STAT 210 Project: Video Game Trends Report

Team 8: Abdullah Aldalaan, Abdulshaheed Alqunber, Hussain Almajed, Rozan Wali

---

Abstract

Video games are a large part of the entertainment industry. Our project analyzes how some less critical factors such as user scores, critics, genre, platforms affect global video game sales. We found out that genres such as Action and Sports are the most famous genres among users in terms of the number of games. Additionally, after creating a linear regression model to evaluate the critic scores and user scores against the sales, we found out that critic scores affect sales more than user scores do. On one hand, we compared the sales in terms of how many games are available in each genre, and we found out that the Action-Adventure and Party games affected the sales positively. On the other hand, some genres, such as Visual Novel and Educational games affected sales negatively. Furthermore, we examined the effects of platforms, and we saw that it has the highest effect on sales compared to other factors, primarily PlayStation, Xbox, and PC platforms. We can see how these statistical findings are informative in terms of growing trends in the industry.

Introduction

The video game industry is valued at 159 billion dollars [1]; with such a big market, companies need to understand consumers' behavior and patterns of use to deliver the optimal product. It is essential for video game companies to analyze key trends and understand what types of games are the most popular in order not to fall behind.

Many factors affect the gaming industry, most obvious ones include consumer purchase power, cultural trends and technological improvements of computers and consoles. In this project,

we are looking at the effects of some less important factors including genre, user and critic satisfaction on sales of video games, and the platform where the game is going to be played. By conducting analysis of past trends in the industry, we create a prediction model for how the "less obvious" factors affect video game sales which can help game developers create games that attract the public and accordingly generate high sales volumes.

The original dataset was outdated (published in 2014) and had few numbers of games. During the first analysis of the dataset, we found out that the dataset is not adequate for our project. Abdulshaheed, a member of our team, figured out where the original dataset was scraped from (VGChartz) and wrote a script to scrape an updated version of the dataset. The process was a bit complex since the server was blocking the script from collecting the data after a few numbers of requests, which is one way to defend against malicious acts. So, we had to go around it to build the new dataset. Finally, after about a week of modifying the code by adding multiprocessing to make the collection process faster and using proxies to forward the requests to prevent them from blocking us, we can scrape the whole website. We reckoned that it was necessary to add such measures as the process would have easily taken a month if it was done normally. i.e. without *multiprocessing or proxies.* The new dataset describes the video games sales up to the year 2019 [2]. This dataset is a continuation of Gregory Smith's efforts who collected about 16,700 records [3]. The updated dataset has 55,792 records of multiple variables, both qualitative and quantitative, almost triple the size of that of the original dataset. This range of information of video games includes Name, Genre, ESRB_Rating, Platform, Publisher, Developer, VGChartz_Score, Critic_Score, User_Score, Year of Release, Sales, Sales by region (North America, Japan, Europe, and Other), and Total_Shipped. Note that sales are in millions, and scores are out of 10. We are focusing on specific variables: Name, Genre, Platform, Critic_Score, User_Score, and other

variables to see how they affect gaming consumption *(appendix 1*). The new dataset Abdulshaheed created was published in Kaggle and to this day there are more than 7000 downloads and 40,000 views.

Statistical Methods

1. **Data Import & Cleaning**

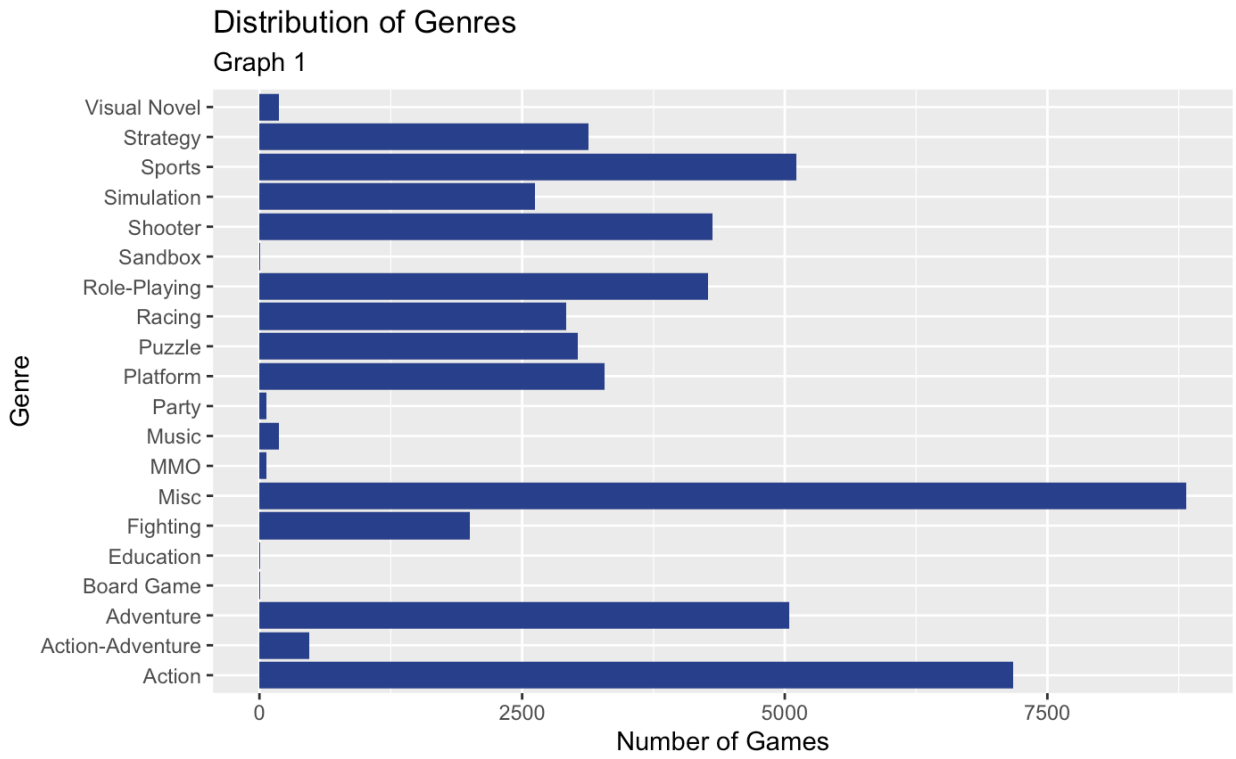
The data was easily imported, however, there were missing values in specific fields of the critic scores, and user scores. This is quite expected and reasonable since not all games in the dataset were critically acclaimed, especially the smaller ones. Also, not all the games had the year of the release, therefore, it was decided to eliminate games that don't have those specific values. Lastly, platforms operating less than 500 games were also eliminated from the dataset for uniformity.

