Case Study: Cardiovascular Disease Risk

Sam Loyd - May 2020

Nano, Heart Image, 2019.

The cardiovascular data set obtained from Kaggle will be analyzed to determine correlations for cardiovascular disease based on features found in the data set.

In part 1, visualizations were the primary focus of my analysis. Statistical information on skew and kurtosis were provided. Rows with invalid data were removed in the data wrangling phase and box plots were used to look at outliers which were only removed based on domain knowledge.

In part 2, I focused on feature selection using random forest regressor and applying one-hot encoding.

In part 3, I focused on selecting a model and then on whether any hyperparameters could help.

In the final Wrap-up and Retention phase, CNN was tested and compared. I saved the best model and reduced data sets in pickle files.

References:

Nano, P. (2019, March 14). 25 cardio-healthy foods and prevention of heart disease. [image]. Retrieved from https://guardian.ng/features/health/25-cardio-healthy-foods-and-prevention-of-heart-disease/

Part One

1. **Select topic.**
    A. *Cardiovascular Disease.*
2. **Obtain the cardiovascular data.**
    A. *Kaggle.com*
        i. https://www.kaggle.com/sid321axn/stacked-ensemble-for-heart-disease-classification
    B. *Renamed file to heart.csv.*
    C. *The dataset consists of 1190 records of patients from US, UK, Switzerland and Hungary.*
        i. 11 features
        ii. 1 target variable
3. **Load the cardio data from the "heart.csv" file into a Pandas dataframe.**
4. **Display the dimensions of the cardio file.**
    A. 1190 records by 12 features
5. **Convert any data to prepare for next step (Data Wrangling).**
    A. *Show data types.*
    B. *Relabel Features.*
    C. *Outlier Analysis.*
        i. Use boxplots.
    D. *Data Wrangling*
        i. Outlier removal based on domain knowledge
            1. Cholesterol
            2. Resting Systolic
        ii. Repeat Boxplots.
    E. *Prepare for statistics functions.*
        i. (Data Science is cyclical) Required moving on to step 7 and returning.
    F. *Show any new data types.*
6. **Isolate target from features in the data set.**
    A. *Notice that target is represented as a 1 or 0*
        i. 1 has cardiovascular disease
        ii. 0 does not
    B. *The Cardio Disease variable will be the "target" and the other variables will be the "features"*
7. **Determine questions for analysis of Cardiovascular Risk data:**
    A. *Explain Variables.*
        i. Gather data types.
            1. Binary
                a. Sex: 0-Female, 1-Male

b. Target (Cardio): 0-None, 1-Cardiovascular Disease
c. Glucose – Elevated: 0-No, 1-Yes
d. Exercise induced angina (Ex Angina): 0-No, 1-Yes

2. Categorical
    a. ST Slope: 0-Nomral, 1-Upsloping, 2-Flat, 3-Downsloping
    b. Chest pain (Chest pain):1-Typical, 2-Angina, 3-Non-Angina, 4-Asymptomatic
    c. Resting ECG: 0-Normal, 1-Abnormal ST-T Wave, Left Ventricular Hypertrophy
3. Numerical
    a. Age
    b. Resting BP (Resting Systolic)
    c. Cholesterol
    d. Max heart rate (Max Heart)
    e. Oldpeak (Old Peak)

ii. Discover distributions of numerical features.
    1. The distribution is typically non-Gaussian

iii. If they are categorical, how many different categories?
    1. See Above, but ranges from 2-4.

B. Preliminary Correlation Analysis?

i. Some correlation, but none were particularly strong.

C. *Evaluate cardiovascular risk rates in different categories?*

i. See graphs.

8. Access summary information about the data set such as total, mean, min, max, frequency, and uniqueness.

A. New questions.

i. The mean cholesterol is high.
    1. Is this data set biased?

ii. The mean blood pressure is high.
    1. Is this further evidence of biased data?

B. Preliminary Conclusions.

i. The mean age is closer to middle age.

9. **Create histograms.**

A. Most of the patients were between 40 and 65. This data set is likely biased towards a mid to older age group.

B. Cholesterol ranges were typically between 200 and 300 which is considered high.

C. Run skew and kurtosis statistics to verify both in charts.

10. Create bar charts for categorical data.

A. *There were more males in the data set than females*.

11. **Create Spearman's Ranking charts.**

A. *There were no strong correlations between numeric values*.

12. **Create Parallel Coordinates visualization to compare the distributions of numerical variables between patients with cardio disease and those with none.**
    A. *Those with higher max heart rates have less cardiovascular disease.*
    B. *Those with a higher old peak have higher rates of cardiovascular disease.*
13. **Create Stack Bar Charts to compare passengers who survived to passengers who didn't survive based on the other variables.**
    A. *Men show a higher percentage of cardio disease*.
    B. *Chest pain categorized as atypical shows a higher percentage of cardio disease*.

Part Two

14. **Use one-hot encoding to convert categorical values.**
    A. *Resting ECG*
    B. *Chest Pain*
15. **Evaluate Missingness.**
    A. *Missingno matrix showed none.*
16. **Repeat correlation analysis on larger set of data which now includes one-hot encoded features.**
17. **Perform Feature Reduction using Random Forrest Regressor from sklearn.**
    A. *Fit and apply model using RandomForrestRegressor.*
    B. *Graph model.*
    C. *Allow SelectFromModel also from sklearn to make the final determination*.
        i. It found 7 – Age, Resting Systolic, Cholesterol, Max Heart Rate, Exercise Induced Angina, Old Peak, and Chest Pain = 4

Part Three

18. **Use model selection function to help select from the following classifier methods using K-Folds.**
    A. *Logistic Regression*
    B. *Random Forrest (fastest)*
    C. *XGBoost (close 2nd)*
    D. *SVM*
19. **Graph the results of previous step.**
20. **Run RandomForrestClassifier and XGB against the data using 67/33 split.**
    A. *Evaluate model for accuracy, precision, F1 and AUC.*
    B. *Random Forrest is slight winner again with an accuracy of .87.*
21. **Graph results from previous step.**

22. Use GridSearchCV function to try to improve accuracy.

A. *Only slight improvement on a few tests and not consistent at that.*
B. *Moving forward with defaults*.

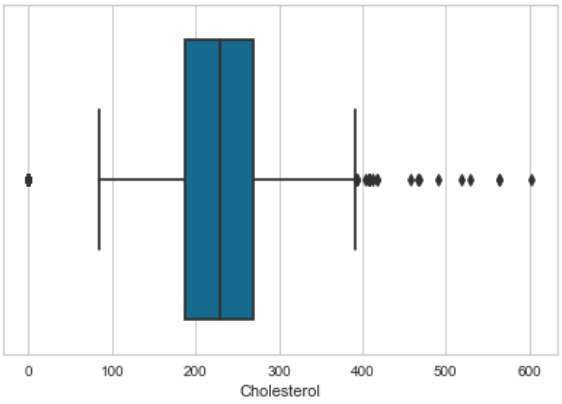
Wrap Up and Data Retention

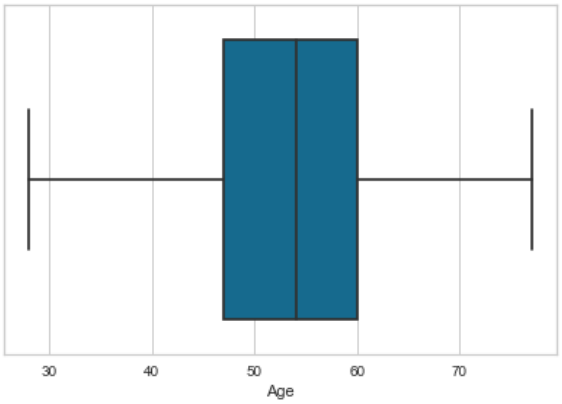
23. Try a CNN.

