Final Report

Noah Estrada-Rand, Brady Hoskins, Charles Filce

12/12/2019

Data Cleaning

```
####cleaning the data
steam <- read.csv("steam.csv")
nrow(steam[!complete.cases(steam),])
```

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

```
steam$price <- steam$price *1.28
steam<- subset(steam, select = -c(appid,english,steamspy_tags,
                                  name,release_date,
                                  platforms,publisher,developer))
steam$genres <- do.call('rbind',strsplit(as.character(steam$genres), ';', fixed=TRUE))[,1]
```

```
## Warning in rbind("Action", "Action", "Action", "Action", "Action",
## "Action", : number of columns of result is not a multiple of vector length
## (arg 23)
```

```
steam$categories <- do.call('rbind',strsplit(as.character(steam$categories), ';', fixed=TRUE))[,1]
```

```
## Warning in rbind(c("Multi-player", "Online Multi-Player", "Local Multi-
## Player", : number of columns of result is not a multiple of vector length
## (arg 1)
```

```
steam$categories <- as.factor(steam$categories)
steam$genres <- as.factor(steam$genres)
steam <- steam[steam$average_playtime < 100000,]
steam <- steam[steam$price <50,]
steam <- steam[steam$average_playtime < 40000,]
steam <- steam[steam$negative_ratings < 2e+05,]
steam <- steam[steam$positive_ratings < 1e+06,]
steam$simple_categories <- fct_lump(steam$categories,n = 4)
steam$categories <- ifelse(steam$categories == "Single-player","SinglePlayer",
                           ifelse(steam$categories == "Multi-player","Multi-Player",
                                  ifelse(steam$categories == "Online Multi-Player","Multi-Player",
                                         ifelse(steam$categories == "Local Multi-Player","Multi-Player"
                                                ifelse(steam$categories == "MMO","MMO",
                                                       ifelse(steam$categories == "Co-op","Co-op",
                                                              ifelse(steam$categories == "Shared/Split
                                                                     ifelse(steam$categories == "Local
                                                                            ifelse(steam$categories ==
                                                                                   ifelse(steam$categor
                                                                                          ifelse(steam$
                                                                                                 ifelse


steam$categories <- as.factor(steam$categories)
steam$genres <- fct_lump(steam$genres,n = 5)
steam$successfulGame <- ifelse(steam$owners == "10000000-20000000",1,
                               ifelse(steam$owners == "20000000-50000000",1,
```

```
                                    ifelse(steam$owners == "50000000-100000000",1,
                                           ifelse(steam$owners == "100000000-200000000",1,
                                                  ifelse(steam$owners == "5000000-10000000",1,
                                                         ifelse(steam$owners == "2000000-5000000",1,
                                                                ifelse(steam$owners == "1000000-20000

steam <- subset(steam, select = -c(owners,categories,positive_ratings,negative_ratings,
                                   median_playtime))
steam$required_age <- as.factor(steam$required_age)
```

Summary Statistics

Sample of Data

```
head(steam,5)
```

```
##   required_age genres achievements average_playtime  price
## 1            0 Action            0            17612 9.2032
## 2            0 Action            0              277 5.1072
## 3            0 Action            0              187 5.1072
## 4            0 Action            0              258 5.1072
## 5            0 Action            0              624 5.1072
##   simple_categories successfulGame
## 1      Multi-player              1
## 2      Multi-player              1
## 3      Multi-player              1
## 4      Multi-player              1
## 5     Single-player              1
```

