Video Game Trends Project

MA415

Group 1
Abdulshaheed Alqunber
Rami Bassil
Ekaterina Gorbunova
Khalid Khumayis

04/24/2019

Table of Contents

Abstract

This project serves to analyze how less obvious factors such as game genre, and critic and user scores affect global video game sales. First, we identified that some genres such as “Action” and “Sports” are the most popular among users, which provided us with useful information on growing trends in the industry. We then created a linear regression model and evaluated that critic scores affected sales to a much greater extent than user score. We also found that several game genres such as “Visual Novel” and “Education” affected sales negatively, while others like “Sports” and “Party” affected sales positively. Finally, sales predictions were made using Genre, Critic and User Score. This analysis ultimately proves that critic scores, while not one of the main factors affecting sales, has a more important impact on sales than user score.

1. Introduction

The video game industry has been around for over 30 years and was worth $43.4 billion in 2018. With so many different genres and platforms to choose from, consumers have leverage over incumbent gaming companies. Therefore, it is essential for those firms to analyze key trends and understand what types of games they should invest in.

A multitude of factors affects global sales of video games. Most obvious ones include consumer spending power, marketing, cultural trends, and the improvement of chips used in computers and consoles. A lot of other factors also have an effect on sales, but to a lesser extent. The goal of this project is to analyze to what extent some of those less important factors such as genre, user and critic satisfaction affect the sales of video games. By inspecting past trends in the industry, especially how genre affects popularity, we created a prediction model to give game publishers an understanding of how "less obvious" factors affect video game sales.

The initial dataset used was found on kaggle.com. Initially, the dataset did not include observations after 2015 and had many missing values in other columns as well. Therefore, we traced back how the dataset was created to a website called VGChartz and made our own video games dataset that includes games up to the first quarter of 2019. The dataset created includes almost 56,000 games, almost triple the size of that of the original dataset.

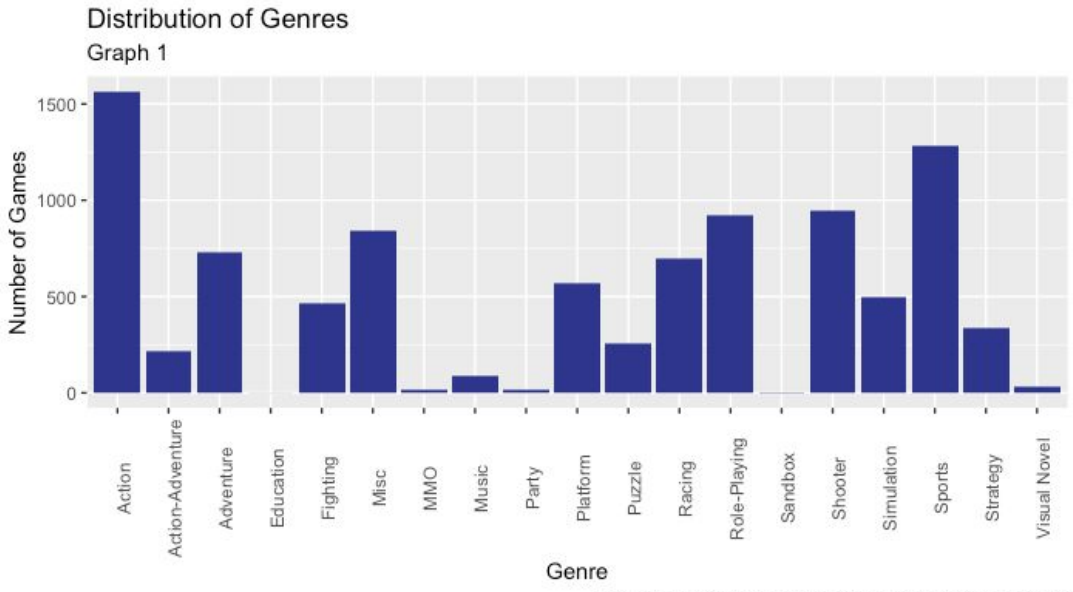
The dataset contains both qualitative and quantitative variables. Those are Name, Genre, ESRB_Rating, Platform, Publisher, Developer, VGChartz_Score, Critic_Score, User_Score, Year of Release, Sales, Sales by region (North America, Japan, Europe, and Other), and Total_Shipped. However, not all variables were used in the research since the focus was on game

