

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,steampy_tags,  
                                name,release_date,  
                                platforms,publisher,developer))
```

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)
```

Make The Above Variables into Factors

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

Removal of Extreme Values

```
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,]
```

Categorizing Various Categories into Less Categories to reduce Dimensionality

```
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 Screen",  
                            ifelse(steam$categories == "Local Co-operative",  
                                ifelse(steam$categories == "Other",  
                                    ifelse(steam$categories == "",  
                                        ifelse(steam$categories == "  
steam$categories <- as.factor(steam$categories)  
unique(steam$categories)
```

```
## [1] Multi-Player SinglePlayer Other      Steam      MMO
## [6] Co-op
## Levels: Co-op MMO Multi-Player Other SinglePlayer Steam
```

Grouping Genres into Larger Umbrella Genres

```
steam$genres <- fct_lump(steam$genres,n = 5)
unique(steam$genres)

## [1] Action      Indie        Other        Casual      Adventure Violent
## Levels: Action Adventure Casual Indie Violent Other
```

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 == "1000000-2000000",1,
                                                                      steam$successfulGame)))
                                                    )))
steam$successfulGame <- as.factor(steam$successfulGame)
```

Removal Of Owners Variable

Since we are now predicting on Successful/Unsuccessful

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

Summary Statistics Table

```
descr(steam)
```

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

```
## Descriptive Statistics
```

```
## steam
```

```
## N: 26876
```

```
##
```

	achievements	average_playtime	median_playtime	negative_ratings
Mean	45.31	117.46	105.56	162.85
Std.Dev	353.96	827.80	769.94	1796.90
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	40.00
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	38.00
CV	7.81	7.05	7.29	11.03
Skewness	13.38	22.16	28.47	50.17
SE.Skewness	0.01	0.01	0.01	0.01
Kurtosis	189.74	676.18	1072.34	3325.75
N.Valid	26876.00	26876.00	26876.00	26876.00
Pct.Valid	100.00	100.00	100.00	100.00