Descriptive Statistics

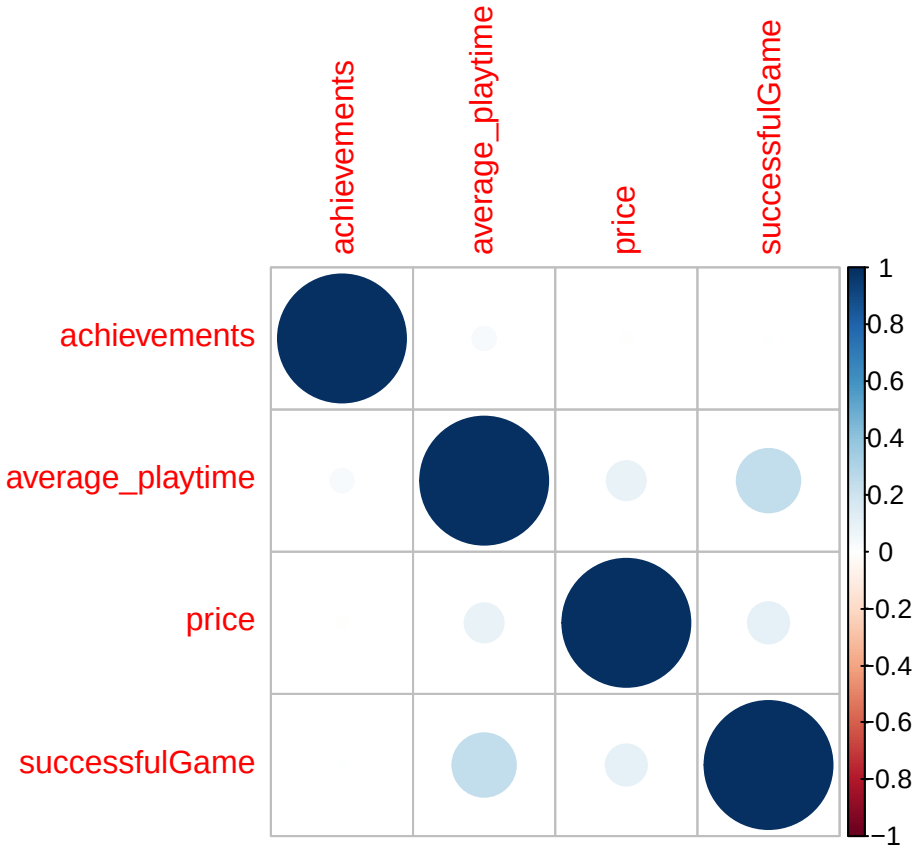
```
descr(steam)
```

```
## Non-numerical variable(s) ignored: required_age, genres, simple_categories
```

```
## Descriptive Statistics
## steam
## N: 26876
##
##                     achievements   average_playtime      price   successfulGame
## ----------------- -------------- ------------------ ---------- ----------------
##              Mean          45.31             117.46       7.33             0.02
##           Std.Dev         353.96             827.80       7.34             0.14
##               Min           0.00               0.00       0.00             0.00
##                Q1           0.00               0.00       2.16             0.00
##            Median           7.00               0.00       5.11             0.00
##                Q3          23.00               0.00       9.20             0.00
##               Max        9821.00           38805.00      49.91             1.00
##               MAD          10.38               0.00       5.69             0.00
##               IQR          23.00               0.00       7.04             0.00
##                CV           7.81               7.05       1.00             7.04
##          Skewness          13.38              22.16       1.89             6.89
##       SE.Skewness           0.01               0.01       0.01             0.01
```

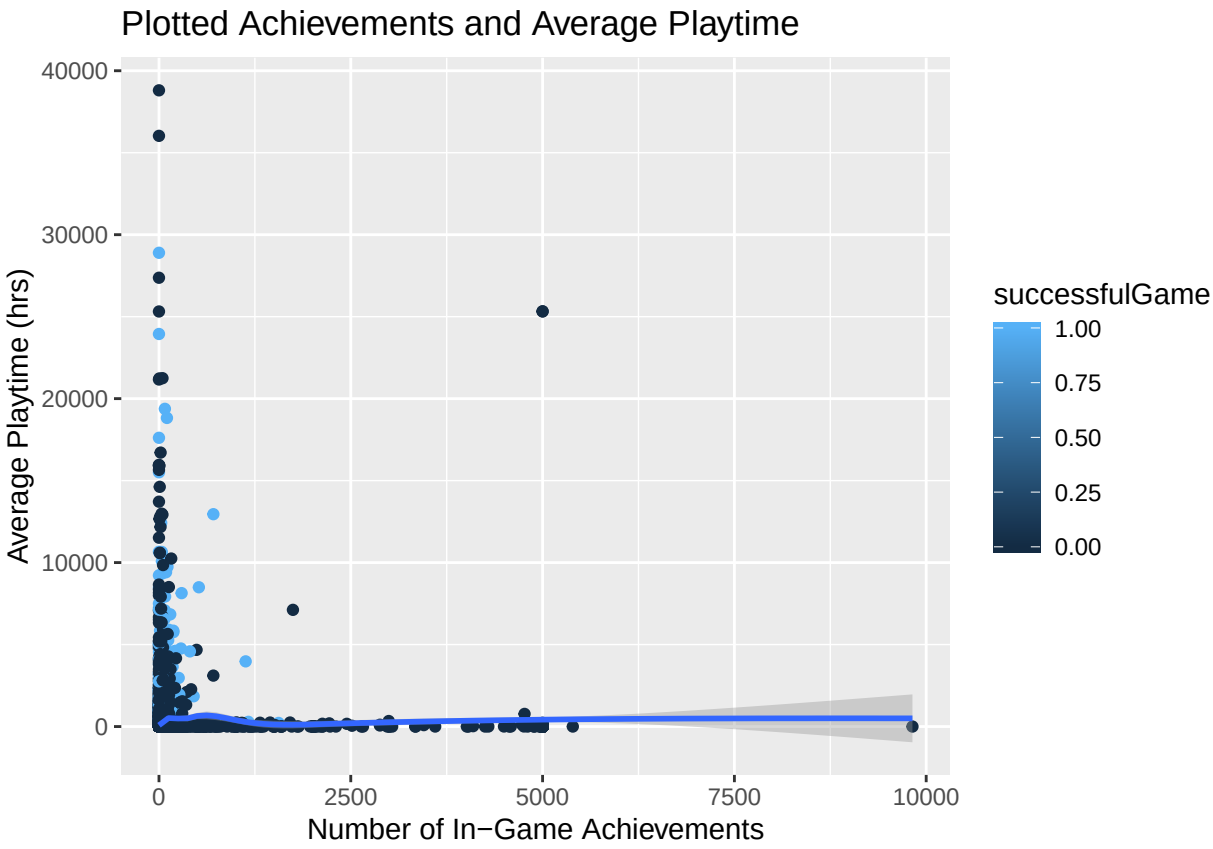
```
##          Kurtosis      189.74      676.18     4.66       45.54
##           N.Valid    26876.00    26876.00 26876.00    26876.00
##         Pct.Valid      100.00      100.00   100.00      100.00
```

