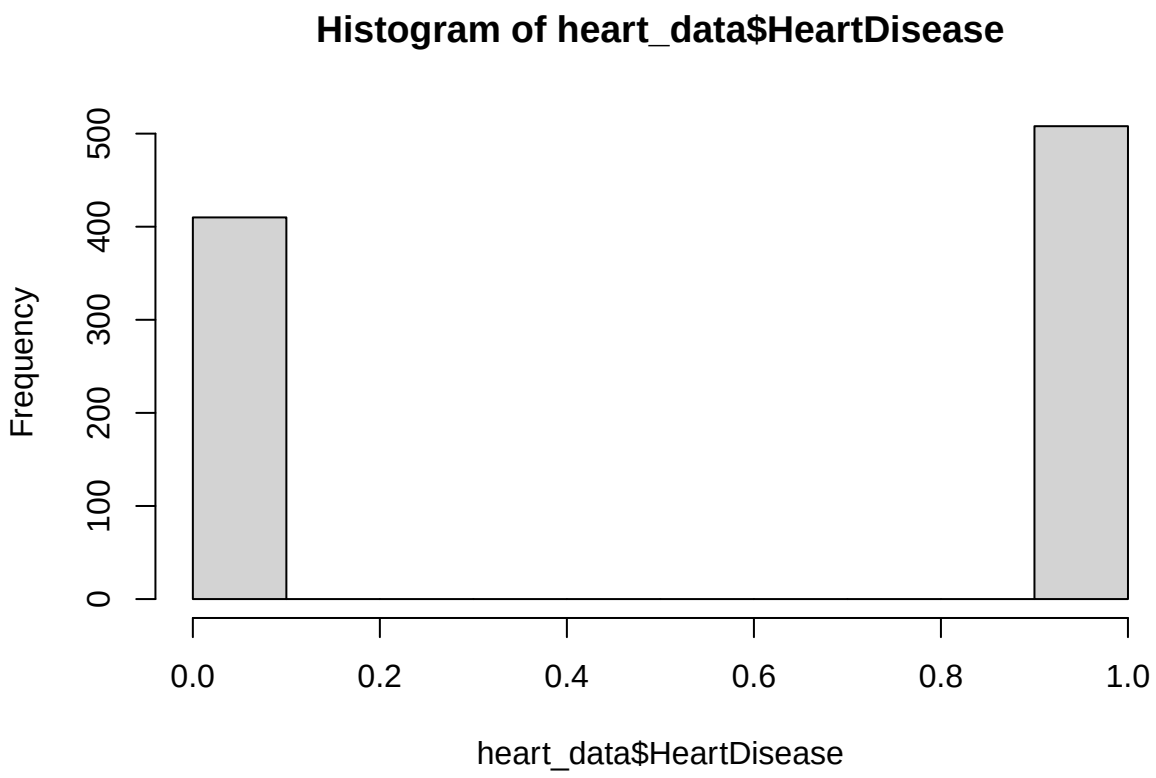
Project Paper 1

Alex Ojemann

2023-03-19

How is your response variable distributed and are there any outliers?

```
hist(heart_data$HeartDisease)
```



My response variable is binary, so there can't be any outliers. We can see that there are more patients that do have heart disease that don't.

What are the real-world implications of the response distribution and outliers?

Outliers are not applicable here because our response variable is binary. The distribution containing more instances of heart disease than instences of no heart disease reflects that this data is not representing the general population of people living near these hospitals, but rather the population of patients at the hospitals, specifically those who are at great enough risk of heart disease to have their biometrics in this data sets measured.