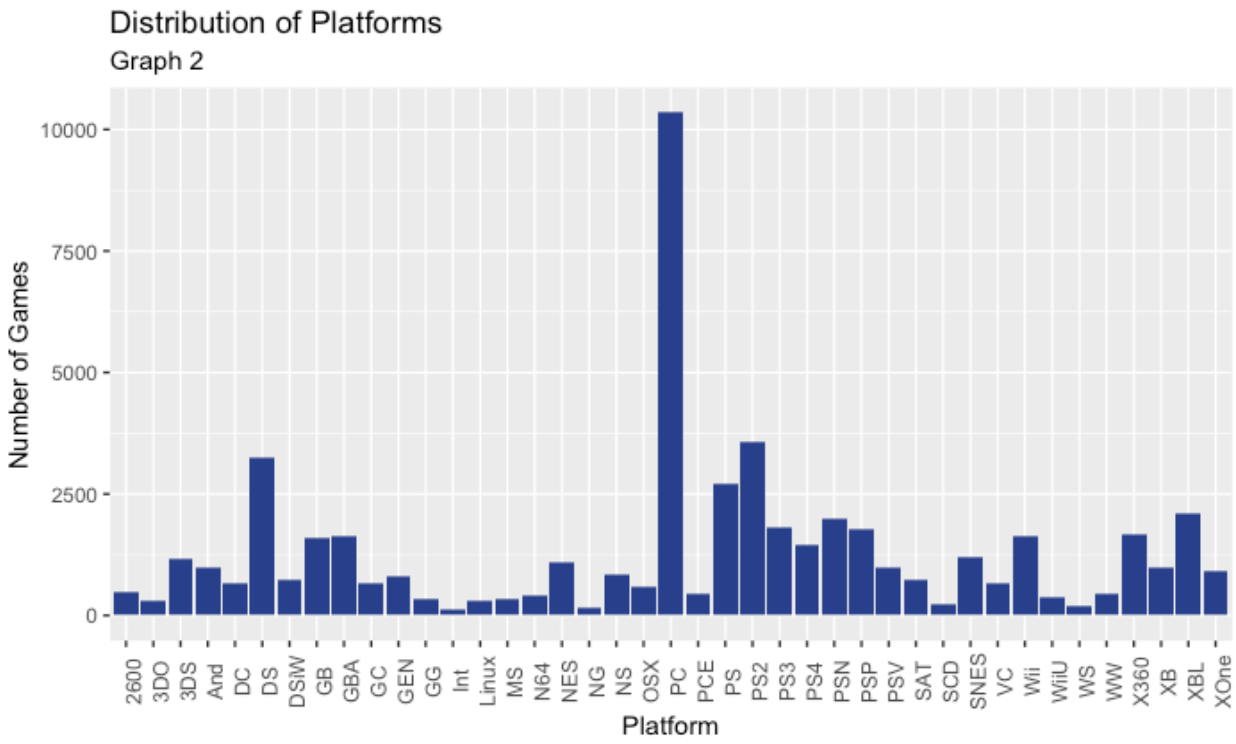
All of the rest of the data was kept for further analysis. More specifically, columns Critic_Score, and User_Score got a lot of null values in them, which will be filtered out for some of the analyses. However, other studies will require us to keep those values. Moreover, some observations were removed in order for the analysis to make sense. For example, observations in the year 2019 and 2020 were eliminated because the dataset was updated in early 2019, hence, it doesn't make sense to include games that were not released back then since they do not have enough information like user_scores and critic_score etc. Also, games older than 1980s were filtered out since they had missing values.

2. **Variation of Single Variables**

To quickly scan the data, we constructed a plot of counts for genres and platforms. This is to study the variation of single variables to understand their patterns of distribution. In graph 1,

genres were shown on the y-axis and the counts for each unique genre are on the x-axis. Based on the plot, we can see that Misc (miscellaneous), which is a compilation of genres with no specific category, has the largest counts, hence, we will not consider it. The most popular genre, therefore, is Action with almost 7000 counts, followed by genres of sports and adventures with roughly similar counts of around 5000. The graph also shows that the least popular genre is to be Education, Board game, and Sandbox with counts less than 10. Similarly, the same analysis was conducted to count for platforms' popularity. In graph 2, we can see that the most popular platform appears to be PC with more than 10,000 counts, and tremendously higher than all the rest of the platforms.



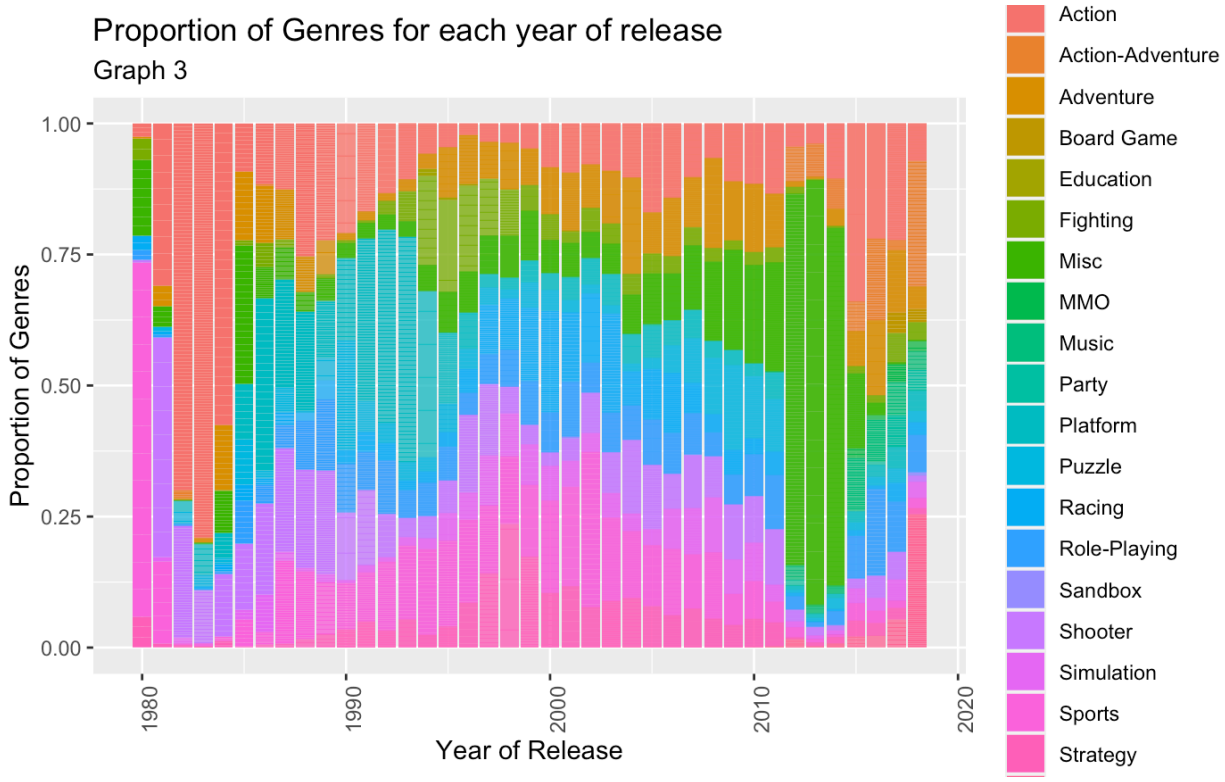


3. Variations of Multiple Variables

Next, we conducted correlations between two variables, Genre and Year of Release, to understand how they affect one other, as seen in Graph 3. The plot highlights genres' distributions, especially pointing out the rise of popularity of "Action" games. Given it was the most popular genre overall in the dataset, we looked at it individually, correlating it to years of release, as shown in Graph 4 below. From Graph 4, we observe a sharp increase in the release of Action games starting from late 1990s, with a peak in 2009. Action games' increased popularity is correlated with the introduction of PlayStation in 1994 which was the first console to popularize Action games such as Battle Arena Toshinden.