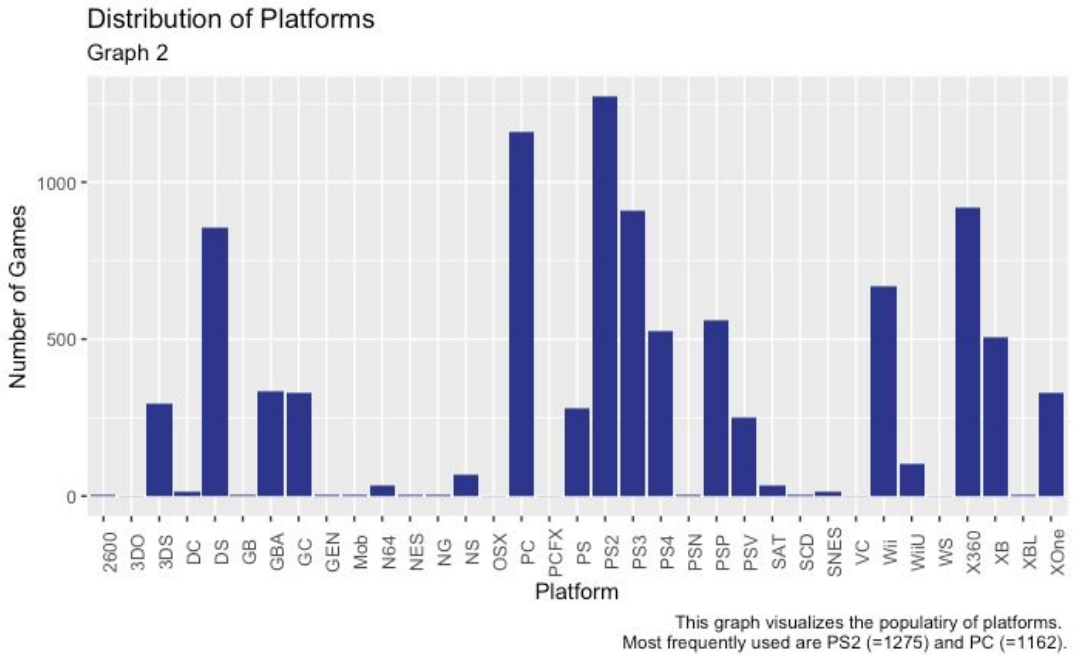
genres, game scores and how they affect gaming consumption. Ultimately, the only variables that were used are Name, Genre, Platform, Critic_Score, User_Score, Year of Release, Sales, and Total_Shipped. Furthermore, the dataset was cleaned from any null values and was left with 9,490 video games used for the analysis. Table 1 in Appendix 1 describes each of those variables.

2. Methods

a. Variation of Single Variables

First, a quick study of the variation of single variables was conducted to see their own patterns and distributions. Graphs 1 and 2 below provide information about counts of genres and platforms, respectively.

