Bharatiya Vidya Bhavan's
Sardar Patel Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai)
Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India

Name:	**Gitanjali Gangurde**
UID:	**2021300034**
Batch:	**COMPS A (Batch G)**
Exp No. :	**3**

Aim:
Design Interactive Dashboards and Storytelling using Tableau / Power BI / R (Shiny) / Python (Streamlit/Flask) / D3.js to be performed on the dataset - Disease spread / Healthcare
Create interactive dashboard - Write observations from each chart given below

- Advanced - Word chart, Box and whisker plot, Violin plot, Regression plot (linear and nonlinear), 3D chart, Jitter, Line, Area, Waterfall, Donut, Treemap, Funnel
- Basic - Bar chart, Pie chart, Histogram, Timeline chart, Scatter plot, Bubble plot)

Dataset Description:

This is a multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis. It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia. This database includes 76 attributes, but all published studies relate to the use of a subset of 14 of them. The Cleveland database is the only one used by ML researchers to date. One of the major tasks on this dataset is to predict based on the given attributes of a patient whether that particular person has heart disease or not and another is the experimental task to diagnose and find out various insights from this dataset which could help in understanding the problem more.

Column Descriptions:

1. id (Unique id for each patient)
2. age (Age of the patient in years)
3. origin (place of study)
4. sex (Male/Female)
5. cp chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])
6. trestbps resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))

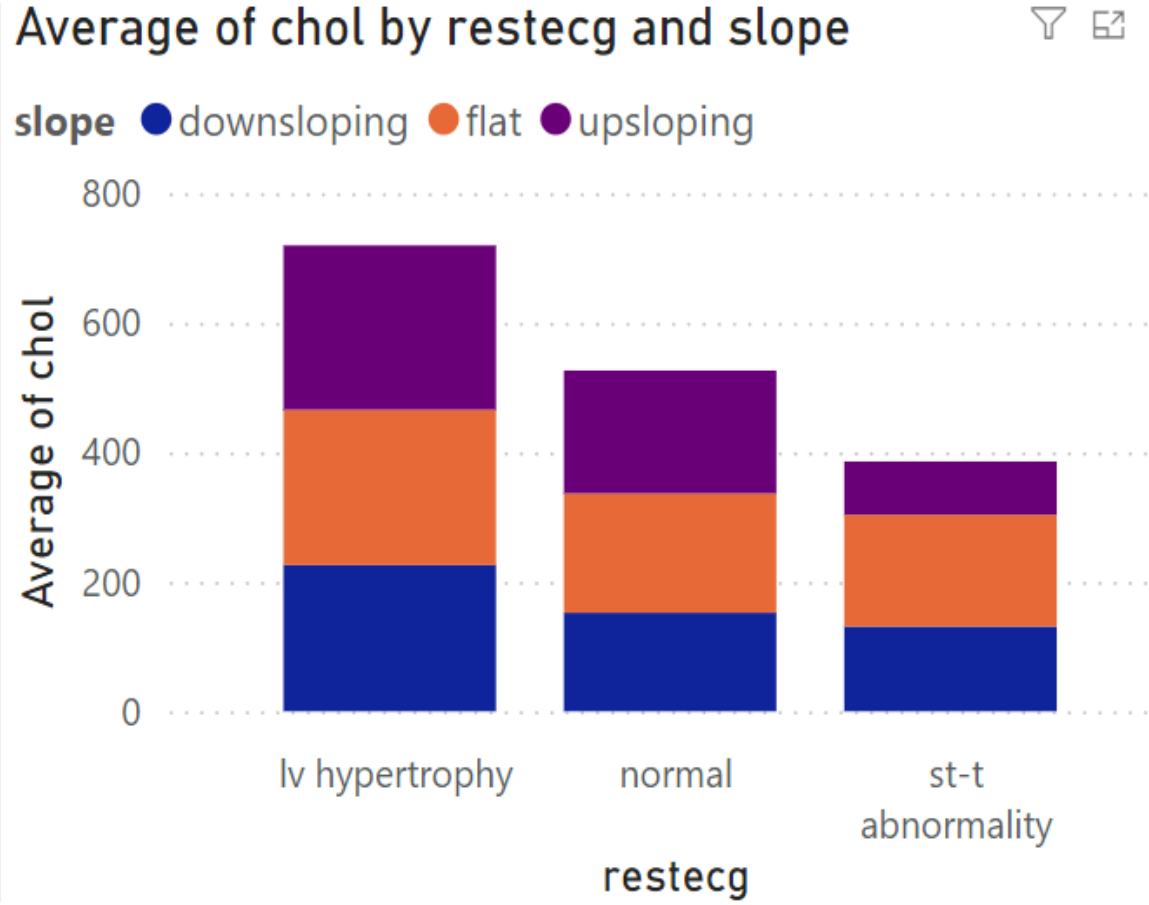
7. chol (serum cholesterol in mg/dl)
8. fbs (if fasting blood sugar > 120 mg/dl)
9. restecg (resting electrocardiographic results)
   -- Values: [normal, st abnormality, lv hypertrophy]
10. thalach: maximum heart rate achieved
11. exang: exercise-induced angina (True/ False)
12. oldpeak: ST depression induced by exercise relative to rest
13. slope: the slope of the peak exercise ST segment
14. ca: number of major vessels (0-3) colored by fluoroscopy
15. thal: [normal; fixed defect; reversible defect]
16. num: the predicted attribute