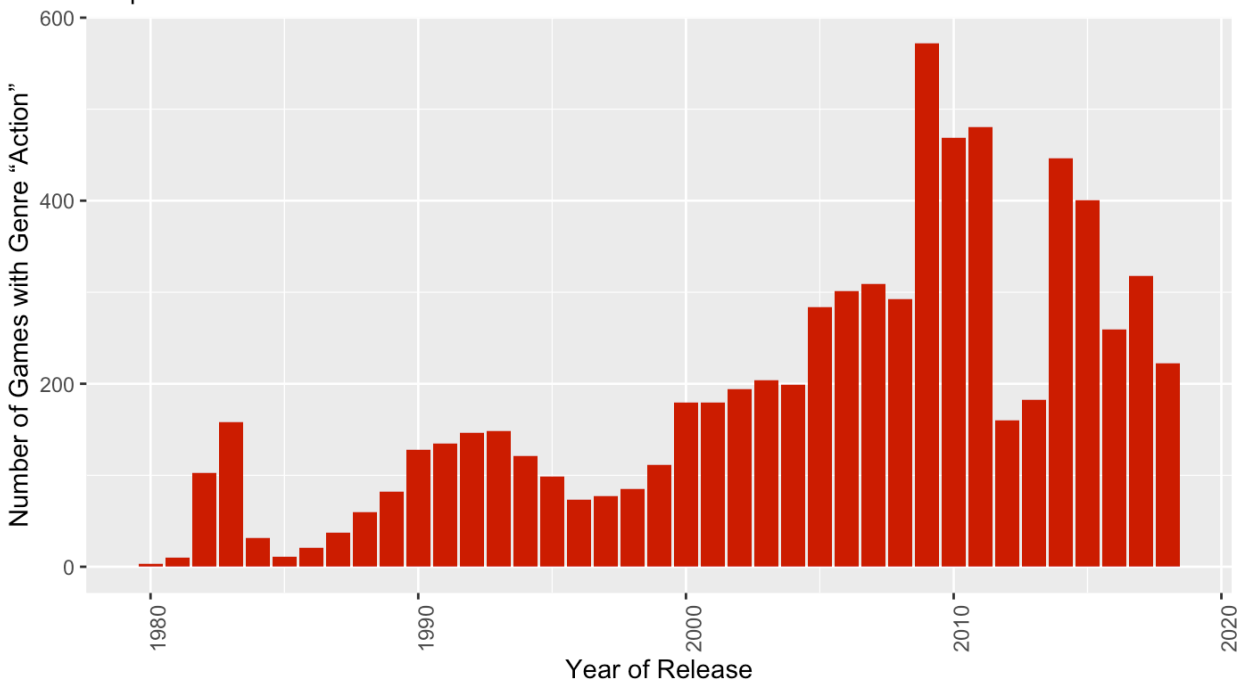
Proportion of Genres for each year of release

Graph 3



Change of action games popularity over years

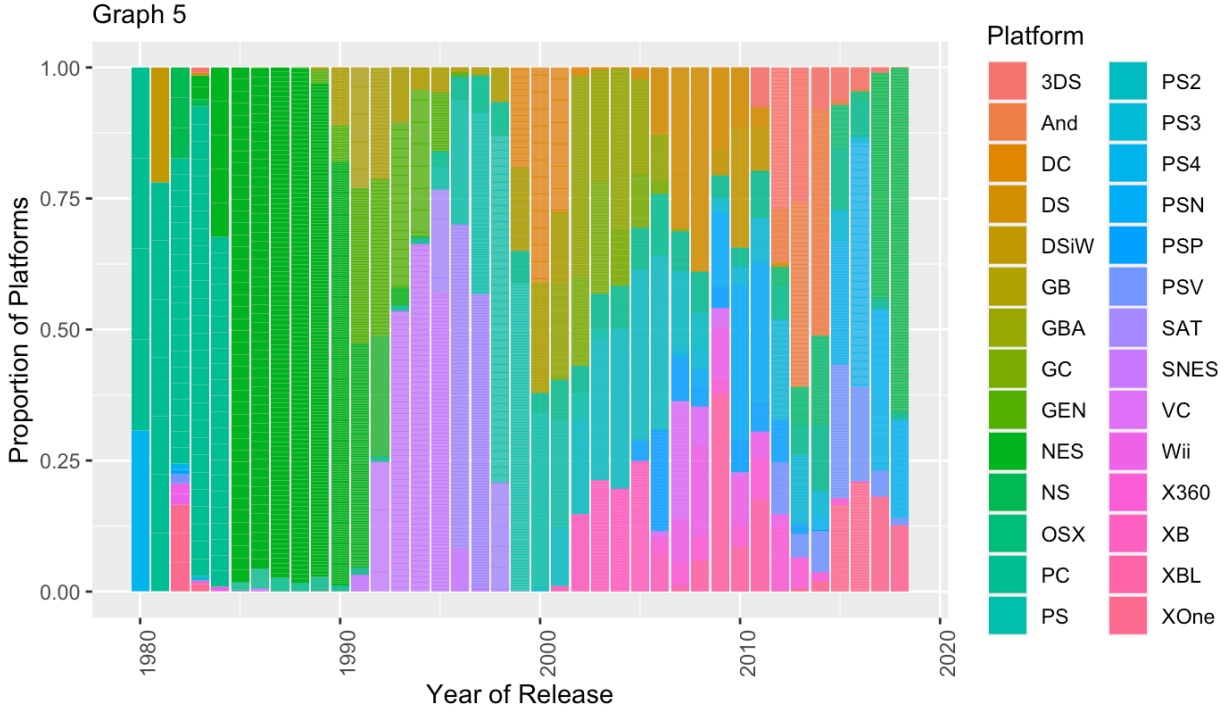
Graph 4



In addition, two analyses were conducted similar to the one in Graph 3, the proportion of platforms vs years and proportion of genre vs platforms, as seen in Graph 5 and 6. In Graph 5, we can see how new platforms were introduced and some became outdated over time, like some of the PS systems in light blue shade. This relates back to the sharp increase in PS popularity mentioned above since the color blue represents the appearance of PS in the market early 1995. This also proves that the correlation between the introduction of PS consoles and the rise in Action games popularity cannot be coincidental, but rather causation effect. In addition, Graph 6 correlates the proportion of genres for the 10 most popular platforms, to investigate if there is an existing relationship between these two variables and if a specific platform causes any particular genre to be more popular. From Graph 6, it should be noted that Actions games only increased in popularity with the later PS versions. The first PS consoles were famous for sports and racing games but shifted towards Actions games with the introduction of PS2 and then PS3. This analysis for popular genres and their trends throughout years in terms of sales can be a start point for a holistic research study for gaming trends for all the rest of the variables.

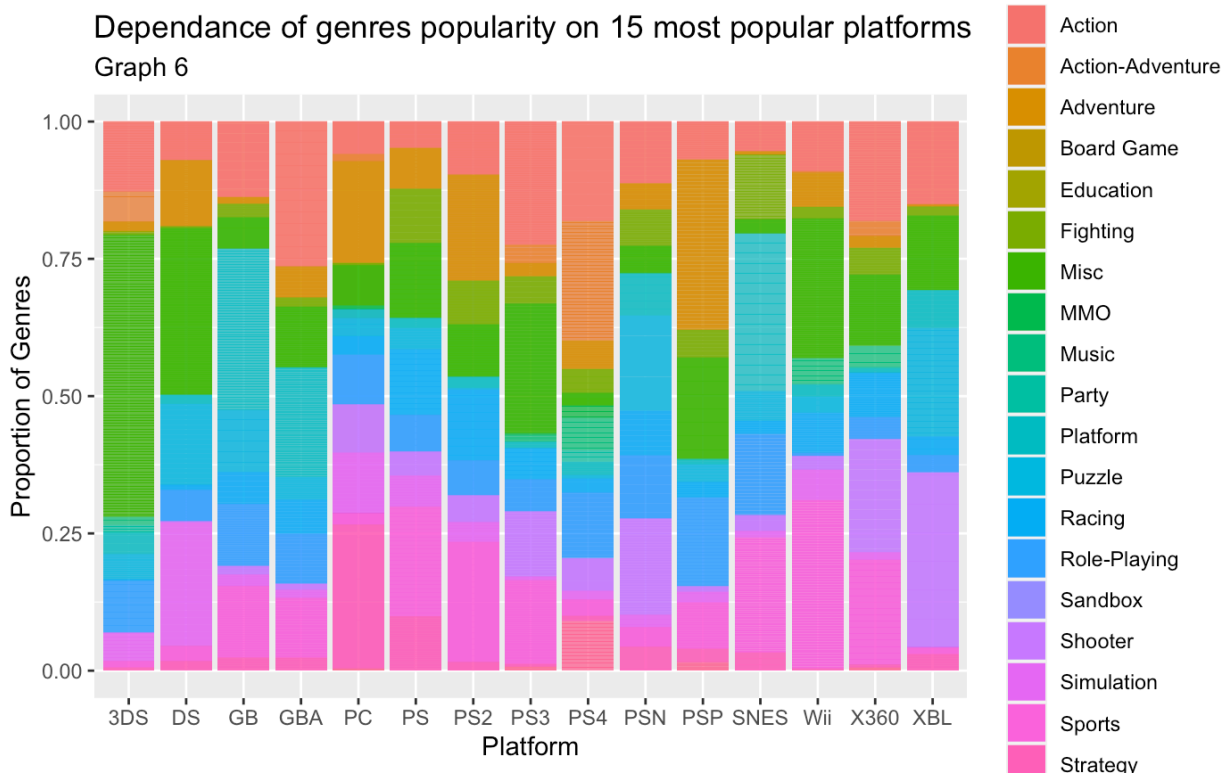
Proportion of platforms for games in each year

