Analysis report on factors impacting heart disease

Analysis of Heart Disease Dataset

Problem Statement

The objective of this dataset is to

1. Develop predictive models for heart disease diagnosis or risk assessment.
2. Explore the relationships between various medical attributes and heart disease.
3. Identifying the most significant risk factors or features contributing to heart disease.
4. Building clinical decision support systems or risk calculators for heart disease.
5. Check if specialised treatment plans are required, such as if gender specific treatments are required
6. Observe the age groups and their impact on this disease
7. Create awareness and educate staff and patients about the risk factors contributing to the heart disease

The problem statement and objective involve leveraging the provided medical data to build accurate and reliable models for heart disease prediction or diagnosis, ultimately aiming to improve patient care and early detection of heart-related issues.

DATA CLEANING AND TRANSFORMATION

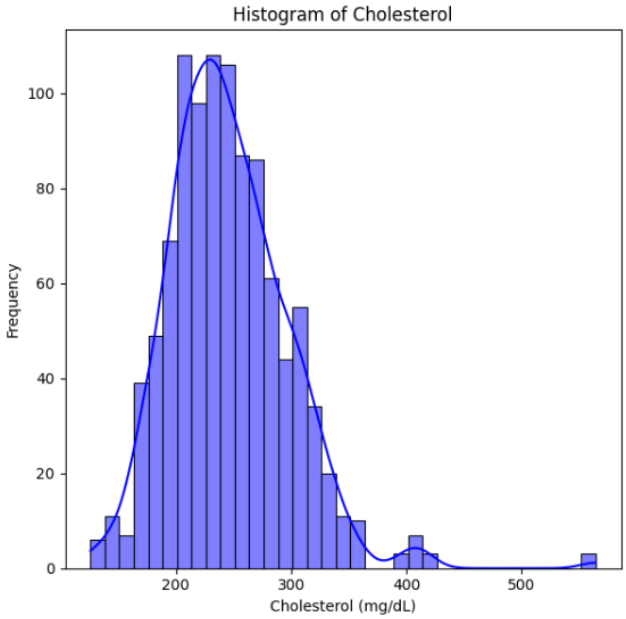
- The target column is balanced and the dataset has no missing values.
- **Descriptive analysis:**
- 1. Age: Ranges from 29 to 77 years. The distribution is fairly normal as indicated by the mean (54.43) being close to the median (56).
- 2. The average age in the dataset is approximately 54 years.
- 3. Chest Pain (cp): Ranges from 0 to 3. The mean is closer to 1, suggesting that most patients have a low level of chest pain.
- 4. The majority of the patients are male (about 70%).
- 5. Resting Blood Pressure (trestbps): Ranges from 94 to 200 mm Hg. The standard deviation is relatively small (17.52), indicating that most values are close to the mean (131.61).
- 6. Cholesterol (chol): Ranges from 126 to 564 mg/dl. The maximum value (564 mg/dl) is quite high and could be considered an outlier or a potential error.
- The average cholesterol level is 246 mg/dl.
- The average maximum heart rate achieved is 149 bpm.
- 7. Fasting Blood Sugar (fbs): Binary variable (0 for < 120 mg/dl, 1 for > 120 mg/dl). Only about 15% of patients have a fasting blood sugar above 120 mg/dl.
- 8. Resting ECG (restecg): Ranges from 0 to 2. Most patients have a value of 0 or 1, indicating normal or having ST-T wave abnormality.

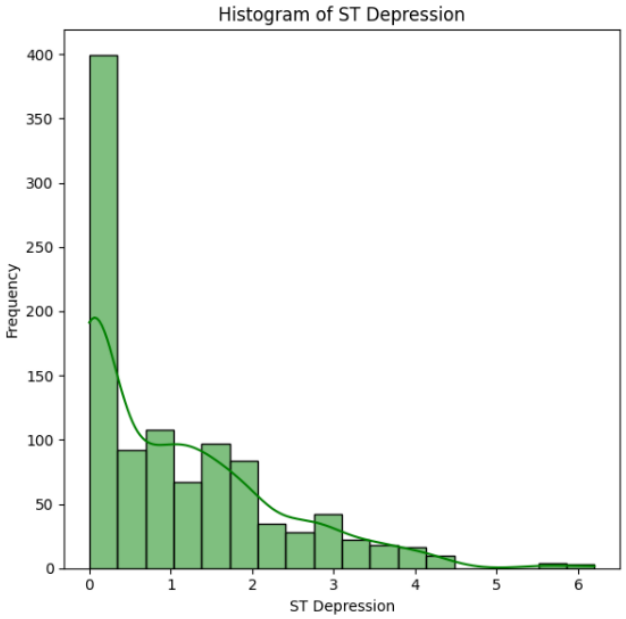
- 9. Maximum Heart Rate (thalach): Ranges from 71 to 202 bpm. The distribution is slightly left-skewed as the mean (149.11) is less than the median (152).
- 10. Exercise Induced Angina (exang): Binary variable (0 for no, 1 for yes). About 34% of patients experience angina due to exercise.
- 11. ST Depression (oldpeak): Ranges from 0 to 6.2. The maximum value (6.2) is significantly higher than the 75th percentile (1.8), suggesting potential outliers.The standard deviation for 'ST Depression (oldpeak)' is relatively high compared to its range (0 to 6.2), suggesting a wide variation in this measurement across patients.
- 12. Slope of the Peak Exercise ST Segment (slope): Ranges from 0 to 2. Most patients have a slope of 1 or 2.
- 13. Number of Major Vessels (ca): Ranges from 0 to 4. The mean (0.75) suggests that most patients have fewer than one visible vessel on fluoroscopy.
- 14. Thalassemia (thal): Ranges from 0 to 3. The distribution is skewed towards higher values, with most patients having a value of 2 or 3.

DATA VISUALIZATION

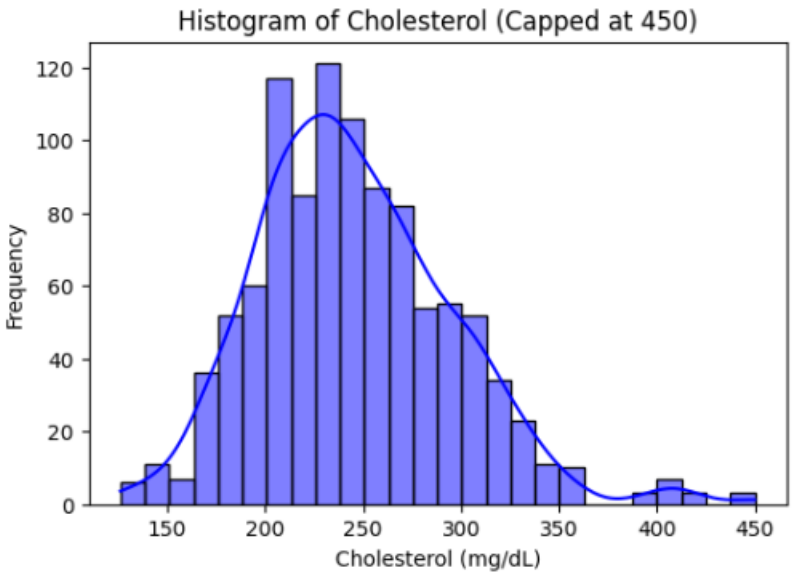
- **Outlier Treatment**

- Cholesterol (chol) and ST Depression (oldpeak) show notable outliers, as indicated by points that lie far outside the upper whiskers. Other variables like Resting Blood Pressure (trestbps) and Maximum Heart Rate (thalach) also display some outliers, but they are closer to the main distribution.

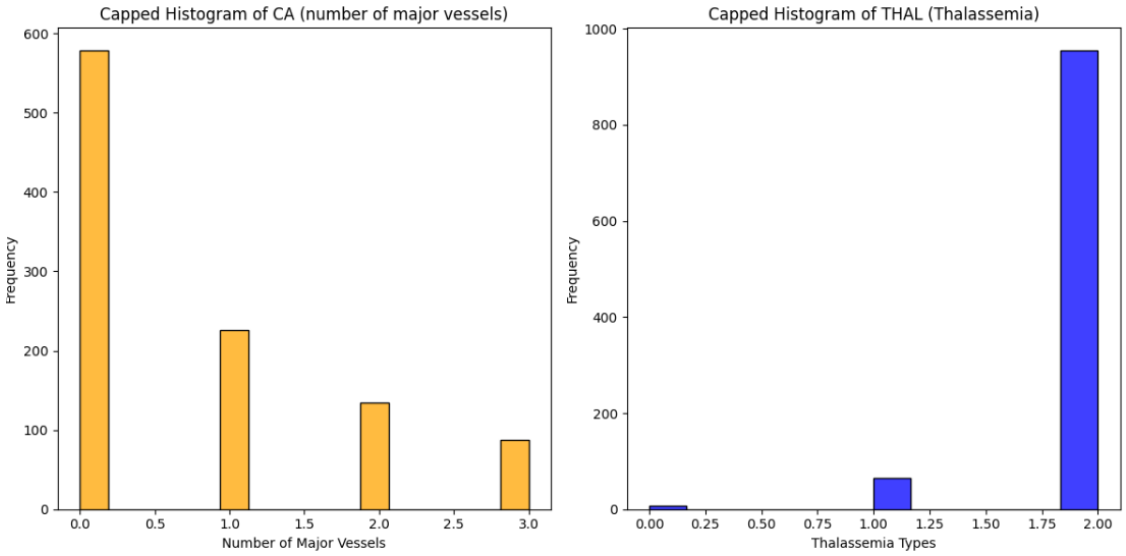




- Capped the values at 450 for chol(cholesterol) column. Note: A human can have 450 mg cholesterol, but is requires urgent attention from doctors, For analysis and to include these cases, the value is capped at 450 mg.