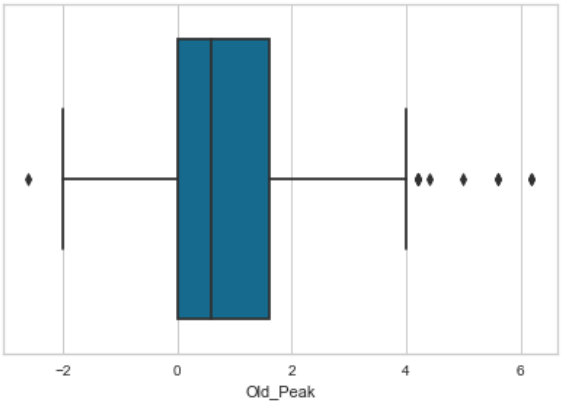
A. *MLPClassifier was less useful than Random Forest at making predictions.*

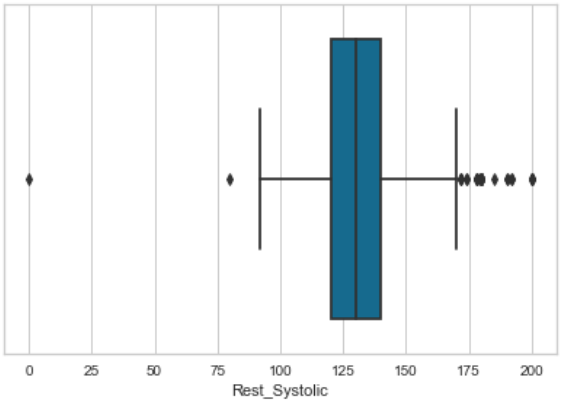
24. Saved reduced data set and Random Forest model for later use.

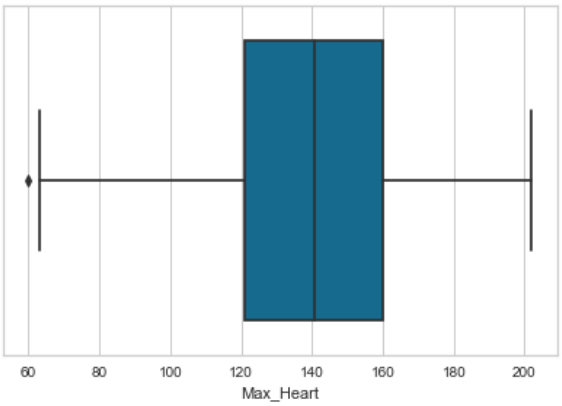
Prior Outlier Removal Analysis

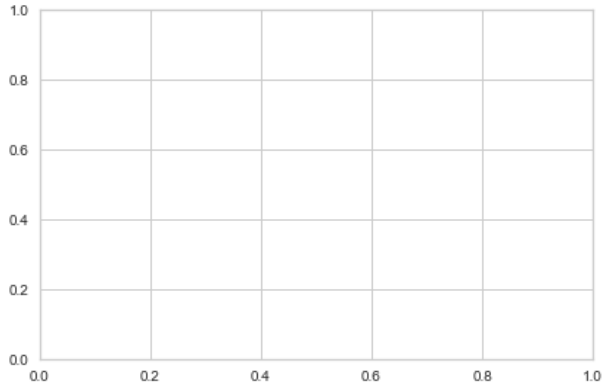






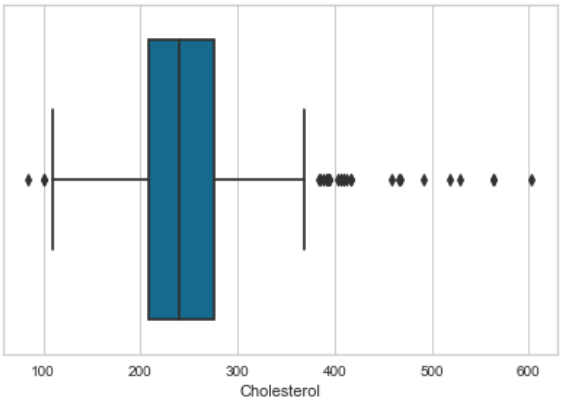


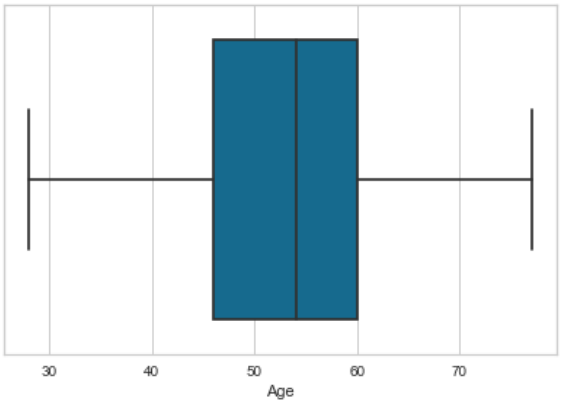


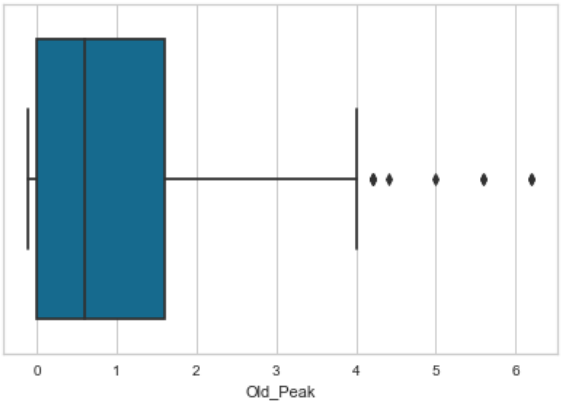


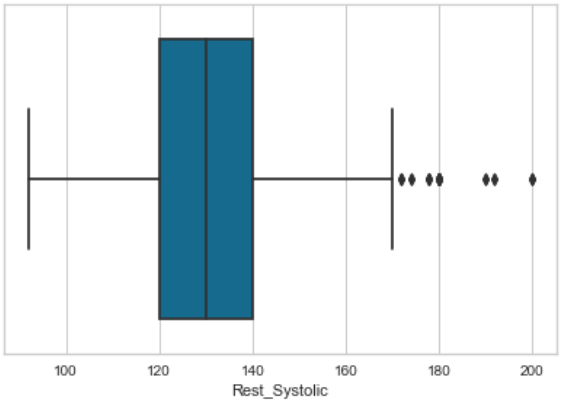
Analysis: While there were outliers in several features, most were realistic and withn an acceptable range. Domain knowledge was used to remove any that were obvious clerical errors such as a systolic of 0 and cholesterol of 0.