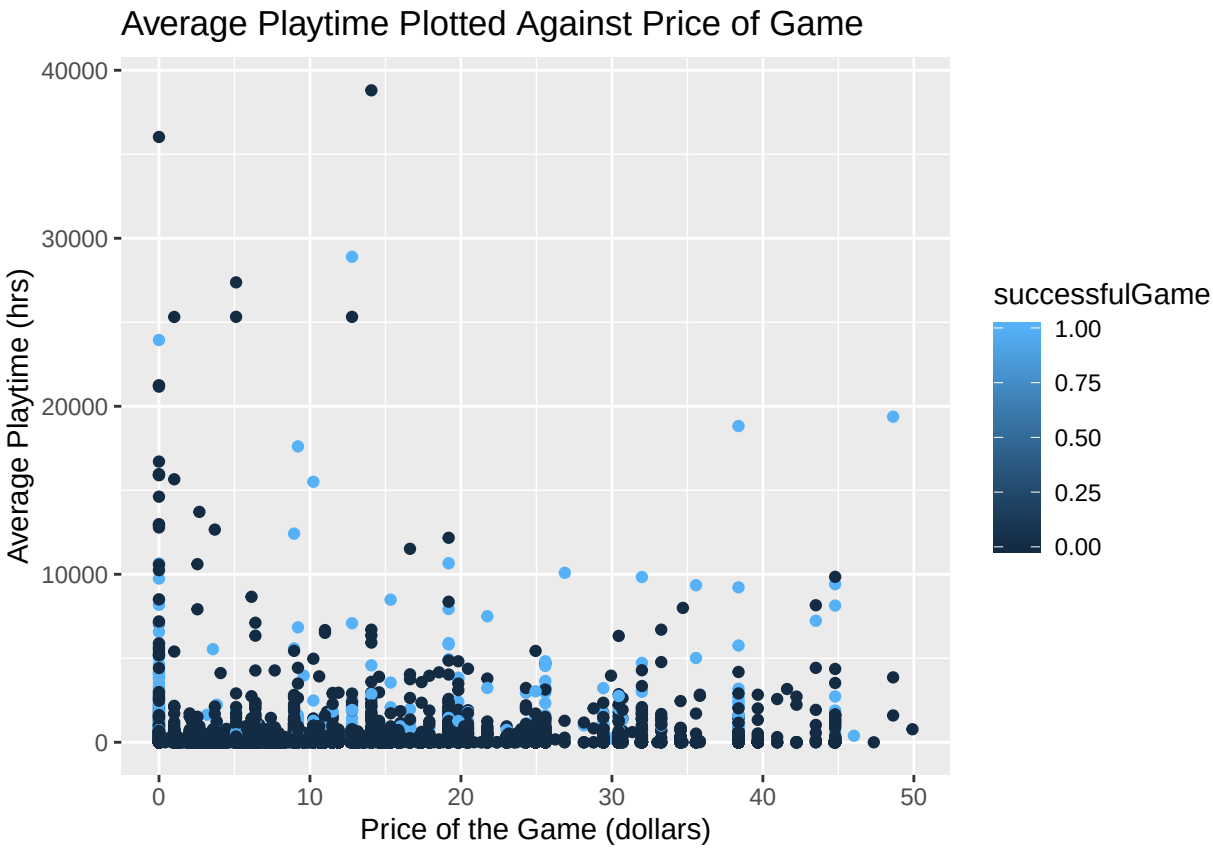
Correlation Matrix

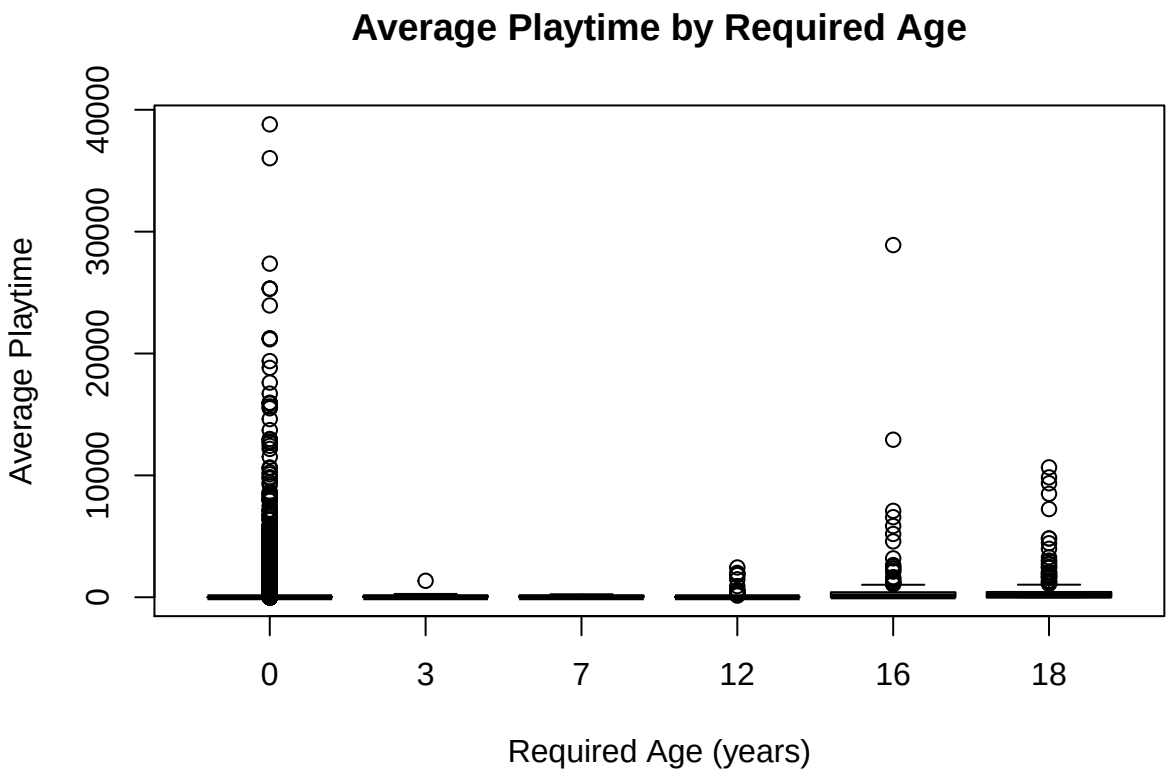


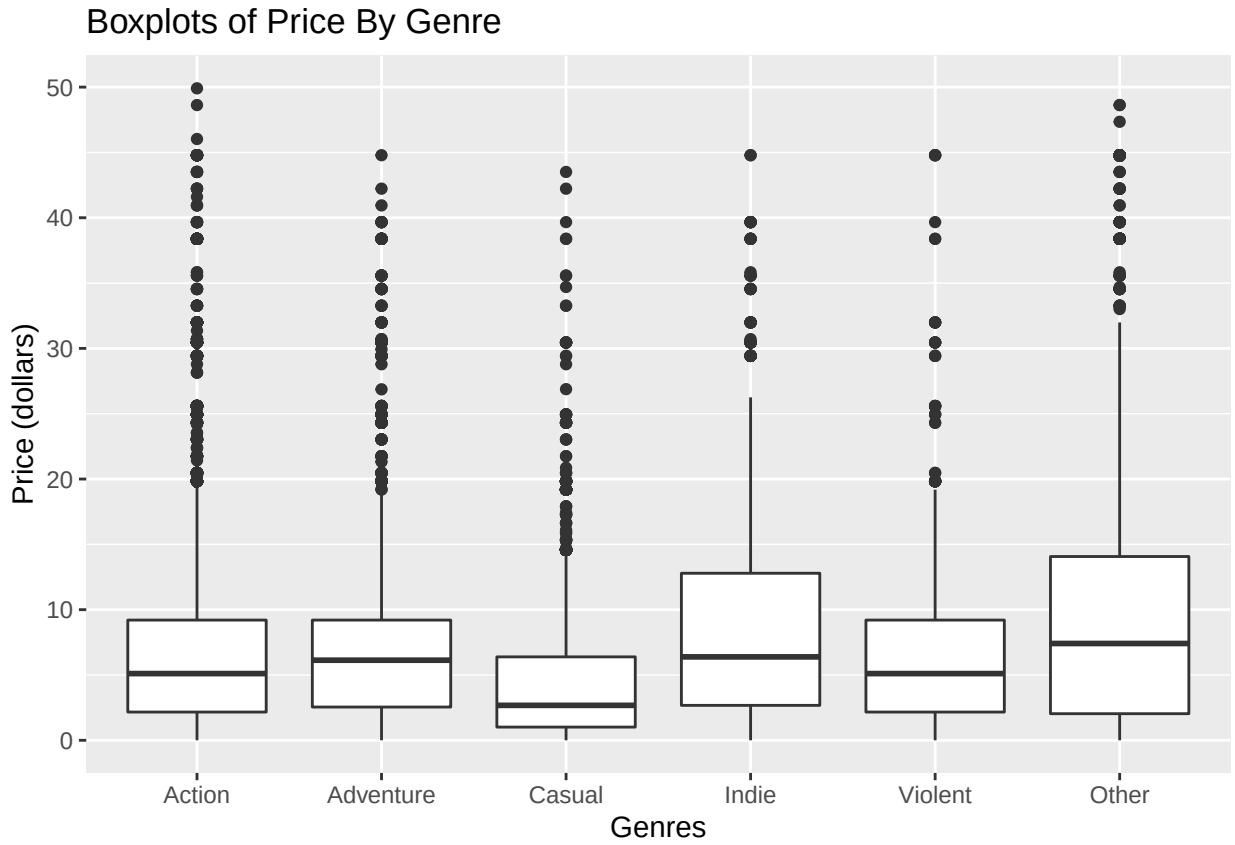
Summary Plots

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```


Plotted Achievements and Average Playtime
Average Playtime (hrs)
40000
30000
20000
10000
0
0
2500
5000
7500
10000
Number of In−Game Achievements
successfulGame
1.00
0.75
0.50
0.25
0.00


