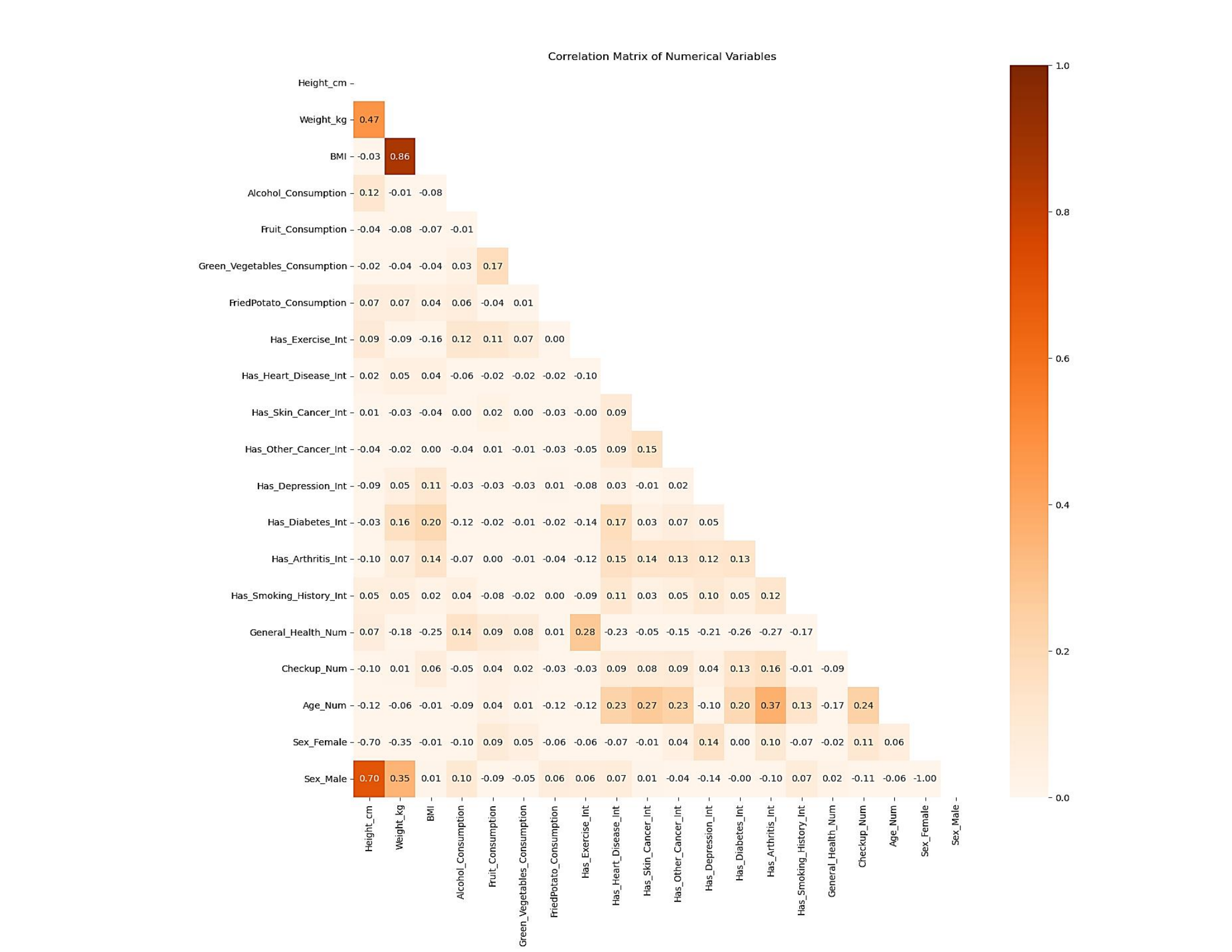
- Most individuals are females.
- The age group with the highest majority is 65-69.
- Most people exercise, have had a checkup within the past year, and describe their health as "Very Good."
- The majority report not having any of the diseases/disorders.
- The majority do not smoke.

I proceeded to generate a heatmap using Python to explore what factors relate most to the target, Heart Disease. This made it easier for me to focus on certain factors during the analysis process. In this step, I have also found it useful to observe gender disparities, if any, so I performed one-hot encoding to include each of the genders within the numerical correlation matrix shown on Page 22. I found the following observations to be the most insightful for the project's objective.

- There is a strong positive correlation (0.86) between BMI and weight, which is expected as BMI is calculated from a person's weight and height. They also show positive but weak correlations with heart disease, which aligns with the understanding that obesity can be a risk factor for cardiovascular issues.

- Age shows a moderate positive correlation (0.23) with having heart disease. This suggests that the likelihood of having heart disease increases with age.
- Age is also positively correlated with having arthritis (0.37), other cancer (0.20), and diabetes (0.17), indicating these conditions are more common in older individuals.
- Regular checkups have a positive correlation (0.16) with better general health, implying that frequent health checkups might be associated with better health outcomes.
- General health shows a moderate negative correlation with having heart disease (-0.26), suggesting that individuals who rate their health poorly have a higher chance of having heart disease.
- Having a routine of exercise is negatively correlated with heart disease (-0.10), which may suggest that individuals who exercise may have a lower likelihood of developing heart disease.

Despite this analysis being comprehensive, it's still important to note that while correlation can indicate an association, it does not confirm causation.


Correlation Matrix of Numerical Variables
Height_cm
Weight_kg 0.47
BMI -0.03 0.86
Alcohol_Consumption 0.12 -0.01 -0.08
Fruit_Consumption -0.04 -0.08 -0.07 -0.01
Green_Vegetables_Consumption -0.02 -0.04 -0.04 0.03 0.17
FriedPotato_Consumption 0.07 0.07 0.04 0.06 -0.04 0.01
Has_Exercise_Int 0.09 -0.09 -0.16 0.12 0.11 0.07 0.00
Has_Heart_Disease_Int 0.02 0.05 0.04 -0.06 -0.02 -0.02 -0.02 -0.10
Has_Skin_Cancer_Int 0.01 -0.03 -0.04 0.00 0.02 0.00 -0.03 -0.00 0.09
Has_Other_Cancer_Int -0.04 -0.02 0.00 -0.04 0.01 -0.01 -0.03 -0.05 0.09 0.15
Has_Depression_Int -0.09 0.05 0.11 -0.03 -0.03 -0.03 0.01 -0.08 0.03 -0.01 0.02
Has_Diabetes_Int -0.03 0.16 0.20 -0.12 -0.02 -0.01 -0.02 -0.14 0.17 0.03 0.07 0.05
Has_Arthritis_Int -0.10 0.07 0.14 -0.07 0.00 -0.01 -0.04 -0.12 0.15 0.14 0.13 0.12 0.13
Has_Smoking_History_Int 0.05 0.05 0.02 0.04 -0.08 -0.02 0.00 -0.09 0.11 0.03 0.05 0.10 0.05 0.12
General_Health_Num 0.07 -0.18 -0.25 0.14 0.09 0.08 0.01 0.28 -0.23 -0.05 -0.15 -0.21 -0.26 -0.27 -0.17
Checkup_Num -0.10 0.01 0.06 -0.05 0.04 0.02 -0.03 -0.03 0.09 0.08 0.09 0.04 0.13 0.16 -0.01 -0.09
Age_Num -0.12 -0.06 -0.01 -0.09 0.04 0.01 -0.12 -0.12 0.23 0.27 0.23 -0.10 0.20 0.37 0.13 -0.17 0.24
Sex_Female -0.70 -0.35 -0.01 -0.10 0.09 0.05 -0.06 -0.06 -0.07 -0.01 0.04 0.14 0.00 0.10 -0.07 -0.02 0.11 0.06
Sex_Male 0.70 0.35 0.01 0.10 -0.09 -0.05 0.06 0.06 0.07 0.01 -0.04 -0.14 -0.00 -0.10 0.07 0.02 -0.11 -0.06 -1.00
Height_cm
Weight_kg
BMI
Alcohol_Consumption
Fruit_Consumption
Green_Vegetables_Consumption
FriedPotato_Consumption
Has_Exercise_Int
Has_Heart_Disease_Int
Has_Skin_Cancer_Int
Has_Other_Cancer_Int
Has_Depression_Int
Has_Diabetes_Int
Has_Arthritis_Int
Has_Smoking_History_Int
General_Health_Num
Checkup_Num
Age_Num
Sex_Female
Sex_Male
1.0
0.8
0.6
0.4
0.2
0.0

Visualizing the insights extracted previously would be helpful in this case.

Univariate, Bivariate, & Multivariate Analysis:

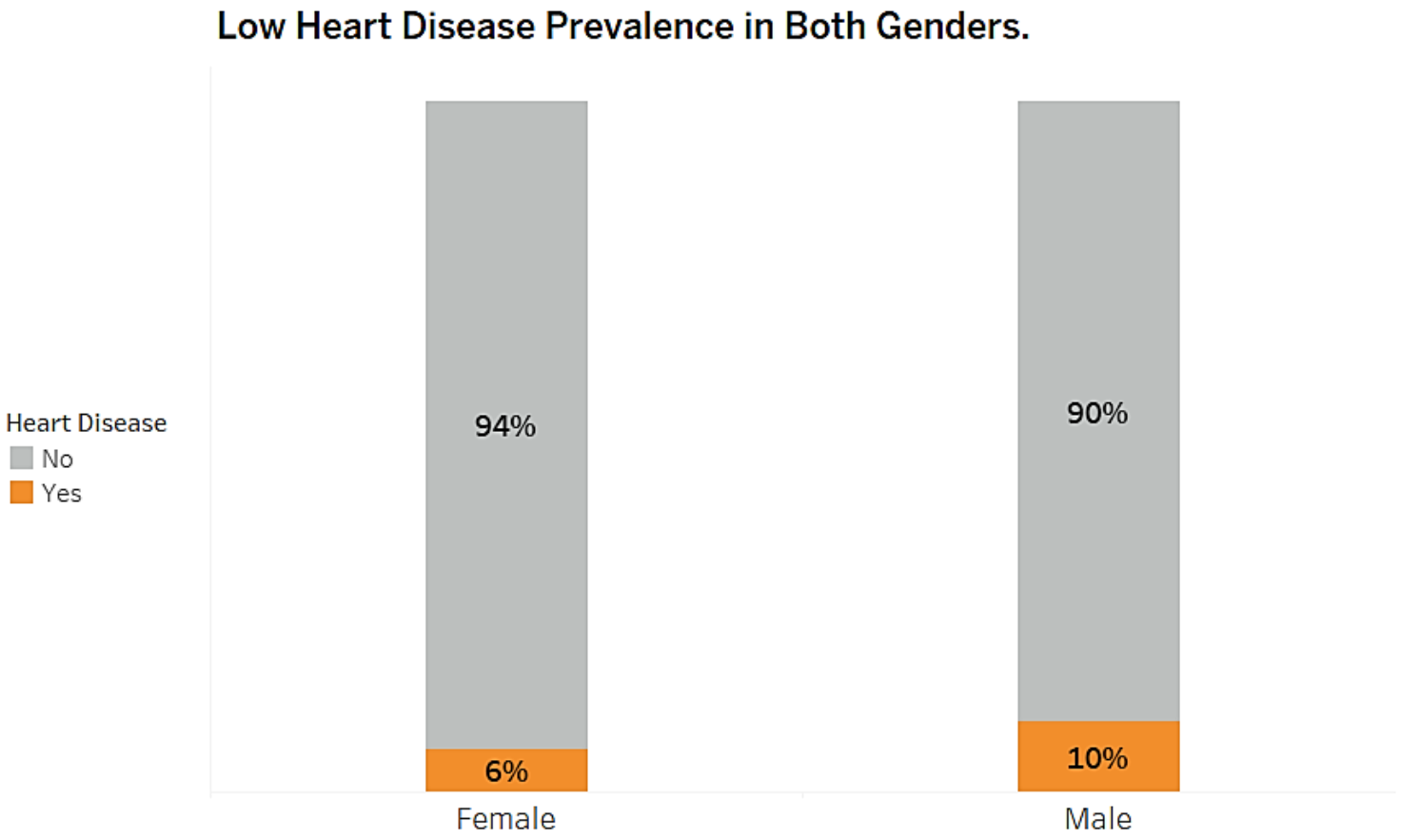
These exploratory analyses are especially useful in answering some curiosity-driven questions based on the insights from the descriptive statistics:

- Is there any gender disparity in the prevalence of heart disease among surveyed individuals?

The bar graphs on the next page show the distribution of the target variables, "*Heart_Disease,*" in addition to the gender *("Sex")* distribution with respect to the target.

The graph below is a bar graph showing the distribution of Heart Disease or Count of heart disease cases across both genders, in which the color shows details about Heart Disease with grey indicating no presence and orange indicating presence.

