Ryan Bolduan

Coding Dojo - Data Science 8.30 Cohort

Oct 31, 2021

Data Source = Kaggle: https://www.kaggle.com/fedesoriano/heart-failure-prediction

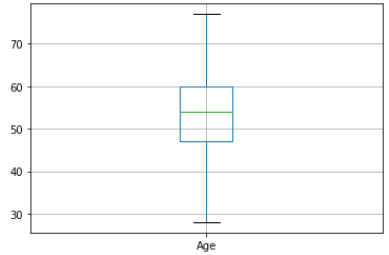
Intro and Data Dictionary

- Cardiovascular diseases (CVD's) are the number one cause of death globally.
- The Kaggle data set contains 11 features that can be used to predict a possible heart attack, including:
    - Age = age of patient
    - Sex = male or female
    - Chest Pain Type = type of chest pain
    - Resting BP = resting blood pressure
    - Cholesterol = serum cholesterol
    - Fasting BS = fasting blood sugar
    - Resting ECG = resting ECG results
    - Max HR = max HR achieved
    - Exercise Angina = exercised induced angina
    - Oldpeak = related to ECG score
    - ST_Slope = slope of peak exercise ST segment

Data Cleaning Steps

- Data set is a combination of international data sources combined
- Missing Data: No missing data in the data set (918, 12)
- Duplicated data: No duplicates were found in the data set
- Outliers: There were several outliers, all of which were explored
    - Biggest challenge with outliers was that many of them were truly outliers; not typical values but within the range of possibility

Age Exploration



	Age
count	918.000000
mean	53.510893
std	9.432617
min	28.000000
25%	47.000000
50%	54.000000
75%	60.000000
max	77.000000

- No outliers in the age group,
- The youngest age was 28 years old, which seemed young for heart disease consideration

Categorical Values

```
# Exploring 'ChestPainType' column
df['ChestPainType'].value_counts()
# Values reflect the values listed on Kaggle
```

```
ASY    496
NAP    202
ATA    173
TA      46
Name: ChestPainType, dtype: int64
```

```
# Exploring 'RestingECG' column
df['RestingECG'].value_counts()
# values are consistent with values listed on Kaggle
```

```
Normal    551
LVH       188
ST        178
Name: RestingECG, dtype: int64
```

```
# Exploring  column
df['Sex'].value_counts()
# Values reflect correct values
# Data set is NOT balanced
```

```
M    724
F    193
Name: Sex, dtype: int64
```

```
# Exploring 'ST_Slope' column
df['ST_Slope'].value_counts()
# values are consistent with values listed on Kaggle
```

```
Flat    459
Up      395
Down     63
Name: ST_Slope, dtype: int64
```

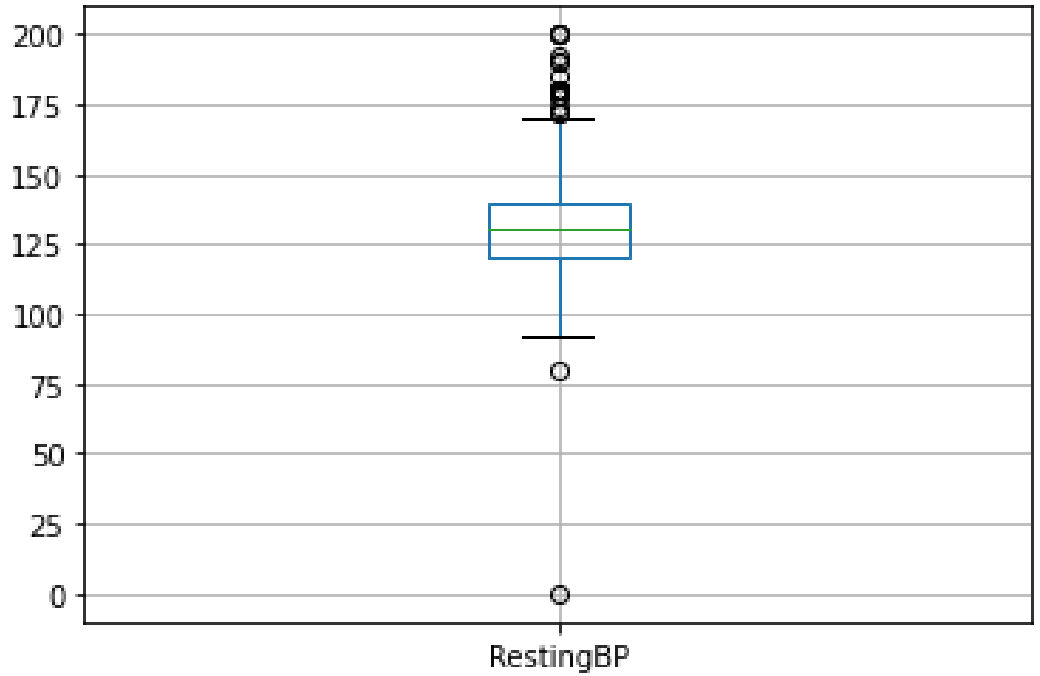
```
# Exploring 'RestingECG' column
df['ExerciseAngina'].value_counts()
# values are consistent with values listed on Kaggle
```

```
N    546
Y    371
Name: ExerciseAngina, dtype: int64
```

