1 Introduction
2 Data Description (Heartdata) & Preprocessing
3 Methods
4 Discussion of Model and Results
5 Conclusion
6 EndNotes
7 Bibliography

Using Support Vector Machines for Binary Classification to Predict Heart Disease

Capstone Project for Harvardx Professional Data Science Certificate

Mike Bryant

March 24, 2019

Abstract

Abstract: Support vector machines (SVM) are powerful machine learning models for binary classification problems. In this exploratory paper, two types of SVM models are investigated: a linear kernal and radial kernal SVM on a UC-Irving heart disease data set with the purpose of predicting if a patient has heart disease or not. The models will be further refined based on gender. The performance of the models is shown in the table below:

Model	cost	gamma	accuracy on test set
SVM linear(all gender)	1	na	90.0%
SVM radial(all gender)	1	.1	71.7%
SVM linear(female)	1	na	67.5%
SWM radial(female)	2	.1	72.58%
SVM linear(male)	1	na	88.9%
SVM radial(male)	2	.1	94.4%

1 Introduction

Heart disease is a leading cause of death in the United States and many parts of the world. As a general rule, early diagnosis of disease improves patient outcomes. For heart disease, early diagnosis may help delay or even prevent progression to heart failure[1]. Machine learning techniques can be used to analyze patient data and create predictive models to help determine if a patient may have a disease or not allowing doctors to diagnose patients earlier in the disease process.

The purpose of this project is to analyze the heart disease patient dataset obtained from the University of California-Irving (see bibliography for link) and create a predictive model to determine if a patient has heart disease based on physiological features that will be described later in the paper. A support vector machine model will be investigated. The application of a predicitive model like this could be used to determine if a patient has heart disease even in the absence of symptoms or suspicion.

The focus of this report is to follow the data science process of reviewing and processing the data, building and analyzing the model and interpreting the outcome. The mathematics of this type of model is far beyond the scope of this report. If the reader is interested in learning the fundemental math of this type of model, I would direct them to MIT OpenCourseWare lectures portal, titled "16. Learning:Suport Vector Machines". Please refer to the bibliography section of this paper for a link.

Support vector machines are nonlinear predictive models that are used in classification problems. Heart disease prediction is a classification problem (either has heart disease or does not have heart disease), therefore a support vector machine is an appropriate model to explore. Essentially, a support vector machine creates a hyperplane (decision) boundry to separate data points, making a distinct boundary that separates the data points into two groups (binary classification). It does so by maximzing the boundry space between the two groups as measured by the distance of vectors (support vectors) from the boundry[2]. Additionaly, spacial transformations may be required, such as changing the linear space into a spherical space (radial coordinates) or other spacial transformation. Transformations applied to the data are commonly referred to as kernals[3]. Two common transformations(kernals) will be explored in this report: the linear & radial kernals.

2 Data Description (Heartdata) & Preprocessing

The dataset consists of 14 physological patient attributes as follows:

1. Attribute Information (variable name in the data set is in quotes below):
   - age ("age", continuous)
   - sex("sex"-categorical, 0=male, 1=female)
   - chest pain type (4 values) ("cp"-ordinal/categorical)
   - resting blood pressure ("trestbps"-continuous)
   - serum cholestoral in mg/dl ("chol"-continuous)
   - fasting blood sugar > 120 mg/dl ("fbs"-categorical is >120 =1 or <120 =0)
   - resting electrocardiographic results ("restecg"-values 0,1,2, categorical)
   - maximum heart rate achieved ("thalack"-continuous)
   - exercise induced angina ("exang", categorical, 1=angina 0 = no agina)
   - oldpeak = ST depression induced by exercise relative to rest ("oldpeak"- continuous)
   - the slope of the peak exercise ST segment ("slope"-continuous)
   - number of major vessels (0-3) colored by flourosopy (factor)
   - thal: 3 = normal; 6 = fixed defect; 7 = reversable defect ("thal"-categorical)
   - target: 1= heart disease present; 0 = no heart disease ("target"-categorical, dependant variable)

