# THE ROLE OF EXERCISE-INDUCED MAXIMUM HEART RATE AND CHEST PAIN TYPES IN PREDICTING HEART DISEASE

Andrea Loizidou (PID: 6271053)

FALL 2023

STA 6244 – DATA ANALYSIS 1
FLORIDA INTERNATIONAL UNIVERSITY

# Contents

## List of Figures

# Introduction and Purpose

For this project, I aimed to investigate the relationship between the various types of chest pain, maximum heart rate, and the likelihood of being diagnosed with heart disease. This research has the potential to offer valuable insights that could aid medical professionals in diagnosing and treating patients, as well as help patients better understand their own risk factors. Using visual aids, I analyzed multiple variables including age, resting blood pressure, and chest pain types. To determine the significance of any differences or associations, I utilized a range of statistical tests such as t-tests, ANOVA, Turkey's range test, and chi-square tests. Lastly, I employed a contingency table to help visualize any variable interactions.

# Methodology

## Data Collection

The data for this project was obtained from a sample of patients at the Cleveland Clinic Foundation. It was downloaded from the UC Irvine Machine Learning Repository[1] and processed to fit the specific needs of this project. The data in the UCI repository was collected by four main institutions, namely:

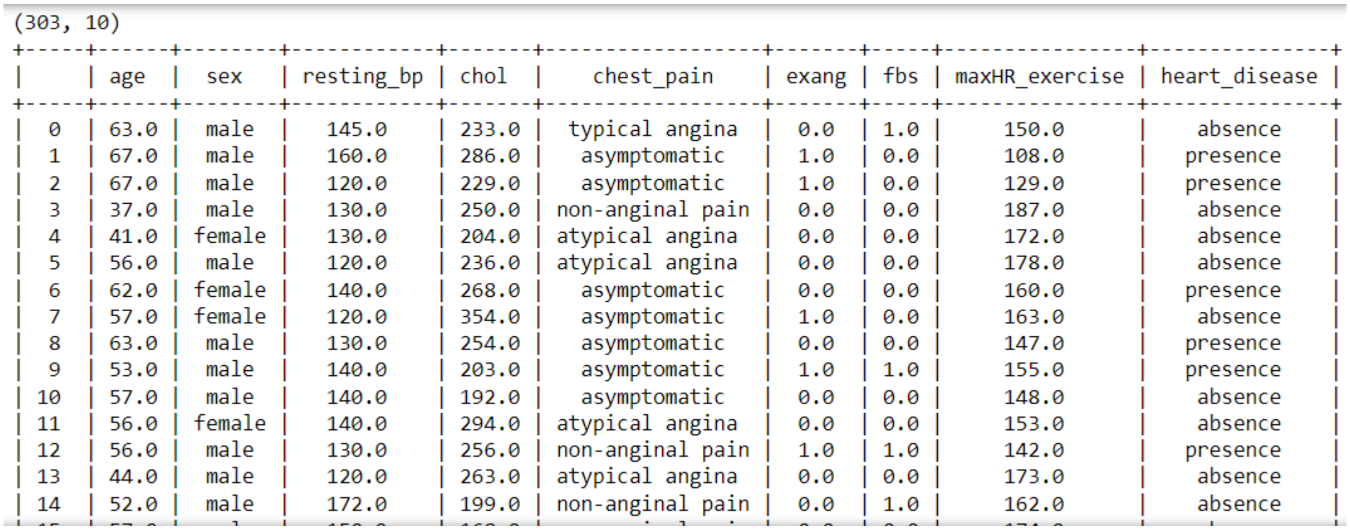
1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

The data used for analysis contained missing values, so only 302 people were analyzed using 9 attributes relevant to the investigation. See the variables table below:

[1] https://archive.ics.uci.edu/dataset/45/heart+disease

| Variable Name | Type | Description | Units |
|---|---|---|---|
| **Age** | Numerical | The age of the patient | Years |
| **Sex** | Categorical | The sex assigned at birth, which can be either male or female | |
| **resting_bp** | Numerical | The resting blood pressure | mm Hg |
| **Chol** | Numerical | The serum cholesterol | mg/dl |
| **chest_pain** | Categorical | The type of chest pain experienced by the patient, which can be typical angina, atypical angina, non-anginal pain, or asymptomatic. | |
| **Exang** | Numerical | A binary value (1 or 0) indicating whether the patient experiences exercise-induced angina (1: yes; 0: no). | |
| **Fbs** | Numerical | A binary value (1 or 0) indicating whether the patients fasting blood sugar is greater than 120 mg/dl | mg/dl |
| **maxHR_exercise** | Numerical | The maximum heart rate achieved during an exercise test | |
| **Heart_disease** | Categorical | A binary value (presence or absence) indicating whether the patient is diagnosed with heart disease. | |

Here is a preview of the first 14 instances of the data after processing:



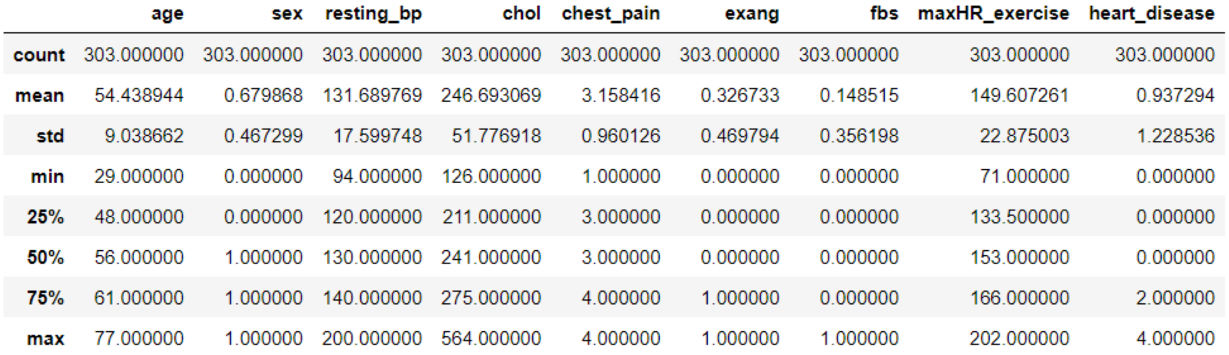
```
(303, 10)
+-----+------+--------+------------+-------+------------------+-------+-----+----------------+---------------+
|     | age  |  sex   | resting_bp | chol  |    chest_pain    | exang | fbs | maxHR_exercise | heart_disease |
+-----+------+--------+------------+-------+------------------+-------+-----+----------------+---------------+
|  0  | 63.0 |  male  |   145.0    | 233.0 |  typical angina  |  0.0  | 1.0 |     150.0      |    absence    |
|  1  | 67.0 |  male  |   160.0    | 286.0 |   asymptomatic   |  1.0  | 0.0 |     108.0      |   presence    |
|  2  | 67.0 |  male  |   120.0    | 229.0 |   asymptomatic   |  1.0  | 0.0 |     129.0      |   presence    |
|  3  | 37.0 |  male  |   130.0    | 250.0 | non-anginal pain |  0.0  | 0.0 |     187.0      |    absence    |
|  4  | 41.0 | female |   130.0    | 204.0 | atypical angina  |  0.0  | 0.0 |     172.0      |    absence    |
|  5  | 56.0 |  male  |   120.0    | 236.0 | atypical angina  |  0.0  | 0.0 |     178.0      |    absence    |
|  6  | 62.0 | female |   140.0    | 268.0 |   asymptomatic   |  0.0  | 0.0 |     160.0      |   presence    |
|  7  | 57.0 | female |   120.0    | 354.0 |   asymptomatic   |  1.0  | 0.0 |     163.0      |    absence    |
|  8  | 63.0 |  male  |   130.0    | 254.0 |   asymptomatic   |  0.0  | 0.0 |     147.0      |   presence    |
|  9  | 53.0 |  male  |   140.0    | 203.0 |   asymptomatic   |  1.0  | 1.0 |     155.0      |   presence    |
| 10  | 57.0 |  male  |   140.0    | 192.0 |   asymptomatic   |  0.0  | 0.0 |     148.0      |    absence    |
| 11  | 56.0 | female |   140.0    | 294.0 | atypical angina  |  0.0  | 0.0 |     153.0      |    absence    |
| 12  | 56.0 |  male  |   130.0    | 256.0 | non-anginal pain |  1.0  | 1.0 |     142.0      |   presence    |
| 13  | 44.0 |  male  |   120.0    | 263.0 | atypical angina  |  0.0  | 0.0 |     173.0      |    absence    |
| 14  | 52.0 |  male  |   172.0    | 199.0 | non-anginal pain |  0.0  | 1.0 |     162.0      |    absence    |
```

*Figure 1: Processed Data Table Preview*

## Data Pre-processing

The analysis of the specific dataset required extensive data processing due to several key operations that needed to be performed before analyzing the dataset. Initially, the data was inspected to understand its structure and types. Then, missing values were replaced with a standard missing marker in pandas, and categorical data was converted to numerical. For example, the 'sex' column had all 'females' labeled as 0 and all 'males' labeled as 1, while other non-numerical values were treated similarly. Conversely, some data, such as the 'chest_pain' variable, had values 1, 2, 3, 4 replaced with 'typical angina', 'atypical angina', 'non-anginal pain', and 'asymptomatic', respectively. The target variable 'heart_disease' was reclassified from binary into 'presence' and 'absence'. Afterward, the data was inspected again, and a comprehensive description of the dataset was provided, including count, unique values, top categories, frequency, mean, standard deviation, and ranges for each column. All data preprocessing was done in a separate Jupyter notebook for code simplification, and both Jupyter notebooks were submitted for review. The data preprocessing and analysis were performed using Python and various Python libraries, including Pandas, NumPy, Matplotlib, and others.

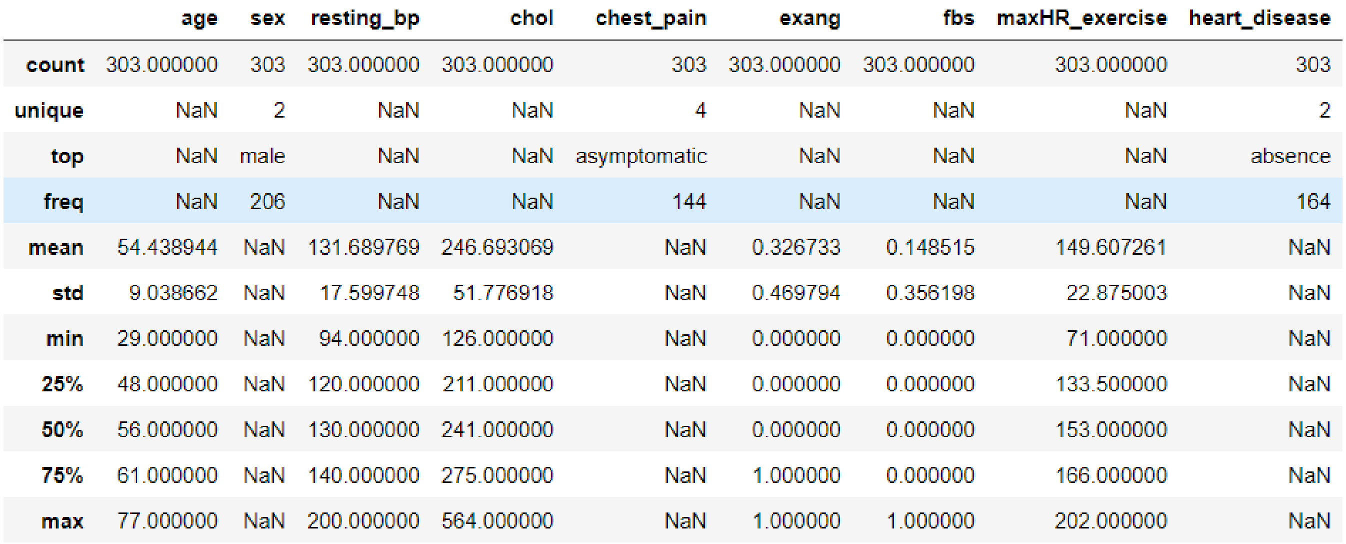
Below is a table of the raw data obtained before any processing. The table has some erratic values, such as the sex variable, which cannot be averaged. After a thorough review of the initial data, I decided to drop some unnecessary columns, rename certain columns to make the data more readable, and convert some numerical variables to categorical. For example, I converted the 'sex' variable from 1 and 0 to 'female' and 'male', and the chest pain values from 1, 2, 3, 4 to the respective type of chest pain. Subsequently, I obtained a summary of all the data variables to aid in my understanding and approach to analyzing the data.



| | age | sex | resting_bp | chol | chest_pain | exang | fbs | maxHR_exercise | heart_disease |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 |
| **mean** | 54.438944 | 0.679868 | 131.689769 | 246.693069 | 3.158416 | 0.326733 | 0.148515 | 149.607261 | 0.937294 |
| **std** | 9.038662 | 0.467299 | 17.599748 | 51.776918 | 0.960126 | 0.469794 | 0.356198 | 22.875003 | 1.228536 |
| **min** | 29.000000 | 0.000000 | 94.000000 | 126.000000 | 1.000000 | 0.000000 | 0.000000 | 71.000000 | 0.000000 |
| **25%** | 48.000000 | 0.000000 | 120.000000 | 211.000000 | 3.000000 | 0.000000 | 0.000000 | 133.500000 | 0.000000 |
| **50%** | 56.000000 | 1.000000 | 130.000000 | 241.000000 | 3.000000 | 0.000000 | 0.000000 | 153.000000 | 0.000000 |
| **75%** | 61.000000 | 1.000000 | 140.000000 | 275.000000 | 4.000000 | 1.000000 | 0.000000 | 166.000000 | 2.000000 |
| **max** | 77.000000 | 1.000000 | 200.000000 | 564.000000 | 4.000000 | 1.000000 | 1.000000 | 202.000000 | 4.000000 |

