Video Game Trends Project: Second Deliverable

Group1: Abdulshaheed Alqunber, Rami Bassil, Ekaterina Gorbunova, Khalid Khumayis

Abstract

One of the main branches of the entertainment industry is video games. The industry has been around for over 30 years and has seen extraordinary growth as it is worth $78.61 billion in 2017. It was decided to investigate statistics related to this growth and look for correlations and trends. The proposed research is based on a dataset of 9,490 (after cleaning) video games and some associated data referring to international sales and scores among others. The purpose of this study is to find a pattern and correlation between ratings of the games, popularity, platforms, genres, and developers. R Studio will be used to analyze the data.

The dataset contains both qualitative and quantitative data. Important variables are Name, Genre, ESRB_Rating, Platform, Publisher, Developer, VGChartz_Score, Critic_Score, User_Score, Year of Release, Sales (number of games sold), Sales by region (NA, Japan, PAL, Other). However, not all variables will be used in the research since the focus will be on game genres and game scores and how they affect gaming consumption. Ultimately, the objective is to find a meaningful correlation between some variables such as genre and score, and Sales to make predictions about future trends in the industry.

Research Question and Motivation

A multitude of factors affects global sales of video games. Most obvious ones include the improvement of chips used in computers and consoles, GDP per capita, marketing, online influencers, and cultural trends. A lot of other factors also have an effect on sales but to a lesser extent. In this project, we analyze the extent to which those “other” factors (Genre, Critic Score, User Score) affect video games sales. This analysis will help establish whether publishers should take those factors into account when creating and marketing new games.

Updates from Deliverable 1

In the first deliverable, we stated that a correlation between game genre and sales would be established. We found out that some genres (such as Action) were more popular than others and that their proportions varied over time. In this deliverable, we hope to establish to what extent do those genres, as well as Critic and User Scores, affect global sales. In order to illustrate these relationships, linear regressions are conducted between Sales and other variables.

Besides our main objectives, the dataset itself was also updated. As mentioned in the first deliverable, there was a lot of data missing data points (especially in the years following 2009) that ultimately resulted in faulty analysis. We searched for an updated version of the same dataset that at least includes the last year or two, but we could not find any. Therefore, we were able to backtrace the source code that used to scrape the data from VGChartz website to build an updated version of the dataset. Because the original dataset is about five years old, the code needed to be modified to make it work with the new structure of HTML document file and CSS style of the website. Eventually, we were able to build a new dataset that includes almost 56,000 games — a multiple of 3.5 in size of the original dataset.

Cleaning the Data

The first step was to make sure that our data was clean and did not have a lot of missing values. However, some games have NAs for certain variables (such as Critic_score) but actual values for other variables (such as Sales). We thus had to carefully estimate whether removing them altogether would affect our results. So, we compared linear fit results before and after removing those games and found out that results did not change significantly and the data still made sense statistically. We, therefore, chose to remove all the games with missing values.

Modeling:

Distribution of Response Variable

The first step of our analysis is 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.

```r
library(tidyverse)
```