What is a reasonable confidence interval and interpretation for the mean response?

```
m <- mean(heart_data$HeartDisease)

sterr <- sqrt((mean(heart_data$HeartDisease)*(1-mean(heart_data$HeartDisease)))/nrow(heart_data))

critical <- qnorm(p=0.025,mean=0,sd=1,lower.tail=FALSE)

m-critical*sterr
```

```
## [1] 0.5212175
```

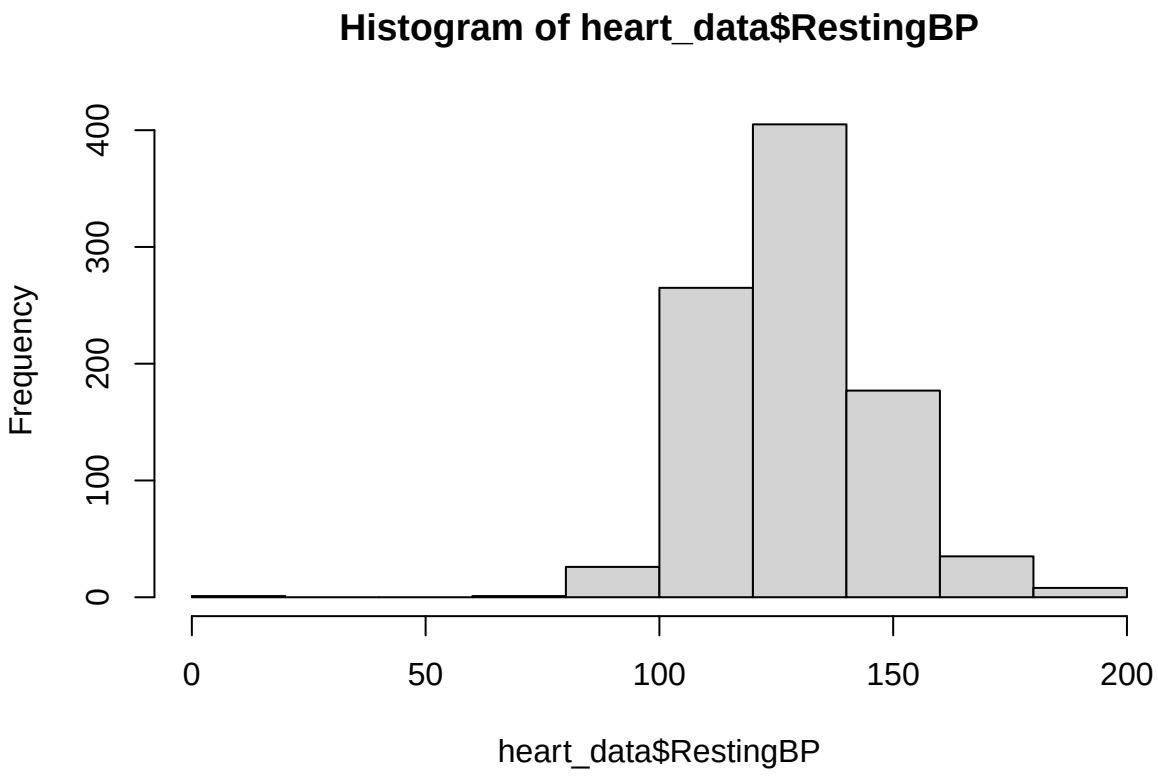
```
m+critical*sterr
```

```
## [1] 0.5855363
```

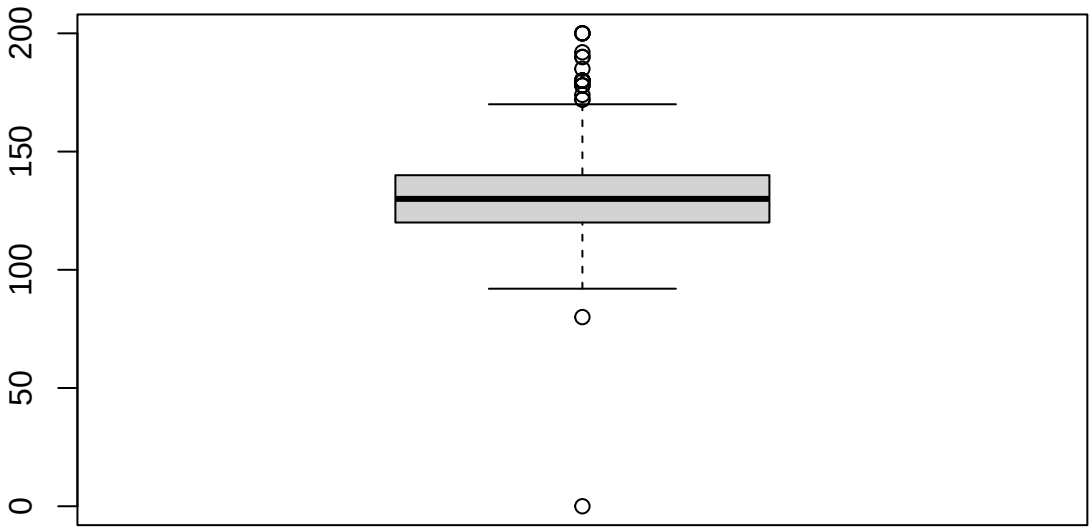
We are 95% confident that the true proportion of patients from these hospitals that have heart disease is between 0.5212175 and 0.5855363.

