

# 210501867 Script

Erika Sim

2023-12-27

## Download all the datasets here

<https://drive.google.com/drive/folders/1m5xdYpjSILphWonXFQVWellzWCfUn34N?usp=sharing>

## Setting working directory

```
setwd("C:/Users/Admin/Desktop/datasets")
```

## Loading library

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
##      filter, lag
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##      intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.3
```

```
##
```

```
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':  
##  
##      select
```

```
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.1.3
```

```
##  
## Attaching package: 'ISLR2'
```

```
## The following object is masked from 'package:MASS':  
##  
##      Boston
```

```
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 4.1.3
```

```
library(Metrics)
```

```
## Warning: package 'Metrics' was built under R version 4.1.3
```

```
library(magrittr)
```

```
##  
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:tidyr':  
##  
##      extract
```

```
library(fastDummies)
```

```
## Thank you for using fastDummies!
```

```
## To acknowledge our work, please cite the package:
```

```
## Kaplan, J. & Schlegel, B. (2023). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.1.3
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##  
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':  
##  
##      combine
```

```
library(xgboost)
```

```
## Warning: package 'xgboost' was built under R version 4.1.3
```

```
##  
## Attaching package: 'xgboost'
```

```
## The following object is masked from 'package:dplyr':  
##  
##      slice
```

```
library(cluster)  
library("ggrepel")
```

```
## Warning: package 'ggrepel' was built under R version 4.1.3
```

```
## Loading required package: ggplot2
```

```
##  
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':  
##  
##      margin
```

```
library("factoextra")
```

```
## Warning: package 'factoextra' was built under R version 4.1.3
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.1.3
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.3
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'

## The following objects are masked from 'package:Metrics':
##
##   precision, recall
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.1.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following object is masked from 'package:Metrics':
##
##   auc
```

```
## The following objects are masked from 'package:stats':
##
##   cov, smooth, var
```

```
library(magrittr)
library(scales)
```

```
## Warning: package 'scales' was built under R version 4.1.3
```

```
library(igraph)
```

```
## Warning: package 'igraph' was built under R version 4.1.3
```

```
##
## Attaching package: 'igraph'
```

```
## The following object is masked from 'package:tidyr':
##
##   crossing
```

```
## The following objects are masked from 'package:dplyr':
##
##   as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##   decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##   union
```

```
library(ggplot2)
library("ggpubr")
```

```
## Warning: package 'ggpubr' was built under R version 4.1.2
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.1.3
```

```
## corrplot 0.92 loaded
```

```
library(RColorBrewer)
```

## Regression Analysis

*Loading and cleaning of dataset*

```
movies <- read.csv("movies.csv")
summary.default(movies)
```

```
##           Length Class  Mode
## name      7658   -none- character
## genre      7658   -none- character
## year       7658   -none- numeric
## score      7658   -none- numeric
## votes      7658   -none- numeric
## budget     7658   -none- numeric
## gross      7658   -none- numeric
## runtime    7658   -none- numeric
## company    7658   -none- character
```

```
sum(is.na(movies))
```

```
## [1] 2366
```

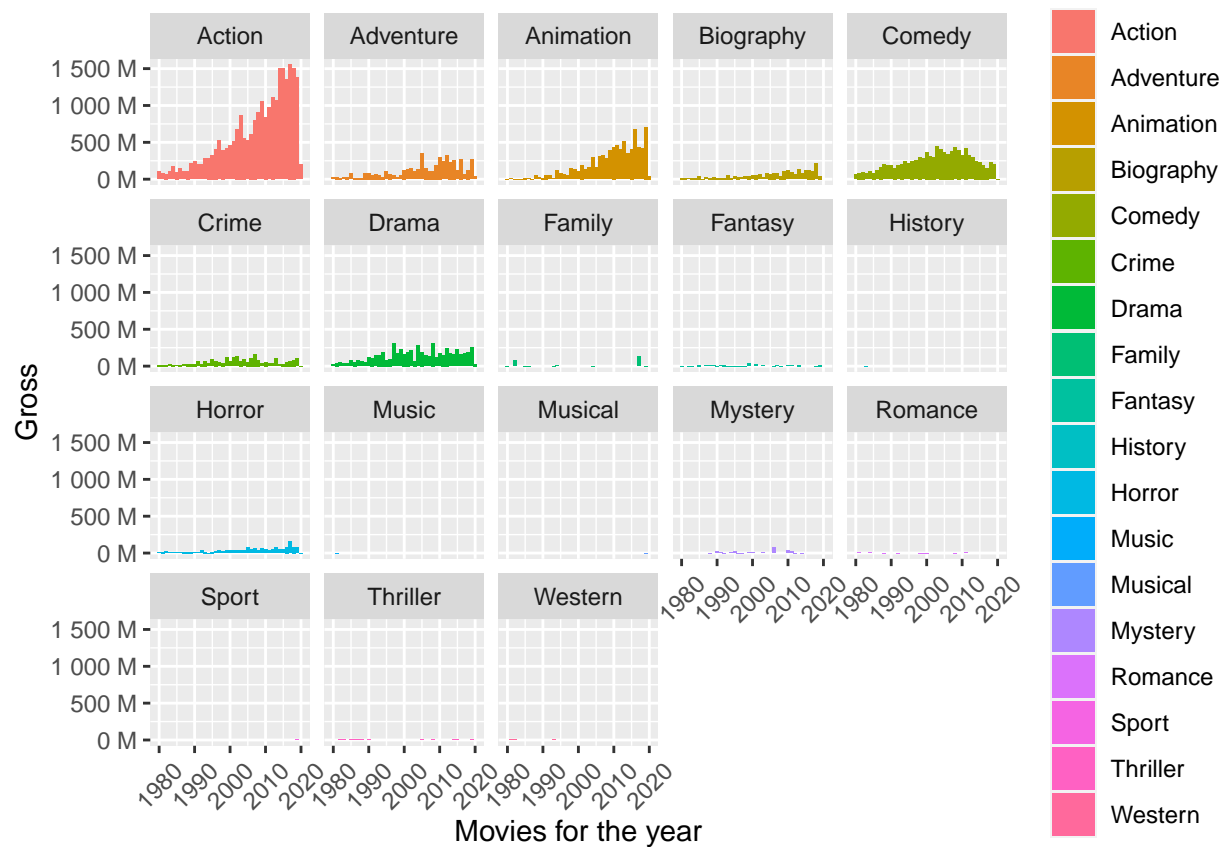
```
sum(duplicated(movies))
```

```
## [1] 0
```

```
movies[is.na(movies)] <- 0
```

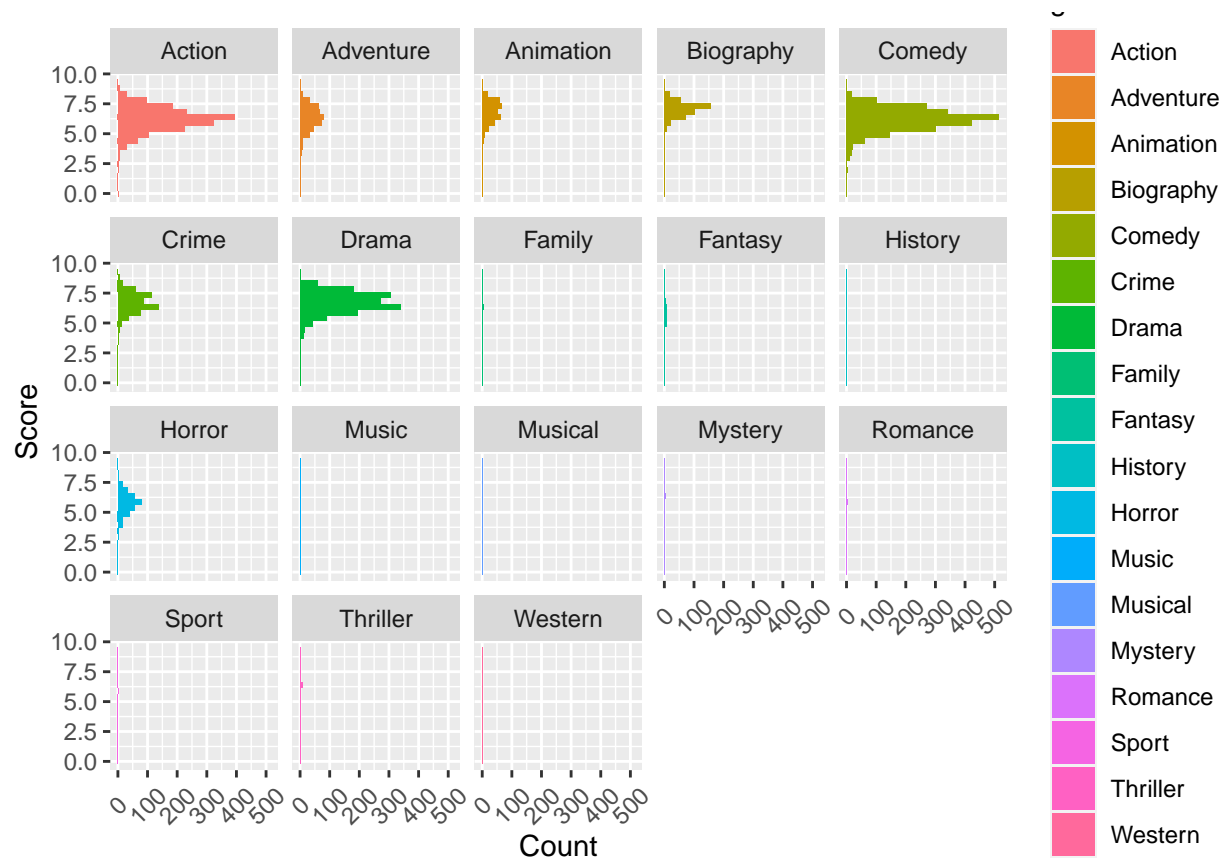
*Exploring data*

```
ggplot(data = movies, aes(x = year, y = gross, fill = genre)) +
  geom_col() +
  xlab("Movies for the year") +
  ylab("Gross") +
  scale_y_continuous(labels = unit_format(unit = "M", scale = 1e-7)) +
  facet_wrap("genre") +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=0.5))
```



Gross by Genre from 1980 to 2020. Action genre made the most revenue while Western made the least.

```
ggplot(data = movies, aes(y = score, fill = genre)) +
  geom_histogram(bins = 20) +
  xlab("Count") +
  ylab("Score") +
  facet_wrap("genre") +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=0.5))
```

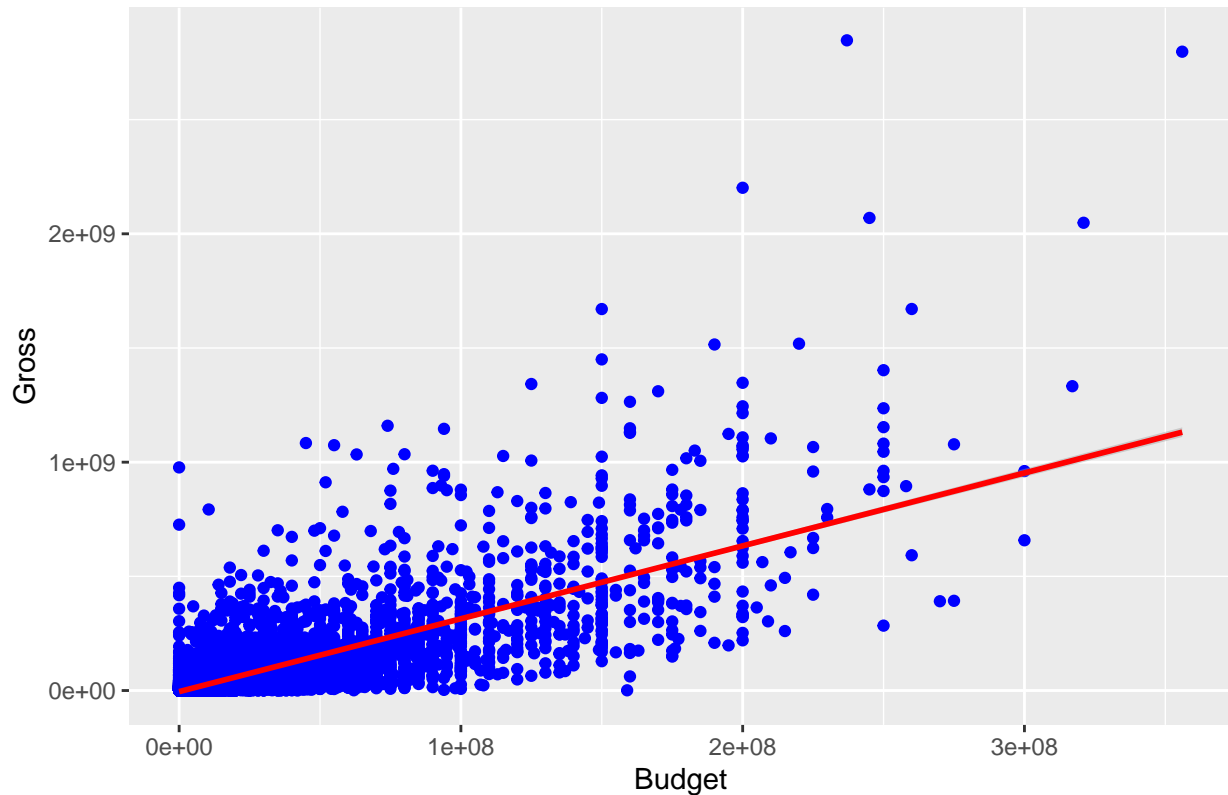


Score by Genre from 1980 to 2020. Comedy genre had the highest score followed by Action genre.

```
ggplot(data = movies, aes(x = budget, y = gross)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red") +
  xlab("Budget") +
  ylab("Gross") +
  labs(title = "Correlation between budget and revenue")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

## Correlation between budget and revenue



Correlation between budget and gross. This showed that there was a positive correlation between the two variables as seen with the regression line.

*Setting dataset into test and train sets* Removing name, year and company from this model as there will not be any significance for these categorical variables. Only exception would be genre.

```
set.seed(5)
movies2 <- subset(movies, select= -c(name, year, company)) # removal of insignificant categorical variables
moviesf <- dummy_cols(movies2, select_columns = c("genre"), remove_selected_columns = TRUE, remove_first_level = FALSE)
movies_set <- sample.split(moviesf, 0.8)
movies_train <- subset(moviesf, movies_set == TRUE)
movies_test <- subset(moviesf, movies_set == FALSE)
```

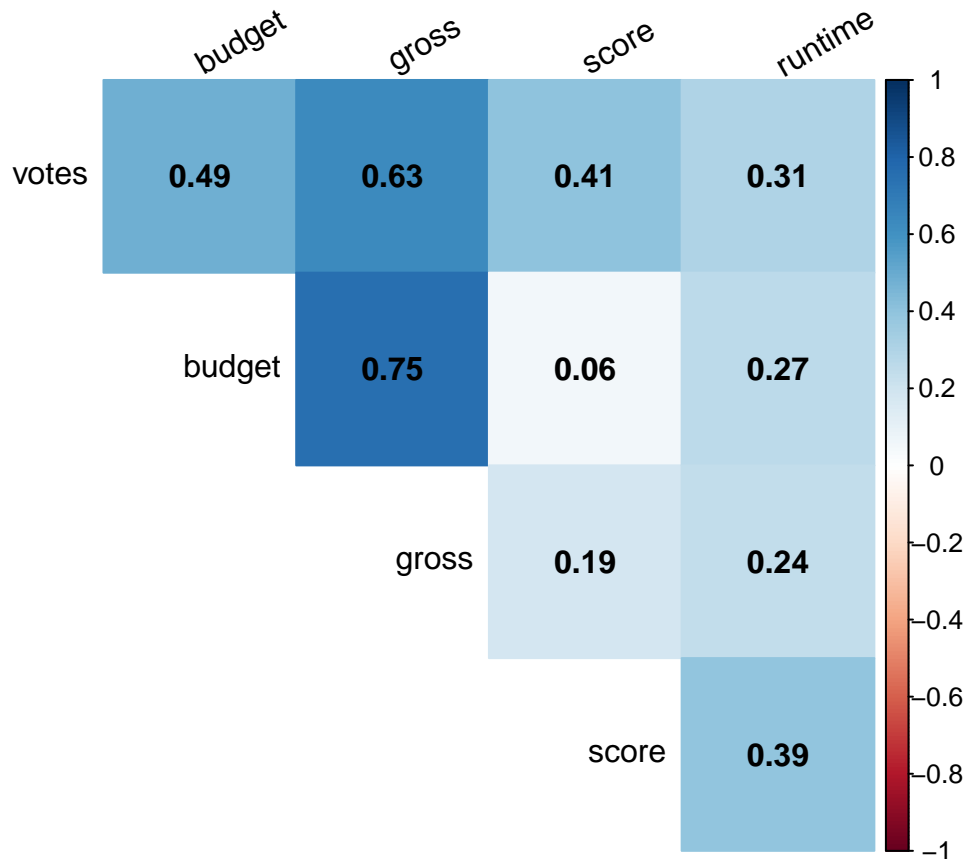
## Correlation Matrix

```
movies3 <- subset(movies, select= -c(name, year, company, genre)) # removal of all categorical variables
rcorr_matrix <- cor(movies3)
round(rcorr_matrix, 4)
```

```
##          score  votes budget  gross runtime
## score    1.0000  0.4068  0.0568  0.1858  0.3945
## votes    0.4068  1.0000  0.4873  0.6330  0.3071
## budget   0.0568  0.4873  1.0000  0.7503  0.2683
## gross    0.1858  0.6330  0.7503  1.0000  0.2442
## runtime  0.3945  0.3071  0.2683  0.2442  1.0000
```



```
corrplot(rcorr_matrix, type = "upper", order = "hclust",
         method = "color",
         tl.col = "black", tl.srt = 30,
         addCoef.col = "black",
         sig.level = 0.01,
         diag=FALSE)
```



We see from this correlation matrix that there is strong correlation between gross and budget.

