Heart Disease Diagnostic Analysis and Visualization

Table of Contents

Problem Statement

As it is rightly said, 'Health is Wealth'. We have realized this fact in the pandemic time after witnessing the brute effects of Covid-19 on people of all age groups. Apart from this, another major contributor to the death rate is heart related diseases.

Located in the chest region of the body, the heart beats at around 80 times per minute. Even though it is just the size of an average human fist, it is the strongest muscle which continuously to pump blood to the body organs, even at rest.

Heart diseases have been known to take a major toll of people's lives. As a layman, we may feel that the common factors for heart related diseases are cardiac arrest or blockages. But the dataset under analysis describes multiple different medical parameters associated with the heart and their typical values. We will be analyzing the relationships between them and study the implications of changes in those parameters. In this project, we will be incorporating most trending and powerful BI tool namely Tableau.

Objective and Benefits

- Objective:
  - The dataset contains the records for the patients and their medical parameters details and the target variable whether they will suffer from heart disease or not.
  - The aim of this project is to use the given data and perform ETL and data analysis to infer key metrics and patterns in the dataset. In addition to this, different visualizations are developed to depict meaningful relationships.

- Benefits:
  - The data analysis will reveal some common and unique patterns in the dataset related to the medical parameters.
  - Data visualizations will enhance the understanding of the effect of the high or low of these features on the chances of heart rate and give a better chance of prediction.

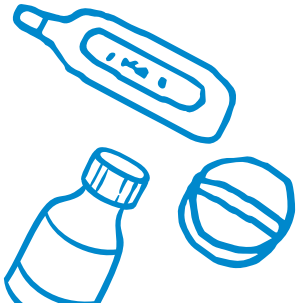
Steps Followed

Data Extraction

Data Exporting

Data Analysis

01 02 03 04 05 06



Data Preprocessing

Data Loading and Modification

Deployment

Steps Followed contd...

1) Data Extraction: This step involves extracting the data from different sources relevant to the problem statement or obtaining data from the client.

2) Data Preprocessing: Once the raw data is obtained, we need to ensure that the data is free from errors. We perform Exploratory Data Analysis followed by Data Cleaning which involves imputing missing values, removing duplicates, finding anomalies or outliers and treating them.

3) Data Exporting: The preprocessed data is exported to a .csv file to be used for analysis.

4) Data Loading and Modification: The preprocessed data in .csv file is loaded into the Tableau Desktop for analysis purpose and modified for simplicity purpose.

5) Data Analysis: Once the data is loaded, we perform the data analysis using Tableau features and store the visualizations in Tableau worksheets.

6) Deployment: The prepared visualizations are deployed on the Tableau Online Software where they will be available publicly.

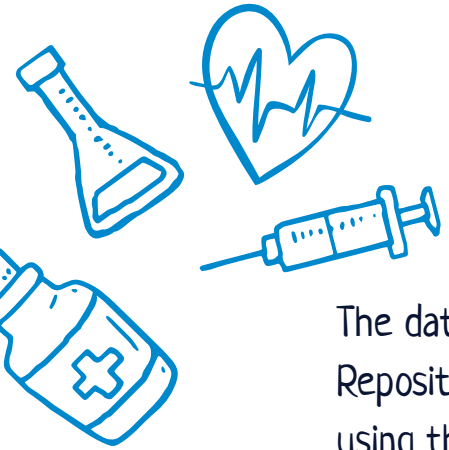
Data Sharing Agreement

Dataset File Name: 'heart_disease_dataset.csv'

Number of columns: 14

Column Names

Column Data Types

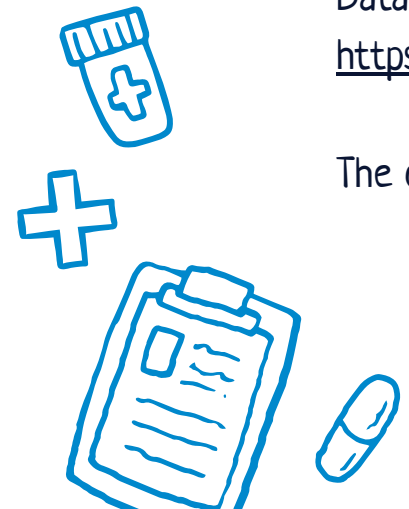


Data Extraction

The dataset used for analysis is the Heart Disease dataset provided by the UCI Repository. It actually contains 76 attributes out of which only 14 are used. We will be using the Cleveland dataset.