Dataset Link:

https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

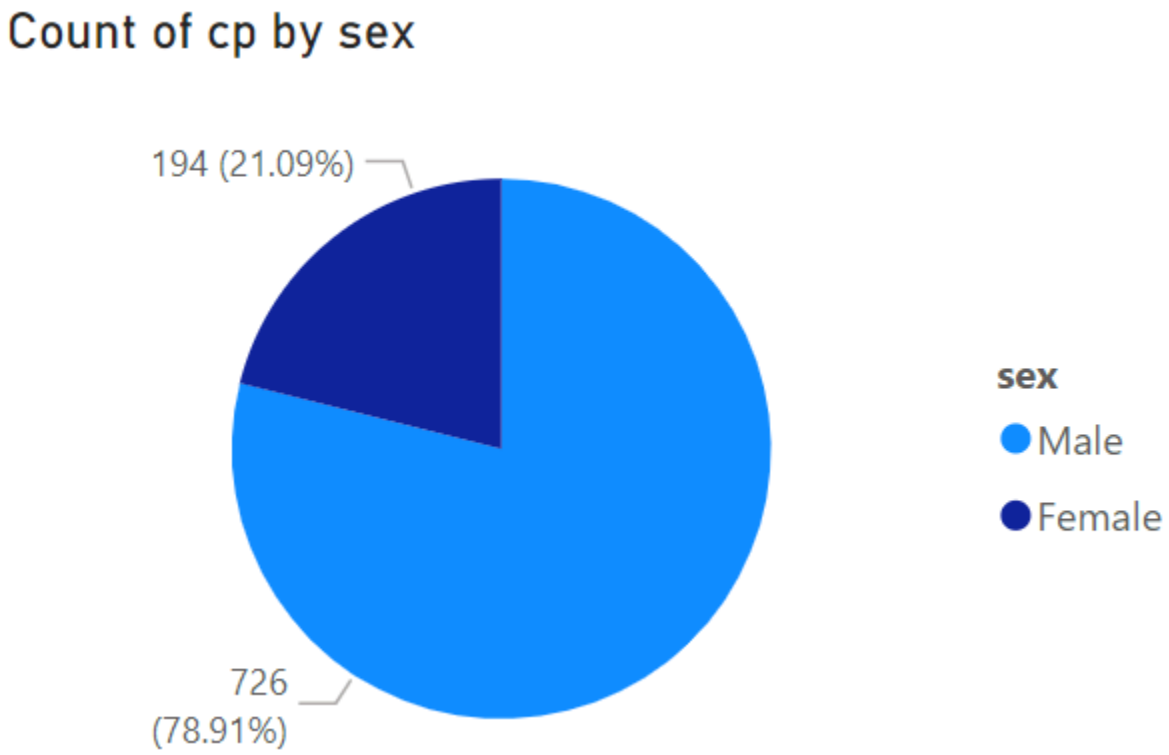
Plots and Inference:

1. Bar chart:



- lv hypertrophy has the highest average cholesterol levels, with a significant contribution from upsloping slopes (purple section) and a relatively smaller contribution from downsloping (blue) and flat slopes (orange).
- Normal restecg has a lower average cholesterol level than lv hypertrophy, with roughly equal contributions from all three slope types.
- st-t abnormality shows the lowest average cholesterol levels, with a larger share of cholesterol associated with downsloping slopes, followed by flat and a smaller contribution from upsloping slopes.

2. Pie chart:



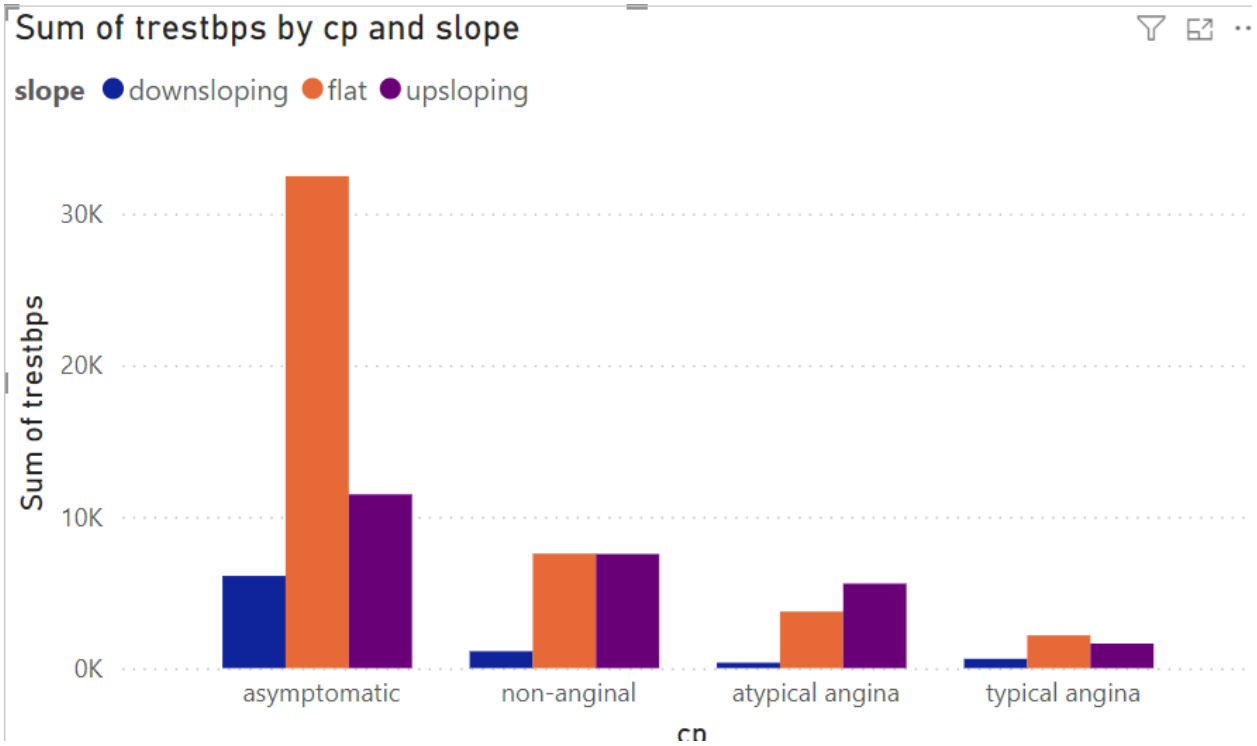
- 78.91% (726 cases) of chest pain cases are reported by males (blue section).
- 21.09% (194 cases) are reported by females (dark blue section).