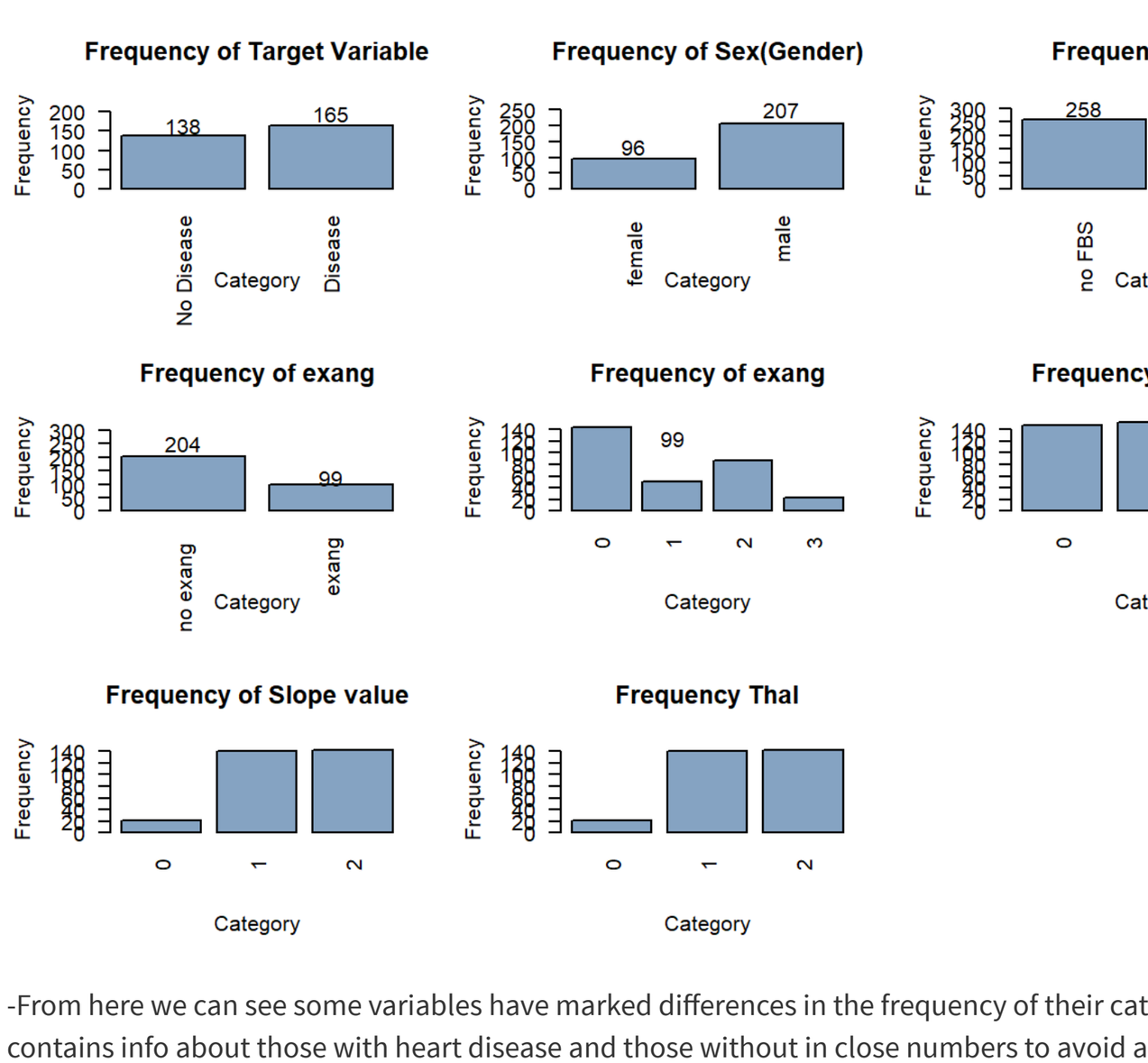
2.1 Attribute Statistcs

Basic statistics about the data are obtained in the below table:

```
##      Ã¯..age          sex               cp            trestbps    
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0  
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0  
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0  
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6  
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0  
##       chol            fbs            restecg          thalach     
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0  
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5  
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0  
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6  
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0  
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0  
##      exang           oldpeak         slope             ca        
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000  
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294  
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000  
##       thal           target      
##  Min.   :0.000   Min.   :0.0000  
##  1st Qu.:2.000   1st Qu.:0.0000  
##  Median :2.000   Median :1.0000  
##  Mean   :2.314   Mean   :0.5446  
##  3rd Qu.:3.000   3rd Qu.:1.0000  
##  Max.   :3.000   Max.   :1.0000
```