Dataset source:
https://archive.ics.uci.edu/ml/datasets/Heart+Disease

The dataset is available is a .csv file - 'heart_disease_dataset.csv'



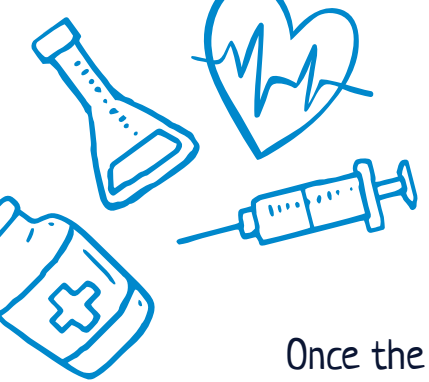
Data Preprocessing

After Exploratory Data Analysis carried out on the dataset we have certain observations with the dataset.

1) There is no column in the dataset with missing values.

2) There are a few columns which actually contain categorical values but have been incorrectly labeled as numeric. As a part of data preprocessing we will convert them to categorical values.

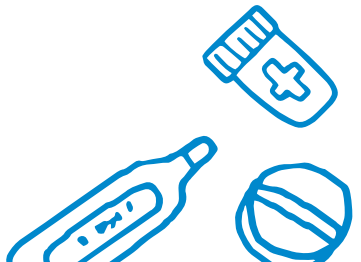
3) There are a few columns which have unusual values / outliers. We will impute these values with the median / mode value obtained from the remaining values of the columns.



Data Exporting

Once the data has been cleaned in the data preprocessing stage, we will export the cleaned dataset into a new file with .csv format.

The new dataset file has name - 'preprocessed_heart_disease_dataset.csv'



Dataset Loading and Modification

The exported .csv dataset file – 'preprocessed_heart_disease_dataset.csv' will be imported into Tableau Public Desktop. Since this a .csv file, we will choose the 'Microsoft Excel' file option when prompted to import dataset into Tableau.

Since the dataset contains many categorical columns which store the categories in the form of integers we will convert these numbers into meaningful phrases which will be understandable to the viewer and also easy to understand the terms used in the visualizations.

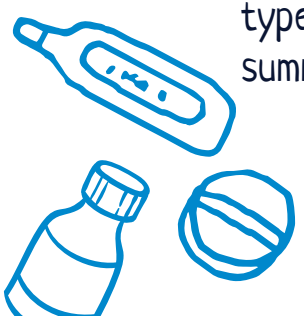
These phrases are called as 'Aliases' and will be provided to the values available in the categorical columns as part of data modification.

Data Analysis

Once the data has been loaded into the Tableau Desktop software, we perform the analysis for the various medical parameters provided in the dataset and study the relationship between them.

Based on these patterns, we try to draw approximate inferences about the data provided on the basis of visualizations created.

We have made use of different aspects of Tableau like different charts, labeling, aliases, filtering, and actions based on user choice. We have created separate worksheets for each type of visualization which contains the chart and a caption as well which contains the summary of analyses drawn.