How are each of your predictor variables distributed (choose 5 of them) and are there any outliers?

```
hist(heart_data$RestingBP)
```




```
boxplot(heart_data$RestingBP)
```



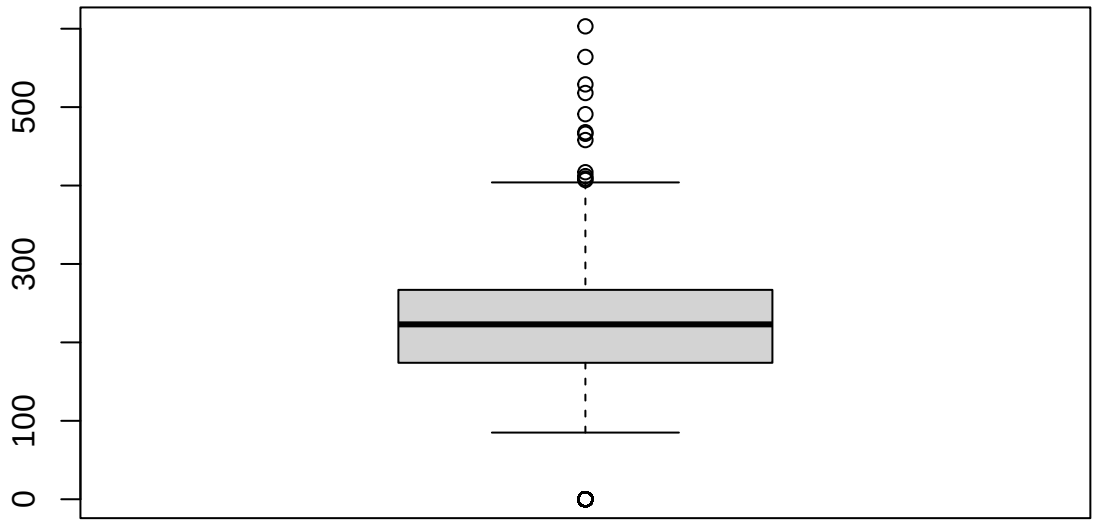
The RestingBP variable is approximately normally distributed but has a number of outliers.

```
heart_data <- dplyr::filter(heart_data, RestingBP > 0)
```

The code above removes the row where RestingBP is 0.

```
hist(heart_data$Cholesterol)
```


```
boxplot(heart_data$Cholesterol)
```