*Figure 2: Statistical Summary for all Attributes before processing*

After completing all pre-processing steps, here is the summary of the data that I analyzed during my investigation. After completing all pre-processing steps, here is the summary of the data that I analyzed during my investigation:

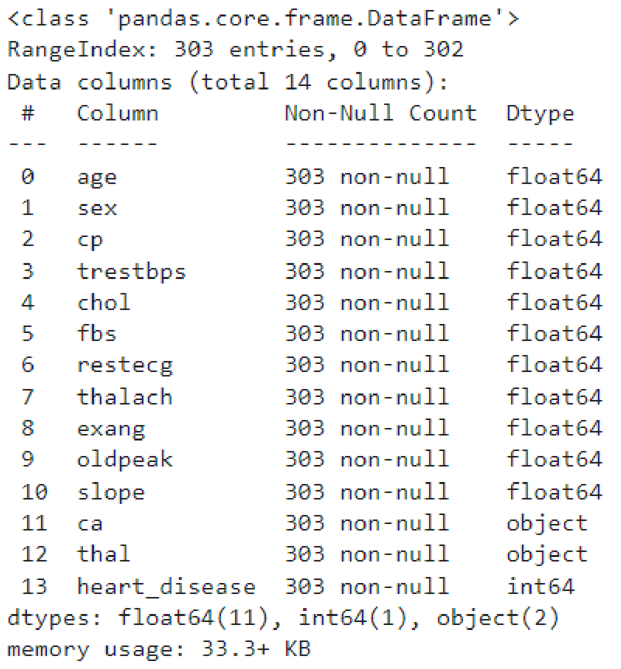


| | age | sex | resting_bp | chol | chest_pain | exang | fbs | maxHR_exercise | heart_disease |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 303.000000 | 303 | 303.000000 | 303.000000 | 303 | 303.000000 | 303.000000 | 303.000000 | 303 |
| **unique** | NaN | 2 | NaN | NaN | 4 | NaN | NaN | NaN | 2 |
| **top** | NaN | male | NaN | NaN | asymptomatic | NaN | NaN | NaN | absence |
| **freq** | NaN | 206 | NaN | NaN | 144 | NaN | NaN | NaN | 164 |
| **mean** | 54.438944 | NaN | 131.689769 | 246.693069 | NaN | 0.326733 | 0.148515 | 149.607261 | NaN |
| **std** | 9.038662 | NaN | 17.599748 | 51.776918 | NaN | 0.469794 | 0.356198 | 22.875003 | NaN |
| **min** | 29.000000 | NaN | 94.000000 | 126.000000 | NaN | 0.000000 | 0.000000 | 71.000000 | NaN |
| **25%** | 48.000000 | NaN | 120.000000 | 211.000000 | NaN | 0.000000 | 0.000000 | 133.500000 | NaN |
| **50%** | 56.000000 | NaN | 130.000000 | 241.000000 | NaN | 0.000000 | 0.000000 | 153.000000 | NaN |
| **75%** | 61.000000 | NaN | 140.000000 | 275.000000 | NaN | 1.000000 | 0.000000 | 166.000000 | NaN |
| **max** | 77.000000 | NaN | 200.000000 | 564.000000 | NaN | 1.000000 | 1.000000 | 202.000000 | NaN |

*Figure 3: Statistical Summary for All Attributes after processing*

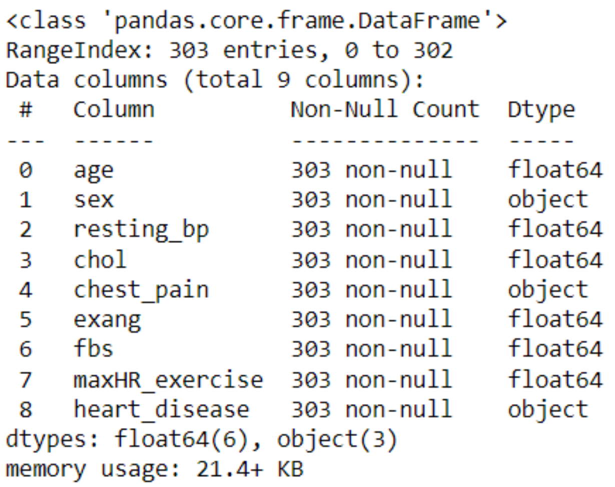
Comparison of Variable before and after processing:

BEFORE:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   age            303 non-null    float64
 1   sex            303 non-null    float64
 2   cp             303 non-null    float64
 3   trestbps       303 non-null    float64
 4   chol           303 non-null    float64
 5   fbs            303 non-null    float64
 6   restecg        303 non-null    float64
 7   thalach        303 non-null    float64
 8   exang          303 non-null    float64
 9   oldpeak        303 non-null    float64
 10  slope          303 non-null    float64
 11  ca             303 non-null    object 
 12  thal           303 non-null    object 
 13  heart_disease  303 non-null    int64  
dtypes: float64(11), int64(1), object(2)
memory usage: 33.3+ KB
```

AFTER:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   age             303 non-null    float64
 1   sex             303 non-null    object 
 2   resting_bp      303 non-null    float64
 3   chol            303 non-null    float64
 4   chest_pain      303 non-null    object 
 5   exang           303 non-null    float64
 6   fbs             303 non-null    float64
 7   maxHR_exercise  303 non-null    float64
 8   heart_disease   303 non-null    object 
dtypes: float64(6), object(3)
memory usage: 21.4+ KB
```

When working with a dataset, it is necessary to identify the variables that are presumed to have an impact on or predict the outcome of interest. These variables are known as independent variables. On the other hand, the dependent variable is the outcome of interest that we aim to predict or explain based on the independent variables. In other words, the dependent variable is affected by the independent variables.

In this scenario, the objective is to predict the presence or absence of heart disease. Therefore, heart disease is considered the dependent variable while all the other variables are independent. These independent variables are factors that may contribute or affect the outcome..

**Dependent Variables:**

- heart_disease

**Independent Variables:**

- age - Age of the individual.
- sex - Gender of the individual.
- resting_bp - Resting blood pressure.
- chol - Cholesterol level.
- chest_pain - Type of chest pain.
- exang - Exercise-induced angina (chest pain).
- fbs - Fasting blood sugar.
- maxHR_exercise - Maximum heart rate achieved during exercise.

## Method of Analysis:

Various statistical methods were used to analyze heart disease data and investigate relationships between variables. The independent two-sample T-test compared means of continuous variables between two groups. ANOVA compared means of more than two groups to assess differences in continuous variables, and Tukey's HSD Test identified which group means were different. The analysis aimed to find patterns and relationships within the data, contributing to a better understanding of factors associated with heart disease.

The analysis will first examine whether heart rate, age, cholesterol, and resting blood pressure are significantly associated with heart disease, using box plots and t-tests.Then, ANOVA will be performed to analyze he relationship between the maximum heart rate during exercise and the different types of chest pain and lastly Tukey's HSD Test was performed to compare the mean differences of specific chest pains

# Statistical Analysis:

## T-tests:

### Heart Rate

In this dataset, each patient underwent a fitness test to monitor their maximum heart rate while exercising (maxHR_exercise), and the highest recorded heart rate was noted. The primary question to consider is whether there is an association between the maximum heart rate while exercising and the diagnosis of heart disease in patients.

I started by creating two boxplots - one for patients diagnosed with heart disease and one for those without. To make the comparison more visual, I placed the boxplot for the presence of heart disease next to the boxplot for the absence of heart disease.

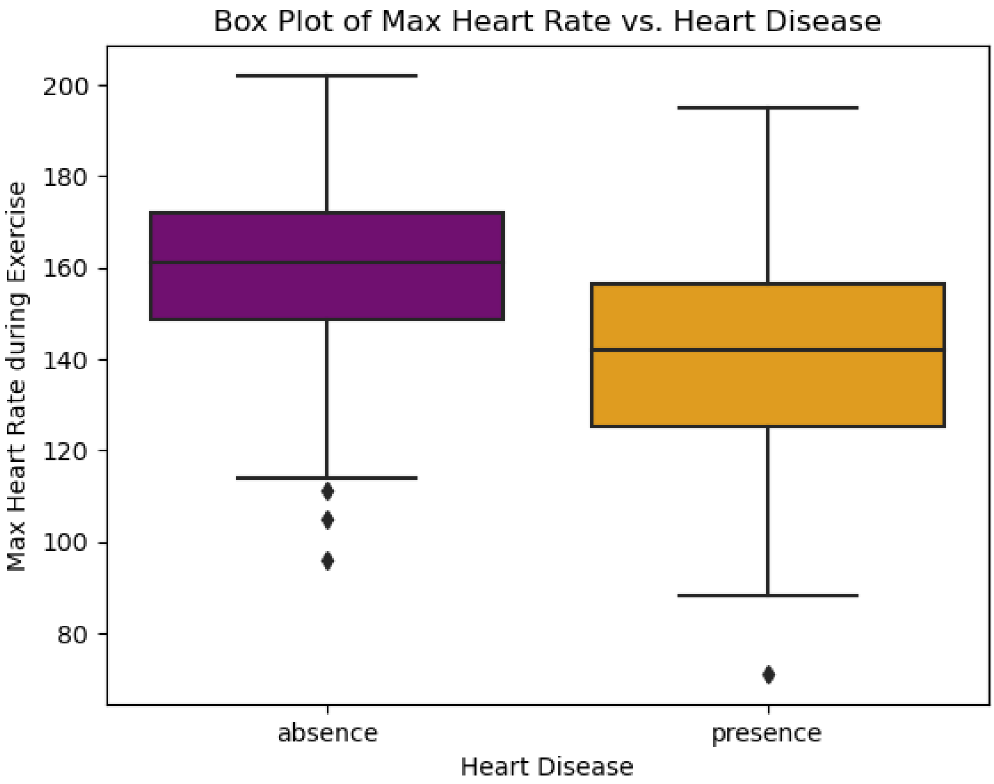
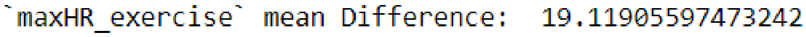


*Figure 4: Box plot illustrating the distribution of maximum heart rate during exercise for individuals with and without heart disease*

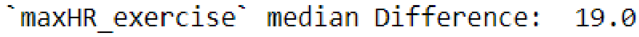
It seems that individuals without heart disease (labeled as 'absence') have a higher median maximum heart rate compared to those with heart disease (labeled as 'presence'). The middle 50% of the data, represented by the Interquartile Range (IQR), appears to be wider for individuals without heart disease than those with heart disease. Both groups have outliers, but there are a few lower outliers in the 'absence' category, indicating that some individuals without heart disease have unusually low maximum heart rates during exercise. The overall spread of maximum heart rates is broad for both groups, which suggests significant variability within each group.

After examining the visualization provided, it appears that there is a correlation between maximum heart rate during exercise and the presence of heart disease. To be more specific, individuals who do not have heart disease tend to have a higher maximum heart rate during exercise compared to those with heart disease. This suggests that a high maximum heart rate during exercise could be an indicator of being free from heart disease. However, just a boxplot is not enough to answer our question, I will need to investigate further with other statistical tests.

In addition to the box plot, we calculated the mean and median difference:



```
`maxHR_exercise` mean Difference:  19.11905597473242
```



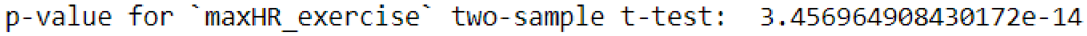
```
`maxHR_exercise` median Difference:  19.0
```

Upon analysis of the maximum heart rate during exercise, it was found that there is a mean difference of approximately 19.12 beats per minute (bpm) and a median difference of 19 bpm between individuals with and without heart disease. This statistical difference reinforces the observation from the box plot, indicating a potential correlation between the maximum heart rate during exercise and the presence of heart disease.

The next step is to determine if there is a significant difference between the average maximum heart rates of patients with heart disease and those without heart disease. We'll be using a two sample t-test to test the following hypothesis:

**Null Hypothesis:** The average maximum heart rate (maxHR_exercise) of patients with heart disease is equal to that of patients without heart disease.

**Alternative Hypothesis:** The average maximum heart rate (maxHR_exercise) of patients with heart disease is not equal to that of patients without heart disease.



```
p-value for `maxHR_exercise` two-sample t-test:  3.456964908430172e-14
```

The p-value that we got is a much smaller number than the 0.05 for statistical significance. Thus, we can reject the Null Hypothesis and conclude that there is a statistically significant difference in the average maximum heart rate during exercise between patients with heart disease and those without it.