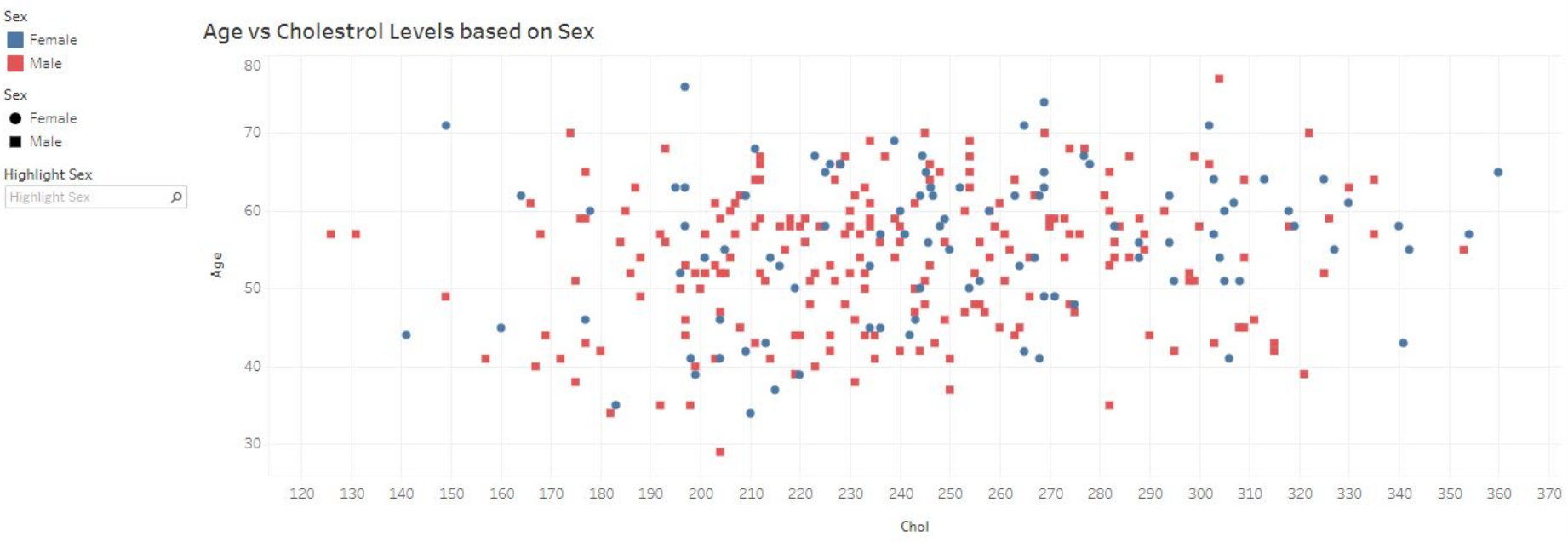
Deployment

All the different worksheets that have been created are compiled together into a Tableau workbook. Each worksheet is named based on the type of visualization performed in the chart.

When we save all the worksheets on Tableau Desktop software, it connects to the Tableau Public Software via personal email id and credentials. All these sheets are uploaded onto the Tableau Public Software on personal profile and this viz. is visible to public.

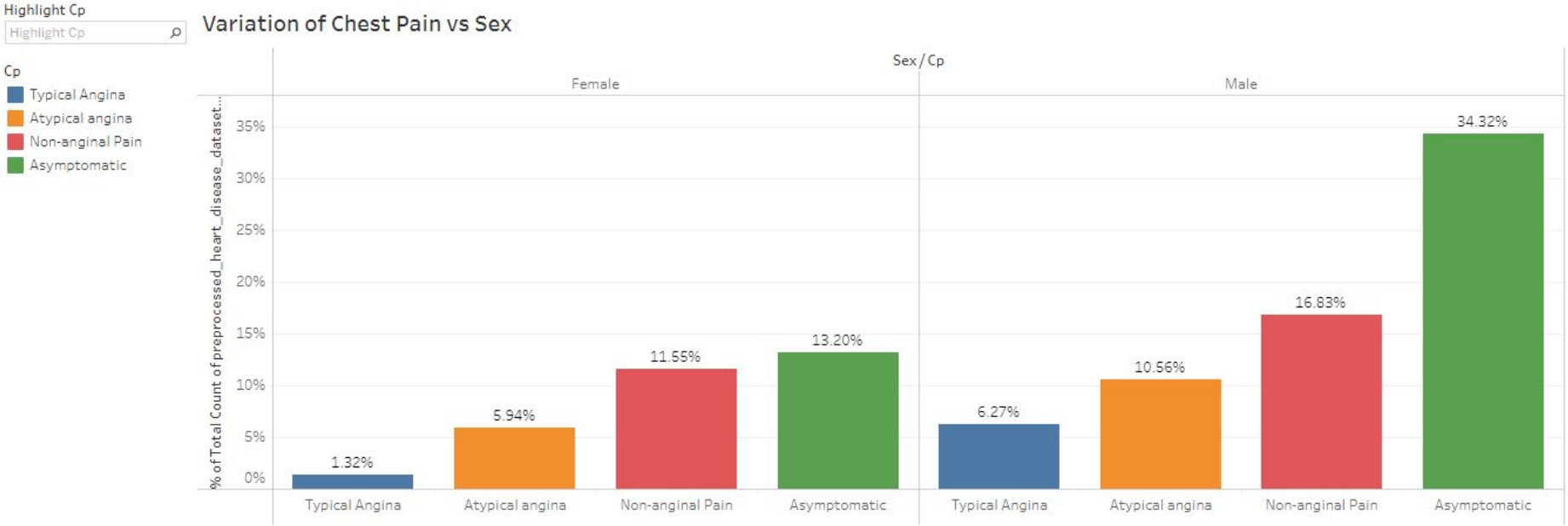
The link for the worksheets is at :
https://public.tableau.com/views/HeartDiseaseDiagnosticAnalysis/AgeandChol?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

Visualizations



Inference:

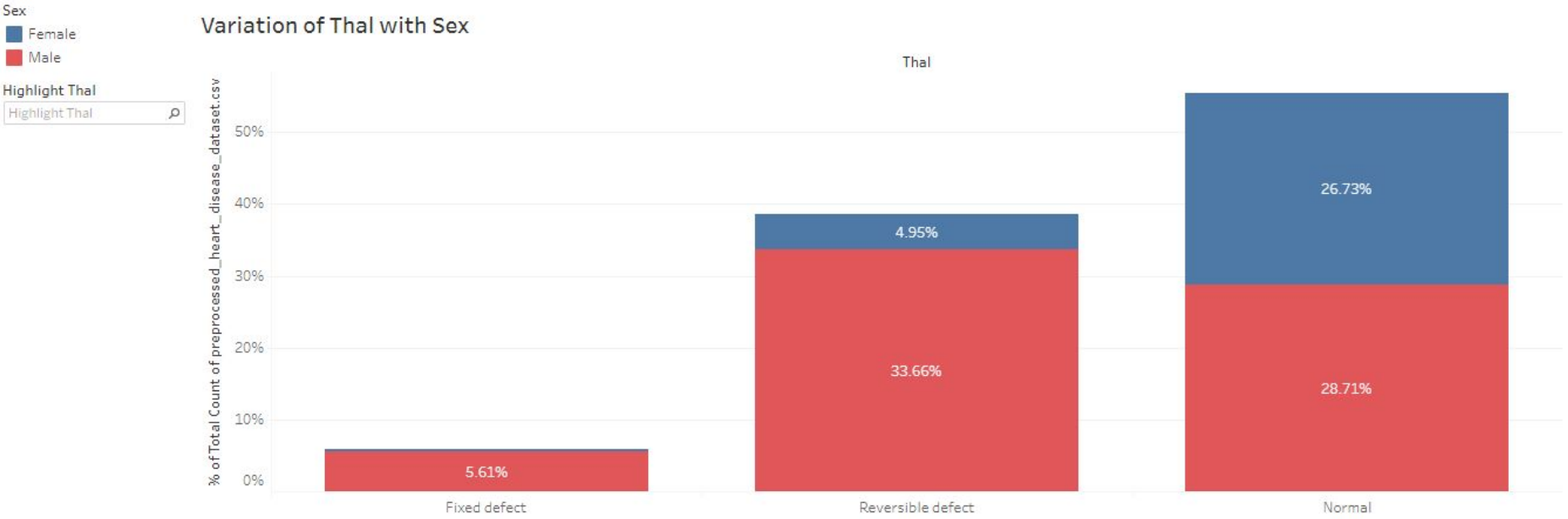
1) The scatterplot of Age vs Chol shows us that there is no specific relationship between the Age and Chol. Low aged people also show moderate to high Chol levels and vice versa.



Inferences:

1) When comparing the individual Chest Pain type percentages for Males and Females, it is clear that Males are more prone to the Chest Pain symptoms.

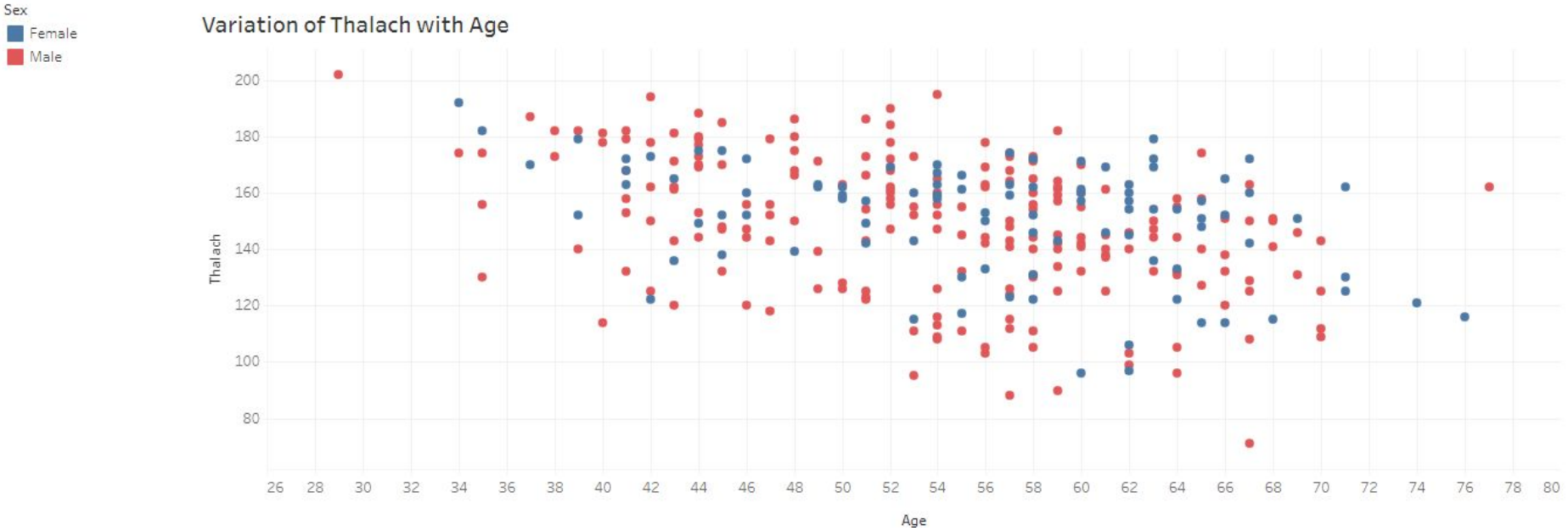
2) The Asymptomatic type of Chest Pain is the one which affects both Males and Females in the higher proportion. But in Males it surpasses the rate as compared to Females by more than double.