```
##
```

```
## Table: Table continues below
```

```
##
```

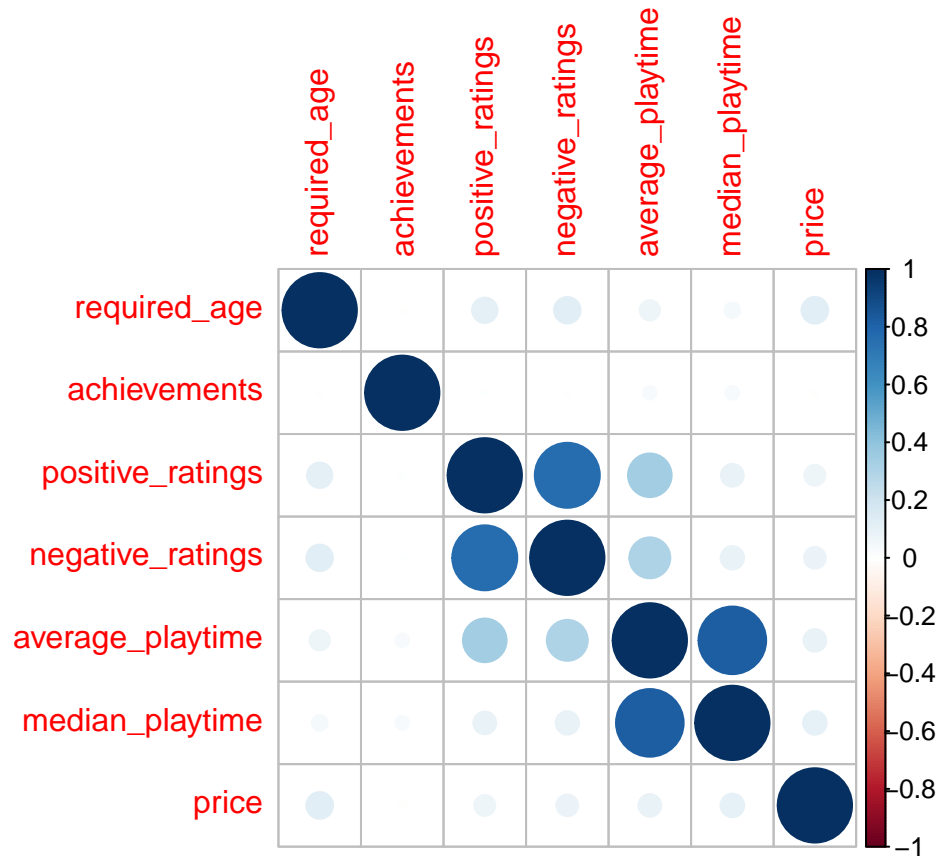
```
##
```

```
##
```

	positive_ratings	price	required_age
Mean	848.56	7.33	0.34
Std.Dev	9594.90	7.34	2.35
Min	0.00	0.00	0.00
Q1	6.00	2.16	0.00
Median	24.00	5.11	0.00
Q3	122.00	9.20	0.00
Max	863507.00	49.91	18.00
MAD	32.62	5.69	0.00
IQR	116.00	7.04	0.00
CV	11.31	1.00	6.93
Skewness	45.17	1.89	6.90
SE.Skewness	0.01	0.01	0.01
Kurtosis	3071.79	4.66	46.24
N.Valid	26876.00	26876.00	26876.00
Pct.Valid	100.00	100.00	100.00

Basic Correlation Matrix

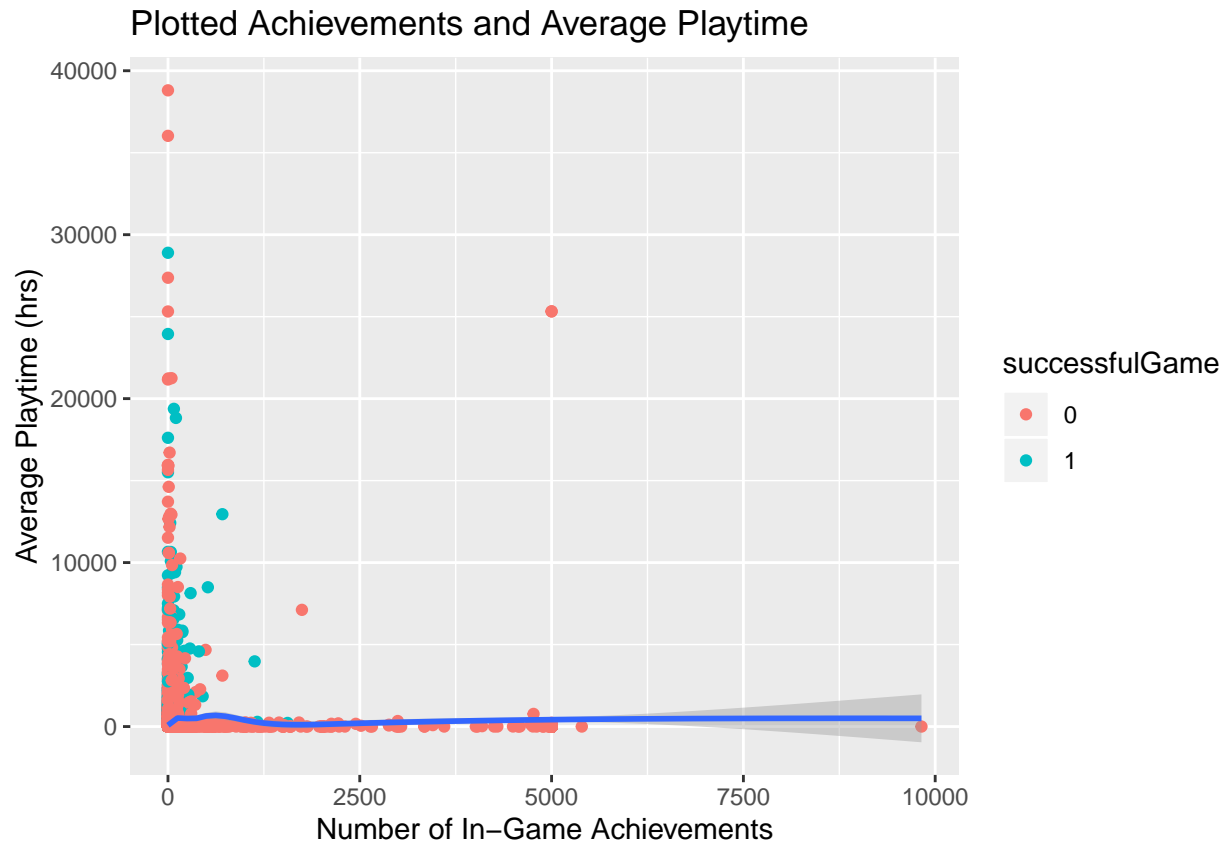
```
numeric_cols <- sapply(steam,is.numeric)
correlations <- cor(steam[,numeric_cols])
corrplot(correlations)
```



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")
```