Graph 5



Dependance of genres popularity on 15 most popular platforms

Graph 6



4. Transformation of Variables

To achieve normality in our results, we used the natural logarithm on our dependent variable Sales, the reason is because Sales has an exponential relationship with the variables Genre, Critic Score, User Score, and Platform. This makes it possible to have smoother curves and more aesthetic ones. Using the natural log for our dependent variable thus just linearizes the curve of the relationship and results in a better suited model.

5. Check for Assumptions using Linear Regression

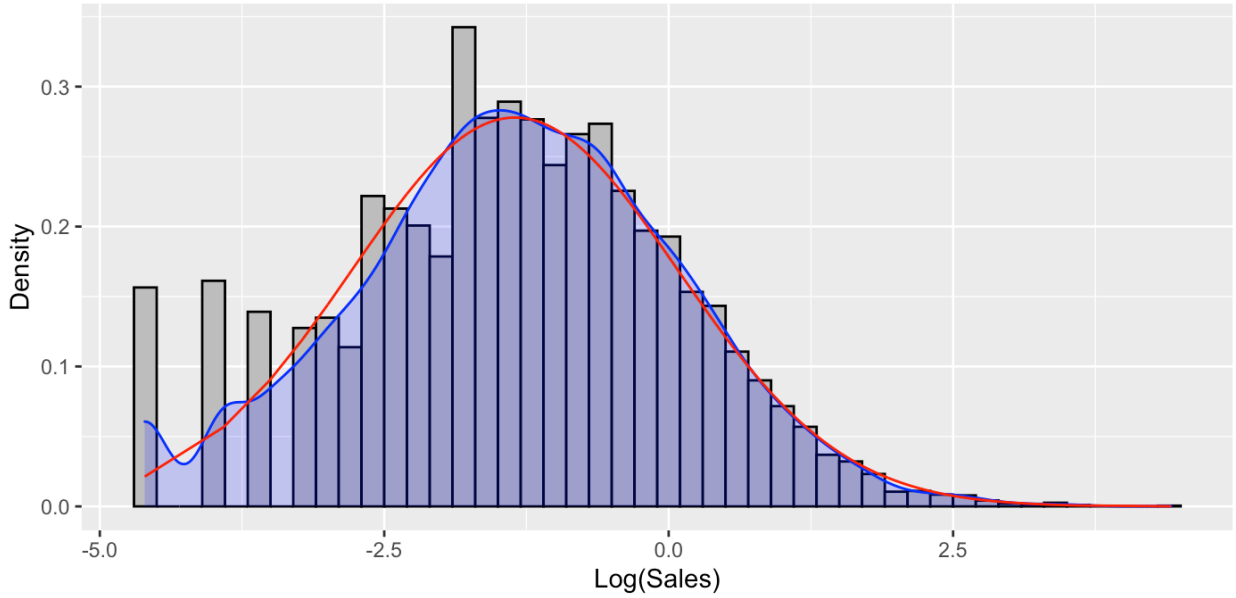
Since we are doing a linear regression model, it is important to verify its assumptions. The first step of our analysis was to check the distribution of the response variable. In this case, we are choosing "Sales" as our response variable as the goal of the project is to evaluate trends and how they affect global sales of video games. Video game Sales have a natural distribution, according to the curve below (graph 7). We can reasonably conclude that the majority of the data points are relatively close, although there are some outliers. The blue curve represents the real distribution of Sales, and the theoretical normal distribution is shown by the red curve. As most games are sold in equal amounts, this makes sense, with some of them being highly successful and others not at all.

The first thing we did after plotting the linear regression model was to verify the assumptions. In Appendix [3], 4 diagnostic charts can be observed. First, to check for the linearity assumption, we check the residuals vs. fitted graph: residuals form a horizontal line around the 0 line. Meaning there is insignificance difference between the observed and predicted values satisfying the linearity condition. Second, to determine if the data comes from a normal distribution, we check the normal QQ plot: since the points fall along the straight

line, we assume normality. Third, to check equal variance assumption we look at the scale location plot checks: the line is horizontal coupled with almost equal spread of data points meeting our homoscedastic assumption. This was also proved by the ncv test results as based on the P-value we got, indicating that it is not statistically significant. Lastly the fitting Residuals vs Leverage graph spots influential cases: two outliers are present, but they don't affect the regression analysis since we have a lot of data and we examined the effect of removing them, but nothing has changed that much, therefore, there's no need to remove them.

Sales Distribution

Graph 7



This graph showcases how close is the actual data relative to the theoretical distribution.
The blue line and its fill showcases the actual Sales' distribution.
The red line is the theoretical normal distribution of Sales.

Analysis and Results

1. Linear Regression Results

We used the critic score, user score, genres, and platforms as independent variables to build the regression model. Then we plotted the model to see how these variables affect sales. The

significance of each variable's influence was given by the coefficients and corresponding p-values we got from fitting the model as seen from *appendix 4*. Since there are more than 30 platforms, we simplified them by grouping them according to the device's manufacturer to avoid having a complex model. To illustrate, instead of having Playstation 1, Playstation 2, Playstation 3, Playstation Network, and Playstation 4, we grouped them under "PlaystationConsole." Same thing was applied to other platforms; at the end, 7 categories of platforms remained.

The R-squared value obtained is 0.1386, meaning that only 13.86% of the variability of the observed variables is explained by the model if it's correct. As we can see from the results in *appendix 4*, there are multiple variables which are statistically significant based on their p-values such as Critic_Score, Adventure genre, Shooter genre, PC platform, PlaystationConsole platform and Xbox platform. These factors are the most influential on affecting the sales. From the cofficents, it seems that Platforms play a major role in affecting the sales, especially Playstation. Some genres also play a high part in affecting the sales positively and negatively, such as Action-Adventure and Puzzles respectively. Critic scores are also an important factor in affecting the sales since it got a very low p-value and a good coefficient.

We did some experiments to improve our model by including more independent variables such as publisher. This increased the R-squared value up to 43%, however, the model became more complicated and difficult to analyze since there are so many publishers, more than 300. Each one of them will get its own estimation, so we decided not to include this variable.