Inferences:

1) We observe that the Normal type of Thalassemia is common in both Male and Female and is approximately equal chances to contract the disorder.

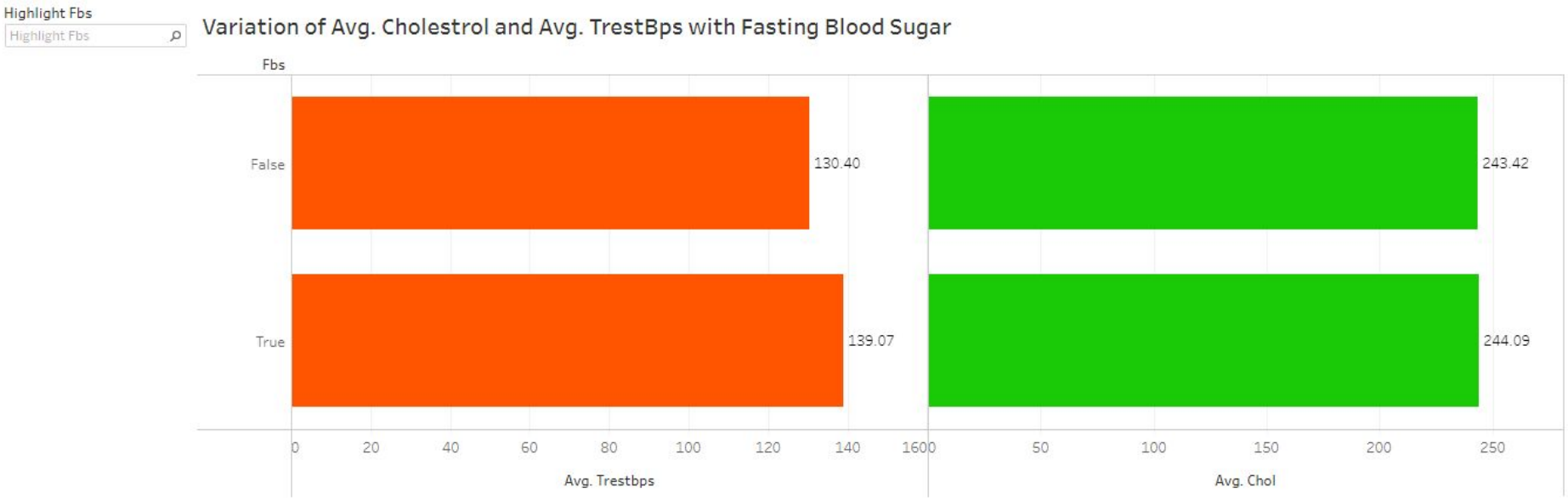
2) As opposed to the Normal type, the Fixed and Reversible types of Thalassemia are more likely to be found in the Male sex. The proportion of the Males having these disorders is much greater than the Females.



Analysis:

1) We can see a general negative correlation between the Age and Thalach parameters.

2) If we highlight the Sex attribute, we will be able to see an approximate downward trend indicating that the as the age increases the thalach (maximum heartrate) starts decreasing.