Average Playtime Plotted Against Price of Game
Average Playtime (hrs)
40000
30000
20000
10000
0
0
10
20
30
40
50
Price of the Game (dollars)
successfulGame
1.00
0.75
0.50
0.25
0.00


Average Playtime by Required Age
Average Playtime
0
10000
20000
30000
40000
0
3
7
12
16
18
Required Age (years)



##Logistic Model

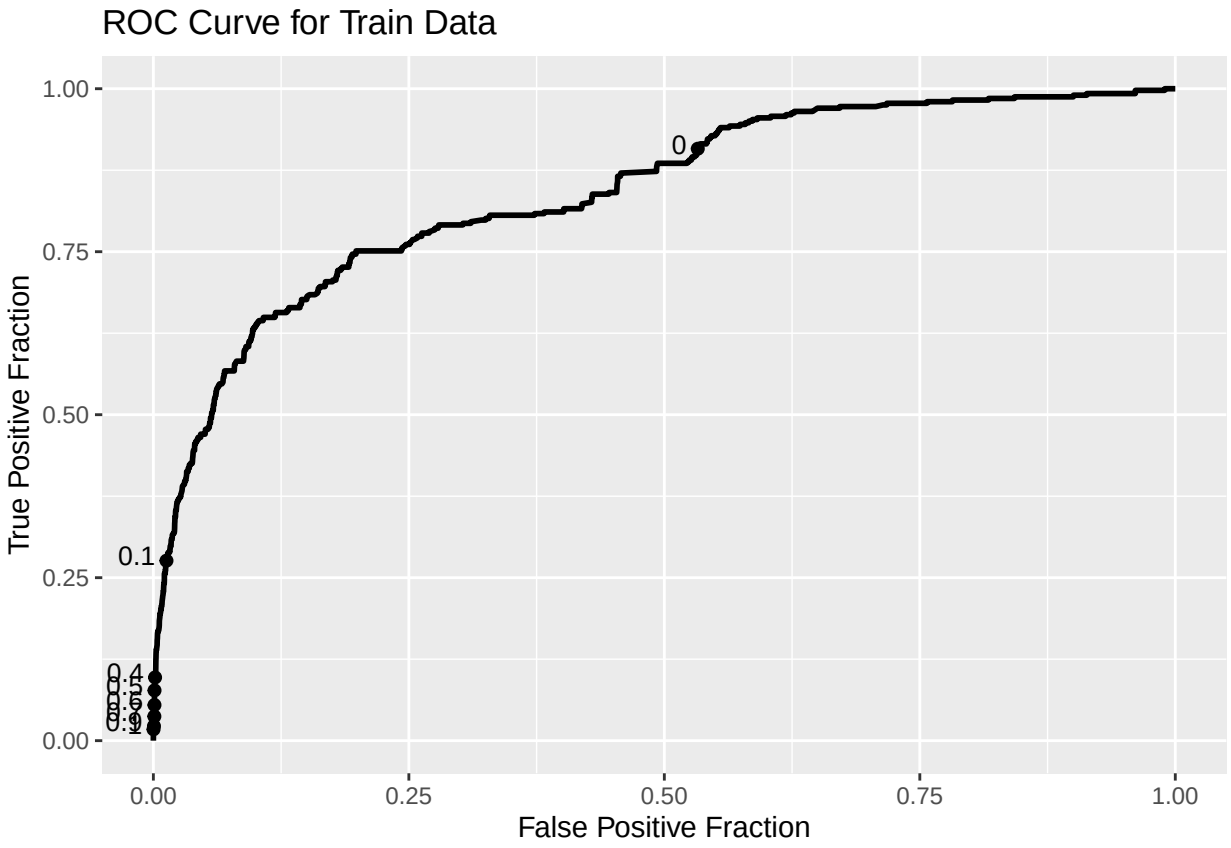
```
##
## Call:
## glm(formula = successfulGame ~ price + relevel(genres, ref = "Other") +
##     required_age + average_playtime, family = "binomial", data = steam_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8585  -0.2032  -0.1778  -0.0981   3.5279
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                             -3.909e+00  1.399e-01 -27.938
## price                                    3.500e-02  5.243e-03   6.675
## relevel(genres, ref = "Other")Action    -2.296e-01  1.364e-01  -1.683
## relevel(genres, ref = "Other")Adventure -1.460e+00  2.200e-01  -6.639
## relevel(genres, ref = "Other")Casual    -1.969e+00  3.093e-01  -6.366
## relevel(genres, ref = "Other")Indie     -1.526e+00  2.872e-01  -5.312
## relevel(genres, ref = "Other")Violent   -2.476e+00  7.244e-01  -3.418
## required_age3                           -9.917e+00  3.002e+02  -0.033
## required_age7                           -1.068e+01  2.950e+02  -0.036
## required_age12                           1.003e+00  6.096e-01   1.645
## required_age16                           1.696e+00  2.798e-01   6.063
## required_age18                           2.345e+00  1.938e-01  12.101
## average_playtime                         4.766e-04  3.618e-05  13.172
##                                         Pr(>|z|)
```