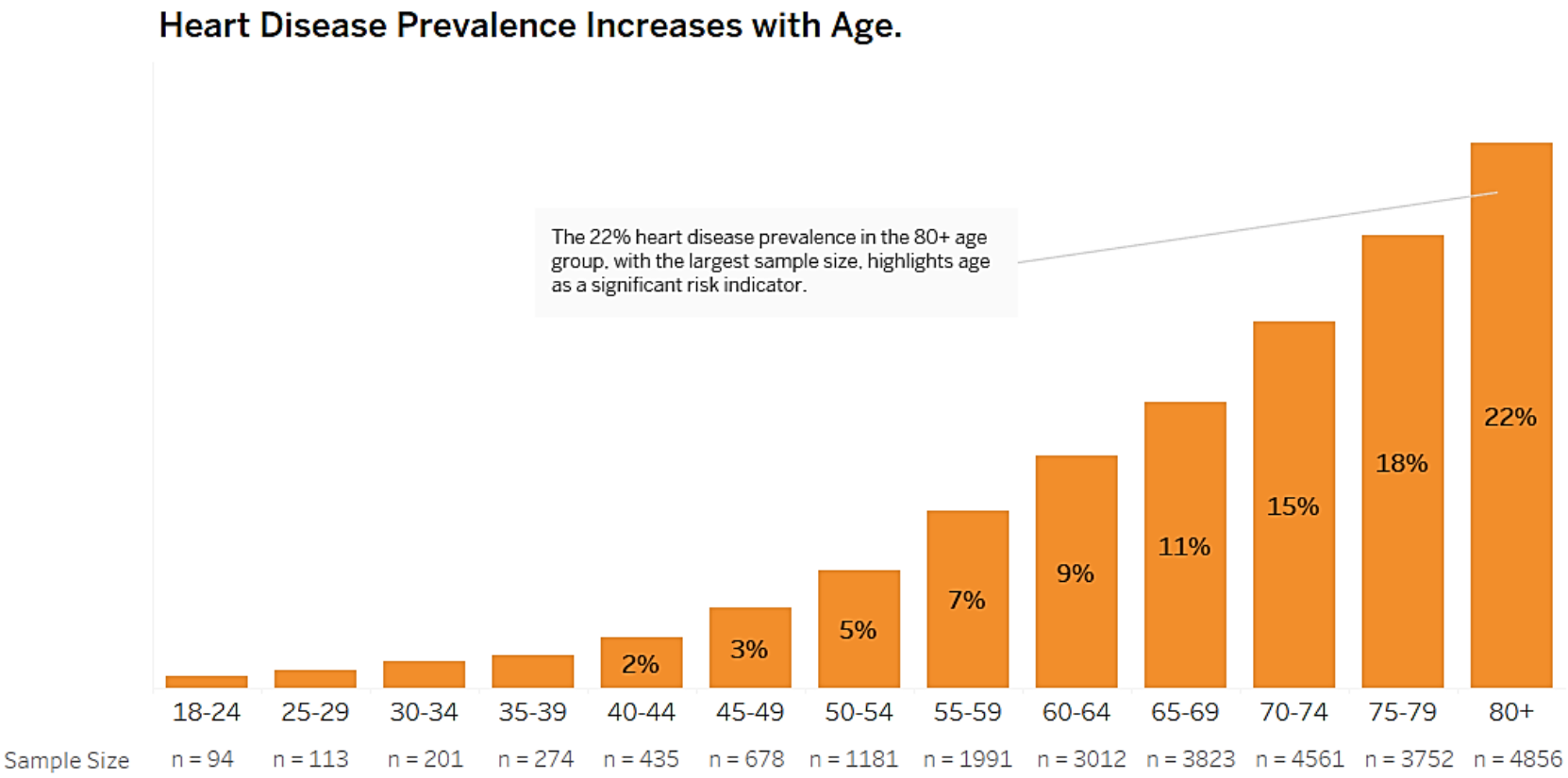
The graph shows that males are more likely to develop CVD but not at a significantly higher rate than females. Therefore, one cannot say that there is a gender disparity in CVD rate observed in this dataset.



Center for Disease Control - Cardiovascular Disease Risk Prediction Dataset
https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?select=CVD_cleaned.csv

- Are there certain life stages where individuals become at higher risk for heart disease?

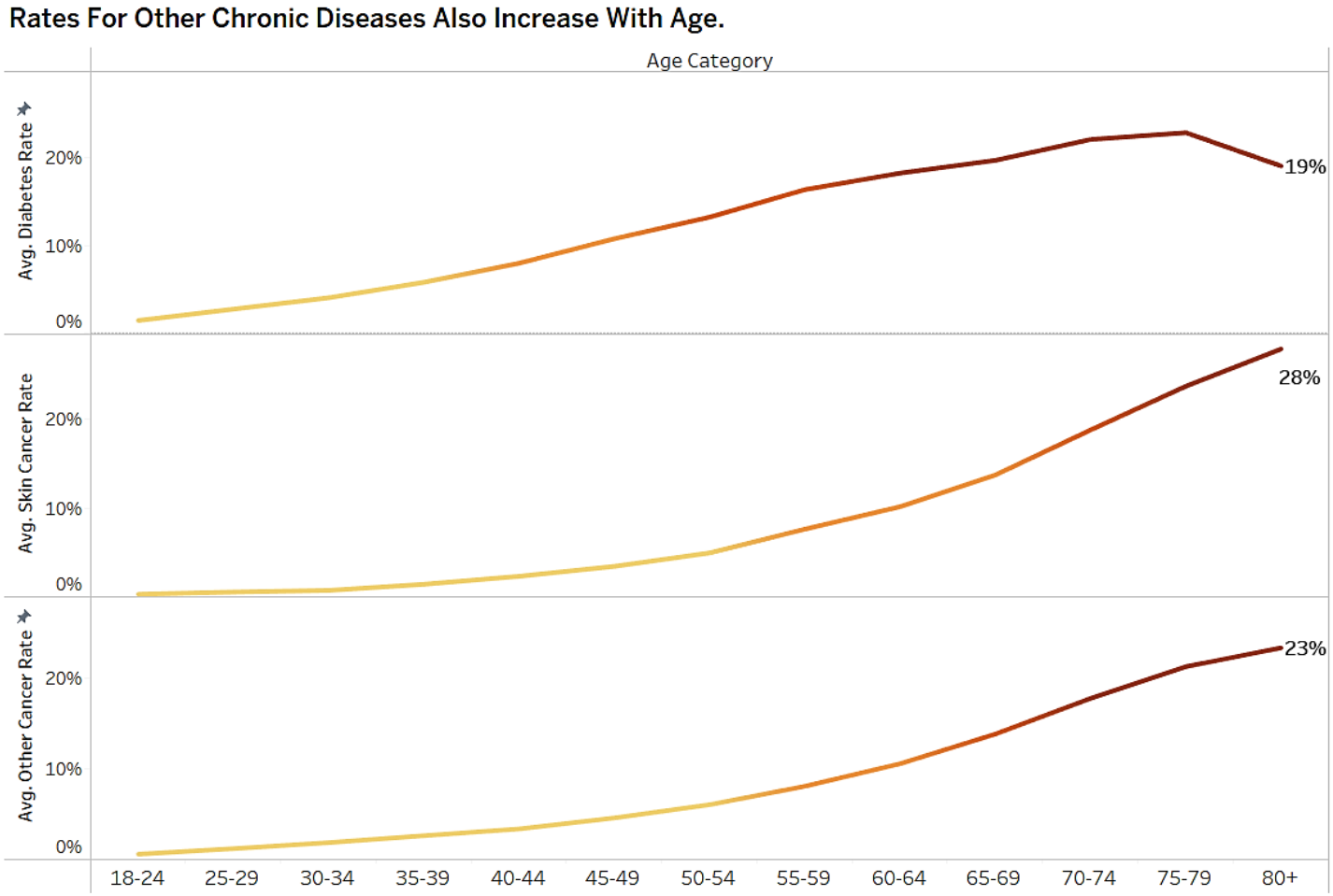
The graph below shows the distribution of age categories in this dataset through percent of the total count in which the orange color represents the cases of those who reported to have CVD. This graph clearly indicates that individuals are more prone to heart disease as they age.



Center for Disease Control - Cardiovascular Disease Risk Prediction Dataset
https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?select=CVD_cleaned.csv

- Is what we observed with CVD also the case for other diseases?

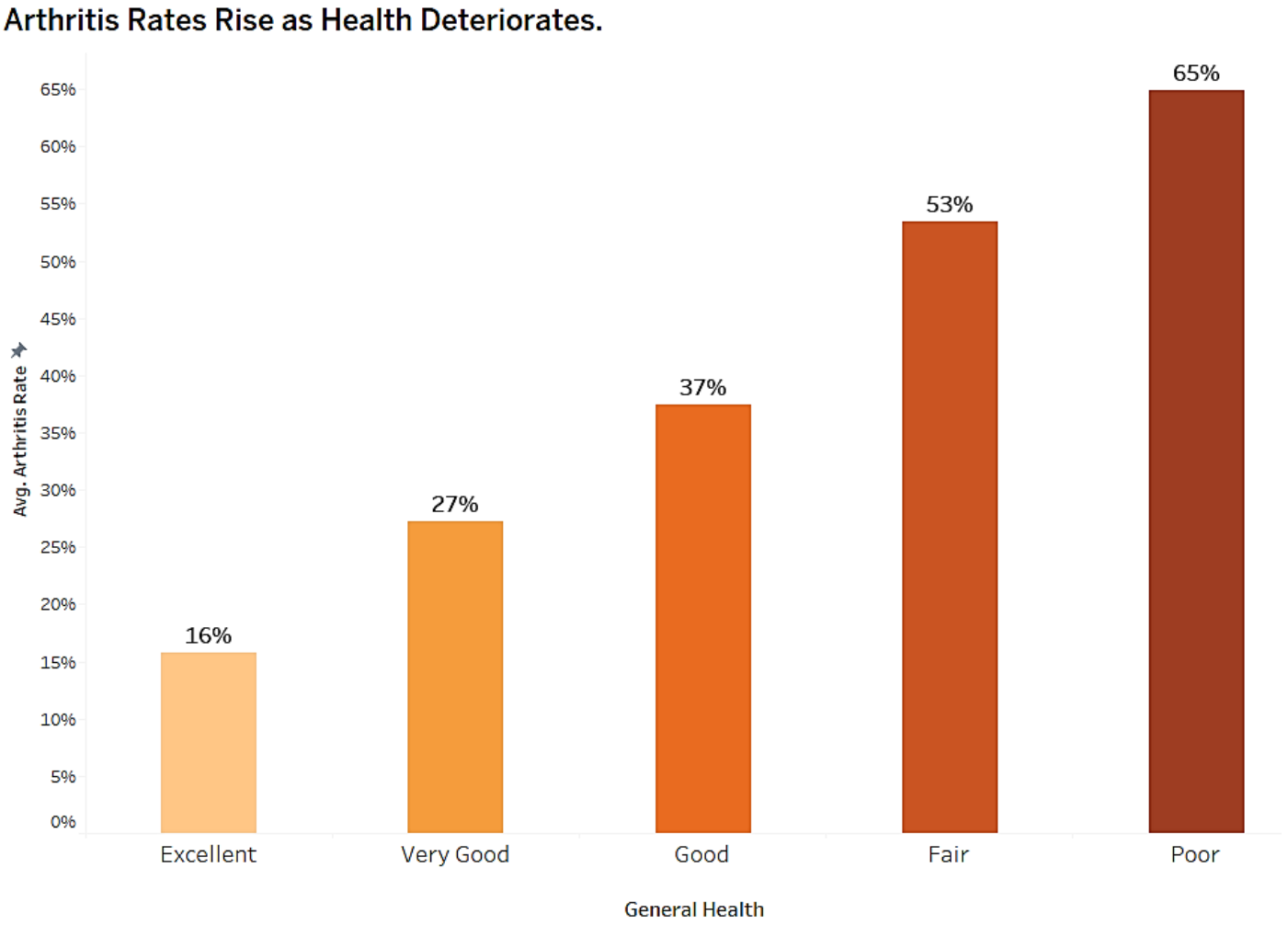
  The visual below shows the trends of Diabetes, Skin Cancer, and Other Cancer types rates across different age categories. A line graph was used for easier interpretation to convey a clear idea that most chronic disease rates increase with age. The use of color further reinforces the message as it goes from low to high intensity. Note that the diseases' rates are aggregated to show the average after being converted from categorical to binary variables.



Center for Disease Control - Cardiovascular Disease Risk Prediction Dataset
https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?select=CVD_cleaned.csv

- How well do people's perceptions of their health reflect actual medical conditions?

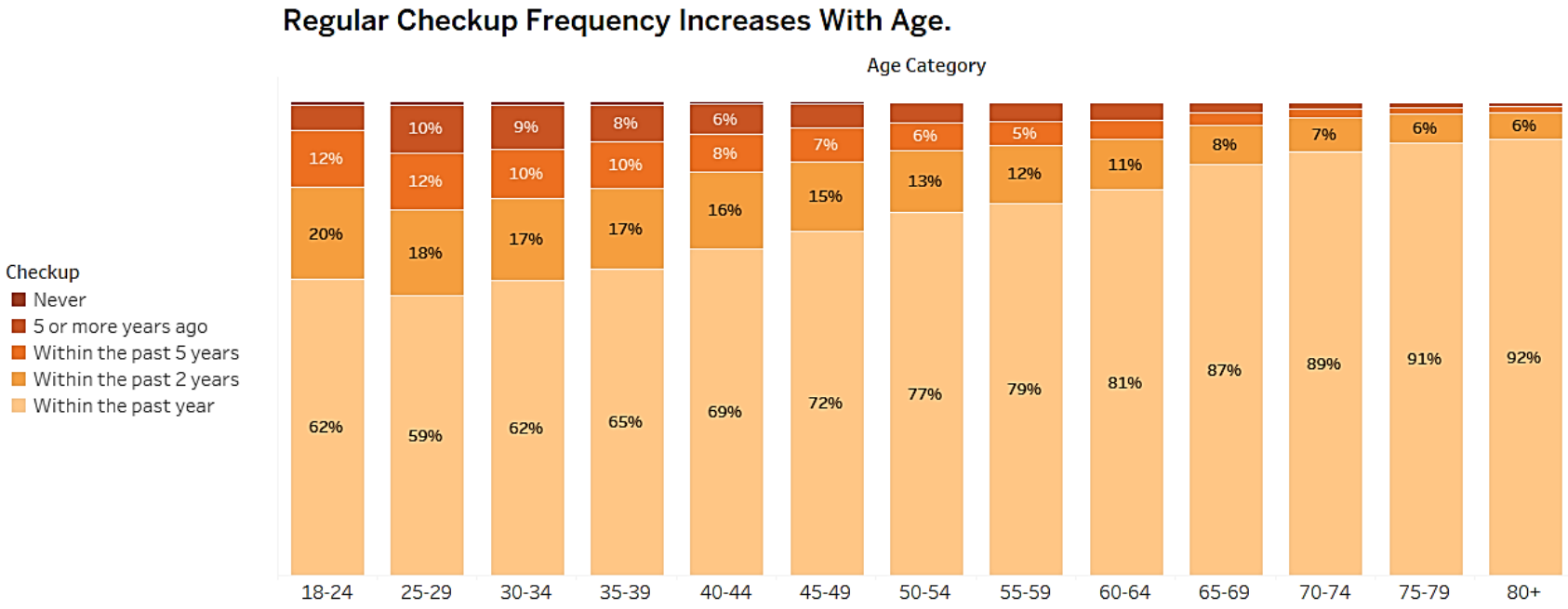
  This visualization retells the insights concluded from the previous visual in a different way by showing the general health categories and how they change with respect to the average arthritis rate. This graph clearly indicates that high average arthritis rates are often related to negative health perceptions.