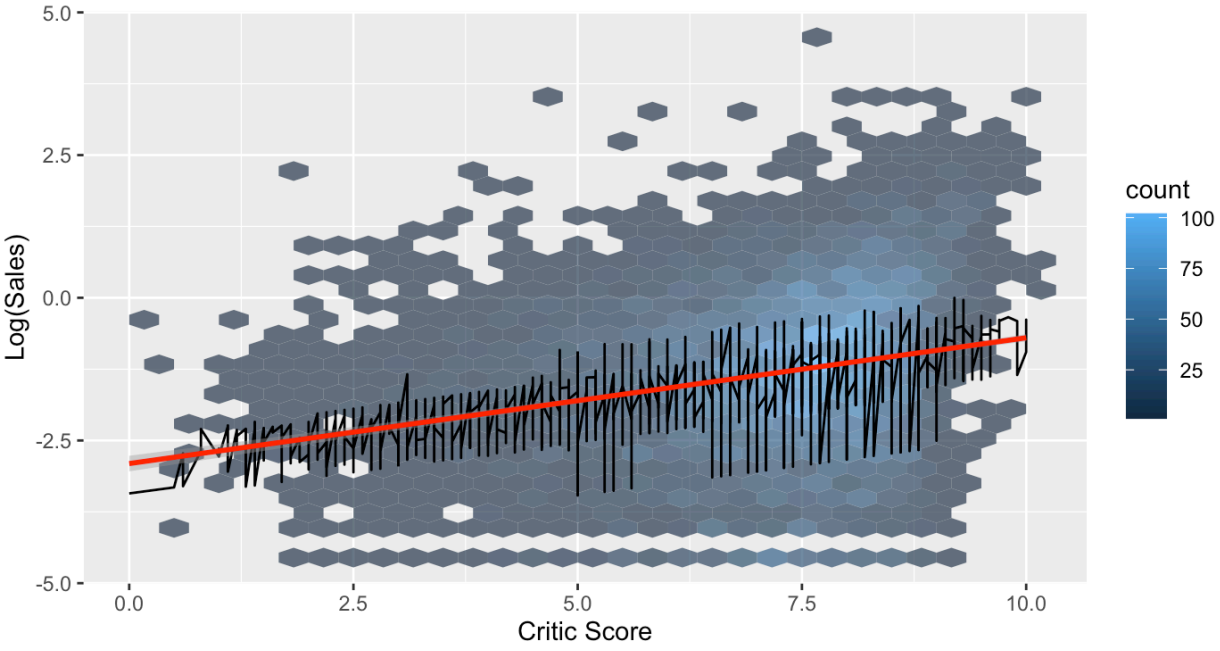
2. Prediction of the Response

We plotted prediction graphs for both Critic and User Scores to forecast global sales using each of the two independent variables after determining to what degree these independent variables impacted global sales.

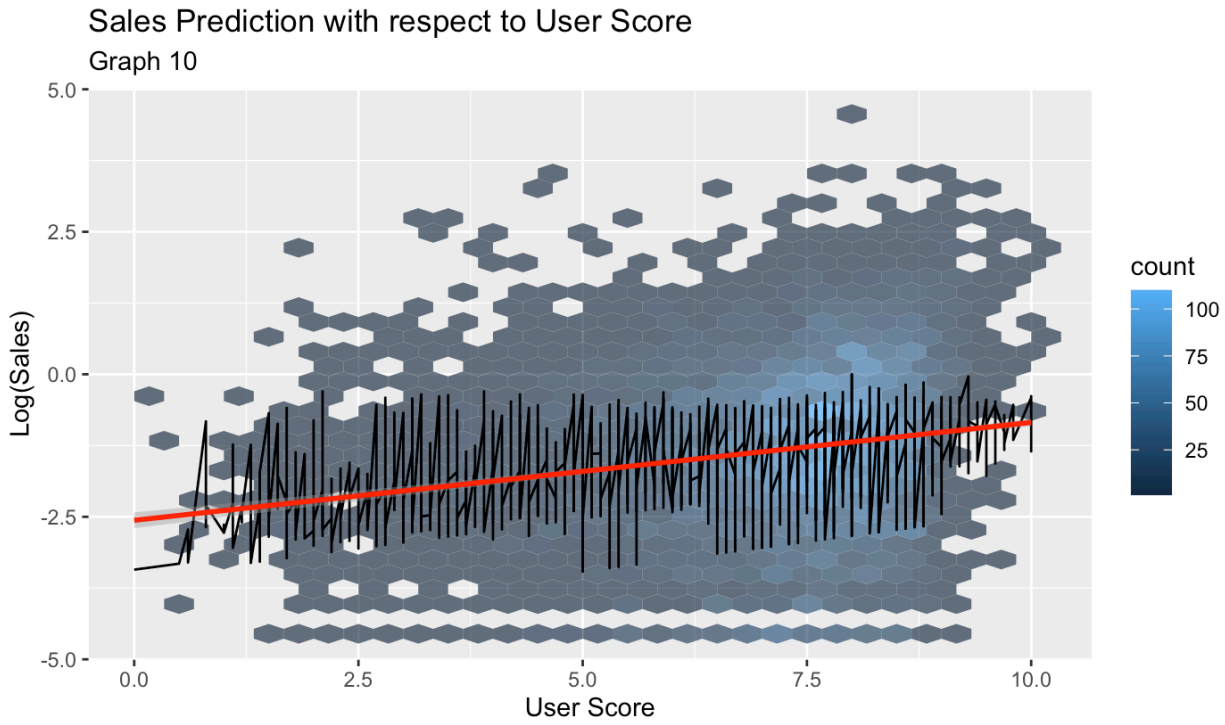
Graphs 9 and 10 illustrate the prediction model of the effects on sales of the critic score and user score, respectively. As the black line is more widespread, there is much more noise in the critic score plot. The trend red line also has a bigger slope. As we have previously known that critic scores have a greater effect on sales than user scores, these findings make sense. The curve of Graph 9 also has more residuals for higher critic scores, and the residuals for Graph 10 are much more staggering than the previous one, especially for lower scores. Ultimately, this implies that critics are more forgiving than users when rating games.

Sales Prediction with respect to Critic Score

Graph 9



retical linear model in red shows the trend of Sales growth related to critic score. The black line shows the actual data.

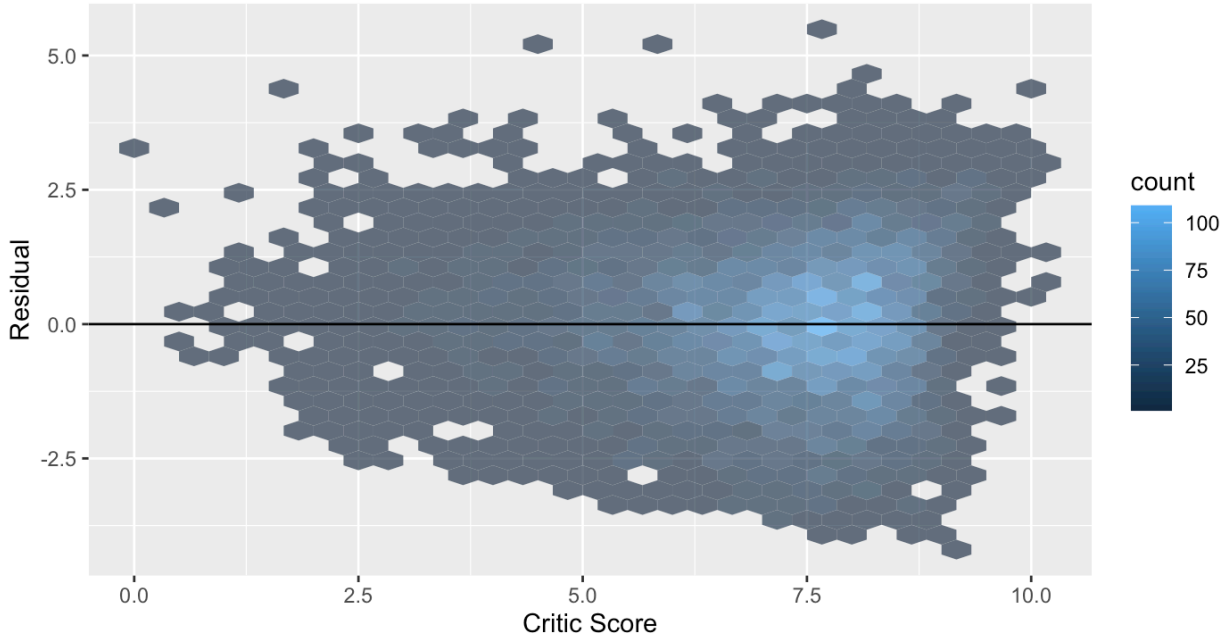


retical linear model in red shows the trend of Sales growth related to critic score. The black line shows the actual data.

