Summary Statistics - Steam Store Data

Noah Estrada-Rand, Brady Hoskins, Charles Filce

11/12/2019

Variable Engineering

Check for Rows with Missing Values

```
nrow(steam[!complete.cases(steam),])
```

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

Transform Price into Dollars

```
steam$price <- steam$price *1.28
```

Subset the Data without Needless Columns

```
steam<- subset(steam, select = -c(appid,english,steamspy_tags,
                                  name,release_date,
                                  platforms))
```

Rid the Data set of All Cells Delimited by ";"

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

```
## Warning in rbind("Valve", "Valve", "Valve", "Valve", "Valve", "Valve",
## "Valve", : number of columns of result is not a multiple of vector length
## (arg 43)
```

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

```
## Warning in rbind("Valve", "Valve", "Valve", "Valve", "Gearbox Software", :
## number of columns of result is not a multiple of vector length (arg 26)
```

Make The Above Variables into Factors

```
steam$developer <- as.factor(steam$developer)
steam$categories <- as.factor(steam$categories)
```

```
steam$genres <- as.factor(steam$genres)
steam$publisher <- as.factor(steam$publisher)
```

Removal of Extreme Values

```
steam <- steam[steam$average_playtime < 100000,]
steam <- steam[steam$price <100,]
steam <- steam[steam$average_playtime < 40000,]
steam <- steam[steam$negative_ratings < 2e+05,]
steam <- steam[steam$positive_ratings < 1e+06,]
```

Creation of New Binary Dummy Variable to Signify A Successful or Unsuccessful Game (Successful defined as any game selling over one million copies)

```
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 == "1000000-200000
steam$successfulGame <- as.factor(steam$successfulGame)
```

Removal Of Owners and Developer Variables Since we are now predicting on Successful/Unsuccessful

```
steam <- subset(steam, select = -c(owners,developer))
```