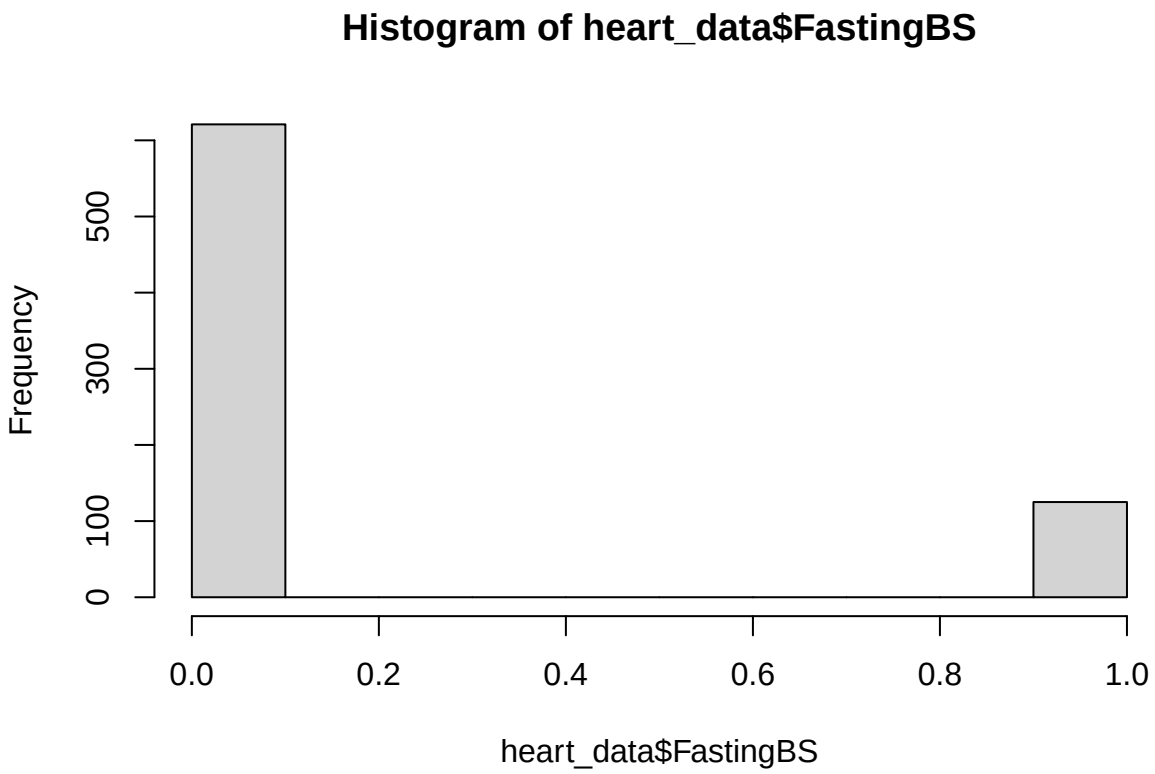


The Cholesterol variable is approximately normally distributed with a slight right skew, however it has many outliers.

```
heart_data <- filter(heart_data, Cholesterol > 0)
```

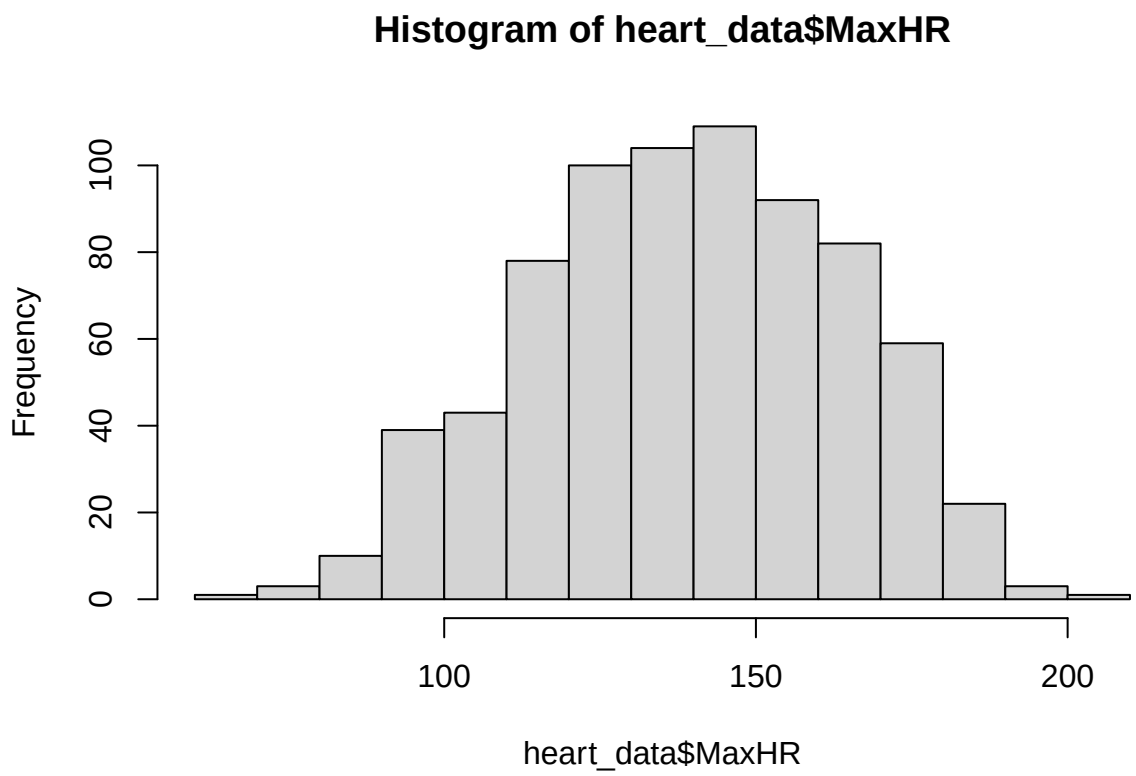
The code above removes the row where Cholesterol is 0.

```
hist(heart_data$FastingBS)
```