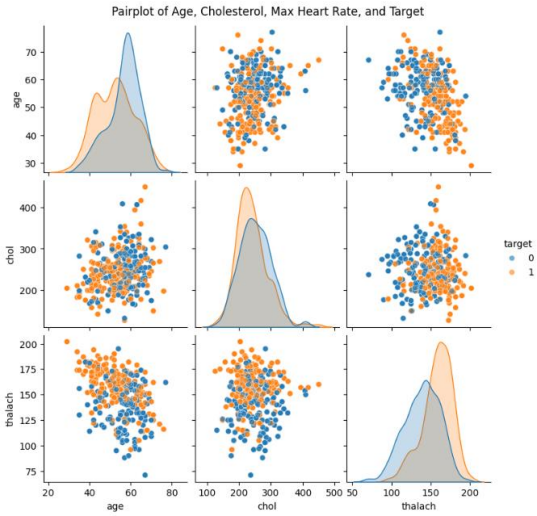


- capped the value number of major vessels (0-3) - 'ca' colored by fluoroscopy - as mentioned in the data dictionary. There is no value as 4, which is mentioned in the dataset
- Thal - 0 = normal; 1 = fixed defect; 2 = reversable defect - this is mentioned in the dictionary, hence the value is capped at 2 even though there was lot of outliers with the value 3. This could be an error or a category not mentioned in the data dictionary.
- Lastly, the value above 180 mm can be an outlier as normal humans do not have anything above 180 mm. Any value above 180 mm is a critical and emergency case. However, for the purpose pf examining the impact of heart attack on such emergency cases as well, I have not capped the value.Used box plot and describe() method to find outliers. Removed outliers for Price, minimum nights and reviews per month column. The capped method with 99 percentile was used to cap the outliers and work with the data



- Age and thalach seem to have a inverse relation old peak and thal have a positive relationship. old peak seems to increase as thal increases.

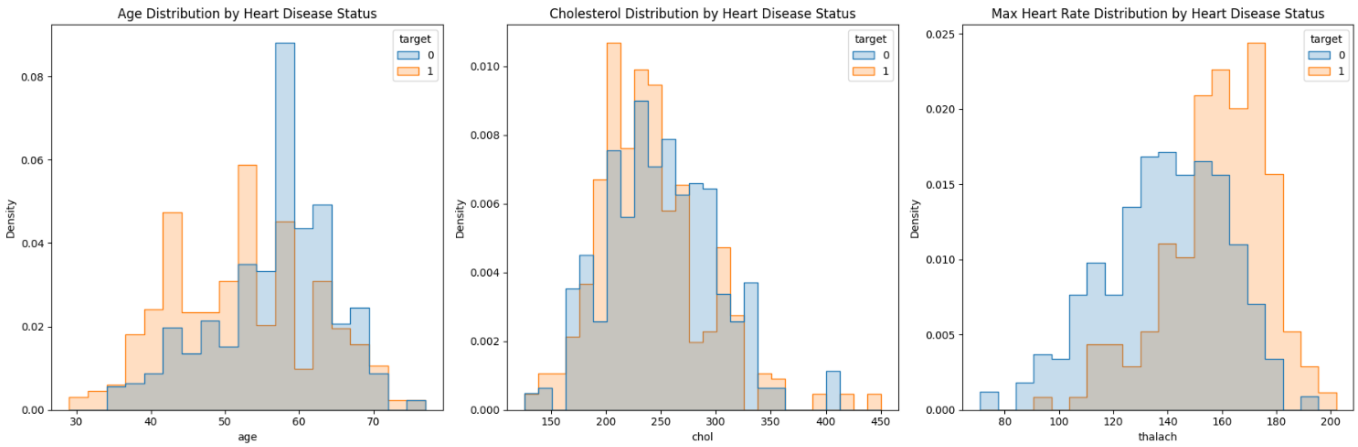
- **Observations from the Pairplot:**
- 1. Age vs. Cholesterol: There appears to be a slight positive trend, suggesting that cholesterol levels might increase with age.
- 2. Age vs. Maximum Heart Rate (thalach): A negative trend is observed, indicating that maximum heart rate tends to decrease as age increases.
- 3. Cholesterol vs. Maximum Heart Rate (thalach): No clear trend is visible, suggesting little to no direct relationship between cholesterol levels and maximum heart rate.
- 4. Target (Presence of Heart Disease): The plots segmented by 'target' show different distributions for patients with and without heart disease, particularly noticeable in the scatter plots involving 'thalach', where individuals with heart disease tend to have lower maximum heart rates.



Distribution plots for 'age', 'chol', and 'thalach' based on the 'target' variable. The plots show that :

- Individuals with heart disease (target = 1) tend to be peak around 50-55 age group. Those without heart disease (target = 0) tend to peak at 55-60 age group

- Cholesterol (Chol): The distribution of cholesterol levels for individuals with heart disease appears slightly higher on average compared to those without heart disease. The peak for target=1, being 200-210 and peal for target =0 being 230-240
- Maximum Heart Rate (Thalach):Individuals with heart disease tend to have a higher maximum heart rate, with a peak around 170-180 beats per minute. The individuals with target = 0, show a blunt peak between 140-150, although those without heart disease generally exhibit lower maximum heart rates, with a more spread out distribution.



- The Average Age by Heart Disease : Those with target=1 (who have heart disease) have an average age of 52 years and those who have none appear to be in age group of 56-57 years

- The counts for sex and target

Sex	**Target & target counts**
0	1 (226)
	0 (86)
1	1 (413)
	0 (300)

- Counts for cp (chest pain) and target

cp	**Target & target counts**

0	0 (375)
	1 (122)
1	0 (33)
	1 (134)
2	0 (65)
	1 (219)
3	0 (26)
	1 (51)

- Average Resting Blood Pressure (trestbps) by Heart Disease Status:
  0 - 134.106212
  1 - 129.245247

- Average Cholesterol Levels by Heart Disease Status:
  0 - 251.292585
  1 - 240.328897

- Exercise-Induced Angina and Heart Disease Risk:

Angina(exang)	**Target & target counts**
0	1 (455)
	0 (125)
1	1 (71)
	0 (274)

- When one compares age and sex, following are the insights:

  _____

  Mean of the dataset: sex
  0 55.849359
  1 53.814867

  _____

  Median of the dataset: sex
  0 57.0
  1 55.0

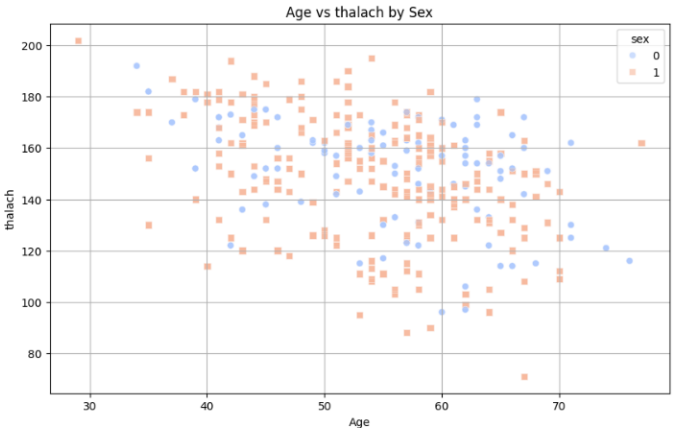
  _____

  Mode of the dataset: sex
  0 62

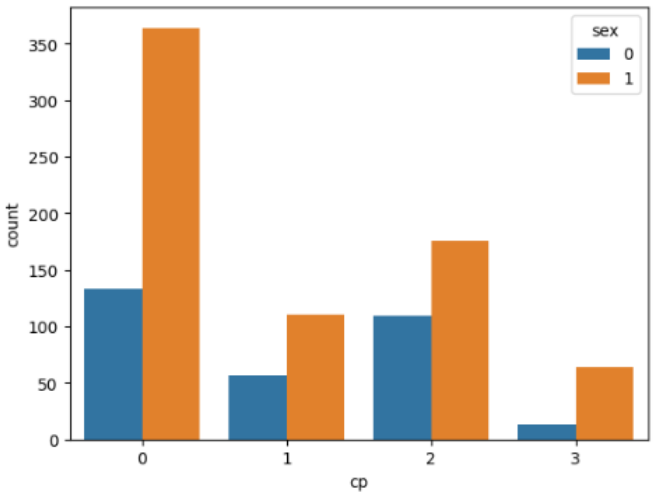
1 58

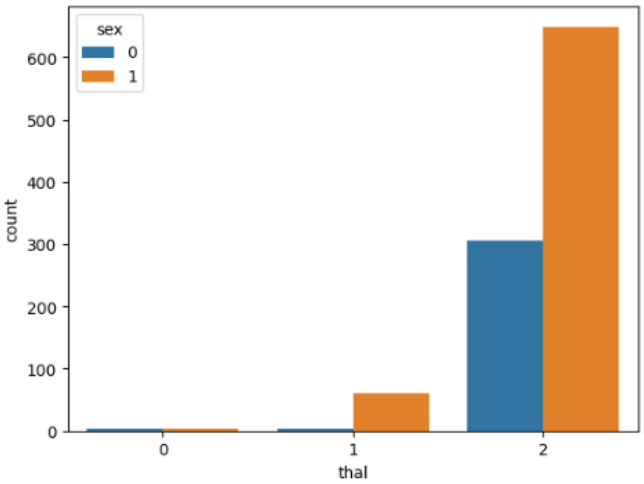
Note: It was also found that the distribution peak was same for both the sexes, which was at 58-59 years age group. The lowest peak for sex=0, was 74-77 years while sex=1 was for 28-29 age group.

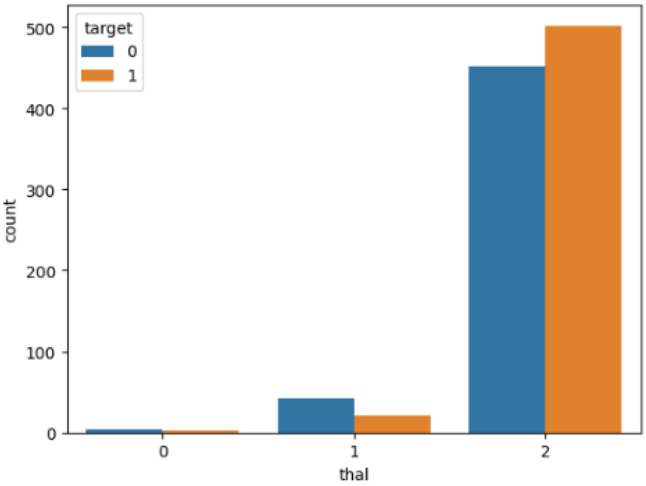
- Age vs trestbps by Sex: A slightly positive or upward trend is observed with her relationship between Age and trestbps, the majority of sex being sex=1. It is most dense in 50-69 age group
- Age vs thalach by sex: A slightly negative or downward trend is observed with her relationship between Age and thalach, the majority of sex being sex=1. It is most dense in 50-69 age group



- **count plot of cp column by sex column:** cp is highest for category 0 for both the sexes and lowest for category 3 for both the sexes.
  count plot of thal column by sex column: thal is highest for category 2 for both the sexes and lowest for category 0 for both the sexes.
  count plot of thal column by target column: thal is highest for category 2 for both the target values and lowest for category 0 for both the target values.







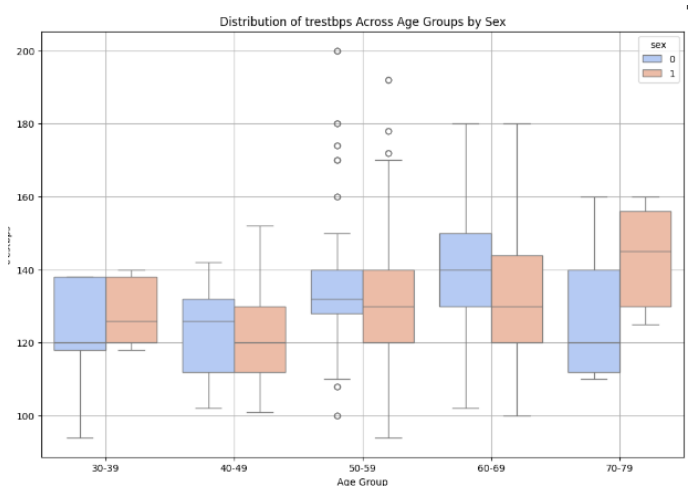
- **cp, age and sex:** The values are generally uniform for the younger age groups (40-69) for both sexes and lower for the older age groups (70-79) for both males and females. The consistent gap between males (blue) and females (orange), with sex=0 having higher values across all age groups, suggests that men tend to experience more chest pain or have a higher severity of chest pain compared to women of the same age group
- **Resting Blood Pressure (trestbps), age and sex**: Blood pressure tends to increase with age for sex=1, with the highest levels observed in the 70-79 age group. The spread of values (represented by the box heights) appears to be quite wide, indicating substantial variability in resting blood pressure within each age and sex group.
- **Cholesterol (chol):** Chol tends to increase with age for sex=1, with the highest levels observed in the 70-79 age group. Sex=0 is at its peak in the 60-69 age group. The spread of values is generally wider in the older age groups, indicating greater variability in cholesterol levels among older individuals.
- **(Thalach - Maximum Heart Rate Achieved, age, sex)** : For both sexes, the maximum heart rate achieved tends to decrease with age, which is expected as cardiovascular fitness and maximum heart rate typically decline with aging.

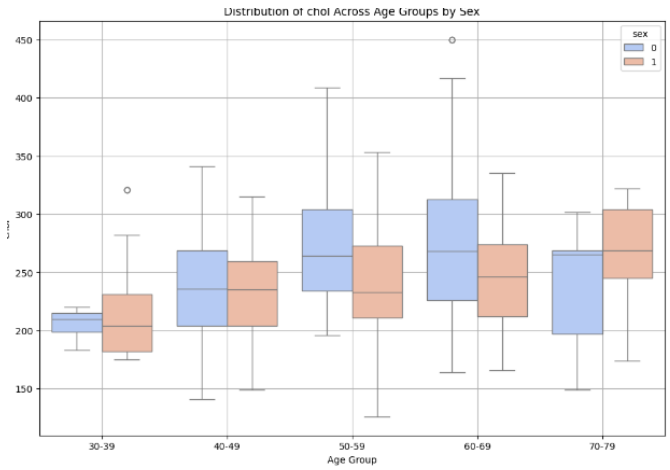
  Across most age groups, males (blue boxes) tend to have higher maximum heart rate values compared to females (orange boxes), potentially reflecting sex differences in cardiovascular fitness or physiology. There are several outliers, particularly in the older age groups, indicating individuals who achieved exceptionally high or low maximum heart rates for their age and sex.

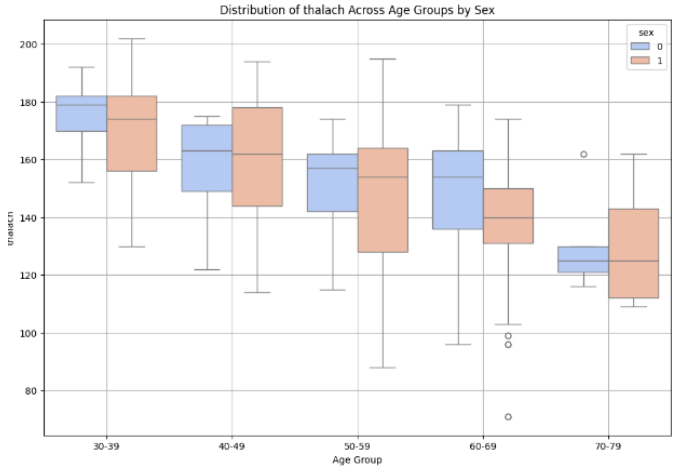
- **restecg, age, sex:** Overall, sex=1 generally exhibit slightly restecg across all age groups compared to sex=0.
- **Distribution of ca across age groups by sex:** For the 30-39 and 40-49 age groups, both males and females have extremely low ca values, suggesting a lower prevalence of major vessels coloured by fluoroscopy (an indicator of heart disease) in these younger age groups.The highest ca values are observed in males aged 70-79, indicating a higher risk of heart disease in this age group for sex=1

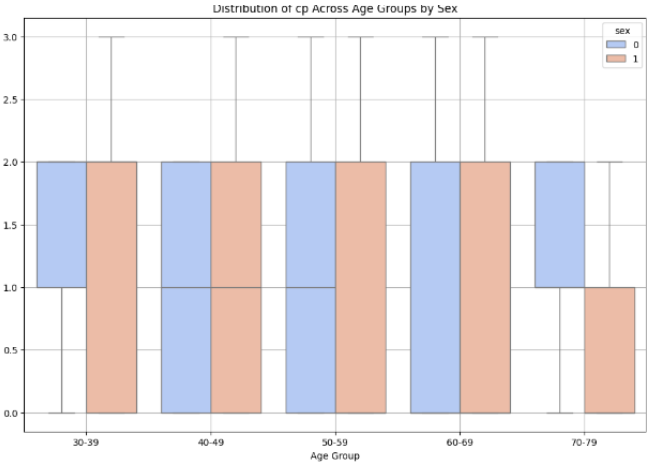
- **Distribution of slope across age groups by sex:** For all age groups, sex=1 tend to have a higher slope value compared to sex=0, indicating a steeper slope of the peak exercise ST segment in sex=1
- **Distribution of oldpeak across age groups by sex:** For both sex=1, the oldpeak value tends to increase with age, indicating a higher ST depression during exercise as people get older.