Linear regression

We will make a dummy variable for genre.

```
lr <- lm(gross ~ . , data = movies_train)
summary(lr)
```

```
##
## Call:
## lm(formula = gross ~ . , data = movies_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -726820373 -30062077  5149112  22539302 1888561673
##
## Coefficients: (1 not defined because of singularities)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.497e+07  1.114e+07  -4.039 5.45e-05 ***
## score         1.392e+06  1.632e+06   0.853  0.3939
## votes         3.617e+02  9.993e+00  36.194 < 2e-16 ***
## budget        2.460e+00  4.308e-02  57.094 < 2e-16 ***
## runtime       7.395e+04  8.197e+04   0.902  0.3670
## genre_Adventure 1.020e+07  5.960e+06   1.711  0.0871 .
## genre_Animation 7.386e+07  6.988e+06  10.570 < 2e-16 ***
## genre_Biography -1.839e+06  6.235e+06  -0.295  0.7681
## genre_Comedy    1.507e+07  3.767e+06   4.000 6.42e-05 ***
## genre_Crime     -6.701e+06  5.547e+06  -1.208  0.2271
## genre_Drama     8.283e+06  4.214e+06   1.965  0.0494 *
## genre_Family    1.799e+08  3.258e+07   5.520 3.54e-08 ***
## genre_Fantasy   7.793e+06  1.747e+07   0.446  0.6556
## genre_History   NA         NA         NA      NA
## genre_Horror    2.438e+07  7.000e+06   3.483  0.0005 ***
## genre_Music     2.207e+07  9.743e+07   0.227  0.8208
## genre_Musical   2.547e+07  6.904e+07   0.369  0.7122
## genre_Mystery   -1.952e+07  2.382e+07  -0.819  0.4126
## genre_Romance   2.961e+06  3.692e+07   0.080  0.9361
## genre_Sport     3.063e+07  9.741e+07   0.314  0.7532
## genre_Thriller  2.256e+07  3.693e+07   0.611  0.5413
## genre_Western   2.260e+07  5.630e+07   0.401  0.6881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97360000 on 5897 degrees of freedom
## Multiple R-squared:  0.6627, Adjusted R-squared:  0.6616
## F-statistic: 579.4 on 20 and 5897 DF, p-value: < 2.2e-16
```

```
radj <- summary(lr)$adj.r.squared # 0.661584
lr_pred <- predict(lr, movies_test, type = "response")
```

```
## Warning in predict.lm(lr, movies_test, type = "response"): prediction from a
## rank-deficient fit may be misleading
```

```
lr_rmse <- rmse(lr_pred, movies_test$gross) # 87368985
lr_result <- c("Adjusted R Square" = radj, "RMSE" = lr_rmse)
```

Our objective here is to find out how is gross (dependent variable) affected by the other independent variables available. As the dummy variables were created, we can now see which genre has significant impact on gross and for the numerical variables as well. The adjusted R-squared is 0.6736.

Linear regression without runtime, score, and certain genres that was not significant (Better r-squared)

```
lr2 <- lm(gross ~ . -runtime -score -genre_Adventure -genre_Biography -genre_Crime -genre_Fantasy -genre
summary(lr2)
```

```
##
## Call:
## lm(formula = gross ~ . - runtime - score - genre_Adventure -
```

```
##      genre_Biography - genre_Crime - genre_Fantasy - genre_History -
##      genre_Music - genre_Musical - genre_Mystery - genre_Romance -
##      genre_Sport - genre_Thriller - genre_Western, data = movies_train)
##
## Residuals:
##      Min        1Q      Median        3Q      Max
## -730679821  -30363390   4977005   22413458  1887786626
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.864e+07  2.352e+06 -12.179  < 2e-16 ***
## votes        3.663e+02  8.865e+00  41.316  < 2e-16 ***
## budget       2.469e+00  3.983e-02  61.990  < 2e-16 ***
## genre_Animation 7.257e+07  6.421e+06  11.302  < 2e-16 ***
## genre_Comedy   1.448e+07  3.127e+06   4.631 3.72e-06 ***
## genre_Drama    9.082e+06  3.543e+06   2.563 0.010390 *
## genre_Family   1.794e+08  3.250e+07   5.518 3.57e-08 ***
## genre_Horror   2.267e+07  6.624e+06   3.423 0.000624 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97330000 on 5910 degrees of freedom
## Multiple R-squared:  0.6622, Adjusted R-squared:  0.6618
## F-statistic: 1655 on 7 and 5910 DF,  p-value: < 2.2e-16
```

```
radj2 <- summary(lr2)$adj.r.squared # 0.661801
lr2_pred <- predict(lr, movies_test, type = "response")
```

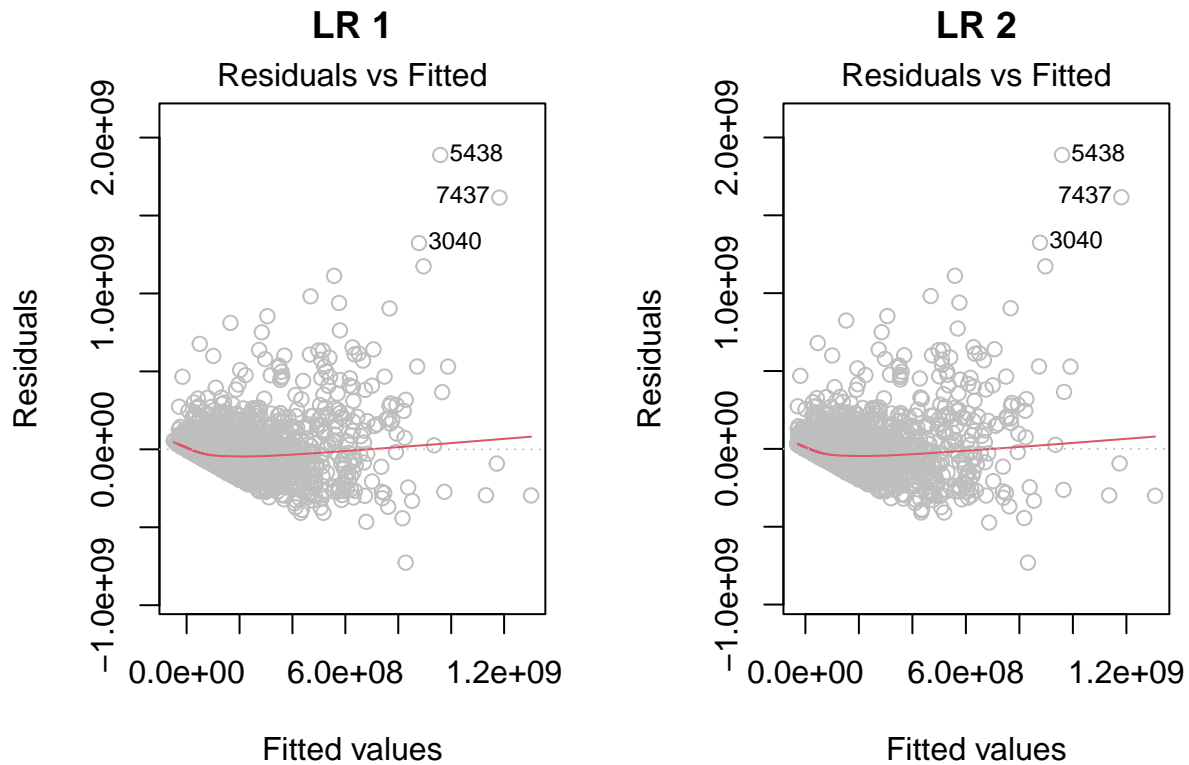
```
## Warning in predict.lm(lr, movies_test, type = "response"): prediction from a
## rank-deficient fit may be misleading
```

```
lr2_rmse <- rmse(lr_pred, movies_test$gross) # 87368985
lr2_result <- cbind("Adjusted R Square" = radj2, "RMSE" = lr2_rmse)
```

After removing the insignificant variables, the adjusted r-squared improved by 0.0002 at 0.6738 from previous model.

## Plots

```
par(mfrow=c(1,2))
plot(lr, c(1), main = "LR 1", col = "grey")
plot(lr2, c(1), main = "LR 2", col = "grey")
```



Non-linear transformation of predictors

```
qlr <- lm(gross ~ votes + I(votes^2) + budget + I(budget^2) + genre_Animation + genre_Comedy + genre_Drama)
summary(qlr)
```

```
##
## Call:
## lm(formula = gross ~ votes + I(votes^2) + budget + I(budget^2) +
##     genre_Animation + genre_Comedy + genre_Drama + genre_Family +
##     genre_Horror, data = movies_train)
##
## Residuals:
```

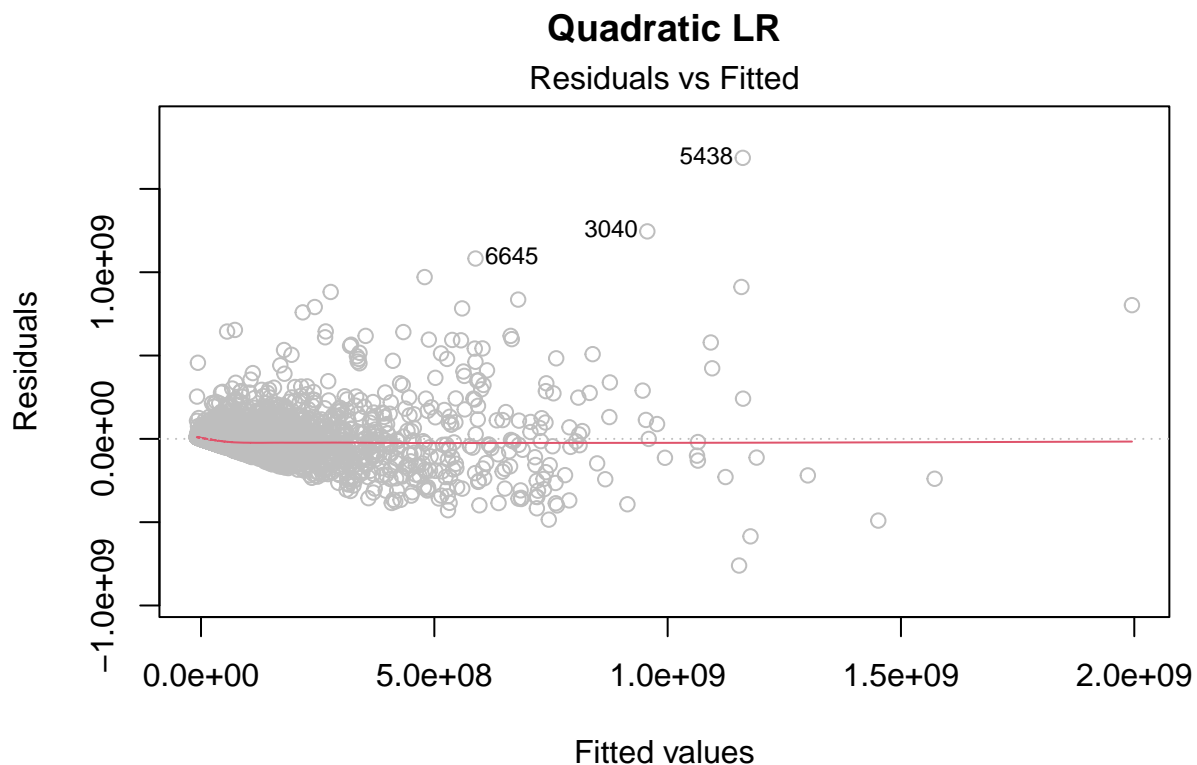
	Min	1Q	Median	3Q	Max
	-760150418	-19153775	-920646	11835068	1686257373

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-8.958e+06	2.399e+06	-3.734	0.000190 ***
votes	5.347e+02	1.573e+01	33.998	< 2e-16 ***
I(votes^2)	-1.539e-04	1.248e-05	-12.338	< 2e-16 ***
budget	4.832e-01	7.620e-02	6.341	2.45e-10 ***
I(budget^2)	1.164e-08	4.036e-10	28.827	< 2e-16 ***
genre_Animation	7.811e+07	5.982e+06	13.057	< 2e-16 ***
genre_Comedy	1.023e+07	2.918e+06	3.505	0.000461 ***
genre_Drama	1.512e+06	3.313e+06	0.456	0.648187

```
## genre_Family      1.608e+08  3.028e+07   5.310 1.14e-07 ***
## genre_Horror      1.133e+07  6.180e+06   1.833 0.066844 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 90630000 on 5908 degrees of freedom
## Multiple R-squared:  0.7072, Adjusted R-squared:  0.7067
## F-statistic: 1585 on 9 and 5908 DF,  p-value: < 2.2e-16
```

```
radj3 <- summary(qlr)$adj.r.squared # 0.7067
qlr_pred <- predict(qlr, movies_test, type = "response")
qlr_rmse <- rmse(qlr_pred, movies_test$gross) # 84423164
qlr_result <- cbind("Adjusted R Square" = radj2, "RMSE" = qlr_rmse)
plot(qlr, c(1), main = "Quadratic LR", col = "grey")
```



`lm(gross ~ votes + I(votes^2) + budget + I(budget^2) + genre_Animation + ge ...`

The non-zero p-value associated with the newly implemented quadratic term resulted in an improved model seen in the increased adjust R-squared.