Center for Disease Control - Cardiovascular Disease Risk Prediction Dataset
https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?select=CVD_cleaned.csv

- How effective are regular health checkups in preventing heart disease?

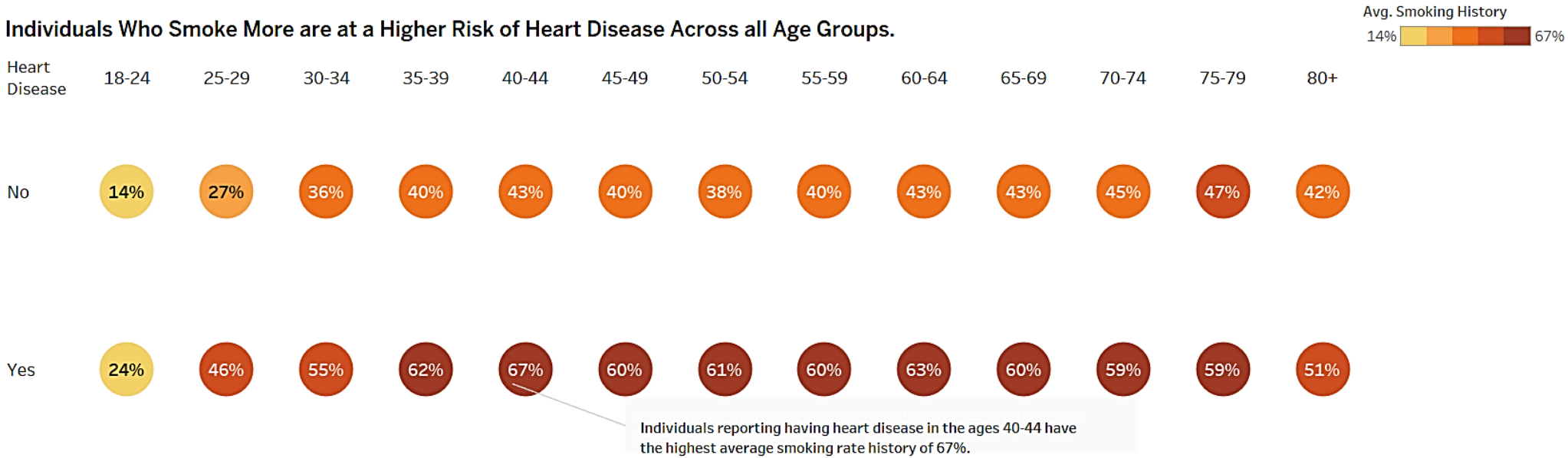
The visual below shows the distribution of the checkup frequency categories through a percentage of total count across all age groups. The colors shown represent each of the checkup categories going from lighter to darker. This graph indicates that individuals are more likely to check up as they get older, which might indicate receipt of treatment as a result of illness as this correlates with what has been observed before.



Center for Disease Control - Cardiovascular Disease Risk Prediction Dataset
https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?select=CVD_cleaned.csv

- How does smoking impact heart disease risk across both genders and different age groups?

The heatmap below shows the Smoking History broken down by age category vs. heart disease prevalence. Color represents Smoking History going lighter to darker to represent the intensity of the case. The marks are labelled by average Smoking History and shown as percentages for easier interpretability.



Center for Disease Control - Cardiovascular Disease Risk Prediction Dataset
https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?select=CVD_cleaned.csv

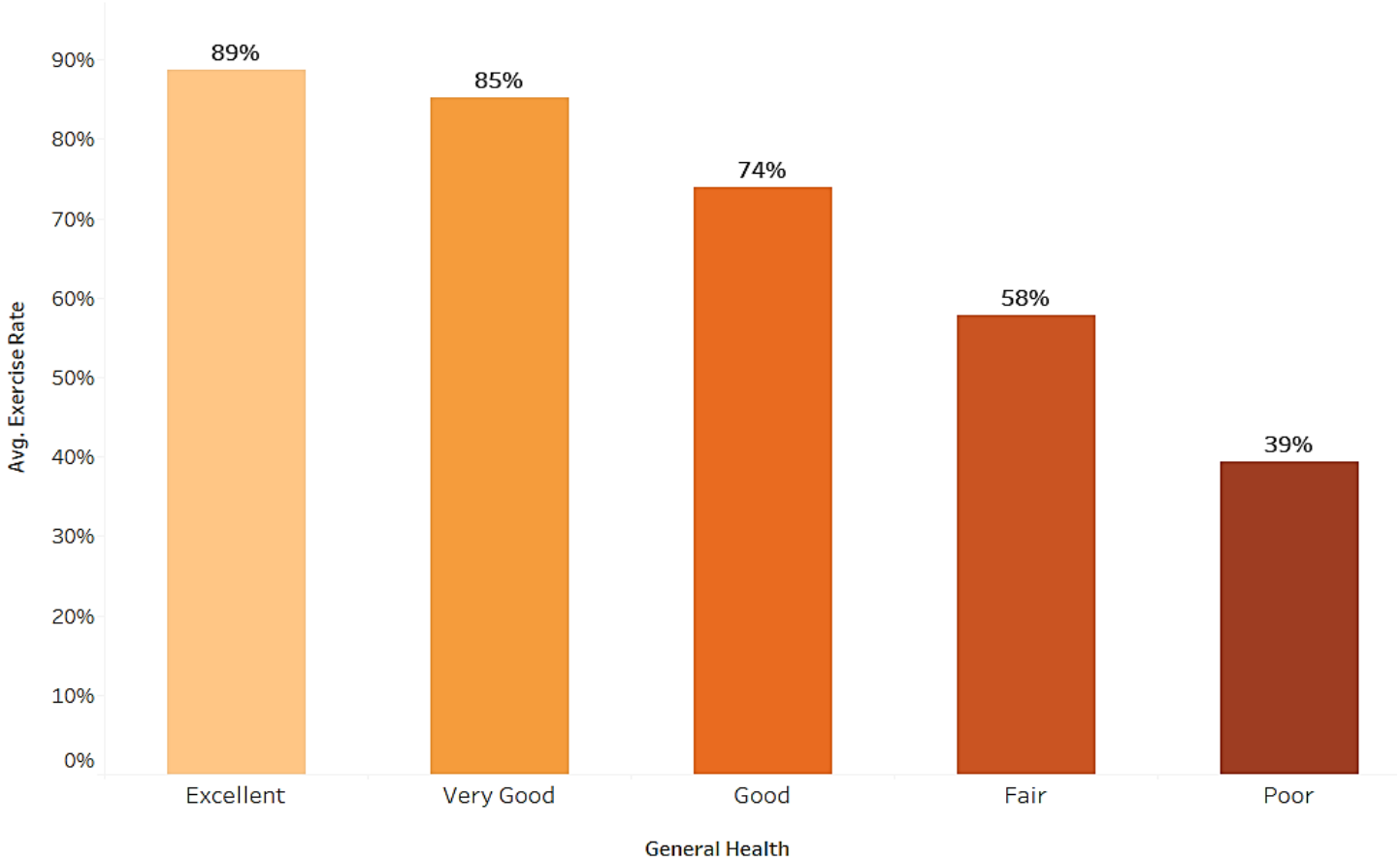
There is an observable trend where the prevalence of heart disease increases with age for both smokers and non-smokers, which aligns with common medical knowledge that risk increases with age. Overall, this visual shows that smoking is a clear indicator of heart disease risk.

Next, it would be useful to explore how body composition helps influence the prevalence of heart disease, in addition to incorporating physical activity into one's daily routine.

- Does regular exercise reduce cardiovascular disease risk?

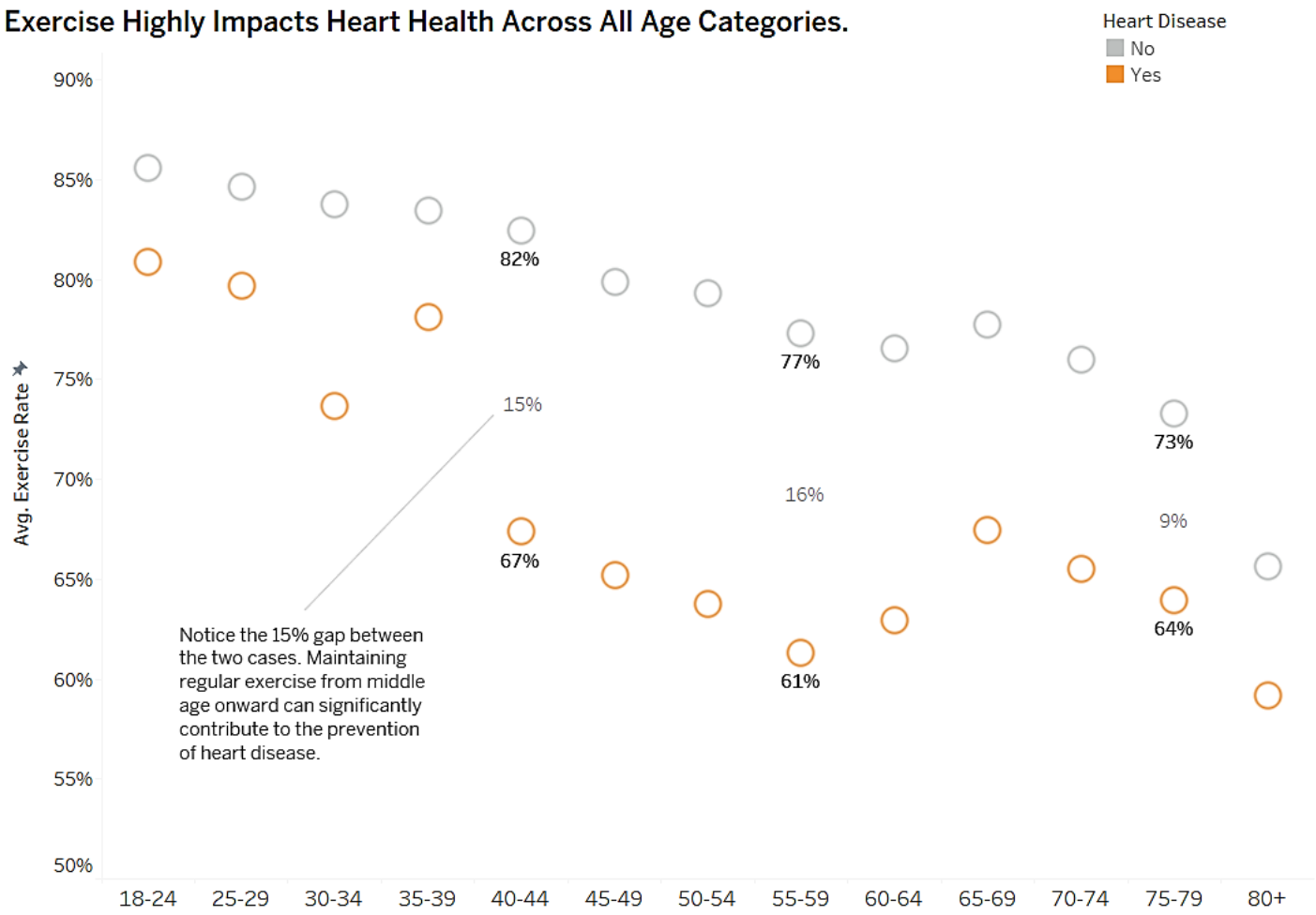
  We know from the descriptive statistics that most people surveyed exercise. The bar graph below shows the average exercise rate for each General Health category. Each color represents a different health category with a lighter color generally signifying a more positive health perception. The graph tells us that exercise leads to better health.

Regular Excercise Helps Improve General Health.



Center for Disease Control - Cardiovascular Disease Risk Prediction Dataset
https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?select=CVD_cleaned.csv

The graph below delivers the same general idea as the previous graph. However, this visual is valuable in showing us the gap in exercise rate between those who report having heart disease (represented in orange) and those who don't (represented in grey) across different age categories.