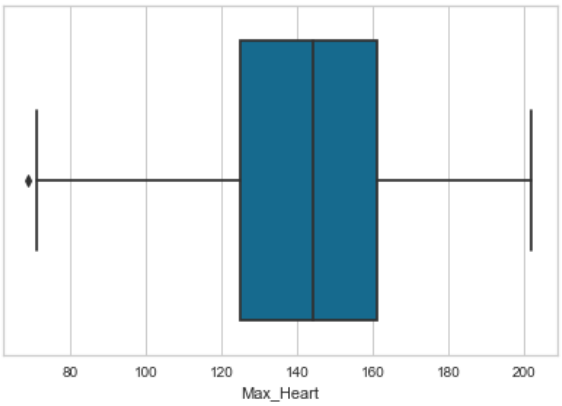
Post Outlier Removal Analysis

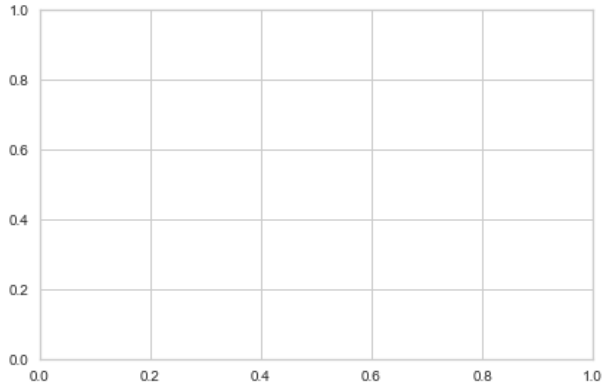






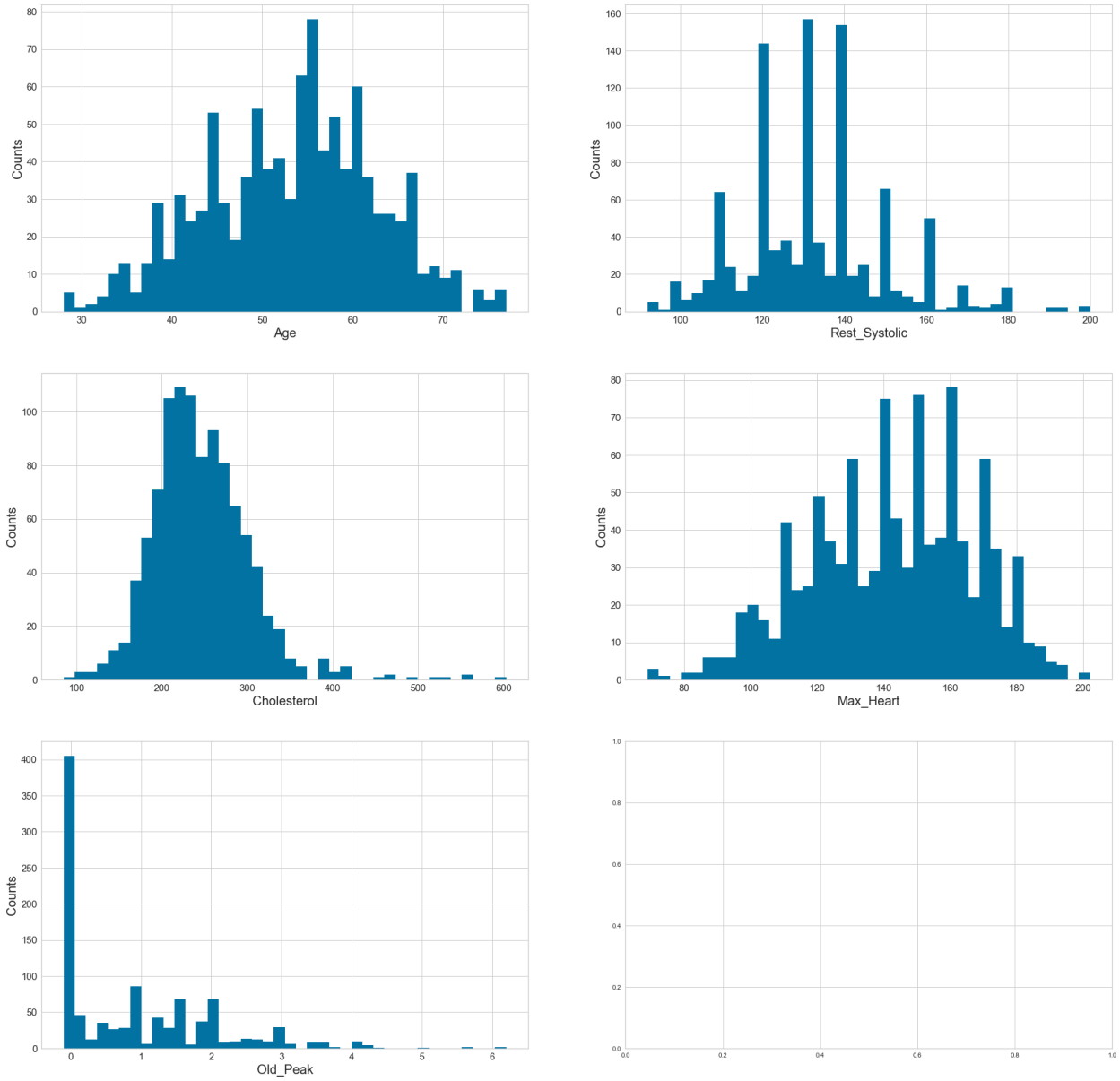






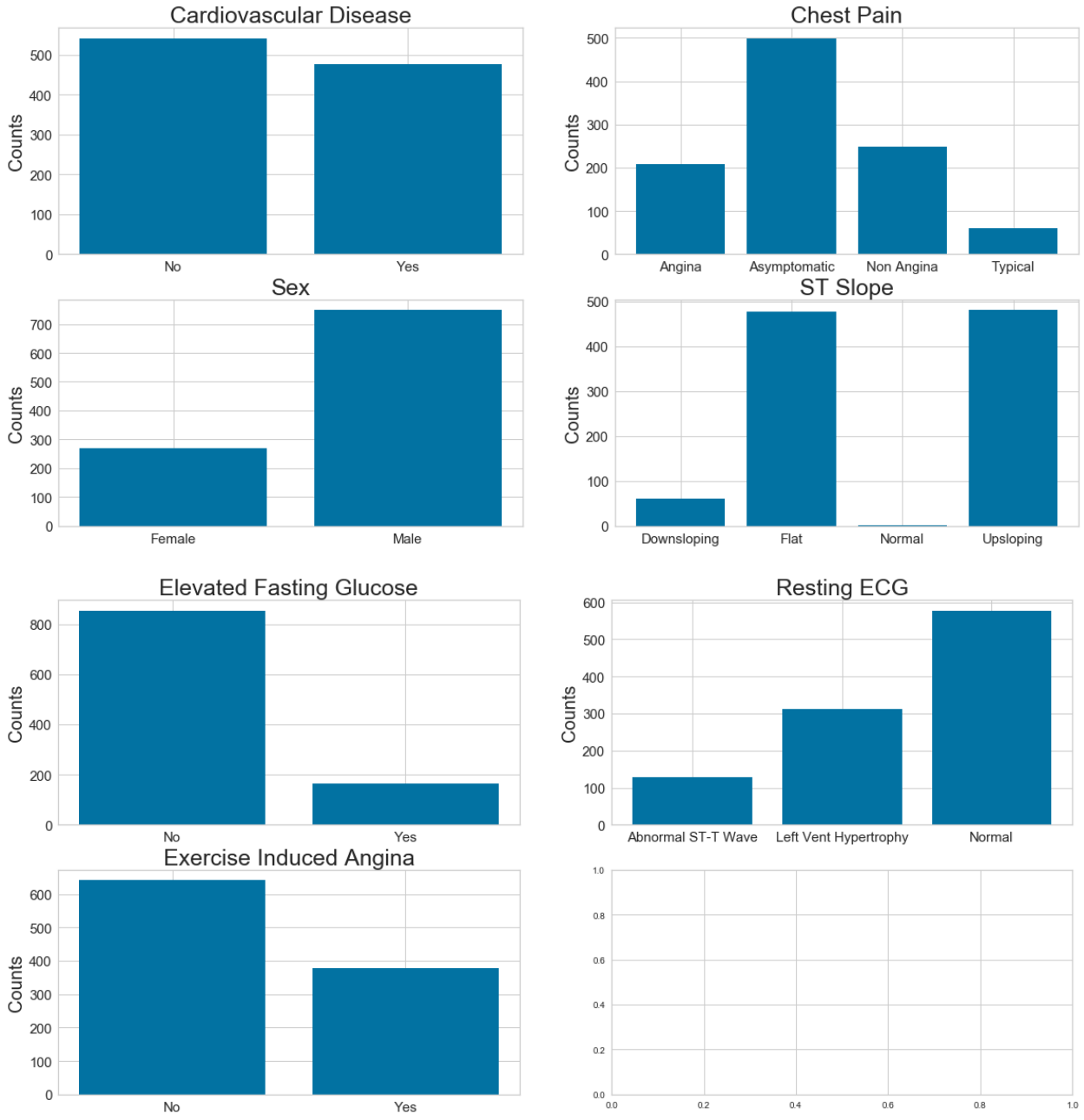
Analysis: Improvement post outlier removal based on domain knowledge. Any further removal could have unintended consequences and skew results incorrectly.

Histograms



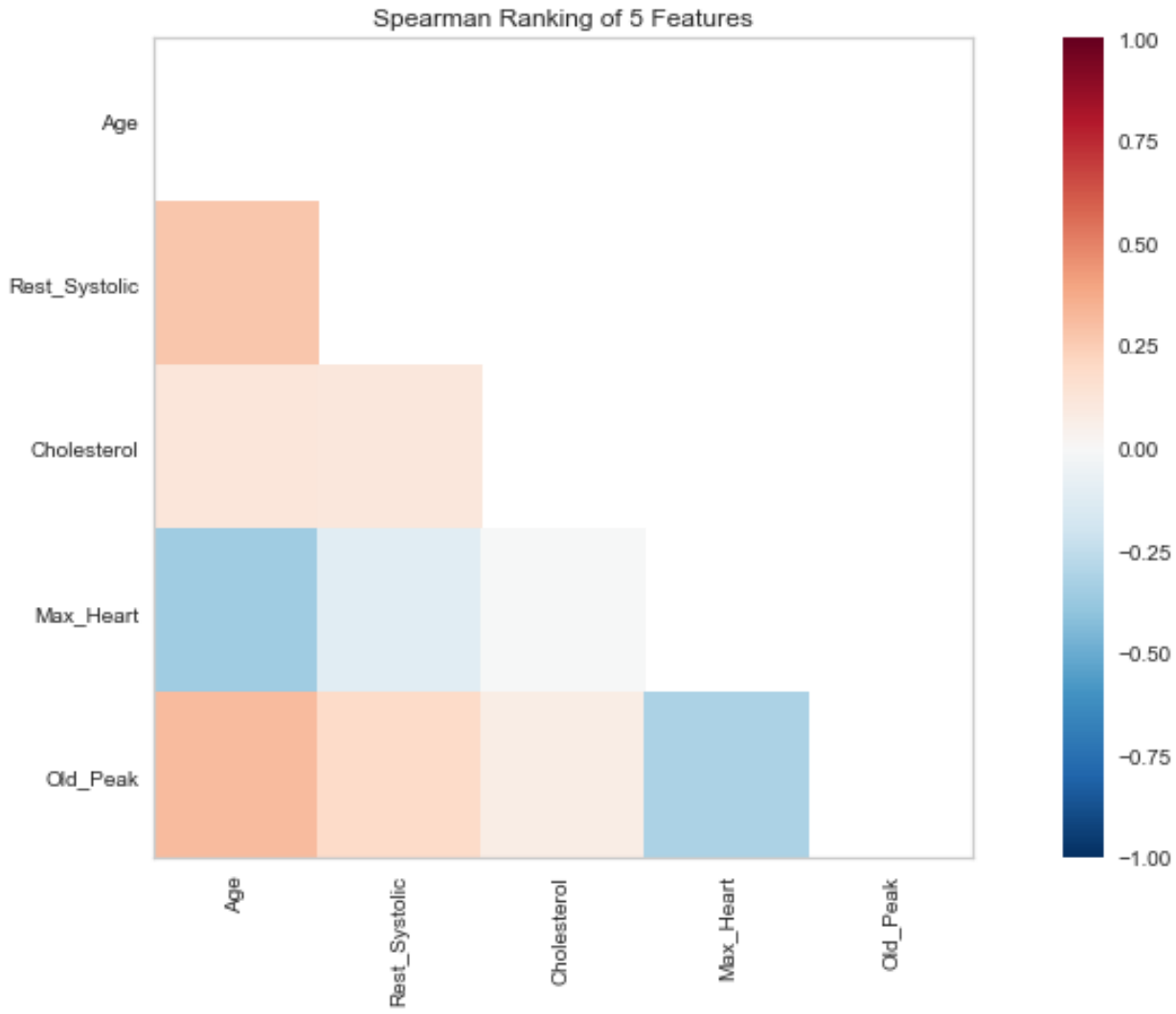
Analysis: The graphs show significant skew and kurtosis in several features. The data set is primarily non-Gaussian or has a non-normal distribution. This would indicate not using Pearson for correlation analysis as it assumes a normal distribution. Of note, much of the data for cholesterol lies in a range that is considered high based on domain knowledge. This might indicate a bias in the data set. The same concern applies to resting blood pressure (systolic).