Random Forest

```
set.seed(15)
rrf <- randomForest(gross ~ . , data = movies_train, type = "regression")
rrf # r-squared = 73.03%
```

```
##
```

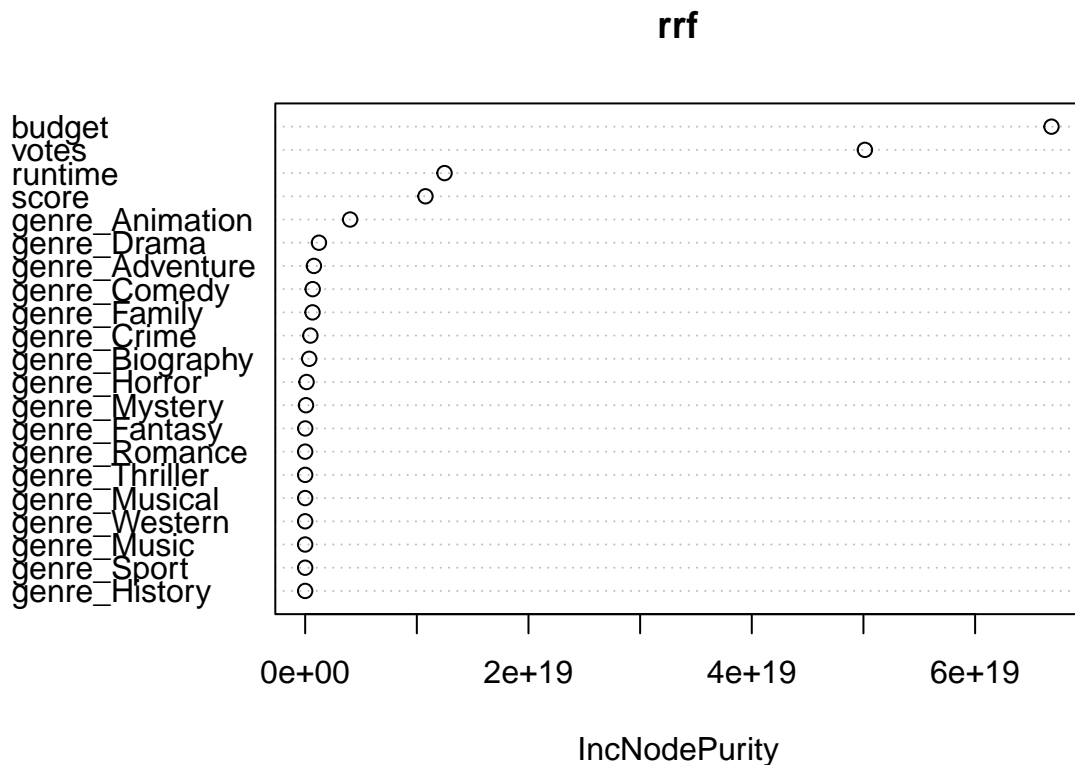
```
## Call:
## randomForest(formula = gross ~ ., data = movies_train, type = "regression")
##           Type of random forest: regression
##           Number of trees: 500
## No. of variables tried at each split: 7
##
##           Mean of squared residuals: 7.553173e+15
##           % Var explained: 73.03
```

```
rrf_pred <- predict(rrf, movies_test)
rrf_rmse <- rmse(rrf_pred, movies_test$gross) # 77096427

#Variables we should take note of
impt_rrf <- varImp(rrf, scale = T)
impt_rrf
```

```
##           Overall
## score          1.077311e+19
## votes          5.013082e+19
## budget         6.684200e+19
## runtime        1.248073e+19
## genre_Adventure 7.829328e+17
## genre_Animation 4.026380e+18
## genre_Biography 3.626013e+17
## genre_Comedy    6.734470e+17
## genre_Crime     4.690549e+17
## genre_Drama     1.235803e+18
## genre_Family    6.596846e+17
## genre_Fantasy   7.967383e+15
## genre_History   0.000000e+00
## genre_Horror    1.254980e+17
## genre_Music     1.091526e+14
## genre_Musical   7.625776e+14
## genre_Mystery   7.366227e+16
## genre_Romance   1.552712e+15
## genre_Sport     1.593942e+13
## genre_Thriller  8.591671e+14
## genre_Western   1.406925e+14
```

```
varImpPlot(rrf)
```



The two most important variables are budget and votes that would have significant impact on gross can be observed from the Random Forest plot.

Support Vector Machine (worse than rf)

```
rsvm <- svm(gross ~ . , data = movies_train, cost = 10, scale = FALSE)
rsvm_pred <- predict(rsvm, movies_test) # predict target label
rsvm_rmse <- rmse(rsvm_pred, movies_test$gross) # 162413205
```

XGBoost

```
set.seed(10)
rxgb <- caret::train(gross ~ . ,
  data = movies_train,
  method = "xgbTree",
  trControl = trainControl("cv", verboseIter = FALSE, number = 10, allowParallel = TRUE),
  verbose = TRUE)
```

```
## [10:48:47] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:48:47] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:48:47] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:48:47] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:48:47] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:48:47] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:48:47] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
```







[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]



[illegible]







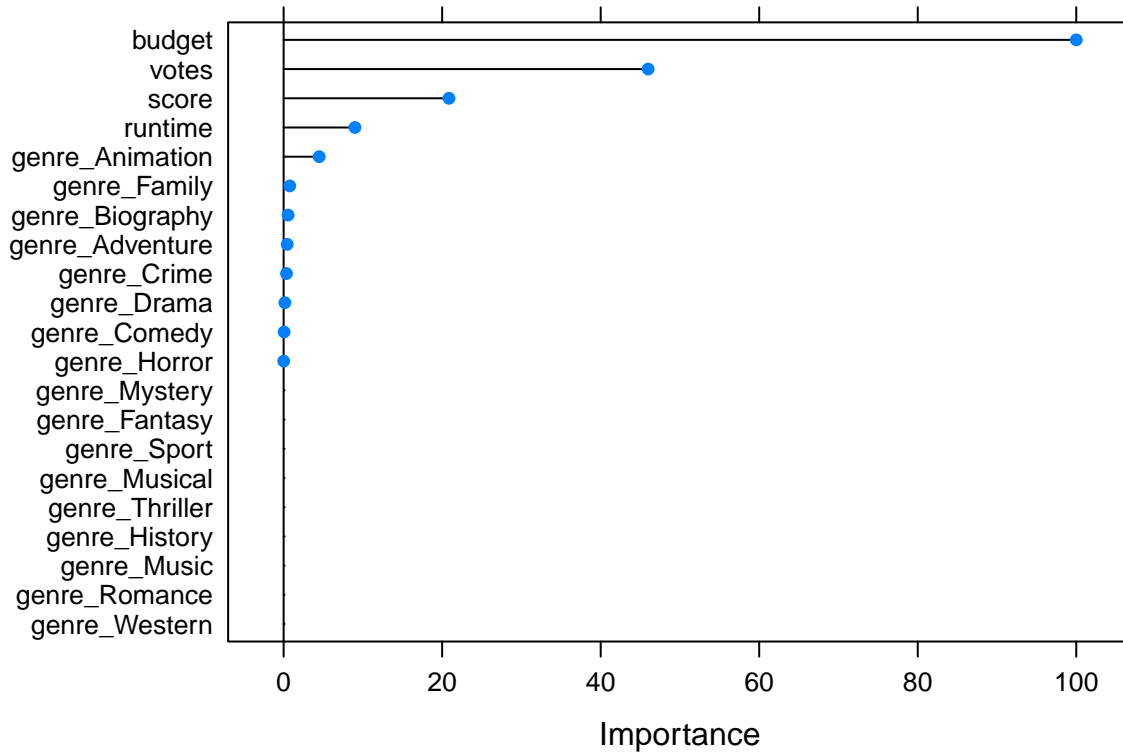
```
## [10:50:01] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:50:01] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:50:01] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:50:02] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:50:02] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:50:02] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:50:02] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:50:02] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:50:02] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:50:03] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
## [10:50:03] WARNING: amalgamation/./src/c_api/c_api.cc:785: `ntree_limit` is deprecated, use `iterat
```

```
rxgb_pred <- predict(rxgb, movies_test)
rxgb_rmse <- rmse(rxgb_pred, movies_test$gross) # 81609432

#Variables we should take note of
impt_rxgb <- varImp(rxgb, scale = T)
impt_rxgb
```

```
## xgbTree variable importance
##
##    only 20 most important variables shown (out of 21)
##
##              Overall
## budget      100.00000
## votes       45.99125
## score       20.86238
## runtime     8.99660
## genre_Animation 4.48795
## genre_Family   0.77761
## genre_Biography 0.56420
## genre_Adventure 0.45136
## genre_Crime    0.36302
## genre_Drama    0.15790
## genre_Comedy   0.06691
## genre_Horror   0.01081
## genre_Fantasy  0.00000
## genre_History  0.00000
## genre_Thriller 0.00000
## genre_Musical  0.00000
## genre_Romance  0.00000
## genre_Mystery  0.00000
## genre_Music    0.00000
## genre_Sport    0.00000
```

```
plot(impt_rxgb)
```

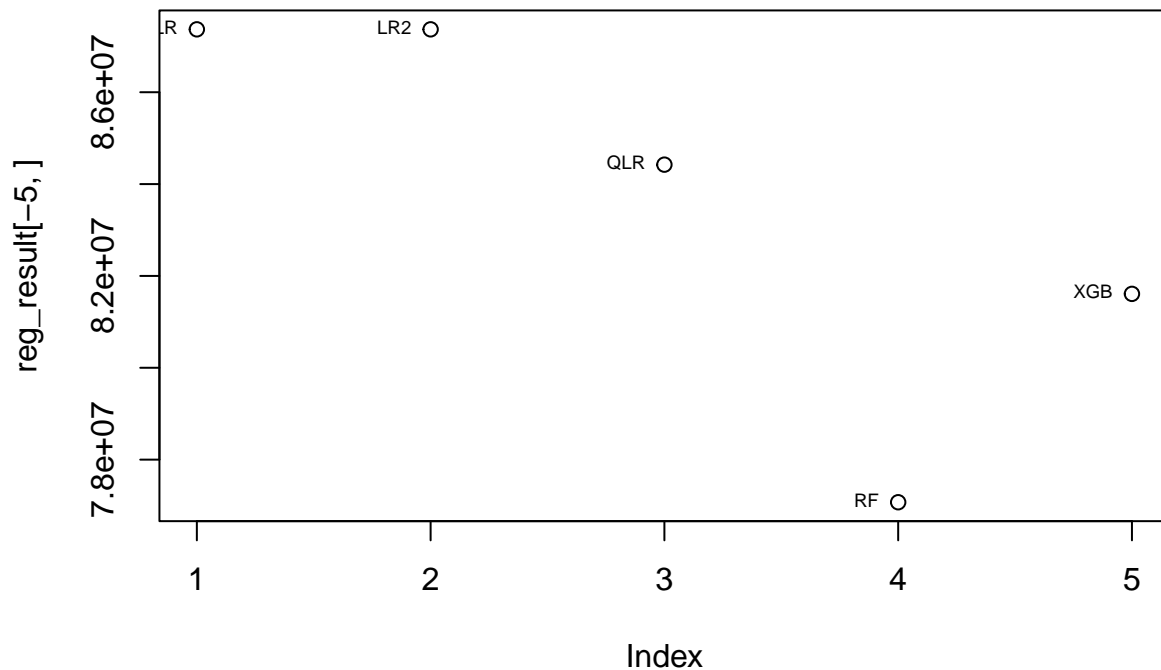


Final result of all model

```
RMSE <- c(lr_rmse, lr2_rmse, qlr_rmse, rrf_rmse, rsvm_rmse, rxgb_rmse)
reg_result <- data.frame(RMSE,
                          row.names = c("Linear Regression", "Linear Regression 2", "Quadratic LR", "Random Forest", "Support Vector Machine", "XGBoost"))
reg_result
```

```
##                                RMSE
## Linear Regression              87368985
## Linear Regression 2            87368985
## Quadratic LR                   84423164
## Random Forest                  77072280
## Support Vector Machine         162413205
## XGBoost                        81609432
```

```
outlier <- reg_result > 90000000
plot(reg_result[-5, ])
text(reg_result[-5, ], labels = c("LR", "LR2", "QLR", "RF", "XGB"), pos = 2, cex = 0.6)
```



From the plot, we can observe that XGBoost has the best model in predicting gross as it has the best AUC of 0.9804 and second lowest RMSE of 81609432.

## Classification Analysis

*Loading and cleaning of dataset*

```
heart_m <- read.csv("heart.csv")
summary(heart_m)
```

```
##      age      sex      cp      trestbps
##  Min.   :29.00  Min.   :0.0000  Min.   :0.0000  Min.    : 94.0
## 1st Qu.:48.00  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:120.0
## Median :56.00  Median :1.0000  Median :1.0000  Median :130.0
## Mean   :54.43  Mean   :0.6956  Mean   :0.9424  Mean   :131.6
## 3rd Qu.:61.00  3rd Qu.:1.0000  3rd Qu.:2.0000  3rd Qu.:140.0
## Max.   :77.00  Max.   :1.0000  Max.   :3.0000  Max.   :200.0
##      chol      fbs      restecg      thalach
##  Min.   :126  Min.   :0.0000  Min.   :0.0000  Min.    : 71.0
## 1st Qu.:211  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:132.0
## Median :240  Median :0.0000  Median :1.0000  Median :152.0
## Mean   :246  Mean   :0.1493  Mean   :0.5298  Mean   :149.1
## 3rd Qu.:275  3rd Qu.:0.0000  3rd Qu.:1.0000  3rd Qu.:166.0
## Max.   :564  Max.   :1.0000  Max.   :2.0000  Max.   :202.0
##      exang      ca      thal      target
```

```
## Min. :0.0000 Min. :0.0000 Min. :0.000 Min. :0.0000
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:2.000 1st Qu.:0.0000
## Median :0.0000 Median :0.0000 Median :2.000 Median :1.0000
## Mean :0.3366 Mean :0.7541 Mean :2.324 Mean :0.5132
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:3.000 3rd Qu.:1.0000
## Max. :1.0000 Max. :4.0000 Max. :3.000 Max. :1.0000
```

```
sum(is.na(heart_m))
```

```
## [1] 0
```

```
sum(duplicated(heart_m))
```

```
## [1] 723
```

No NAs to omit. Clean data.

*Exploring data*

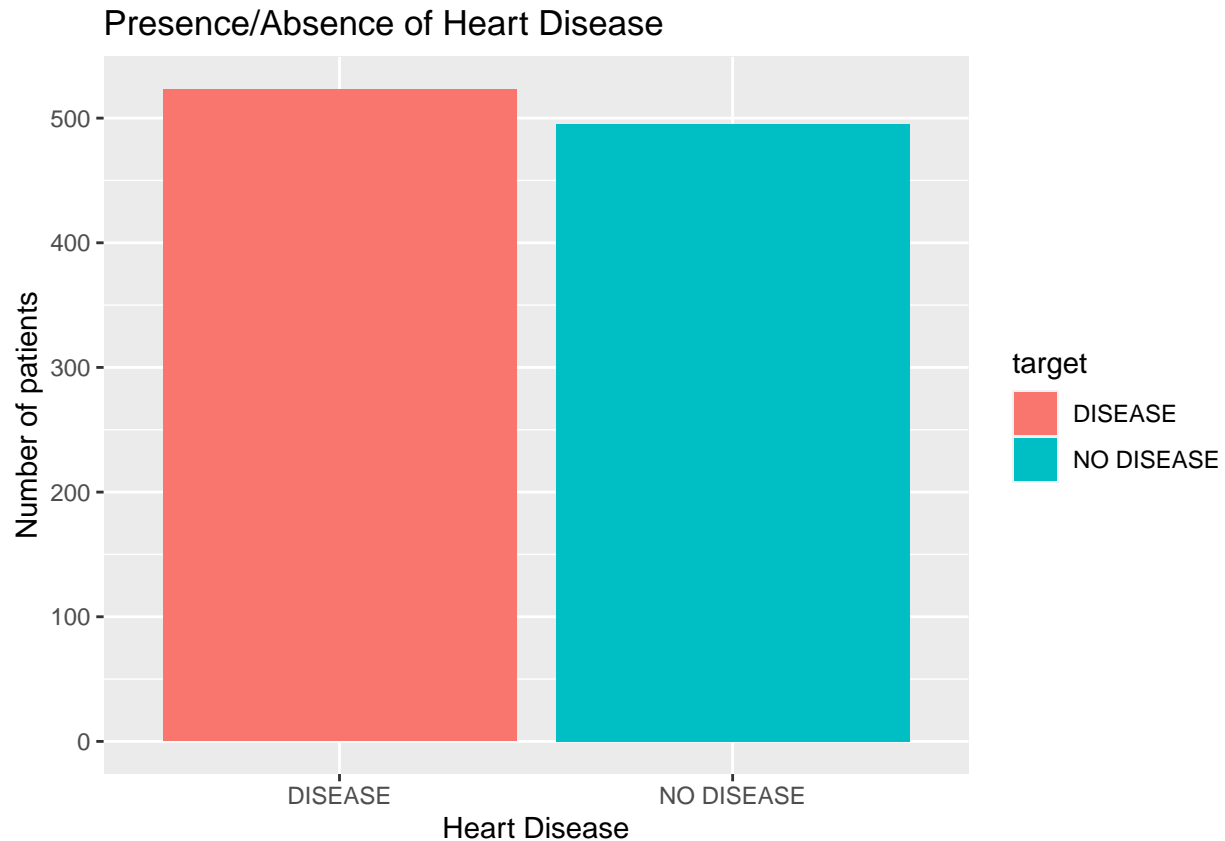
```
heart_m["thal"][heart_m["thal"] == 0] <- NA
heart_dropna <- heart_m %>% drop_na(thal)
heart <- heart_dropna %>%
  mutate(sex = if_else(sex == 1, "MALE", "FEMALE"),
         fbs = if_else(fbs == 1, ">120", "<=120"),
         exang = if_else(exang == 1, "YES", "NO"),
         thal = if_else(thal == 1, "NORMAL", if_else(thal == 2, "FIXED", "REVERSABLE")),
         target = if_else(target == 1, "DISEASE", "NO DISEASE"))
  ) %>%
  mutate_if(is.character, as.factor) # changing to factor
summary(heart)
```

```
##      age      sex      cp      trestbps      chol
## Min.   :29.00  FEMALE:309  Min.   :0.000  Min.   : 94.0  Min.   :126.0
## 1st Qu.:48.00  MALE  :709  1st Qu.:0.000  1st Qu.:120.0  1st Qu.:211.0
## Median :56.00                      Median :1.000  Median :130.0  Median :240.0
## Mean   :54.45                      Mean   :0.943  Mean   :131.6  Mean   :246.3
## 3rd Qu.:61.00                      3rd Qu.:2.000  3rd Qu.:140.0  3rd Qu.:275.8
## Max.   :77.00                      Max.   :3.000  Max.   :200.0  Max.   :564.0
##      fbs      restecg      thalach      exang      ca
## <=120:869  Min.   :0.0000  Min.   : 71.0  NO :677  Min.   :0.0000
## >120 :149  1st Qu.:0.0000  1st Qu.:132.0  YES:341  1st Qu.:0.0000
##                      Median :1.0000  Median :152.0                      Median :0.0000
##                      Mean    :0.5295  Mean    :149.2                      Mean    :0.7593
##                      3rd Qu.:1.0000  3rd Qu.:166.0                      3rd Qu.:1.0000
##                      Max.    :2.0000  Max.    :202.0                      Max.    :4.0000
##      thal      target
## FIXED   :544  DISEASE :523
## NORMAL  : 64  NO DISEASE:495
## REVERSABLE:410
##
##
##
```