Summary Statistics Table

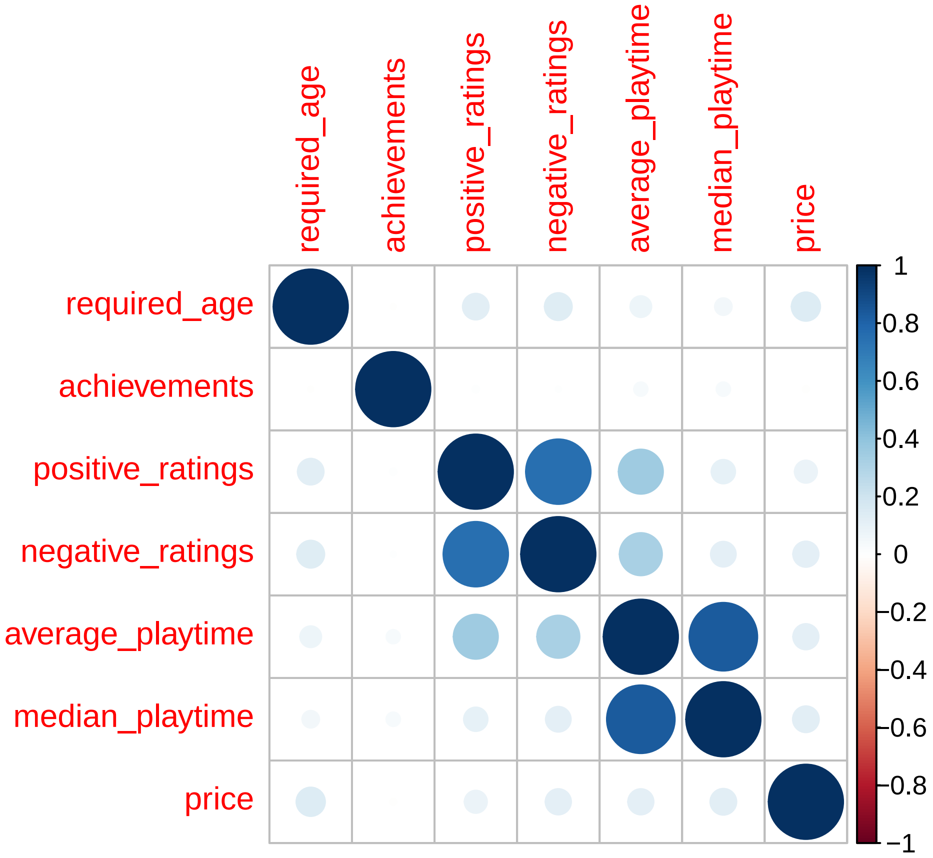
```
descr(steam)
```

```
## Non-numerical variable(s) ignored: publisher, categories, genres, successfulGame
```

```
## Descriptive Statistics
## steam
## N: 27048
##
##                     achievements   average_playtime   median_playtime   negative_ratings
## ----------------- -------------- ------------------ ----------------- ------------------
##              Mean          45.28             122.87            110.42             178.25
##           Std.Dev         352.84             844.46            782.64            1906.84
##               Min           0.00               0.00              0.00               0.00
##                Q1           0.00               0.00              0.00               2.00
##            Median           7.00               0.00              0.00               9.00
##                Q3          23.00               0.00              0.00              41.00
##               Max        9821.00           38805.00          38805.00          142079.00
##               MAD          10.38               0.00              0.00              11.86
##               IQR          23.00               0.00              0.00              39.00
##                CV           7.79               6.87              7.09              10.70
##          Skewness          13.42              21.32             27.43              44.68
##       SE.Skewness           0.01               0.01              0.01               0.01
##          Kurtosis         190.95             630.19           1007.48            2693.31
##           N.Valid       27048.00           27048.00          27048.00           27048.00
##         Pct.Valid         100.00             100.00            100.00             100.00
##
## Table: Table continues below
##
##
##
##                     positive_ratings      price   required_age
## ----------------- ------------------ ---------- --------------
##              Mean             885.05       7.66           0.36
##           Std.Dev            9669.62       8.42           2.41
##               Min               0.00       0.00           0.00
##                Q1               6.00       2.16           0.00
##            Median              24.00       5.11           0.00
##                Q3             125.00       9.20           0.00
##               Max          863507.00      97.27          18.00
##               MAD              32.62       5.69           0.00
##               IQR             119.00       7.04           0.00
##                CV              10.93       1.10           6.78
##          Skewness              44.04       2.77           6.75
##       SE.Skewness               0.01       0.01           0.01
##          Kurtosis            2960.71      12.42          44.04
##           N.Valid           27048.00   27048.00       27048.00
##         Pct.Valid             100.00     100.00         100.00
```

Basic Correlation Matrix

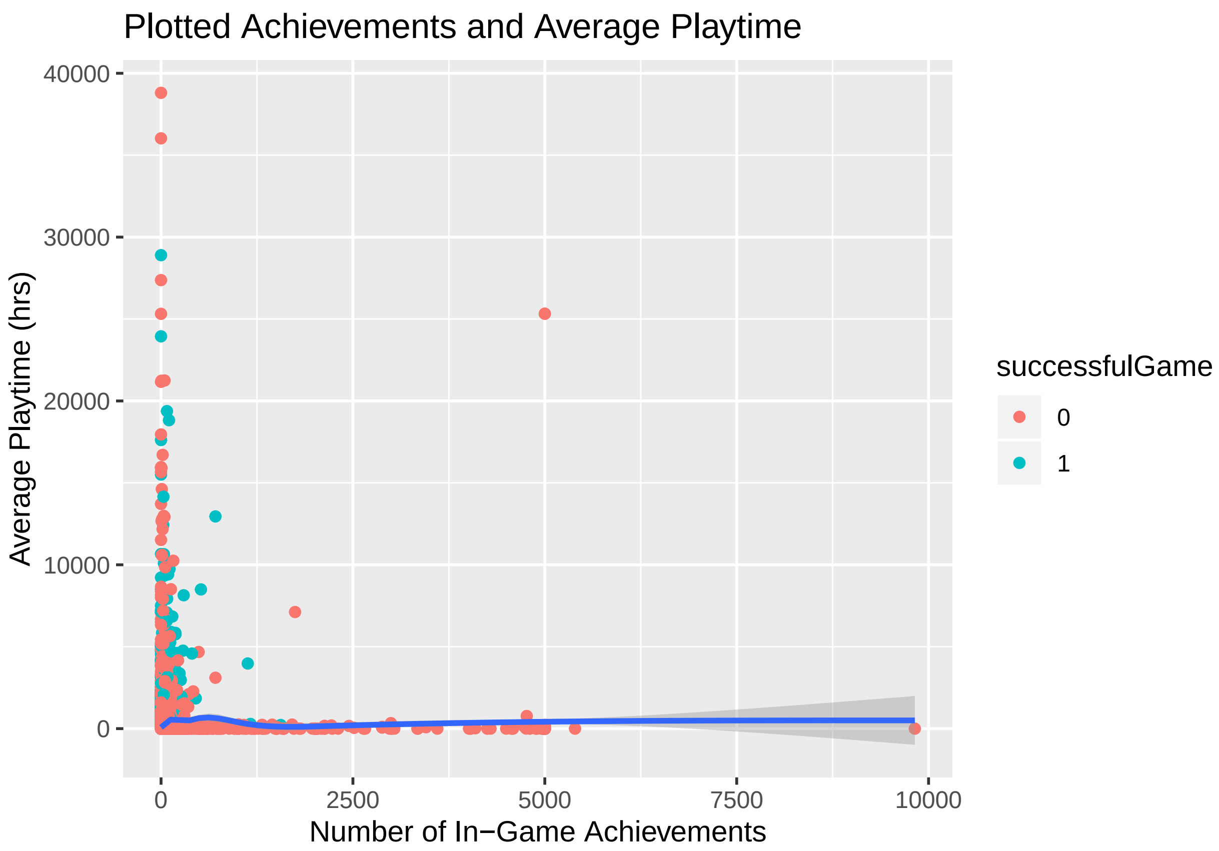
```
numeric_cols <- sapply(steam,is.numeric)
correlations <- cor(steam[,numeric_cols])
corrplot(correlations)
```