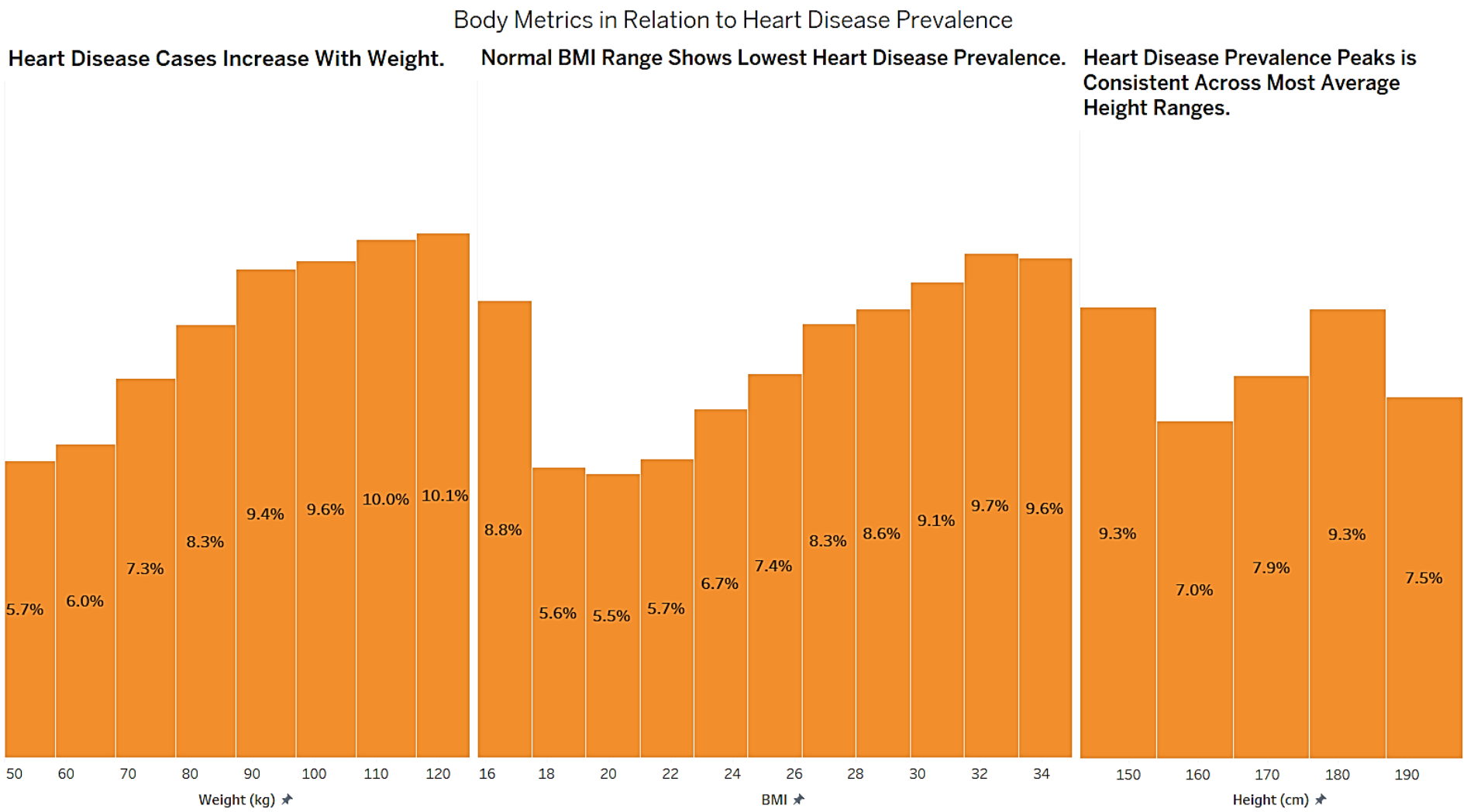


Center for Disease Control - Cardiovascular Disease Risk Prediction Dataset
https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?select=CVD_cleaned.csv

According to research published by Johns Hopkins Medical Department, “people who exercise regularly are less likely to suffer a sudden heart attack,” while exercise has benefits in and of itself, the best way to prevent heart disease is to combine exercise with a healthy diet. The following analyses will explore body metrics and diet in relation to CVD (John Hopkins Medicine, 2023).

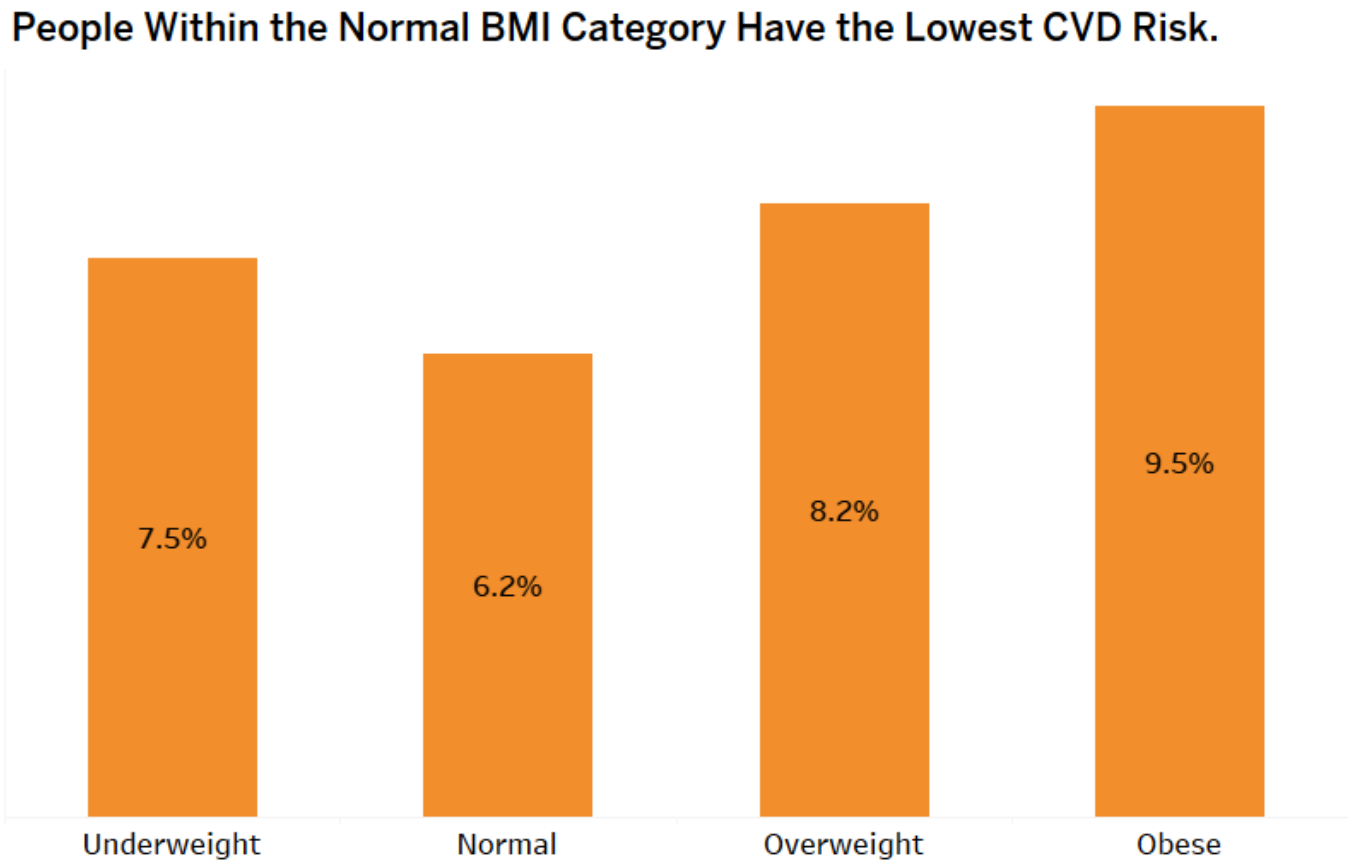
- How do weight, BMI, and height correlate with heart disease risk?

  The visual below shows the general distribution of different body metrics (weight, BMI, and height) color-coded to report only heart disease cases. This helps me understand the body types of individuals in this dataset as it is commonly known to influence health. It was observed that most follow an average weight, height, and BMI but there is a significant number of extreme cases/outliers.


Body Metrics in Relation to Heart Disease Prevalence
Heart Disease Cases Increase With Weight.
5.7%
6.0%
7.3%
8.3%
9.4%
9.6%
10.0%
10.1%
50
60
70
80
90
100
110
120
Weight (kg)
Normal BMI Range Shows Lowest Heart Disease Prevalence.
8.8%
5.6%
5.5%
5.7%
6.7%
7.4%
8.3%
8.6%
9.1%
9.7%
9.6%
16
18
20
22
24
26
28
30
32
34
BMI
Heart Disease Prevalence Peaks is Consistent Across Most Average Height Ranges.
9.3%
7.0%
7.9%
9.3%
7.5%
150
160
170
180
190
Height (cm)

- How does BMI affect heart disease risk?

  The graph below shows the distribution of BMI categories in this dataset through the percent of total count in which the orange color represents the cases of those who reported to have CVD. This graph clearly indicates that individuals are more prone to developing heart disease with an increase in BMI and weight and are also at high risk in an underweight range as compared to being in the normal category.



Center for Disease Control - Cardiovascular Disease Risk Prediction Dataset
https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?select=CVD_cleaned.csv

- How does BMI relate to certain disease rates such as Diabetes?

The heatmap below shows the Average Diabetes rate broken down by BMI Category vs. Age Category. Color shows the average number of cases reported for each age and BMI category with a darker color indicating a higher average percentage of cases.

This visual indicates that individuals suffering from obesity are more likely to develop Diabetes and are thus, at a higher risk of CVD confirming what was observed previously.

Individuals Suffering From Obesity Experience Higher Rates of Diabetes, Which Increase Significantly With Age.

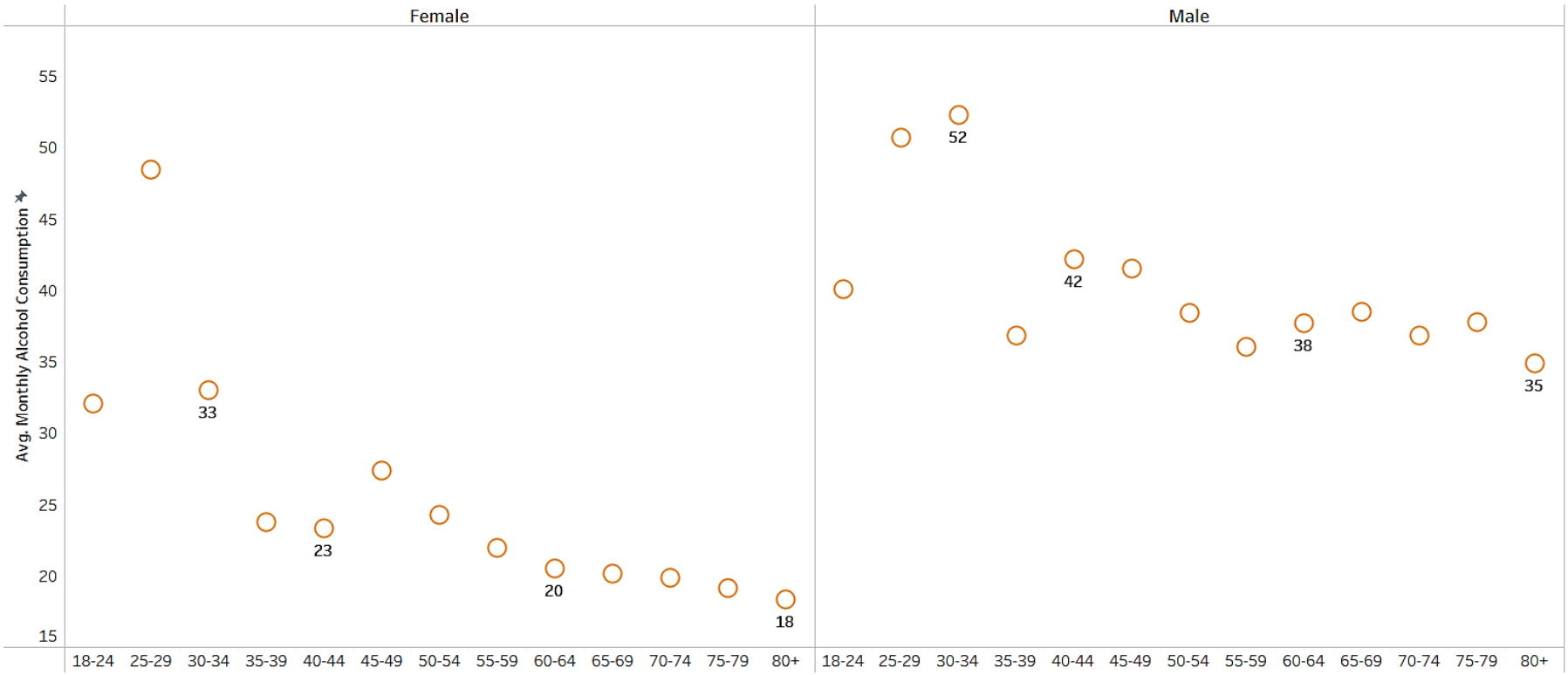
	BMI Category			
Age Category	Underweight	Normal	Overweight	Obese
18-24	1.2%	0.9%	1.1%	3.0%
25-29	2.1%	1.5%	2.0%	5.0%
30-34	2.3%	2.2%	2.6%	7.1%
35-39	3.2%	2.7%	3.7%	10.1%
40-44	4.4%	3.9%	4.9%	13.2%
45-49	4.8%	4.4%	7.2%	17.2%
50-54	7.0%	6.0%	9.2%	20.5%
55-59	9.2%	7.1%	12.0%	25.6%
60-64	8.7%	8.2%	14.2%	28.4%
65-69	8.1%	8.8%	16.5%	30.7%
70-74	8.5%	10.8%	19.6%	33.8%
75-79	7.0%	12.2%	22.1%	34.5%
80+	7.5%	13.0%	20.6%	28.7%

Center for Disease Control - Cardiovascular Disease Risk Prediction Dataset
https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?select=CVD_cleaned.csv

- How do consumption choices relate to cardiovascular disease prevalence across different age groups and genders?

  This visual shifts to individuals' diets by shedding light on how alcohol consumption can influence heart disease development. The graph shows the average consumption across different age categories in both genders. Most importantly, the use of color further helps make the idea clear by showing heart disease prevalence. Males, regardless of CVD prevalence, are observed to drink more than their female counterparts across all age groups.

Out of All Heart Disease Cases, Alcohol Consumption is Notably Higher in Males Across All Age Groups.