Summary of the data after data transformation

```
ggplot(heart, aes(x = target, fill = target)) +  
  geom_bar() +  
  xlab("Heart Disease") +  
  ylab("Number of patients") +  
  ggtitle("Presence/Absence of Heart Disease")
```



```
table(heart$target)
```

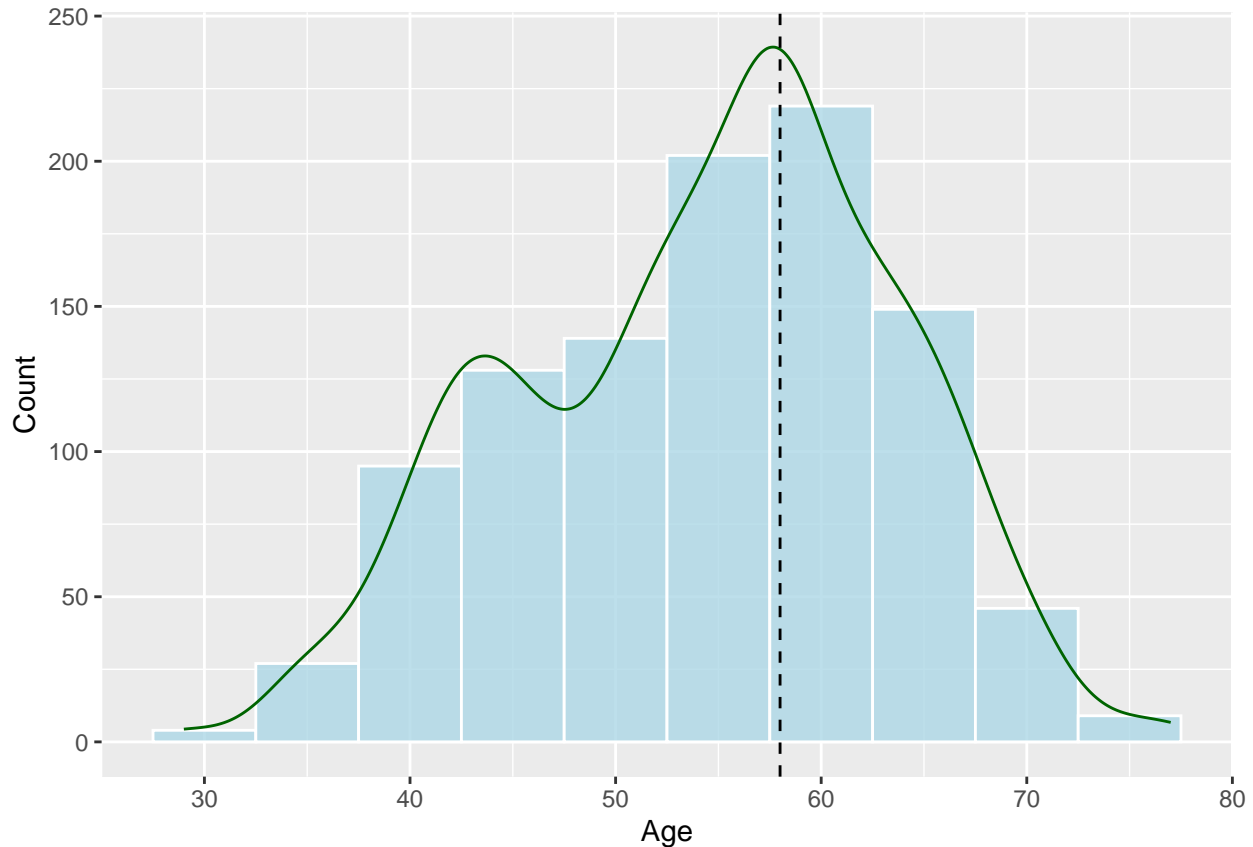
```
##  
##      DISEASE NO DISEASE  
##          523      495
```

Visualization of target showing no imbalance issue.

```
ggplot(heart, aes(x = age)) +  
  geom_histogram(binwidth = 5, colour = "white", fill = "lightblue", alpha = 0.8) +  
  geom_density(eval(bquote(aes(y=..count..*5))), colour="darkgreen", alpha=0.3) +  
  geom_vline(xintercept = 58, linetype="dashed") +  
  xlab("Age") +  
  ylab("Count")
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
```

```
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Showing a normal distribution of age.

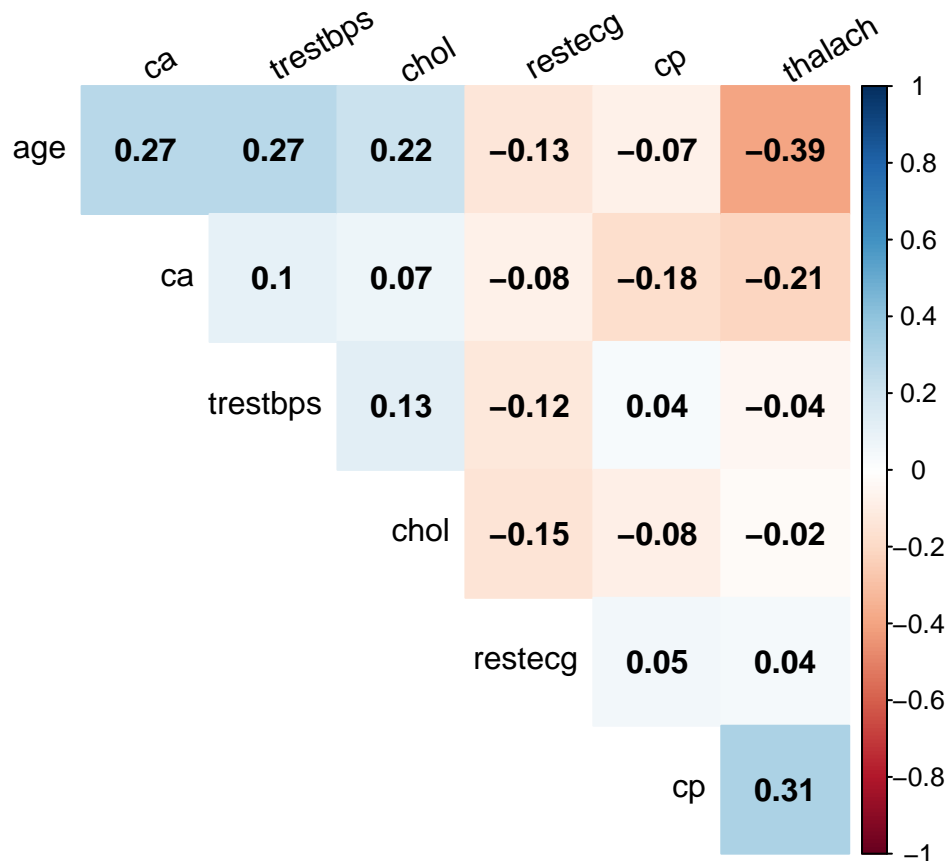
#### Continuous Variables

```
num.cols <- sapply(heart, is.numeric)
ccorr_matrix <- cor(heart[,num.cols])
round(ccorr_matrix, 4)
```

```
##          age      cp trestbps      chol restecg thalach      ca
## age      1.0000 -0.0727  0.2709  0.2191 -0.1327 -0.3920  0.2710
## cp       -0.0727  1.0000  0.0382 -0.0832  0.0501  0.3144 -0.1775
## trestbps 0.2709  0.0382  1.0000  0.1272 -0.1241 -0.0401  0.1037
## chol     0.2191 -0.0832  0.1272  1.0000 -0.1470 -0.0235  0.0709
## restecg -0.1327  0.0501 -0.1241 -0.1470  1.0000  0.0433 -0.0781
## thalach -0.3920  0.3144 -0.0401 -0.0235  0.0433  1.0000 -0.2113
## ca       0.2710 -0.1775  0.1037  0.0709 -0.0781 -0.2113  1.0000
```

```
corrplot(ccorr_matrix, type = "upper", order = "hclust",
         method = "color",
         tl.col = "black", tl.srt = 30,
```

```
addCoef.col = "black",
sig.level = 0.01,
diag=FALSE)
```



Correlation plot. Barely shown any correlation between the numeric variables, meaning it may require the categorical variables for further analysis.

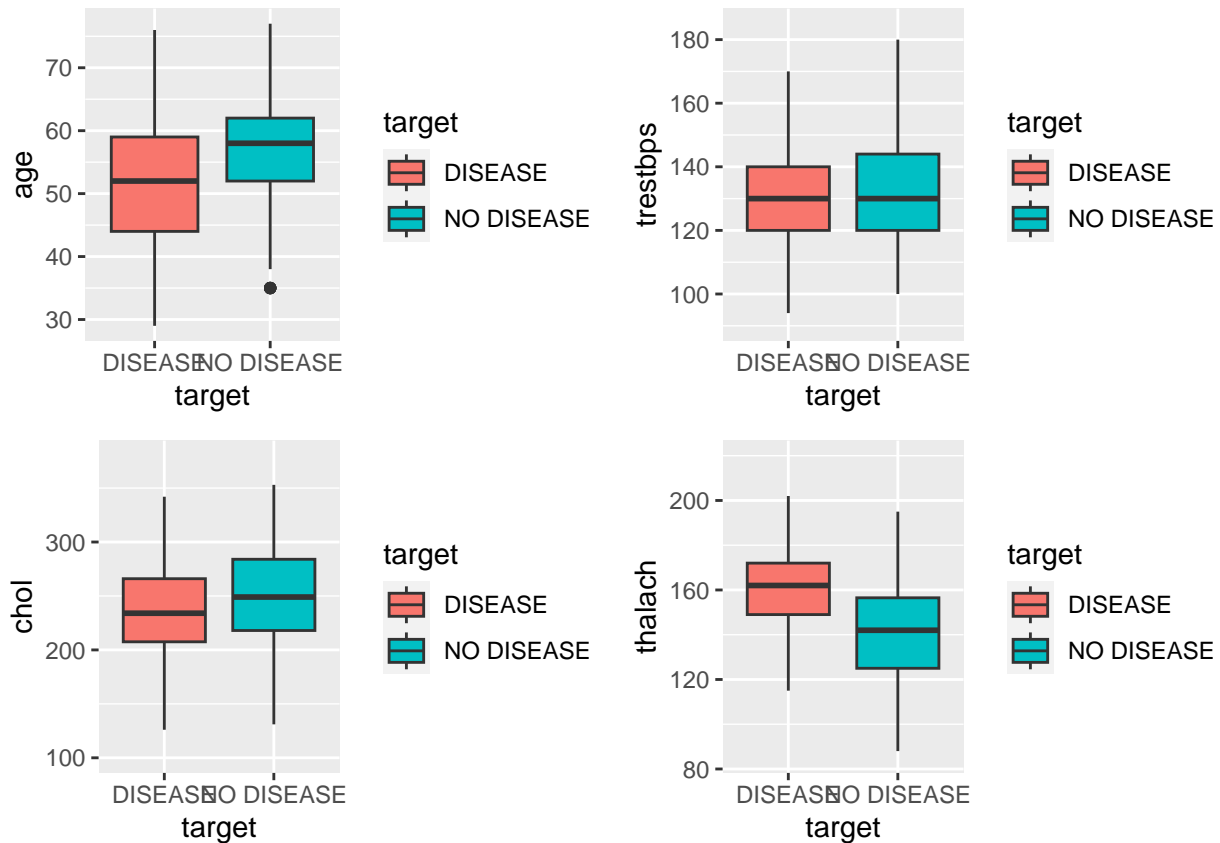
```
tar_age <- ggplot(heart, aes(target, age, fill = target)) +
  geom_boxplot()

tar_trest <- ggplot(heart, aes(target, trestbps, fill = target)) +
  geom_boxplot(outlier.shape = NA) +
  coord_cartesian(ylim = c(90,185))

tar_chol <- ggplot(heart, aes(target, chol, fill = target)) +
  geom_boxplot(outlier.shape = NA) +
  coord_cartesian(ylim = c(100,380))

tar_thalach <- ggplot(heart, aes(target, thalach, fill = target)) +
  geom_boxplot(outlier.shape = NA) +
  coord_cartesian(ylim = c(85,220))

tar_conplots <- ggarrange(tar_age, tar_trest, tar_chol, tar_thalach,
  ncol = 2, nrow = 2)
tar_conplots
```



Categorical variables

All target related EDA