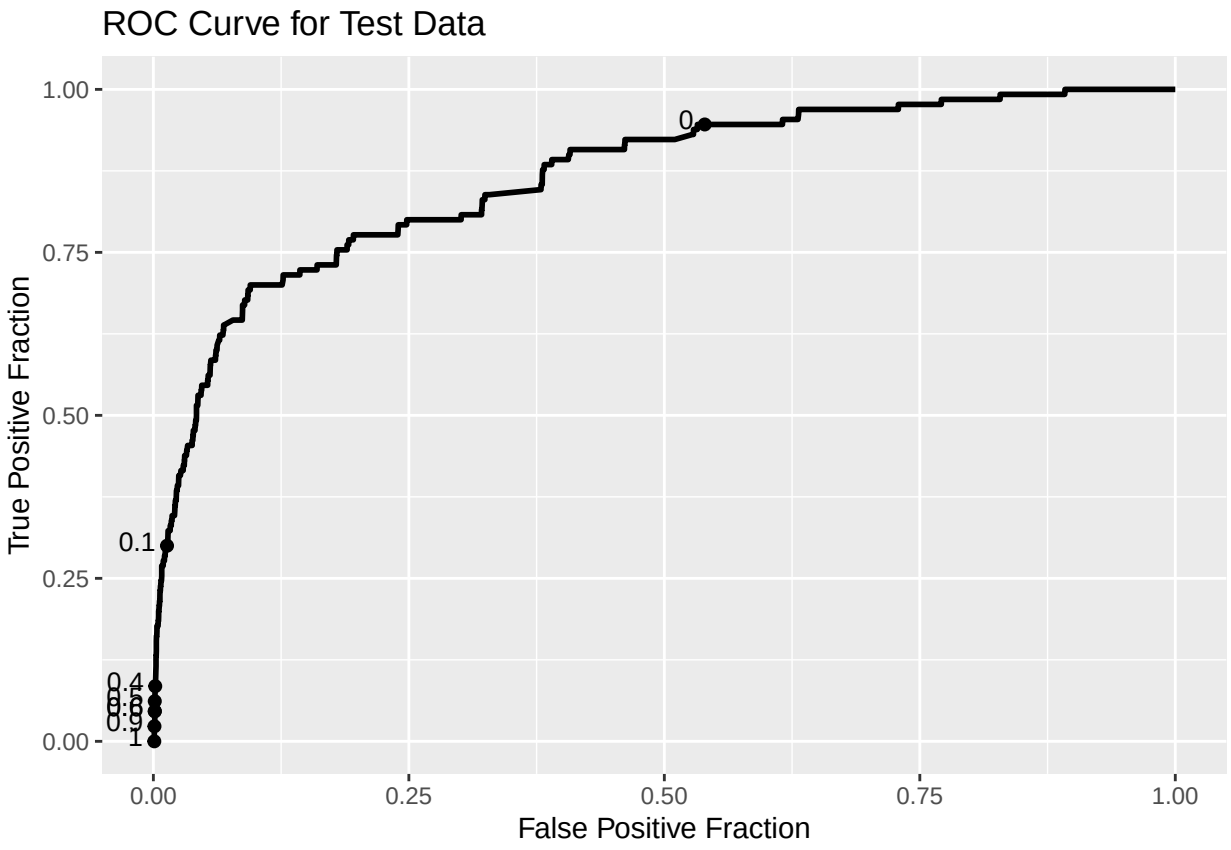
```
## (Intercept)                                < 2e-16 ***
## price                                     2.47e-11 ***
## relevel(genres, ref = "Other")Action      0.092345 .
## relevel(genres, ref = "Other")Adventure 3.17e-11 ***
## relevel(genres, ref = "Other")Casual      1.95e-10 ***
## relevel(genres, ref = "Other")Indie       1.08e-07 ***
## relevel(genres, ref = "Other")Violent     0.000632 ***
## required_age3                             0.973649
## required_age7                             0.971114
## required_age12                            0.099876 .
## required_age16                            1.34e-09 ***
## required_age18                             < 2e-16 ***
## average_playtime                           < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3943.5  on 20156  degrees of freedom
## Residual deviance: 3219.4  on 20144  degrees of freedom
## AIC: 3245.4
##
## Number of Fisher Scoring iterations: 13
```

```
exp(steam_logit$coefficients)
```

```
##                              (Intercept)
##                             2.005074e-02
##                                    price
##                             1.035616e+00
##    relevel(genres, ref = "Other")Action
##                             7.948691e-01
## relevel(genres, ref = "Other")Adventure
##                             2.321590e-01
##    relevel(genres, ref = "Other")Casual
##                             1.396069e-01
##     relevel(genres, ref = "Other")Indie
##                             2.174737e-01
##   relevel(genres, ref = "Other")Violent
##                             8.408791e-02
##                            required_age3
##                             4.935156e-05
##                            required_age7
##                             2.293985e-05
##                           required_age12
##                             2.726689e+00
##                           required_age16
##                             5.454452e+00
##                           required_age18
##                             1.042955e+01
##                          average_playtime
##                             1.000477e+00
```

