Andrew Napolitano

Tyler Beard

Zach Carlson

DSCI 631: Applied Machine Learning Project Proposal

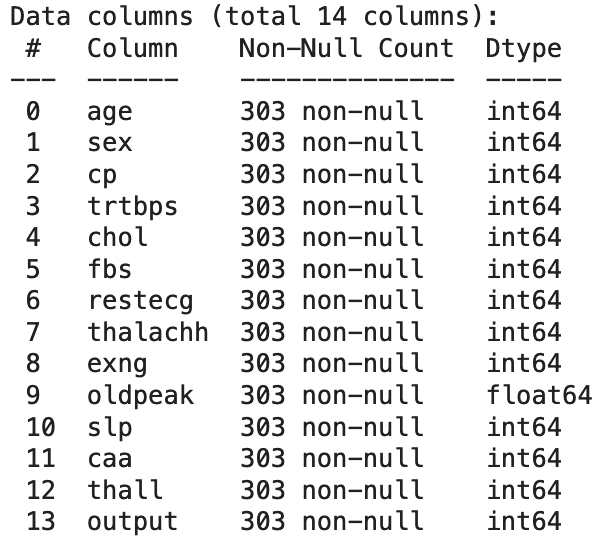
# DATASET DESCRIPTION

The dataset of study is the [Heart Attack Analysis & Prediction Dataset](https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset) made available via Kaggle. The dataset features a series of independent variables (age, sex, exercise induced angina flag, major vessel count, chest pain indicators, resting blood pressure, cholesterol measures, fasting blood sugar, resting electrocardiographic results, and maximum heart rate achieved) and a binary indicator of low or high risk of heart attack. There are 14 columns and 303 observations in this dataset. The dataset is optimally structured for supervised machine learning algorithms since the target outcome variable is provided.

# EXPLORATORY DATA ANALYSIS

## General Feature Analysis

This dataset consists of 12 integer independent variables and 1 floating point independent variable to predict 1 binary outcome (Figure 1).