Age:

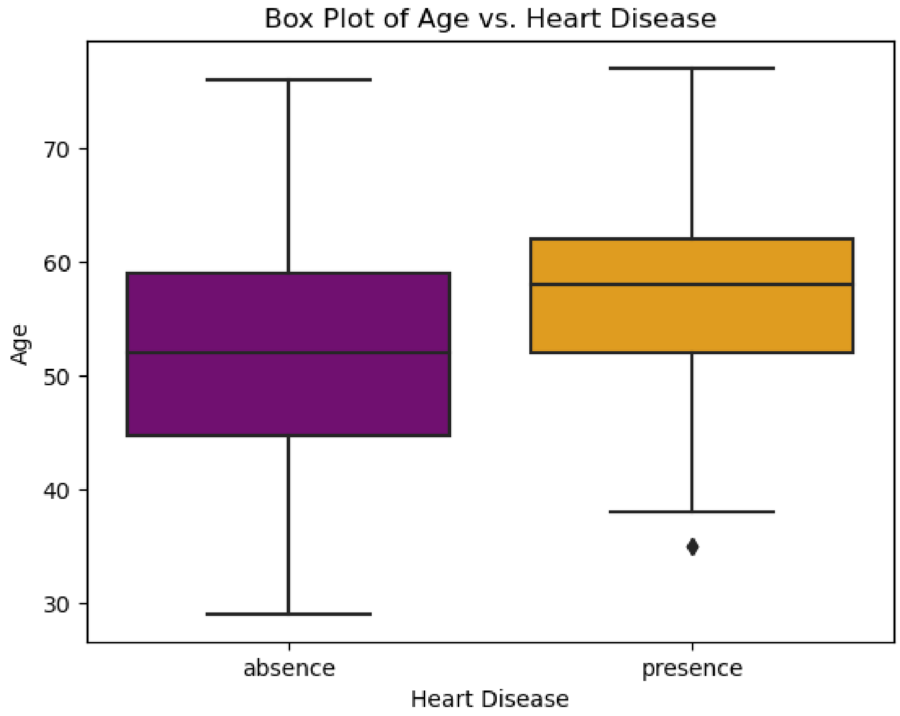
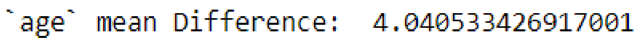


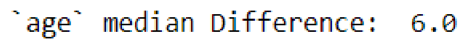
*Figure 5: Box Plot illustrating the distribution of Age for individuals with and without heart disease*

Individuals with heart disease are older than those without, as shown by the median age line. The interquartile range is slightly wider for those with heart disease, suggesting more variability in age. The 'whiskers' on the plot extend further for those without heart disease, indicating a broader age range, while the whiskers for those with heart disease are shorter, indicating a less varied age range. At least one outlier in the group with heart disease suggests an individual significantly younger than the rest within that group.

The plot suggests that there might be a relation between age and heart disease, as older individuals seem to be more prone to the condition. This observation is supported by medical knowledge that establishes a correlation between age and the risk of heart disease. However, it's important to note that the box plot only shows a visual association and doesn't prove causality or the strength of the relationship. To determine the extent and significance of the link between age and heart disease, the mean and median difference was calculated



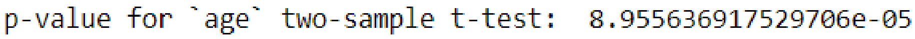
```
`age` mean Difference:  4.040533426917001
```



```
`age` median Difference:  6.0
```

These statistics provide additional information to support the conclusions drawn from the box plot. The median age difference, in particular, is a reliable measure because it is less affected by outliers and skewed data. The fact that both the mean and median ages are higher for people with heart disease reinforces the conclusion that age is a significant factor in the risk of heart disease.

**Null Hypothesis:** No difference in average age between individuals with and without heart disease**.**
**Alternative Hypothesis:** There is a difference in average age between individuals with and without heart disease.



```
p-value for `age` two-sample t-test:  8.955636917529706e-05
```

The p-value obtained from the statistical test is significantly smaller than 0.05 (or 5%). Therefore, we can confidently reject the null hypothesis stating that there is no difference in the average age between individuals with and without heart disease. Based on the test results, we can conclude that the observed difference in average age is statistically significant, and it is highly unlikely that it occurred by chance.

### Resting blood pressure:

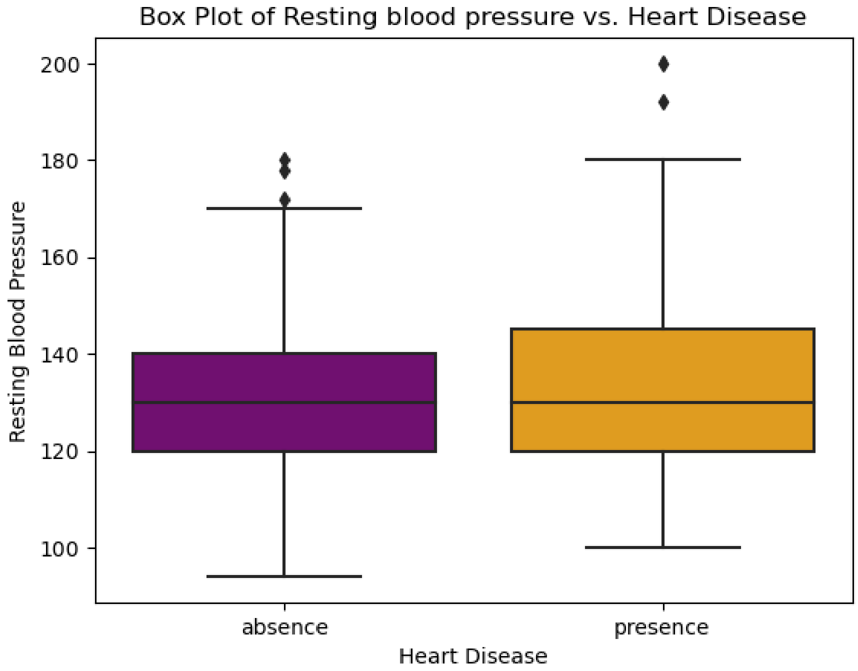
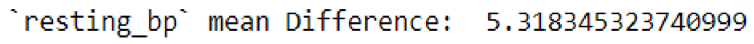


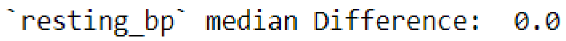
*Figure 6: Box Plot illustrating the distribution of resting blood pressure for individuals with and without heart disease*

Both groups have a median resting blood pressure above the normal range limit of 120 mmHg. The interquartile range for both groups is similar, indicating a similar range of blood pressure values. Outliers suggest individuals with unusually high blood pressure in both groups, while there is some variation between individuals with and without heart disease, the bulk of the data for both groups lies within a similar range. This suggests that resting blood pressure alone may not provide a significant difference between the two groups.

Based on the plot, it is not immediately evident if resting blood pressure is a strong indicator of the presence of heart disease. The overlap in the interquartile ranges (IQRs) and the range of values imply that while resting blood pressure may play a role in heart disease, it is probably not the sole contributing factor.



```
`resting_bp` mean Difference:  5.318345323740999
```

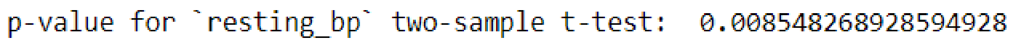


```
`resting_bp` median Difference:  0.0
```

Resting blood pressure is similar for individuals with and without heart disease, as the median difference is zero. However, the mean difference shows that those with heart disease have a higher average resting blood pressure. This might be due to outliers, particularly in the heart disease group. Therefore, this difference may not be significant for most people.

**Null Hypothesis:** No difference in mean resting blood pressure between the two groups

**Alternative Hypothesis**: There is a difference in mean resting blood pressure between the two groups



```
p-value for `resting_bp` two-sample t-test:  0.008548268928594928
```

The p-value for the observed difference in mean resting blood pressure between individuals with and without heart disease is less than the significance level of 0.05. This means that we reject the null hypothesis and conclude that the difference is statistically significant.

Even though the median difference is zero, indicating no difference in the central tendency of the distribution, the mean difference and the corresponding p-value from the t-test indicate that the overall averages of the two groups are significantly different. This difference is unlikely to have occurred by chance.

## Cholesterol

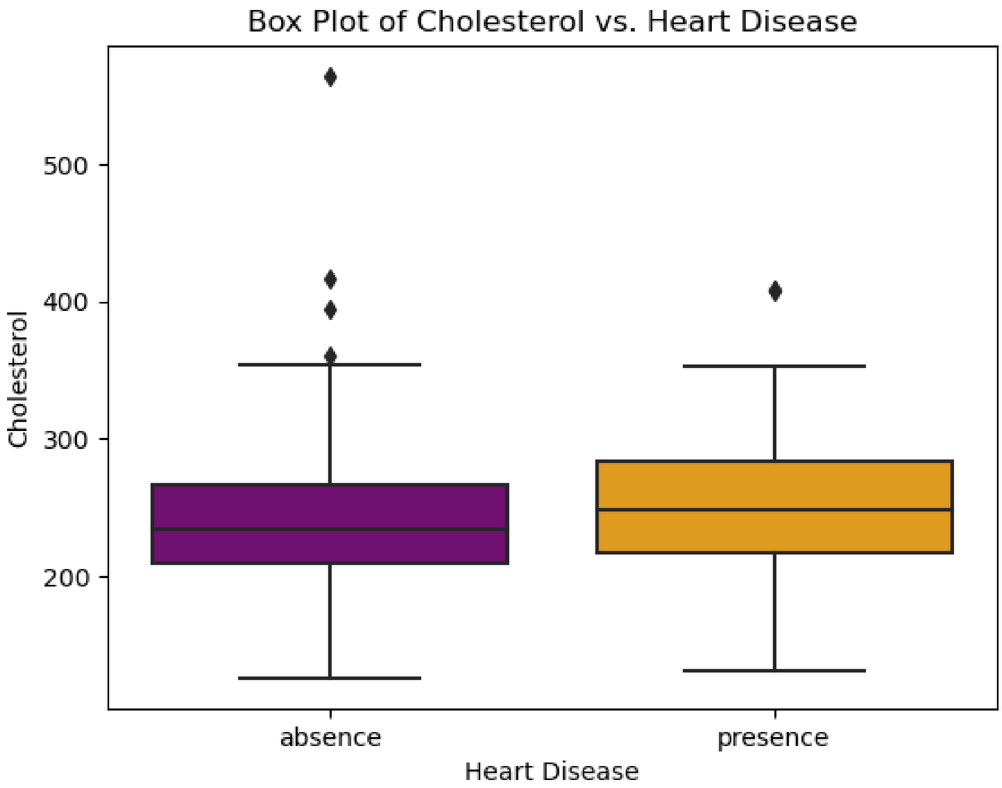
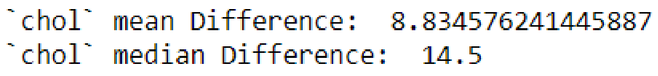


*Figure 7: Box Plot illustrating the distribution of Cholesterol for individuals with and without heart disease*

The median cholesterol level is slightly higher in the group with heart disease compared to the group without, suggesting a potential association. Both groups have a substantial interquartile range, with slightly less variability among individuals with heart disease. The range of cholesterol levels is similar for both groups, with several outliers in both.

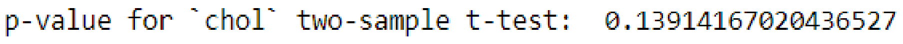
The plot suggests that although individuals with heart disease tend to have higher median cholesterol levels, there is significant overlap between the two groups. This implies that other factors may also play a role in heart disease. Outliers in both groups indicate that high cholesterol levels occur in individuals regardless of heart disease status.



```
`chol` mean Difference:  8.834576241445887
`chol` median Difference:  14.5
```

These findings suggest that individuals with heart disease tend to have higher cholesterol levels compared to those without. On average, the mean cholesterol level of individuals with heart disease is approximately 8.83 mg/dL higher than those without. The median cholesterol level is also 14.5 mg/dL higher in individuals with heart disease. This difference between the two groups indicates that there is indeed a correlation between heart disease and cholesterol levels. It is worth noting that the median difference is greater than the mean difference, which may suggest that some individuals with significantly high or low cholesterol readings are skewing the overall distribution.

**Null Hypothesis**: No difference in mean cholesterol levels between the two groups
**Alternative Hypothesis:** There is a difference in mean cholesterol levels between the two groups



```
p-value for `chol` two-sample t-test:  0.13914167020436527
```

This p-value suggests that the observed differences in mean cholesterol levels between individuals with and without heart disease are not statistically significant since it is greater than 0.05.

## ANOVA AND TURKEY'S TEST

### Chest Pain and Max Heart Rate

The objective is to conduct a thorough investigation into the relationship between the maximum heart rate during exercise and the different types of chest pain. The aim is to determine whether there is a significant difference in the maximum heart rate during exercise for any particular type of chest pain. This will help us understand the relationship between the two variables and identify any potential trends or patterns that may exist.To do so, we will perform an ANOVA test and continue by creating a pairwise turkey test.

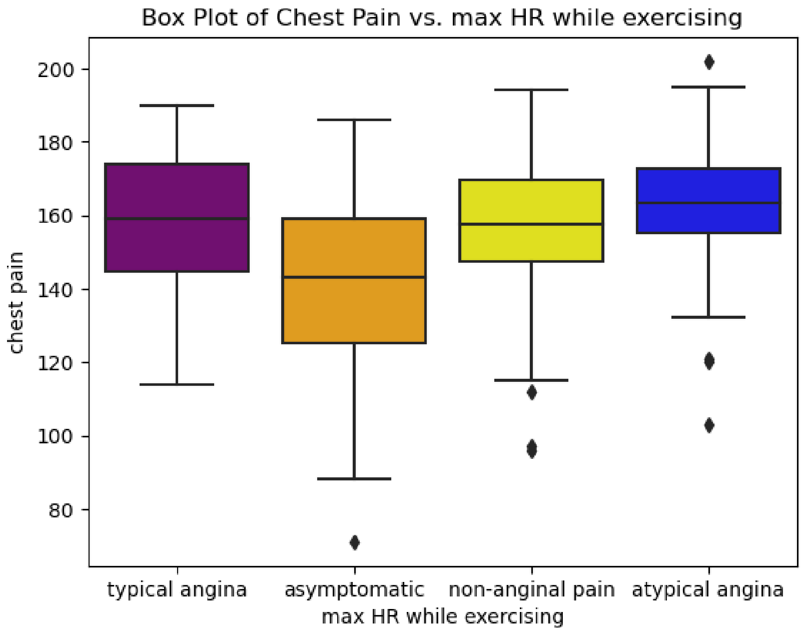
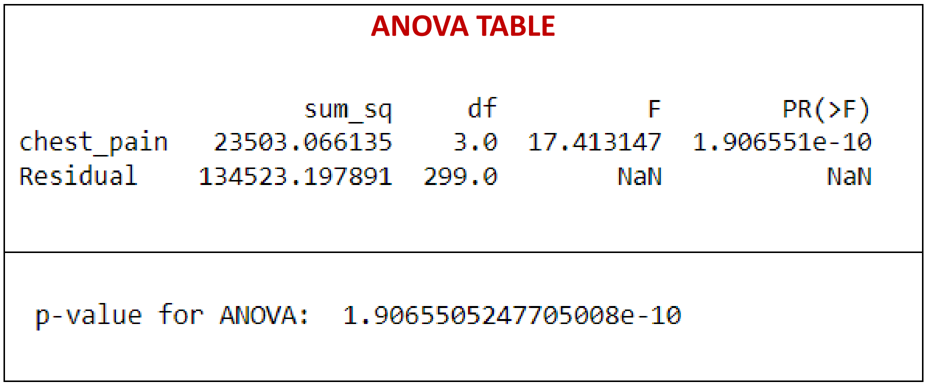


*Figure 8: Box plot illustrating the significance between different types of chest pains and max HR*

The box plot shows that people with non-anginal and atypical angina types of chest pain achieve higher median maximum HR during exercise compared to those with typical angina or asymptomatic chest pain. The IQR varies across the different chest pain categories, with the asymptomatic group having more variability and the atypical angina group having less. Outliers are present in all categories, indicating individuals with significantly different maximum HRs. The typical angina group has lower overall maximum HR values, indicating more limited exercise tolerance. People with non-anginal and atypical angina types of chest pain tend to have higher maximum HRs on average, implying a better exercise tolerance or less severe heart disease.