This indicates that males experience a higher count of chest pain compared to females in this dataset

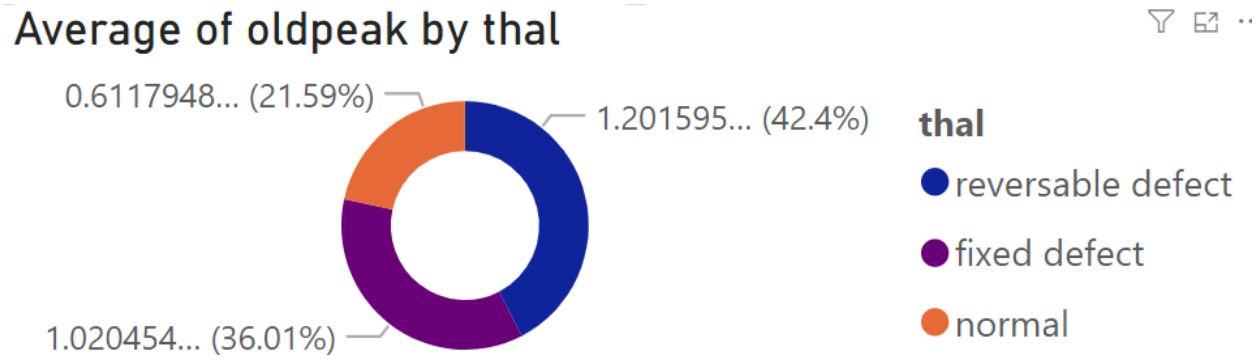
3. Histogram:

Asymptomatic chest pain has the highest total resting blood pressure, with a significant portion coming from patients with flat slopes (orange). The contributions from upsloping and downsloping are also substantial, but the flat slope dominates.

Non-anginal chest pain shows a lower overall blood pressure sum compared to asymptomatic cases, with flat and upsloping slopes contributing almost equally, while downsloping is minimal.



4. Donut Chart:

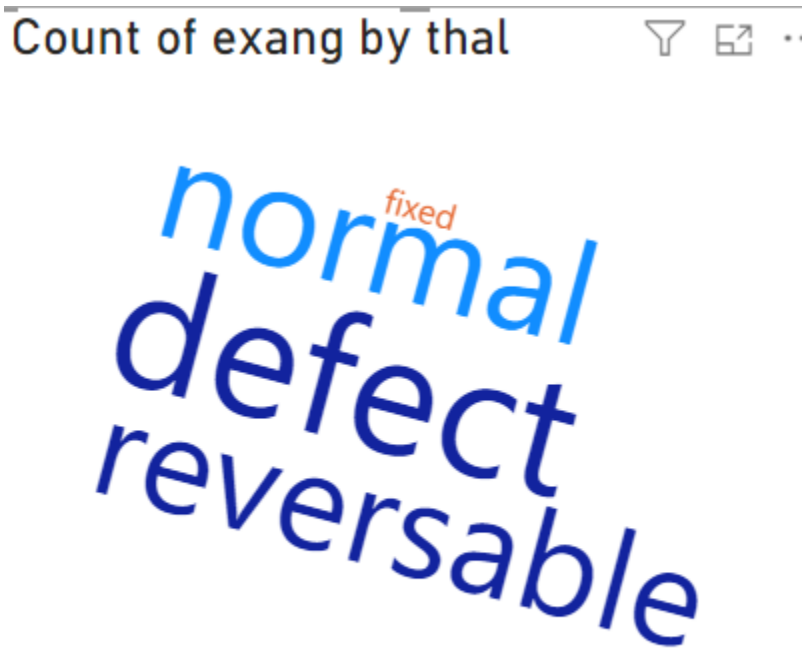


The highest average oldpeak is seen in the "Reversible defect" category, indicating that individuals in this group tend to have more significant ST depression.

The "Fixed defect" group has the second-highest average oldpeak, suggesting moderate ST depression.

The "Normal" category has the lowest average oldpeak, which is expected as these individuals likely exhibit less cardiac stress or abnormality.

5. Word Chart:



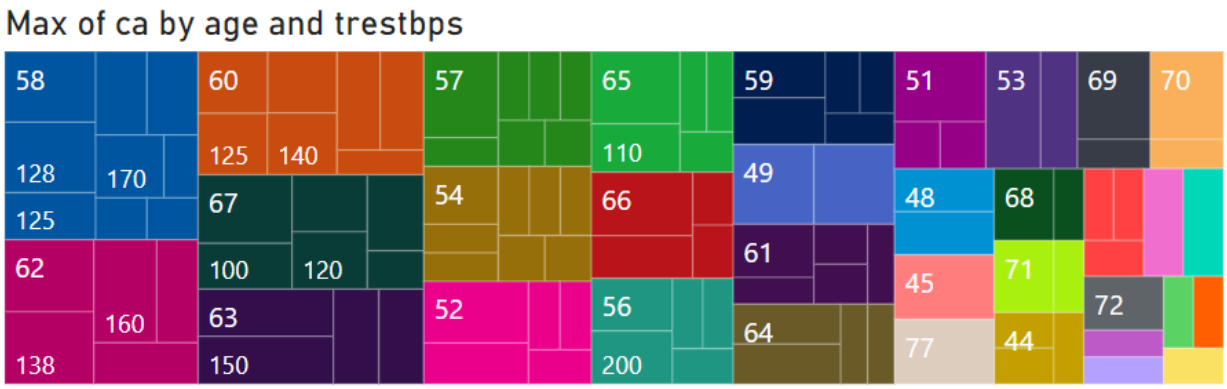
Normal appears as the largest, indicating that the majority of individuals in the dataset have the "thal" category marked as "normal."

Reversible defect and defect (likely related to the "fixed defect") are also prominent, suggesting that these categories have a significant number of cases but are less frequent than "normal."