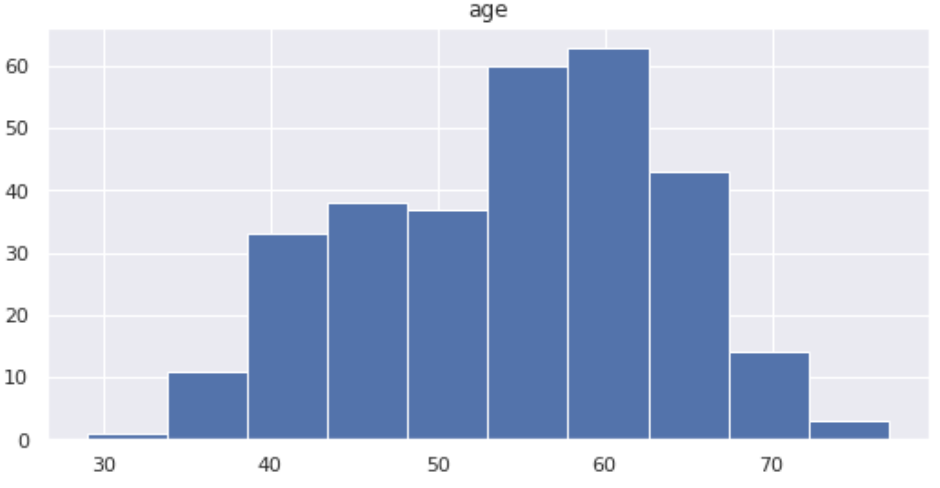
*Figure 1: Column Data Types*

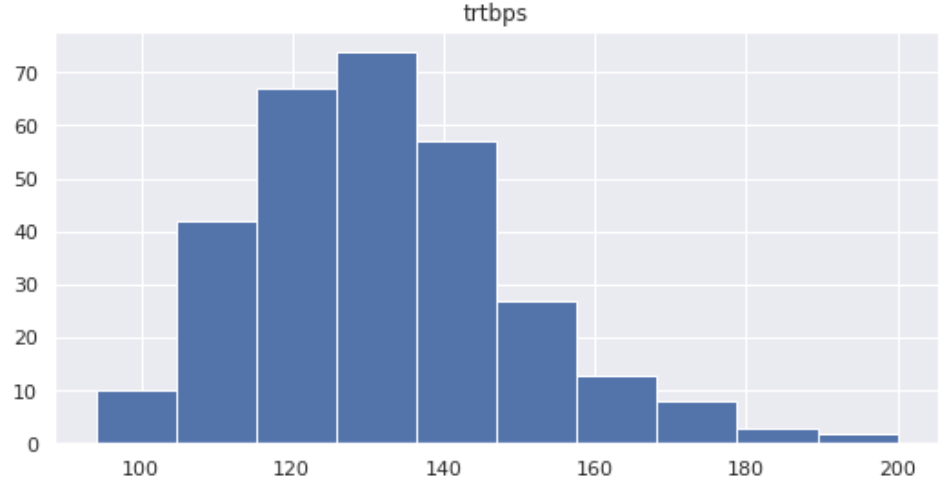
None of the 14 total columns feature any null values (Figure 1, Non-Null Count column).

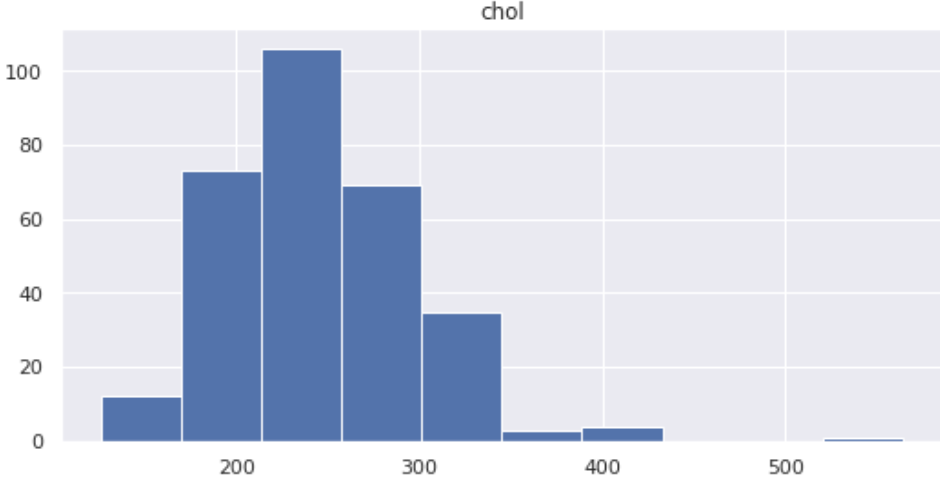
## Continuous Feature Distribution Analysis

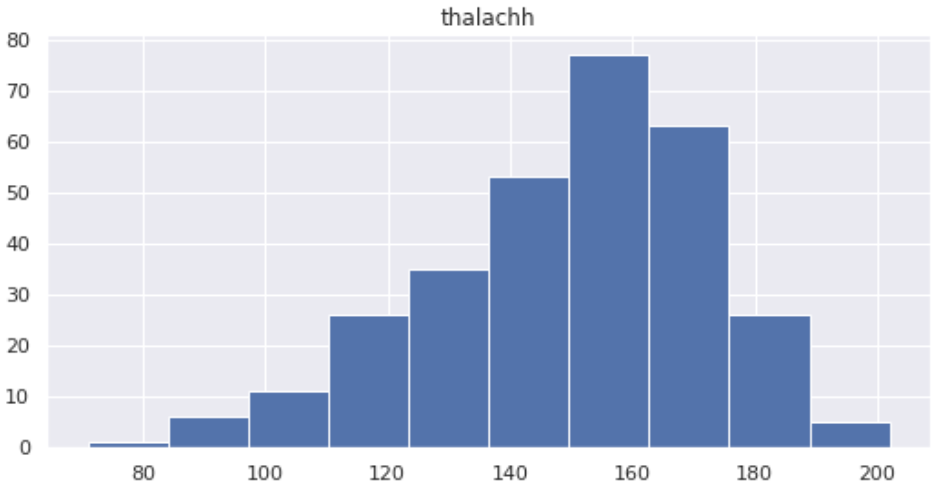
We first plotted the distribution of each continuous feature (Figure 3). Clearly these features have different scales and distributions. None of the features have a perfect Gaussian distribution. Feature scaling will be required later during processing.