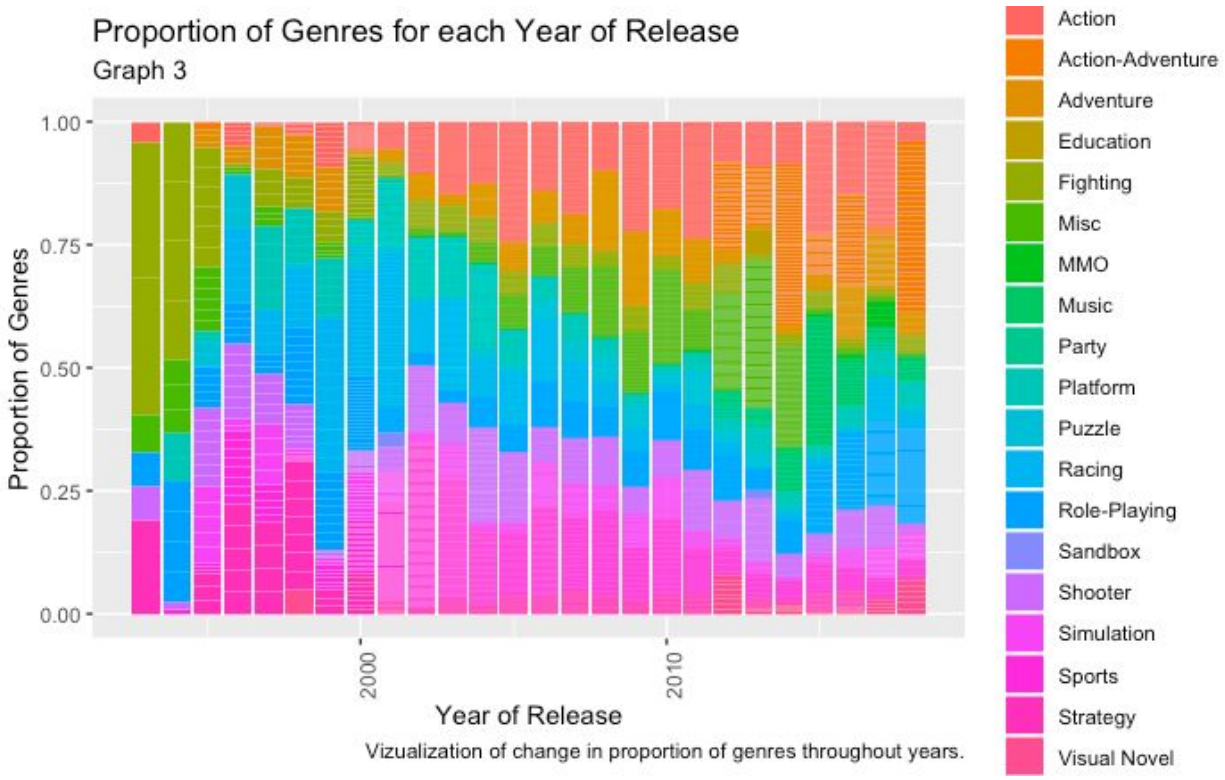




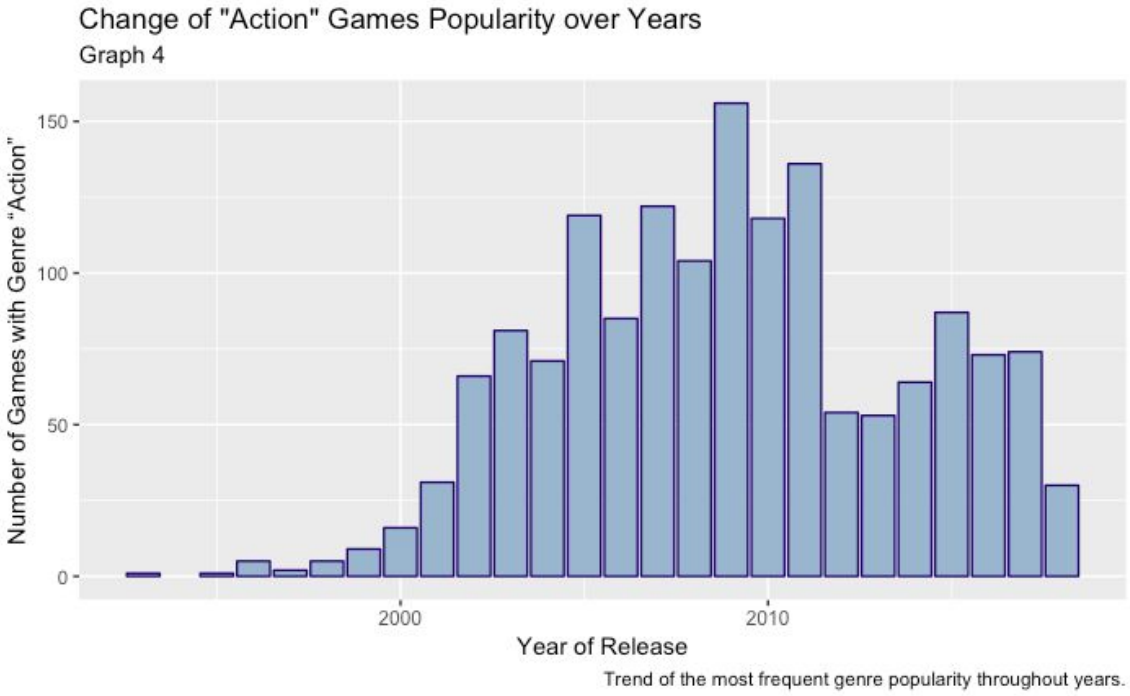
Graph 1 takes unique values of genres and creates a plot with a count for each on Y-axis and genres themselves on X-axis. Based on this plot, it looks like the most popular genre is "Action" with 1567 entries, followed by "Sports" with 1283 records. The least frequent ones appeared to be "Education" and "Sandbox". The same analysis was done with platforms to investigate their popularity (Graph 2). Because the data contains information from the year 1980, two most popular platforms appeared to be Nintendo DS and PlayStation 2.

b. Variation of Multiple Variables

Next, we conducted correlations between two variables, Genre and Year of Release, to understand how they affect each other, as seen in Graph 3 below.