- All categorical values were checked and there were no inconsistencies with categories las listed on Kaggle's data dictionary
- Categorical values are not balanced

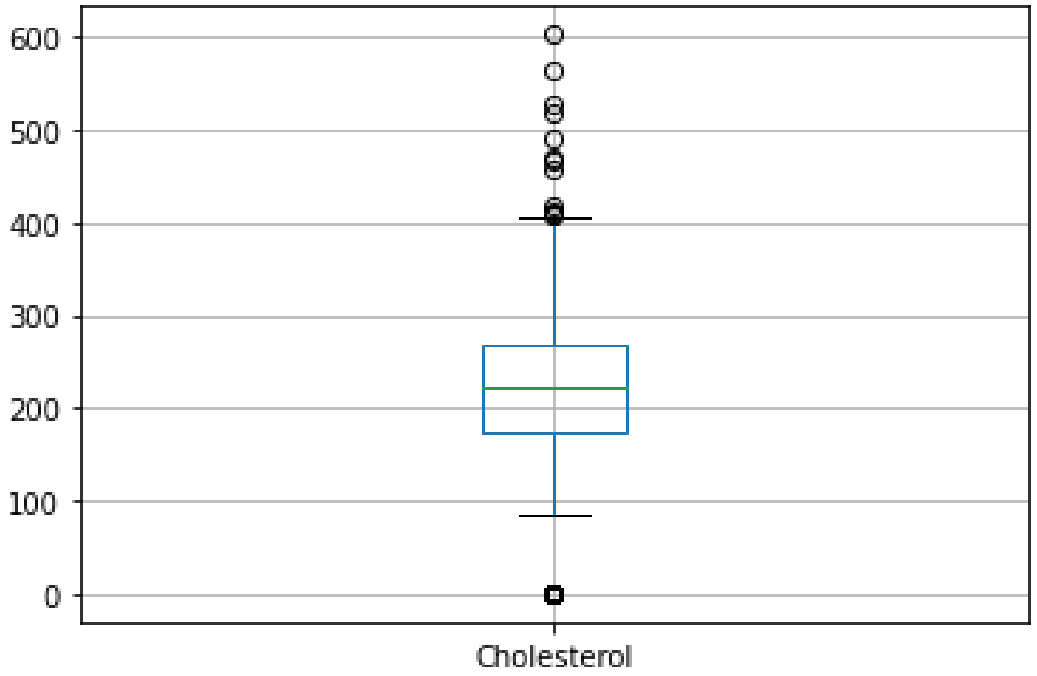
Resting BP Data Visualization

- Resting BP had one data point with a Resting BP of zero, which was dropped
- All other outliers were kept