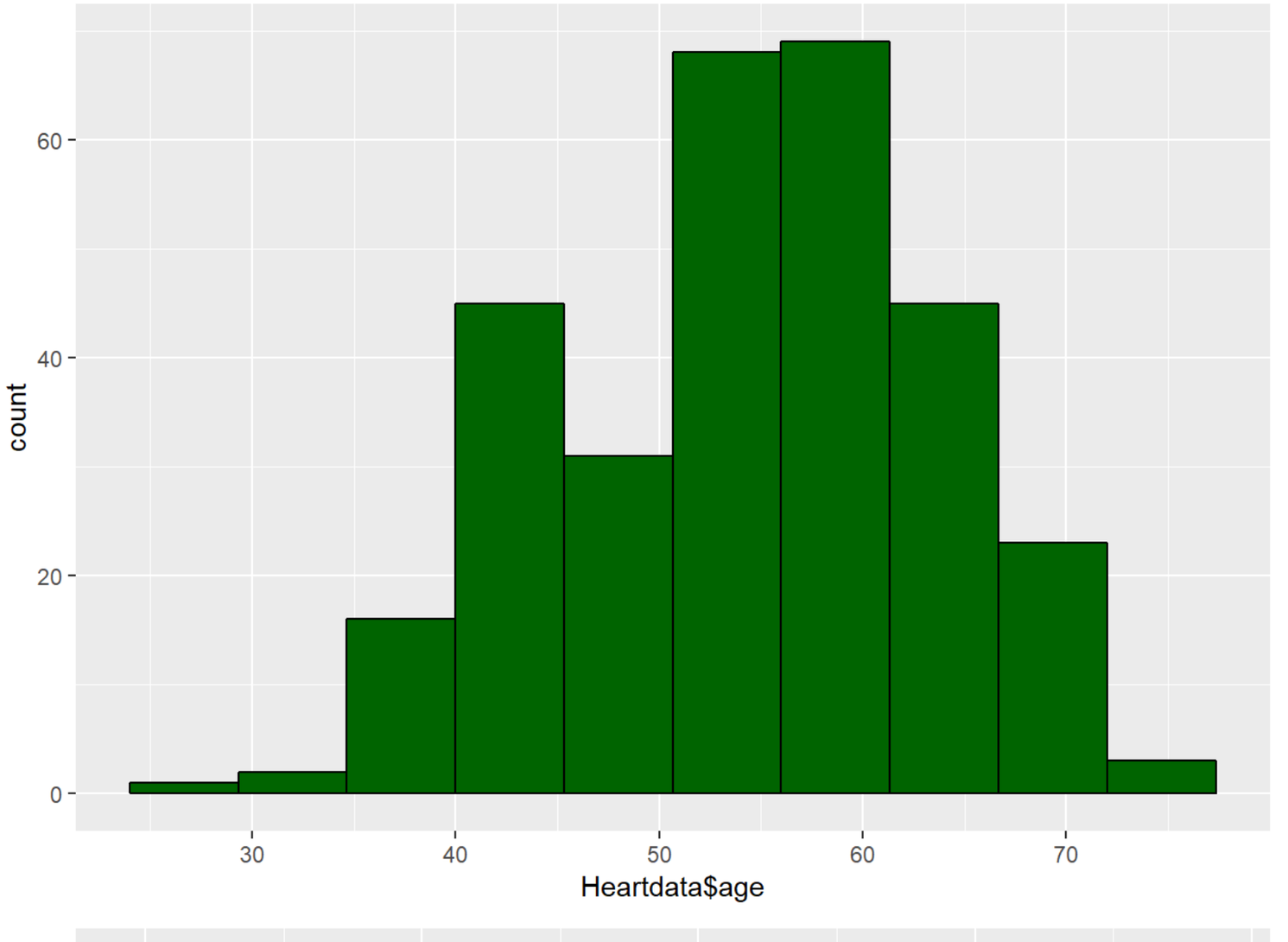
From the summary, we can conclude there are no common issues with unclean data. There are no "N/A" values and no negative values where one would