**Null Hypothesis**: all types of chest pain have the same average effect on the maximum heart rate.
**Alternative Hypothesis**: all types of chest pain do not have the same average effect on the maximum heart rate.



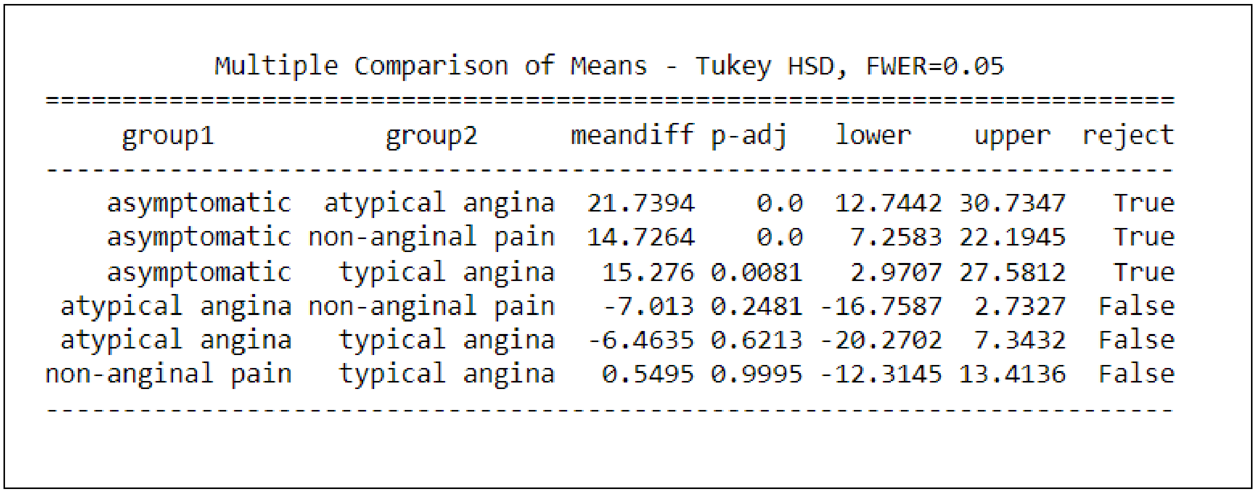
ANOVA TABLE

```
                   sum_sq     df          F        PR(>F)
chest_pain   23503.066135    3.0  17.413147  1.906551e-10
Residual    134523.197891  299.0        NaN           NaN
```

```
p-value for ANOVA:  1.9065505247705008e-10
```

The chest pain sum of squares is about 23503.07, which explains the variability due to different types of chest pain. The residual sum of squares is about 134523.20, representing the variability not explained by chest pain types. The F-statistic value of 17.413 indicates a significant difference in mean maximum heart rate during exercise between at least two types of chest pain. The P-value indicates strong evidence against the null hypothesis since it is well below 0.5. thus, we reject the Null and conclude that the type of chest pain has a statistically significant effect on maximum heart rate achieved during exercise.

To continue, we perform a pst-hoc test to determine which chest pain means are significantly different from each other.



```
           Multiple Comparison of Means - Tukey HSD, FWER=0.05
=========================================================================
     group1          group2      meandiff p-adj   lower    upper   reject
-------------------------------------------------------------------------
    asymptomatic  atypical angina  21.7394    0.0  12.7442 30.7347   True
    asymptomatic non-anginal pain  14.7264    0.0   7.2583 22.1945   True
    asymptomatic   typical angina   15.276 0.0081   2.9707 27.5812   True
 atypical angina non-anginal pain   -7.013 0.2481 -16.7587  2.7327  False
 atypical angina   typical angina  -6.4635 0.6213 -20.2702  7.3432  False
non-anginal pain   typical angina   0.5495 0.9995 -12.3145 13.4136  False
-------------------------------------------------------------------------
```

Columns 'group1' and 'group2' show pairs of chest pain types compared. 'meandiff' is the mean difference in maximum heart rate values during exercise between two groups. 'p-adj' is the p-value adjusted for multiple comparisons, with a p-value below 0.05 indicating a statistically significant difference. 'lower' and 'upper' are the bounds of the 95% confidence interval, and 'reject' indicates if the null hypothesis is rejected, with True indicating a statistically significant difference.

From the table, we can conclude that there is a significant difference in the maximum heart rate during exercise between the asymptomatic group and both the atypical angina and typical angina groups. Similarly, there is a significant difference between the asymptomatic group and the non-anginal pain group. However, there is no significant difference between the atypical angina and non-anginal pain groups, the atypical angina and typical angina groups, as well as the non-anginal pain and typical angina groups.

# Results

The data from the Cleveland Clinic Foundation was analyzed to find out which factors were associated with heart disease. The results showed that people with heart disease had different maximum heart rates during exercise, ages, and resting blood pressures compared to people without heart disease. Chest pain types also had an impact on maximum heart rate during exercise. The findings provide a better understanding of the factors related to heart disease and how they are connected to the condition.

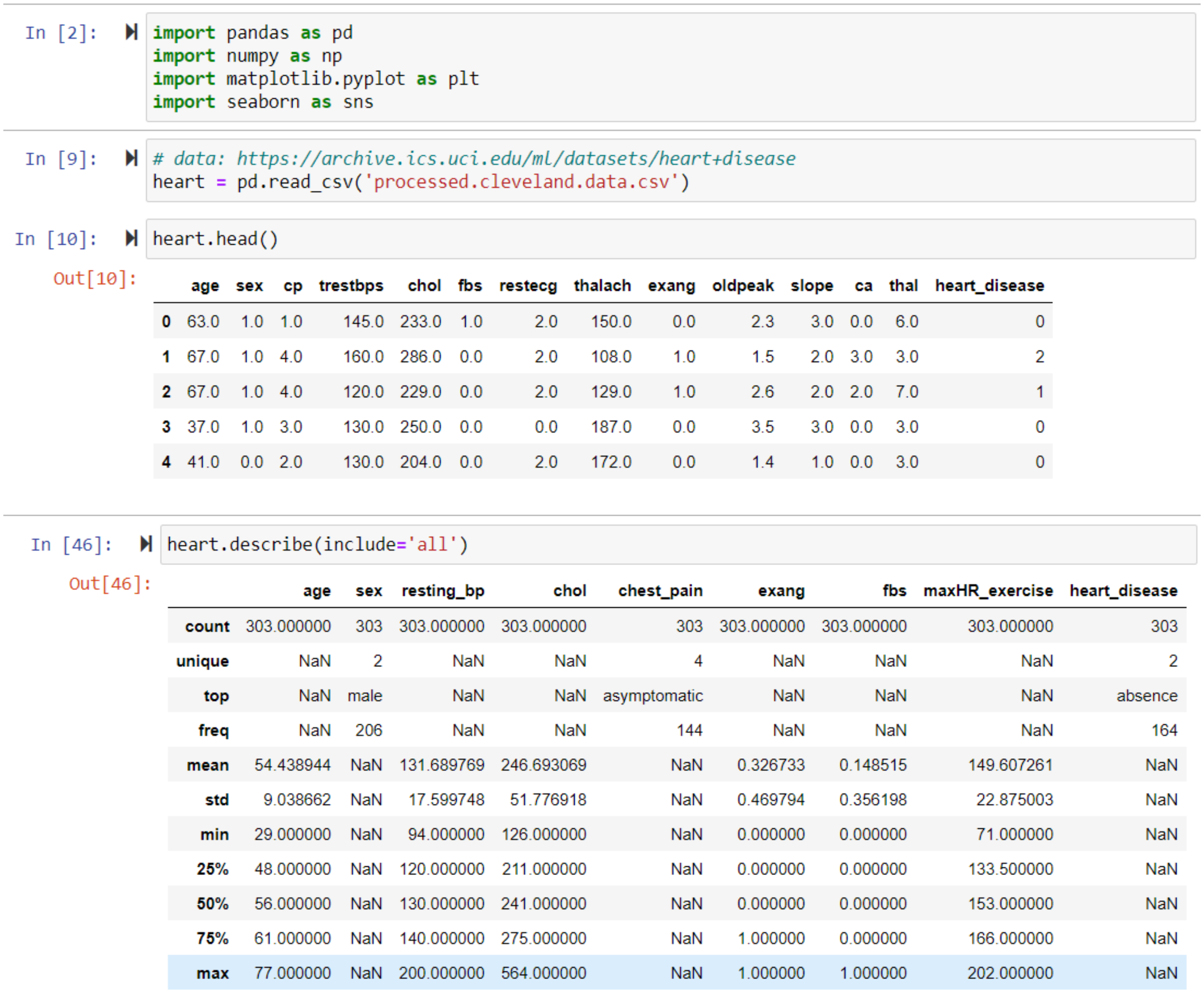
# Conclusion

The study shows that heart disease is associated with lower maximum heart rates during exercise, higher age, and increased resting blood pressure. The type of chest pain experienced by patients also affects their maximum heart rate during exercise, with certain types of pain indicating heart disease more than others. These findings highlight the need for further research into the mechanisms underlying these associations and the development of targeted interventions for better heart health.

# Python Code:

All Jupyter notebooks containing pre-processing and data analysis code, as well as the processed dataset, will be submitted with the report.

## Data Processing Code:



```
In [2]:  import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import seaborn as sns
```

```
In [9]:  # data: https://archive.ics.uci.edu/ml/datasets/heart+disease
         heart = pd.read_csv('processed.cleveland.data.csv')
```

```
In [10]: heart.head()
```

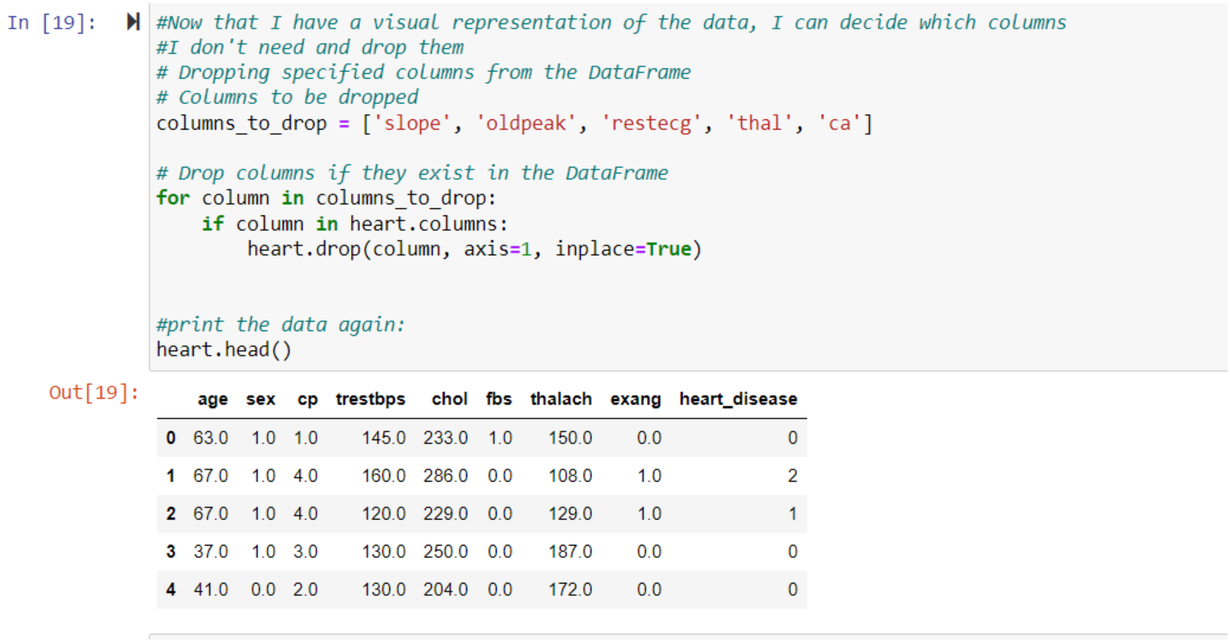
Out[10]:

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | heart_disease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 63.0 | 1.0 | 1.0 | 145.0 | 233.0 | 1.0 | 2.0 | 150.0 | 0.0 | 2.3 | 3.0 | 0.0 | 6.0 | 0 |
| **1** | 67.0 | 1.0 | 4.0 | 160.0 | 286.0 | 0.0 | 2.0 | 108.0 | 1.0 | 1.5 | 2.0 | 3.0 | 3.0 | 2 |
| **2** | 67.0 | 1.0 | 4.0 | 120.0 | 229.0 | 0.0 | 2.0 | 129.0 | 1.0 | 2.6 | 2.0 | 2.0 | 7.0 | 1 |
| **3** | 37.0 | 1.0 | 3.0 | 130.0 | 250.0 | 0.0 | 0.0 | 187.0 | 0.0 | 3.5 | 3.0 | 0.0 | 3.0 | 0 |
| **4** | 41.0 | 0.0 | 2.0 | 130.0 | 204.0 | 0.0 | 2.0 | 172.0 | 0.0 | 1.4 | 1.0 | 0.0 | 3.0 | 0 |

```
In [46]: heart.describe(include='all')
```