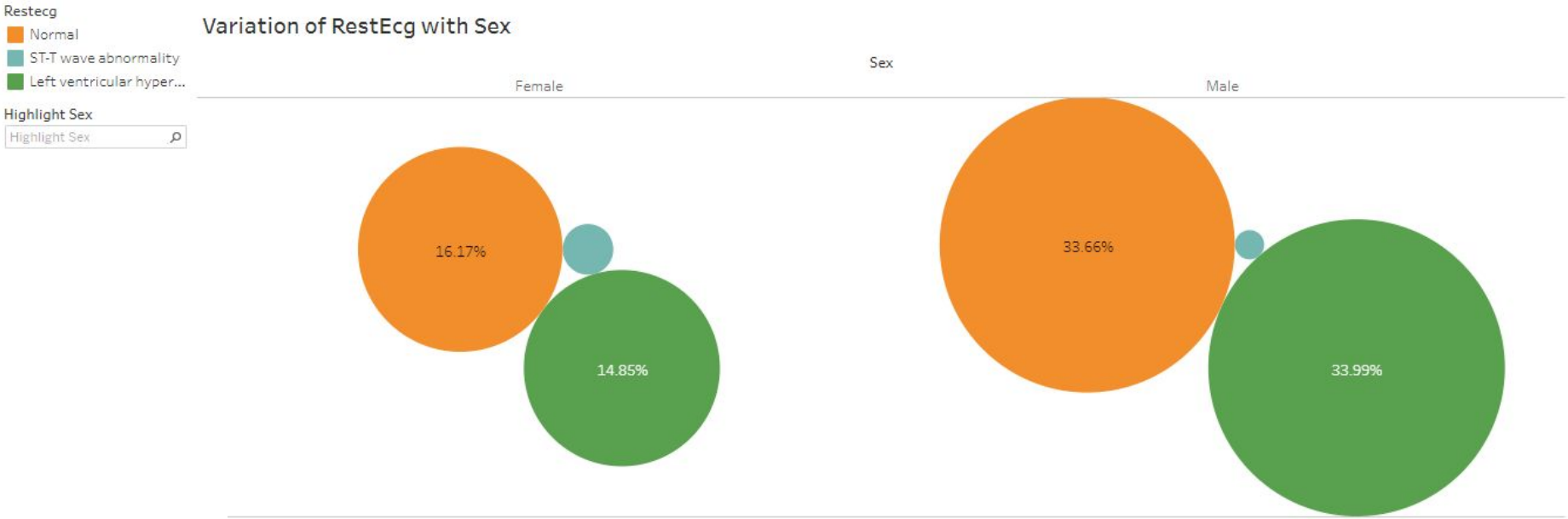


Inferences:

1) This plot shows us the variation and relationship between the Cholesterol levels, Fasting Blood Sugar and the Resting Blood Pressure.

2) When the Fasting Blood Sugar of a person > 120 mg/dl, then the Cholesterol level is slightly higher than when the Fasting Blood Sugar level < 120 mg/dl. In this case the Blood pressure of the person is higher than the Blood Pressure of the person whose Blood Sugar Level < 120 mg/dl.

3) This indicates that when the person fasts, the blood sugar level is less as the body uses the existing body sugar and the blood pressure is also low due to low activity.

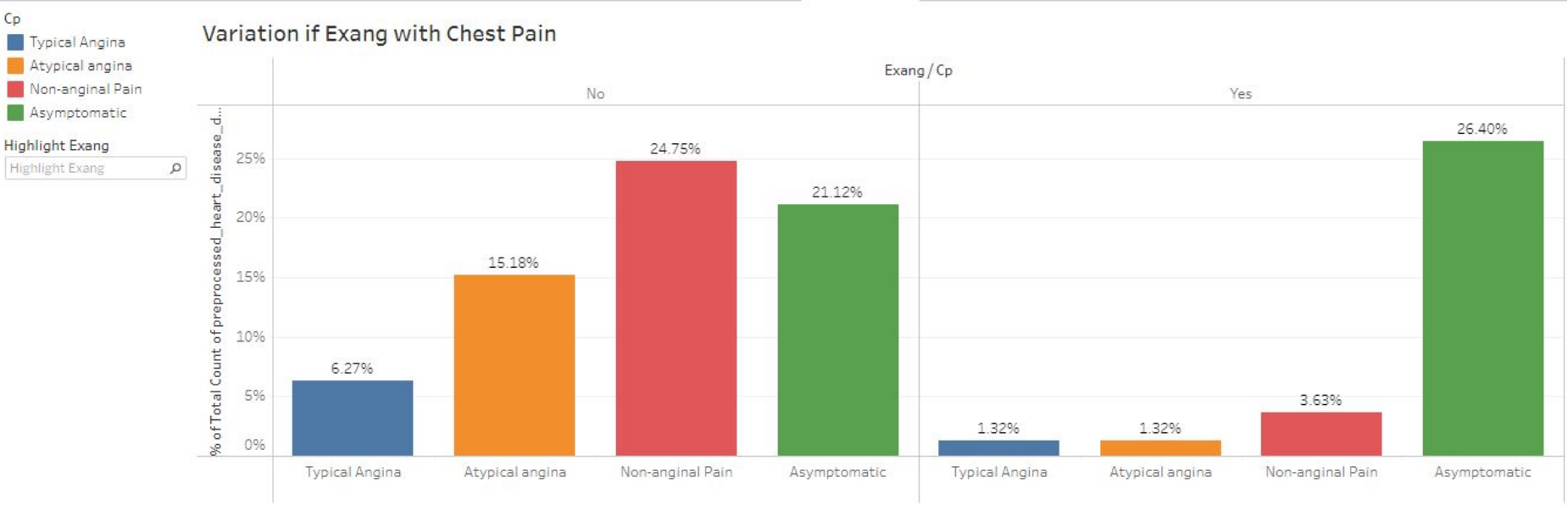


Inferences:

1) A very small percentage of the total population ~2% shows the ST-T wave abnormality.