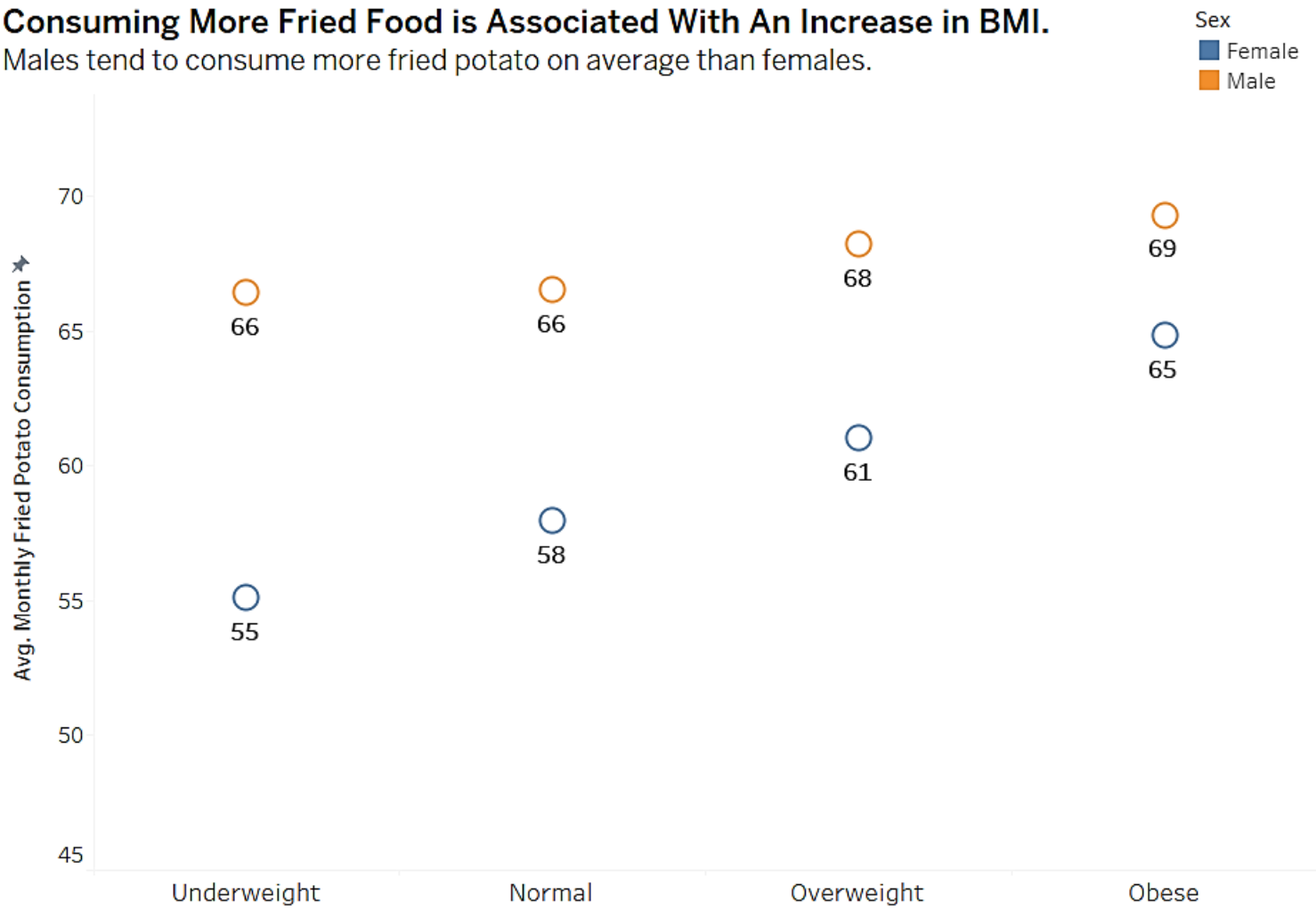
Center for Disease Control - Cardiovascular Disease Risk Prediction Dataset
https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?select=CVD_cleaned.csv

In the graph on the previous page on Alcohol Consumption and CVD, the average monthly alcohol intake is notably higher in males than females, which might be influencing why they tend to have a higher prevalence of CVD.

However, this does not fully explain the gender disparity, indicating that there might be other dietary or non-dietary factors affecting it. This can also suggest potential biological or lifestyle factors beyond alcohol consumption influencing CVD rates, especially considering that those reporting not having heart disease have high alcohol consumption as well.

So far in the analysis, smoking and drinking are habits that influence the rise of CVD prevalence in most individuals, especially males. It would be useful to observe how other dietary choices influence CVD health.

- How do dietary choices correlate with cardiovascular disease prevalence and other factors?



Key Takeaways:

These findings, derived from my analysis, highlight the multifactorial nature of heart disease risk:

- There is an increased likelihood of heart disease with advancing age. Older individuals also have a higher prevalence of conditions like arthritis, cancer, and diabetes.
- Males exhibit a higher prevalence CVD compared to females. This trend is consistent across various analyses, indicating potential biological or lifestyle factors influencing CVD rates differently across genders.
- BMI shows a positive correlation with heart disease, aligning with the understanding that obesity can be a risk factor for cardiovascular issues.
- Regular exercise is negatively correlated with heart disease suggesting a protective role against CVD.
- Moreover, regular health checkups show a positive correlation with better general health, potentially leading to better heart health outcomes.
- Smoking is a clear indicator of heart disease risk, with male smokers having a higher prevalence of heart disease across all age groups.
- Alcohol consumption and fried potato consumption also show a higher average monthly intake in males, which might contribute to their higher prevalence of CVD, but it doesn't fully explain the gender disparity. This data suggests that while diet is important, other lifestyle or genetic factors might play a more significant role in CVD development.

2) ***The University of California at Irvine (UCI) Machine Learning Repository (Cleveland's Database):*** Heart Disease Dataset

Data Cleaning & Pre-processing:

The following is a remodification of the data after imputing the NA values with the mean and changing the variable names:

```
> glimpse(cleveland_raw)
Rows: 303
Columns: 14
$ Age                                 <int> 63, 67, 67, 37, 41, 56, 62, 57, 63, 53, 57, 56, 56, 44, 52, 57, 48, 54, 48, 49, 64, …
$ Sex                                 <int> 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, …
$ ChestPain_Type                      <int> 1, 4, 4, 3, 2, 2, 4, 4, 4, 4, 4, 2, 3, 2, 3, 3, 2, 4, 3, 2, 1, 1, 2, 3, 4, 3, 3, 1, …
$ Blood_Pressure_Level                <int> 145, 160, 120, 130, 130, 120, 140, 120, 130, 140, 140, 140, 130, 120, 172, 150, 110,…
$ Cholestrol_Level                    <int> 233, 286, 229, 250, 204, 236, 268, 354, 254, 203, 192, 294, 256, 263, 199, 168, 229,…
$ Sugar_Level                         <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, …
$ Electrocardiogram_Results           <int> 2, 2, 2, 0, 2, 0, 2, 0, 2, 2, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 0, 0, 0, …
$ Max_Heart_Rate                      <int> 150, 108, 129, 187, 172, 178, 160, 163, 147, 155, 148, 153, 142, 173, 162, 174, 168,…
$ Exercise_Angina_Presence            <int> 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, …
$ Exercise_STsegment_Depression_Value <dbl> 2.3, 1.5, 2.6, 3.5, 1.4, 0.8, 3.6, 0.6, 1.4, 3.1, 0.4, 1.3, 0.6, 0.0, 0.5, 1.6, 1.0,…
$ SlopePeak_Exercise_STsegment_Value  <int> 3, 2, 2, 3, 1, 1, 3, 1, 2, 3, 2, 2, 2, 1, 1, 1, 3, 1, 1, 1, 2, 1, 2, 1, 2, 2, 1, 3, …
$ Visible_Arteries                    <int> 0, 3, 2, 0, 0, 0, 2, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, …
$ Thalassemia_Type                    <int> 6, 3, 7, 3, 3, 3, 3, 3, 7, 7, 6, 3, 6, 7, 7, 3, 7, 3, 3, 3, 3, 3, 3, 7, 7, 3, 3, 3, …
$ Heart_Disease                       <int> 0, 2, 1, 0, 0, 0, 3, 0, 2, 1, 0, 0, 2, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 3, 4, 0, 0, 0, …
>
```

Statistical Summary:

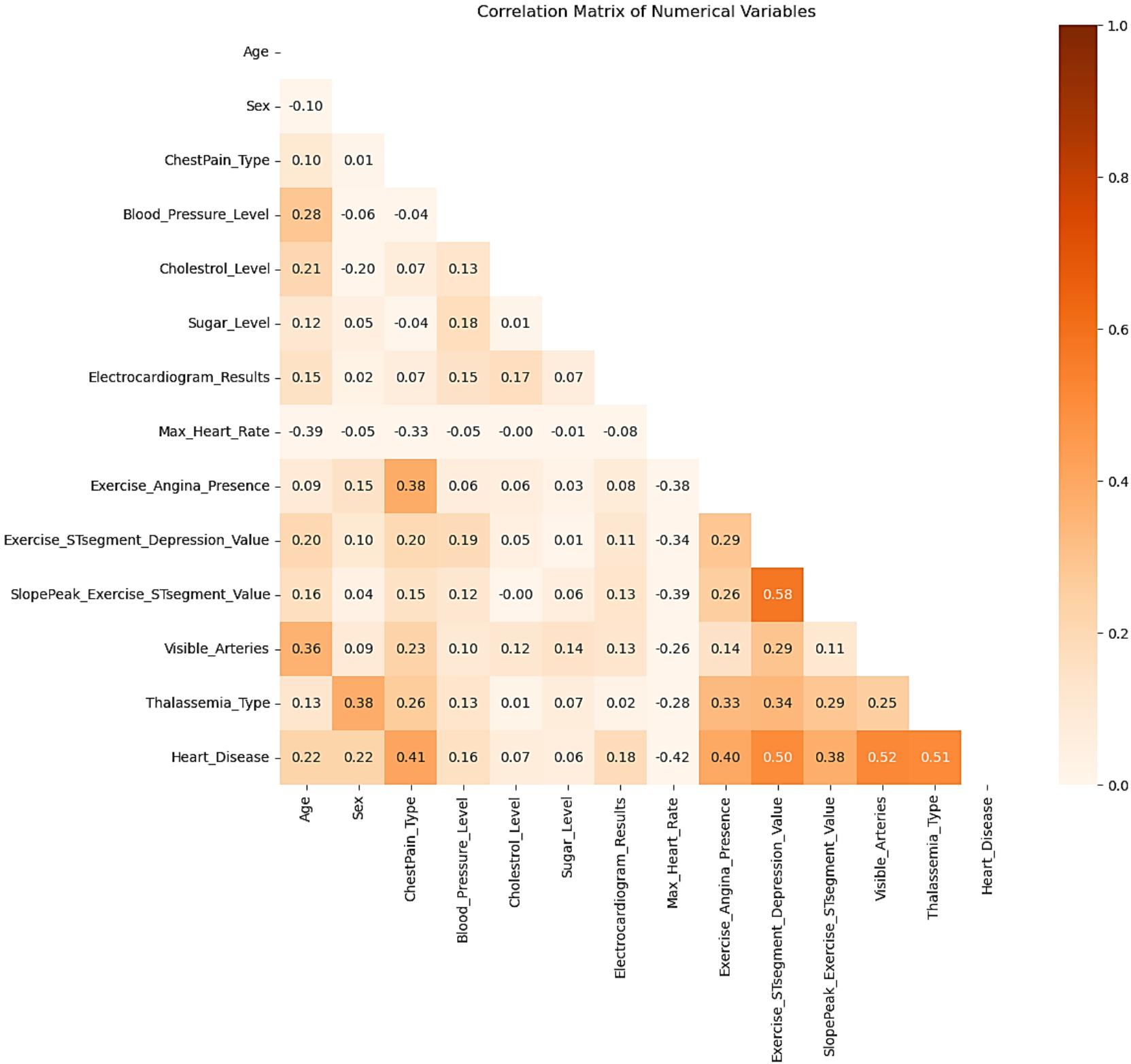
Before performing basic statistical analysis to understand the distribution of each variable, I also ran other simple analyses to confirm that there are no missing or duplicate values in the dataset and that it actually contains 303 cases and 14 variables.

	count	mean	std	min	25%	50%	75%	max
Age	303.0	54.438944	9.038662	29.0	48.0	56.0	61.0	77.0
Sex	303.0	0.679868	0.467299	0.0	0.0	1.0	1.0	1.0
ChestPain_Type	303.0	3.158416	0.960126	1.0	3.0	3.0	4.0	4.0
Blood_Pressure_Level	303.0	131.689769	17.599748	94.0	120.0	130.0	140.0	200.0
Cholestrol_Level	303.0	246.693069	51.776918	126.0	211.0	241.0	275.0	564.0
Sugar_Level	303.0	0.148515	0.356198	0.0	0.0	0.0	0.0	1.0
Electrocardiogram_Results	303.0	0.990099	0.994971	0.0	0.0	1.0	2.0	2.0
Max_Heart_Rate	303.0	149.607261	22.875003	71.0	133.5	153.0	166.0	202.0
Exercise_Angina_Presence	303.0	0.326733	0.469794	0.0	0.0	0.0	1.0	1.0
Exercise_STsegment_Depression_Value	303.0	1.039604	1.161075	0.0	0.0	0.8	1.6	6.2
SlopePeak_Exercise_STsegment_Value	303.0	1.600660	0.616226	1.0	1.0	2.0	2.0	3.0
Visible_Arteries	303.0	0.672241	0.931209	0.0	0.0	0.0	1.0	3.0
Thalassemia_Type	303.0	4.734219	1.933272	3.0	3.0	3.0	7.0	7.0
Heart_Disease	303.0	0.937294	1.228536	0.0	0.0	0.0	2.0	4.0

The table above summarizes the dataset and conveys some important insights that I will consider through my EDA:

- Age: The average age of the participants is approximately 54. The age ranges from 29 to 77, showing a wide distribution of participants' ages.
- Resting Blood Pressure Level: The mean resting blood pressure level is about 131.69 mmHg, which is within the typical range for adults.
- Cholesterol Level: The mean cholesterol level is approximately 246.69 mg/dL. Also, some participants have high cholesterol levels.
- Fasting Blood Sugar: On average, fasting blood sugar doesn't indicate diabetes (mean < 0.15), given that the threshold for diabetes is a fasting blood sugar level of 120 mg/dL.
- Resting Electrocardiogram Results: The mean value is close to 1, indicating that ST-T wave abnormality is the most common finding in the ECG results among the participants.
- Exercise-Induced Angina: On average, exercise-induced angina is present in about 32.67% of the participants.

I will proceed to generate a correlation matrix to further understand how different variables correlate with the target variable and with each other.