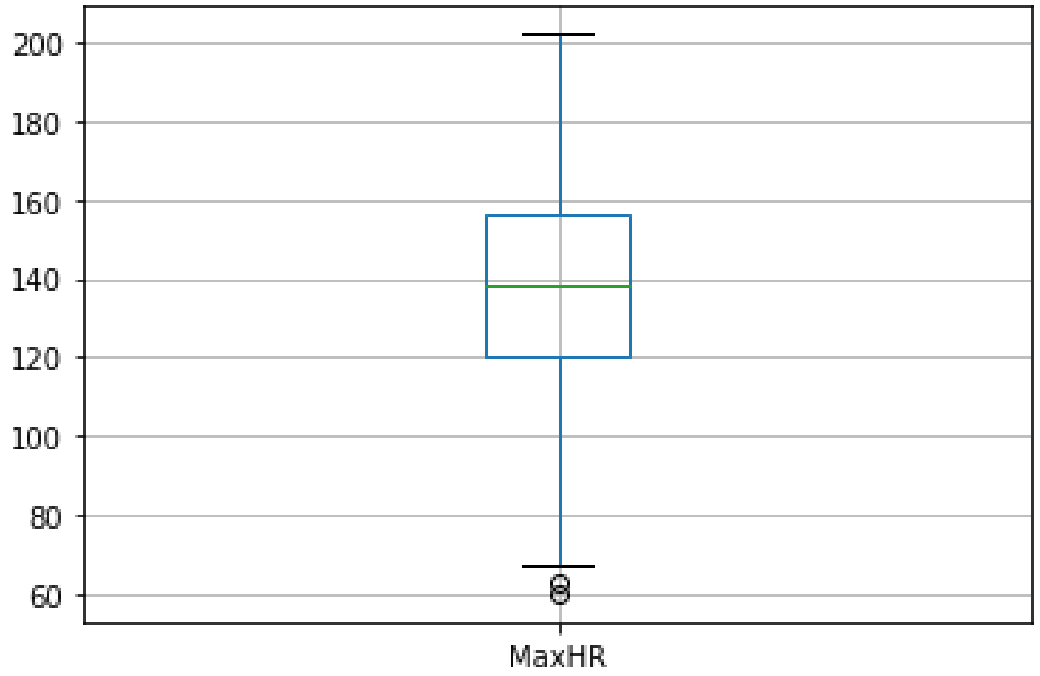
Cholesterol Data Visualization

- Cholesterol had several values (171) with a value of zero
- Cholesterol value of zero represents almost 20% of the data points, so they were left in
- Other outliers, despite seeming incredibly high, were left in



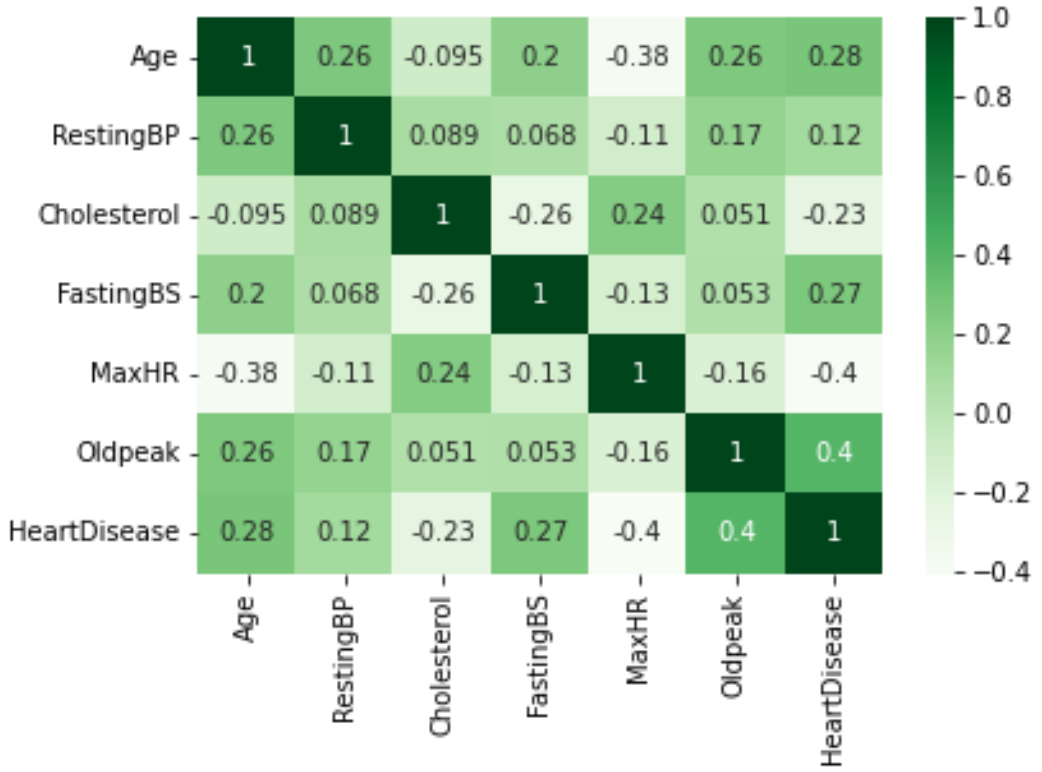
Max HR Data Visualization

- Five data points had MaxHR of <70bpm
- Four of the five had heart disease, so decided to keep them in the data set

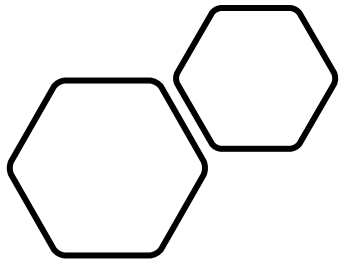


Heat Map

- Very little correlation between any of the columns and heart disease
- Next steps will be to turn to ML models to be able to combine the impacts of multiple factors in order to better predict heart disease



Unique Challenges with Data set



- Data set was relatively clean
- Numerous outliers were the most challenging part of the data
- Some outlier data points were obvious drops
- Most of the outliers ended up being possible after doing some internet research on possible value range for each attribute

- Good practice on how data scientists end up making judgement calls as part of their data cleaning and exploration!