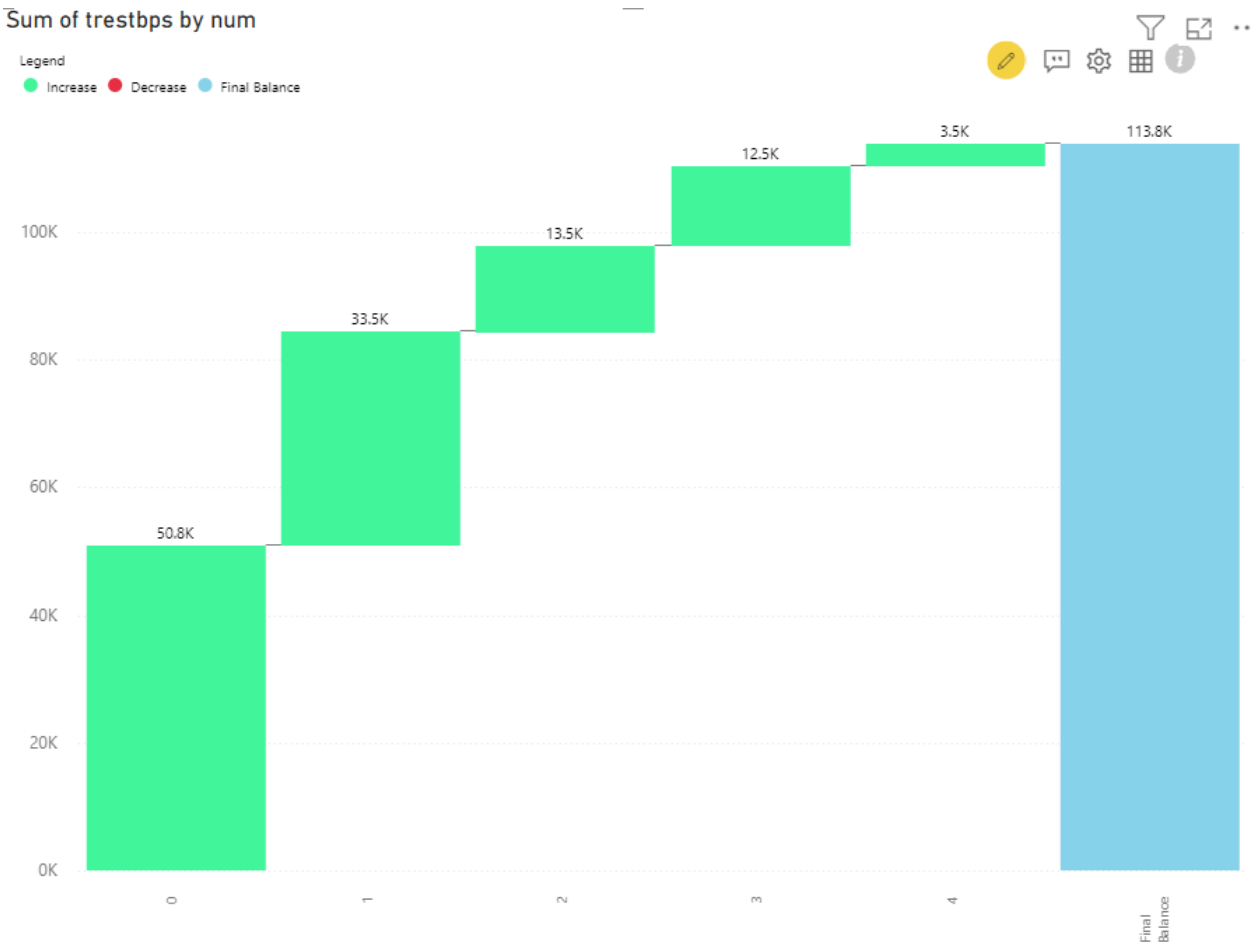
Fixed defect appears small, meaning that it has the fewest counts in the dataset compared to other thalassemia types.

6. Tree map:

Large Boxes: Age-resting blood pressure pairs such as 58-128, 60-125, and 67-100 show larger rectangles, which indicates these combinations likely have higher "ca" values. This may suggest these individuals have more prominent coronary artery disease risk.

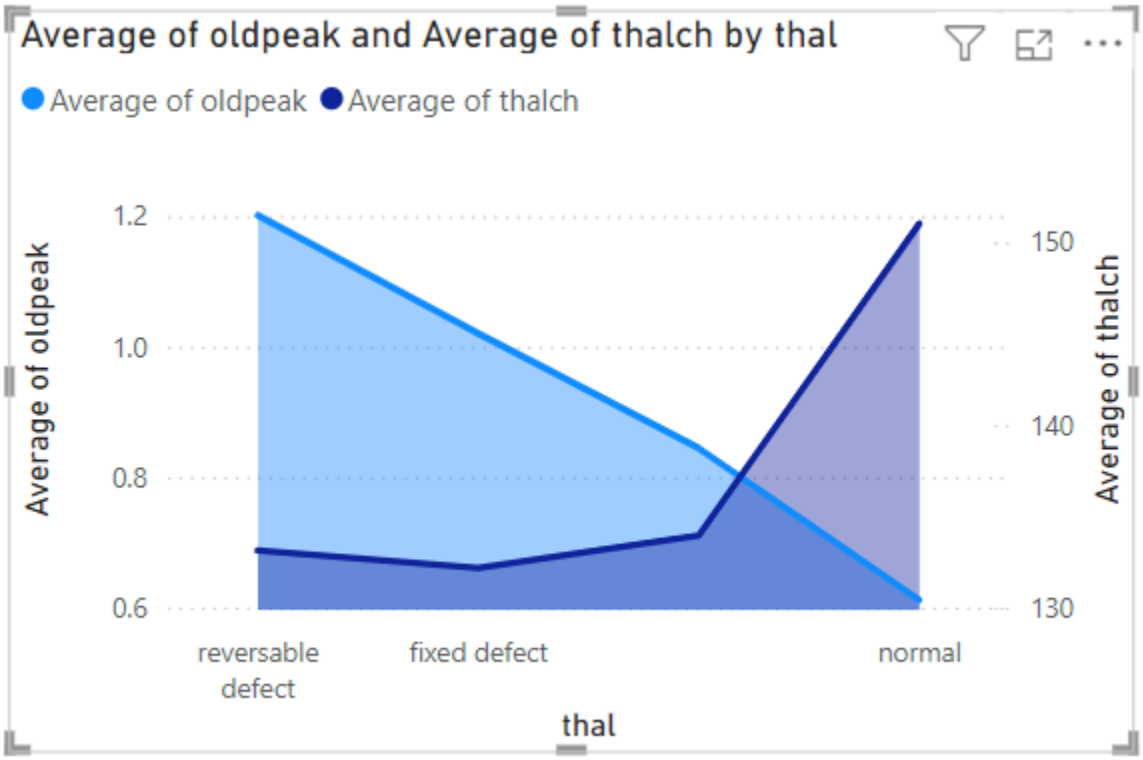


7. Waterfall chart:



The trend suggests a potential positive correlation between the predicted attribute (num) and resting blood pressure (trestbps). This could indicate that as the predicted attribute (possibly risk or severity) increases, the resting blood pressure tends to increase as well, which could be critical for risk assessment, patient management, and treatment planning in a medical context.