Plots to Illustrate Data and Relationship to SuccessfulGame Variable

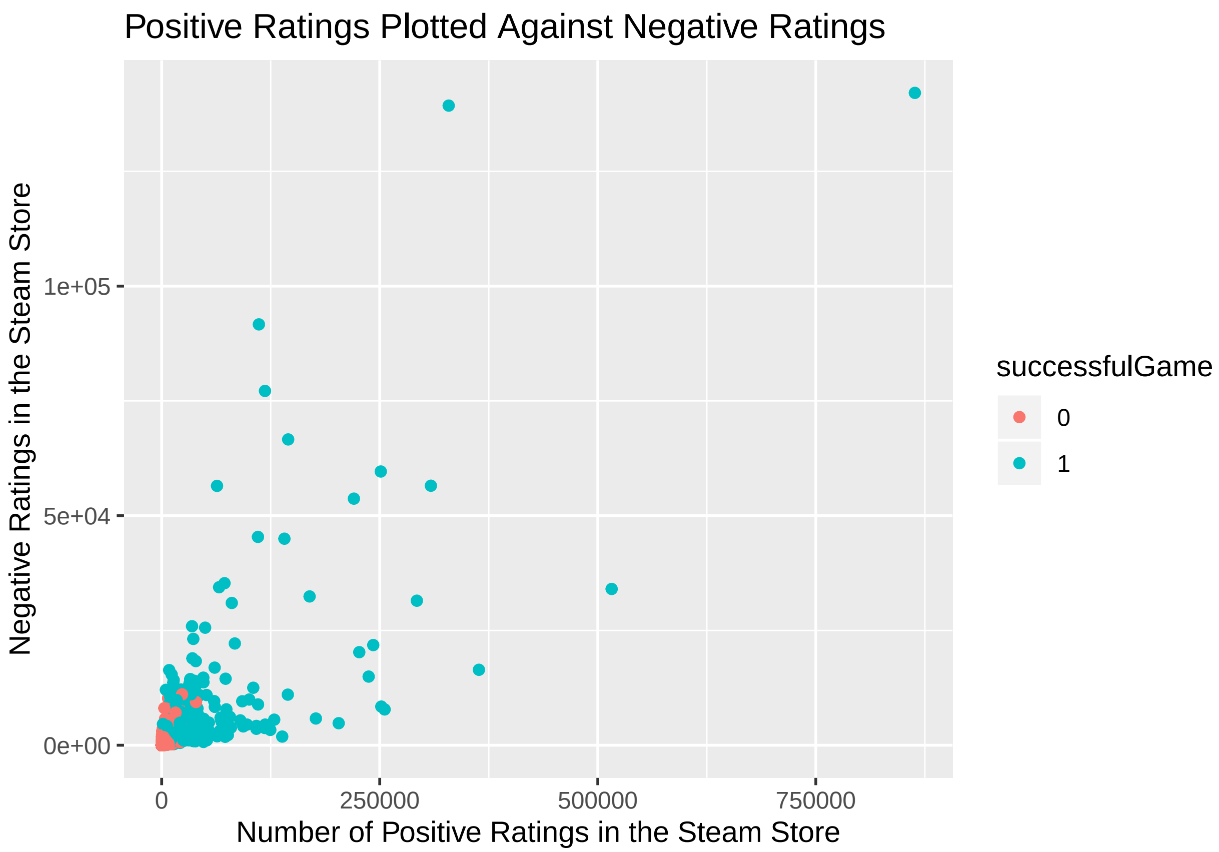
```
ggplot(steam,aes(x = achievements,y = average_playtime)) + geom_point(aes(color = successfulGame))+
  geom_smooth() + labs(x = "Number of In-Game Achievements",y = "Average Playtime (hrs)",
                      title = "Plotted Achievements and Average Playtime")
```