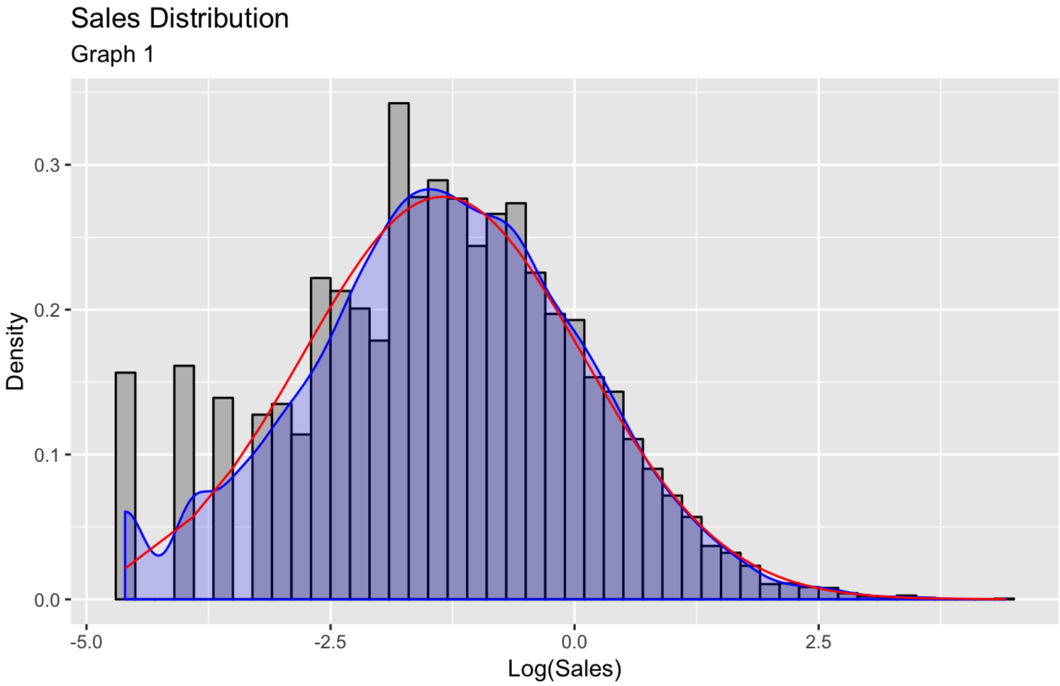
```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.2.1 ──
```

```
## ✔ ggplot2 3.1.0       ✔ purrr   0.3.0
## ✔ tibble  2.1.1       ✔ dplyr   0.8.0.1
## ✔ tidyr   0.8.2       ✔ stringr 1.3.1
## ✔ readr   1.3.1       ✔ forcats 0.3.0
## ── Conflicts ──────────────────────────────────────────────────────────────── tidyverse_confl
icts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

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

```
## [1] row      col      expected actual
## <0 rows> (or 0-length row.names)
```

```
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 1",
       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.")
```



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.

According to the above curve, video game sales data can be seen to have a normal distribution. While there are some outliers, we can safely assume that the majority of the data points are relatively similar. The blue curve portrays actual sales distribution, and the red curve shows the theoretical normal distribution. This makes sense, as most games are sold in similar numbers, with some of them being extremely popular and others not at all.

Transformation of Variables

In order to achieve normality and clearer results, we chose to use the natural logarithm of our dependent variable “Sales”. The reason for this choice is mainly because video game sales (dependent variable) grew exponentially over the past two decades, which might be due to the rise of technological trends. As for the independent variables, they do not grow at such a rate since they are not closely related to Sales. Therefore, using the natural log for our dependent variable only linearizes the relationship curve and results in a better fitted model.

Model Parameter Estimates and their Interpretations

When the regression model was plotted, we used critic score, user score and genres as independent variables to see how they influence sales. The coefficients and corresponding p-values that we got from fitting the model told us about the significance of each variable’s effect. Consequently, we were able to state that critic score, user score, and some particular genres had a significant influence on the dependent variable. However, even though each of those arguments is important in the model, the overall goodness of fit is only 10.5%, as our R^2 value shows us. Therefore, our parameters explain only 10% of the variability in the data, which means that those independent variables do not have a great impact on global sales, as expected. That being said, it does show a clear correlation between genre and score, and sales, something that game publishers might not have thought about.

```
coef(mod <- lm(log(Sales) ~ Critic_Score + User_Score + Genre, data = data_filtered))
```

```
##           (Intercept)          Critic_Score            User_Score 
##           -2.69662006            0.25772027           -0.04760354 
## GenreAction-Adventure        GenreAdventure        GenreEducation 
##            0.21219916           -0.73042247           -1.37764771 
##         GenreFighting             GenreMisc              GenreMMO 
##            0.02117136           -0.46826130            0.70457272 
##            GenreMusic            GenreParty         GenrePlatform 
##           -0.10565214            0.75043436            0.12813749 
##           GenrePuzzle           GenreRacing     GenreRole-Playing 
##           -0.88596590           -0.11019555           -0.15261686 
##          GenreSandbox          GenreShooter       GenreSimulation 
##            0.14798986            0.22905121           -0.29440051 
##           GenreSports         GenreStrategy     GenreVisual Novel 
##            0.10092681           -0.58089587           -1.81731530
```