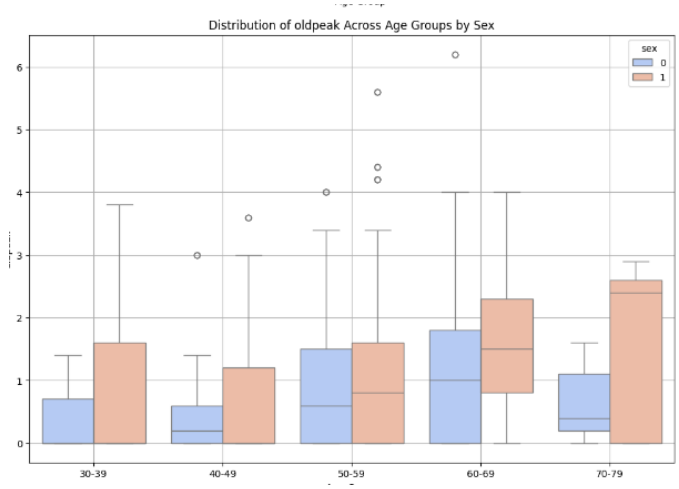
- For the age groups 30-39, 40-49, and 50-59, sex=1 generally have a upward trend. The highest oldpeak values are observed in the 70-79 age group for sex=1 and for sex=0, the highest peak is in the age group of 60-69.

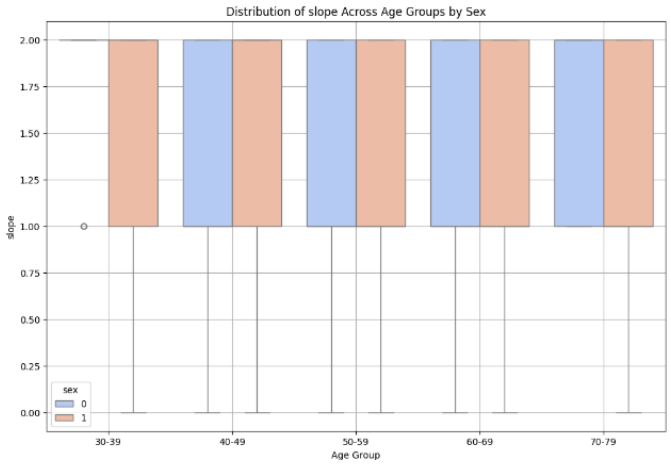


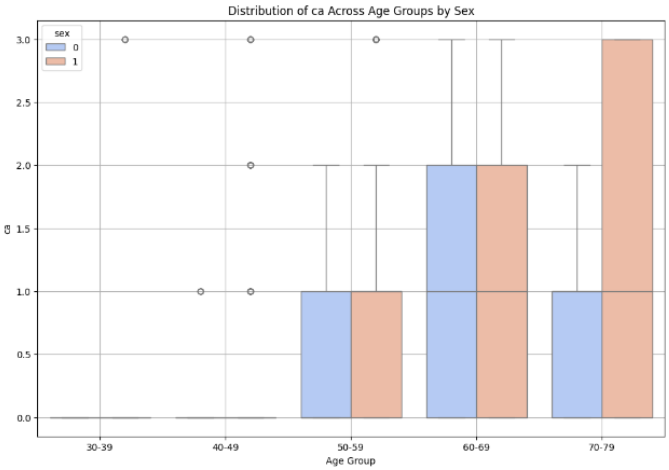










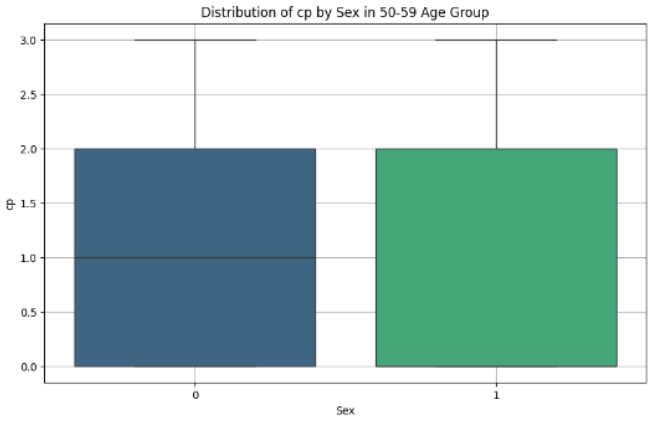


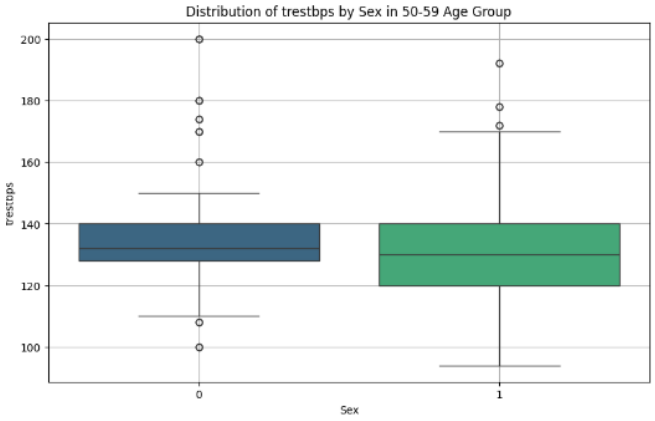
As the average age group for heart disease target = 0 is 56.569138 and target = 1 - 52.408745. We can conclude that, on ana average, the majority if heart disease occurs between the age group of 50-59 years. Hence decided to analyze this age group in detail

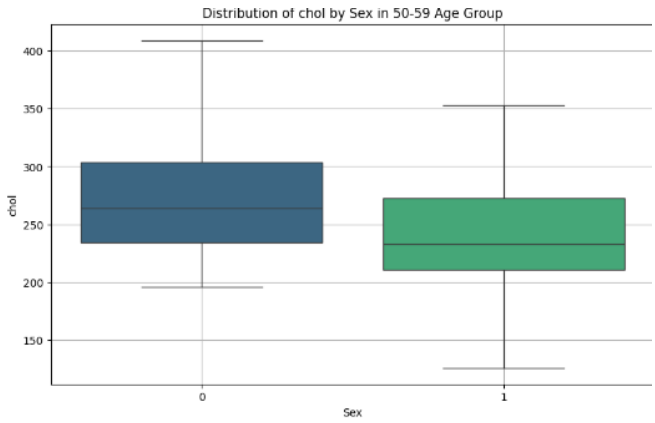
['CP', 'TRESTBPS', 'CHOL', 'FBS', 'RESTECG', 'THALACH', 'EXANG', 'OLDPEAK', 'SLOPE', 'CA', 'THAL', 'TARGET'] FOR THE 50-59 AGE GROUP

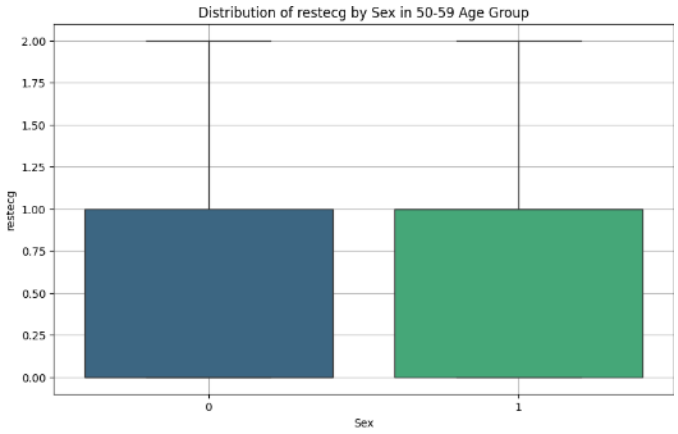
- **CHOLESTEROL (CHOL): Sex=0 (median 260) in this age group tend to have higher cholesterol levels compared to sex=1 (median 230) .**
- Distribution of **TRESTBPS** by Sex in the 50-59 Age Group: Sex=0, the median resting blood pressure (trestbps) is approximately 140 mmHg, as indicated by the middle line of the box. For sex=1 (represented by 1 on the x-axis), the median resting blood pressure is slightly lower, around 130 mmHg. Thus suggesting greater variability in resting blood pressure among sex=0 in this age group. Blood pressure distribution, indicating some individuals with elevated resting blood pressure.