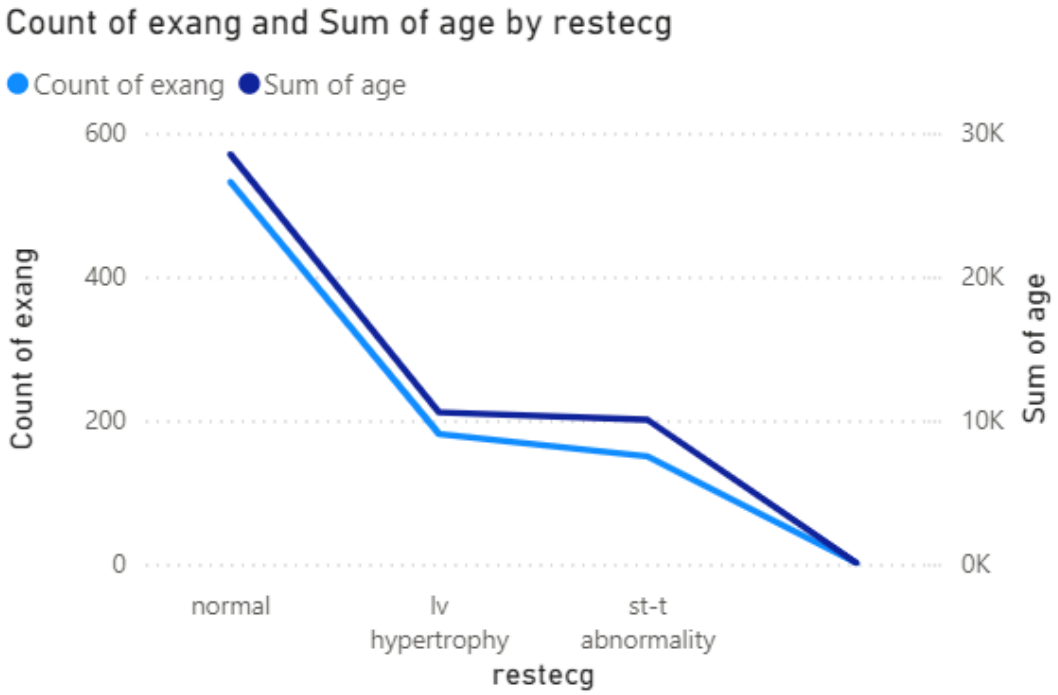
8. Area chart:



There is a clear inverse relationship between oldpeak (ST depression) and thalach (maximum heart rate achieved). As oldpeak decreases from "reversible defect" to "normal," thalach increases. This suggests that patients with better heart health (indicated by a "normal" condition) have higher maximum heart rates and lower ST depression.

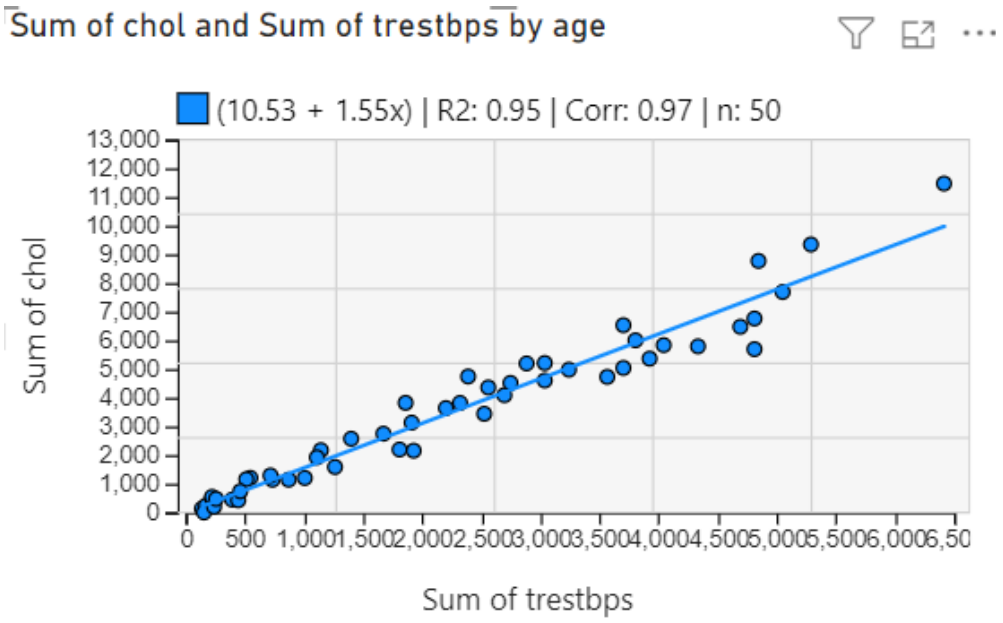
Patients categorized as "normal" show the lowest oldpeak (~0.6) and the highest thalach (~150). This indicates an optimal cardiac response to stress, suggesting no significant heart defects or ischemia. It serves as a benchmark for healthy heart performance during stress tests.

9. Line chart:



The chart provides insights into the distribution of exercise-induced angina cases and the total age of patients across different resting ECG results. It shows that patients with a "normal" ECG result have a higher prevalence of exercise-induced angina and are generally older compared to those with "lv hypertrophy" or "st-t abnormality".