ROC Plots


ROC Curve for Train Data
True Positive Fraction
False Positive Fraction
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
0
0.1


ROC Curve for Test Data
True Positive Fraction
False Positive Fraction
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
0
0.1

Confusion Matrices

```
## [1] "Train Confusion Matrix"

##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  20157
##
##
##                          | steam_train$successfulGame
## steam_train$pred_class |         0 |         1 | Row Total |
## -----------------------|-----------|-----------|-----------|
##                      0 |     18916 |       217 |     19133 |
## -----------------------|-----------|-----------|-----------|
##                      1 |       839 |       185 |      1024 |
## -----------------------|-----------|-----------|-----------|
##           Column Total |     19755 |       402 |     20157 |
## -----------------------|-----------|-----------|-----------|
##
##

## [1] "Accuracy: 94.8%"

## [1] "Sensitivity: 46%"

## [1] "Specificity: 95.7%"

## [1] "Test Confusion Matrix"

##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  6719
##
##
##                          | steam_test$successfulGame
## steam_test$pred_class |         0 |         1 | Row Total |
## ----------------------|-----------|-----------|-----------|
##                     0 |      6310 |        63 |      6373 |
## ----------------------|-----------|-----------|-----------|
##                     1 |       279 |        67 |       346 |
## ----------------------|-----------|-----------|-----------|
##          Column Total |      6589 |       130 |      6719 |
## ----------------------|-----------|-----------|-----------|
##
##
```

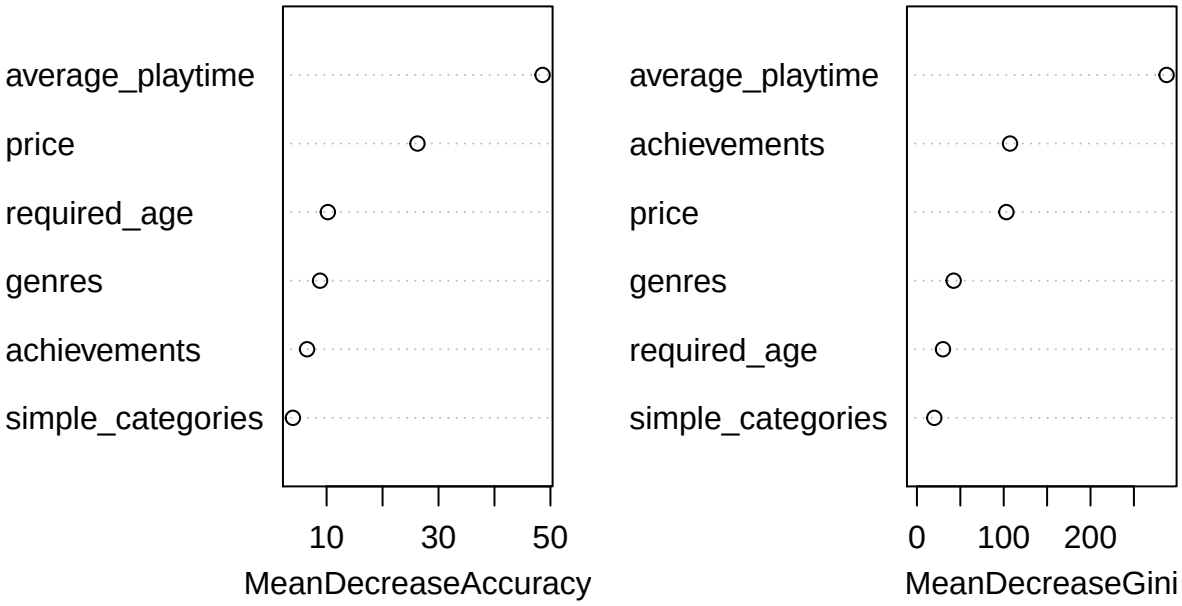
```
## [1] "Accuracy: 94.9%"
```

```
## [1] "Sensitivity: 51%"
```

```
## [1] "Specificity: 95.7%"
```