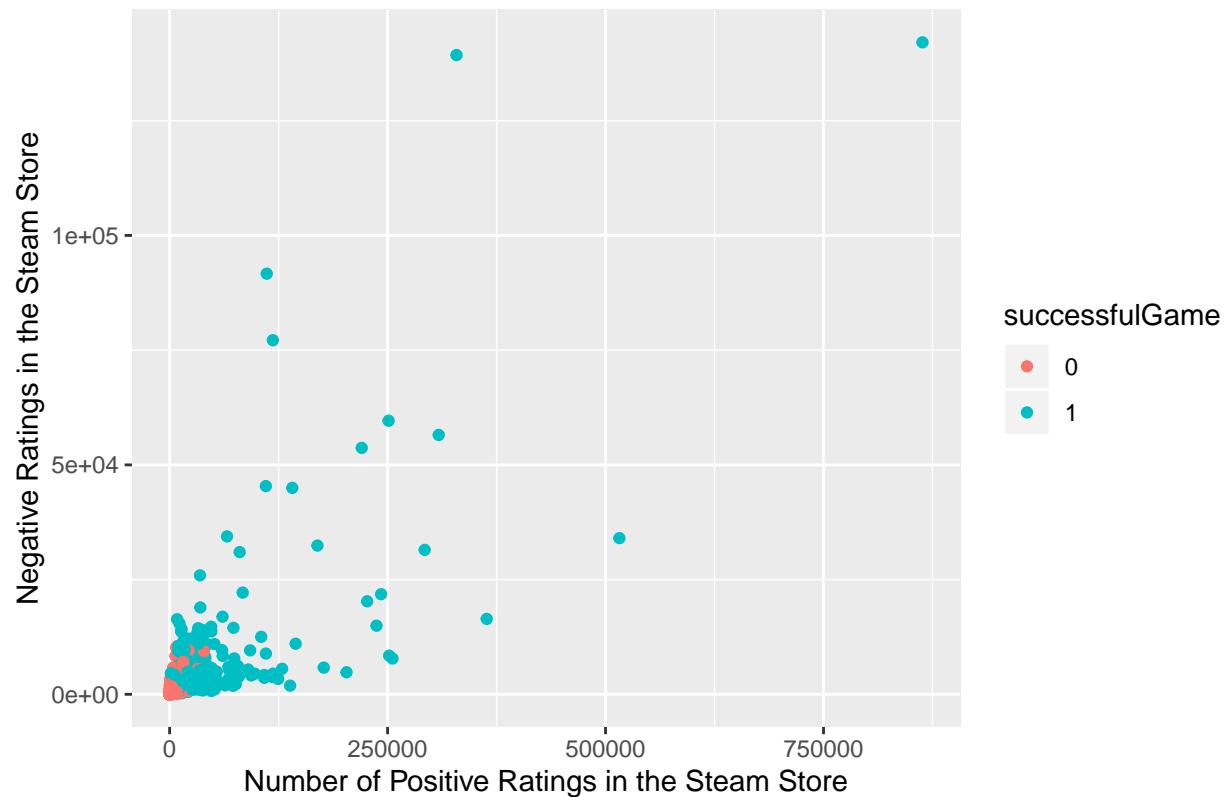
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



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 an 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.