Out[46]:

| | age | sex | resting_bp | chol | chest_pain | exang | fbs | maxHR_exercise | heart_disease |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 303.000000 | 303 | 303.000000 | 303.000000 | 303 | 303.000000 | 303.000000 | 303.000000 | 303 |
| **unique** | NaN | 2 | NaN | NaN | 4 | NaN | NaN | NaN | 2 |
| **top** | NaN | male | NaN | NaN | asymptomatic | NaN | NaN | NaN | absence |
| **freq** | NaN | 206 | NaN | NaN | 144 | NaN | NaN | NaN | 164 |
| **mean** | 54.438944 | NaN | 131.689769 | 246.693069 | NaN | 0.326733 | 0.148515 | 149.607261 | NaN |
| **std** | 9.038662 | NaN | 17.599748 | 51.776918 | NaN | 0.469794 | 0.356198 | 22.875003 | NaN |
| **min** | 29.000000 | NaN | 94.000000 | 126.000000 | NaN | 0.000000 | 0.000000 | 71.000000 | NaN |
| **25%** | 48.000000 | NaN | 120.000000 | 211.000000 | NaN | 0.000000 | 0.000000 | 133.500000 | NaN |
| **50%** | 56.000000 | NaN | 130.000000 | 241.000000 | NaN | 0.000000 | 0.000000 | 153.000000 | NaN |
| **75%** | 61.000000 | NaN | 140.000000 | 275.000000 | NaN | 1.000000 | 0.000000 | 166.000000 | NaN |
| **max** | 77.000000 | NaN | 200.000000 | 564.000000 | NaN | 1.000000 | 1.000000 | 202.000000 | NaN |

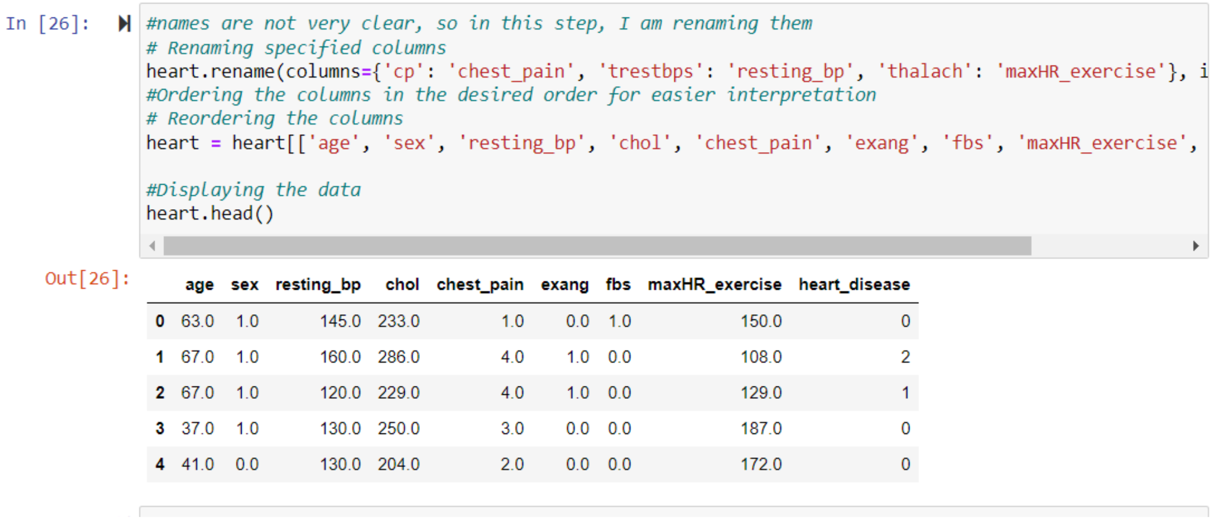


In [19]:

```
#Now that I have a visual representation of the data, I can decide which columns
#I don't need and drop them
# Dropping specified columns from the DataFrame
# Columns to be dropped
columns_to_drop = ['slope', 'oldpeak', 'restecg', 'thal', 'ca']

# Drop columns if they exist in the DataFrame
for column in columns_to_drop:
    if column in heart.columns:
        heart.drop(column, axis=1, inplace=True)


#print the data again:
heart.head()
```

Out[19]:

```
    age  sex   cp  trestbps   chol  fbs  thalach  exang  heart_disease
0  63.0  1.0  1.0     145.0  233.0  1.0    150.0    0.0              0
1  67.0  1.0  4.0     160.0  286.0  0.0    108.0    1.0              2
2  67.0  1.0  4.0     120.0  229.0  0.0    129.0    1.0              1
3  37.0  1.0  3.0     130.0  250.0  0.0    187.0    0.0              0
4  41.0  0.0  2.0     130.0  204.0  0.0    172.0    0.0              0
```

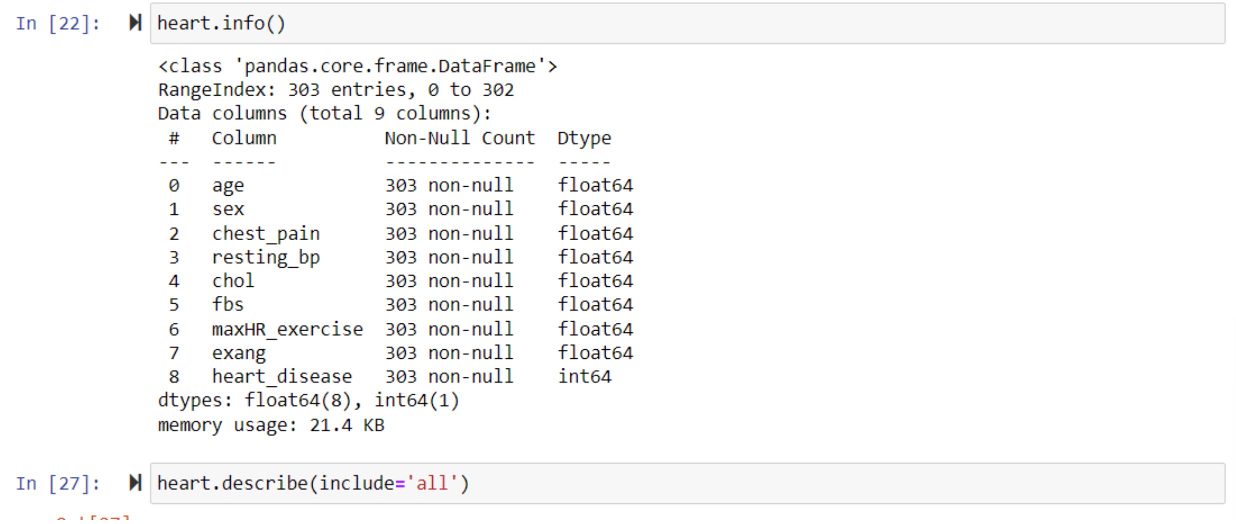


In [26]:

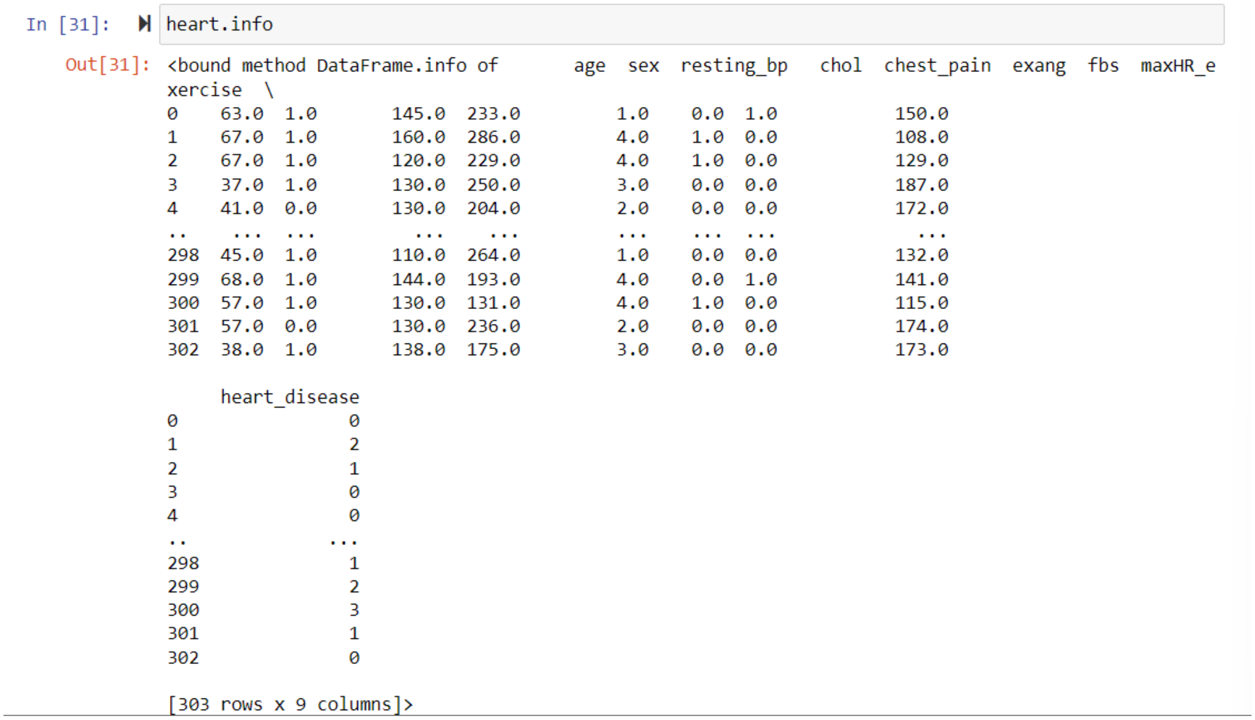
```
#names are not very clear, so in this step, I am renaming them
# Renaming specified columns
heart.rename(columns={'cp': 'chest_pain', 'trestbps': 'resting_bp', 'thalach': 'maxHR_exercise'}, i
#Ordering the columns in the desired order for easier interpretation
# Reordering the columns
heart = heart[['age', 'sex', 'resting_bp', 'chol', 'chest_pain', 'exang', 'fbs', 'maxHR_exercise',

#Displaying the data
heart.head()
```

Out[26]:

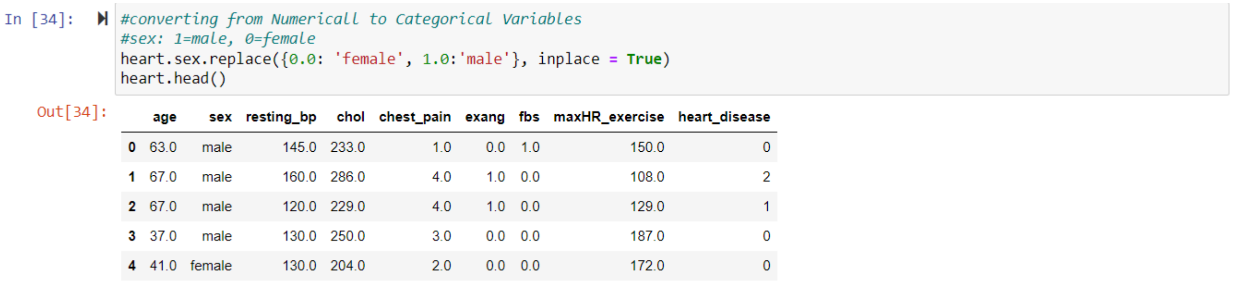
```
    age  sex  resting_bp   chol  chest_pain  exang  fbs  maxHR_exercise  heart_disease
0  63.0  1.0       145.0  233.0         1.0    0.0  1.0           150.0              0
1  67.0  1.0       160.0  286.0         4.0    1.0  0.0           108.0              2
2  67.0  1.0       120.0  229.0         4.0    1.0  0.0           129.0              1
3  37.0  1.0       130.0  250.0         3.0    0.0  0.0           187.0              0
4  41.0  0.0       130.0  204.0         2.0    0.0  0.0           172.0              0
```



```
In [22]:  heart.info()

          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 303 entries, 0 to 302
          Data columns (total 9 columns):
           #   Column          Non-Null Count  Dtype  
          ---  ------          --------------  -----  
           0   age             303 non-null    float64
           1   sex             303 non-null    float64
           2   chest_pain      303 non-null    float64
           3   resting_bp      303 non-null    float64
           4   chol            303 non-null    float64
           5   fbs             303 non-null    float64
           6   maxHR_exercise  303 non-null    float64
           7   exang           303 non-null    float64
           8   heart_disease   303 non-null    int64  
          dtypes: float64(8), int64(1)
          memory usage: 21.4 KB

In [27]:  heart.describe(include='all')
```



```
In [31]:  heart.info

Out[31]:  <bound method DataFrame.info of        age  sex  resting_bp   chol  chest_pain  exang  fbs  maxHR_e
          xercise  \
          0    63.0  1.0       145.0  233.0         1.0    0.0  1.0           150.0   
          1    67.0  1.0       160.0  286.0         4.0    1.0  0.0           108.0   
          2    67.0  1.0       120.0  229.0         4.0    1.0  0.0           129.0   
          3    37.0  1.0       130.0  250.0         3.0    0.0  0.0           187.0   
          4    41.0  0.0       130.0  204.0         2.0    0.0  0.0           172.0   
          ..    ...  ...         ...    ...         ...    ...  ...             ...   
          298  45.0  1.0       110.0  264.0         1.0    0.0  0.0           132.0   
          299  68.0  1.0       144.0  193.0         4.0    0.0  1.0           141.0   
          300  57.0  1.0       130.0  131.0         4.0    1.0  0.0           115.0   
          301  57.0  0.0       130.0  236.0         2.0    0.0  0.0           174.0   
          302  38.0  1.0       138.0  175.0         3.0    0.0  0.0           173.0   

               heart_disease  
          0                0  
          1                2  
          2                1  
          3                0  
          4                0  
          ..             ...  
          298              1  
          299              2  
          300              3  
          301              1  
          302              0  

          [303 rows x 9 columns]>
```