not expect to see them. The summary function in R would show those if they existed in the data.

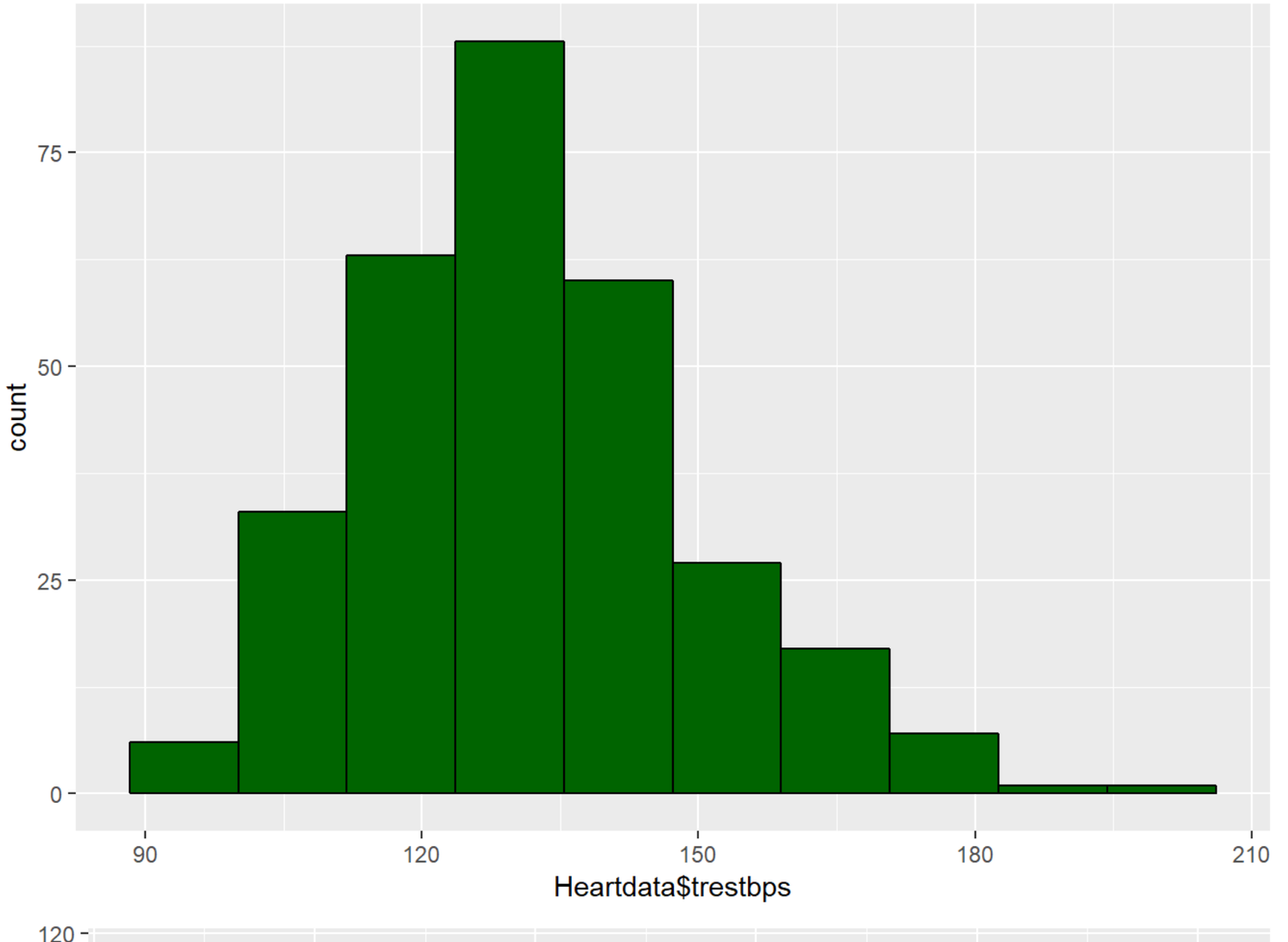
2.2 Visual exploration of Categorical Variables