Bar Chart



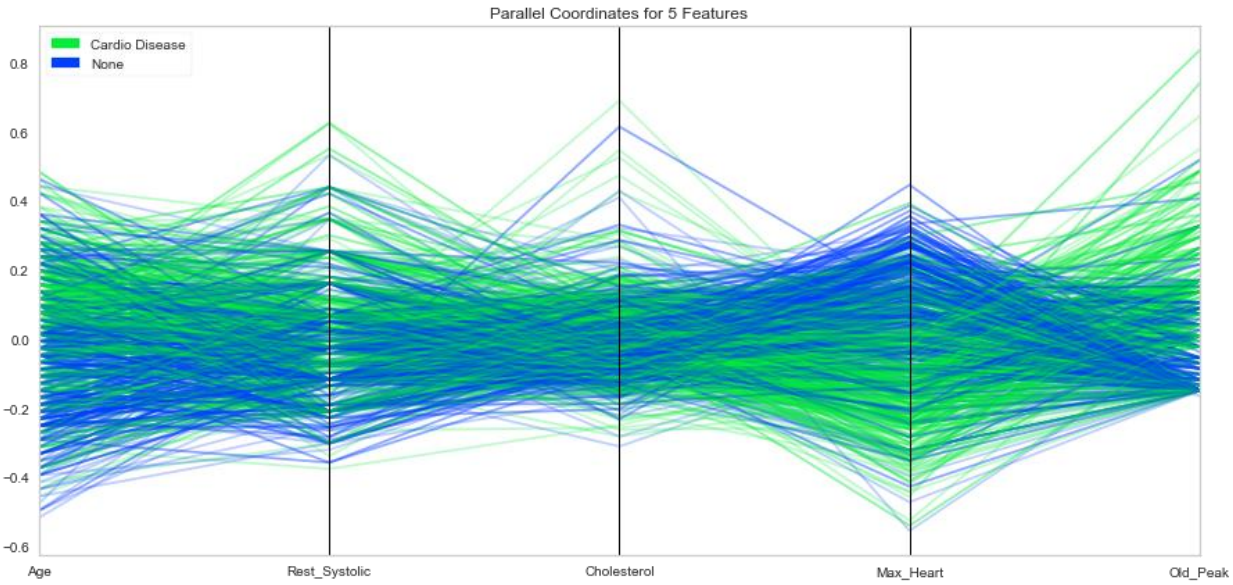
Analysis: The data set includes many more males than females. *This is not representative of the population at large and could indicate bias.*

Spearman's Correlation



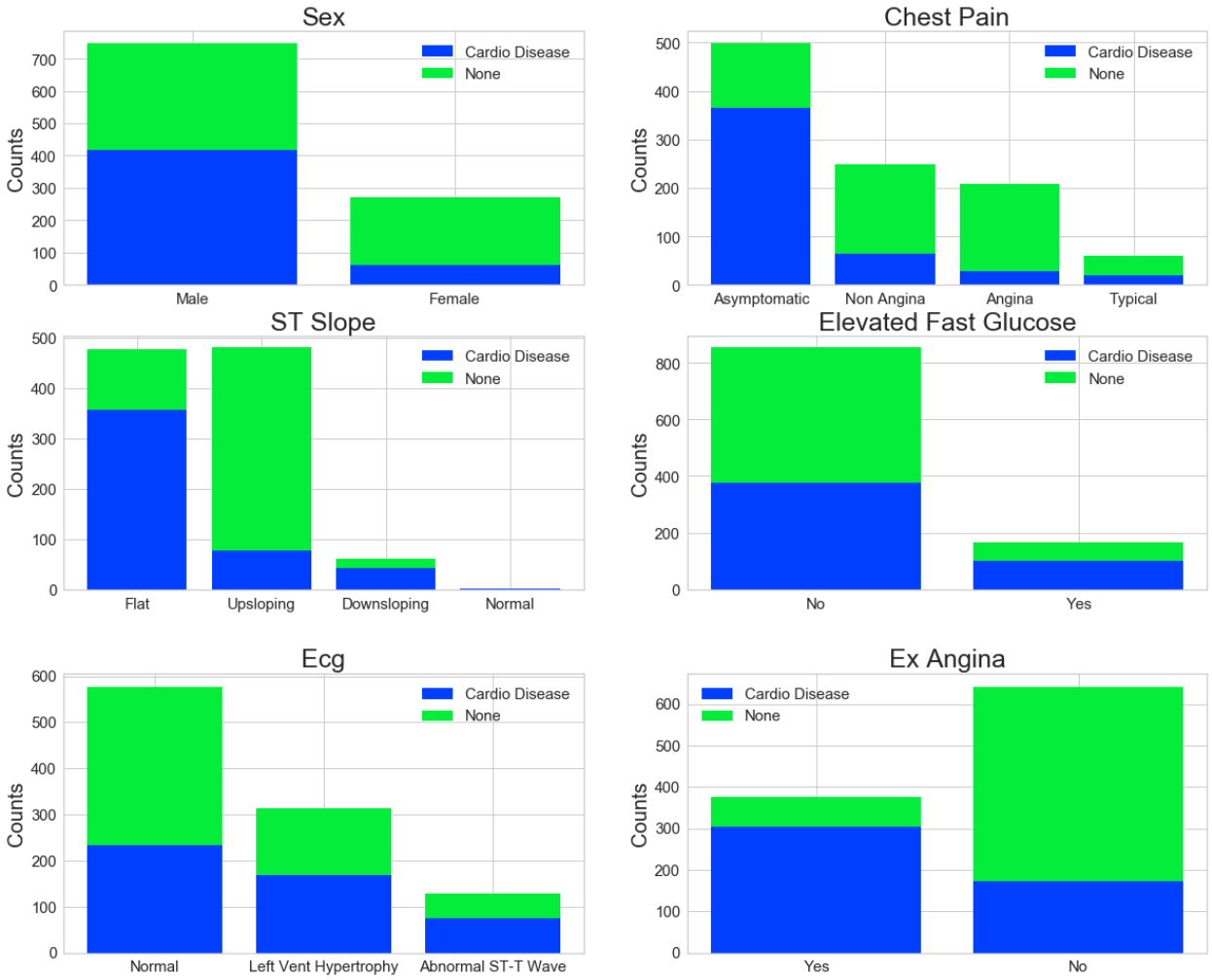
Analysis: No strong correlations are shown from this preliminary analysis. Spearman method was applied given the non-normal distribution shown in previous graphs and statistical measures that can be found in the html file from the source python code. Note the target variable was not shown so further correlation analysis is required post data wrangling.