2) The percentage of Females with Normal RestEcg is more as compared to percentage of Females with Left Ventricular Hypertrophy.

3) The percentage of Males with Normal RestEcg is almost same as compared to percentage of Males with Left Ventricular Hypertrophy.

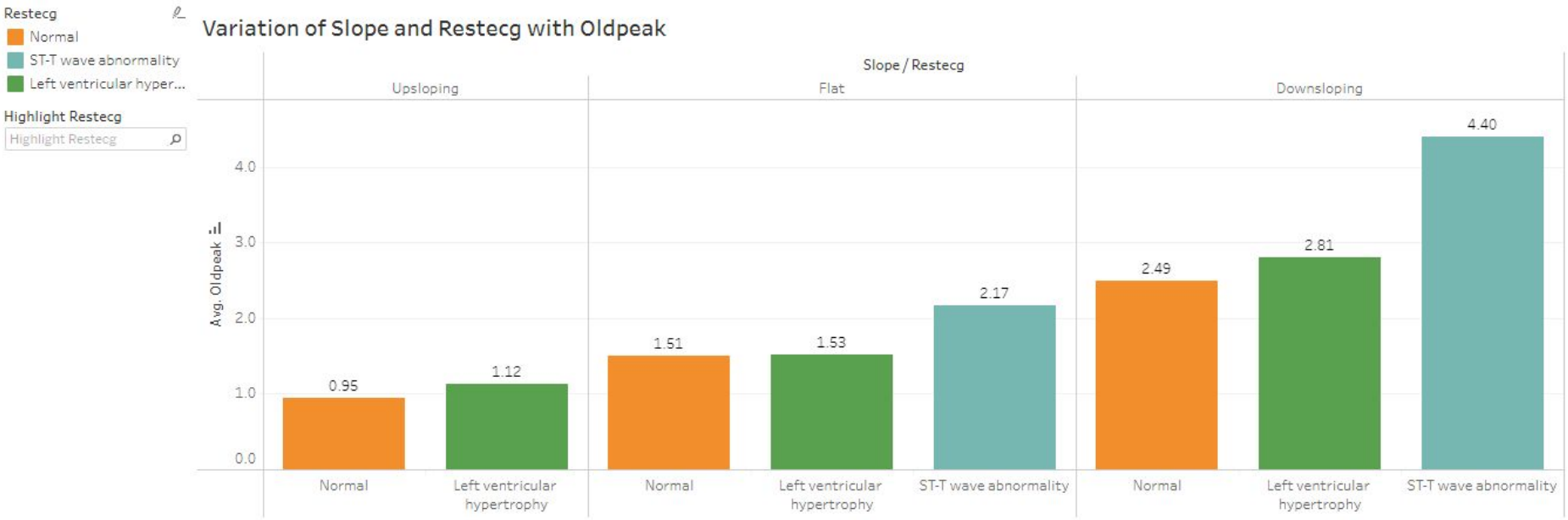


Inferences:

1) This is a graph showing the variation of the exercise induced angina against Chest Pain type and the count in each category.

2) It can be clearly seen that the Asymptomatic angina type has shown increase in those people for whom the exang variable is Yes. Meaning that people who have performed exercise have a higher chance of experiencing the Asymptomatic angina type.

3) The remaining Chest Pain types have shown a great downfall when the people have performed exercise.