```
summary(mod)
```

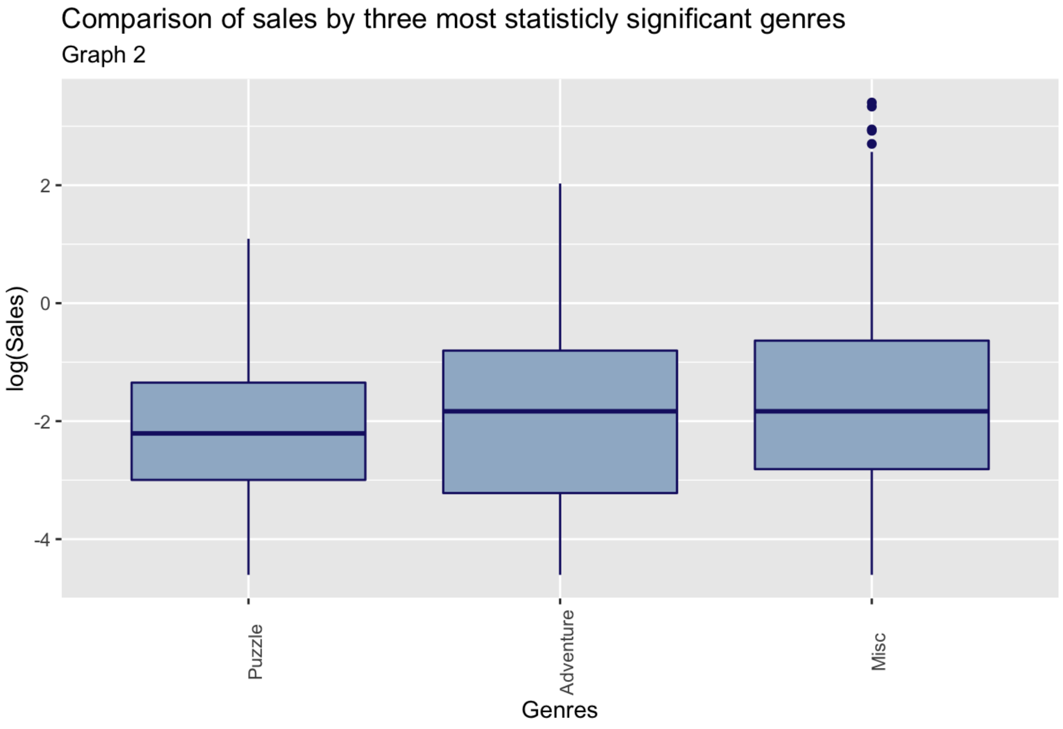
```
## 
## Call:
## lm(formula = log(Sales) ~ Critic_Score + User_Score + Genre, 
##     data = data_filtered)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.1762 -0.8647  0.0359  0.9204  5.4092 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           -2.69662    0.07622 -35.378  < 2e-16 ***
## Critic_Score           0.25772    0.01789  14.408  < 2e-16 ***
## User_Score            -0.04760    0.01795  -2.652  0.00802 ** 
## GenreAction-Adventure  0.21220    0.09809   2.163  0.03054 *  
## GenreAdventure        -0.73042    0.06094 -11.986  < 2e-16 ***
## GenreEducation        -1.37765    1.36033  -1.013  0.31121    
## GenreFighting          0.02117    0.07197   0.294  0.76863    
## GenreMisc             -0.46826    0.05822  -8.043 9.79e-16 ***
## GenreMMO               0.70457    0.32230   2.186  0.02883 *  
## GenreMusic            -0.10565    0.14896  -0.709  0.47819    
## GenreParty             0.75043    0.33151   2.264  0.02362 *  
## GenrePlatform          0.12814    0.06653   1.926  0.05412 .  
## GenrePuzzle           -0.88597    0.09137  -9.697  < 2e-16 ***
## GenreRacing           -0.11020    0.06204  -1.776  0.07571 .  
## GenreRole-Playing     -0.15262    0.05657  -2.698  0.00699 ** 
## GenreSandbox           0.14799    0.78564   0.188  0.85059    
## GenreShooter           0.22905    0.05600   4.090 4.35e-05 ***
## GenreSimulation       -0.29440    0.06985  -4.215 2.52e-05 ***
## GenreSports            0.10093    0.05152   1.959  0.05013 .  
## GenreStrategy         -0.58090    0.08200  -7.084 1.50e-12 ***
## GenreVisual Novel     -1.81732    0.23241  -7.819 5.88e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.359 on 9469 degrees of freedom
## Multiple R-squared:  0.1052, Adjusted R-squared:  0.1033 
## F-statistic: 55.66 on 20 and 9469 DF,  p-value: < 2.2e-16
```

Hypothesis Tests for Significant Main Effects

After fitting a linear regression model and getting p-values for each variable, we selected the genres that were the most statistically significant and compared them in terms of their sales means. We made a plot and also conducted Tukey's Test which allowed us to make pairwise comparisons of means.