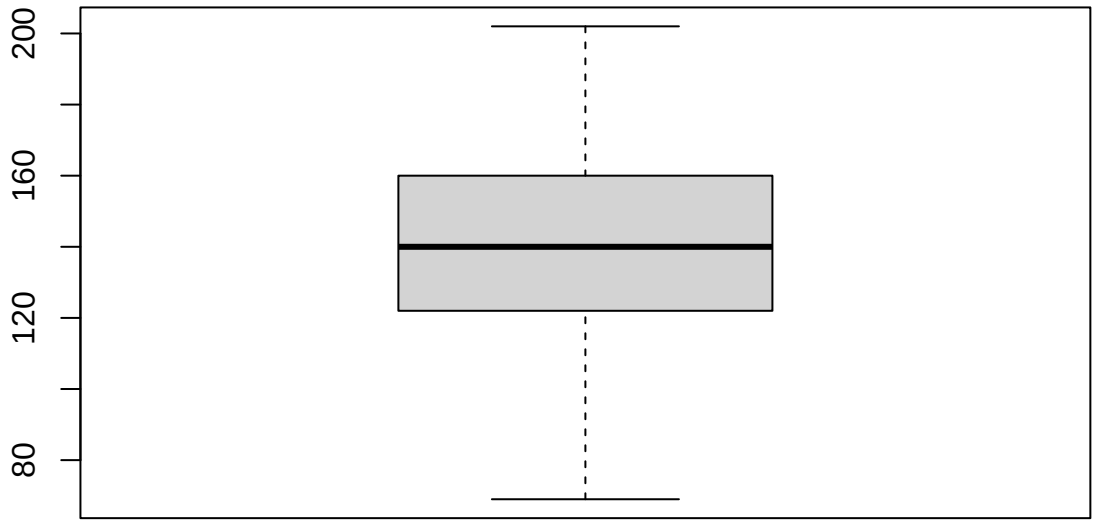


The FastingBS variable represents whether a patient's fasting blood sugar is above 120 mg/dl which is binary so outliers are not possible. The distribution is heavily slanted towards not having a fasting blood sugar above 120 mg/dl.

```
hist(heart_data$MaxHR)
```