Parallel Coordinate Chart



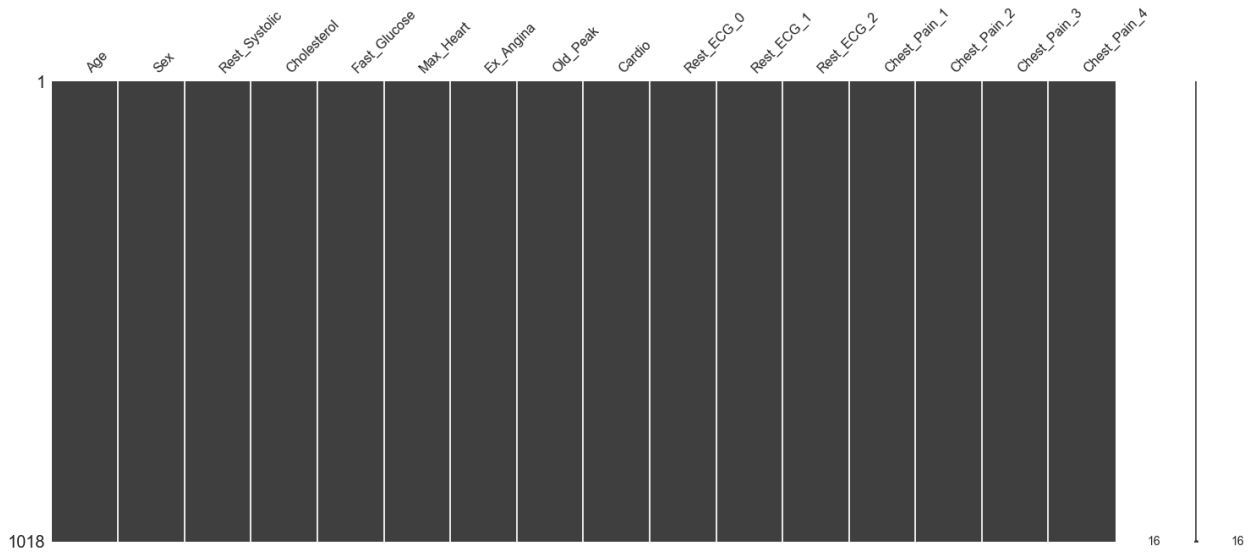
Analysis: There are two standouts based on this chart. Those with higher max heart rates have less cardiovascular disease. Those with a higher old peak have higher rates of cardiovascular disease.

Stacked Bar Charts



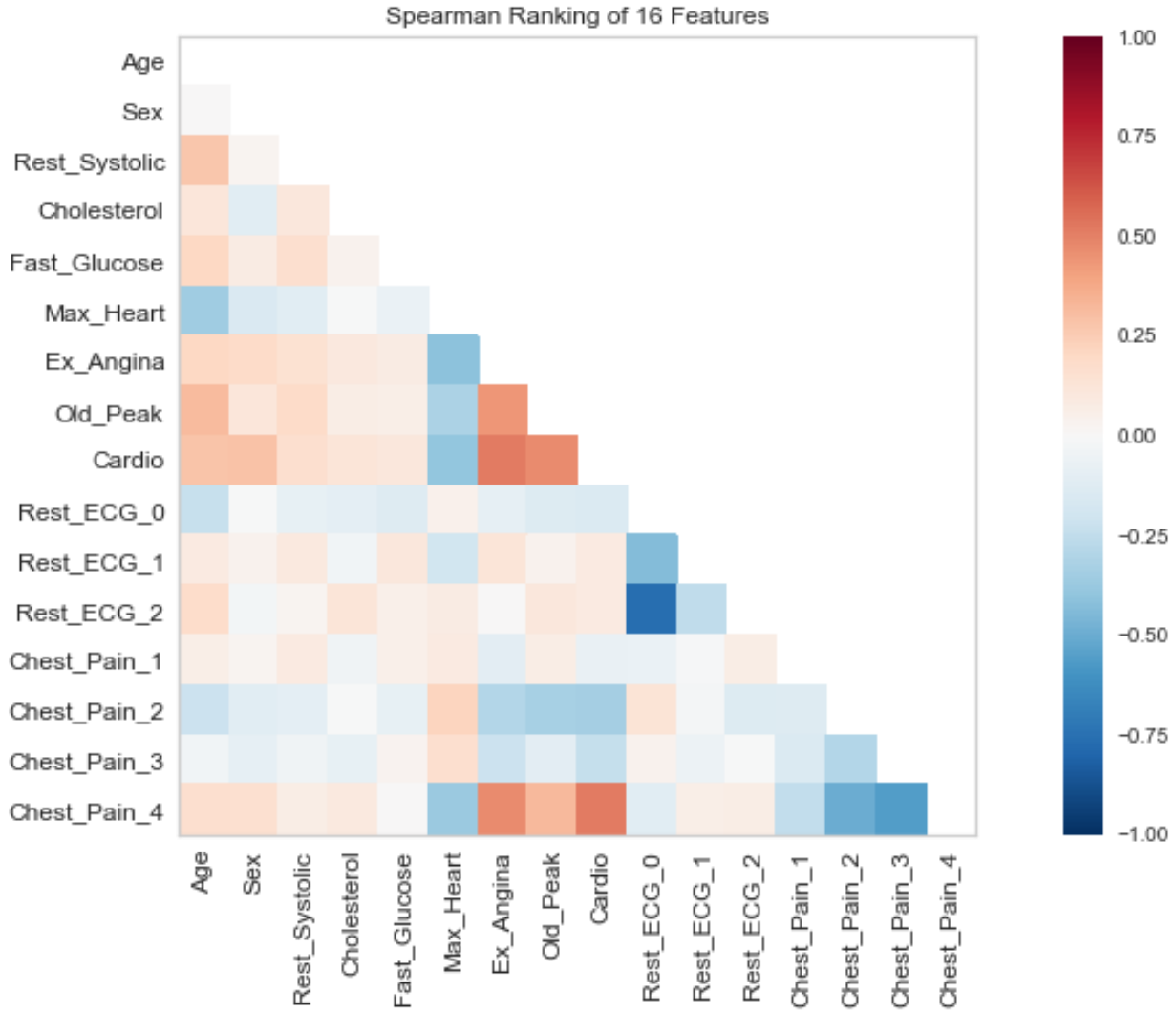
Analysis: While several inferences can be drawn from these charts, key standouts were the higher rates of cardiovascular disease in males in the data set compared to females. Cardio disease showed prevalence in those with exercised induced angina and having a flat ST Slope.

Missingness Matrix



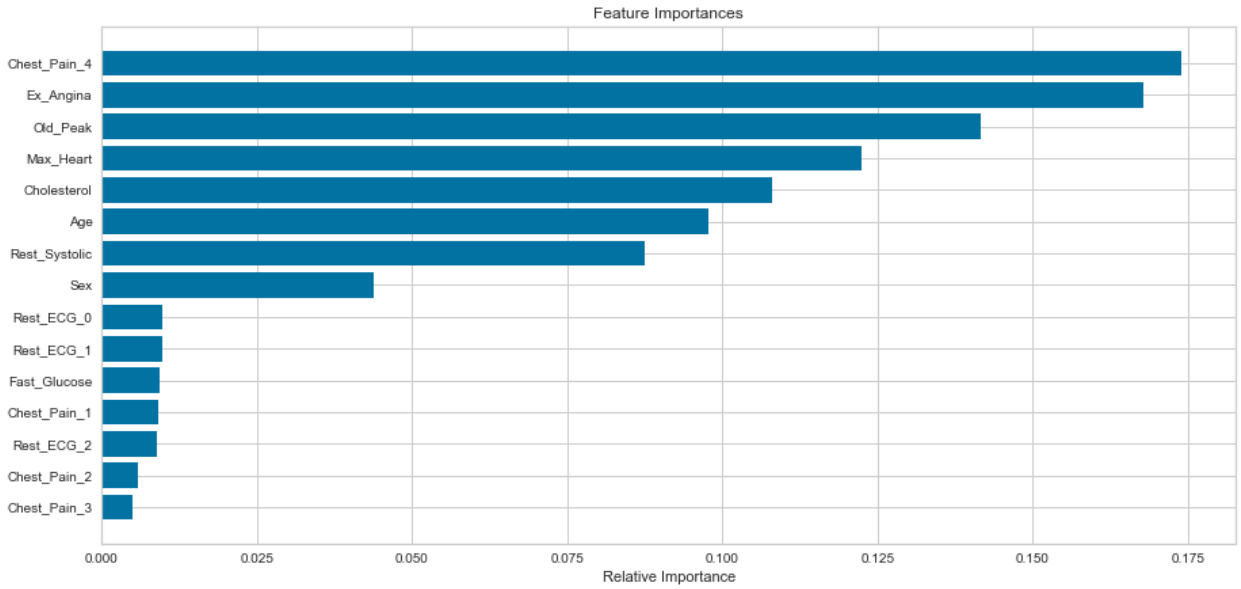
Analysis: There is no significant missingness in this dataset discovered by the missingno function.