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



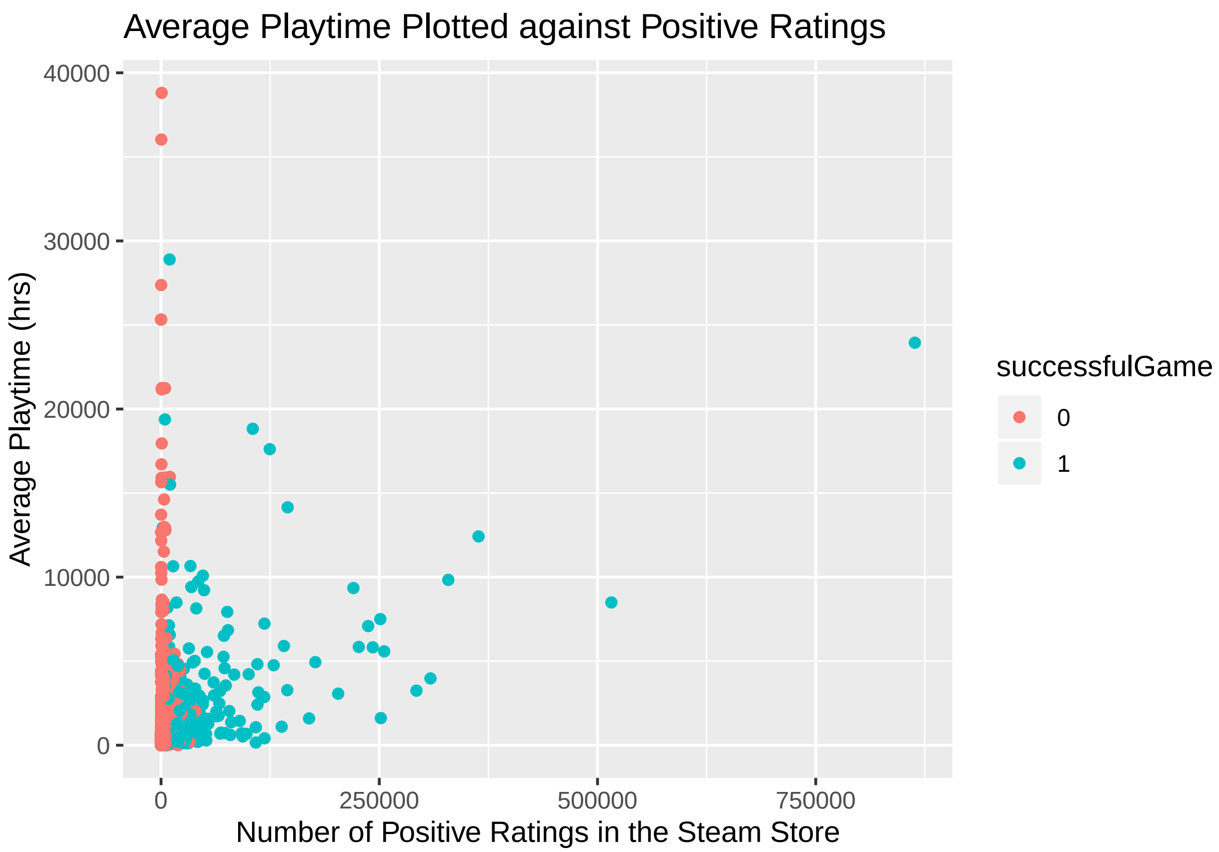
Here we observe that a large portion of successful games have lower number of in-game achievements, yet high amounts of playtime, indicating that perhaps loading the game with and overabundance of achievements and goals does not necessarily incentivize the player to continue playing. This is important when developing a game to be appealing to consumers.

```r
ggplot(steam,aes(x = positive_ratings,y = negative_ratings)) +
  geom_point(aes(color = successfulGame))+
  labs(x = "Number of Positive Ratings in the Steam Store",y = "Negative Ratings in the Steam Store",
       title = "Positive Ratings Plotted Against Negative Ratings")
```