10. Regression chart:

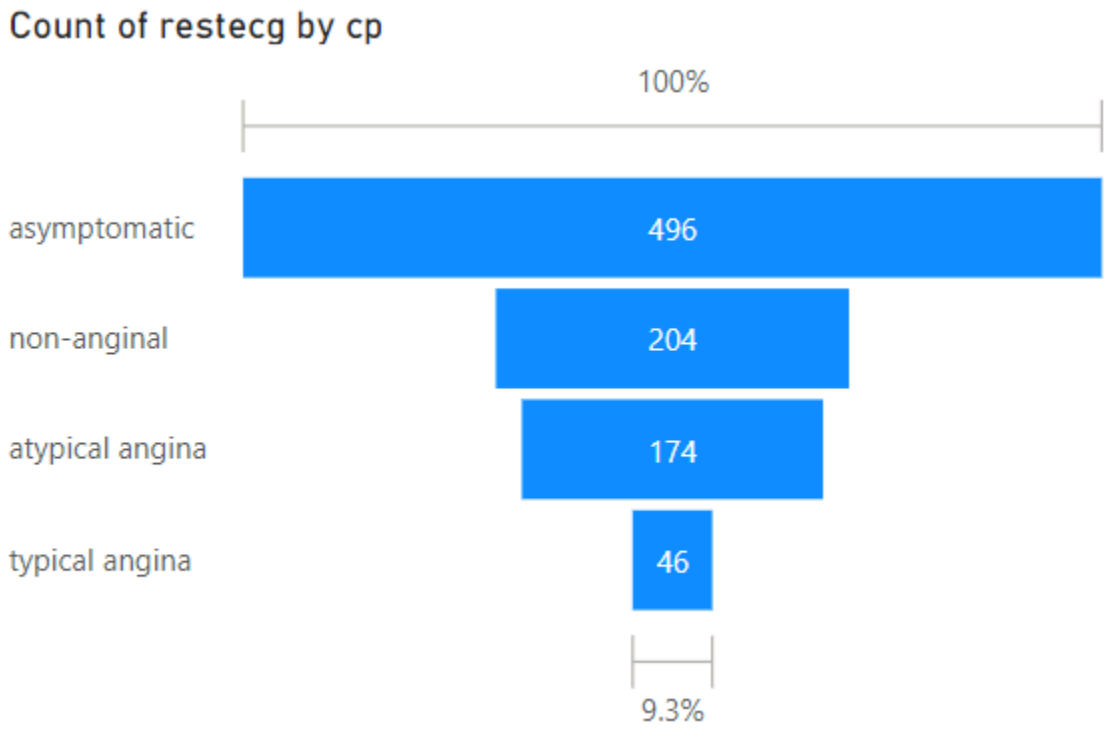


The plot shows a strong positive linear relationship between the Sum of chol (cholesterol) and Sum of trestbps (resting blood pressure), as indicated by the correlation coefficient (Corr =

0.97). This suggests that as resting blood pressure increases, cholesterol levels also tend to increase.

The scatter plot suggests a very strong linear relationship between cholesterol levels and resting blood pressure across age groups, with higher blood pressure being associated with higher cholesterol. This could indicate that individuals with higher resting blood pressure might be at an increased risk of elevated cholesterol, which could be an important consideration for cardiovascular risk assessments and management.

11. Funnel chart:

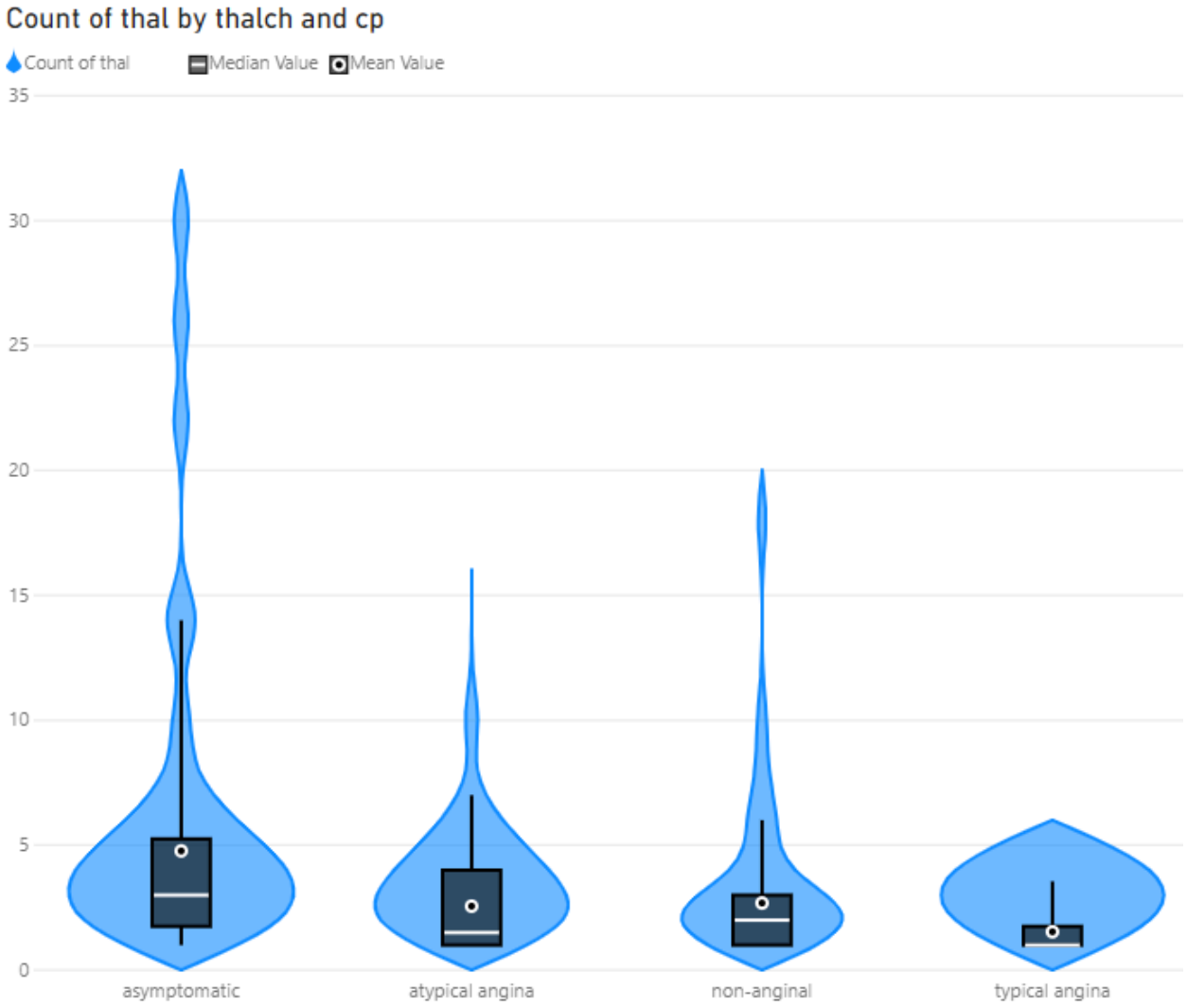


Asymptomatic cases are by far the most common, suggesting that many individuals may not exhibit typical symptoms of angina or chest pain, despite abnormal ECG results.