```
data_filtered2 <- data_filtered %>% filter(Genre == "Adventure" | Genre == 'Misc' | Genre == 'Puzzle')

data_filtered2 %>% group_by(Genre) %>% ggplot() + geom_boxplot(aes(reorder(Genre, log(Sales), FUN = median),
log(Sales)), fill = "slategray3", color = 'midnightblue') + labs(title = "Comparison of sales by three most
statisticly significant genres", x = "Genres", y = "log(Sales)", subtitle = "Graph 2", caption = 'This plot
shows difference of sales amount in terms of three most significant genres in our linear regression model')
+ theme(axis.text.x = element_text(angle = 90))
```



This plot shows difference of sales amount in terms of three most significant genres in our linear regression model

In the plot we observe the three most statistically significant genres in terms of our regression model, Puzzle, Adventure, and Miscellaneous. The box plots represent the natural logarithm of sales. As we can see from the black horizontal lines, the medians for those genres are approximately the same. However, it is hard to tell if there is a statistically significant difference between them, so we conduct a Tukey comparison to obtain p-values and evaluate whether the combination of genres has a greater effect.

```
plant.lm <- lm(log(Sales) ~ Genre, data = data_filtered2)
plant.av <- aov(plant.lm)
summary(plant.av)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## Genre           2     36  17.770   7.526 0.000556 ***
## Residuals    1824   4307   2.361                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
tukey.test <- TukeyHSD(plant.av)
tukey.test
```

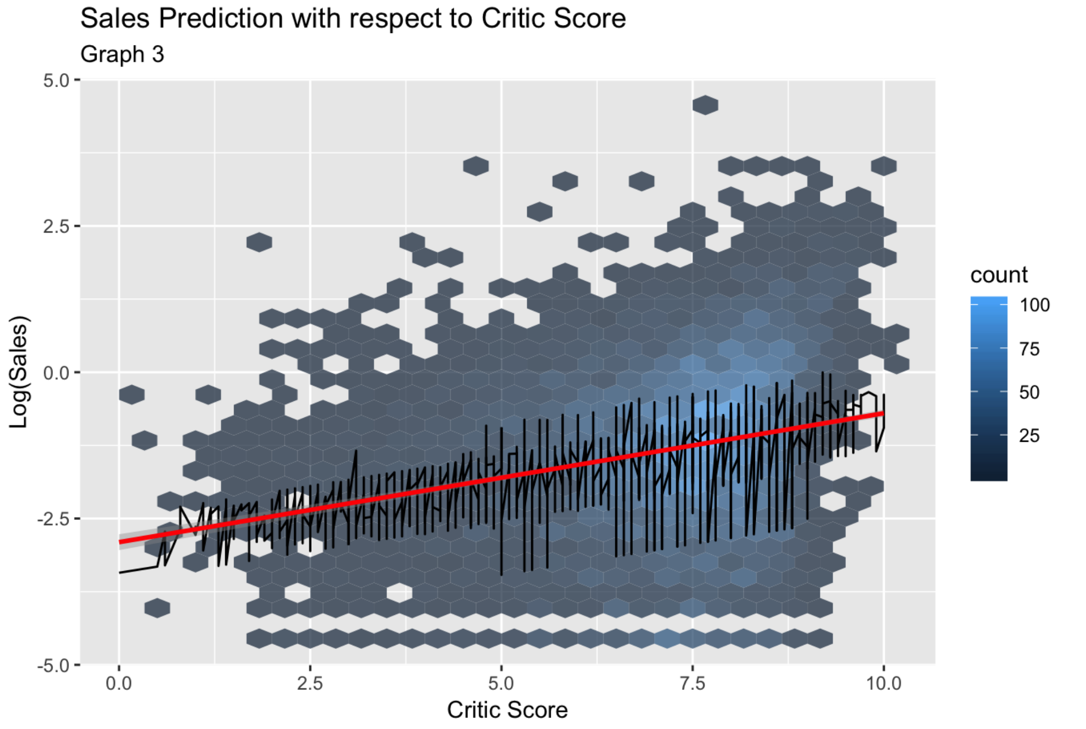
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = plant.lm)
##
## $Genre
##                        diff         lwr         upr     p adj
## Misc-Adventure    0.2058703  0.02343979  0.38830082 0.0223013
## Puzzle-Adventure -0.1827238 -0.44378093  0.07833339 0.2283857
## Puzzle-Misc      -0.3885941 -0.64518477 -0.13200338 0.0011392
```