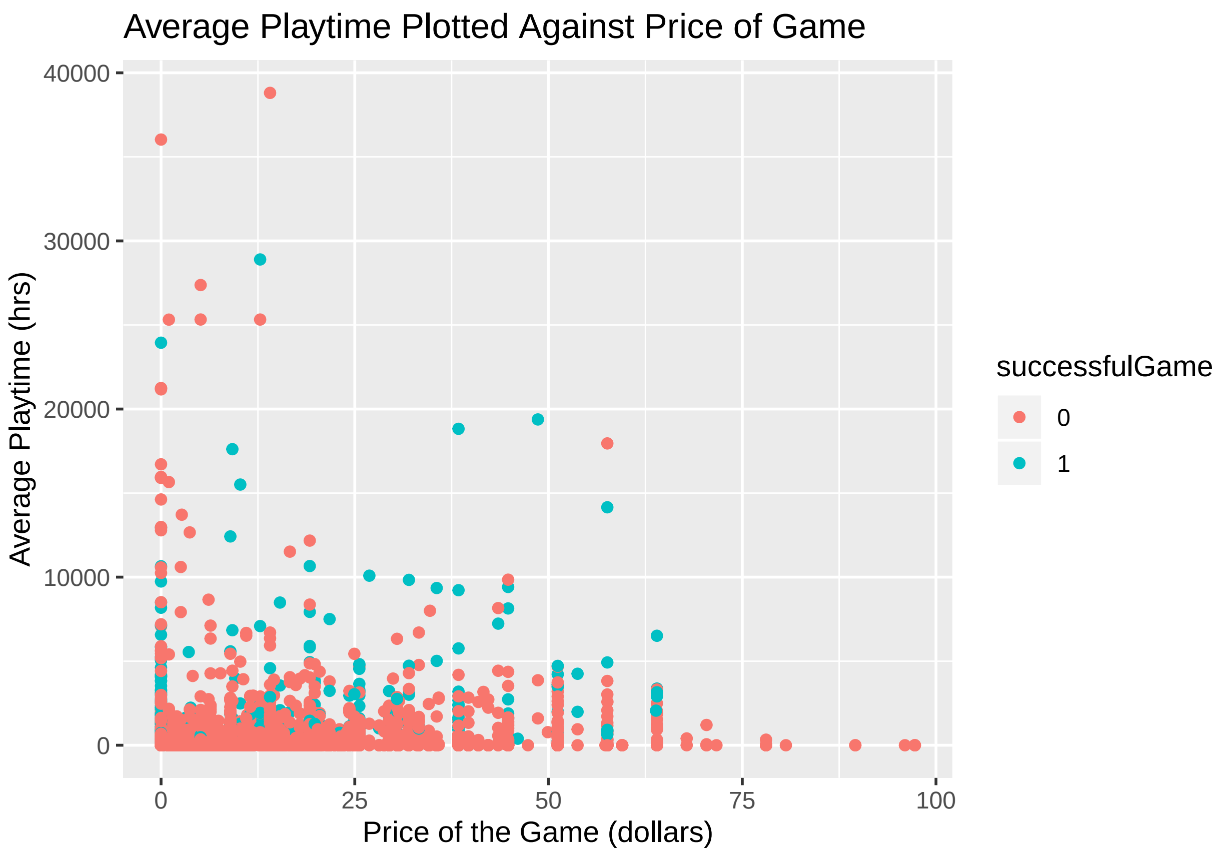
Here we observe that successful games have more of both types of ratings than unsuccessful games. The successful games gain significant comments and ratings on Steam whereas the lesser games do not apparently garner such attention.

```
ggplot(steam,aes(x = positive_ratings,y = average_playtime)) +geom_point(aes(color = successfulGame))+
  labs(x = "Number of Positive Ratings in the Steam Store",y = "Average Playtime (hrs)",
       title = "Average Playtime Plotted against Positive Ratings")
```