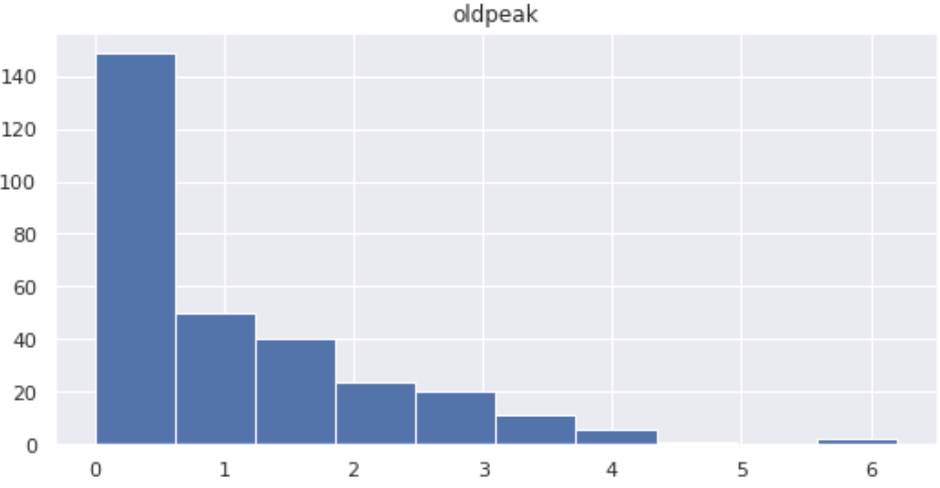
| Feature Name | Feature Description |
| --- | --- |
| age | Age of patient |
| thalachh | Maximum heart rate achieved (BPM) |
| trtbps | Resting blood pressure (in mmHg) |
| oldpeak | Previous peak |
| chol | Cholesterol in mg/dL (via BMI sensor) |











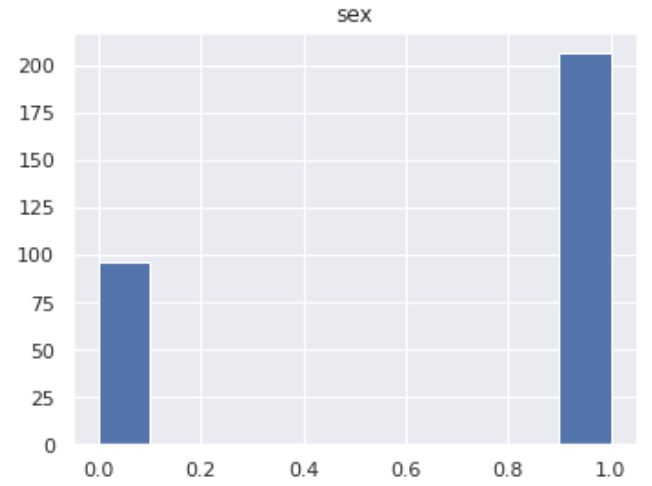
*Figure 3: Continuous Feature Distribution Analysis*

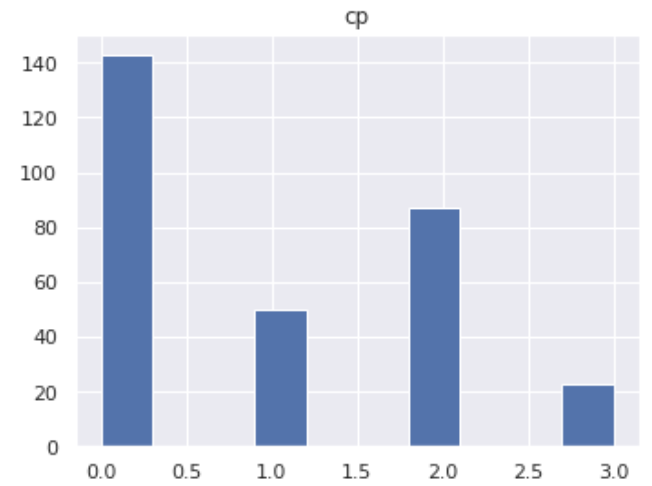
## Categorical Feature Distribution Analysis