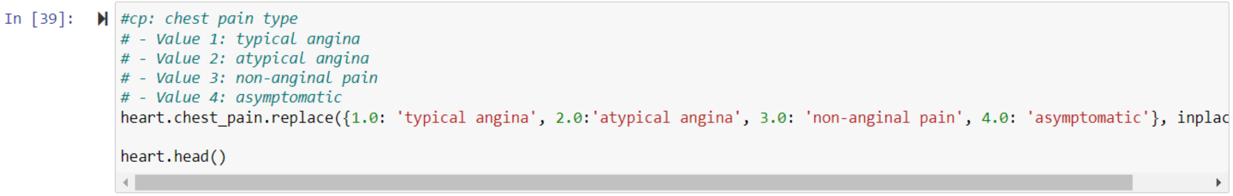


In [34]:

```
#converting from Numericall to Categorical Variables
#sex: 1=male, 0=female
heart.sex.replace({0.0: 'female', 1.0:'male'}, inplace = True)
heart.head()
```

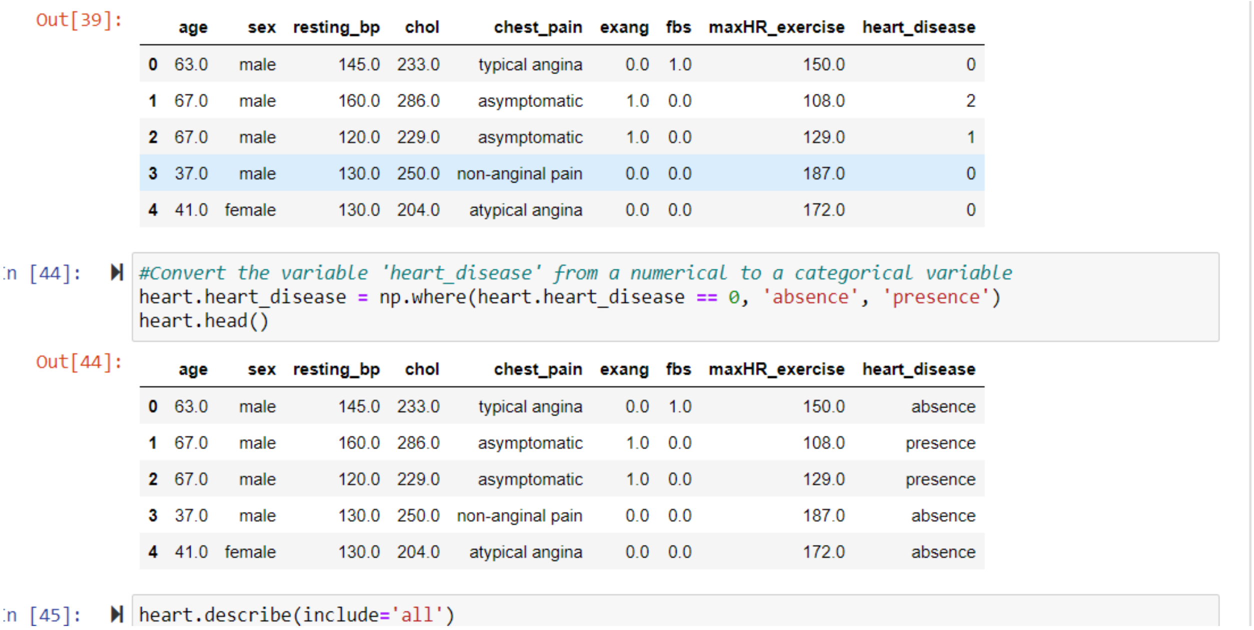
Out[34]:

|  | age | sex | resting_bp | chol | chest_pain | exang | fbs | maxHR_exercise | heart_disease |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 63.0 | male | 145.0 | 233.0 | 1.0 | 0.0 | 1.0 | 150.0 | 0 |
| 1 | 67.0 | male | 160.0 | 286.0 | 4.0 | 1.0 | 0.0 | 108.0 | 2 |
| 2 | 67.0 | male | 120.0 | 229.0 | 4.0 | 1.0 | 0.0 | 129.0 | 1 |
| 3 | 37.0 | male | 130.0 | 250.0 | 3.0 | 0.0 | 0.0 | 187.0 | 0 |
| 4 | 41.0 | female | 130.0 | 204.0 | 2.0 | 0.0 | 0.0 | 172.0 | 0 |



In [39]:

```
#cp: chest pain type
# - Value 1: typical angina
# - Value 2: atypical angina
# - Value 3: non-anginal pain
# - Value 4: asymptomatic
heart.chest_pain.replace({1.0: 'typical angina', 2.0:'atypical angina', 3.0: 'non-anginal pain', 4.0: 'asymptomatic'}, inplac

heart.head()
```



Out[39]:

|  | age | sex | resting_bp | chol | chest_pain | exang | fbs | maxHR_exercise | heart_disease |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 63.0 | male | 145.0 | 233.0 | typical angina | 0.0 | 1.0 | 150.0 | 0 |
| 1 | 67.0 | male | 160.0 | 286.0 | asymptomatic | 1.0 | 0.0 | 108.0 | 2 |
| 2 | 67.0 | male | 120.0 | 229.0 | asymptomatic | 1.0 | 0.0 | 129.0 | 1 |
| 3 | 37.0 | male | 130.0 | 250.0 | non-anginal pain | 0.0 | 0.0 | 187.0 | 0 |
| 4 | 41.0 | female | 130.0 | 204.0 | atypical angina | 0.0 | 0.0 | 172.0 | 0 |

:n [44]:

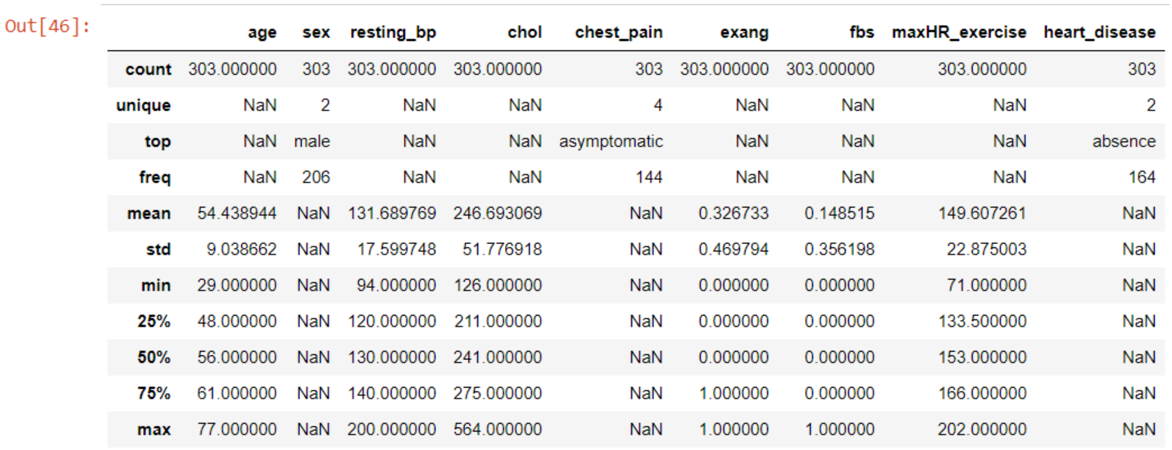
```
#Convert the variable 'heart_disease' from a numerical to a categorical variable
heart.heart_disease = np.where(heart.heart_disease == 0, 'absence', 'presence')
heart.head()
```

Out[44]:

|  | age | sex | resting_bp | chol | chest_pain | exang | fbs | maxHR_exercise | heart_disease |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 63.0 | male | 145.0 | 233.0 | typical angina | 0.0 | 1.0 | 150.0 | absence |
| 1 | 67.0 | male | 160.0 | 286.0 | asymptomatic | 1.0 | 0.0 | 108.0 | presence |
| 2 | 67.0 | male | 120.0 | 229.0 | asymptomatic | 1.0 | 0.0 | 129.0 | presence |
| 3 | 37.0 | male | 130.0 | 250.0 | non-anginal pain | 0.0 | 0.0 | 187.0 | absence |
| 4 | 41.0 | female | 130.0 | 204.0 | atypical angina | 0.0 | 0.0 | 172.0 | absence |

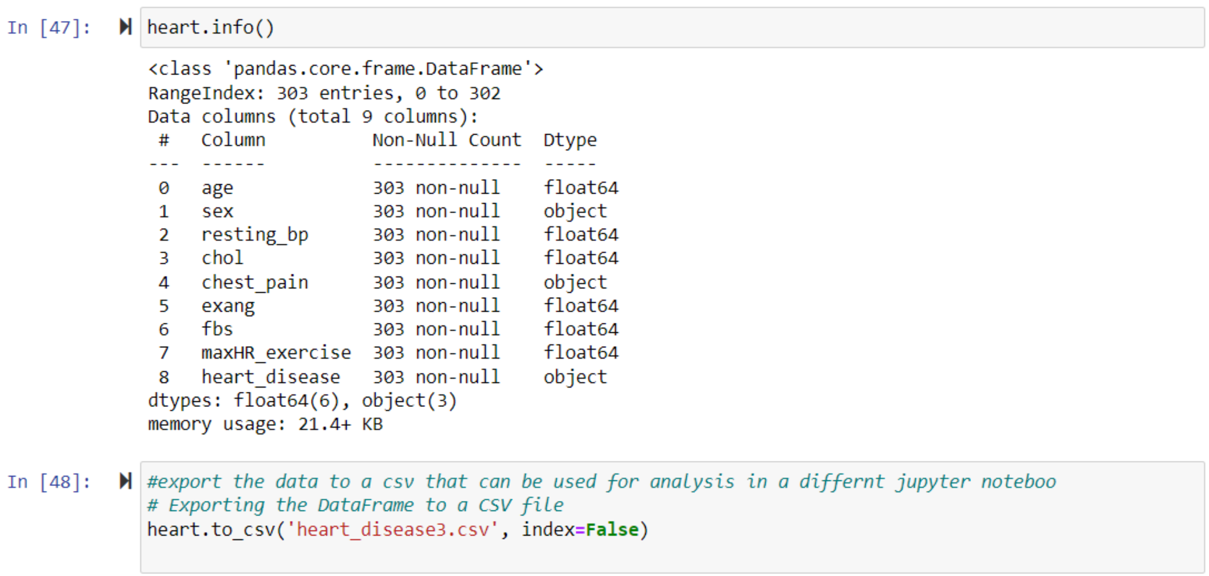
:n [45]:

```
heart.describe(include='all')
```



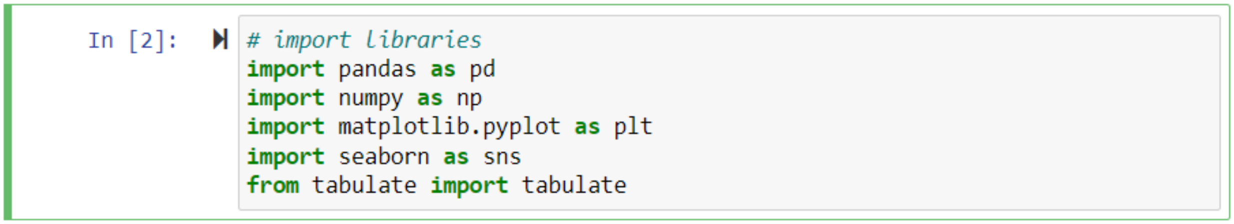
Out[46]:

|  | age | sex | resting_bp | chol | chest_pain | exang | fbs | maxHR_exercise | heart_disease |
|---|---|---|---|---|---|---|---|---|---|
| count | 303.000000 | 303 | 303.000000 | 303.000000 | 303 | 303.000000 | 303.000000 | 303.000000 | 303 |
| unique | NaN | 2 | NaN | NaN | 4 | NaN | NaN | NaN | 2 |
| top | NaN | male | NaN | NaN | asymptomatic | NaN | NaN | NaN | absence |
| freq | NaN | 206 | NaN | NaN | 144 | NaN | NaN | NaN | 164 |
| mean | 54.438944 | NaN | 131.689769 | 246.693069 | NaN | 0.326733 | 0.148515 | 149.607261 | NaN |
| std | 9.038662 | NaN | 17.599748 | 51.776918 | NaN | 0.469794 | 0.356198 | 22.875003 | NaN |
| min | 29.000000 | NaN | 94.000000 | 126.000000 | NaN | 0.000000 | 0.000000 | 71.000000 | NaN |
| 25% | 48.000000 | NaN | 120.000000 | 211.000000 | NaN | 0.000000 | 0.000000 | 133.500000 | NaN |
| 50% | 56.000000 | NaN | 130.000000 | 241.000000 | NaN | 0.000000 | 0.000000 | 153.000000 | NaN |
| 75% | 61.000000 | NaN | 140.000000 | 275.000000 | NaN | 1.000000 | 0.000000 | 166.000000 | NaN |
| max | 77.000000 | NaN | 200.000000 | 564.000000 | NaN | 1.000000 | 1.000000 | 202.000000 | NaN |