```
tar_sex <- ggplot(heart, aes(target)) +
  geom_bar(aes(fill = factor(sex)), position = "dodge") +
  labs(fill = "Sex") +
  theme(legend.title = element_text(size = 10))

tar_cp <- ggplot(heart, aes(target)) +
  geom_bar(aes(fill = factor(cp)), position = "dodge") +
  labs(fill = "Cp") +
  theme(legend.title = element_text(size = 10))

tar_fbs <- ggplot(heart, aes(target)) +
  geom_bar(aes(fill = factor(fbs)), position = "dodge") +
  labs(fill = "Fbs") +
  theme(legend.title = element_text(size = 10)) # though not that accurate as research says that if its

tar_restecg <- ggplot(heart, aes(target)) +
  geom_bar(aes(fill = factor(restecg)), position = "dodge") +
  labs(fill = "Restecg") +
  theme(legend.title = element_text(size = 10))

tar_exang <- ggplot(heart, aes(target)) +
```

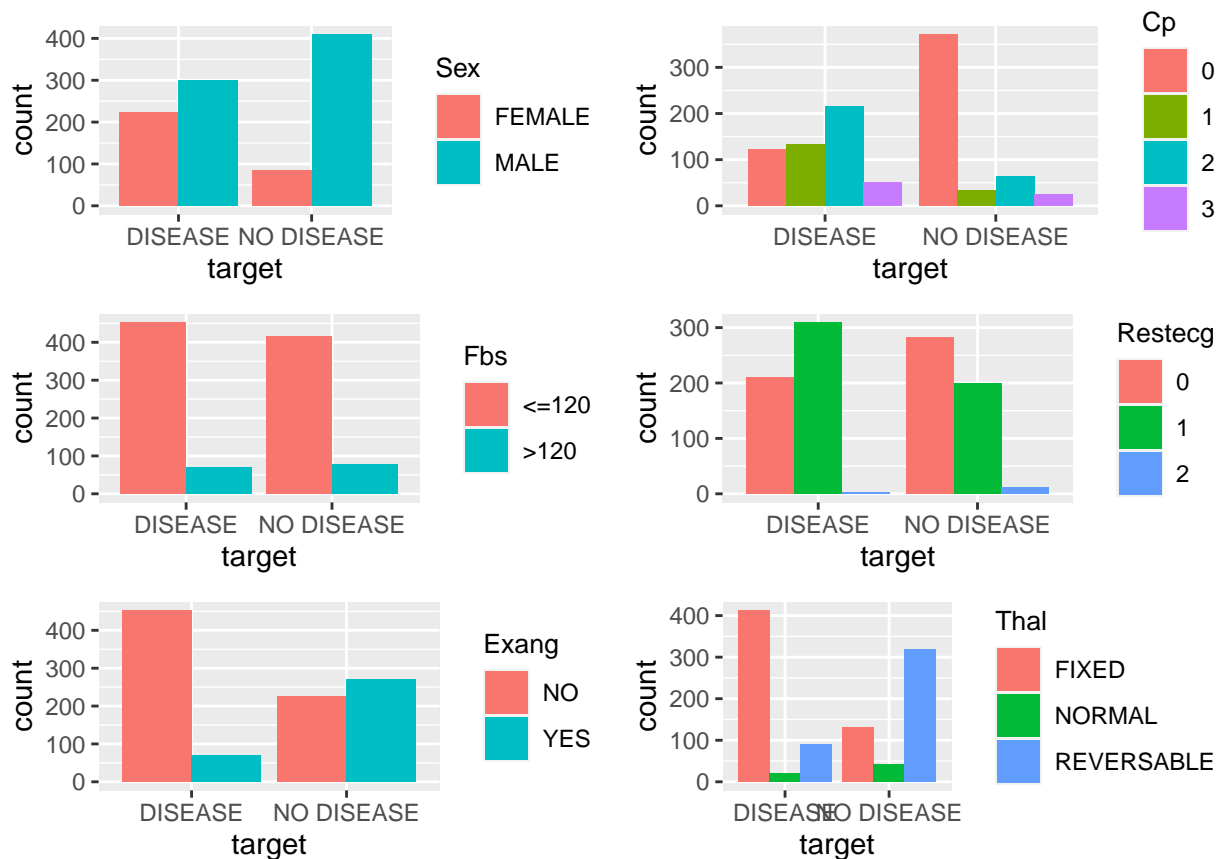
```

geom_bar(aes(fill = factor(exang)), position = "dodge") +
labs(fill = "Exang") +
theme(legend.title = element_text(size = 10))

tar_thal <- ggplot(heart, aes(target)) +
geom_bar(aes(fill = factor(thal)), position = "dodge") +
labs(fill = "Thal") +
theme(legend.title = element_text(size = 10))

tar_catplots <- ggarrange(tar_sex, tar_cp, tar_fbs, tar_restecg, tar_exang, tar_thal,
ncol = 2, nrow = 3)
print(tar_catplots)

```



*Setting dataset into training and testing sets*

```

set.seed(10)
heart_set <- sample.split(heart_dropna, 0.8)
heart_train <- subset(heart_dropna, heart_set == TRUE)
heart_test <- subset(heart_dropna, heart_set == FALSE)

```

## Logistic Regression

*Training Log model*

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.1.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.2
```

```
##
```

```
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
```

```
##
```

```
##      recode
```

```
lgm <- glm(target ~ ., data = heart_train, family = "binomial")
summary(lgm) # AIC = 610.61
```

```
##
```

```
## Call:
```

```
## glm(formula = target ~ ., family = "binomial", data = heart_train)
```

```
##
```

```
## Deviance Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2.2525  -0.4998   0.1115   0.5927   2.7783
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.413207   1.533294   2.226  0.02601 *
## age          0.001366   0.014220   0.096  0.92349
## sex         -1.597452   0.272327  -5.866 4.47e-09 ***
## cp           0.722906   0.105650   6.842 7.78e-12 ***
## trestbps     -0.023991   0.006070  -3.953 7.73e-05 ***
## chol        -0.006054   0.002203  -2.748  0.00599 **
## fbs          0.183344   0.310814   0.590  0.55527
## restecg      0.506864   0.201258   2.518  0.01179 *
## thalach      0.029793   0.005887   5.061 4.18e-07 ***
## exang       -1.137073   0.244701  -4.647 3.37e-06 ***
## ca          -0.809376   0.110747  -7.308 2.70e-13 ***
## thal        -0.922307   0.176899  -5.214 1.85e-07 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
```

```
##      Null deviance: 1059.00  on 763  degrees of freedom
```

```
## Residual deviance:  586.61  on 752  degrees of freedom
```

```
## AIC: 610.61
```

```
##
```

```
## Number of Fisher Scoring iterations: 5
```

```
vif(lgm)
```

```
##      age      sex      cp trestbps      chol      fbs restecg  thalach
## 1.414884 1.308671 1.146207 1.141986 1.261657 1.092062 1.083157 1.278136
##      exang      ca      thal
## 1.137667 1.101462 1.042433
```

AIC is at 610.10 (smaller the AIC the better). Since all VIF are less than three, there is no multicollinearity between the variables, only insignificant variables for this model.

```
lgm2 <- glm(target ~ . - fbs - age, data = heart_train, family = "binomial")
summary(lgm2)
```

```
##
## Call:
## glm(formula = target ~ . - fbs - age, family = "binomial", data = heart_train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.2586  -0.4846   0.1100   0.5835   2.7740
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.506511   1.295967   2.706  0.00682 **
## sex          -1.608548   0.270981  -5.936 2.92e-09 ***
## cp             0.731408   0.104966   6.968 3.21e-12 ***
## trestbps      -0.023509   0.005854  -4.016 5.92e-05 ***
## chol          -0.006037   0.002156  -2.800  0.00512 **
## restecg       0.491305   0.197758   2.484  0.01298 *
## thalach       0.029496   0.005450   5.412 6.22e-08 ***
## exang        -1.127526   0.244309  -4.615 3.93e-06 ***
## ca           -0.797647   0.107294  -7.434 1.05e-13 ***
## thal         -0.930064   0.175857  -5.289 1.23e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1059.00  on 763  degrees of freedom
## Residual deviance:  586.98  on 754  degrees of freedom
## AIC: 606.98
##
## Number of Fisher Scoring iterations: 5
```

```
vif(lgm2)
```

```
##      sex      cp trestbps      chol restecg  thalach      exang      ca
## 1.291700 1.130583 1.062990 1.215954 1.048399 1.100107 1.132012 1.037208
##      thal
## 1.033195
```

We removed fbs and age as they were insignificant from the previous model and the AIC improved at 606.98. This would be our final model as there are nothing else to be removed.

Making predictions

```
clgm_pred <- predict(lgm2, heart_test, type = "response")
clgm_pred <- ifelse(clgm_pred >= 0.5, 1, 0)

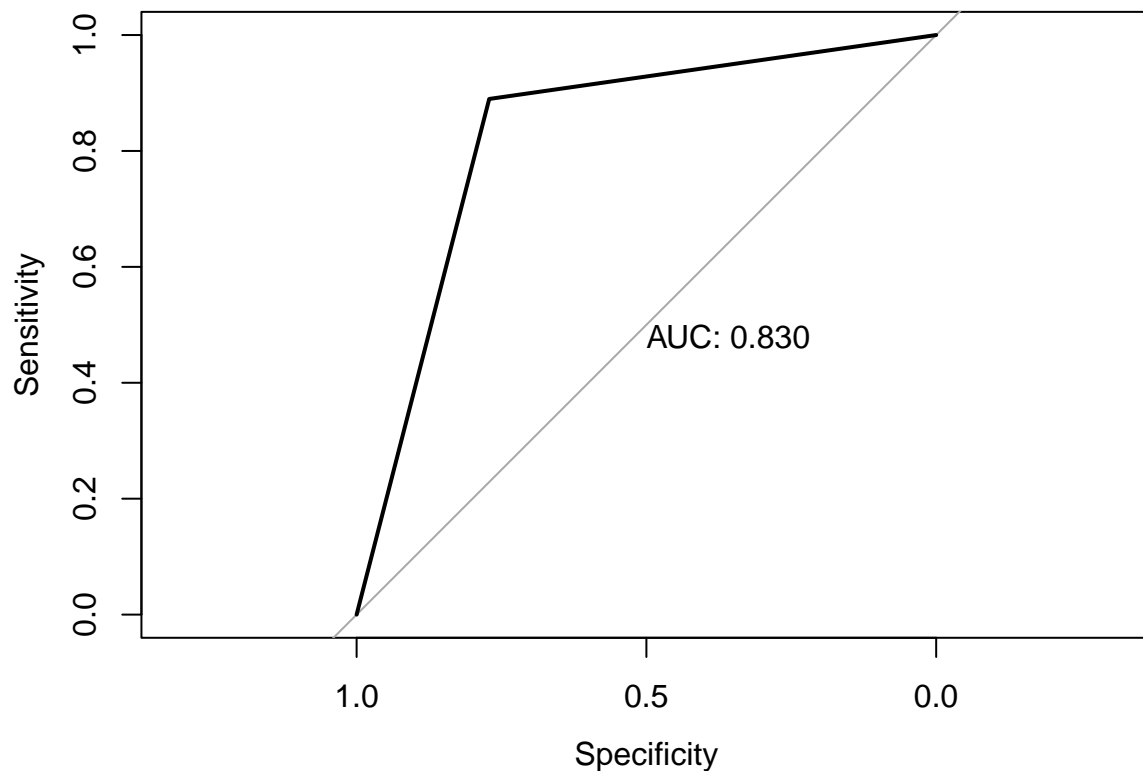
# Confusion Matrix
clgm_cm <- confusionMatrix(factor(heart_test$target), factor(clgm_pred))

# ROC-AUC & Accuracy
clgm_roc <- roc(heart_test$target, clgm_pred)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(clgm_roc, plot = TRUE, print.auc = TRUE)
```



```
clgm_acc <- clgm_cm$overall[1]
clgm_auc <- clgm_roc$auc
clgm_result <- cbind("Accuracy" = clgm_acc, "AUC" = clgm_auc) # acc = 0.8346 / auc = 0.8304
```



# Random Forest

*Training RF model*

```
set.seed(20)
crf <- randomForest(factor(target) ~ . , data = heart_train)
```

Making predictions

```
crf_pred <- predict(crf, heart_test, type = "class")

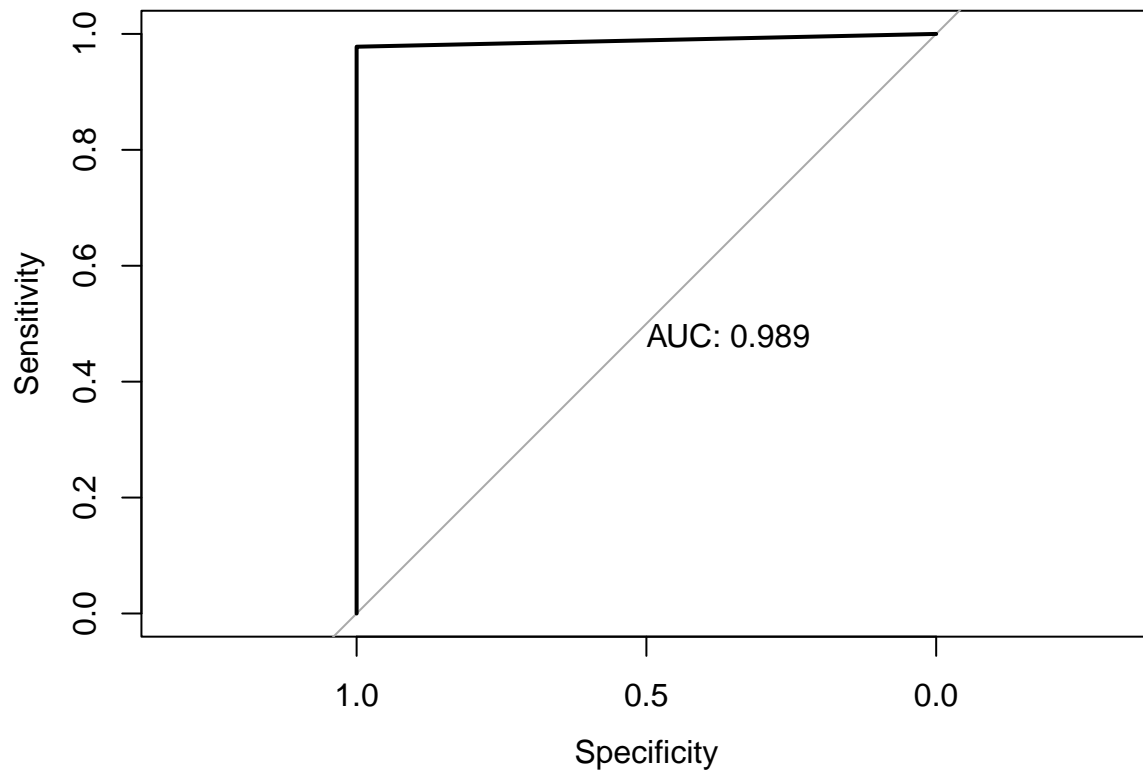
# Confusion Matrix
crf_cm <- confusionMatrix(factor(heart_test$target), factor(crf_pred))

# ROC-AUC & Accuracy
crf_roc <- roc(heart_test$target, as.numeric(crf_pred))
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(crf_roc, plot = TRUE, print.auc = TRUE)
```



```

crf_acc <- crf_cm$overall[1]
crf_auc <- crf_roc$auc
crf_result <- cbind("Accuracy" = crf_acc, "AUC" = crf_auc) # acc = 0.9882 / auc = 0.98

# Variables we should take note of
impt_crf <- varImp(crf, scale = T)
impt_crf

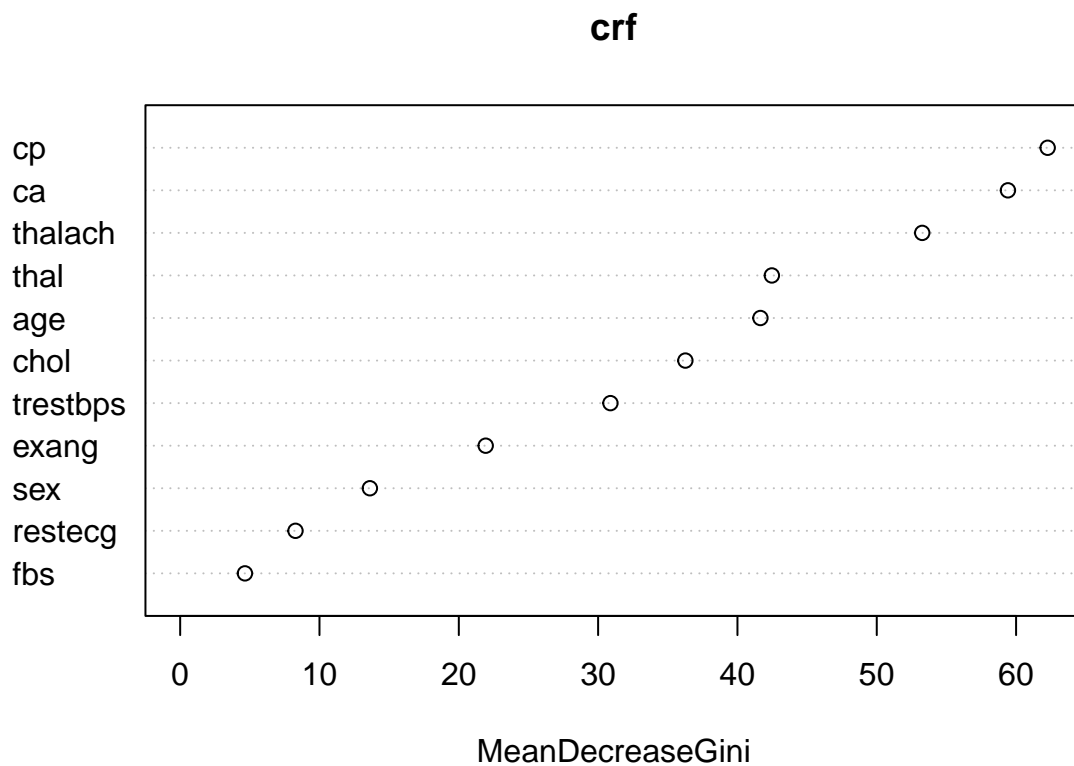
```

```

##           Overall
## age      41.649579
## sex      13.617770
## cp       62.266709
## trestbps 30.883876
## chol     36.263029
## fbs       4.649603
## restecg  8.277402
## thalach  53.262774
## exang     21.930416
## ca       59.415013
## thal     42.469137