Graphs 11 and 12 display the residuals, respectively, with regard to critic scores and user scores. Although they do not include quite many details, they do illustrate how much of the noise is around critics and user scores of around 7.5. This suggests that the majority of games from different genres are scored as such.

Residual with respect to Critic Score

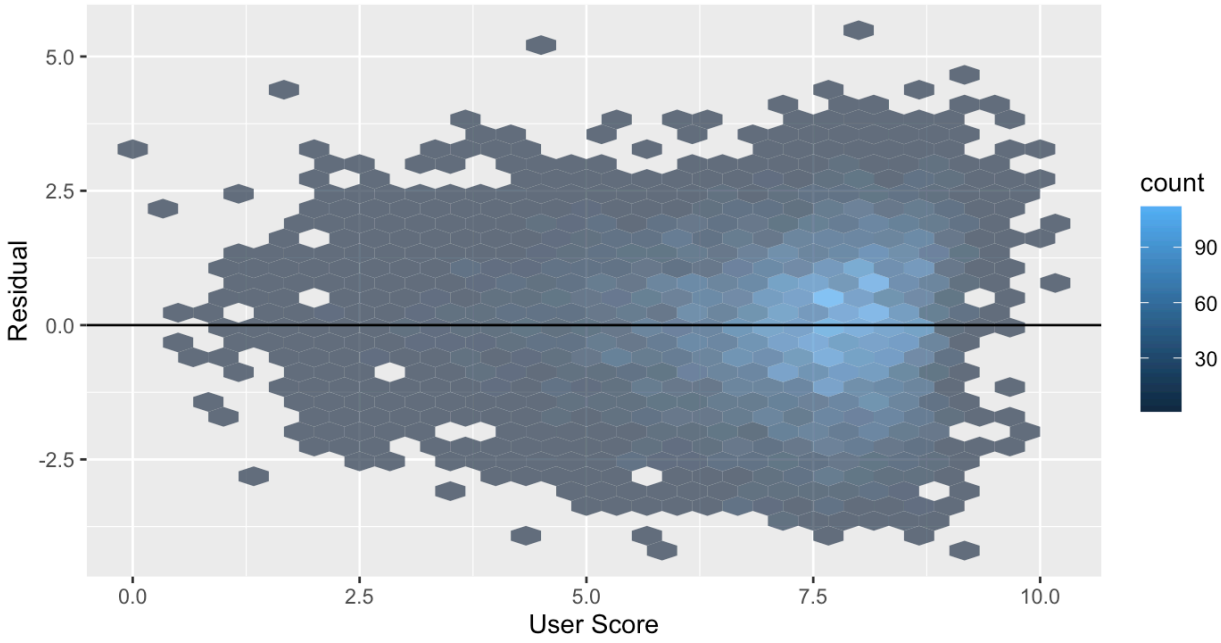
Graph 11



The black line portrays the residual error of the model.

Residual with respect to User Score

Graph 12



The black line portrays the residual error of the model.

Discussion

The goal of this research report was to identify the extent of "lesser important" factors on impacting video games sales. The analysis was conducted using linear regression models and hypothesis tests. Based on the results that we got as we can see from appendix [4], there are multiple variables that have a high effect on sales. For instance, Critic_Score, some types of genres such as Adventures, Action-Adventure, Puzzle, Shooter, Party, and Visual Novel, and multiple platforms such as PlaystationConsole, Xbox, and PC.

Starting with Platforms, we can see that they have the highest effect on sales from the coefficients we got. Based on the P-values, we see that PlaystationConsole, Xbox, and PC all have very low P-values indicating that they are all statistically significant. The coefficient for PlaystationConsole is 2.81, which is the highest among all of them. Then comes next Xbox, with a coefficient equals 2.73, and finally PC with a coefficient equals 2.62. There are other platforms that also have high effect, but we will focus on the ones we mentioned since they are the most statistically significant in our case.

Game genres also had several significant effects on sales. Our Analysis indicated that the category: Adventure, Puzzle, and Visual Novel are the most statistically significant genres affecting sales. In specific, they have a negative impact on sales meaning games in those genres would have lower sales revenue. On the contrary, the most popular genres: Action-Adventure, and Shooter, have positive effects on the sales.

Critic_Score does also have an effect on sales based on its coefficient and P-value. It is considered statistically significant since it is very small. The coefficient is equal to 0.23, which means it has a positive effect on sales.

Also, from the prediction model, we can see that critic score data is much cleaner as the critics tend to use specific parameters and metrics. From these results we can infer that consumers trust critic scores more than user scores when it comes to buying video games.

One limitation to our study is the limited sources of data regarding the video games industry. Since the data comes from the real-world opposed to the ones we used in the class, the data has a lot of noise and many missing values. Accordingly, the critic scores and the user scores data were collected from a different source, *Metacritics* [4], for the missing values as opposed to the rest of the data which were collected from *Vgchartz*. Using multiple sources, we had challenges with data refining and cleaning to make sure the homogeneity of the data set as discussed above. Furthermore, we tried to improve the model by adding more independent variables. However, the model became more complex and harder to analyze. We even tried to reduce complexity by grouping some of the data, like platforms, but that would be hard to do for other independent variables like publishers.

For future work, many things could be done to improve the model. We could add more parameters and try to simplify them to better understand the model and make more accurate analysis. For instance, as discussed before, we could add another independent variable, which is Publisher, that will increase the R-square value of the model. However, we will need to find a way to simplify it so we could analyze the model easily. Another thing that could be done is to find more resources to gather more data in this area to fill the missing values, so we could make better predictions.

Based on what we discussed, it seems that the most factors that are affecting the sales are some specific platforms, genres, and critic scores, with platforms having the highest effect.

Companies should take into consideration the platform they will publish their game on, and the type of game they are working on because that will be one of many factors that will help them in getting better sales. It seems like platforms like PlayStation, Xbox, and PC are the top platforms that most players are using. At the same time, genres like Action-Adventure, and Shooters are what most players are playing.