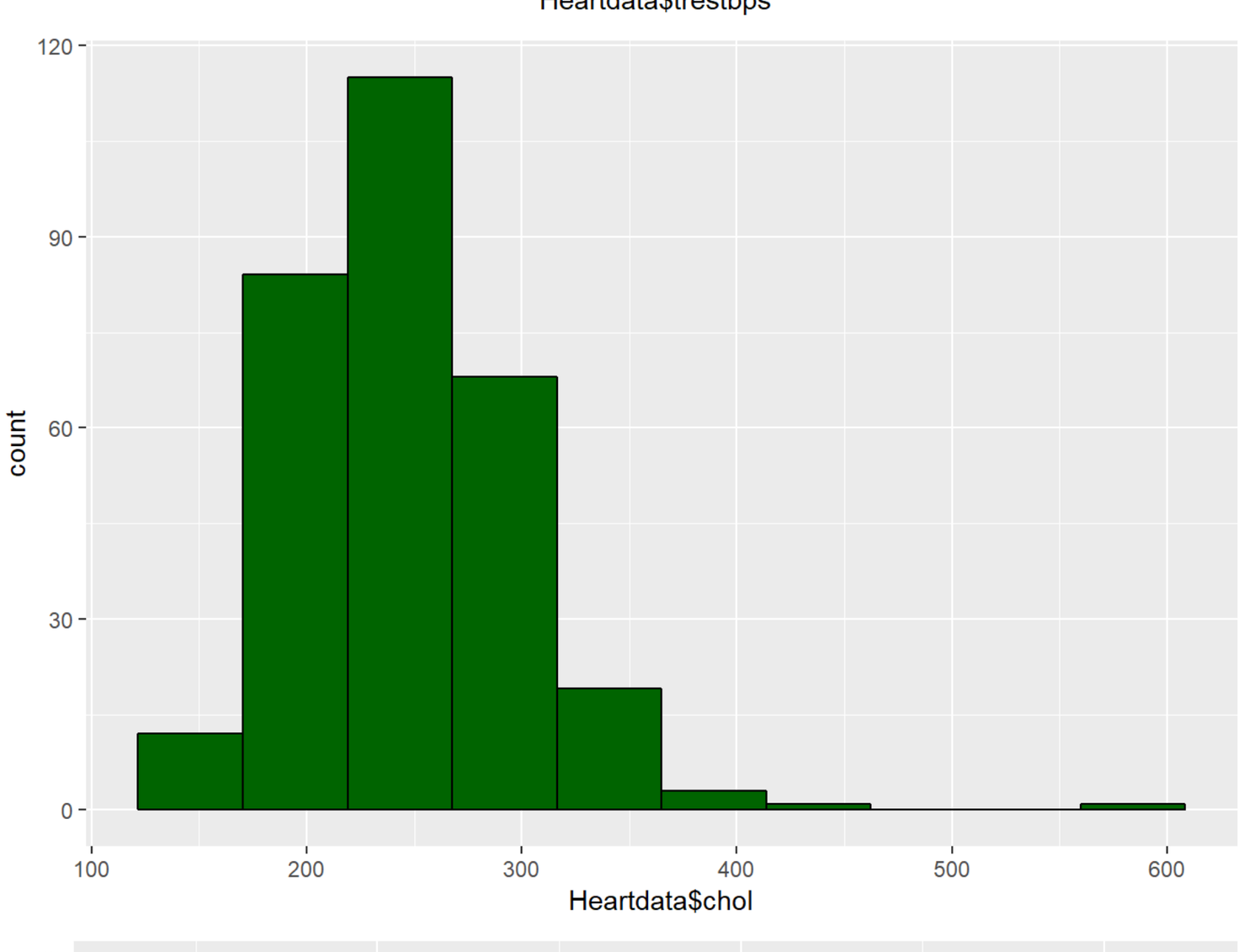


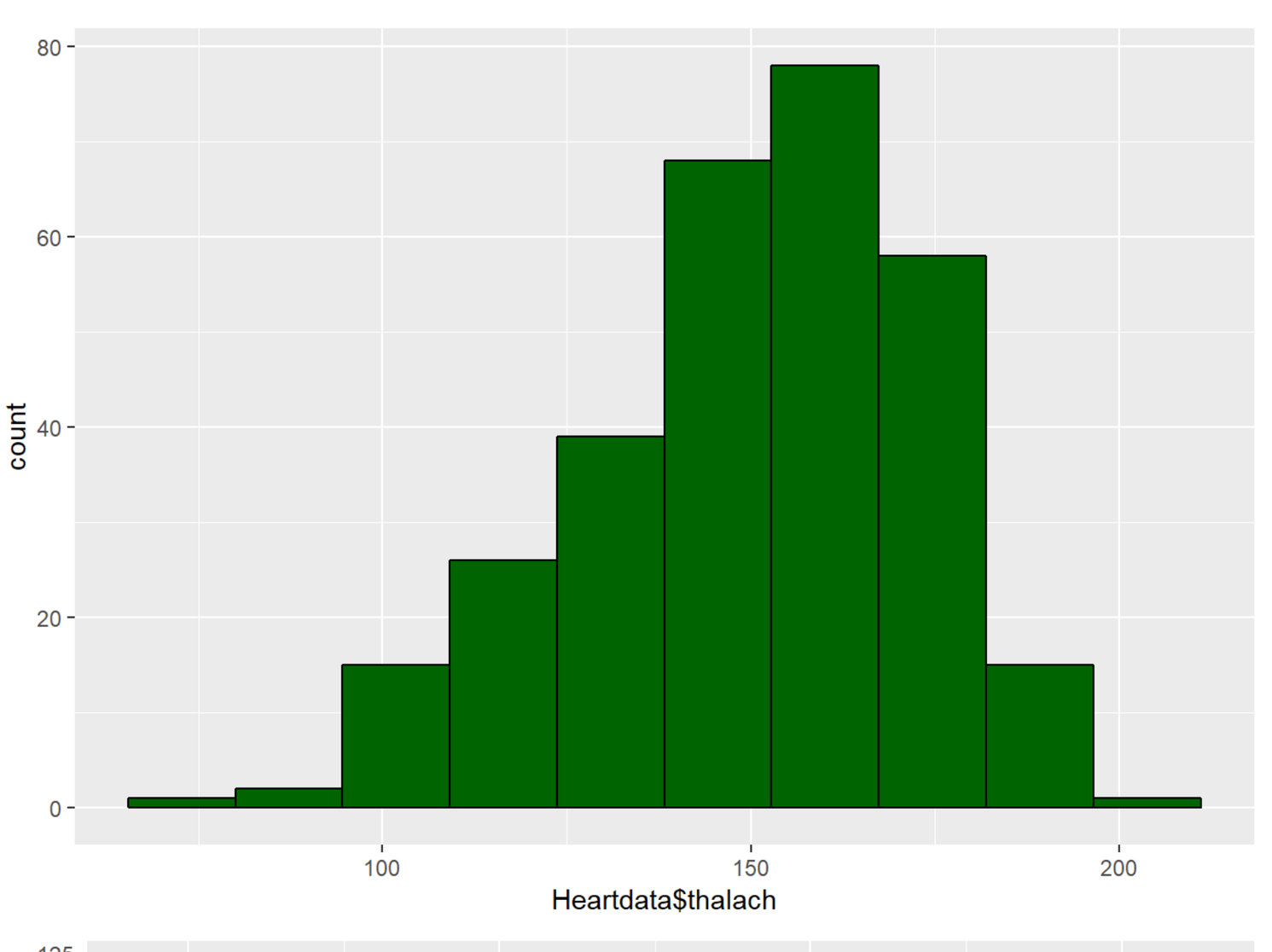
-From here we can see some variables have marked differences in the frequency of their categories. It is good that the data contains info about those with heart disease and those without in close numbers to avoid accuracy paradox. Interestingly, the number of males to females varies drastically and future data collection may want to try and obtain an equal number of males and females to accurately represent the population.