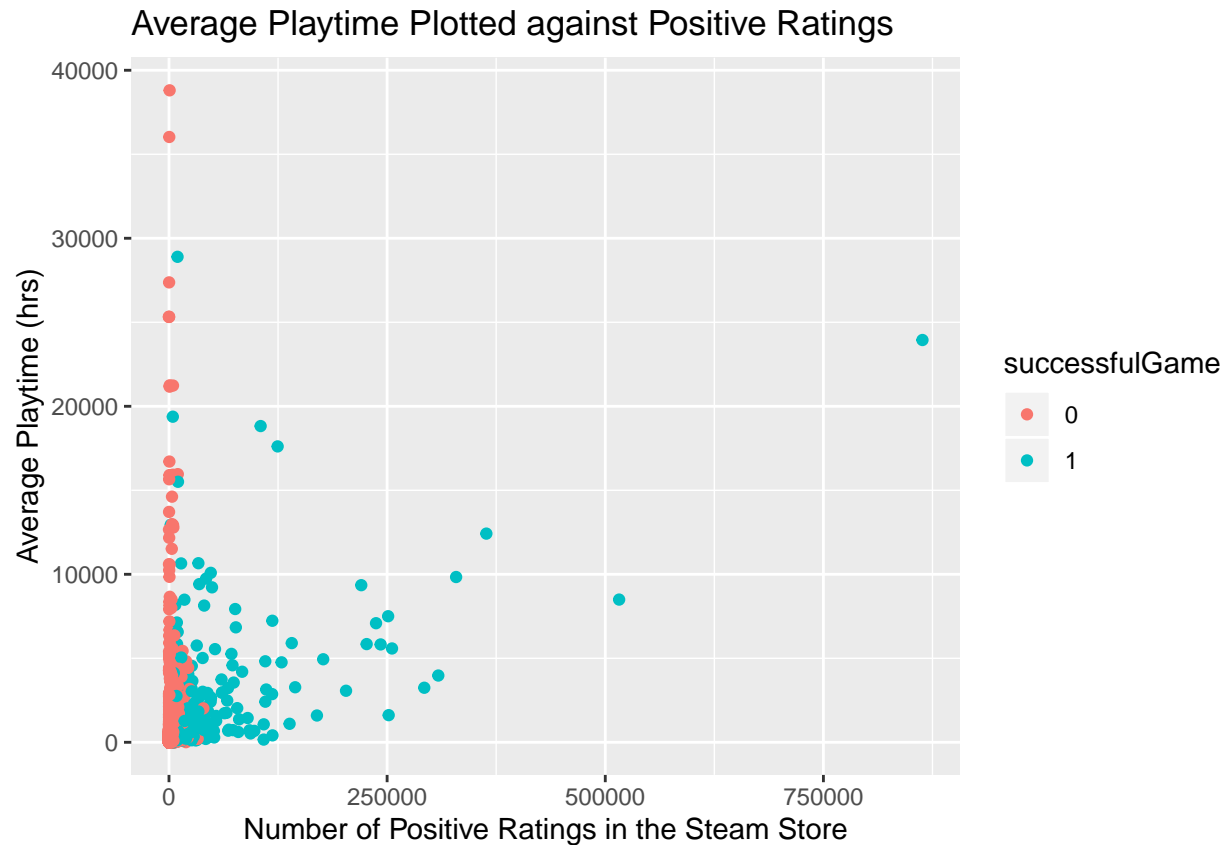
```
ggplot(steam,aes(x = positive_ratings,y = negative_ratings)) +
  geom_point(aes(color = successfulGame))+
  labs(x = "Number of Positive Ratings in the Steam Store",y = "Negative Ratings in the Steam Store",
    title = "Positive Ratings Plotted Against Negative Ratings")
```