Non-anginal pain is the second most frequent, indicating that a significant number of individuals experience chest pain that is not related to angina.

Typical angina, often considered a classic symptom of heart disease, is relatively rare compared to other types.

12. Violin chart:



1. Violin Shape: The shape of the violin represents the density of the data at different values. Wider sections indicate a higher frequency of observations. For example, the "asymptomatic" category has the widest violin, suggesting that more data points exist for this chest pain type compared to the others.
2. Box Plot Inside: Each violin contains a box plot. The black rectangle shows the interquartile range (IQR), which contains the middle 50% of the data. The white circle represents the median value, and the black bar represents the range of data within 1.5 times the IQR from the quartiles.
3. Mean Values: The chart also includes a marker for the mean value (marked with a diamond or similar shape).

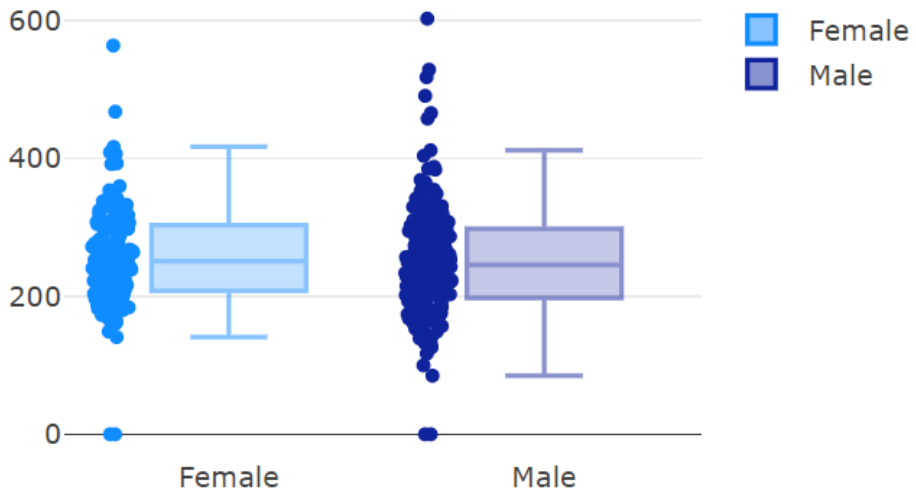
Most "thal" values are associated with "asymptomatic" cases of chest pain, which has the largest distribution.

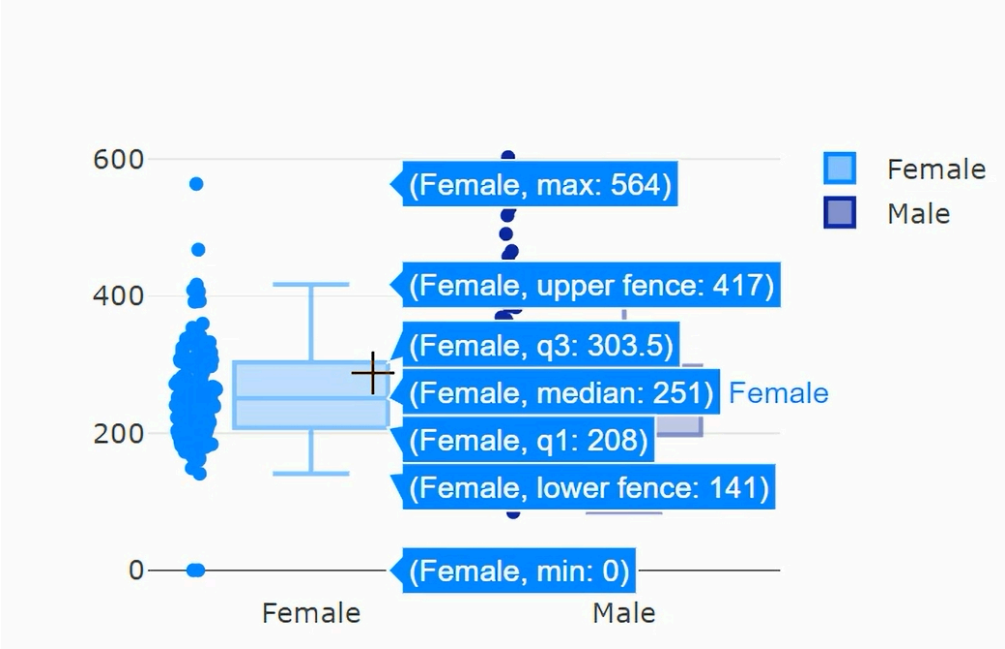
Other types of chest pain have fewer "thal" values.

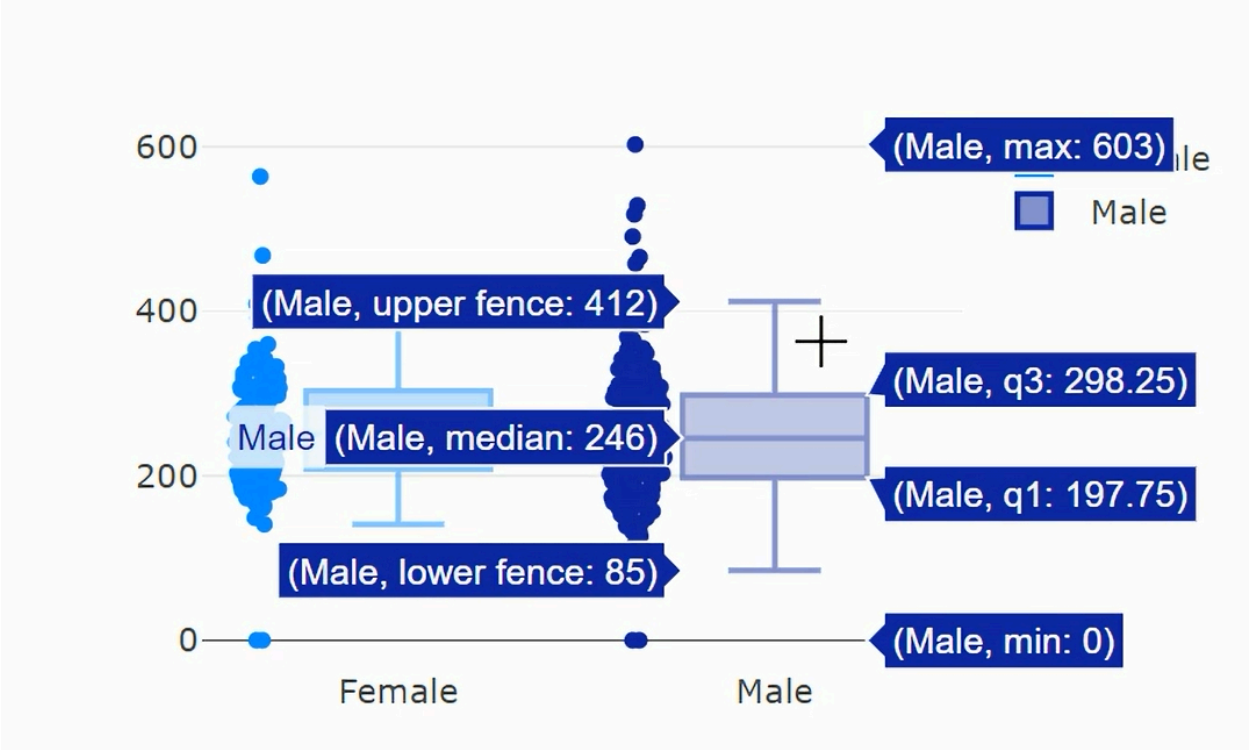
The median and mean values, as well as the overall distribution, vary between the different categories of chest pain.