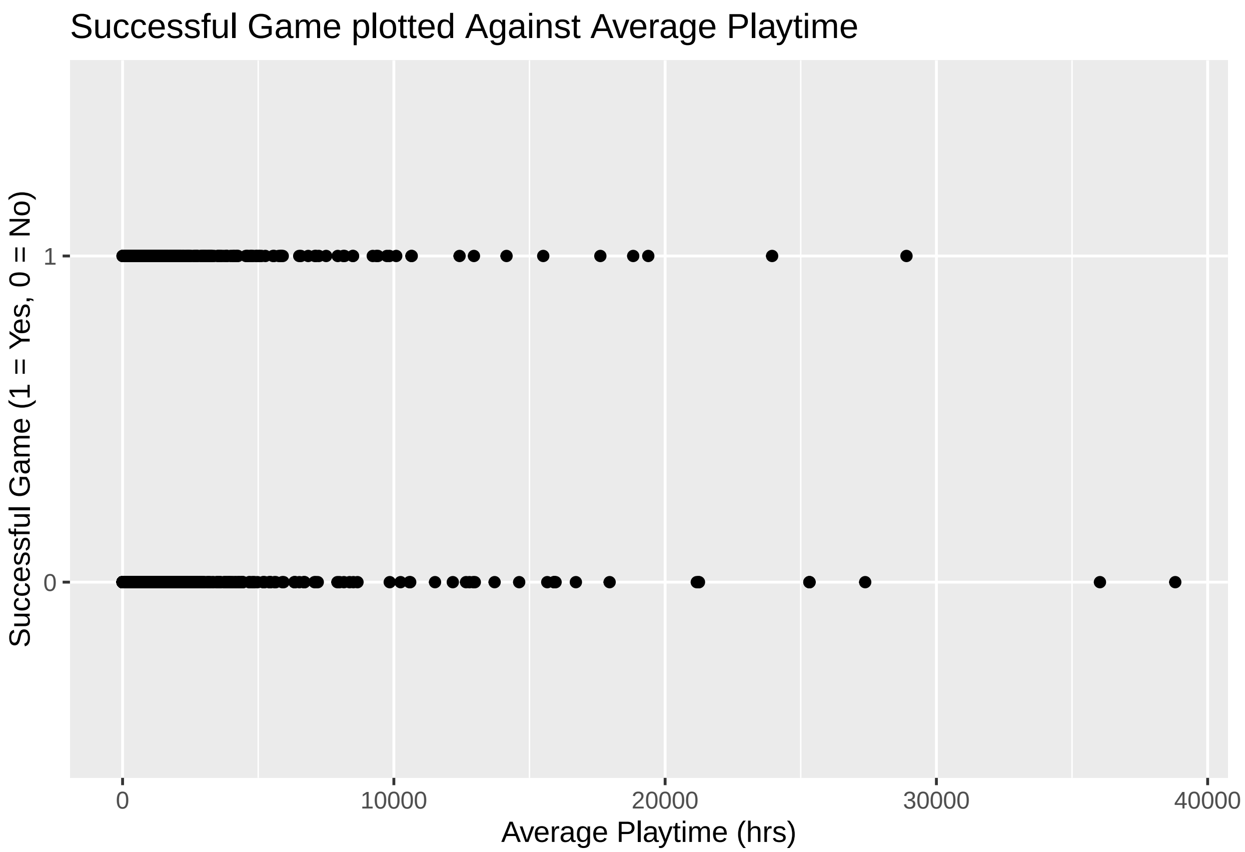
Interestingly, when considering the relationship between average playtime and positive ratings on the Steam store, it appears that higher positive ratings do not necessarily translate to more playtime on a particular game. Moreover, it appears that the successful games have some of the highest positive ratings but not necessarily the highest play time.

```
ggplot(steam,aes(x = price,y = average_playtime)) +
  geom_point(aes(color = successfulGame)) +
  labs(x = "Price of the Game (dollars)",y ="Average Playtime (hrs)",
       title = "Average Playtime Plotted Against Price of Game")
```