Our null hypothesis for Tukey’s test is that the medians for two genres are the same, in our alternative hypothesis, we say that they are not. The only statistically significant difference is observed between puzzle and adventure, as the p-value is bigger than 0.05. Another reason to conclude that would be 0 that is included in the confidence interval, which means that the potential difference between those groups can be 0. Other than that, we conclude that there are significant differences between Misc and Adventure, and Puzzle and Misc. Therefore, sales will be affected differently by Misc and other genres. Whereas, there won’t be a huge difference in the effect on sales between puzzle and adventure.

Prediction of the Response

After assessing to what extent the independent variables affected global sales, we plotted prediction graphs for both Critic and User Scores to predict global sales using each of the two independent variables.

```
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 3",
  caption = "The theoretical linear model  in red shows the trend of Sales growth related to critic score. T
he black line shows the actual data.")
```

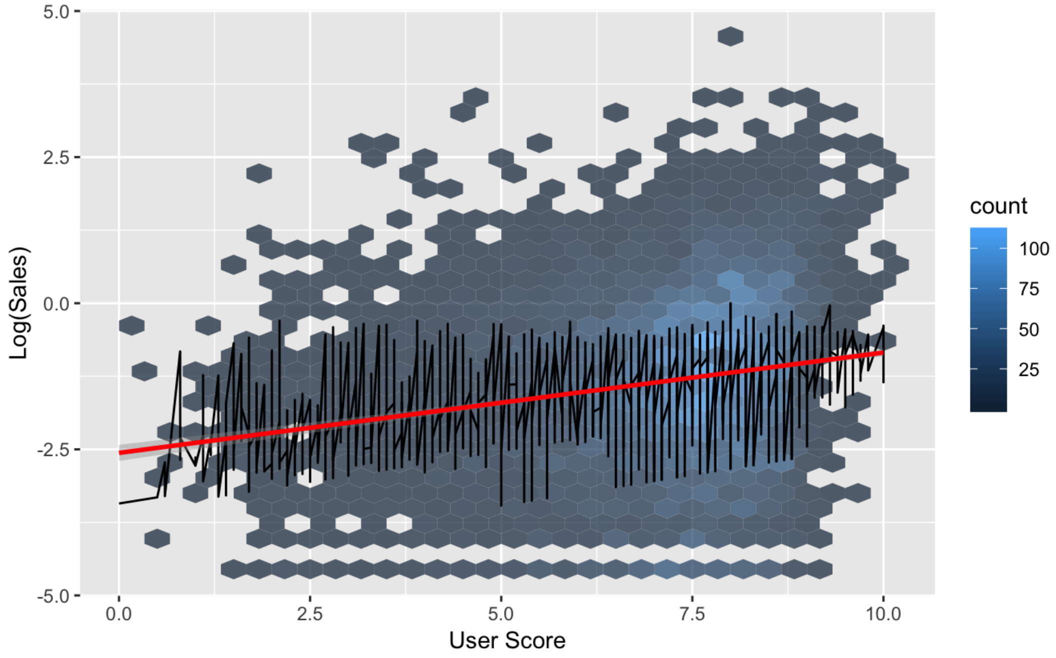


l linear model in red shows the trend of Sales growth related to critic score. The black line shows the actual data.

```
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 4",
  caption = "The theoretical linear model  in red shows the trend of Sales growth related to critic score. T
he black line shows the actual data.")
```