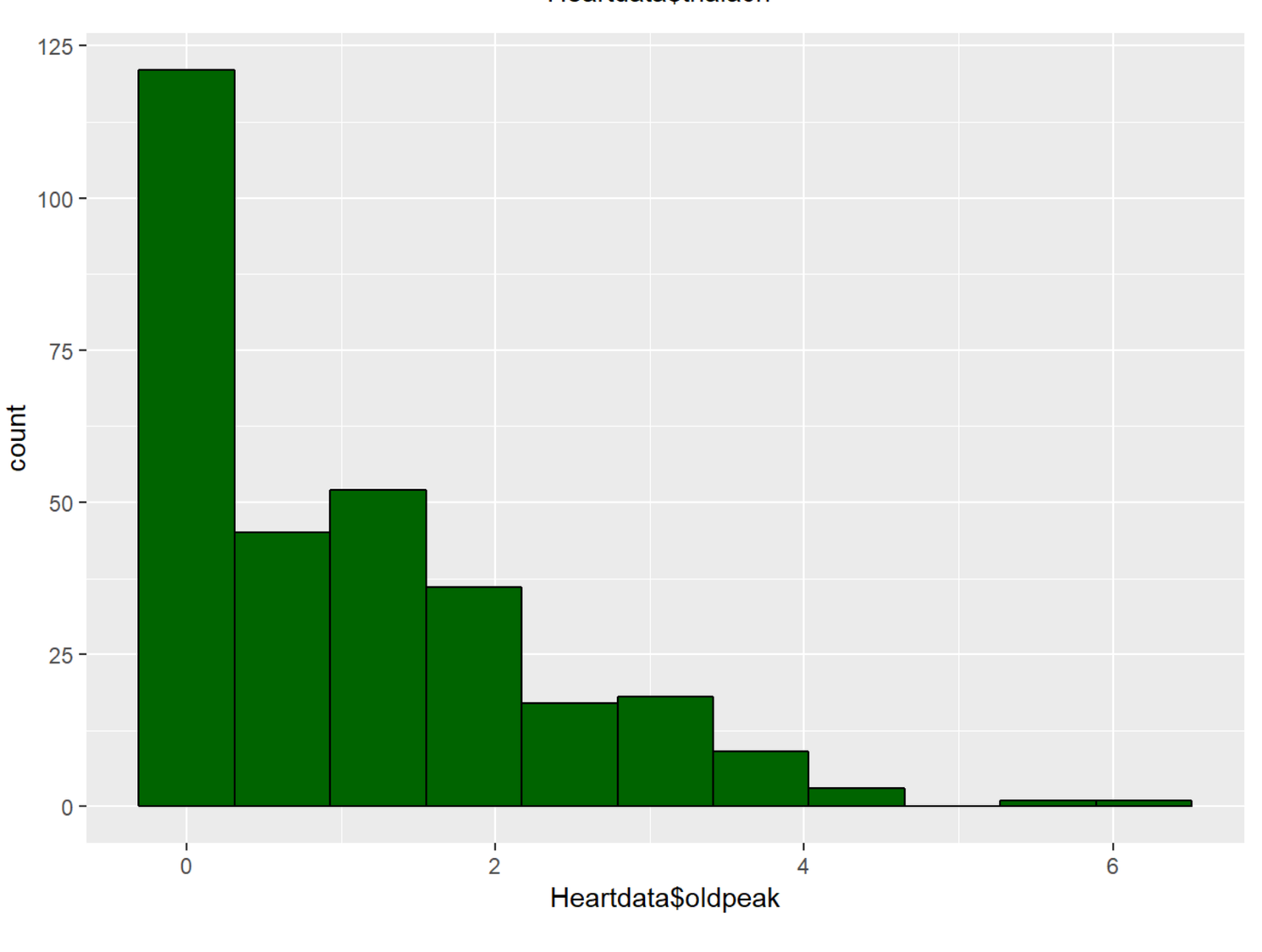
2.3 Visual exploration of Categorical Variables











2.4 Summary of Data and Data processing

After review we can conclude there are no common issues with unclean data. There are no "N/A" values and no negative values where one would not expect to see them. There are a few points in some of the features that may be considered outliers, such as the values exceeding 500 in the cholesterol data. However, for this report all the data will be inlcuded.

Some of the data is not normally distributed, but this does not affect SVM models. Also, some of the data values are clearly on different magnitudes, but the e1071 package that will be used to create SVM type models resolves this issue. This is explained in the proceeding section.

3 Methods

The SVM package e1071 will be used to create SVM models. This package standarizes all feature values to be on the same scale (zero mean & unit variance), this addresses featuers with larger magnitudes drowning out the effect of features with smaller magnitudes[4]. Supporting the e1071 package are the caret package and purrr package. All package documentation can be viewed on the Cran server by searching the package name (see bibliography for link).

Using these packages, test and training models will be created;tuning parameters of the SVM models (cost and gamma parameters) will be tuned using K-fold cross Validation (built into the e1071 package using the tune argument, k-cross is defaulted unless another is specified). Linear and radial kernals will be tested for all the data. Other types of kernals, such as polynomial or sigmoid kernals, will not be explored but could serve as further research.

Finally, the data will be split into gender groups and SVM models will be created using K-fold cross validation. This is explored due to the greater number of males than females in the data set and because of the obvious physological differences between males and females, it may be better to have seperate models for the genders. Exploring this idea in more detail could also be a subject for further research.