Positive Ratings Plotted Against Negative Ratings



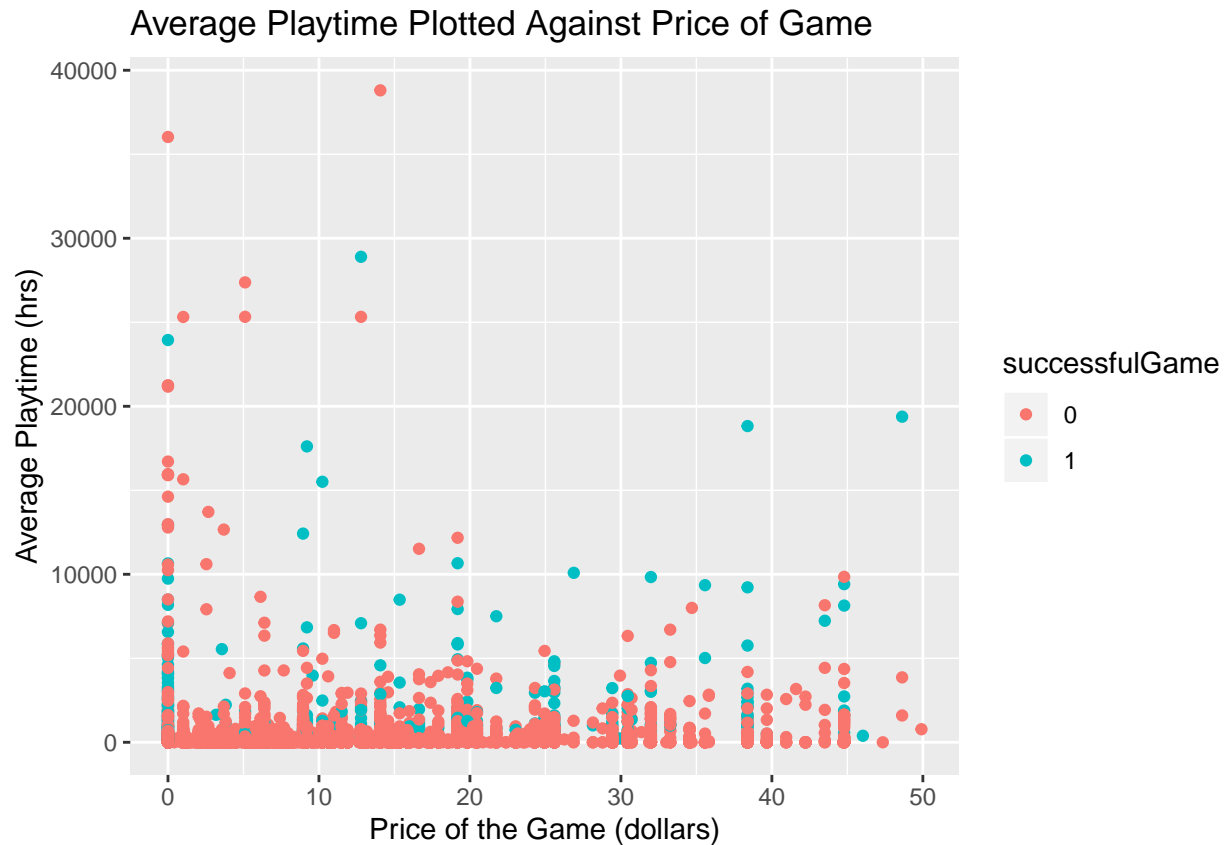
Here we observe that successful games have more of both types of ratings than unsuccessful games. The successful games gain significant comments and ratings on Steam whereas the lesser games do not apparently garner such attention.

```
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.

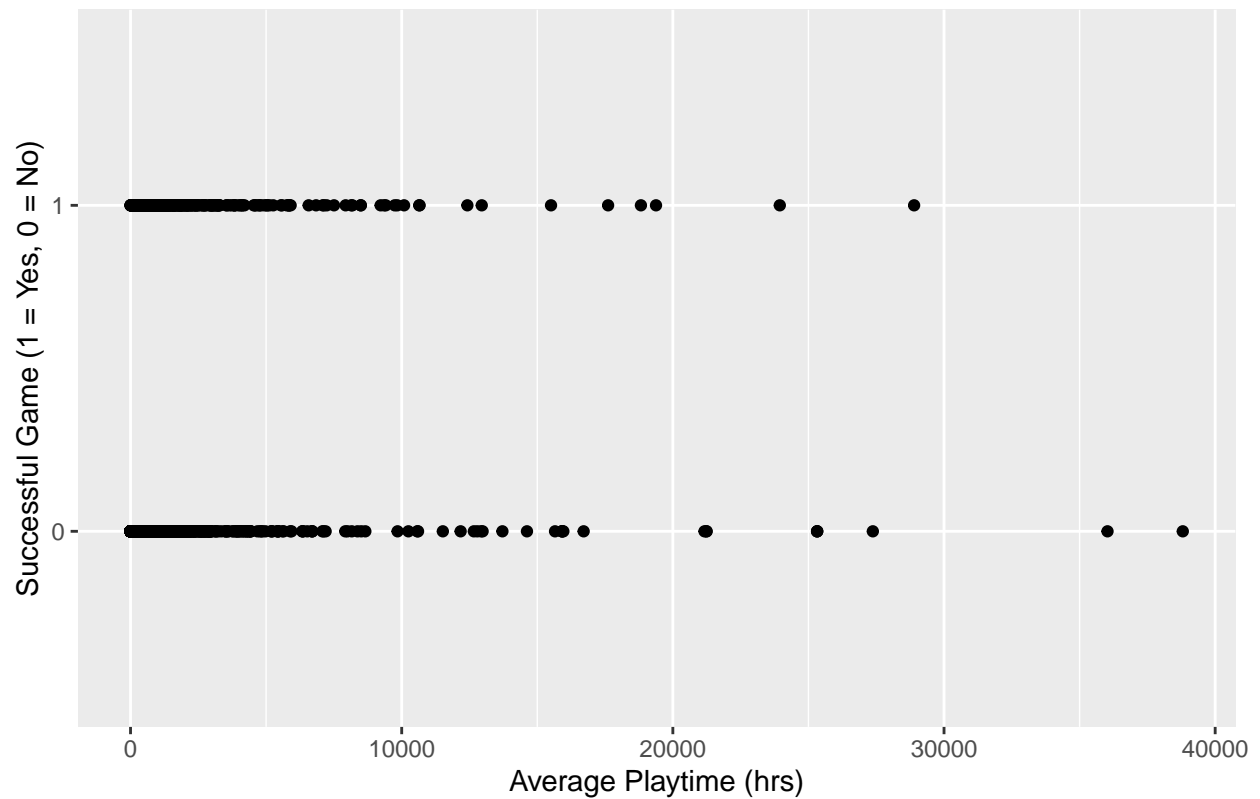
```
ggplot(steam,aes(x = price,y = average_playtime)) +
  geom_point(aes(color = successfulGame)) +
  labs(x = "Price of the Game (dollars)",y ="Average Playtime (hrs)",
       title = "Average Playtime Plotted Against Price of Game")
```