```

```
varImpPlot(crf)
```



Important variables to take note of are chest pain and ca

# Decision Tree

```
library(rpart)
cdt <- rpart(factor(target) ~ . , data = heart_train, method = "class")
```

Making predictions

```
cdt_pred <- predict(cdt, heart_test, type = "class")

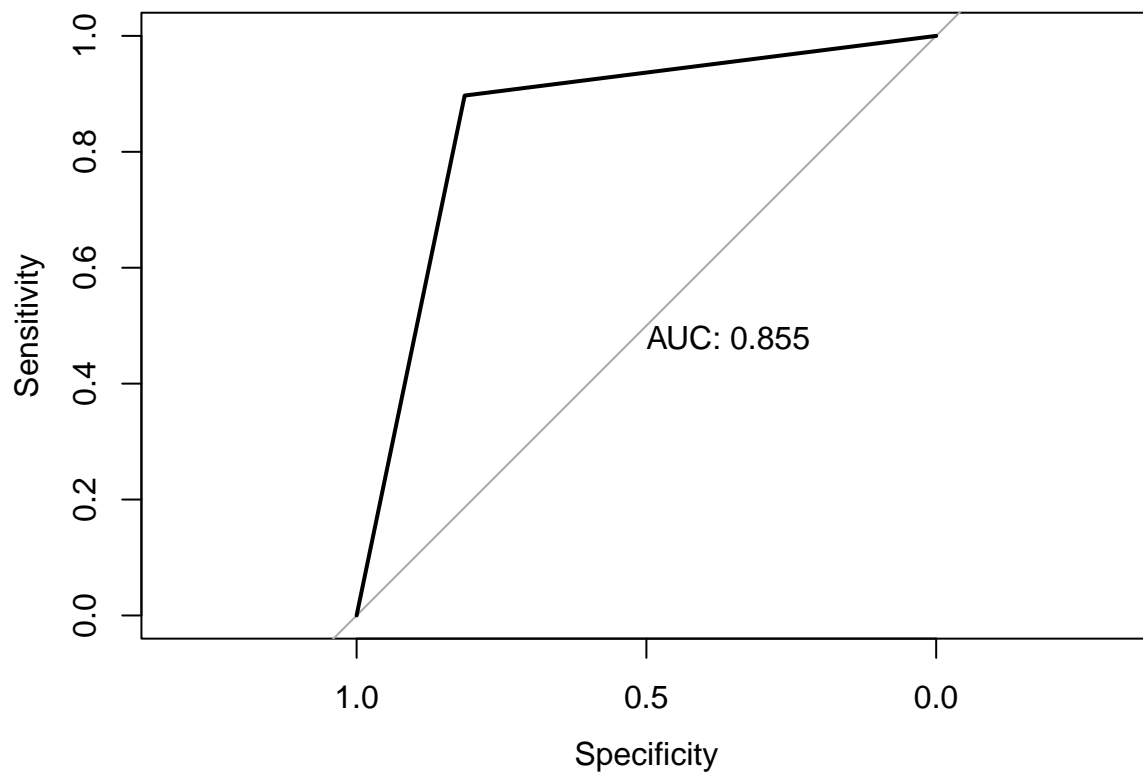
# Confusion Matrix
cdt_cm <- confusionMatrix(factor(heart_test$target), factor(cdt_pred))

# ROC-AUC & Accuracy
cdt_roc <- roc(heart_test$target, as.numeric(cdt_pred))
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(cdt_roc, plot = TRUE, print.auc = TRUE)
```



```
cdt_acc <- cdt_cm$overall[1]
cdt_auc <- cdt_roc$auc
cdt_result <- cbind("Accuracy" = cdt_acc, "AUC" = cdt_auc) # acc = 0.8582677 / auc = 0.8553091
```

Accuracy and AUC is lower than RF

## Support Vector Machine

```
csvm <- svm(target ~ . ,
            data = heart_train,
            kernel = "linear",
            type = "C-classification",
            cost = 10,
            scale = FALSE)
```

Making predictions

```
csvm_pred <- predict(csvm, heart_test)

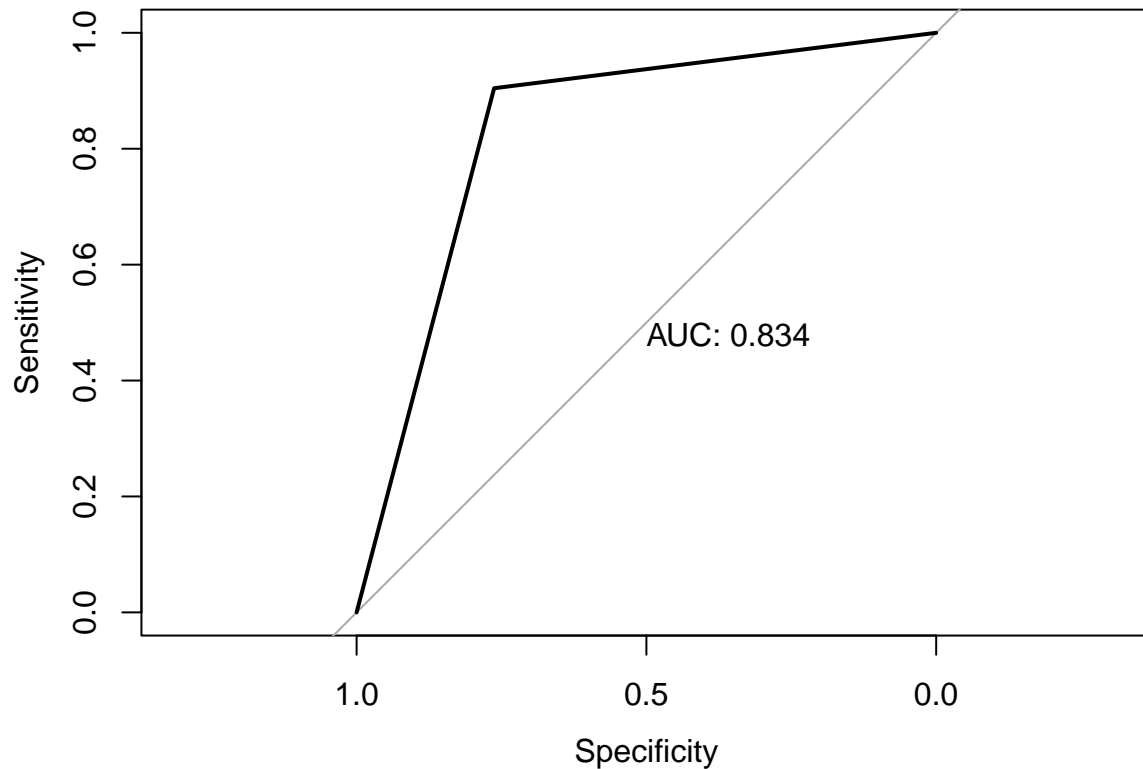
# Confusion Matrix
csvm_cm <- confusionMatrix(as.factor(heart_test$target), csvm_pred)

# ROC-AUC & Accuracy
csvm_roc <- roc(heart_test$target, as.numeric(csvm_pred))
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(csvm_roc, plot = TRUE, print.auc = TRUE)
```



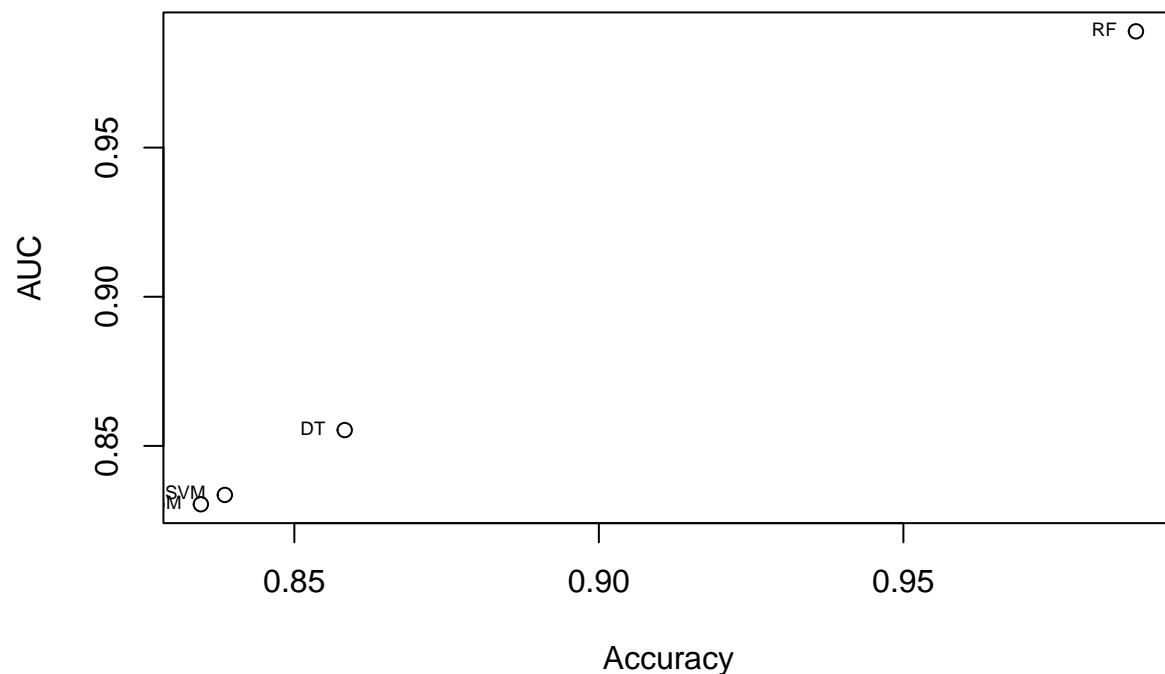
```
csvm_acc <- csvm_cm$overall[1]
csvm_auc <- csvm_roc$auc
csvm_result <- cbind("Accuracy" = csvm_acc, "AUC" = csvm_auc) # acc = 0.8385827 / auc = 0.8335618
```

Final result of all model

```
cres <- rbind(clgm_result, crf_result, cdt_result, csvm_result)
cat_result <- data.frame(cres,
                        row.names = c("Log Regression", "Random Forest", "Decision Tree", "Support Vector Machine"))
```

##	Accuracy	AUC
## Log Regression	0.8346457	0.8304462
## Random Forest	0.9881890	0.9889706
## Decision Tree	0.8582677	0.8553091
## Support Vector Machine	0.8385827	0.8335618

```
plot(cat_result)
text(cat_result, labels = c("LGM", "RF", "DT", "SVM"), pos = 2, cex = 0.6)
```



From the plot, we can observe that Random Forest has the best model in targeting patients with heart disease as it has the best Accuracy of 0.9882 and AUC of 0.9890.

## Unsupervised Learning

*Loading and cleaning dataset*

```
customer <- read.csv("Mall_Customers.csv")
summary(customer)
```

```
##      CustomerID      Gender      Age      Annual.Income..k..
##  Min.   :  1.00  Length:200  Min.   :18.00  Min.   : 15.00
## 1st Qu.: 50.75  Class :character 1st Qu.:28.75 1st Qu.: 41.50
## Median :100.50  Mode  :character  Median :36.00 Median : 61.50
## Mean   :100.50          Mean   :38.85 Mean   : 60.56
## 3rd Qu.:150.25          3rd Qu.:49.00 3rd Qu.: 78.00
## Max.   :200.00          Max.   :70.00 Max.   :137.00
## Spending.Score..1.100.
##  Min.   :  1.00
## 1st Qu.:34.75
## Median :50.00
## Mean   :50.20
## 3rd Qu.:73.00
## Max.   :99.00
```

```
sum(is.na(customer))
```

```
## [1] 0
```

```
sum(duplicated(customer))
```

```
## [1] 0
```

No NAs or duplicates found.

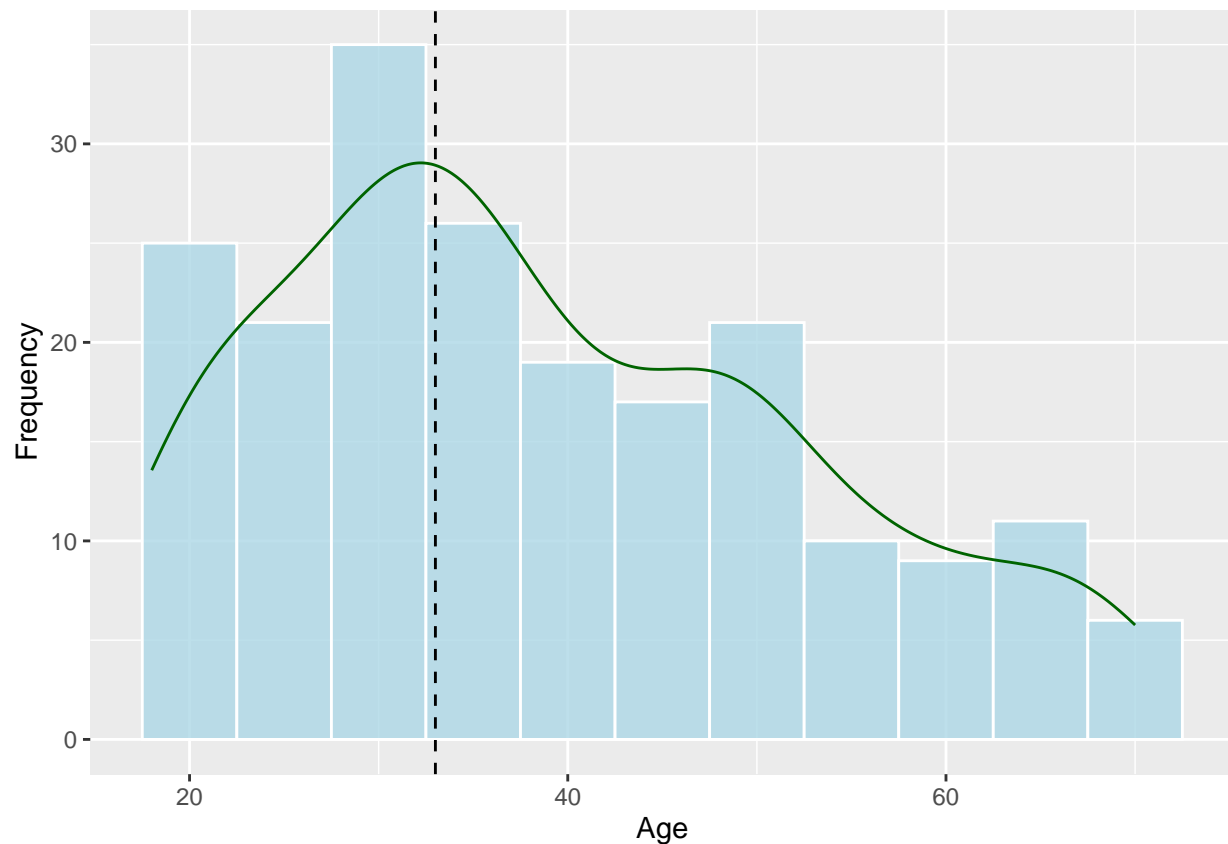
*Exploring data*

```
avg_age <- summary(customer$Age) # 38.85
```

```
avg_income <- summary(customer$Annual.Income..k..) # 60.56
```