13. Box and whisker plot:

sex and chol



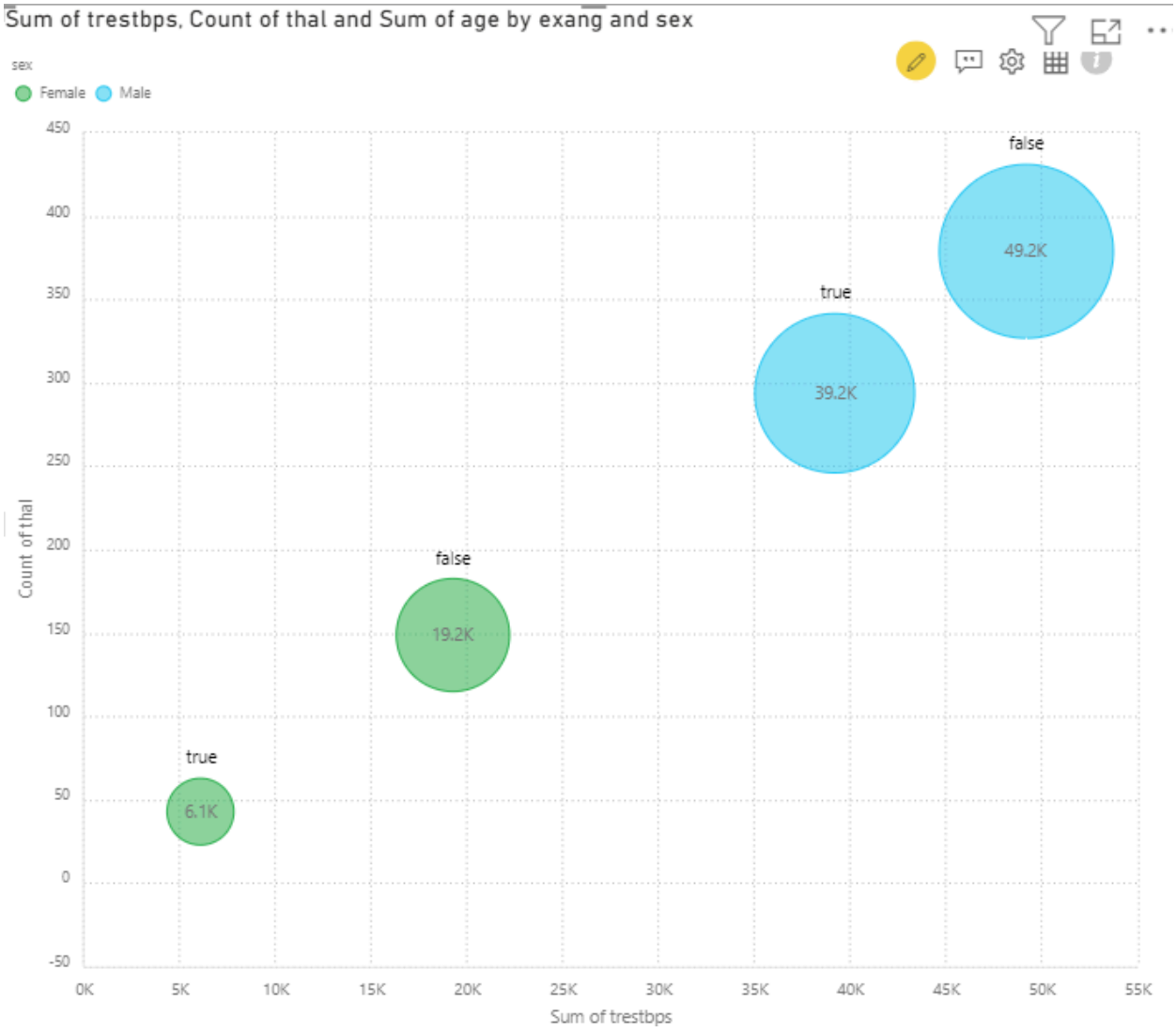




Both distributions are fairly symmetric, with whiskers extending evenly on both sides of the box. However, the male group has a denser clustering of data points around the median.

Females tend to have more variability in cholesterol levels compared to males, and their median cholesterol level seems to be slightly higher. Both genders show some extreme outliers with very high cholesterol values.

14. Bubble plot:



The largest bubble (male group, no exercise-induced angina) has a high sum of age (49.2k), a high count of thal (around 400), and a large sum of trestbps (around 50k).

The smallest bubble (female group, with exercise-induced angina) has the lowest sum of age (6.1k) and is also positioned at the lower end of the y-axis and x-axis, with a lower count of thal and lower sum of systolic blood pressure.

Overall, males tend to have higher values across all three metrics compared to females.

This chart suggests that males generally have higher systolic blood pressure, a greater count of thal results, and a larger cumulative age than females. Additionally, the presence of exercise-induced angina varies between the groups, affecting these metrics differently.