Sales Prediction with respect to User Score

Graph 4

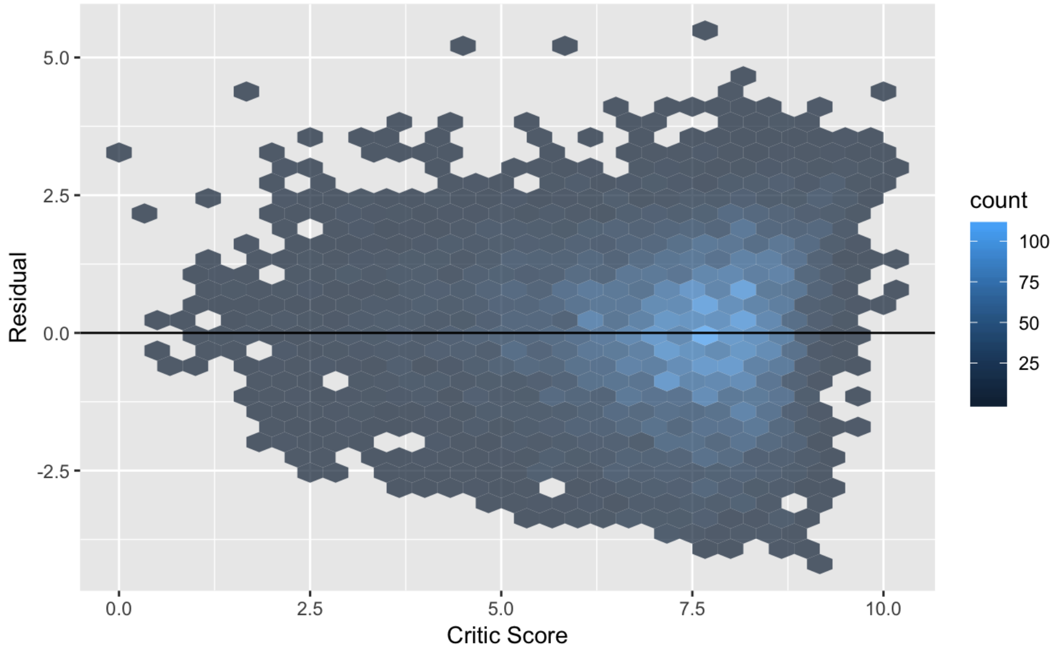


l linear model in red shows the trend of Sales growth related to critic score. The black line shows the actual data.

```
ggplot(data_filtered %>% add_residuals(mod), aes(x = Critic_Score, y = resid)) + geom_hex(alpha = 0.7) + geo
m_hline(yintercept = 0) + labs(
  x = "Critic Score",
  y =  "Residual",
  title = "Residual with respect to Critic Score",
  subtitle = "Graph 5",
  caption = "The black line portrays the residual error of the model.")
```