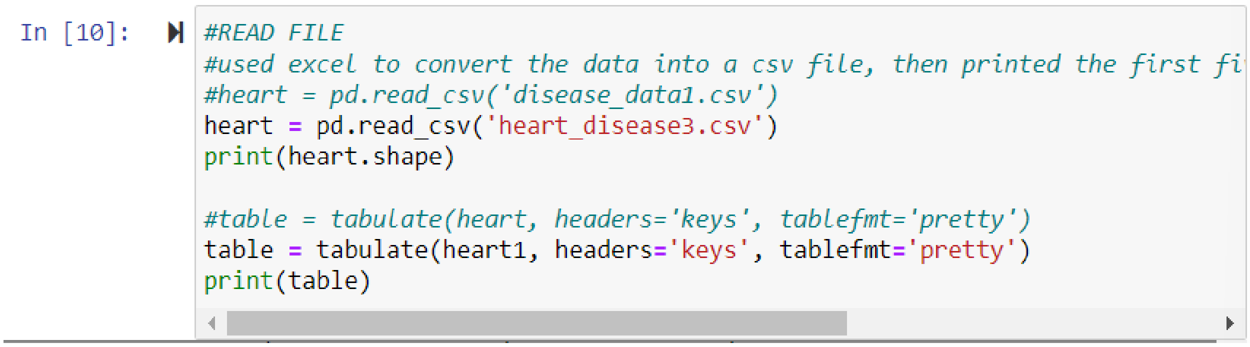


```
In [47]:  heart.info()

          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 303 entries, 0 to 302
          Data columns (total 9 columns):
           #   Column          Non-Null Count  Dtype
          ---  ------          --------------  -----
           0   age             303 non-null    float64
           1   sex             303 non-null    object
           2   resting_bp      303 non-null    float64
           3   chol            303 non-null    float64
           4   chest_pain      303 non-null    object
           5   exang           303 non-null    float64
           6   fbs             303 non-null    float64
           7   maxHR_exercise  303 non-null    float64
           8   heart_disease   303 non-null    object
          dtypes: float64(6), object(3)
          memory usage: 21.4+ KB

In [48]:  #export the data to a csv that can be used for analysis in a differnt jupyter noteboo
          # Exporting the DataFrame to a CSV file
          heart.to_csv('heart_disease3.csv', index=False)
```

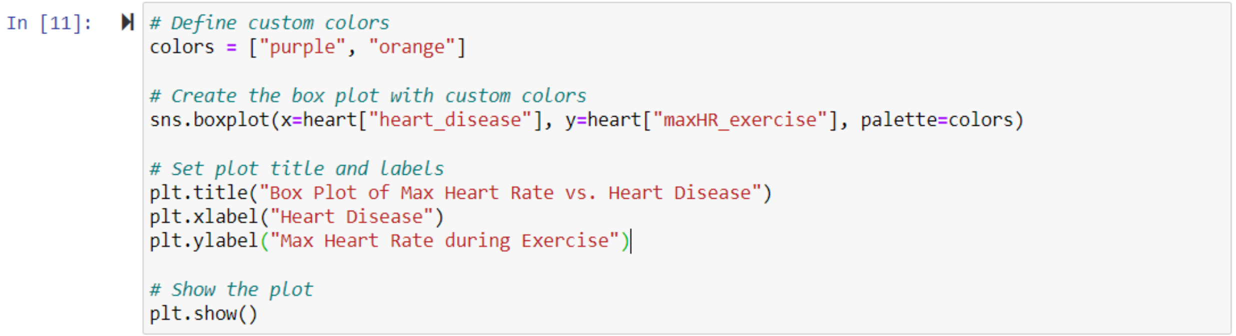
## Statistical Analysis Code:



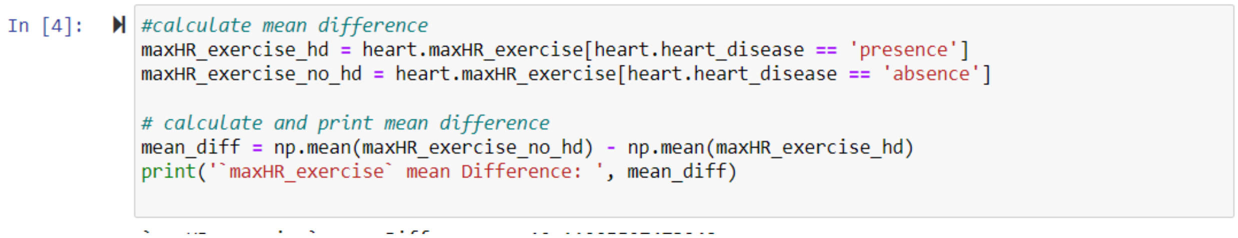
```
In [2]:  # import libraries
         import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import seaborn as sns
         from tabulate import tabulate
```



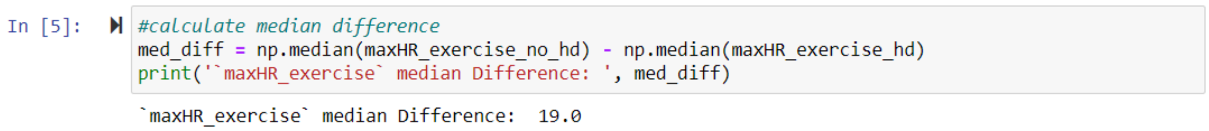
```
In [10]:  #READ FILE
          #used excel to convert the data into a csv file, then printed the first fi
          #heart = pd.read_csv('disease_data1.csv')
          heart = pd.read_csv('heart_disease3.csv')
          print(heart.shape)

          #table = tabulate(heart, headers='keys', tablefmt='pretty')
          table = tabulate(heart1, headers='keys', tablefmt='pretty')
          print(table)
```



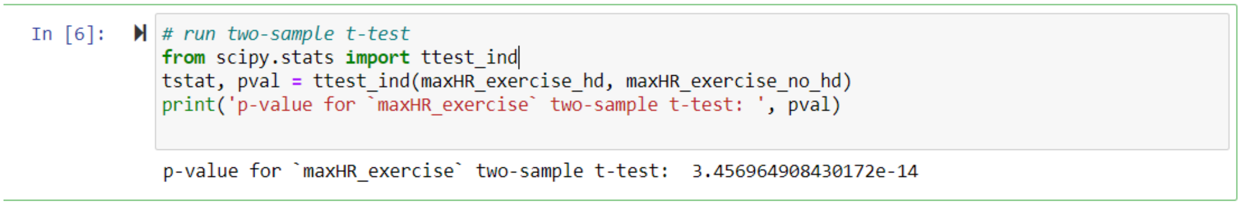
```
In [11]:  # Define custom colors
          colors = ["purple", "orange"]

          # Create the box plot with custom colors
          sns.boxplot(x=heart["heart_disease"], y=heart["maxHR_exercise"], palette=colors)

          # Set plot title and labels
          plt.title("Box Plot of Max Heart Rate vs. Heart Disease")
          plt.xlabel("Heart Disease")
          plt.ylabel("Max Heart Rate during Exercise")

          # Show the plot
          plt.show()
```



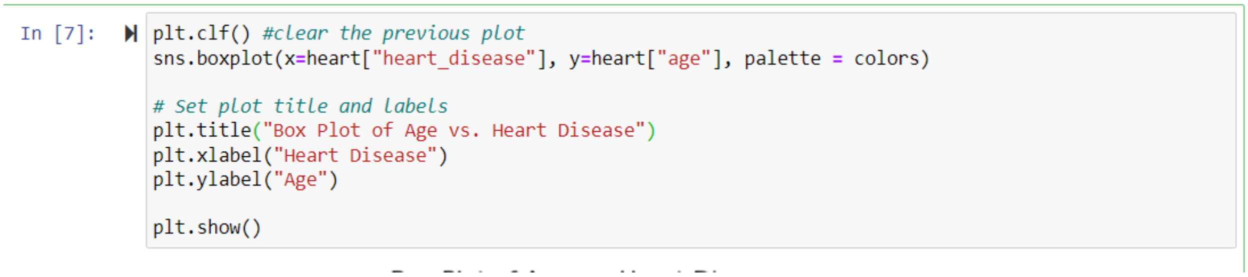
```
In [4]:  #calculate mean difference
         maxHR_exercise_hd = heart.maxHR_exercise[heart.heart_disease == 'presence']
         maxHR_exercise_no_hd = heart.maxHR_exercise[heart.heart_disease == 'absence']

         # calculate and print mean difference
         mean_diff = np.mean(maxHR_exercise_no_hd) - np.mean(maxHR_exercise_hd)
         print('`maxHR_exercise` mean Difference: ', mean_diff)
```



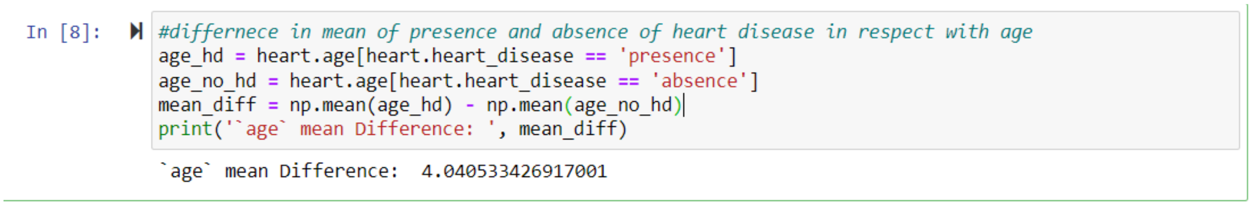
```
In [5]:  #calculate median difference
         med_diff = np.median(maxHR_exercise_no_hd) - np.median(maxHR_exercise_hd)
         print('`maxHR_exercise` median Difference: ', med_diff)

         `maxHR_exercise` median Difference:  19.0
```



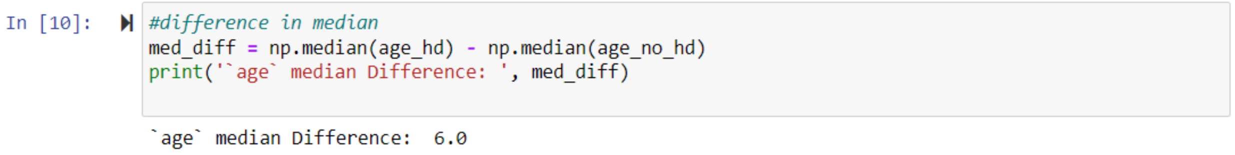
```
In [6]:  # run two-sample t-test
         from scipy.stats import ttest_ind
         tstat, pval = ttest_ind(maxHR_exercise_hd, maxHR_exercise_no_hd)
         print('p-value for `maxHR_exercise` two-sample t-test: ', pval)

         p-value for `maxHR_exercise` two-sample t-test:  3.456964908430172e-14
```



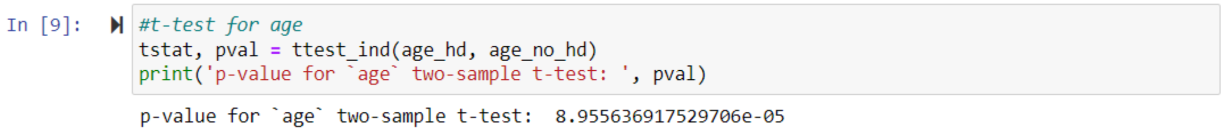
```
In [7]:  plt.clf() #clear the previous plot
         sns.boxplot(x=heart["heart_disease"], y=heart["age"], palette = colors)

         # Set plot title and labels
         plt.title("Box Plot of Age vs. Heart Disease")
         plt.xlabel("Heart Disease")
         plt.ylabel("Age")

         plt.show()
```



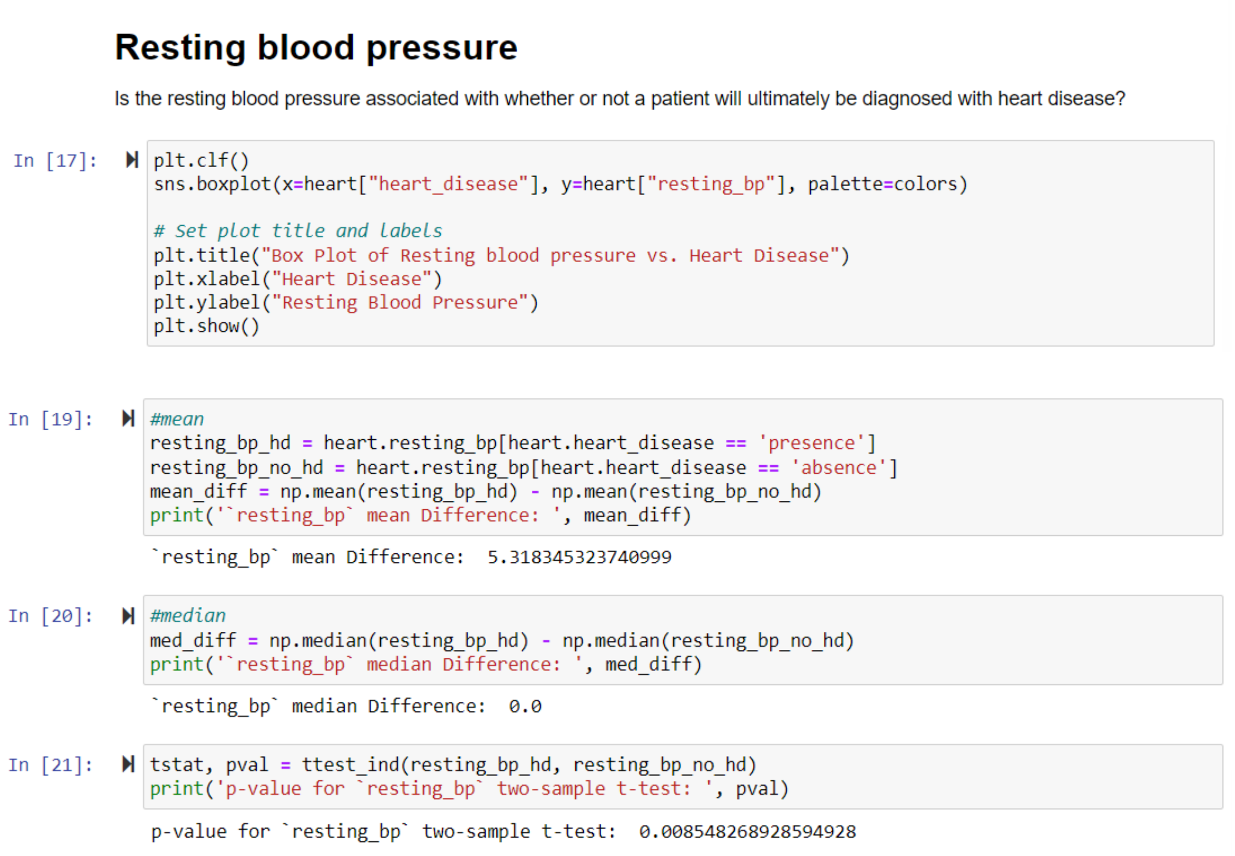
```
In [8]:  #differnece in mean of presence and absence of heart disease in respect with age
         age_hd = heart.age[heart.heart_disease == 'presence']
         age_no_hd = heart.age[heart.heart_disease == 'absence']
         mean_diff = np.mean(age_hd) - np.mean(age_no_hd)
         print('`age` mean Difference: ', mean_diff)

         `age` mean Difference:  4.040533426917001
```



```
In [10]: #difference in median
         med_diff = np.median(age_hd) - np.median(age_no_hd)
         print('`age` median Difference: ', med_diff)

         `age` median Difference:  6.0
```



```
In [9]:  #t-test for age
         tstat, pval = ttest_ind(age_hd, age_no_hd)
         print('p-value for `age` two-sample t-test: ', pval)

         p-value for `age` two-sample t-test:  8.955636917529706e-05
```