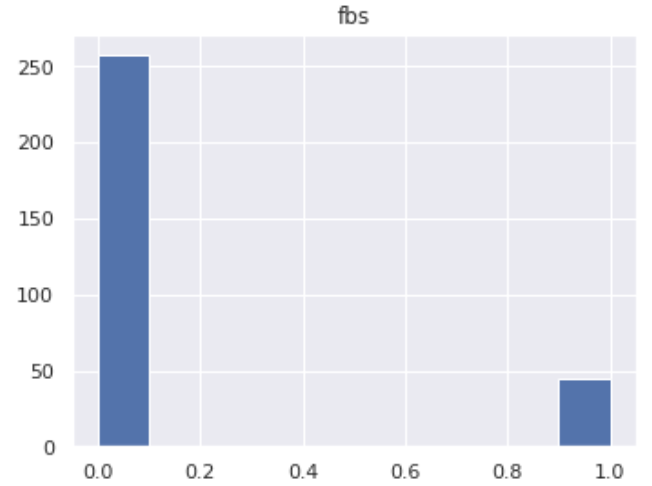
After plotting the continuous features, we plotted the categorical feature distributions (Figure 4). Fortunately, these categorical features were already encoded with numerical values. We will not need to utilize OrdinalEncoder or OneHotEncoder classes.

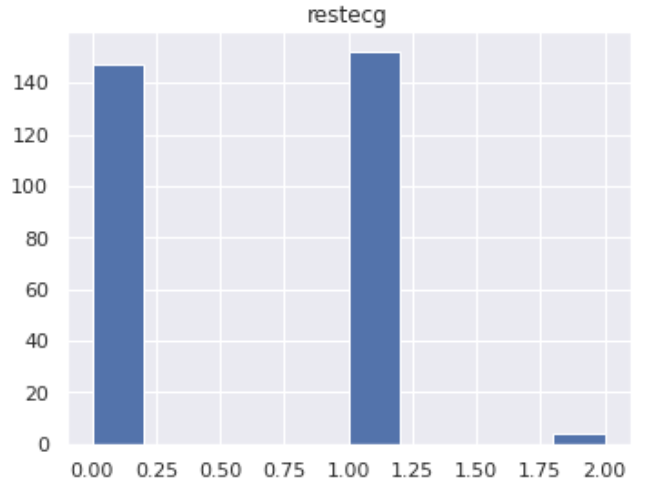
These categories may have an impact on heart attack risk and are not evenly-distributed. In the event they do, we will need to utilize the StratifiedShuffleSplit class. The stratification wouldn’t be necessary if we had a major number of instances. Unfortunately with only 300 instances, random sampling may not be possible and using the StratifiedShuffleSplit class will be required, assuming one of the categorical variables has a large impact on the target feature. The categorical variables are:

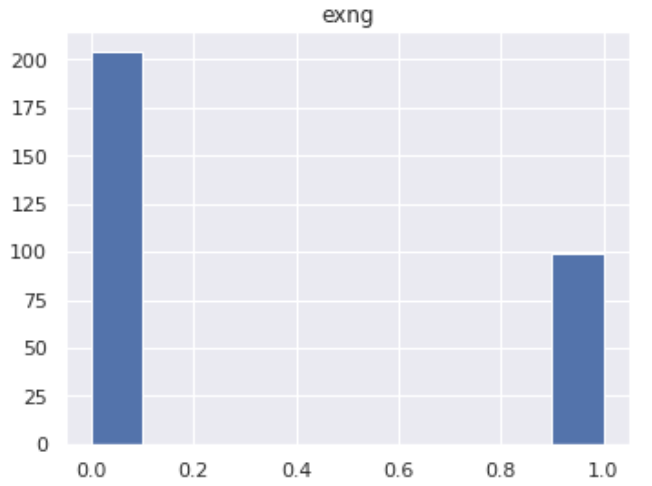
| Feature Name | Feature Description |
| --- | --- |
| sex | Gender of patient |
| cp | Chest pain type (0-3) |
| fbs | Fasting blood sugar > 120 mg/dL (0 = false, 1 = true) |
| slp | Slope |
| restecg | Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy) |
| thall | Thal stress test result (0-3) |
| exng | Exercise induced angina (0 = no, 1 = yes) |
| caa | Number of major blood vessels (0-4) |

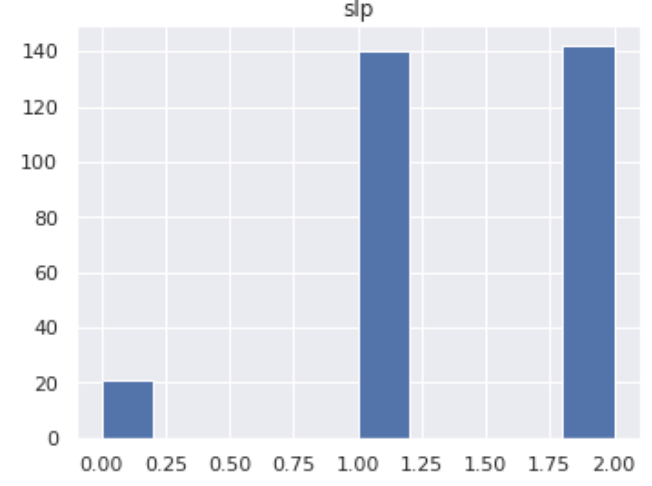
**

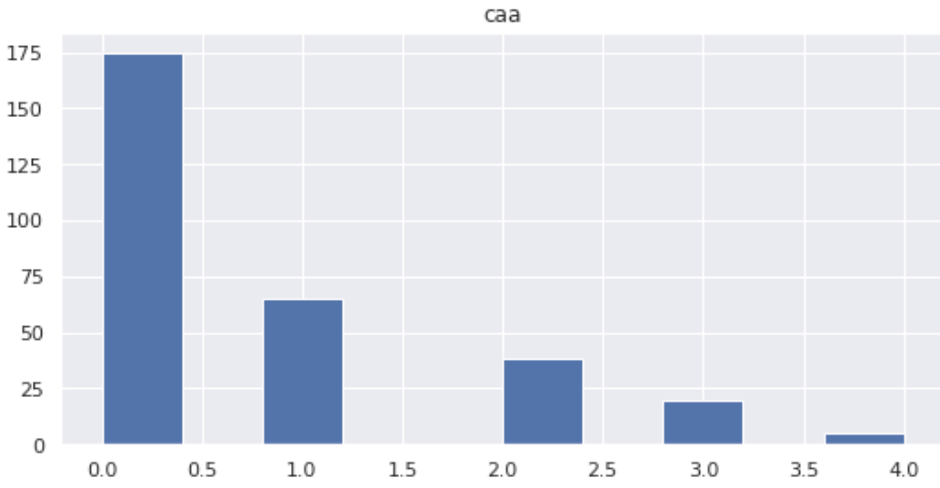
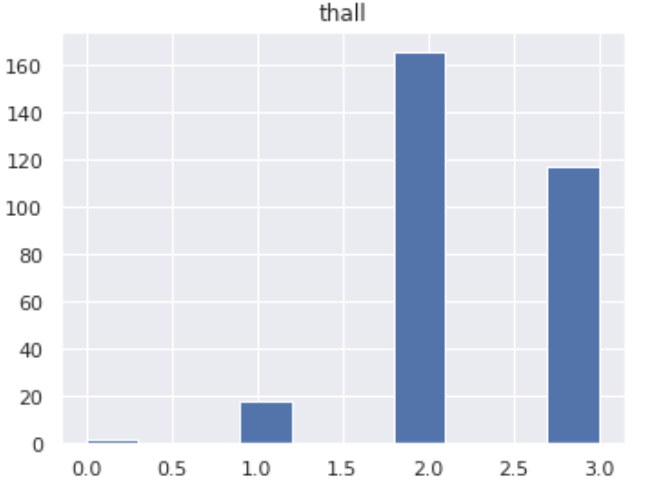
**