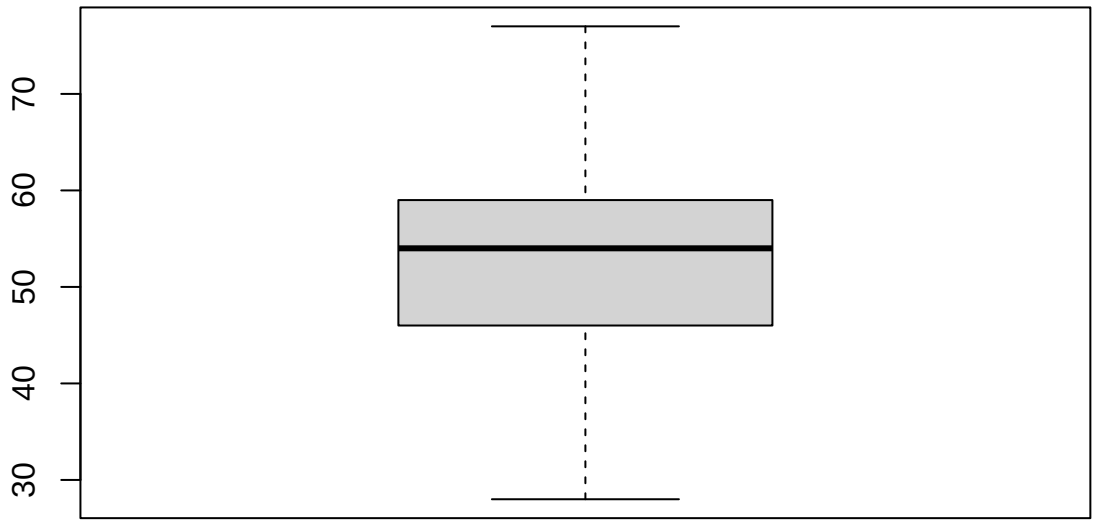

```
boxplot(heart_data$MaxHR)
```



The MaxHR variable is very close to normally distribued with no outliers.

```
hist(heart_data$Age)
```


```
boxplot(heart_data$Age)
```



The Age variable is very close to normally distribued with no outliers.

What are the real-world implications of the predictor distributions and outliers?

Most of the outliers of the RestingBP variable as shown on the boxplot are within reason, however one of them is at 0. This entry will be removed as a blood pressure of 0 mm Hg cannot exist in a living being because blood flow wouldn't be possible.

The outliers of the cholesterol variable above the top whisker in the boxplot aren't a concern for our purposes because given a patient is in this database they were likely deemed enough of a heart disease risk to have these measurements taken so higher cholesterol than normal is possible. However it's not possible to have a cholesterol level of 0 mm/dl so those points must be removed.

The distribution being heavily slanted towards not having a fasting blood sugar above 120 mg/dl means a lot of nuance is likely being left out in the group of patients below 120 mg/dl. For example, a patient with 50 mg/dl of fasting blood sugar is treated the same as a patient with 110 mg/dl, which isn't ideal for prediction.

The MaxHR and Age variables being normally distributed with no outliers is ideal for prediction. However, the age variable having a normal distribution centered around 55 doesn't align with the ages of the entire human population. It shows that the patients in this data set are generally older than the average person.

What are reasonable confidence intervals and interpretations for the mean predictors?

```
m <- mean(heart_data$RestingBP)

sterr <- sd(heart_data$RestingBP)/nrow(heart_data)

critical <- qnorm(p=0.025,mean=0,sd=1,lower.tail=FALSE)

m-critical*sterr
```

```
## [1] 132.9774
```

```
m+critical*sterr
```

```
## [1] 133.0682
```

We are 95% confident that the true mean resting blood pressure of patients from these hospitals is between 132.9774 and 133.0682.

```
m <- mean(heart_data$Cholesterol)

sterr <- sd(heart_data$Cholesterol)/nrow(heart_data)

critical <- qnorm(p=0.025,mean=0,sd=1,lower.tail=FALSE)

m-critical*sterr
```

```
## [1] 244.48
```

```
m+critical*sterr
```

```
## [1] 244.7908
```

We are 95% confident that the true mean cholesterol of patients from these hospitals is between 244.48 and 244.7908.

```
m <- mean(heart_data$FastingBS)

sterr <- sqrt((mean(heart_data$FastingBS)*(1-mean(heart_data$FastingBS)))/nrow(heart_data))

critical <- qnorm(p=0.025,mean=0,sd=1,lower.tail=FALSE)

m-critical*sterr
```

```
## [1] 0.14076
```

```
m+critical*sterr
```

```
## [1] 0.1943607
```

We are 95% confident that the true proportion of patients from these hospitals that have fasting blood sugar levels above 120 mg/dl is between 0.14076 and 0.1943607.

```
m <- mean(heart_data$MaxHR)

sterr <- sd(heart_data$MaxHR)/nrow(heart_data)

critical <- qnorm(p=0.025,mean=0,sd=1,lower.tail=FALSE)

m-critical*sterr
```

```
## [1] 140.1621
```

```
m+critical*sterr
```

```
## [1] 140.291
```

We are 95% confident that the true mean maximum heart rate of patients from these hospitals is between 140.1621 and 140.291.

```
m <- mean(heart_data$Age)

sterr <- sd(heart_data$Age)/nrow(heart_data)

critical <- qnorm(p=0.025,mean=0,sd=1,lower.tail=FALSE)

m-critical*sterr
```

```
## [1] 52.85706
```

```
m+critical*sterr
```

```
## [1] 52.90701
```

We are 95% confident that the true mean age of patients from these hospitals is between 52.85706 and 52.90701.