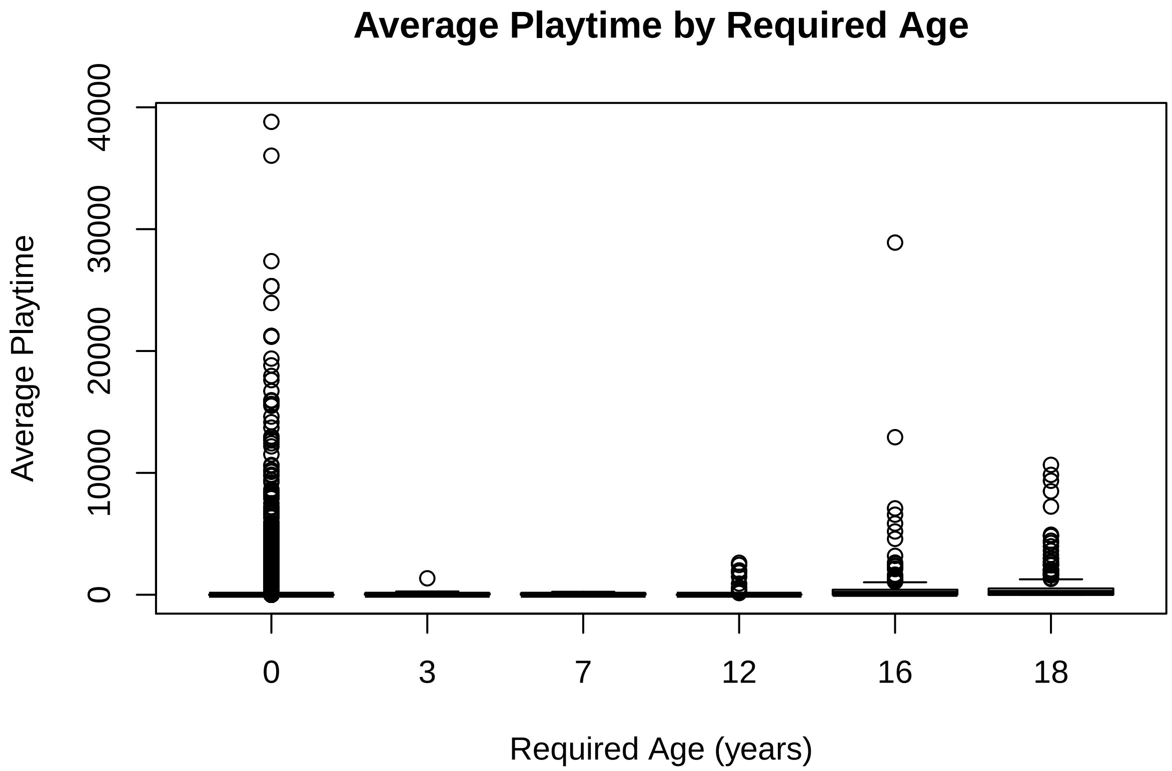
When plotting average playtime against price of a game, no clear relationship appears between the two variables. However, when considering the average playtime of successful games, it appears that, regardless of price, there is greater average playtime than non-successful games.

```
ggplot(steam,aes(x = average_playtime,y = successfulGame)) +
  geom_point()+
  labs(title = "Successful Game plotted Against Average Playtime",
       x = "Average Playtime (hrs)",y = "Successful Game (1 = Yes, 0 = No)")
```