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.

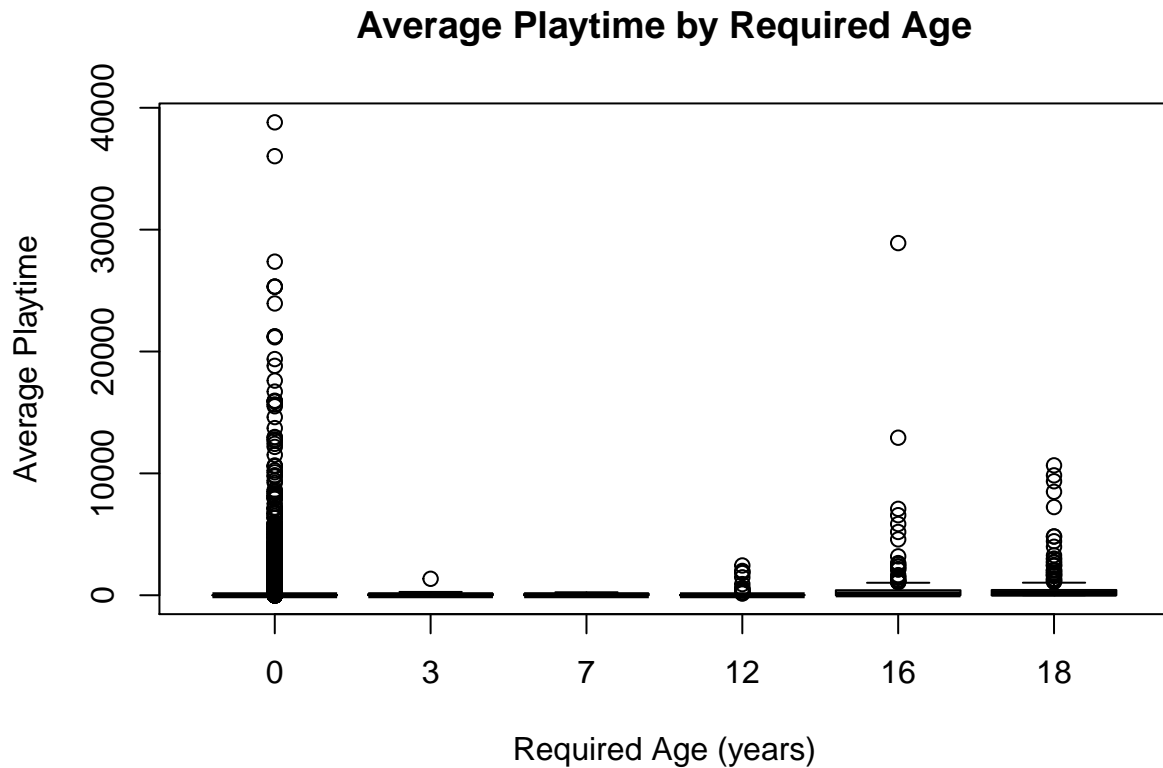
```
ggplot(steam,aes(x = average_playtime,y = successfulGame)) +
  geom_point()+
  labs(title = "Successful Game plotted Against Average Playtime",
        x = "Average Playtime (hrs)",y = "Successful Game (1 = Yes, 0 = No)")
```


Successful Game plotted Against Average Playtime



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.

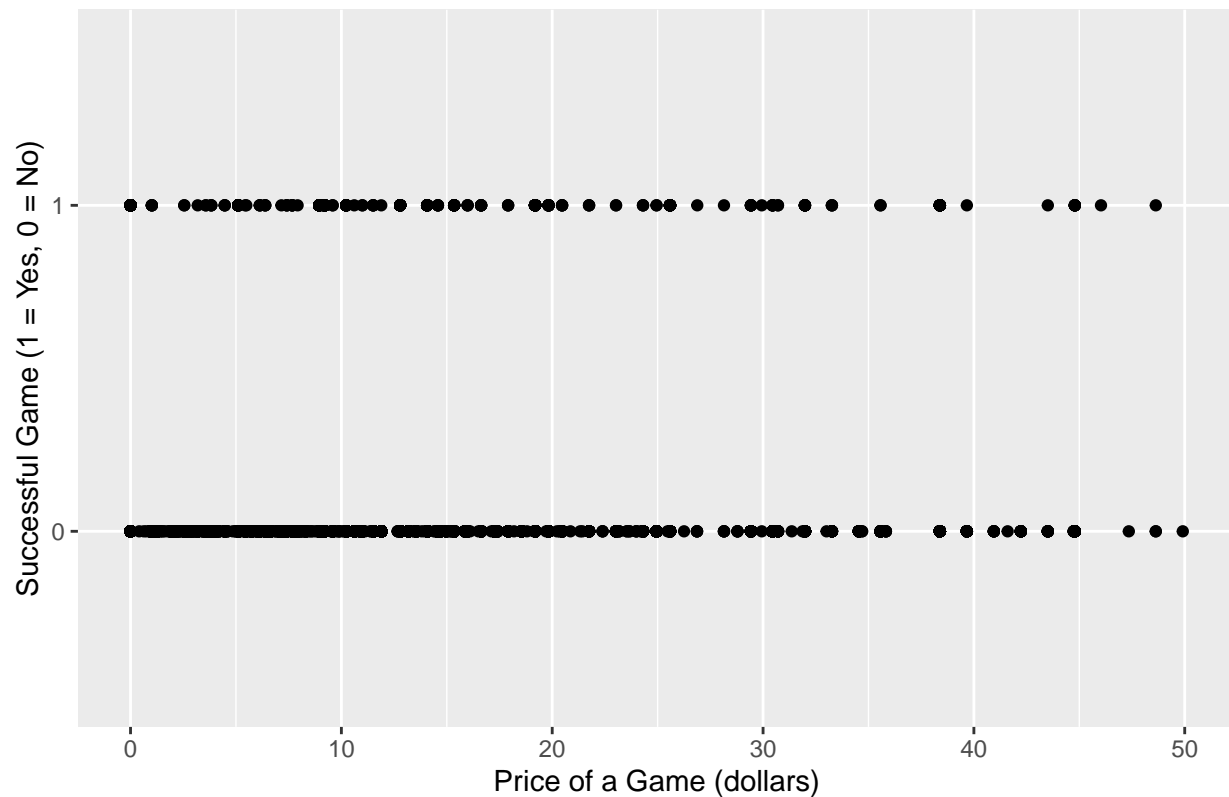
```
boxplot(steam$average_playtime~steam$required_age,
        main = "Average Playtime by Required Age",xlab = "Required Age (years)",
        ylab = "Average Playtime")
```



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)")
```