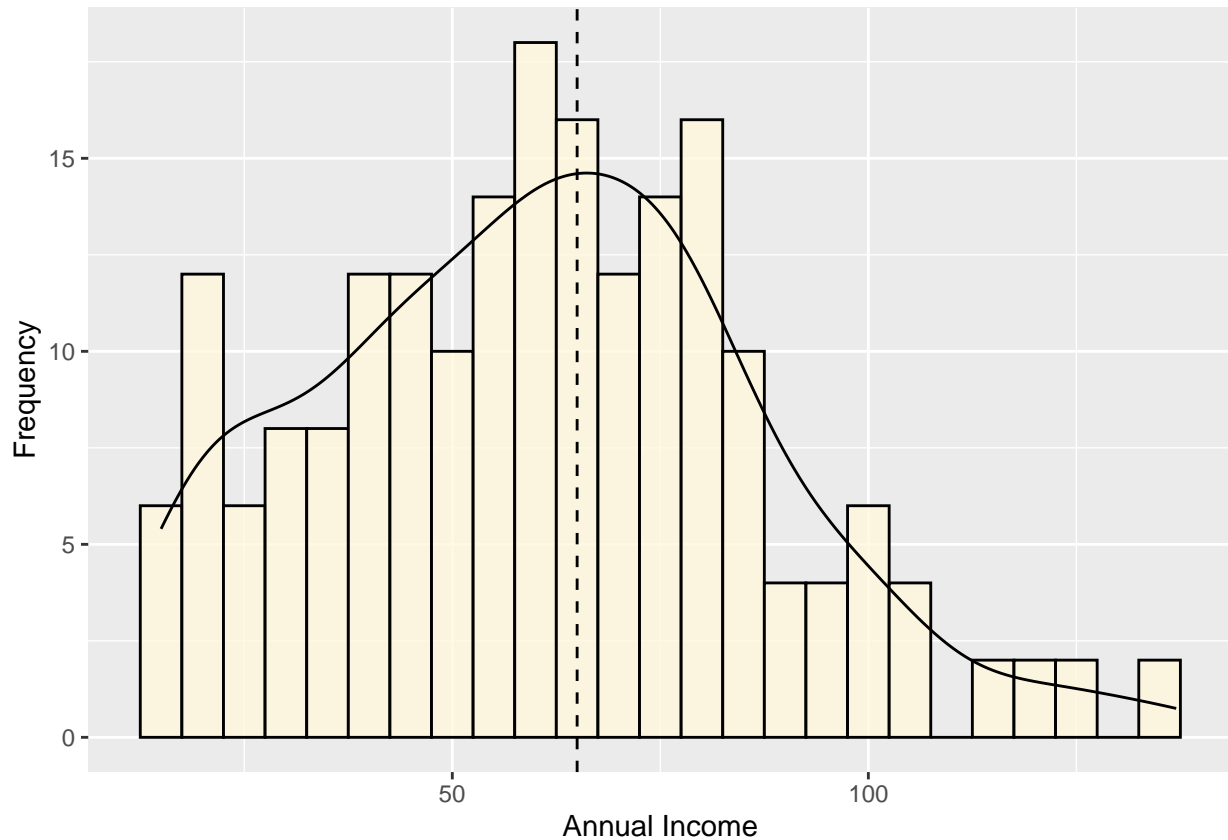
```
avg_spending_score <- summary(customer$Spending.Score..1.100..) #50.20
```

```
ggplot(customer, aes(Age)) +  
  geom_histogram(binwidth = 5, colour = "white", fill = "lightblue", alpha = 0.8) +  
  geom_density(eval(bquote(aes(y=..count..*5))), colour="darkgreen", alpha=0.3) +  
  geom_vline(xintercept = 33, linetype="dashed") +  
  xlab("Age") +  
  ylab("Frequency")
```



The highest count of customer's age were between 30 to 35.

```
ggplot(customer, aes(Annual.Income..k..)) +
  geom_histogram(binwidth = 5, colour = "black", fill = "cornsilk", alpha = 0.8) +
  geom_density(eval(bquote(aes(y=..count..*5))), colour="black", alpha=0.3) +
  geom_vline(xintercept = 65, linetype="dashed") +
  xlab("Annual Income") +
  ylab("Frequency")
```



Customer's annual income is around 55k to 75k.

```
ggplot(customer, aes(Spending.Score..1.100..)) +
  geom_histogram(binwidth = 5, colour = "white", fill = "darkseagreen", alpha = 0.8) +
  geom_density(eval(bquote(aes(y=..count..*5))), colour="black", alpha=0.3) +
  geom_vline(xintercept = 50, linetype="dashed") +
  xlab("Spending Score") +
  ylab("Frequency")
```





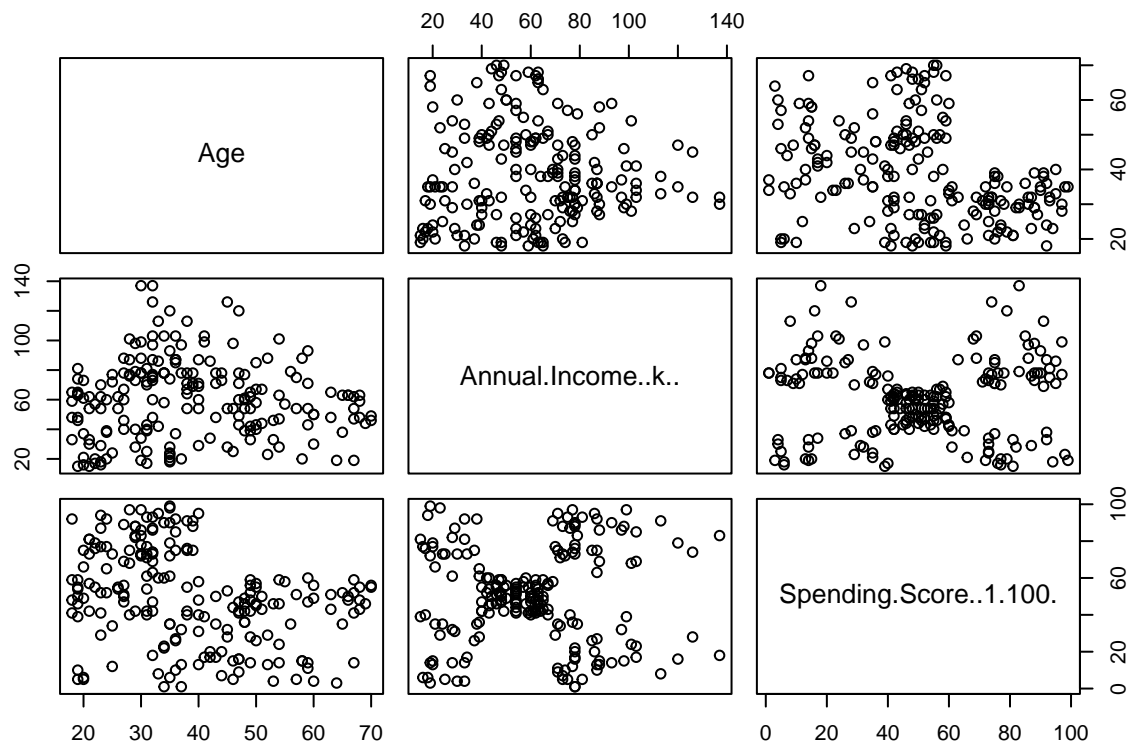
Most of the customer's spending score are around 40 to 60 out of 100

## K-means clustering

*Correlation plot & matrix*

```
# Removing CustomerID from dataframe
customer <- subset(customer, select = -c(CustomerID))

# converting Gender to binary variables
customer <- customer %>%
  mutate(Gender = if_else(Gender == "Male", 0, 1))
plot(customer[, 2:4])
```



### Principal Component Analysis

```
pca <- prcomp(customer, scale = F)
summary(pca)
```

```
## Importance of components:
##              PC1      PC2      PC3      PC4
## Standard deviation 26.4625 26.1597 12.9317 0.49548
## Proportion of Variance 0.4512 0.4409 0.1077 0.00016
## Cumulative Proportion 0.4512 0.8921 0.9998 1.00000
```

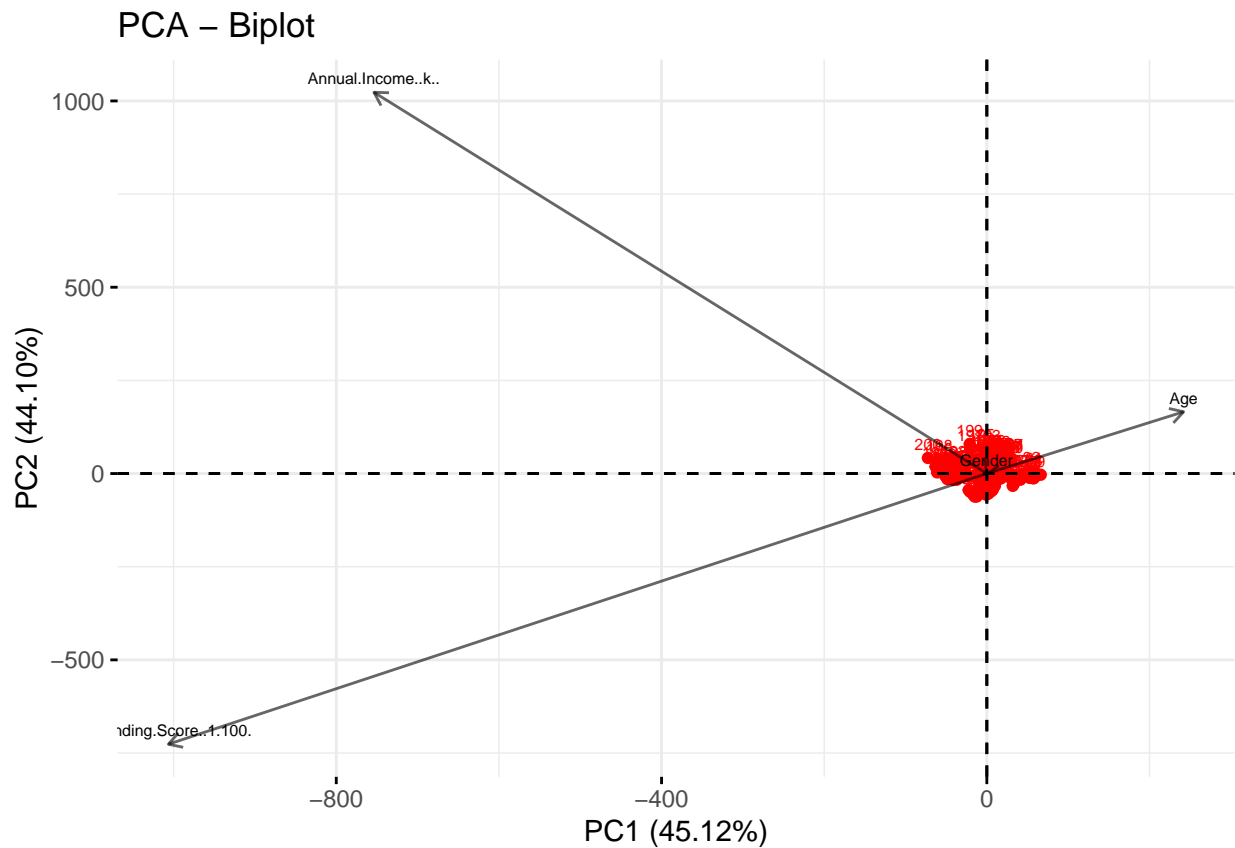
```
pca$rotation
```

```
##              PC1      PC2      PC3      PC4
## Gender      -0.0003327282 -0.001578712 0.001462466 0.9999976291
## Age          0.1889772912 0.130961404 -0.973208009 0.0016929157
## Annual.Income..k.. -0.5886227558 0.808388308 -0.005514570 0.0010884284
## Spending.Score..1.100. -0.7860093664 -0.573894557 -0.229855656 -0.0008313877
```

We are only using PC1 and PC2 because PC3 variance is close to 0.1 suggesting that it does not have much interpretive value.

### Biplot of PCA

```
biplot <- fviz_pca_biplot(pca,
  col.var = "black",
  col.ind = "red",
  alpha.var = 0.6,
  labelsize = 2) +
  labs(x = "PC1 (45.12%)",
    y = "PC2 (44.10%)")
biplot
```



### Elbow Method

```
# Taking out Gender to proceed with formula as they cannot calculate binary values
customer_f <- customer[, 2:4]
kmeans(customer_f,3) #Initial cluster sum of squares by cluster = 49.2%
```

```
## K-means clustering with 3 clusters of sizes 76, 86, 38
##
## Cluster means:
##      Age Annual.Income..k.. Spending.Score..1.100.
## 1 28.75000      65.36842      75.06579
## 2 47.09302      44.62791      42.17442
## 3 40.39474      87.00000      18.63158
##
## Clustering vector:
```

```
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
## 2 1 2 1 2 1 2 1 2 2 2 1 2 1 2 1 2 2 2 1
## 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
## 2 1 2 2 2 1 2 2 2 1 2 1 2 1 2 1 2 1 2 1
## 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
## 2 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
## 2 1 2 2 2 1 2 2 1 2 2 2 2 2 2 1 2 2 1 2
## 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
## 2 2 2 2 1 2 2 1 1 2 2 2 2 2 2 1 2 1 2 1
## 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
## 2 2 2 1 2 2 2 2 2 2 2 1 2 1 1 1 2 1 2 2
## 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
## 1 2 1 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1
## 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
## 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1
## 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
## 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1
## 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200
## 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1
##
```

```
## Within cluster sum of squares by cluster:
```

```
## [1] 77572.61 58007.73 18993.92
```

```
## (between_SS / total_SS = 49.9 %)
```

```
##
```

```
## Available components:
```

```
##
```

```
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"
```

```
## [6] "betweenss" "size" "iter" "ifault"
```

```
pca_score <- data.frame(pca$x[, 1:2])
```

```
set.seed(25)
```

```
ssw <- vector()
```

```
for(i in 1:10){
  x <- sum(kmeans(pca_score, i)$withinss)
  ssw[i] = x
  print(x)
}
```

```
## [1] 275534.7
```

```
## [1] 186865.9
```

```
## [1] 111540.7
```

```
## [1] 74701.15
```

```
## [1] 45092.05
```

```
## [1] 37876.15
```

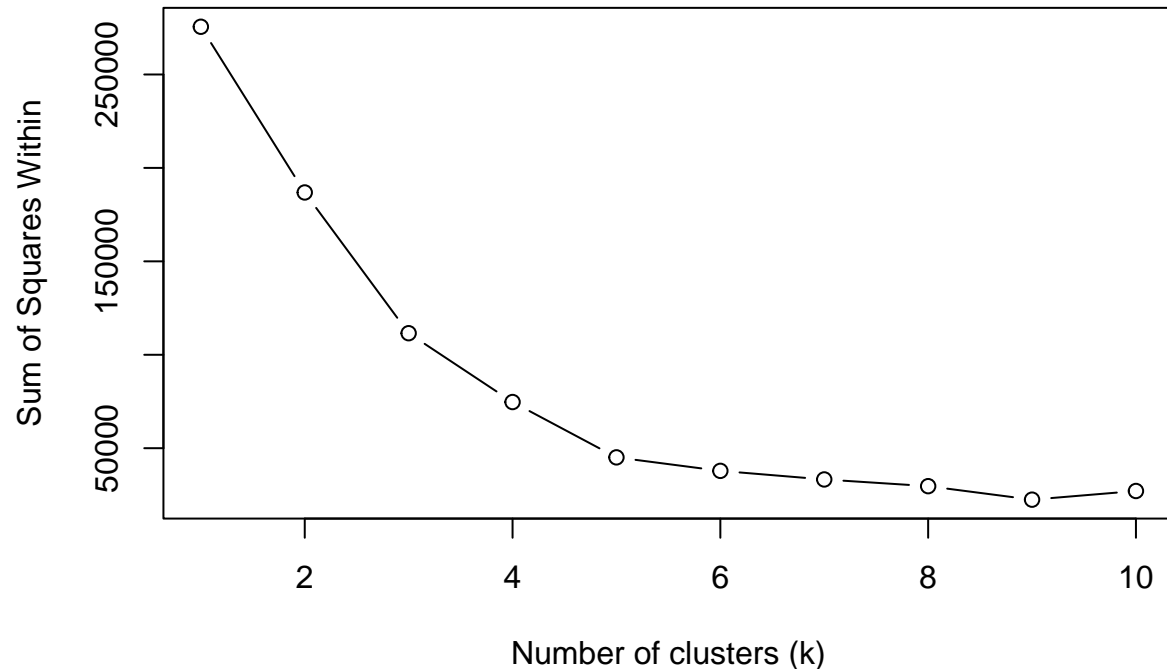
```
## [1] 33277.53
```

```
## [1] 29685.22
```

```
## [1] 22467.23
```

```
## [1] 27096.42
```

```
plot(x = 1:10, ssw,
     type = "b",
     xlab = "Number of clusters (k)",
     ylab = "Sum of Squares Within")
```



From the plot, we conclude that 5 is the optimal number of clusters as its observed that the bent starts from in this elbow plot.