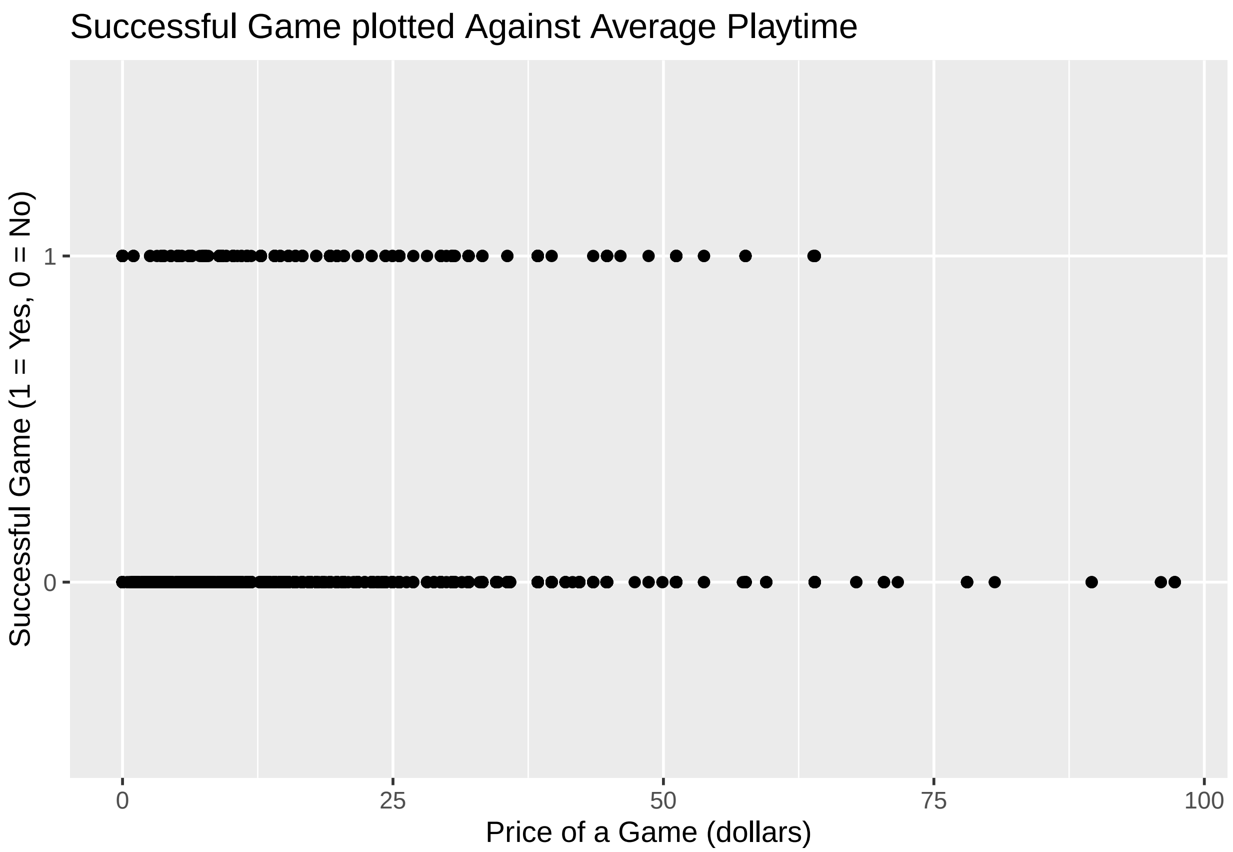
In the plot above, there is no clear separation between the successful and unsuccessful games. However, this plot is deceiving as the mean average playtime of successful games is much higher than that of non-successful games.

```
boxplot(steam$average_playtime~steam$required_age,
        main = "Average Playtime by Required Age",xlab = "Required Age (years)",
        ylab = "Average Playtime")
```



When looking at the boxplots of average playtime against required age for all games, it appears that those with no age requirements garner the highest play time. This intuitively makes sense as games with a broader age range have a large audience to have play their games.

```
ggplot(steam,aes(x = price,y = successfulGame)) +
  geom_point()+
  labs(title = "Successful Game plotted Against Average Playtime",
       x = "Price of a Game (dollars)",y = "Successful Game (1 = Yes, 0 = No)")
```



Here it appears as though non successful games have a larger spread of price while the more successful games stay in the lower price ranges. This leads us to believe that having too expensive of a game deters people from purchasing the game whil having the game at lower prices allows people to buy more of the game.