References:

[1] F. (2020, May 12). Report: Gaming revenue to top $159B in 2020. Retrieved October 29, 2020, from https://www.reuters.com/article/esports-business-gaming-revenues-idUSFLM8jkJMl

[2] Alqunber, A. (2019, April 13). Video Games Sales 2019. Retrieved October 29, 2020, from https://www.kaggle.com/ashaheedq/video-games-sales-2019

[3] Smith, G. (2016, October 26). Video Game Sales. Retrieved October 29, 2020, from https://www.kaggle.com/gregorut/videogamesales

[4] Movie Reviews, TV Reviews, Game Reviews, and Music Reviews. (n.d.). Retrieved December 06, 2020, from https://www.metacritic.com/

Appendix 1:

Table 1: Study variables and their meaning

Variable	Meaning
Name	Name of game
Genre	Classification of type of game
Platform	Gaming system used (i.e. PC, PS4, XOne, etc.)
Critic_Score	Critic score from metacritic
User_Score	User score from metacritic
Year of Release	Year the game was released
Sales	Number of games sold in millions
Total_Shipped	Number of games sold by Nintendo

Appendix 2:

[1] Data cleaning code

```
```{r test & clean}

library(tidyverse)
data <- read.csv("vgsales-12-4-2019.csv")
problems(data)

data <- data %>%
  filter(Genre != '', Year != 'N/A', Year != '2020', Year !='2019' )
data <- data %>%  group_by(Platform) %>% filter(n() > 100)
```
```

[2] Per genre distribution code graph 1

```
```{r Variation Single 1}
data %>% group_by(Genre, Name) %>% ggplot()
+ geom_bar(aes(Genre), fill = 'royalblue4')
+ coord_flip()+ labs(title = 'Distribution of Genres',
x = 'Genre', y = 'Number of Games', subtitle = 'Graph 1')

```
```

[3] Per platform distribution code graph 2

```
```{r Variation Single 2}
data %>% group_by(Platform, Name) %>% ggplot()
+ geom_bar(aes(Platform), fill = 'royalblue4')
+ labs(title = 'Distribution of Platforms', x = 'Platform',
      y = 'Number of Games', subtitle = 'Graph 2')
+ theme(axis.text.x = element_text(angle = 90, size=9))

```
```

[4] graph 3

```
data1 <- data %>% group_by(Genre) %>% mutate(sum_each = n()) %>% ungroup() %>% group_by(Year, Genre) %>%
mutate(Proportion_of_Genre = n()/sum_each)

data1 %>% ggplot(aes(Year, Proportion_of_Genre, fill = Genre)) + geom_bar(stat="identity", position = 'fill') + labs(title =
'Proportion of Genres for each year of release', subtitle = 'Graph 3', x = 'Year of Release', y = 'Proportion of Genres') +
theme(axis.text.x = element_text(angle = 90, size=9))
```

[5] graph 4

```
#Genre 'Action' over time
data1 %>% filter(Genre == "Action") %>%
  ggplot() + geom_bar(aes(Year), fill = 'red3') + labs(title = 'Change of action games popularity over years', x = 'Year of
Release', y = 'Number of Games with Genre "Action"', subtitle = 'Graph 4') + theme(axis.text.x = element_text(angle = 90,
size=9))
```

[6] graph 5

```
data2 <- data %>% group_by(Platform) %>% filter(n() > 500) %>% mutate(sum_each1 = n()) %>% ungroup() %>% group_by(Platform,
Year) %>% mutate(Proportion_of_Plat = n()/sum_each1)

data2 %>% ggplot(aes(Year, Proportion_of_Plat, fill = Platform)) + geom_bar(stat="identity", position = 'fill') + labs(title =
'Proportion of platforms for games in each year', subtitle = 'Graph 5', x = 'Year of Release', y = 'Proportion of Platforms') +
theme(axis.text.x = element_text(angle = 90, size=9))
```

[7] graph 6

```
data3 <- data %>% group_by(Genre) %>% mutate(sum_each2 = n()) %>% group_by(Genre, Platform) %>%  mutate(prop2 = n()/sum_each2)

data3 %>%  group_by(Platform) %>% filter(n() > 1100) %>% ggplot(aes(Platform,prop2, fill=Genre)) + geom_bar(stat = "identity",
position = 'fill') + labs(title = 'Dependance of genres popularity on 15 most popular platforms', subtitle = 'Graph 6', x =
'Platform', y = 'Proportion of Genres')
```

[8] graph 7

```
library(tidyverse)
data <- read_csv("vgsales_metacritic.csv", col_types =  cols(Critic_Score = col_double(), User_Score = col_double()))
data_filtered <- data %>% filter((Global_Sales != 0 | Total_Shipped != 0) &!is.na(User_Score) & !is.na(Critic_Score)) %>%
mutate(Sales = Global_Sales + Total_Shipped, Sales_log = log(Global_Sales + Total_Shipped), Log_CS = log(Critic_Score), Log_US
= log(User_Score))
problems(data_filtered)

data_filtered %>% mutate(density_th = dnorm(log(Sales), mean = mean(log(Sales)), sd = sd(log(Sales)))) %>%
  ggplot() +
  geom_histogram(aes(x = log(Sales), y = ..density..), fill = "gray", color = "black", binwidth = 0.2) +
  geom_density(aes(x = log(Sales)), colour = "blue", fill = "blue", alpha = 0.2) +
  geom_line(aes(x = log(Sales), y = density_th), colour = "red") +
  labs(x = "Log(Sales)",
       y =  "Density",
       title = "Sales Distribution",
       subtitle = "Graph 7",
       caption = "This graph showcases how close is the actual data relative to the theoretical distribution.
       The blue line and its fill showcases the actual Sales' distribution.
       The red line is the theoretical normal distribution of Sales.")
```

[9] graph 9

```
library(modelr)
library(hexbin)
mod <- lm(log(Sales) ~ Critic_Score + User_Score + Genre, data = data_filtered)
ggplot(data_filtered %>% add_predictions(mod), aes(x = Critic_Score, y = log(Sales))) + geom_hex(alpha = 0.7) + geom_line(aes(y
= pred)) + geom_smooth(method = "lm", color = "red") + labs(
  x = "Critic Score",
  y =  "Log(Sales)",
  title = "Sales Prediction with respect to Critic Score",
  subtitle = "Graph 9",
  caption = "The theoretical linear model  in red shows the trend of Sales growth related to critic score. The black line shows
the actual data.")
```

[11] graph 10

```
ggplot(data_filtered %>% add_predictions(mod), aes(x = User_Score, y = log(Sales))) + geom_hex(alpha = 0.7) + geom_line(aes(y =
pred))  + geom_smooth(method = "lm", color = "red") + labs(
  x = "User Score",
  y =  "Log(Sales)",
  title = "Sales Prediction with respect to User Score",
  subtitle = "Graph 10",
  caption = "The theoretical linear model  in red shows the trend of Sales growth related to critic score. The black line shows
the actual data.")
```

[12] graph 11

```
ggplot(data_filtered %>% add_residuals(mod), aes(x = Critic_Score, y = resid)) + geom_hex(alpha = 0.7) + geom_hline(yintercept
= 0) + labs(
  x = "Critic Score",
  y =  "Residual",
  title = "Residual with respect to Critic Score",
  subtitle = "Graph 11",
  caption = "The black line portrays the residual error of the model.")
```

[13] graph 12

```
ggplot(data_filtered %>% add_residuals(mod), aes(x = User_Score, y = resid)) + geom_hex(alpha = 0.7) + geom_hline(yintercept =
0) + labs(
  x = "User Score",
  y =  "Residual",
  title = "Residual with respect to User Score",
  subtitle = "Graph 12",
  caption = "The black line portrays the residual error of the model.")
```