4 Discussion of Model and Results

4.1 Intial SVM model with Arbitrary Parameter values

The initial model is run with arbitrary tuning parameters and results are found using the following code:

```
library(caret)
library(e1071)
library(purrr)#support vector machine package
Heartdata<-read.csv(file="heart.csv", header = TRUE)
Heartdata$target<-as.factor(Heartdata$target) #ensures dependant variable is a factor with 2 levels, a
nd not an integer

set.seed(1)
Heartdata_sampling_vector <- createDataPartition(Heartdata$target, p=0.8, list=FALSE)
Heartdata_train<- Heartdata[Heartdata_sampling_vector,]
Heartdata_test<- Heartdata[-Heartdata_sampling_vector,]

set.seed(1)
Linearfirst<-svm(target ~., data=Heartdata_train, kernel = "linear", cost=10)
  Linear_Kernal_acc<-confusionMatrix(Linearfirst$fitted, Heartdata_train[,"target"])$overall[1]
test_predictions<-predict(Linearfirst, Heartdata_test[,c(1:13)])
 First_pass_linear<-mean(Heartdata_test[,14] == test_predictions)


set.seed(1)
Radialfirst<-svm(target ~., data=Heartdata_train, kernel = "radial", cost=10, gamma=0.5)
  Radial_Kernal_acc<-confusionMatrix(Radialfirst$fitted, Heartdata_train[,"target"])$overall[1]
test_predictions<-predict(Radialfirst, Heartdata_test[,c(1:13)])
 First_pass_Radial<-mean(Heartdata_test[,14] == test_predictions)

Firstpassresults<-data.table(model_name=c("Linear Kernal", "Radial Kernal"), trainset_accuracy=c(Linea
r_Kernal_acc,Radial_Kernal_acc),Test_set_accuarcy=c(First_pass_linear,First_pass_Radial), Cost=c(Linea
rfirst$cost,Radialfirst$cost), Gamma=c("na", Radialfirst$gamma))

Firstpassresults
```

```
##       model_name trainset_accuracy Test_set_accuarcy Cost Gamma
## 1: Linear Kernal         0.8312757         0.9000000   10    na
## 2: Radial Kernal         1.0000000         0.7166667   10   0.5
```

The radial kernal produced the most accurate results on this first pass for the training section. However, for the test predictions the linear model produced the more accurate predictions.

The next step is to tune the parameters for the model. In this particular case we have a cost parameter and a gamma parameter (relevant only for radial kernal). This report will not go into detail about these parameters as it is beyond the scope of this project.

4.2 K-fold Cross Validation to Determine Tuning Parameters

K-fold cross validation is a procedure defined as "randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k folds[5]." In other words, it creates many training sets to train the model and picks the tuning parameter based on the best performance as measured against a validation set from the data, in this case the training set. Therefore, parameters will be chosen ultimately using the training data (and a validation set within the training data), and then an external test set will be used to measure the accuracy of the model.

The e1071 package has a built-in k-fold cross validation method that will choose the best tuning parameters (gamma, cost).

The code to run such a model and its output is as follows:

```
Heartdata<-read.csv(file="heart.csv", header = TRUE)
set.seed(1)
Heartdata$age<-Heartdata$Ã¯..age
Heartdata<-Heartdata[,-1]
Heartdata<-Heartdata[,-1]
Heartdata$target<-as.factor(Heartdata$target)
Heartdata_sampling_vector <-createDataPartition(Heartdata$target, p=0.8, list=FALSE)
Heartdata_train<- Heartdata[Heartdata_sampling_vector,]
Heartdata_test<- Heartdata[-Heartdata_sampling_vector,]



set.seed(1)
yes<-seq(0,1,.1)
tuneradial<-tune(svm, target ~., data=Heartdata_train, kernal="radial", ranges=list(cost=seq(1:10), ga
mma= yes))

set.seed(1)
tunelinear<-tune(svm, target ~.,data=Heartdata_train, kernal="linear", ranges=list(cost=seq(1:10)))

set.seed(1)
model_SVM<-svm(target ~., data=Heartdata_train, kernel = "linear", cost=tunelinear$best.parameters$cos
t)

test_predictions<-predict(model_SVM, Heartdata_test[,c(1:11,13)])
TEST_SET_result_linear<-mean(Heartdata_test[,12] == test_predictions)

set.seed(1)
model_SVM<-svm(target ~., data=Heartdata_train, kernel = "radial", cost=tuneradial$best.parameters$cos
t, gamma=tuneradial$best.parameters$gamma)

 test_predictions<-predict(model_SVM, Heartdata_test[,c(1:11,13)])
 TEST_SET_result_radial<-mean(Heartdata_test[,12] == test_predictions)


Bestparameters<-data.table(model_name=c("Linear Kernal SVM", "Radial Kernal SVM"),Performance_train_se
t=c(1-tunelinear$best.performance, 1-tuneradial$best.performance), Cost=c(tunelinear$best.parameters$c
ost, tuneradial$best.parameters$cost),Gamma=c( "NA", tuneradial$best.parameters$gamma),Performance_tes
t_set=c(TEST_SET_result_linear,TEST_SET_result_radial))

Bestparameters
```

```
##           model_name Performance_train_set Cost Gamma Performance_test_set
## 1: Linear Kernal SVM             0.8183333    1    NA            0.9000000
## 2: Radial Kernal SVM             0.8141667    1   0.1            0.8166667
```

The linear kernal model had the best perfomrance with an 90% accuracy on the training set using cross validation technique with the associated found tuning parameters. Its performance on the test set is far greater than its performance on the training set, while for the radial kernal, the performance is about the same. Therefore, if one were to pick a model for all genders to predict heart disease it would be the Linear Kernal model.