Correlation Matrix of Numerical Variables
Age
Sex -0.10
ChestPain_Type 0.10 0.01
Blood_Pressure_Level 0.28 -0.06 -0.04
Cholestrol_Level 0.21 -0.20 0.07 0.13
Sugar_Level 0.12 0.05 -0.04 0.18 0.01
Electrocardiogram_Results 0.15 0.02 0.07 0.15 0.17 0.07
Max_Heart_Rate -0.39 -0.05 -0.33 -0.05 -0.00 -0.01 -0.08
Exercise_Angina_Presence 0.09 0.15 0.38 0.06 0.06 0.03 0.08 -0.38
Exercise_STsegment_Depression_Value 0.20 0.10 0.20 0.19 0.05 0.01 0.11 -0.34 0.29
SlopePeak_Exercise_STsegment_Value 0.16 0.04 0.15 0.12 -0.00 0.06 0.13 -0.39 0.26 0.58
Visible_Arteries 0.36 0.09 0.23 0.10 0.12 0.14 0.13 -0.26 0.14 0.29 0.11
Thalassemia_Type 0.13 0.38 0.26 0.13 0.01 0.07 0.02 -0.28 0.33 0.34 0.29 0.25
Heart_Disease 0.22 0.22 0.41 0.16 0.07 0.06 0.18 -0.42 0.40 0.50 0.38 0.52 0.51
Age
Sex
ChestPain_Type
Blood_Pressure_Level
Cholestrol_Level
Sugar_Level
Electrocardiogram_Results
Max_Heart_Rate
Exercise_Angina_Presence
Exercise_STsegment_Depression_Value
SlopePeak_Exercise_STsegment_Value
Visible_Arteries
Thalassemia_Type
Heart_Disease
1.0
0.8
0.6
0.4
0.2
0.0

From the matrix, I found the insights below to be useful:

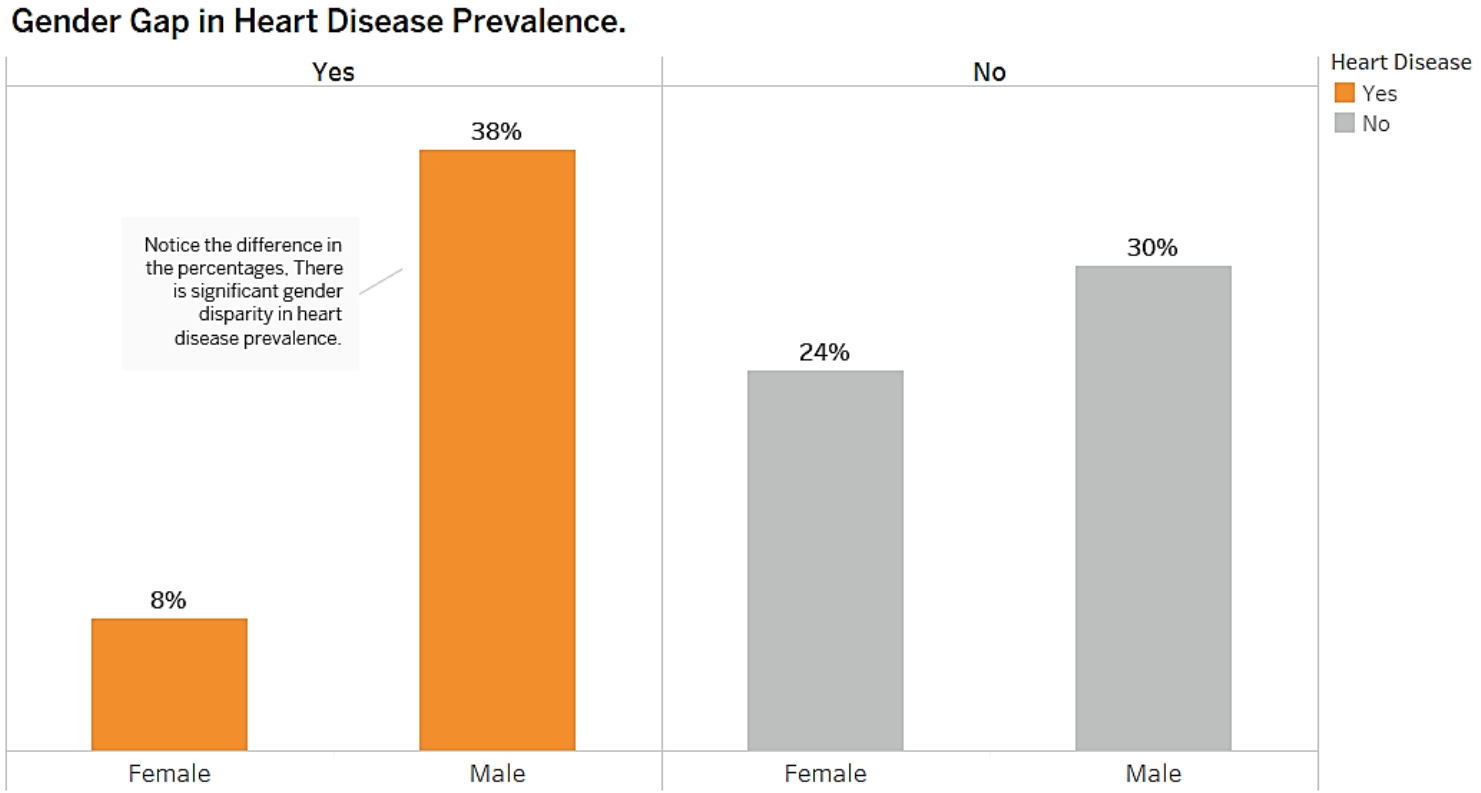
- Maximum Heart Rate Achieved has a negative correlation (-0.42) with heart disease, indicating that a higher maximum heart rate achieved is associated with a lower likelihood of heart disease.
- Exercise ST segment Depression Value shows a moderate positive correlation (0.50) with heart disease, meaning that higher ST segment depression is associated with a higher likelihood of heart disease.
- Slope Peak Exercise ST segment Value and Unblocked Coronary Arteries also show positive correlations (0.38 and 0.52, respectively) with heart disease.
- Unblocked Coronary Arteries have a strong positive correlation (0.36) with age, indicating that older participants tend to have more unblocked coronary arteries.
- Thalassemia Type shows strong positive correlations with Chest Pain Type (0.26) and Unblocked Coronary Arteries (0.29), suggesting some association between these conditions.

Univariate, Bivariate, & Multivariate Analysis:

I started off by exploring if there are similarities in the observations I had in this dataset with the first dataset from the BRFSS.

The first question I wanted to answer is if there is also any gender gap observed in the heart disease cases reported.

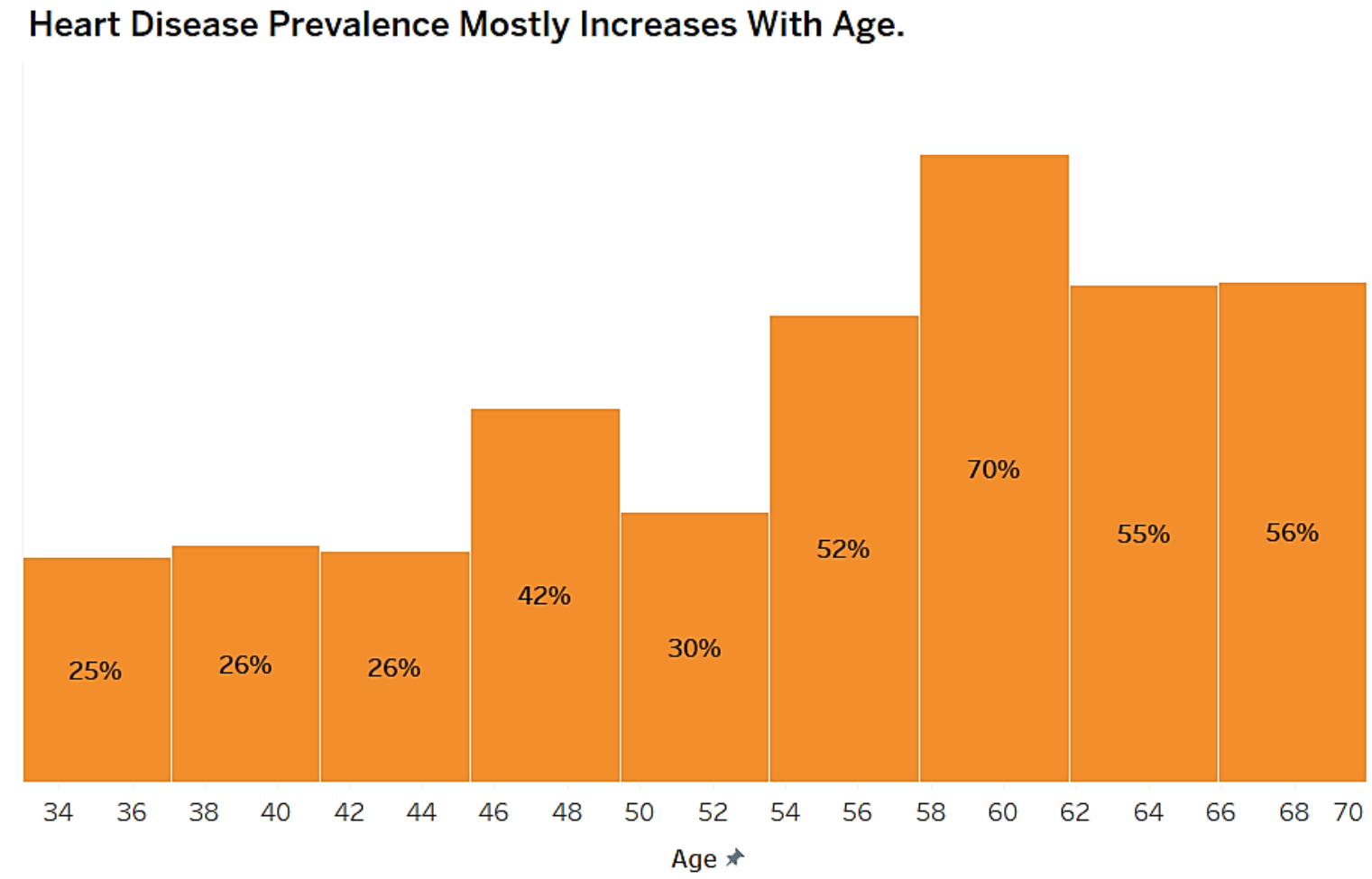
Notice the stark contrast below; there are 54% reported heart disease cases, 38% of which are males.



University of California Irvine - Cleveland's Database Heart Disease Dataset
https://archive.ics.uci.edu/dataset/45/heart+disease

I proceeded to observe if aging is also a factor influencing CVD prevalence in this dataset.

From the visualization below, age is, indeed, a risk factor. As age increases, an individual is more likely to develop CVD.



University of California Irvine - Cleveland's Database Heart Disease Dataset
https://archive.ics.uci.edu/dataset/45/heart+disease

Moving forward, it would be more useful to explore new factors to leverage our insights from the first dataset.

Chest Pain has shown a moderate positive correlation (about 0.4) with Heart Disease as shown from the correlation matrix and other analyses I have performed using Python.

Remember that, there are four chest pain types some of which are more likely to signal heart disease prevalence.

This is the order of how directly related each type is to CVD: Typical Anginal, Non-Anginal, Atypical, and Asymptomatic (Nakas et al., 2019).

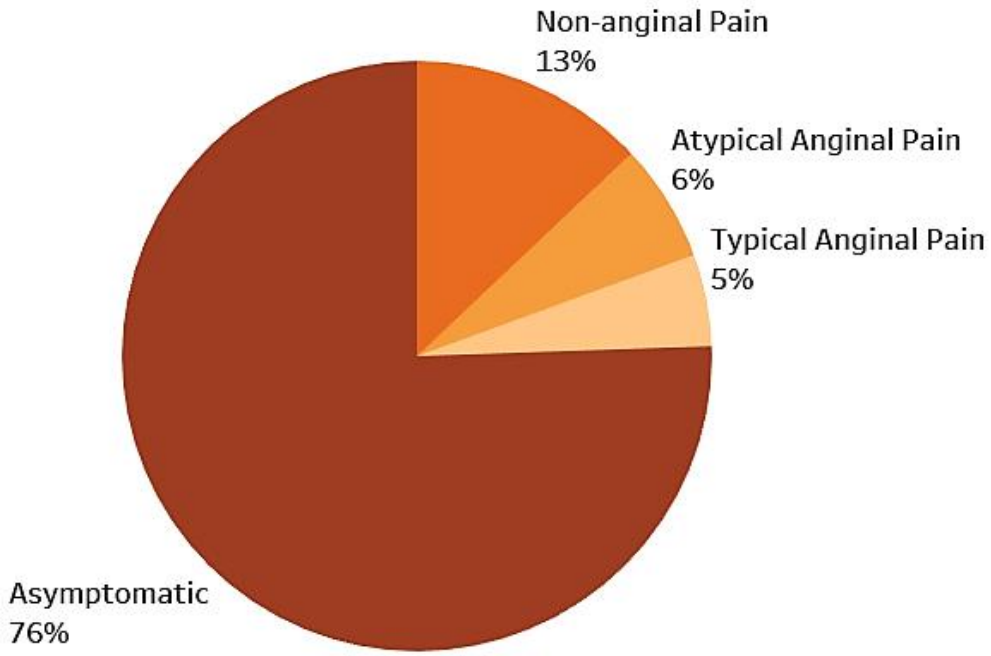
The pie chart on the following page filters the data to only show heart disease cases. The different show details about the chest pain type and the size is relative to the percent of the total count of cases reporting to have each of the 4 chest pain types.

Despite typical anginal pain being most medically proven to be related to coronary issues, asymptomatic pain proves to be the most dangerous and indicative of heart disease.

The largest proportion of individuals report having either Asymptomatic or Non-Anginal pain, which may suggest that other chest pain types are rare cases.

76% of heart disease cases report having experienced asymptomatic chest pain.