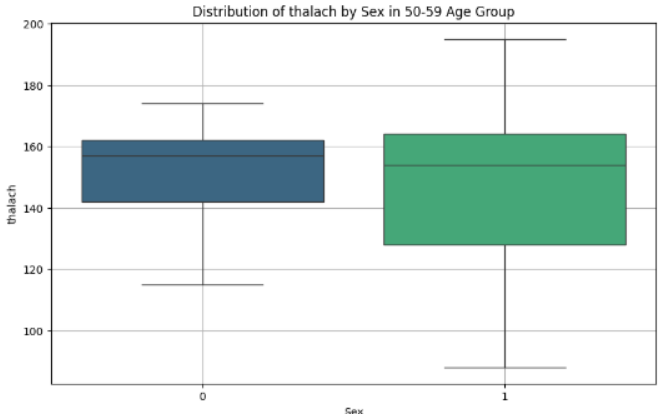
- **CP:** Both sexes portray a uniform distribution in this age group.
- **RESTECG:** Both sexes portray a uniform distribution in this age group.
- **EXANG:** Both sexes portray a uniform distribution in this age group
- Distribution of **THALACH** by Sex in the 50-59 Age Group: For sex=0, the median maximum heart rate achieved (thalach) is around 155 beats per minute (bpm), as indicated by the middle line of the box. For sex=1, the median thalach is slightly lower, around 150 bpm.The box representing the IQR for sex=0 is larger than that of sex=1, suggesting greater variability in maximum heart rate achieved among males in this age group
- **ST Depression (oldpeak):** sex=1 exhibit higher levels of ST depression induced by exercise relative to rest, indicating more severe ischemic changes during stress.

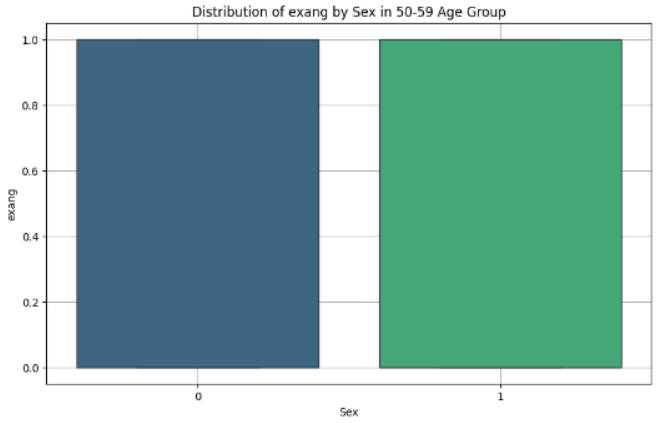


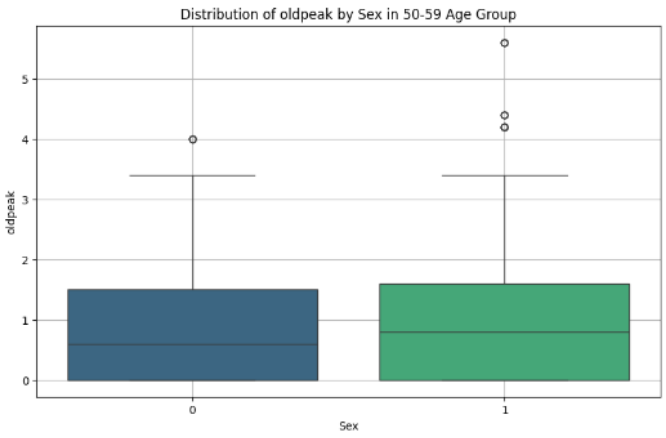




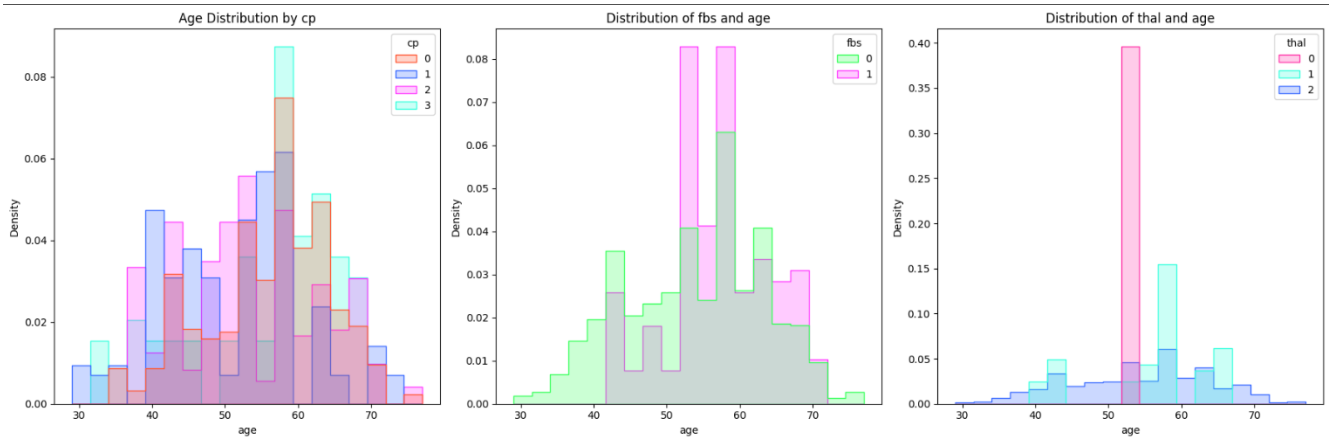




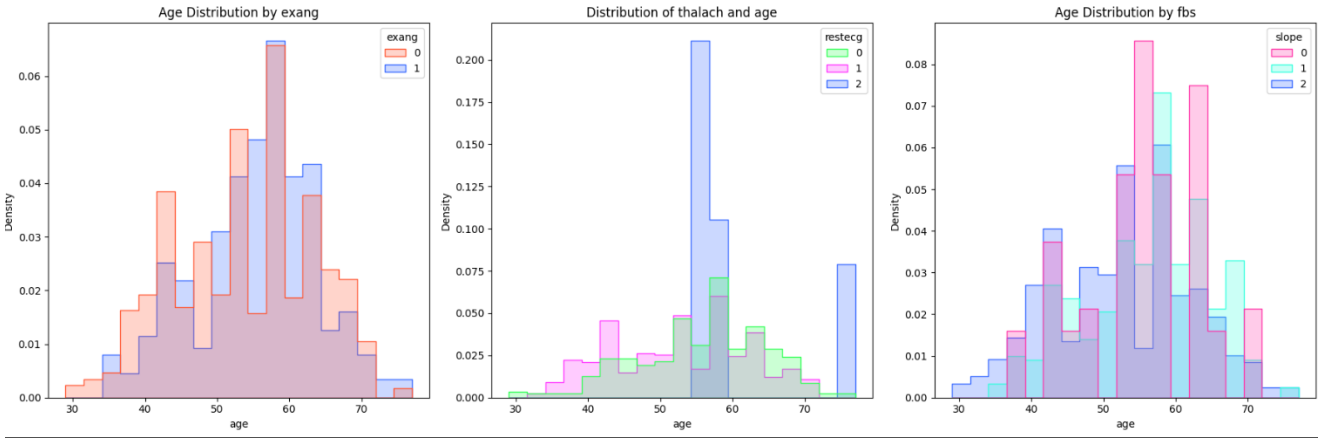




- The first plot for cp shows the highest frequency around ages 58-60 for all categories. Category 3 is the highest category
  The second plot for fbs has a bimodal distribution for category 1of fbs across ages. The highest frequency can be seen with age 57-60 for both categories. Category 1 has higher frequency than category 0
  The third plot for thal has peaks around ages 52-55 for category 0 and 58-60 for category 1 and 2. Category 0 has higher frequency than category as compared to category 1 and 2



- The first plot for exang shows the highest frequency around ages 58-60 for all categories. Category 1 has higher frequency as compared to category 0
  The second plot for thalach shows highest frequency for thalach category for category 1of fbs across ages. The highest frequency can be seen with age 57-60 for both categories. Category 2 has higher frequency as compared to category 0 and 1
  The third plot for thal has peaks around ages 52-55 for category 0 and 58-60 for category 1 and 2. Category 0 has higher frequency as compared to category 1 and 2