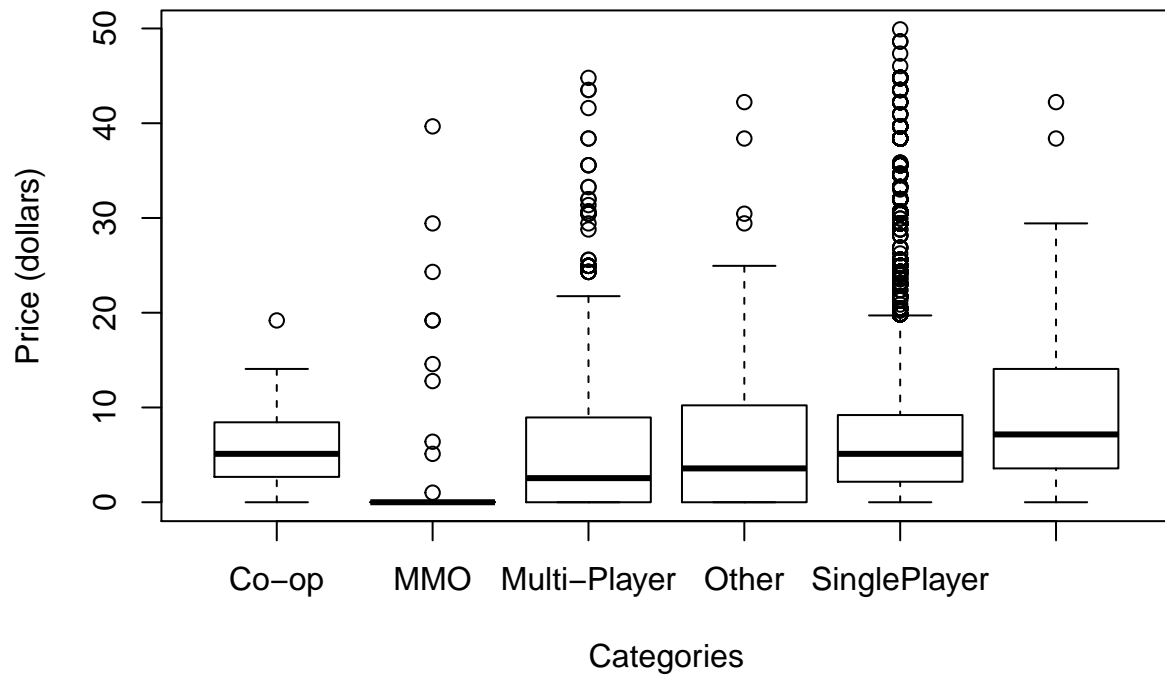
Successful Game plotted Against Average Playtime