This indicates the importance of regular checkups.



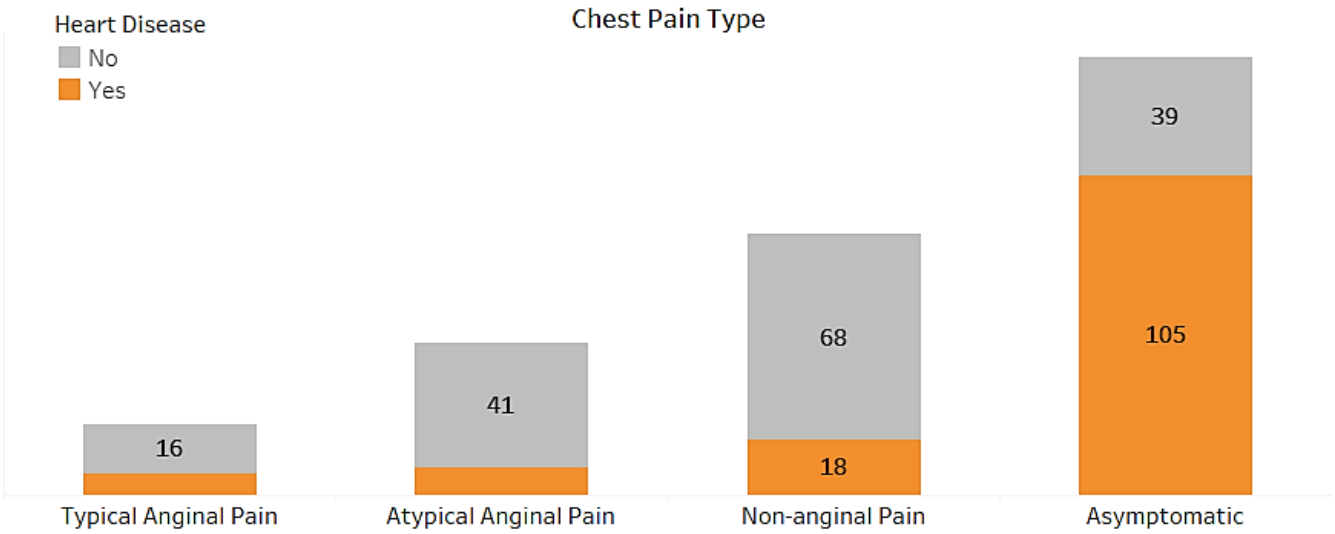
University of California Irvine - Cleveland's Database Heart Disease Dataset
https://archive.ics.uci.edu/dataset/45/heart+disease

Upon further analysis as shown in the stacked bar graph below, asymptomatic pain shows 105 individuals reporting having CVD out of a total of 144 Asymptomatic cases. With that in mind, one can conclude that Asymptomatic pain, despite being the least predictable in regard to heart disease is, indeed, the most dangerous if experienced by an individual.

Asymptomatic pain refers to experiencing no symptoms of CVD. Researchers from the Mayo Clinic reported that despite this being more of a silent heart attack, it is no less deadly than the other types (Mankad, 2022).

This can further suggest that regular checkups are indeed necessary for better overall health.

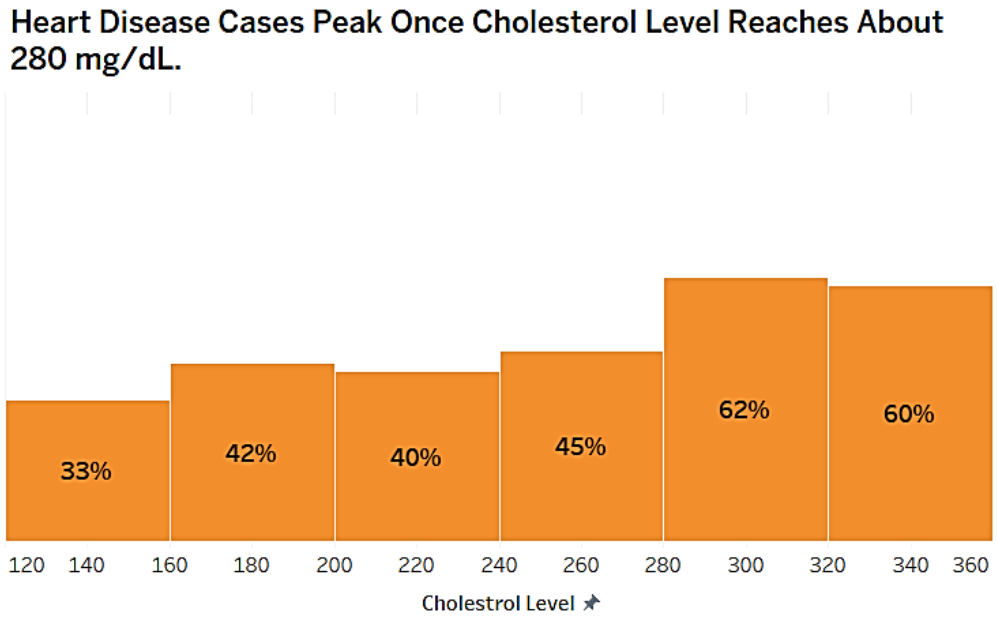
The importance of this graph lies in the idea that an individual could be experiencing signs of heart disease while he/she may not be aware of it due to their experience being asymptomatic.

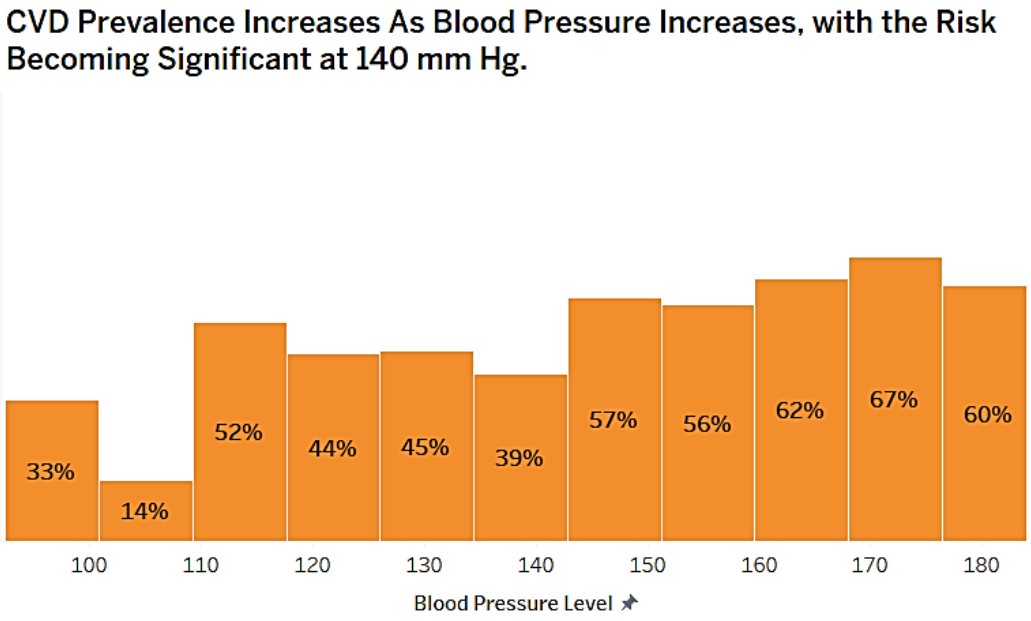


University of California Irvine - Cleveland's Database Heart Disease Dataset
https://archive.ics.uci.edu/dataset/45/heart+disease

I proceeded to explore more on blood pressure and cholesterol levels.

Blood pressure greater than 120/80 mm Hg and a cholesterol level greater than 200 mg/dL are considered high and require attention. From the graph below, higher levels of both of these factors are linked to a higher risk of CVD.





University of California Irvine - Cleveland's Database Heart Disease Dataset
https://archive.ics.uci.edu/dataset/45/heart+disease

Both high cholesterol and blood pressure levels lead to coronary artery damage besides other factors such as diabetes or insulin resistance, not getting enough exercise, and smoking (Mayo Clinic, 2022).

Also, according to the Mayo Clinic staff, “With high cholesterol, you can develop fatty deposits in your blood vessels. Eventually, these deposits grow, making it difficult for enough blood to flow through your arteries. Sometimes, those deposits can break suddenly and form a clot that causes a heart attack or stroke” (2023).

Therefore, it would be useful to explore how the blockage of arteries relates to the reported heart disease cases in this data set. This is important because “High cholesterol has no symptoms. A blood test is the only way to detect if you have it” (Mayo Clinic, 2023).

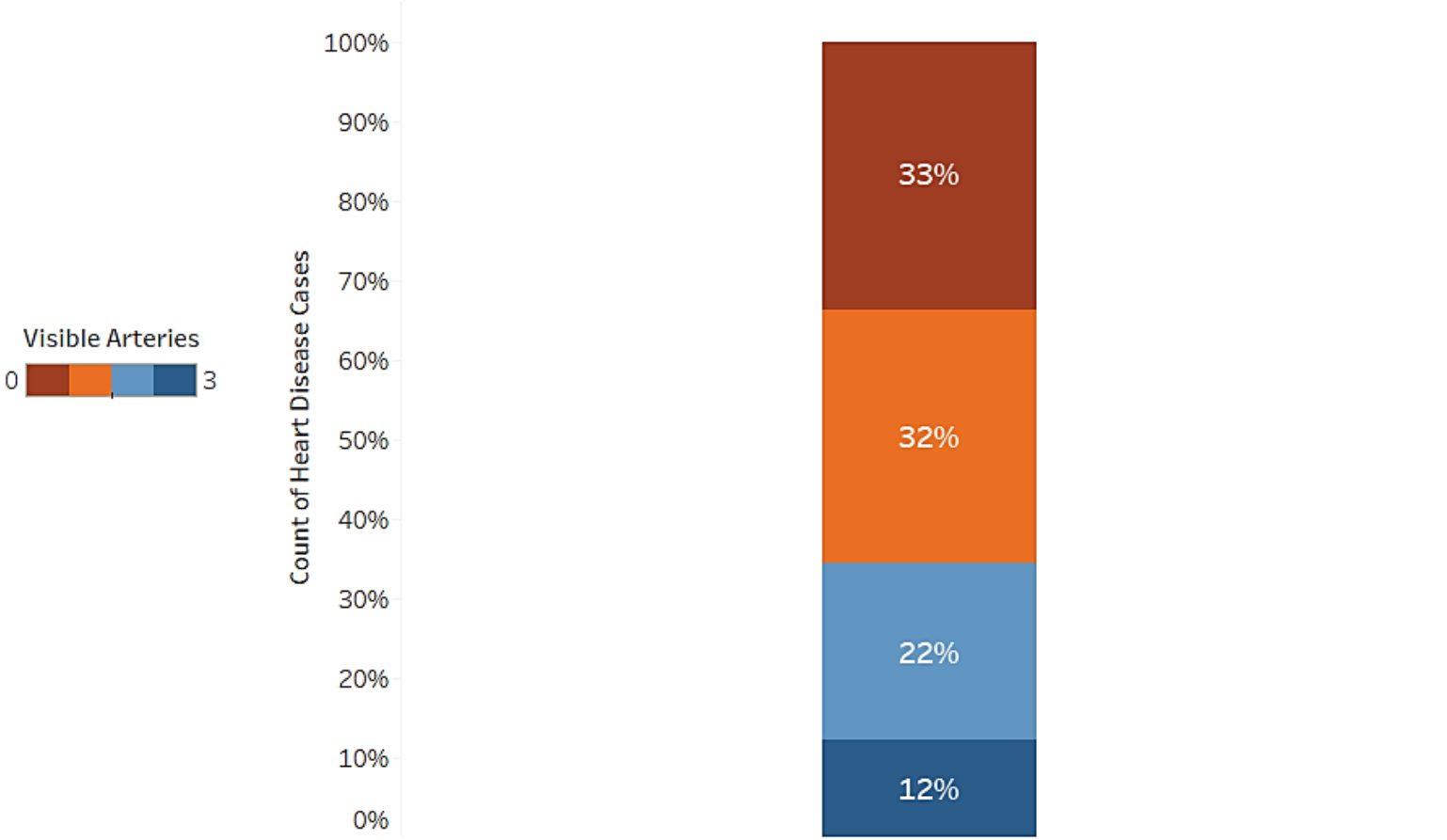
In this data set, the “Visible_Arteries.” variable shows how many arteries were visible during fluoroscopy, an imaging technique that uses X-rays to obtain real-time moving images of the interior of an object, in this case, the human heart.

The fewer the number of visible arteries shown, the higher the risk of CVD due to artery blockage.

The bar graph on the following page visualizes the number of visible arteries and the heart disease cases reported where the highest reported cases of heart disease are also for individuals who showed 0 or 1 visible arteries during their fluoroscopy examination.

The Highest Reported Cases of Heart Disease Are Observed to be In Individuals Who Showed None or Only 1 Visible Artery During Fluoroscopy, Indicating Significant Coronary Blockage.