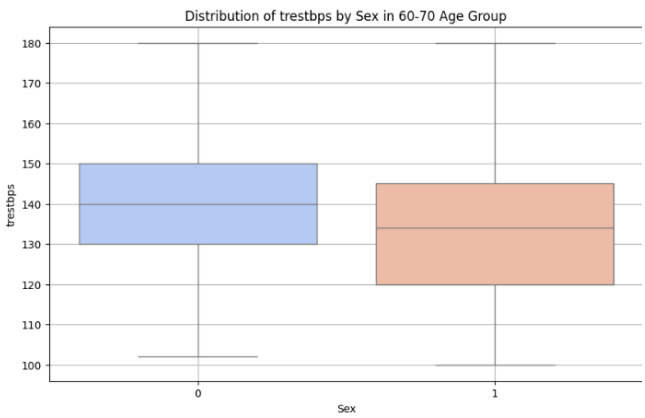


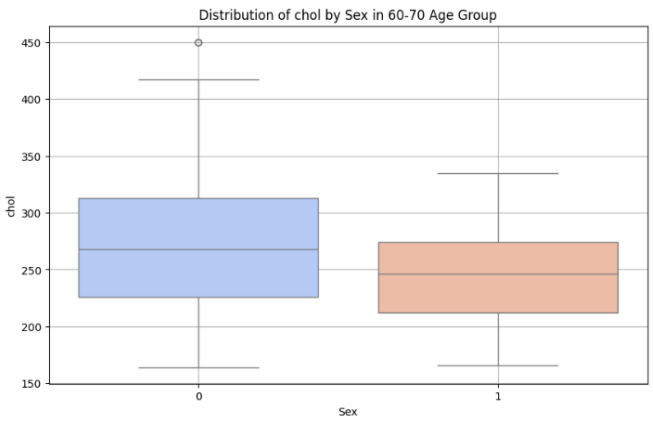
- The age distribution for sex=0, is 50-63 years with exang being at peak for category 0 as compared to category 1. The age distribution for sex=1, 52-62, with exang category 0 at its peak

- **60-70 age group**

  1. 60-70 age group for Thestbps: Thestbps is higher for sex=0 as compared to sex=1

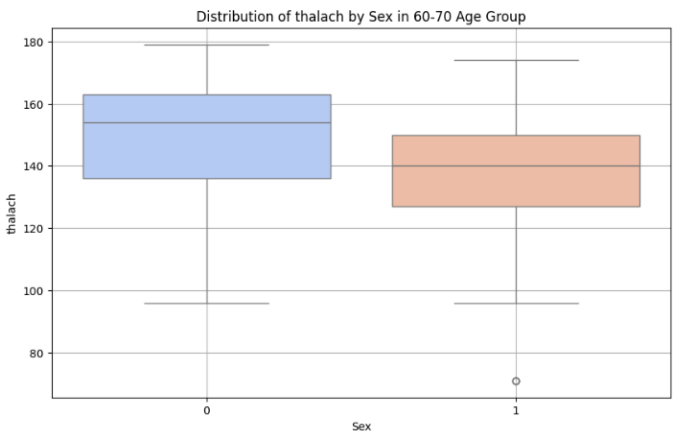
  2. 60-70 age group for cholesterol: Cholesterol is higher for sex=1 as compared to sex=0

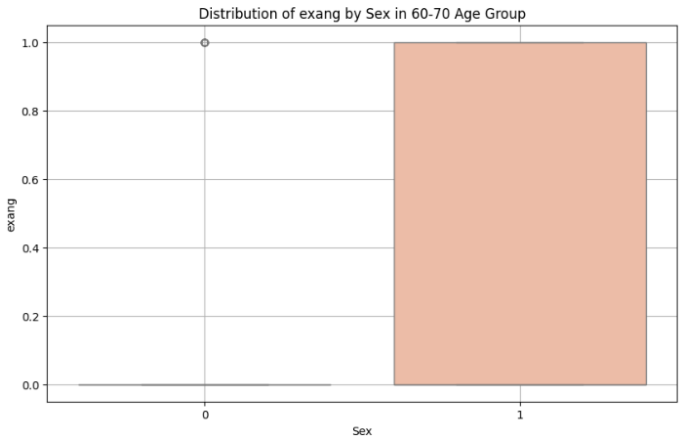




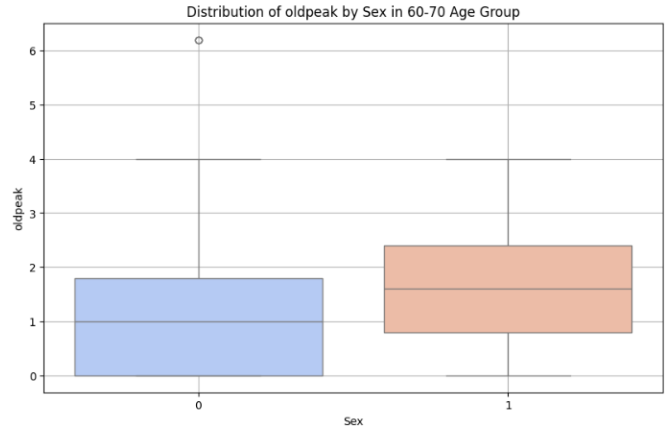
  3. 60-70 age group for thalach: Thalach is higher for sex=0 as compared to sex=1

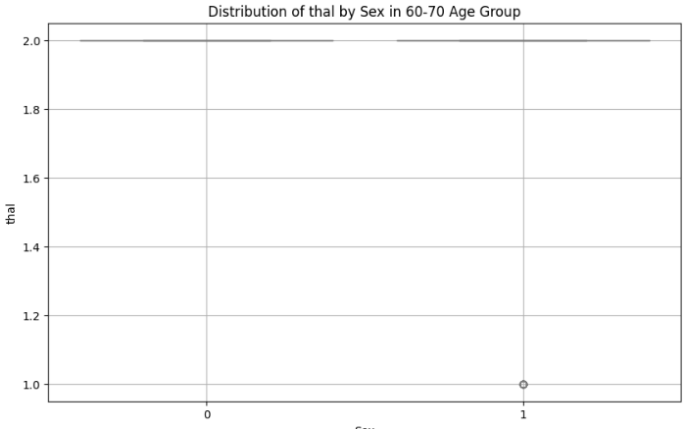
  4. 60-70 age group for exang: Cholesterol is higher for sex=1 and there is no sex=1 for this metrics





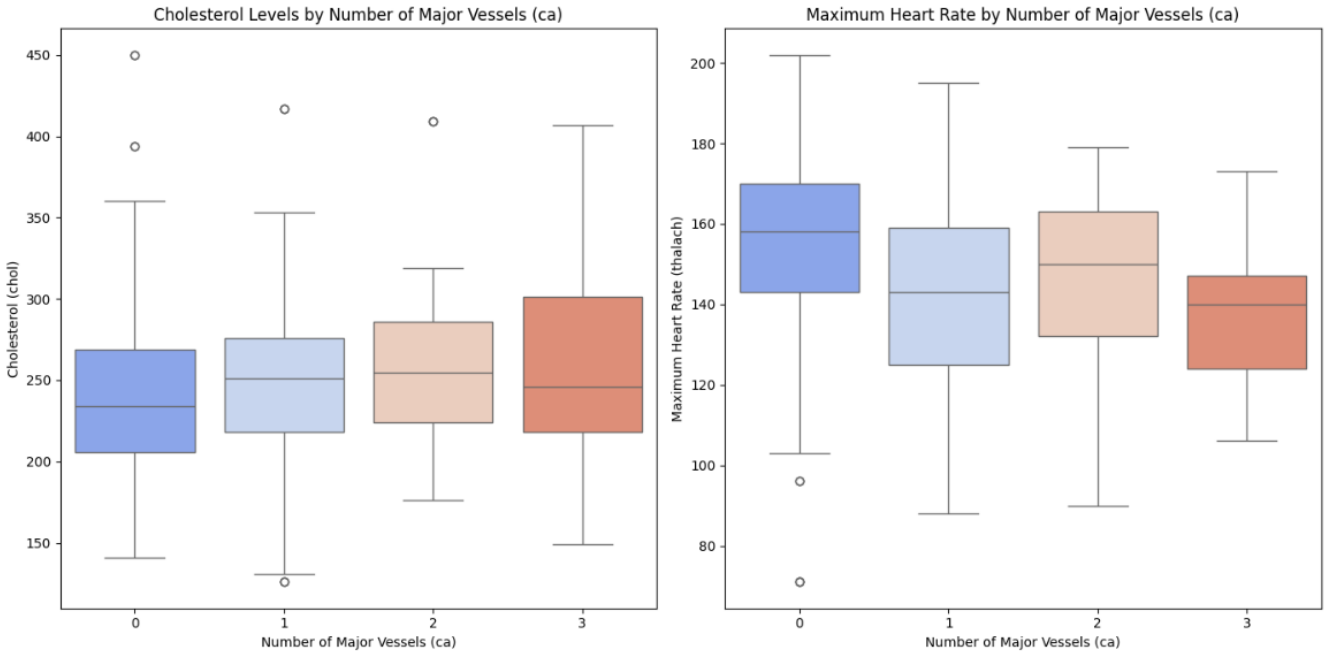
5. 60-70 age group for oldpeak: oldpeak is higher for sex=1 as compared to sex=0
6. 60-70 age group for thal: There is no or negligible record of thal for this age group
7. 60-70 age group for target: There is no or negligible record of target for sex=1 for this age group. One can only see a majority for sex=0