Random Forest Model

```
##                 Length Class  Mode
## call                 7  -none- call
## type                 1  -none- character
## predicted        20157  factor numeric
## err.rate          1500  -none- numeric
## confusion            6  -none- numeric
## votes            40314  matrix numeric
## oob.times        20157  -none- numeric
## classes              2  -none- character
## importance          24  -none- numeric
## importanceSD        18  -none- numeric
## localImportance      0  -none- NULL
## proximity            0  -none- NULL
## ntree                1  -none- numeric
## mtry                 1  -none- numeric
## forest              14  -none- list
## y                20157  factor numeric
## test                 0  -none- NULL
## inbag                0  -none- NULL
## terms                3  terms  call
```



##Confusion Matrices

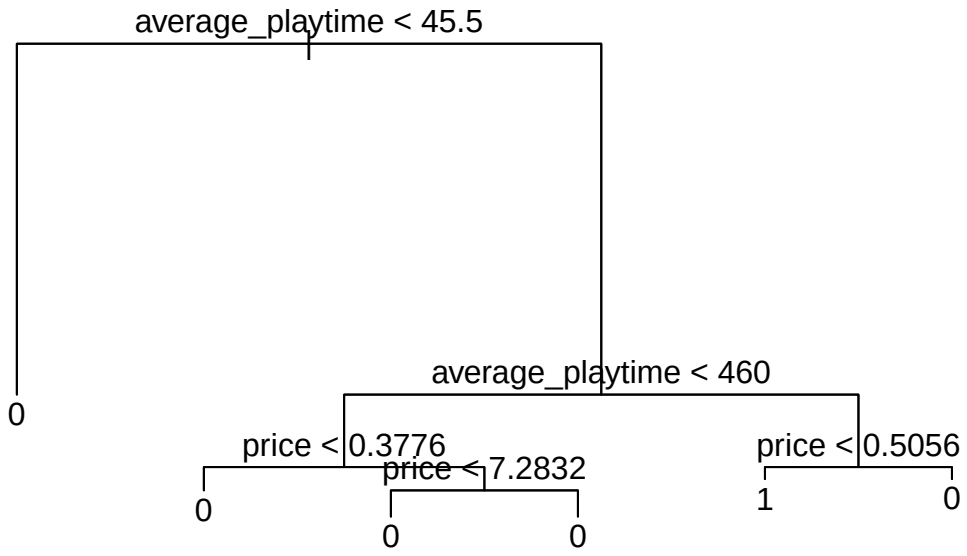
```
## [1] "Train Confusion Matrix"
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  20157
##
##
##                          | preds_2$preds
## preds_2$successfulGame |         0 |         1 | Row Total |
## -----------------------|-----------|-----------|-----------|
##                      0 |     19705 |        50 |     19755 |
## -----------------------|-----------|-----------|-----------|
##                      1 |       325 |        77 |       402 |
## -----------------------|-----------|-----------|-----------|
##           Column Total |     20030 |       127 |     20157 |
## -----------------------|-----------|-----------|-----------|
##
##
```

```
## [1] "Accuracy: 98.1%"
```

```
## [1] "Sensitivity: 60.6%"

## [1] "Specificity: 98.4"

## [1] "Test Confusion Matrix"

##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  6719
##
##
##                              | preds_test_2$preds
## preds_test_2$successfulGame |         0 |         1 | Row Total |
## ----------------------------|-----------|-----------|-----------|
##                           0 |      6581 |         8 |      6589 |
## ----------------------------|-----------|-----------|-----------|
##                           1 |       106 |        24 |       130 |
## ----------------------------|-----------|-----------|-----------|
##                Column Total |      6687 |        32 |      6719 |
## ----------------------------|-----------|-----------|-----------|
##
##

## [1] "Accuracy: 98.3%"

## [1] "Sensitivity: 75%"

## [1] "Specificity: 98.4"
```

Pruned Tree Model



Confusion Matrices

```
## [1] "Train Confusion Matrix"
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  20157
##
##
##                                  | basic_preds_train$preds
## basic_preds_train$successfulGame |         0 |         1 | Row Total |
## ---------------------------------|-----------|-----------|-----------|
##                                0 |     19699 |        56 |     19755 |
## ---------------------------------|-----------|-----------|-----------|
##                                1 |       337 |        65 |       402 |
## ---------------------------------|-----------|-----------|-----------|
##                     Column Total |     20036 |       121 |     20157 |
## ---------------------------------|-----------|-----------|-----------|
##
##
```

```
## [1] "Accuracy: 98%"

## [1] "Sensitivity: 42.9%"

## [1] "Specificity: 98.3"

## [1] "Test Conusion Matrix"

##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  6719
##
##
##                                 | basic_preds_test$preds
## basic_preds_test$successfulGame |         0 |         1 | Row Total |
## --------------------------------|-----------|-----------|-----------|
##                               0 |      6569 |        20 |      6589 |
## --------------------------------|-----------|-----------|-----------|
##                               1 |       115 |        15 |       130 |
## --------------------------------|-----------|-----------|-----------|
##                    Column Total |      6684 |        35 |      6719 |
## --------------------------------|-----------|-----------|-----------|
##
##

## [1] "Accuracy: 98.1%"

## [1] "Sensitivity: 53.7%"

## [1] "Specificity: 98.3"
```