Are any of the predictor variables strongly correlated with one another?

```
heart_data_encoded <- model.matrix(~.-1, data = heart_data)

cor(heart_data_encoded)
```

```
##                          Age         SexF         SexM ChestPainTypeATA
## Age               1.00000000 -0.040916665  0.040916665      -0.22981979
## SexF             -0.04091667  1.000000000 -1.000000000       0.13883493
## SexM              0.04091667 -1.000000000  1.000000000      -0.13883493
## ChestPainTypeATA -0.22981979  0.138834927 -0.138834927       1.00000000
## ChestPainTypeNAP -0.02227073  0.080301808 -0.080301808      -0.28953128
## ChestPainTypeTA   0.05067094 -0.013732192  0.013732192      -0.12901417
## RestingBP         0.25986472 -0.034363208  0.034363208      -0.07609206
## Cholesterol       0.05875824  0.107044935 -0.107044935      -0.01594523
## FastingBS         0.24133789 -0.096074630  0.096074630      -0.09332479
## RestingECGNormal -0.23859165  0.002763163 -0.002763163       0.13124161
## RestingECGST      0.10985264 -0.037573963  0.037573963      -0.03292047
## MaxHR            -0.38211212  0.157001819 -0.157001819       0.24332922
## ExerciseAnginaY   0.24590819 -0.192579093  0.192579093      -0.31704862
## Oldpeak           0.28600628 -0.125742907  0.125742907      -0.30319982
## ST_SlopeFlat      0.20254751 -0.102287982  0.102287982      -0.31477319
## ST_SlopeUp       -0.27229567  0.136713359 -0.136713359       0.36390111
## HeartDisease      0.29861668 -0.292778703  0.292778703      -0.37563365
##                  ChestPainTypeNAP ChestPainTypeTA   RestingBP Cholesterol
## Age                   -0.022270730     0.050670938  0.25986472  0.05875824
## SexF                   0.080301808    -0.013732192 -0.03436321  0.10704494
## SexM                  -0.080301808     0.013732192  0.03436321 -0.10704494
## ChestPainTypeATA      -0.289531282    -0.129014171 -0.07609206 -0.01594523
## ChestPainTypeNAP       1.000000000    -0.130512712 -0.05429930 -0.06844057
## ChestPainTypeTA       -0.130512712     1.000000000  0.06506908 -0.05025269
## RestingBP             -0.054299295     0.065069076  1.00000000  0.09593929
## Cholesterol           -0.068440573    -0.050252694  0.09593929  1.00000000
## FastingBS              0.005850402     0.049294628  0.17376511  0.05401235
## RestingECGNormal       0.027343588    -0.053436980 -0.08608479 -0.04694976
## RestingECGST          -0.062745275     0.002047785  0.08669048 -0.02880555
## MaxHR                  0.152320727     0.091130903 -0.12577393 -0.01985579
## ExerciseAnginaY       -0.197575791    -0.118155331  0.16103496  0.08676806
## Oldpeak               -0.135818565     0.042978131  0.19857506  0.05848813
## ST_SlopeFlat          -0.091037440    -0.017147416  0.08174556  0.10399566
## ST_SlopeUp             0.115119619    -0.002133238 -0.12707718 -0.09906255
## HeartDisease          -0.222145942    -0.053765149  0.17324158  0.10386560
##                     FastingBS RestingECGNormal RestingECGST       MaxHR
## Age                0.241337886     -0.238591647  0.109852640 -0.38211212
## SexF              -0.096074630      0.002763163 -0.037573963  0.15700182
## SexM               0.096074630     -0.002763163  0.037573963 -0.15700182
## ChestPainTypeATA  -0.093324794      0.131241610 -0.032920472  0.24332922
## ChestPainTypeNAP   0.005850402      0.027343588 -0.062745275  0.15232073
## ChestPainTypeTA    0.049294628     -0.053436980  0.002047785  0.09113090
## RestingBP          0.173765114     -0.086084792  0.086690481 -0.12577393
## Cholesterol        0.054012354     -0.046949760 -0.028805547 -0.01985579
## FastingBS          1.000000000     -0.165081142  0.135072464 -0.10270984
## RestingECGNormal  -0.165081142      1.000000000 -0.545514080  0.03915396
## RestingECGST       0.135072464     -0.545514080  1.000000000 -0.15997283
## MaxHR             -0.102709842      0.039153964 -0.159972833  1.00000000
## ExerciseAnginaY    0.109995227     -0.090977942  0.132126785 -0.39628903
## Oldpeak            0.055568286     -0.130778326  0.062598463 -0.25953263
## ST_SlopeFlat       0.105542473     -0.066571228  0.062415934 -0.34615243
## ST_SlopeUp        -0.147309008      0.103039343 -0.068182365  0.38278640
## HeartDisease       0.160594102     -0.133254588  0.095921070 -0.37721219
```

```
##                  ExerciseAnginaY     Oldpeak ST_SlopeFlat   ST_SlopeUp
## Age                   0.24590819  0.28600628   0.20254751 -0.272295667
## SexF                 -0.19257909 -0.12574291  -0.10228798  0.136713359
## SexM                  0.19257909  0.12574291   0.10228798 -0.136713359
## ChestPainTypeATA     -0.31704862 -0.30319982  -0.31477319  0.363901112
## ChestPainTypeNAP     -0.19757579 -0.13581856  -0.09103744  0.115119619
## ChestPainTypeTA      -0.11815533  0.04297813  -0.01714742 -0.002133238
## RestingBP             0.16103496  0.19857506   0.08174556 -0.127077185
## Cholesterol           0.08676806  0.05848813   0.10399566 -0.099062552
## FastingBS             0.10999523  0.05556829   0.10554247 -0.147309008
## RestingECGNormal     -0.09097794 -0.13077833  -0.06657123  0.103039343
## RestingECGST          0.13212678  0.06259846   0.06241593 -0.068182365
## MaxHR                -0.39628903 -0.25953263  -0.34615243  0.382786405
## ExerciseAnginaY       1.00000000  0.46549054   0.44035435 -0.503960135
## Oldpeak               0.46549054  1.00000000   0.37064189 -0.553175497
## ST_SlopeFlat          0.44035435  0.37064189   1.00000000 -0.890996622
## ST_SlopeUp           -0.50396014 -0.55317550  -0.89099662  1.000000000
## HeartDisease          0.55183450  0.49569625   0.59155358 -0.653758736
##                  HeartDisease
## Age                 0.29861668
## SexF               -0.29277870
## SexM                0.29277870
## ChestPainTypeATA   -0.37563365
## ChestPainTypeNAP   -0.22214594
## ChestPainTypeTA    -0.05376515
## RestingBP           0.17324158
## Cholesterol         0.10386560
## FastingBS           0.16059410
## RestingECGNormal   -0.13325459
## RestingECGST        0.09592107
## MaxHR              -0.37721219
## ExerciseAnginaY     0.55183450
## Oldpeak             0.49569625
## ST_SlopeFlat        0.59155358
## ST_SlopeUp         -0.65375874
## HeartDisease        1.00000000
```

None of the predictor variables have a correlation > 0.5 (moderate) other than one hot encoded variables within the same category.

What are the real-world implications for any multicollinearity you discover?

Most of the relatively high correlation values are between ExerciseAngina, OldPeak, and ST_Slope. These variables may be more advanced and associated with one another, however their correlations are all less than 0.5 thus it's not worth holding them out of feature selection. One of the few pairs of features outside these three that have a correlation magnitude above 0.35 is age and maximum heart rate, which corroborates domain knowledge, but once again the correlation isn't strong enough to assume they're redundant.