Correlation Analysis Post One-Hot using Spearman



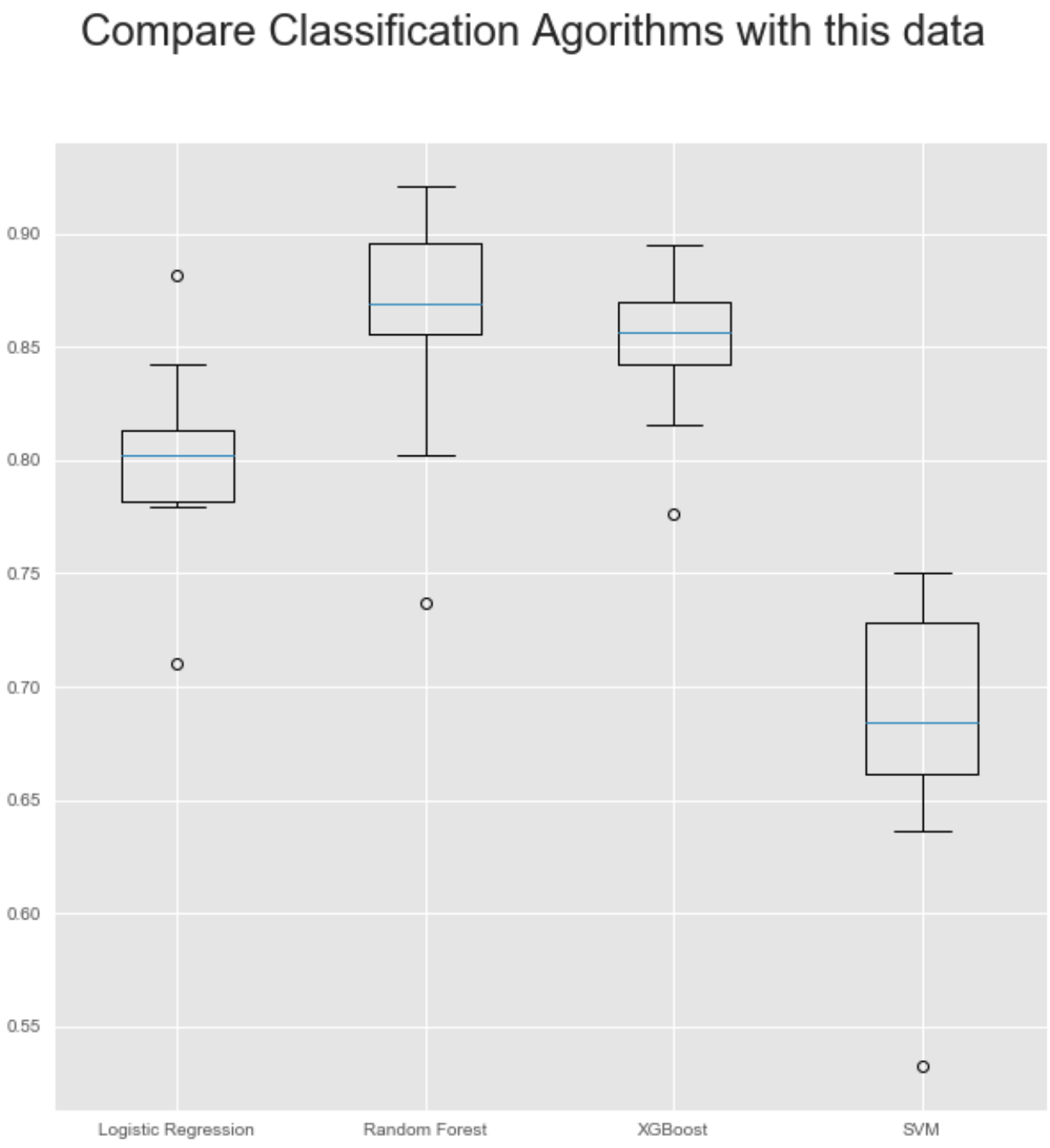
Analysis: This graph was created after one-hot encoding was used. Given the target variable of cardiovascular disease or cardio, the most significant correlation was noticed in chest_pain_4, ex_angina, old_peak and max_heart. This correlation was created using Spearman's method due to the non-normal data distribution.

Feature Reduction Analysis – Random Forrest



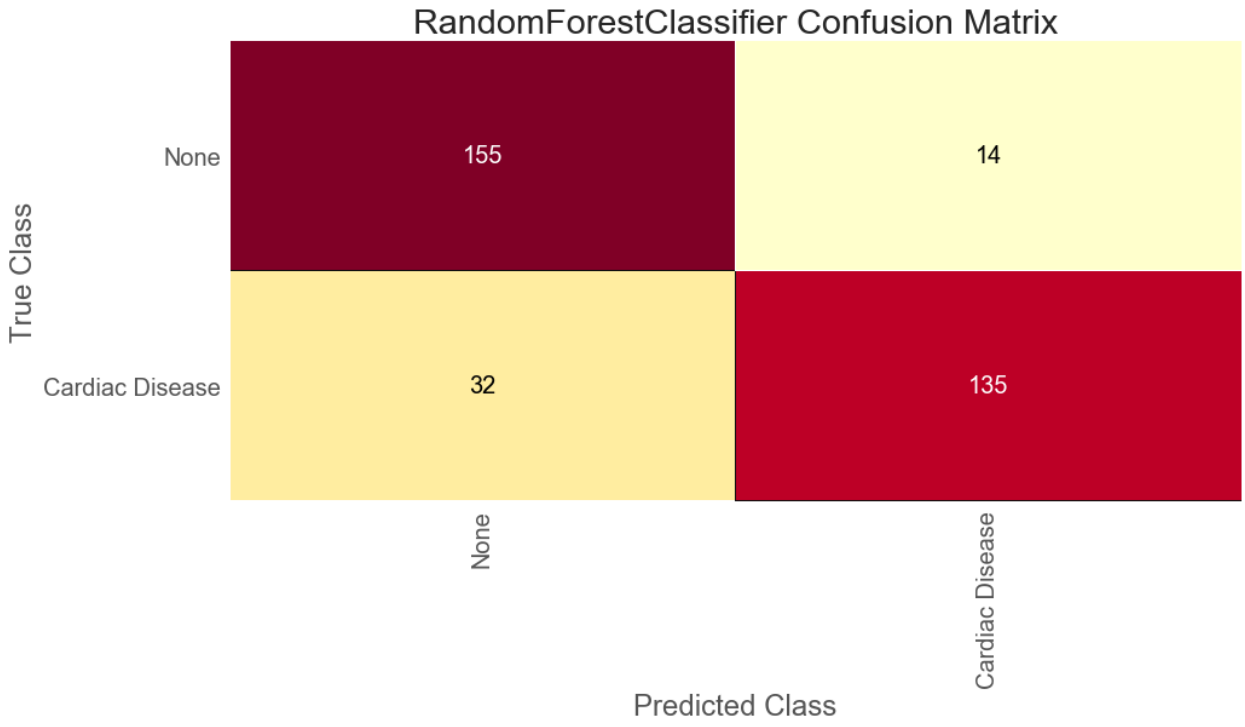
Analysis: The RandomForrestRegressor function and SciKit Learn's feature_selection functions were used for feature selection. K-folds were created. It selected the top seven features from the graph above as the most important to use for modeling.