*Taking  $k = 5$  as our optimal cluster into the model*

```
kmeans_model <- kmeans(customer_f, 5) # css went to 75.6%
kmeans_model
```

```
## K-means clustering with 5 clusters of sizes 23, 23, 39, 79, 36
```

```
##
```

```
## Cluster means:
```

```
##      Age Annual.Income..k.. Spending.Score..1.100.
```

```
## 1 45.21739      26.30435      20.91304
```

```
## 2 25.52174      26.30435      78.56522
```

```
## 3 32.69231      86.53846      82.12821
```

```
## 4 43.08861      55.29114      49.56962
```

```
## 5 40.66667      87.75000      17.58333
```

```
##
```

```
## Clustering vector:
```

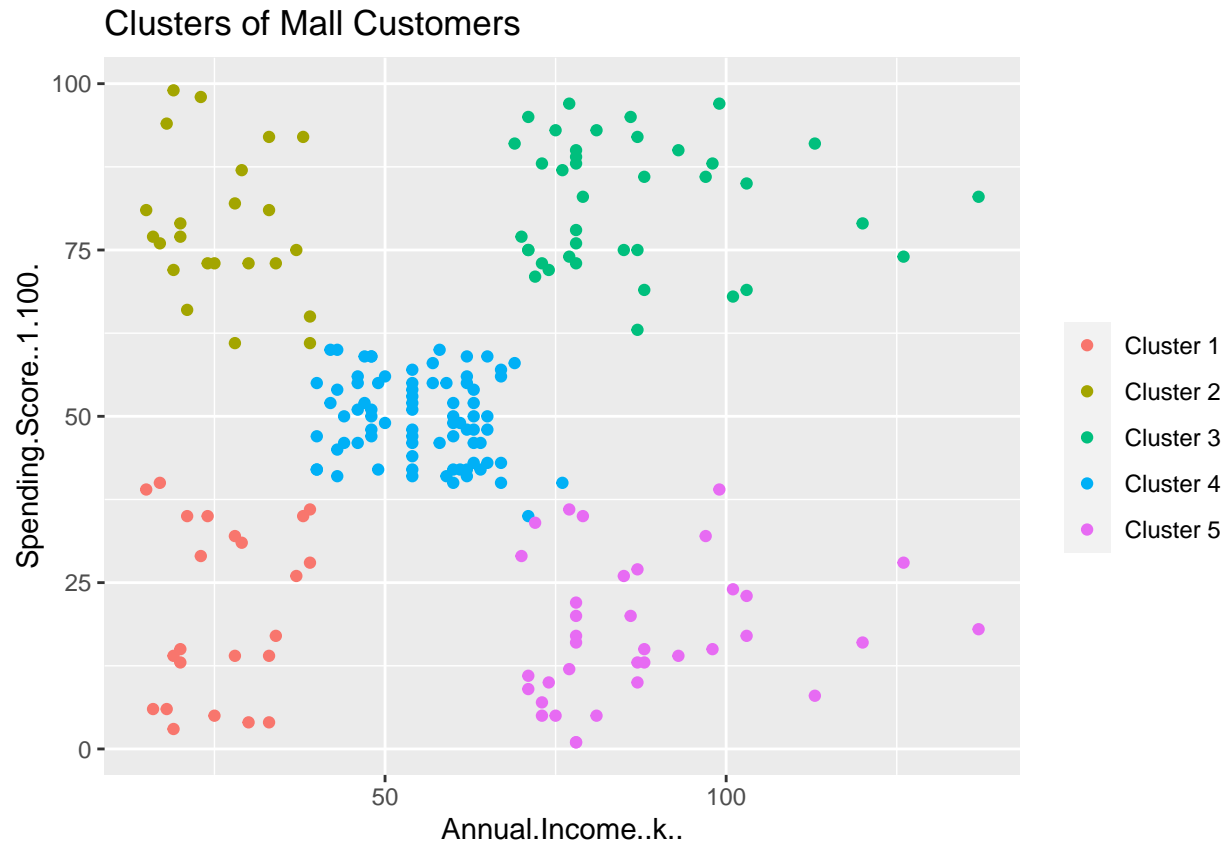
```
##   1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
```

```
##   1  2  1  2  1  2  1  2  1  2  1  2  1  2  1  2  1  2  1  2
```

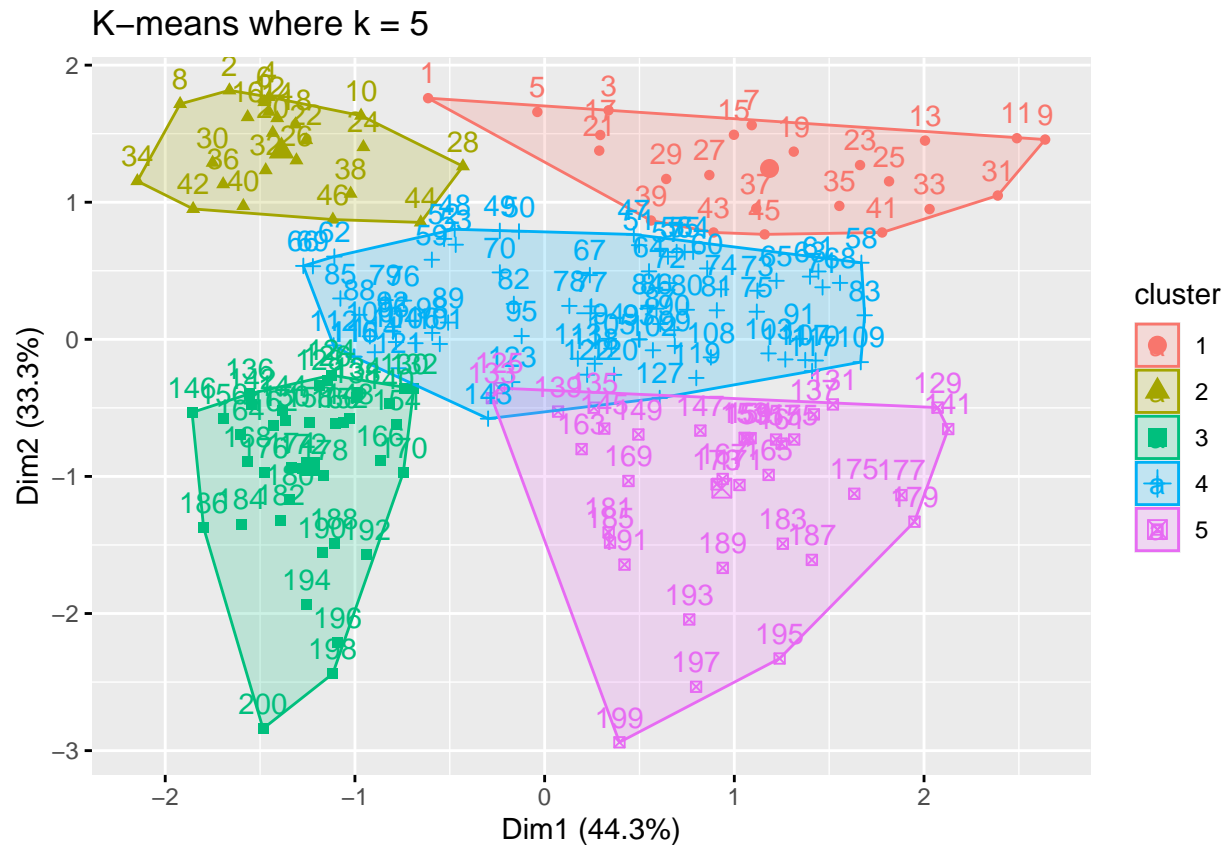
```
## 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
## 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
## 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
## 1 2 1 2 1 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
## 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
## 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
## 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
## 4 4 4 3 5 3 4 3 5 3 5 3 5 3 5 3 5 3 5 3
## 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
## 5 3 4 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3
## 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
## 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3
## 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200
## 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3 5 3
##
## Within cluster sum of squares by cluster:
## [1] 8948.609 4622.261 13972.359 30138.051 17669.500
## (between_SS / total_SS = 75.6 %)
##
## Available components:
##
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"
## [6] "betweenss" "size" "iter" "ifault"
```

*Visualizing the clusters*

```
set.seed(30)
ggplot(customer_f, aes(x = Annual.Income..k.., y = Spending.Score..1.100.)) +
  geom_point(stat = "identity", aes(color = as.factor(kmeans_model$cluster))) +
  scale_color_discrete(name = " ",
                       breaks = c("1", "2", "3", "4", "5"),
                       labels = c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4", "Cluster 5")) +
  ggtitle("Clusters of Mall Customers")
```



```
fviz_cluster(kmeans_model, data = customer_f) +  
  ggtitle("K-means where k = 5")
```



## Hierarchical Clustering (Agglomerative Method)

We will be using Euclidean distance metric

*Best linkage method to use*

```
x <- c("ward", "single", "complete", "average")
names(x) <- c("ward", "single", "complete", "average")

coeff <- function(x){
  agnes(customer_f, method = x)$ac
}

sapply(x, coeff) # 0.9835
```

```
##      ward    single complete  average
## 0.9835084 0.7031423 0.9535158 0.8978845
```

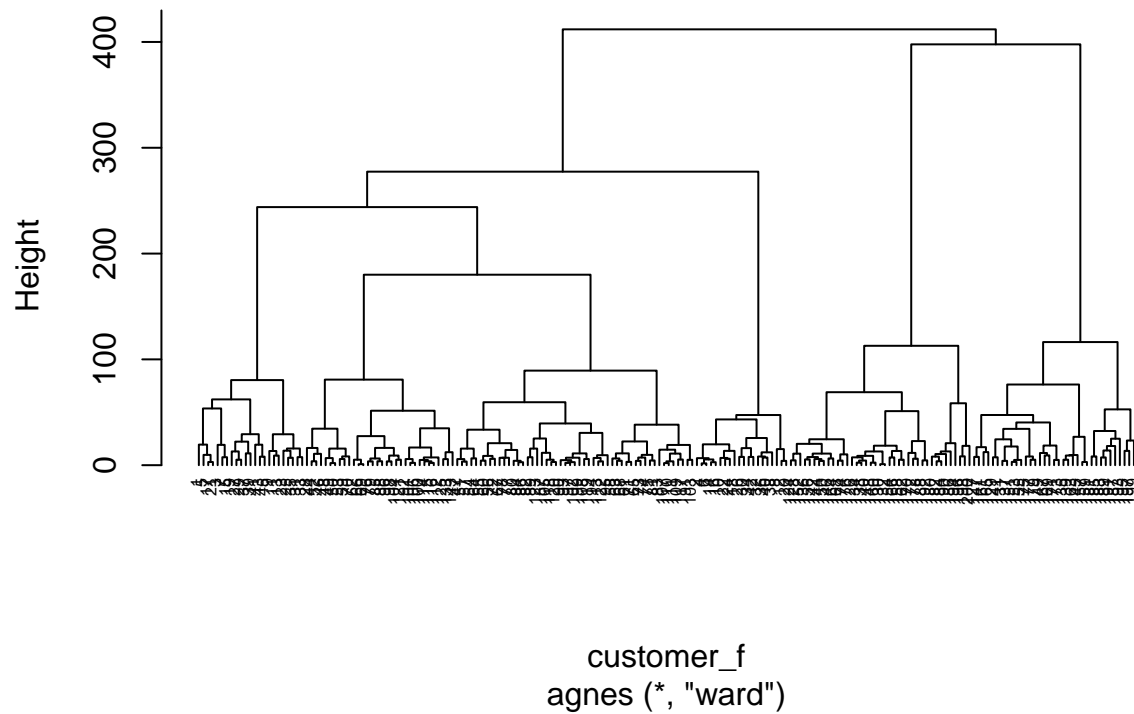
The best linkage method to use would be Ward's method so we will be using this for our clustering

```
hclust <- agnes(customer_f, method = "ward")

pltree(hclust, cex = 0.5, hang = -1, main = "Hierarchical Clustering Dendrogram")
```

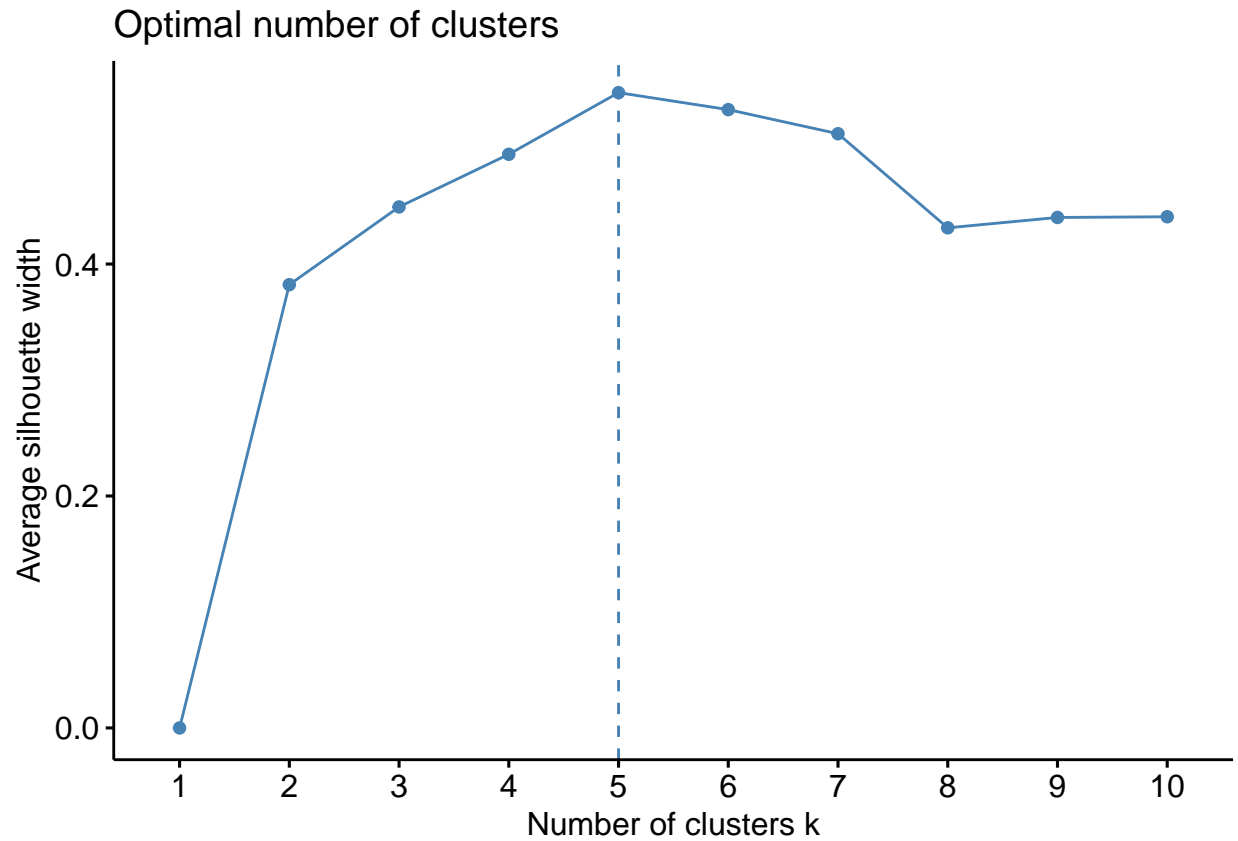


## Hierarchical Clustering Dendrogram



*Optimal cluster*

```
fviz_nbclust(pca_score, FUN = hcut, method = "silhouette")
```



We are using silhouette method here instead of elbow method, both of which shows to use  $k = 5$ .

*Cutting the dendrogram at  $k = 5$*

```
final_hclust <- cutree(hclust, k = 5)

fviz_cluster(list(cluster = final_hclust, data = customer_f)) +
  ggtitle("Hierarchical Cluster where k = 5")
```

Hierarchical Cluster where k = 5