Boxplot distribution

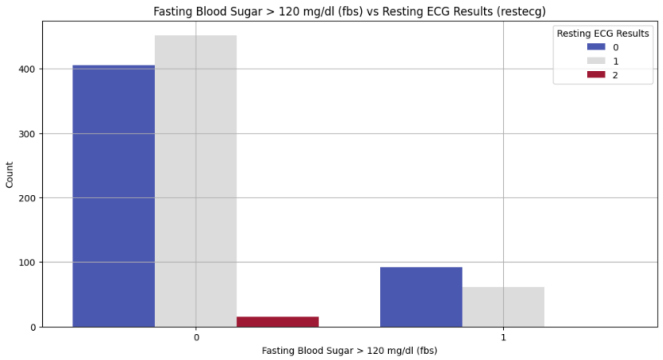
Cholesterol by ca: The cholesterol is highest for ca category 3 and lowest for ca category 0
Thalach by ca: The thalach is highest for ca category 0 and lowest for ca category 3

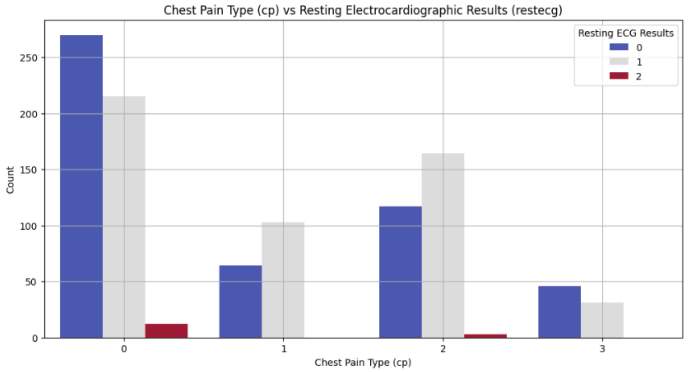


- **Bar chart distribution**

1. Fbs vs restecg: The fbs =0 is highest for restecg category 1 and lowest for restecg category 2. For Fbs=1, the restecg category 0 is the highest

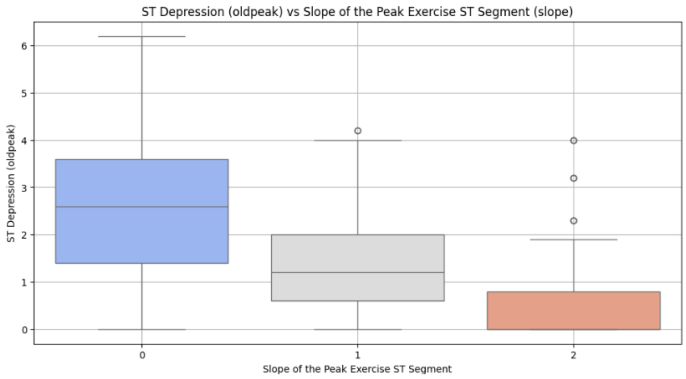
2. Chest pain cp vs restecg: The cp =0 is highest for restecg category 0 and lowest for restecg category 2. For cp=1, the restecg category 1 is the highest, for cp=2, the restecg category 1 is the highest. It is observed that restech category 1 is the highest across all cp category while categpry 2 is negligible or non existent across all cp categories

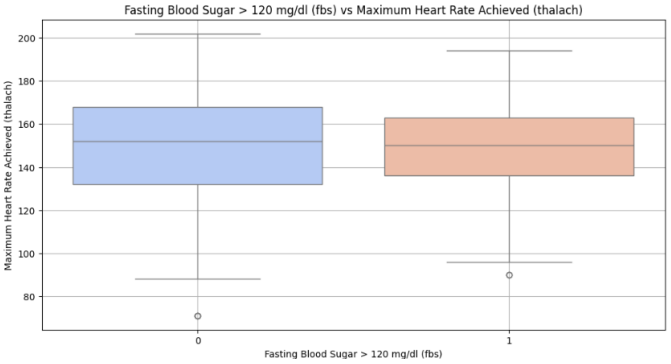




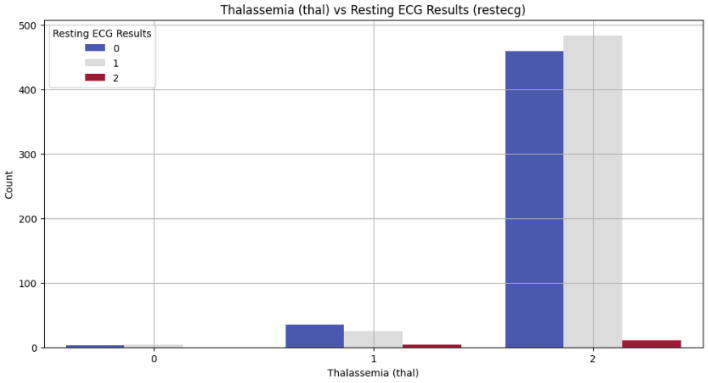
- **Old peak vs slope:** The slope =0 is highest in oldpeak category and lowest for slope category 2.

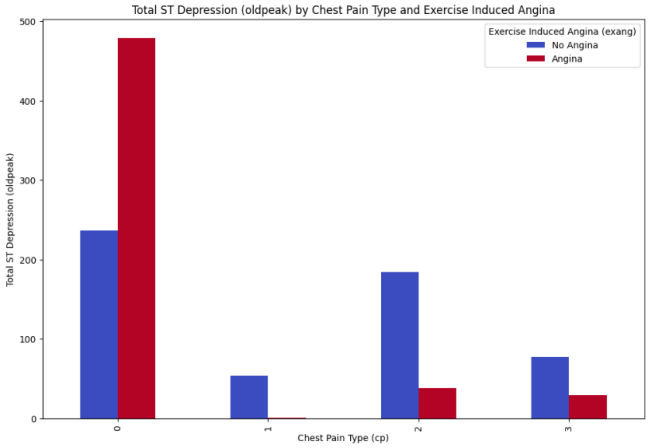
- **Fbs VS thalach:** Fbs =0 is highest in thalach





- **Thalach vs Restecg:** The oldpeak category 0 is highest across restecg category 0 and 1. The old peak category 1 is highest for restech category 2
- **Cp, exang and oldpeak:** Majority of cp cases have no angina except for cp category 0, where angina is highest across than cp all categories and has the highest old peak. Overall, the old peak is high for no angina cases across all cp categories except category 0.

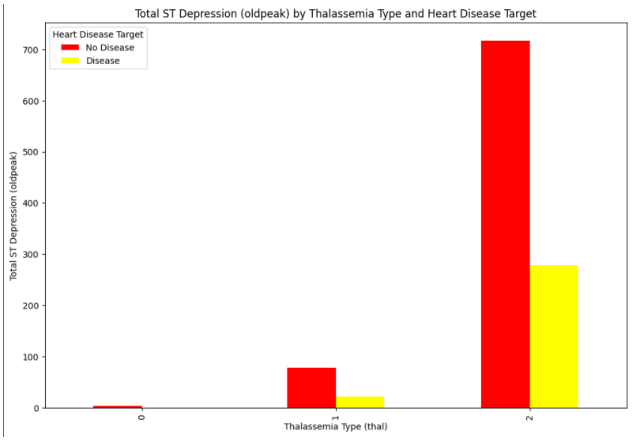


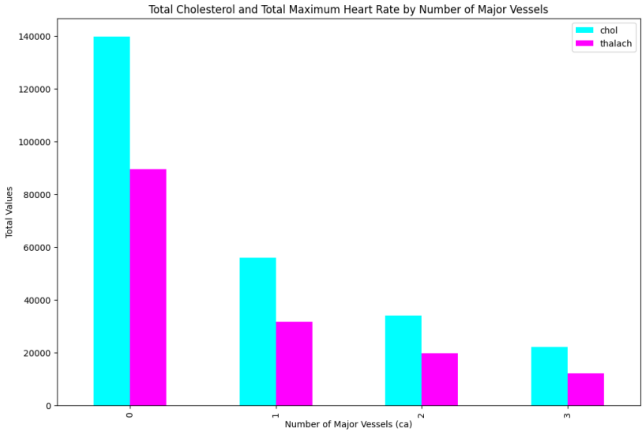


Bar distribution

- **Old peak VS thal vs target:** Majority of the thal cases have a high target =no disease and high old peak. The likelihood for disease is highest at category 2 of thal and at old peak close to 280