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 while having the game at lower prices allows people to buy more of the game.

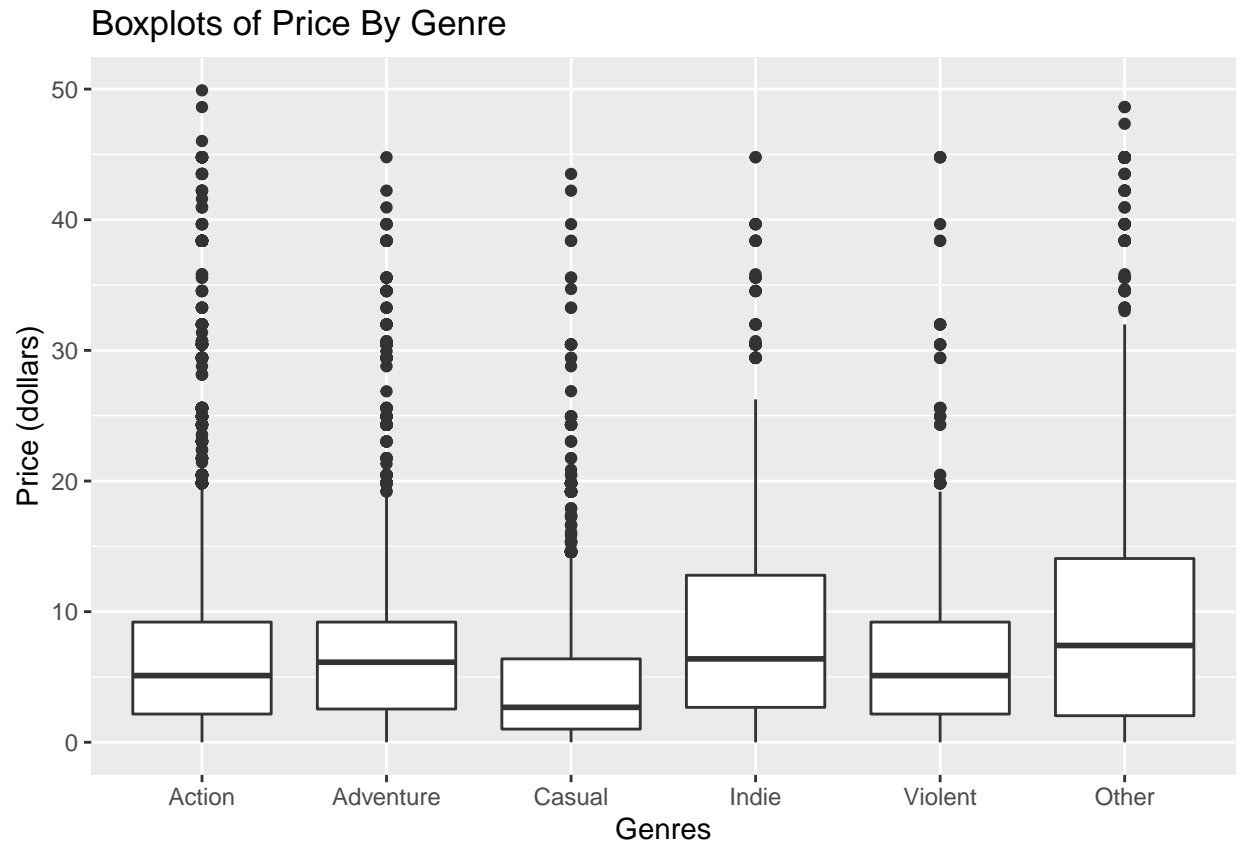
```
boxplot(steam$price~steam$categories,
        main = "Boxplot of Price By Category",
        xlab = "Categories",
        ylab = "Price (dollars)")
```

Boxplot of Price By Category



The above boxplot shows how the spread of price values for video games differ across the main categories of games. It appears that the spread of prices are generally the same, yet single player video games appear to have the most amount of outliers compared to other main categories.

```
ggplot(steam,aes(x = genres,y = price)) + geom_boxplot() +  
  labs(title = "Boxplots of Price By Genre",x = "Genres",y = "Price (dollars)")
```



From the above boxplot we can determine that the spread of prices are generally larger for indie games and “other” games than when compared with the remaining main genres. This is interesting, as perhaps the increased price of indie games and miscellaneous games may have an effect on the amount of copies bought, thus effecting its ability to sell over 1,000,000 copies. However, previous plots illustrating the relationship between price and successful game suggest there may not be a significant relationship between the two variables.