## Resting blood pressure

Is the resting blood pressure associated with whether or not a patient will ultimately be diagnosed with heart disease?

```
In [17]:
plt.clf()
sns.boxplot(x=heart["heart_disease"], y=heart["resting_bp"], palette=colors)

# Set plot title and labels
plt.title("Box Plot of Resting blood pressure vs. Heart Disease")
plt.xlabel("Heart Disease")
plt.ylabel("Resting Blood Pressure")
plt.show()
```

```
In [19]:
#mean
resting_bp_hd = heart.resting_bp[heart.heart_disease == 'presence']
resting_bp_no_hd = heart.resting_bp[heart.heart_disease == 'absence']
mean_diff = np.mean(resting_bp_hd) - np.mean(resting_bp_no_hd)
print('`resting_bp` mean Difference: ', mean_diff)
```

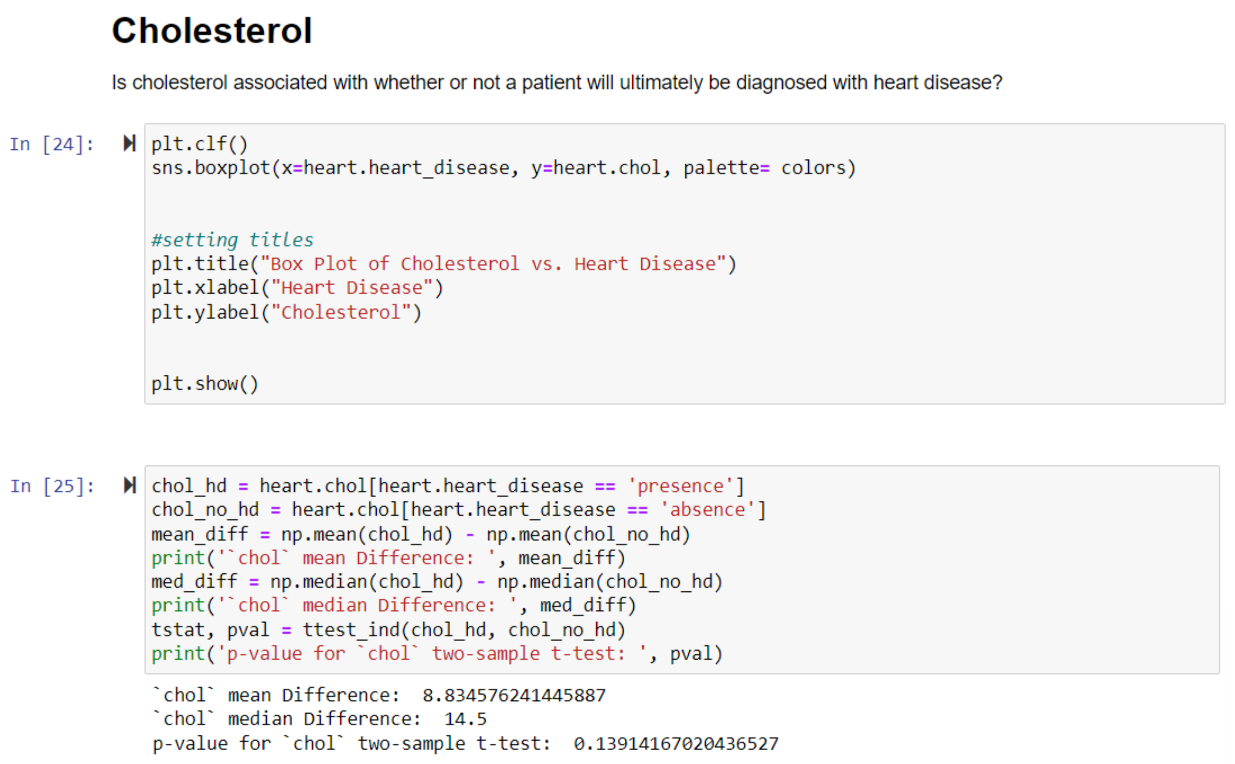
```
`resting_bp` mean Difference:  5.318345323740999
```

```
In [20]:
#median
med_diff = np.median(resting_bp_hd) - np.median(resting_bp_no_hd)
print('`resting_bp` median Difference: ', med_diff)
```

```
`resting_bp` median Difference:  0.0
```

```
In [21]:
tstat, pval = ttest_ind(resting_bp_hd, resting_bp_no_hd)
print('p-value for `resting_bp` two-sample t-test: ', pval)
```

```
p-value for `resting_bp` two-sample t-test:  0.008548268928594928
```



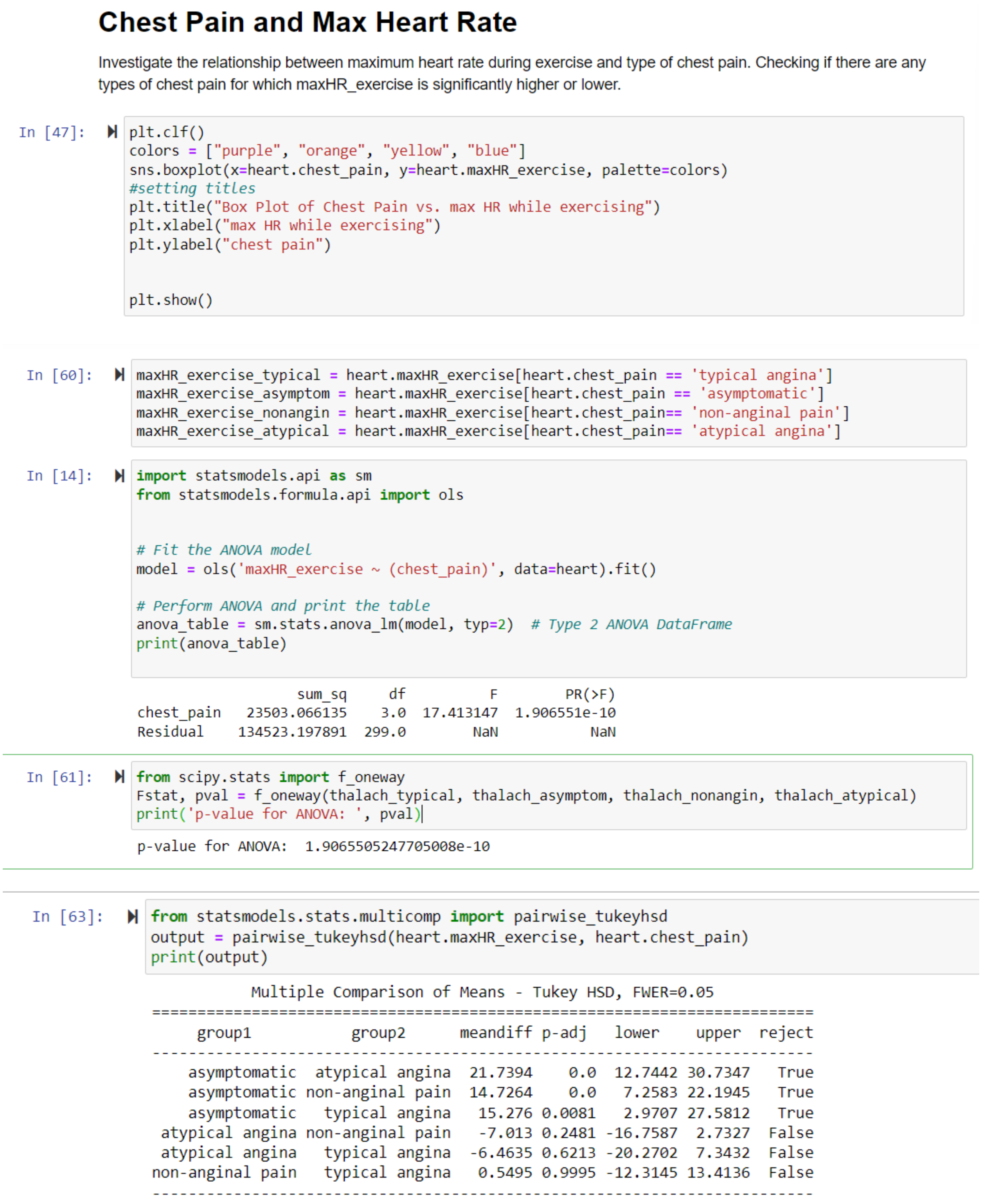
## Cholesterol

Is cholesterol associated with whether or not a patient will ultimately be diagnosed with heart disease?

```
In [24]:
plt.clf()
sns.boxplot(x=heart.heart_disease, y=heart.chol, palette= colors)


#setting titles
plt.title("Box Plot of Cholesterol vs. Heart Disease")
plt.xlabel("Heart Disease")
plt.ylabel("Cholesterol")


plt.show()
```

```
In [25]:
chol_hd = heart.chol[heart.heart_disease == 'presence']
chol_no_hd = heart.chol[heart.heart_disease == 'absence']
mean_diff = np.mean(chol_hd) - np.mean(chol_no_hd)
print('`chol` mean Difference: ', mean_diff)
med_diff = np.median(chol_hd) - np.median(chol_no_hd)
print('`chol` median Difference: ', med_diff)
tstat, pval = ttest_ind(chol_hd, chol_no_hd)
print('p-value for `chol` two-sample t-test: ', pval)
```

```
`chol` mean Difference:  8.834576241445887
`chol` median Difference:  14.5
p-value for `chol` two-sample t-test:  0.13914167020436527
```



## Chest Pain and Max Heart Rate

Investigate the relationship between maximum heart rate during exercise and type of chest pain. Checking if there are any types of chest pain for which maxHR_exercise is significantly higher or lower.

In [47]:

```
plt.clf()
colors = ["purple", "orange", "yellow", "blue"]
sns.boxplot(x=heart.chest_pain, y=heart.maxHR_exercise, palette=colors)
#setting titles
plt.title("Box Plot of Chest Pain vs. max HR while exercising")
plt.xlabel("max HR while exercising")
plt.ylabel("chest pain")


plt.show()
```

In [60]:

```
maxHR_exercise_typical = heart.maxHR_exercise[heart.chest_pain == 'typical angina']
maxHR_exercise_asymptom = heart.maxHR_exercise[heart.chest_pain == 'asymptomatic']
maxHR_exercise_nonangin = heart.maxHR_exercise[heart.chest_pain== 'non-anginal pain']
maxHR_exercise_atypical = heart.maxHR_exercise[heart.chest_pain== 'atypical angina']
```

In [14]:

```
import statsmodels.api as sm
from statsmodels.formula.api import ols


# Fit the ANOVA model
model = ols('maxHR_exercise ~ (chest_pain)', data=heart).fit()

# Perform ANOVA and print the table
anova_table = sm.stats.anova_lm(model, typ=2)  # Type 2 ANOVA DataFrame
print(anova_table)
```

```
                  sum_sq     df          F        PR(>F)
chest_pain   23503.066135    3.0  17.413147  1.906551e-10
Residual    134523.197891  299.0        NaN           NaN
```

In [61]:

```
from scipy.stats import f_oneway
Fstat, pval = f_oneway(thalach_typical, thalach_asymptom, thalach_nonangin, thalach_atypical)
print('p-value for ANOVA: ', pval)
```

```
p-value for ANOVA:  1.9065505247705008e-10
```

In [63]:

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
output = pairwise_tukeyhsd(heart.maxHR_exercise, heart.chest_pain)
print(output)
```

```
           Multiple Comparison of Means - Tukey HSD, FWER=0.05
===============================================================================
     group1            group2      meandiff p-adj   lower    upper  reject
-------------------------------------------------------------------------------
    asymptomatic   atypical angina  21.7394    0.0  12.7442 30.7347   True
    asymptomatic non-anginal pain  14.7264    0.0   7.2583 22.1945   True
    asymptomatic    typical angina   15.276 0.0081   2.9707 27.5812   True
 atypical angina non-anginal pain   -7.013 0.2481 -16.7587  2.7327  False
 atypical angina    typical angina  -6.4635 0.6213 -20.2702  7.3432  False
non-anginal pain    typical angina   0.5495 0.9995 -12.3145 13.4136  False
-------------------------------------------------------------------------------
```