**

**

**

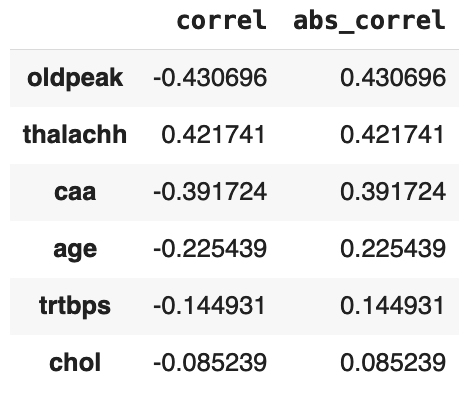
**

**

*Figure 4: Continuous Feature Distribution Analysis*

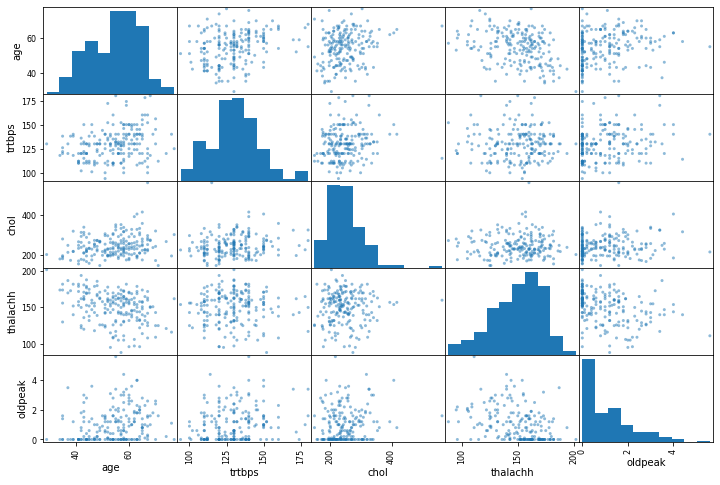
## Correlation Analysis

Correlation analysis was conducted between each of the five continuous features in the dataset and the outcome variable (Figure 7). The Previous Peak Heart Rate (oldpeak) is strongest correlated with higher risk of heart attack, followed by Maximum Heart Rate Achieved (thalachh), Age, Resting Blood Pressure (trtbps), and lastly Cholesterol levels. The Maximum Heart Rate Achieved (thalachh) is the only variable that is positively correlated with higher risk of heart attack. The rest of the variables are negatively correlated.



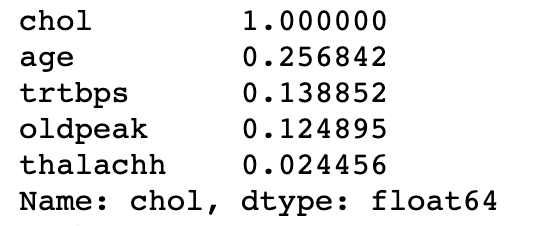
*Figure 7: Continuous Feature Correlation Analysis*

We then used the scatter\_matrix function to see if any linear or polynomial relationships existed between the continuous features (Figure 8). If any are highly correlated, we can utilize feature extraction to combine them and optimize our model.



*Figure 8: scatter\_matrix of Continuous Features*

The most correlated continuous features are age and chol (Figure 9). The correlation is only 0.256842.