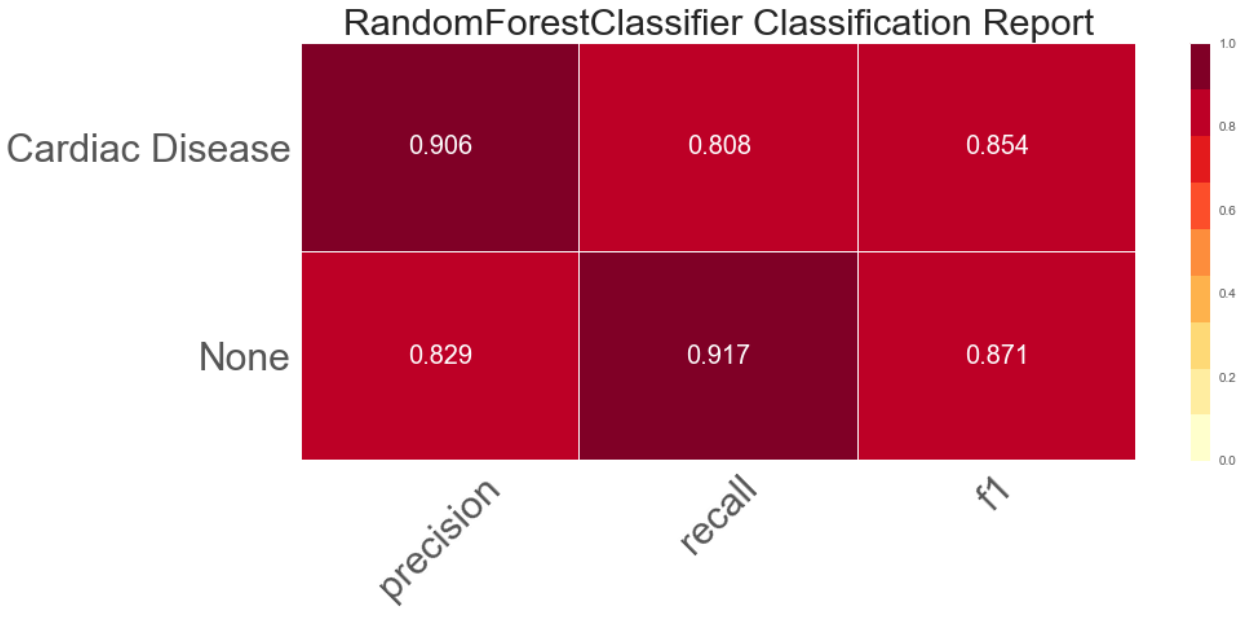
Model Selection



Analysis: Random Forest and XGBoost were worth further exploration based on the results of running the model_selection function.

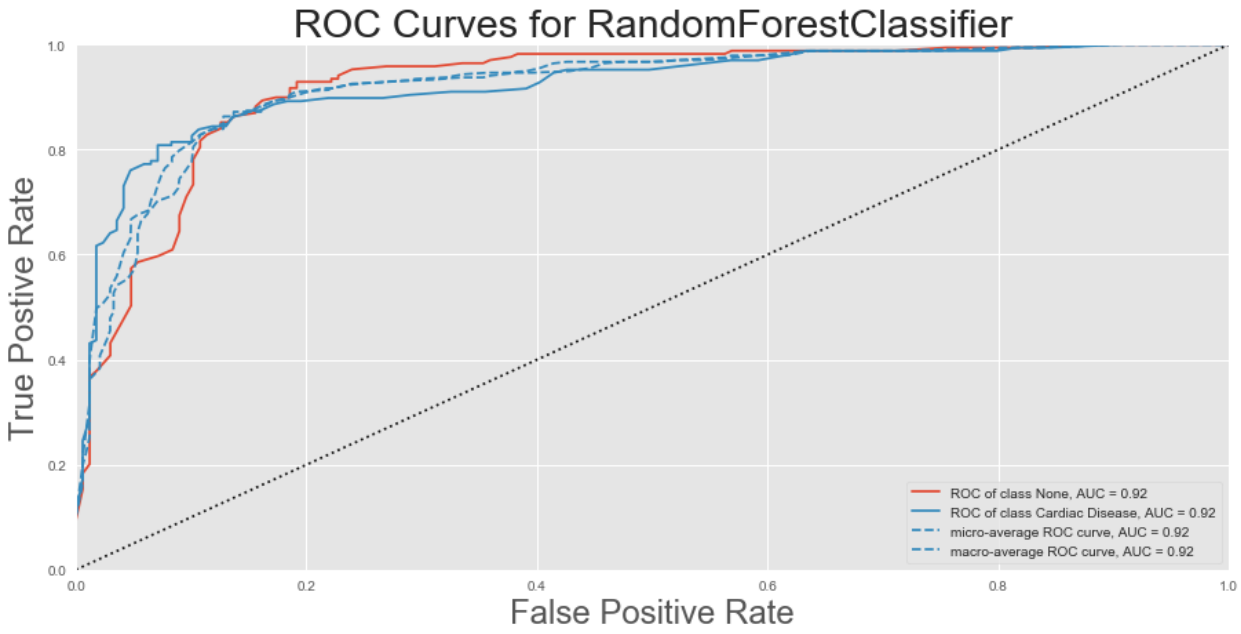
Model Evaluation





Analysis: **Random Forest Classifier** provided for the best predictions overall considering precision, recall and accuracy.

Conclusions and Final Model Evaluation



Conclusions: As noted, **Random Forest Classifier** provided the best predictive results with this data set. CNN was added and compared as well, but Random Forest still provided the best results. The reduced data set and best performing model was saved for later use in pickle format.

Disclaimer*: As noted in earlier charts, some bias, including gender bias, has been noted in the population used for this data set. This could lead to issues with accuracy when applied to the larger population.*