4.3 Exploring Gender Difference in the SVM models

As seen earlier, there are many more males than females in the dataset. The following code produces an ouput the compares outcomes of the sVM models for each gender.

```
## 
## Parameter tuning of 'svm':
## 
## - sampling method: 10-fold cross validation 
## 
## - best parameters:
##  cost gamma
##     2   0.1
## 
## - best performance: 0.2091912
```

```
## 
## Parameter tuning of 'svm':
## 
## - sampling method: 10-fold cross validation 
## 
## - best parameters:
##  cost
##     2
## 
## - best performance: 0.1856618
```

```
##                  model_name Performance_on_train_set Cost Gamma
## 1:  Linear Kernal SVM Female                0.8143382    2    NA
## 2: Radial Kernal SVM Female                0.7908088    2   0.1
## 3:    Linear Kernal SVM Male                0.9071429    2    NA
## 4:    Radial Kernal SVM Male                0.9071429    2   0.1
##    Performance_on_test_set
## 1:               0.6750000
## 2:               0.7250000
## 3:               0.8888889
## 4:               0.9444444
```

The table above shows a stark contrast between male and female SVM models of both types (radial and kernal). Surprisingly, performance on the training set for linear and radial male models was the same, but the performance on the test set was more accurate for the radial kernel. The above table shows that the radial kernal has better accuracy on test sets for radial models and similar performance on the training set.

5 Conclusion

This report explored support vector machine models (SVM models) for the purposes of building a model to predict heart disease. The following process was executed:

1. Process Review:
   - visually explore features and obtain basic summary statistics
   - Use e1071 package to build SVM model using two types of kernals, linear and radial using arbitrary tunning values for cost and gamma parameters
   - Use e1071 package to execute K-fold cross validation and automatically pick the best tunning parameters, k=10 as default value
   - Compare accuracy of the linear and radial kernals
   - Compare accuracy of each type of model between gender groups (male and female)

The accuracy of the models tested can be sumarized in the following table:

Model	cost	gamma	accuracy on test set
SVM linear(all gender)	1	na	90.0%
SVM radial(all gender)	1	0.2	81.7%
SVM linear(female)	1	na	67.5%
SWM radial(female)	2	.1	72.58%
SVM linear(male)	1	na	88.9%
SVM radial(male)	2	.1	94.4%

From the table, it can be seen that the SVM Radial model for the male gender is the most accurate model. Males dominate the data set (68.3% of total data), and as an improvement to future research, more data on females should be collected. The linear model for all genders produced decent results too, with 90% predictive accuracy on the test set.

To improve upon the models built in this project the following could be executed: feature engineering (i.e a new variable that is a ratio of two current variables), adding more features (like alcohol usage); different models could also be explored such as an SVM using a polynomial kernal, or multilogistic regression. Also, combination models could be explored (decision by committee). The reader is encouraged to explore these suggestions to improve upon the predictive outcomes generated in this project.

6 EndNotes

1.Heart Failure Fact Sheet|Data & Statistics|DHDSP|CDC." Centers for Disease Control and Prevention. Accessed May 1, 2019. https://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_heart_failure.htm.

2.Forte, Rui Miguel. *Mastering Predictive Analytics with R*, (Packt Publishing, 2015), 164-166.

3.Forte, Rui Miguel. *Mastering Predictive*, 172-173.

4.Forte, Rui Miguel. *Mastering Predictive*, 175.

5.A Gentle Introduction to K-fold Cross-Validation." Machine Learning Mastery. May 08, 2019. Accessed May 1, 2019. https://machinelearningmastery.com/k-fold-cross-validation/.

7 Bibliography

1."A Gentle Introduction to K-fold Cross-Validation." Machine Learning Mastery. May 08, 2019. Accessed May 9, 2019. https://machinelearningmastery.com/k-fold-cross-validation/.

2.Heart Failure Fact Sheet|Data & Statistics|DHDSP|CDC." Centers for Disease Control and Prevention. Accessed May 1, 2019. https://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_heart_failure.htm.

3."Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [R Package E1071 Version 1.7-1]." The Comprehensive R Archive Network. Accessed May 1, 2019. https://cran.r-project.org/web/packages/e1071/.

4."UCI Machine Learning Repository: Heart Disease Data Set." Accessed March 26, 2019. https://archive.ics.uci.edu/ml/datasets/heart disease.

5.Henry, Patrick. "Lecture 16: Learning: Support Vector Machines." MIT OpenCourseWare, Massachusetts Institute of Technology. Accessed May 1, 2019. https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-034-artificial-intelligence-fall-2010/lecture-videos/lecture-16-learning-support-vector-machines/.

5.Forte, Rui Miguel. 2015. *Mastering Predictive Analytics with R.* Packt Publishing.