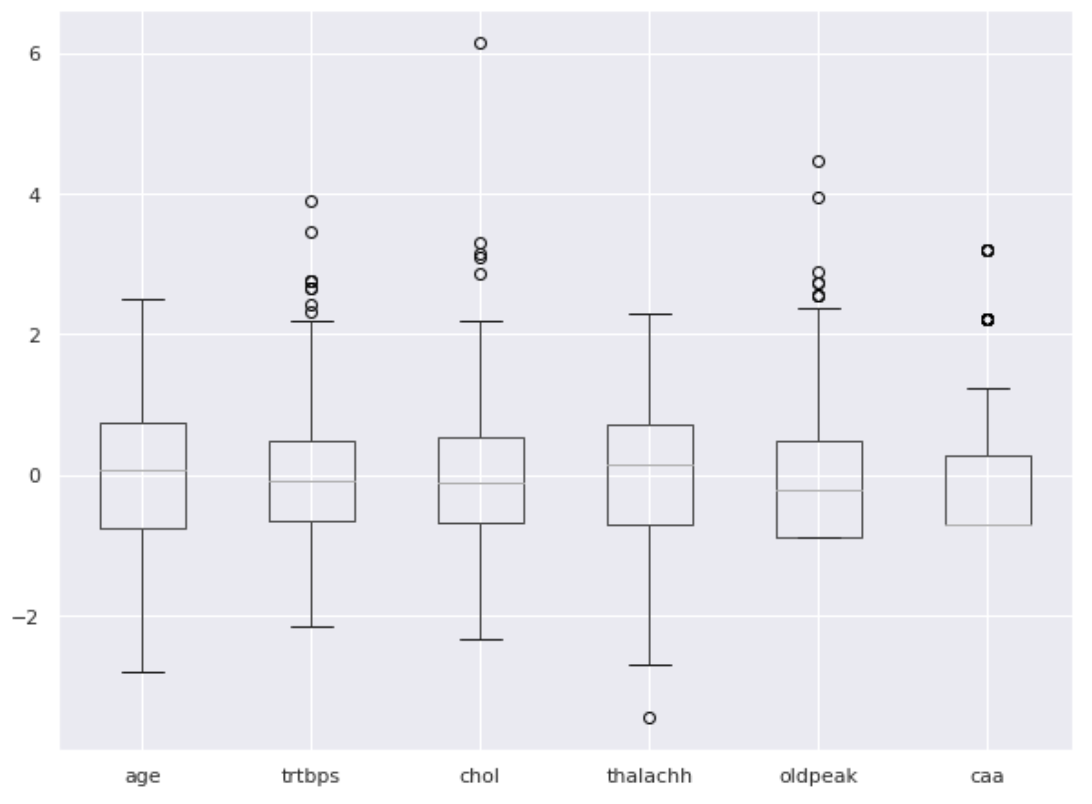
*Figure 9: Correlation Matrix of Cholesterol*

## 

## Outlier Analysis

We then analyzed the five continuous features for outliers using the StandardScaler class (Figure 5). We are able to see outliers present in the Resting Blood Pressure (trtbps), Cholesterol (chol), Maximum Heart Rate Achieved (thalachh), and Oldpeak (oldpeak).

Recognizing these outlier values will guide several key decisions such as which feature scaling and modelling methods to utilize. For example, suppose some of these outliers are genuine errors in measurement and show not be included in our training model - this may inform a decision to use the RobustScaler or MinMaxScaler classes that are less or more sensitive to outliers, respectively. However, after examination, we are able to conclude that these values are not likely to be the result of errors in measurement and therefore should not be excluded from our model. The RobustScaler class might be a good Feature Scaling class to utilize considering these outliers are important for training.



*Figure 6: Continuous Feature Outlier Analysis*

# ALGORITHM SELECTION

* Algorithms intend to apply: regression or classification task? what algorithms planned to use (at least 3 algorithms, the algorithms can be changed in phase 2, can also incorporate deep learning algorithms).

For EDA, pipeline in book is:

* .info(), .describe(), .head()
* Plot distributions of each feature
* Split into training and test set (stratified sampling)
* Correlations
  + Scatter\_matrix
* Combine attributes (optional)
* Remove nulls
* Categorical encoding (we can skip)
* Feature scaling (we will use robustscaler)
* etc