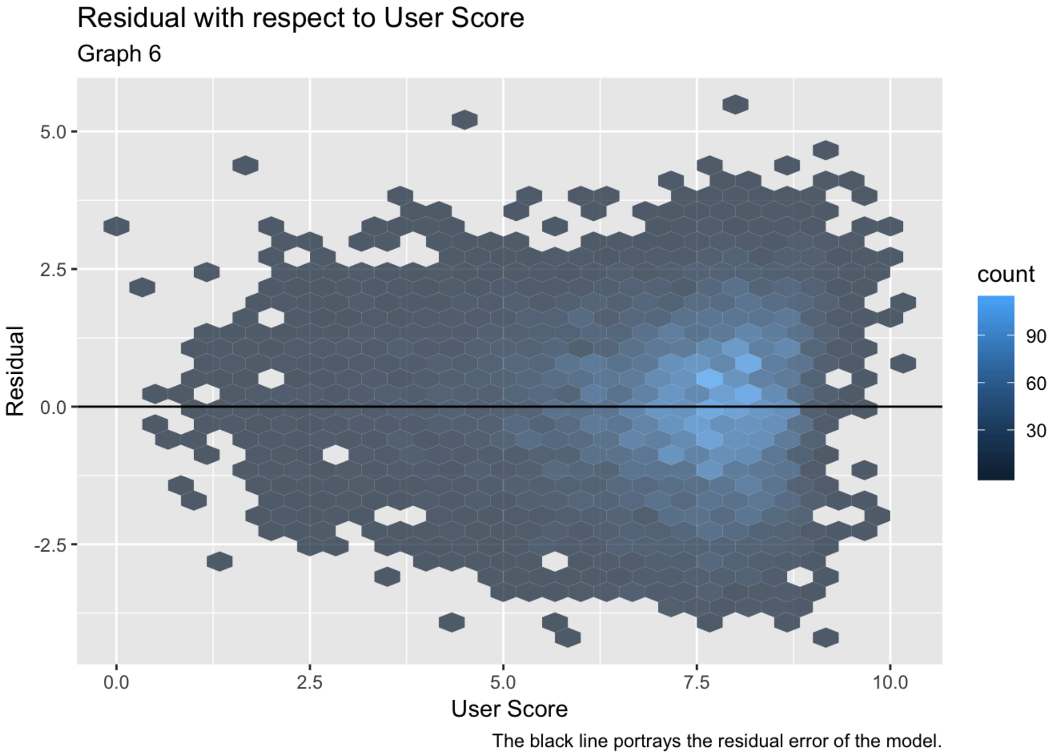
Residual with respect to Critic Score

Graph 5



The black line portrays the residual error of the model.

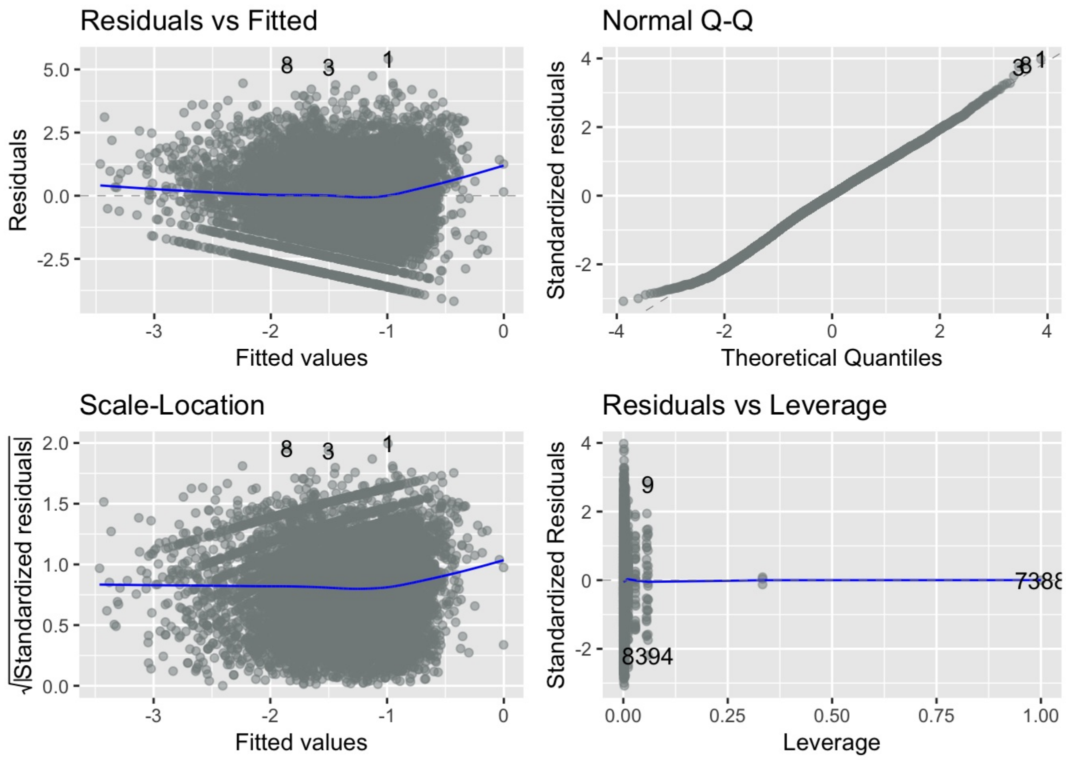
```
ggplot(data_filtered %>% add_residuals(mod), aes(x = User_Score, y = resid)) + geom_hex(alpha = 0.7) + geom_
hline(yintercept = 0) + labs(
  x = "User Score",
  y =  "Residual",
  title = "Residual with respect to User Score",
  subtitle = "Graph 6",
  caption = "The black line portrays the residual error of the model.")
```