[13] Grouping the platforms

```
data_filtered$Platform[data_filtered$Platform == "Wii"] <- "NintendoConsole"
data_filtered$Platform[data_filtered$Platform == "GBA" ] <- "NintendoPortable"
data_filtered$Platform[data_filtered$Platform == "N64"] <- "NintendoConsole"
data_filtered$Platform[data_filtered$Platform == "NS"] <- "NintendoConsole"
data_filtered$Platform[data_filtered$Platform == "DS"] <- "NintendoConsole"
data_filtered$Platform[data_filtered$Platform == "WiiU"] <- "NintendoConsole"
data_filtered$Platform[data_filtered$Platform == "3DS"] <- "NintendoPortable"
data_filtered$Platform[data_filtered$Platform == "GC"] <- "NintendoConsole"
data_filtered$Platform[data_filtered$Platform == "GB"] <- "NintendoPortable"
data_filtered$Platform[data_filtered$Platform == "SNES"] <- "NintendoConsole"
data_filtered$Platform[data_filtered$Platform == "NES"] <- "NintendoConsole"
data_filtered$Platform[data_filtered$Platform == "VC"] <- "NintendoConsole"

data_filtered$Platform[data_filtered$Platform == "PS"] <- "PlaystationConsole"
data_filtered$Platform[data_filtered$Platform == "PS2"] <- "PlaystationConsole"
data_filtered$Platform[data_filtered$Platform == "PS3"] <- "PlaystationConsole"
data_filtered$Platform[data_filtered$Platform == "PS4"] <- "PlaystationConsole"
data_filtered$Platform[data_filtered$Platform == "PSN"] <- "PlaystationConsole"
data_filtered$Platform[data_filtered$Platform == "PSV"] <- "PlaystationPortable"
data_filtered$Platform[data_filtered$Platform == "PSP"] <- "PlaystationPortable"

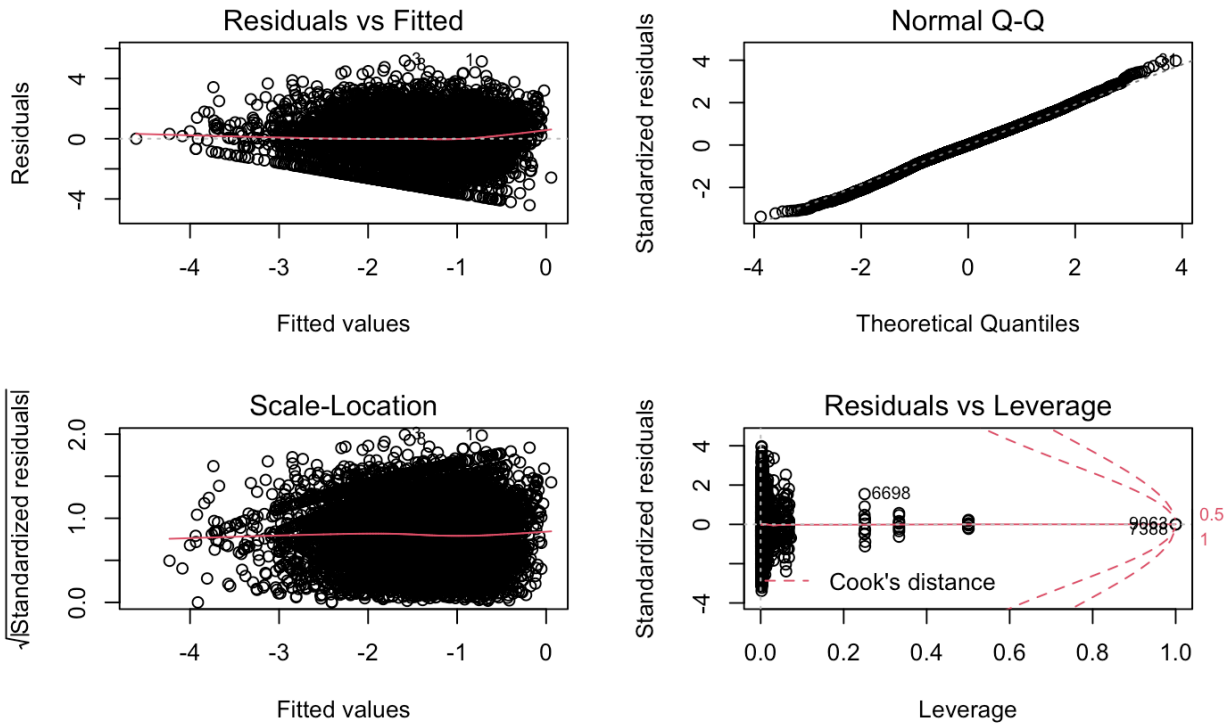

data_filtered$Platform[data_filtered$Platform == "X360"] <- "Xbox"
data_filtered$Platform[data_filtered$Platform == "XOne"] <- "Xbox"
data_filtered$Platform[data_filtered$Platform == "XB"] <- "Xbox"
data_filtered$Platform[data_filtered$Platform == "XBL"] <- "Xbox"

data_filtered$Platform[data_filtered$Platform == "OSX"] <- "PC"


data_filtered$Platform[data_filtered$Platform == "2600"] <- "Others"
data_filtered$Platform[data_filtered$Platform == "3DO"] <- "Others"
data_filtered$Platform[data_filtered$Platform == "DC"] <- "Others"
data_filtered$Platform[data_filtered$Platform == "3DO"] <- "Others"
data_filtered$Platform[data_filtered$Platform == "SAT"] <- "SEGA"
data_filtered$Platform[data_filtered$Platform == "SCD"] <- "SEGA"
data_filtered$Platform[data_filtered$Platform == "GEN"] <- "SEGA"

data_filtered$Platform[data_filtered$Platform == "WS"] <- "Others"
data_filtered$Platform[data_filtered$Platform == "NG"] <- "Others"
```

Appendix 3:




```
not plotting observations with leverage one:
  8648, 8702, 8775, 9317Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.1337326, Df = 1, p = 0.71459
```

```{r}

par(mfrow = c(2, 2))
plot(mod)
ncvTest(mod)
```

Appendix 4:

```
lm(formula = log(Sales) ~ Critic_Score + User_Score + Genre + 
    Platform, data = data_filtered)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1232 -0.8487  0.0240  0.8971  5.5884 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -5.26527    0.77559  -6.789 1.20e-11 ***
Critic_Score                0.23096    0.01770  13.051  < 2e-16 ***
User_Score                 -0.02581    0.01772  -1.457 0.145174    
GenreAction-Adventure       0.21463    0.09657   2.223 0.026271 *  
GenreAdventure             -0.63378    0.06009 -10.547  < 2e-16 ***
GenreEducation             -1.11372    1.33475  -0.834 0.404077    
GenreFighting               0.02984    0.07083   0.421 0.673508    
GenreMisc                  -0.37698    0.05765  -6.539 6.49e-11 ***
GenreMMO                    0.62262    0.31680   1.965 0.049400 *  
GenreMusic                 -0.08912    0.14611  -0.610 0.541934    
GenreParty                  0.83205    0.32520   2.559 0.010525 *  
GenrePlatform               0.17554    0.06574   2.670 0.007595 ** 
GenrePuzzle                -0.75402    0.09063  -8.320  < 2e-16 ***
GenreRacing                -0.13371    0.06098  -2.193 0.028350 *  
GenreRole-Playing          -0.07204    0.05590  -1.289 0.197518    
GenreSandbox                0.12479    0.77127   0.162 0.871464    
GenreShooter                0.15824    0.05523   2.865 0.004178 ** 
GenreSimulation            -0.26424    0.06917  -3.820 0.000134 ***
GenreSports                 0.08028    0.05073   1.583 0.113558    
GenreStrategy              -0.56600    0.08223  -6.883 6.23e-12 ***
GenreVisual Novel          -1.56967    0.22870  -6.863 7.15e-12 ***
PlatformNintendoConsole     2.44182    0.77133   3.166 0.001552 ** 
PlatformNintendoPortable    2.38340    0.77269   3.085 0.002045 ** 
PlatformOthers              1.54528    0.81785   1.889 0.058864 .  
PlatformPC                  2.62465    0.77196   3.400 0.000677 ***
PlatformPlaystationConsole  2.81499    0.77131   3.650 0.000264 ***
PlatformPlaystationPortable 1.95595    0.77230   2.533 0.011337 *  
PlatformSEGA                1.62501    0.80043   2.030 0.042367 *  
PlatformXbox                2.73770    0.77162   3.548 0.000390 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.333 on 9433 degrees of freedom
Multiple R-squared:  0.1386,    Adjusted R-squared:  0.1361 
F-statistic: 54.22 on 28 and 9433 DF,  p-value: < 2.2e-16
```

```{r}
coef(mod <- lm(log(Sales) ~ Critic_Score + User_Score + Genre + Platform, data = data_filtered))
summary(mod)


```