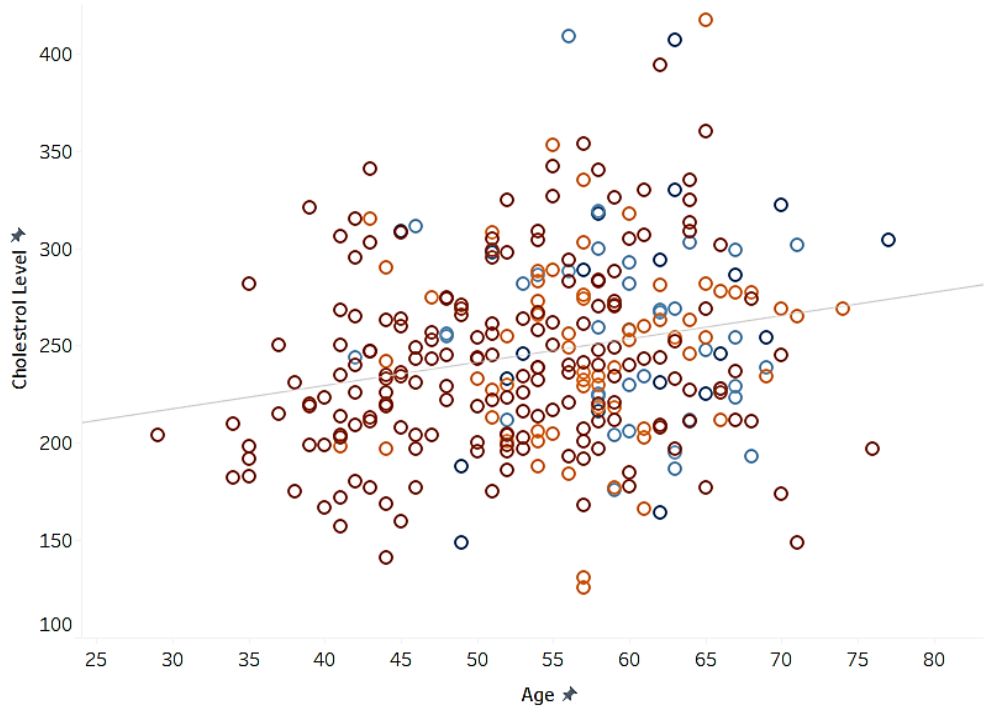
Fluoroscopy is an imaging technique that uses X-rays to obtain real-time moving images of the interior of an object, in this case, the human heart.



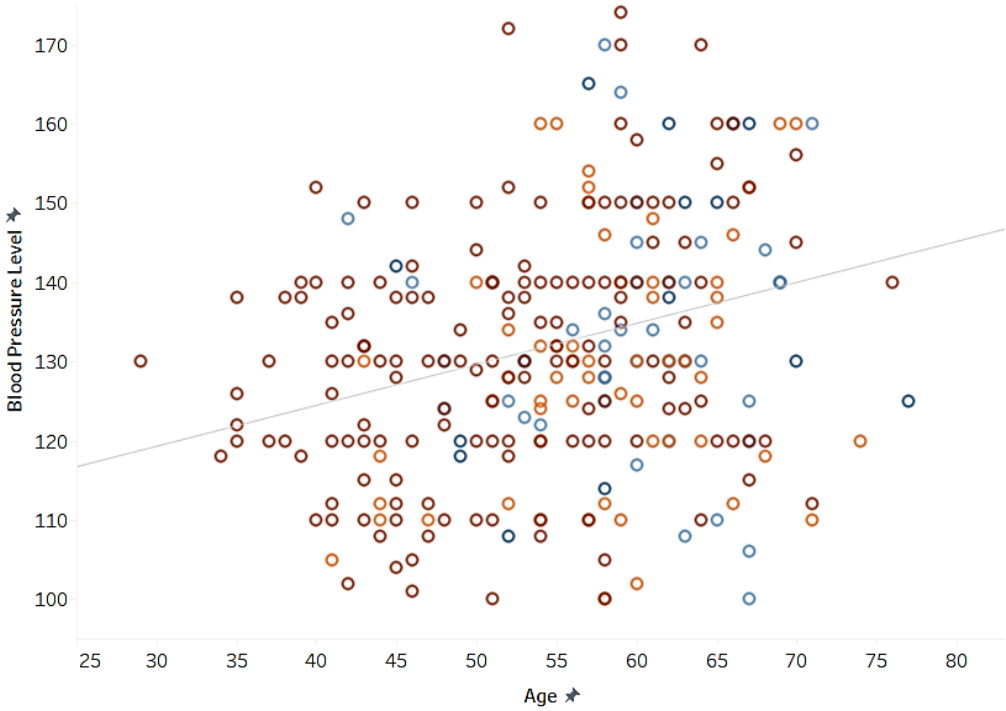
University of California Irvine - Cleveland's Database Heart Disease Dataset
https://archive.ics.uci.edu/dataset/45/heart+disease

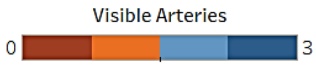
With less visible arteries showing higher CVD cases, let's explore if a similar pattern is evident in relation to cholesterol and blood pressure.

Individuals are more likely to have cholestrol levels increase with age as well as artery blockage due to plaque build up.



Older individuals not only have higher blood pressure on average but also a greater likelihood of artery blockage.





University of California Irvine - Cleveland's Database Heart Disease Dataset
https://archive.ics.uci.edu/dataset/45/heart+disease

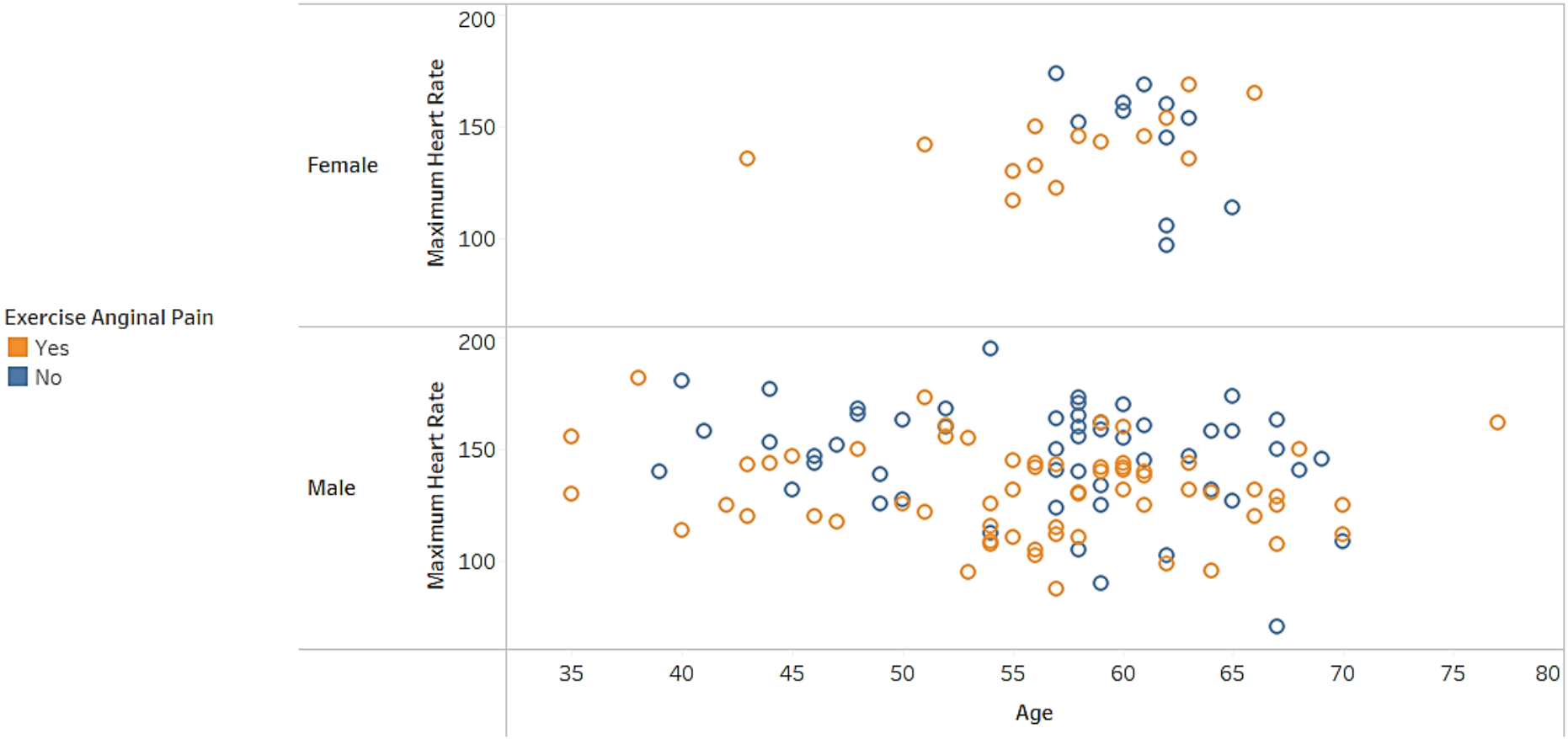
The scatter plots on the previous page show that some of the two significant causes of heart disease, high blood pressure and cholesterol, increase as we age. Not only that but also most cases that show high blood pressure or cholesterol report showing no visible arteries during fluoroscopy examinations, indicating artery blockage. This can significantly decrease blood flow to the heart causing heart disease or other coronary-related issues.

Now that we know that less exercise or physical activity puts us at a higher risk of heart disease.

I thought it would be interesting to see how it relates to the maximum heart rate (MHR) achieved during exercise. This is the highest heart rate an individual can achieve through exercise stress without causing severe physiological distress. It is " a range of numbers that reflect how fast your heart should be beating when you exercise" (John Hopkins Medicine, 2019).

During exercise, a healthy heart typically increases the rate of pumping more blood and oxygen to the muscles. If the heart doesn't increase rate properly, or if the MHR is lower than what's expected for a person's age and fitness level, it may signal underlying cardiovascular issues (John Hopkins Medicine, 2023).

Out of all heart disease cases, males are not only more likely to develop CVD but are also more likely to experience a decline in maximum heart rate with age.



University of California Irvine - Cleveland's Database Heart Disease Dataset
https://archive.ics.uci.edu/dataset/45/heart+disease

From the scatter plot on the previous page, one can clearly observe that exercise seems to be more beneficial for males considering that they are at a higher risk of maximum heart rate decline with age.

Key Takeaways:

- Age is positively correlated with heart disease, indicating the need for increased monitoring and preventive measures for older age groups.
- Both high blood pressure and high cholesterol are linked to an increased risk of heart disease. Regular monitoring and lifestyle modifications are recommended for individuals with elevated levels.
- Certain types of chest pain, particularly asymptomatic pain, are highly indicative of heart disease. This suggests the importance of checkups, even in the absence of symptoms.
- The presence of exercise-induced angina in about a third of the participants points to the necessity of stress testing and lifestyle advice for individuals with this condition. Exercise Anginal Pain also increases as the maximum heart rate decreases which can cause reduced blood flow to the heart.
- The negative correlation between maximum heart rate achieved and heart disease suggests promoting physical activity and exercise as preventive measures.

- Considering the dataset, there is an implication of the relevance of diabetes management in the context of heart disease prevention.
- Given the variety of factors influencing heart disease, regular health checkups are essential for early detection and management.
- As seen in both data sets, heart disease manifests differently in males and females, with males proving to need more attention to manage their heart health.

IV. Final Takeaways & Recommendations

While certain risk factors for heart disease lie beyond our control, such as advancing age, male gender, and inherited genetic predispositions, it is crucial to focus on what we can influence. Unmodifiable elements like family history and potentially overlooked genetic markers extracted from data analysis highlight the significance of vigilance in heart health management.

There are actionable insights deduced from the data which are key to mitigating heart disease risks. These include:

- Stop Smoking: Smoking is a clear indicator of heart disease risk. If you smoke, seek help to quit. Avoid exposure to secondhand smoke as well.
- Limit Alcohol Consumption: Limit alcohol to moderate levels.
- Checkup Regularly: Regular monitoring of blood pressure, cholesterol levels, and other risk factors in case of pain or other symptoms absence can help in early detection and management of heart disease.
- Manage Blood Pressure: Regular monitoring and a balanced diet aid in maintaining optimal blood pressure levels.
- Check Cholesterol Level: Routine screenings and dietary adjustments help keep cholesterol in check.
- Increase Physical Activity: Individuals as they age, particularly males, are Encouraged to maintain regular physical activity, which can include anything from brisk walking, cycling, and swimming, to more structured exercises. Regular

exercise not only combats obesity but also aids in maintaining a healthy Body Mass Index (BMI), contributing to overall wellness.

- Maintain Healthy Body Weight: This involves balancing the calories consumed with calories burned through physical activities. Avoid obesity and overweight, which are significant risk factors for heart disease.
- Balance Diet: Opting for a diet rich in nutrients and low in fried and fatty foods benefits heart health and reduces sugar consumption to decrease the risk of diabetes and obesity, both of which are linked to heart disease.

By integrating these practices into daily life, we can significantly improve our heart health and avoid many problems.

V. References

American Heart Association. (2017, May 31). What is cardiovascular disease? https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease

Baystate Health. (2022, February 7). The history of heart disease dates back to Egyptian pharaohs – What will the future look like? https://www.baystatehealth.org/news/2022/02/history-of-heart-disease

Cardiovascular Diseases Risk Prediction Dataset. (2023). Kaggle. https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?select=CVD_cleaned.csv

Centers for Disease Control and Prevention. (2021). *Behavioral Risk Factor Surveillance System*. CDC. https://www.cdc.gov/brfss/index.html

Centers for Disease Control and Prevention. (2021). Behavioral risk factor surveillance system. https://www.cdc.gov/brfss/index.html

Lupague, R. M. J. M., Mabborang, R. C., Bansil, A. G., & Lupague, M. M. (2023). Integrated machine learning model for comprehensive heart disease risk assessment based on multi-dimensional health factors. European Journal of Computer Science and Information Technology, 11(3), 44-58.

Mankad, R. (2022, May 22). Silent heart attack: What are the risks? Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/heart-attack/expert-answers/silent-heart-attack/faq-20057777

Mayo Clinic. (2023, January 11). High cholesterol: Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/symptoms-causes/syc-20350800

Mayo Clinic. (2022, May 25). Coronary artery disease: Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/coronary-artery-disease/symptoms-causes/syc-20350613

Nakas, G., Bechlioulis, A., Marini, A., Vakalis, K., Bougiakli, M., Giannitsi, S., & Naka, K. (2019). The importance of characteristics of angina symptoms for the prediction of coronary artery disease in a cohort of stable patients in the modern era. Hellenic Journal of Cardiology, 60(4), 241–246. https://doi.org/10.1016/j.hjc.2018.06.003

National Institute on Aging. (2018). Heart health and aging. https://www.nia.nih.gov/health/heart-health-and-aging