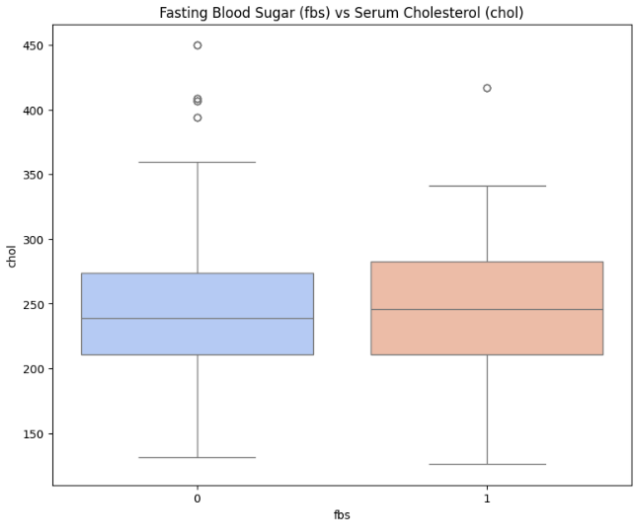
- **Thalach vs Old peak:** The oldpeak category 0 is highest across restecg category 0 and 1. The old peak category 1 is highest for restech category 2
- **Ca vs chol vs thalach:** ca and cholesterol is highest in category 0 and lowest in category 3.

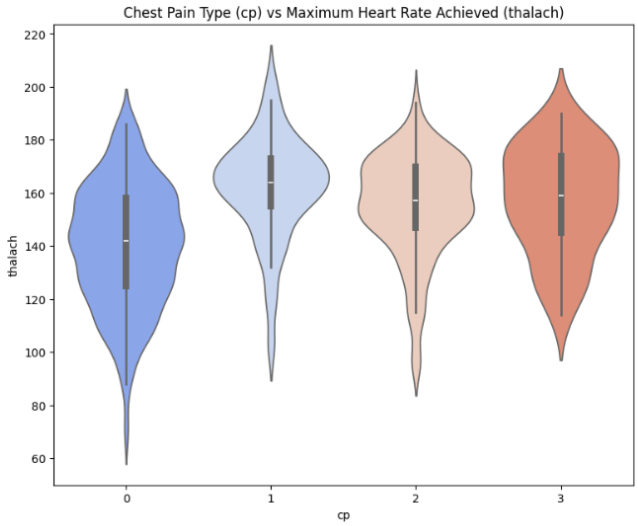




- **Fbs vs Chol:** There is no disease as a majority for all categories for thal and majority of distribution is on the higher side for old peak across all categories for no disease. We can conclude that thal and old peak on the higher side have no heart disease. The highest likelihood of heart disease is found for category 2 for thal, the peak being around 280 for old peak.

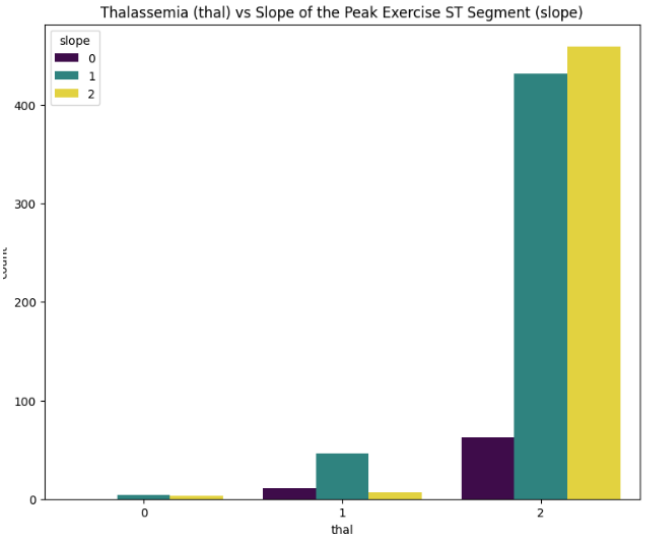
- **Cp vs thalach**: Thalach is highest for cp category 1 and lowest for category 0

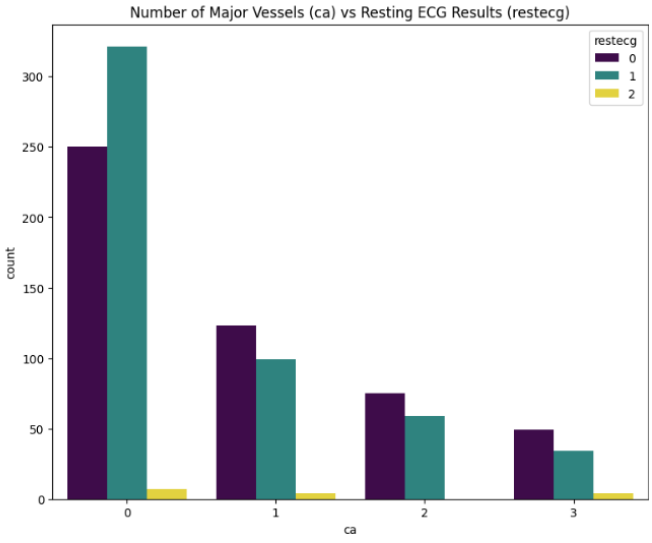




Thal vs slope: Slope category 1 is high across all levels thal levels accept thal category 2, which sees the highest distribution peak among all the thal categories.

Ca vs restecg: restecg category 0 is high across all levels ca levels accept ca category 0, which sees the highest distribution peak among all the ca categories. restecg category 2 is lowest across all ca categories.

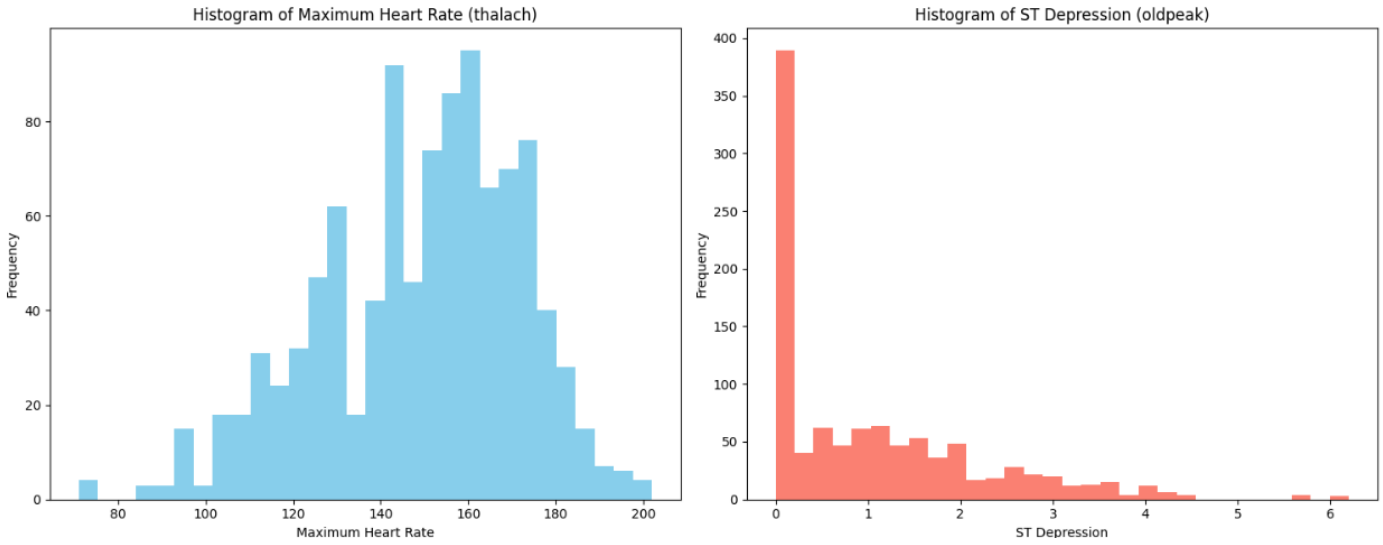




Histogram distribution

- Thalch distribution: Thalach is at its highest frequency between 158-162 and lowest between 85-100

- Oldpeak distribution: Olapeak is at its highest frequency between 0- 0.2 and lowest between beyond 5.5-6.2



DATA MODELLING

- **Logistic regression and stats model** is used to understand the statistical summary of the features and dataset
- Used the **VIF method for feature engineering**
- Dropped the column **thal and thalach** in the train and test data.
- Have found the confusion matrix for test and train features and the ROC curve
- **Top features for test are: sex, oldpeak, cp, exang, ca , fbs**
- **Top features for train are: cp, restecg, exang, ca , fbs**
- **we can conclude ca, fbs and exang are the top features for the target column**

The accuracy for train data is **82.4%** whereas for test data is **85%**