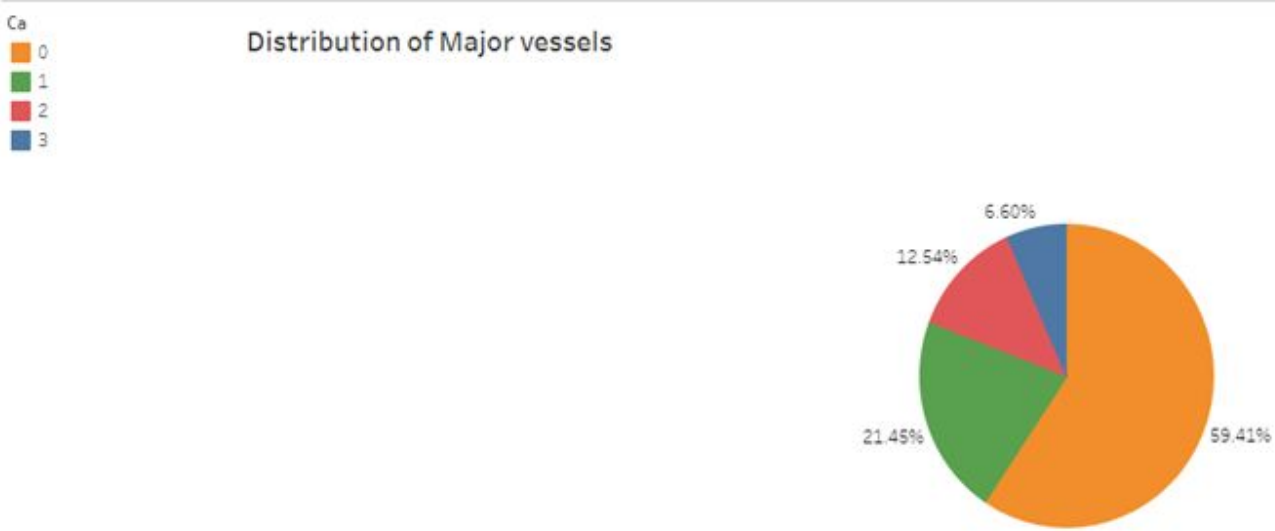


Inferences:

1) This graph shows the variation of the Slope and RestEcg against Oldpeak value.

2) We can infer that the people for whom the Slope is Upsloping, they experience the Normal or Left ventricular hypertrophy type of Restecg.

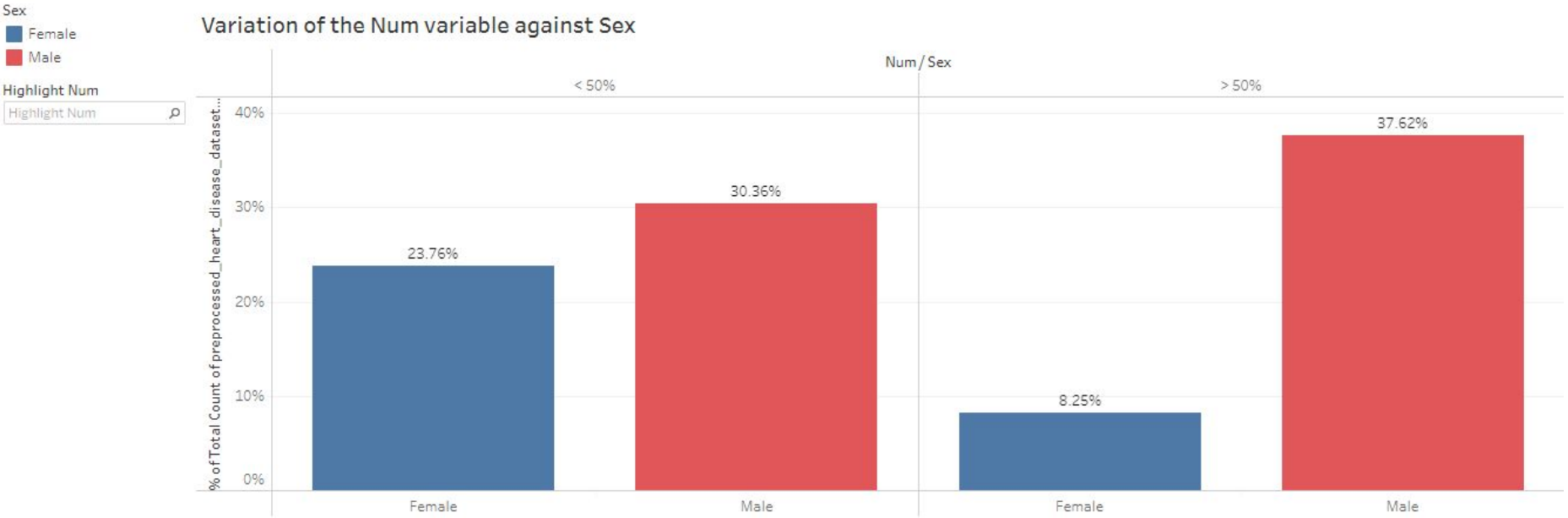
3) But for people with Flat or Downsloping Slope, they experience all 3 different types of Restecg categories.



Inferences:

1) The major contribution in the pie chart is by the 0 number of major vessels.

2) There is a trend in the chart - as the number of major vessels increases, the percentage of population contributing to the same decreases.



Inferences:

1) We can see that out of the total Female population, close to 24% have less than 50% diameter narrowing and hence less chances of angiographic disease and around 8% have more than 50% diameter narrowing.

2) But in Males, this trend is reversed. The Males have higher chances of suffering the angiographic disease owing to around 38% of Male population having more than 50% diameter narrowing.

THANK YOU