Graphs 3 and 4 show the prediction model of the effect of critic score and user score on sales, respectively. There is much more noise in the critic score plot as the black line is more scattered. The trend line in red has a bigger slope as well. These observations makes sense as we have established earlier that critic scores have more effect on sales than user scores. Graph 3’s curve also has more residuals for higher critic scores and graph 4’s residuals are much more staggering than the previous one, especially for lower scores. This ultimately means that critics are more forgiving when rating games than users. Graphs 5 and 6 portray the residuals with respect to critic scores and user scores, respectively. While they do not deliver too much information, they do show how most of the noise is around critic and user scores of around 7.5. This means that most games from various genres are rated as such.

Residual Analysis to Check Model Assumptions

```
library(ggfortify)
par(mfrow = c(2, 2))
autoplot(mod, alpha = 0.5, colour = 'azure4')
```



After fitting a linear regression model, we checked our assumptions about the normality of the model by plotting the above graphs. These 4 plots also serve to detect potential problems and include: Residual vs Fitted, Normal Q-Q, Scale-Location and Residual vs Leverage.

The first one checks for the linearity assumption, i.e. whether our model has a linear relationship between dependent and independent variables. It does so by showing us the relationship between residual and fitted values. A residual is the difference between the observed value of the dependent variable (y) and the predicted value (ŷ). As seen, the residual plot shows almost no fitted pattern, as our data is generally symmetrically distributed around y=0 (blue line). This indicates that there is almost no difference in the gathered data and predicted values, which satisfies a linearity condition.

The second graph is a normal Q-Q (Quantile-Quantile), which indicates to what extent is our data normally distributed. Two lines indicate a comparison between our data and normal distribution by plotting their quantiles against each other. Because the plotted “experimental” line almost identically mimics a straight fitted line, we can safely assume normality. The third plot portrays a Scale-Location relationship, which provides us with information on whether residuals are spread equally along the ranges of predictors. The horizontal blue line coupled with an equal spread of data points means that all of the independent variables have the same variance (deviation of a variable from its mean). In other words, the variability of the residual points stays the same across all values of the fitted outcome variable. We can therefore conclude that variances across independent variables are very similar.

The last plot we created was fitting Residuals vs Leverage (a measure of how far away independent values are from each other). We then proceed to spot any outliers that would affect the normality of the data. As we can see, there are some extreme points like “394” and “7388”, but in terms of residuals, none of those values exceed 3 standard deviations in absolute value, which allows us to conclude that there are no extreme outliers that need to be omitted.

Discussion

The main goal of this research was to find the most significant factors that have an impact on video game sales. We conducted this analysis using linear regression models and hypothesis tests. We found out that both critic score and user score were influential to a limited extent. Critic score, however, appeared to be much more significant than user scores, which means that higher critic scores could result in higher sales. Critic Score data is also much cleaner as the prediction models portray. This makes sense as critics tend to use specific parameters and metrics to rate games, unlike actual gamers who are more subjective. These results are interesting and prove the way critics try to stay objective as much as possible. Another important finding was related to game genres. Several of them that had significant effects on sales, but only a few of those had a positive effect. According to our analysis, “adventure”, “miscellaneous”, “puzzle”, “simulation” and “visual novel” categories have a negative impact on sales, which implies that sales would be lower for games of those genres. The best genre for the video games industry ended up to be “shooter”, with “action” taking the second place. This finding is aligned with the result we found in the first deliverable as we saw how “action” category grew in popularity over the years.