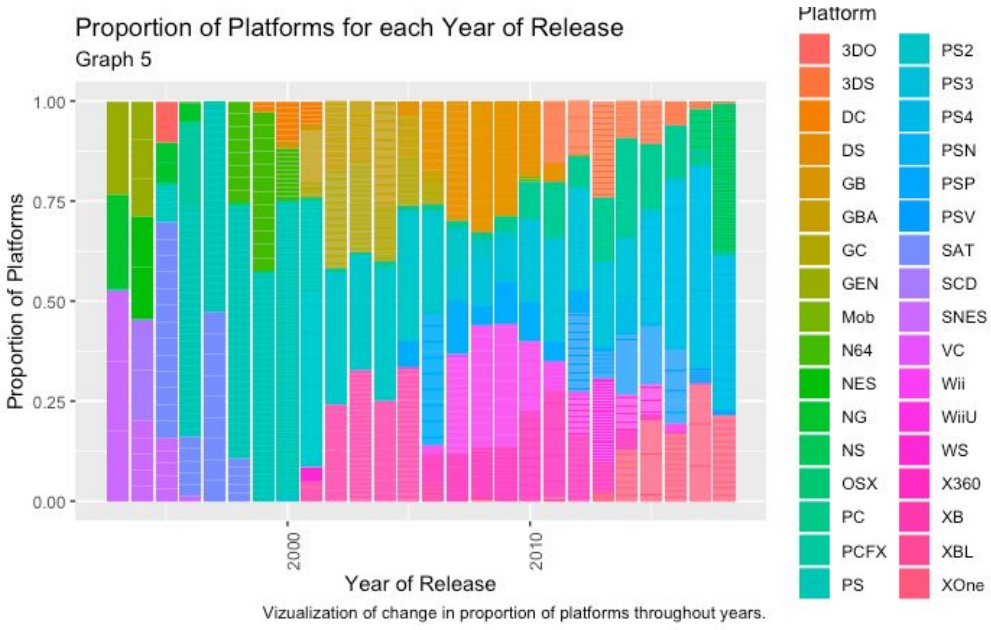


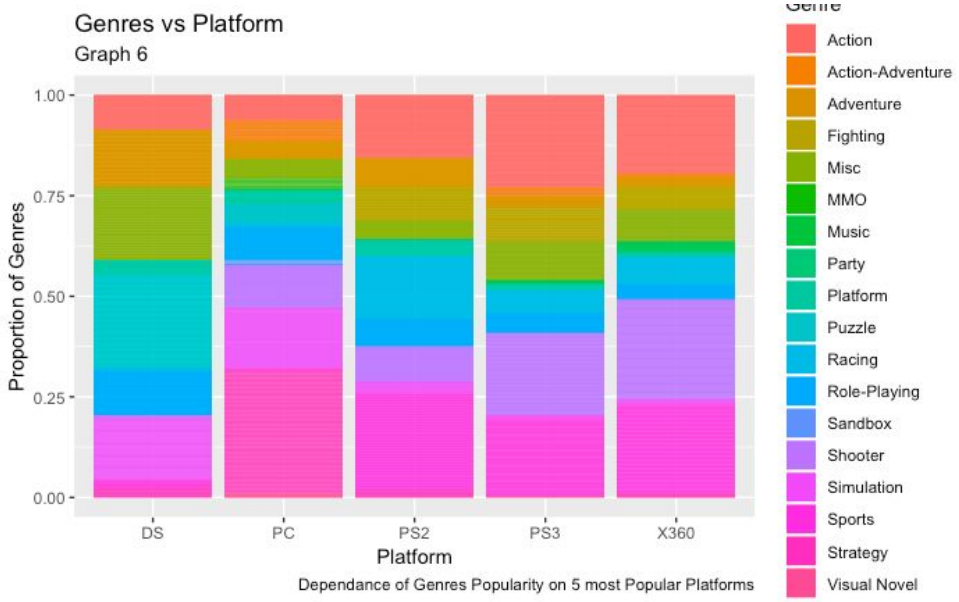
The plot says a lot about the distribution of genres, especially pointing out the rise of popularity “Action” games. Since it was the most popular genre overall in the dataset, we decided to look at it individually, correlating it to years of release, as shown in Graph 4 below.



We observed a quick increase in releases for “Action” games starting in the late 1990s, with a peak in 2009. This sharp increase of Action games is correlated with the introduction of the PlayStation console in 1994, which was the first console to popularize Action games such as Battle Arena Toshinden and Street Fighter.

Two analysis similar to the one in Graph 3 were done: the proportion of platforms vs years and proportion of genre vs platforms, as seen in Graph 5 and 6 below.





Graph 5 clearly shows how new platforms were introduced and some became obsolete, like some of the PS systems in blue.In graph 6, we correlated the proportion of genres for the 10 most popular platforms, to see whether a relationship exists between those two variables and if a specific platform causes any particular genre to be more popular.

c. Discussion

After analyzing and visualizing the data, we found that variation of single variables gave us information about the most popular genres and platforms in our dataset. As it turns out, "Action" is the most popular genre, as seen in Graph 1. However, it only started to increase in popularity in 1995, when the number of games for this genre released doubled and then started growing rapidly (Graph 4). There is a clear explanation for that as the first PlayStation console, the PS, was released on December 3, 1994. Meanwhile, publishers started creating games for that new console, which started being released in 1995. We can also see statistical evidence of this fact on the graph 5, as the blue color that represents PS starts to appear in the year 1995. This relationship between the introduction of PlayStation and action games cannot be a coincidence and shows that the introduction of PlayStation consoles ignited the development of a new genre.

That being said, action games were not as high in popularity with the PS as they are with later versions. Instead, the main focus of the first PS platform were sports and racing games, as can be seen on graph 6, but shifted more towards action games with PS2, then PS3. This makes sense because as with every other new trend, it takes some time for them to take off.

Overall, this analysis was a helpful starting point for our research, indicating most popular genres and their trends throughout years in terms of sales.

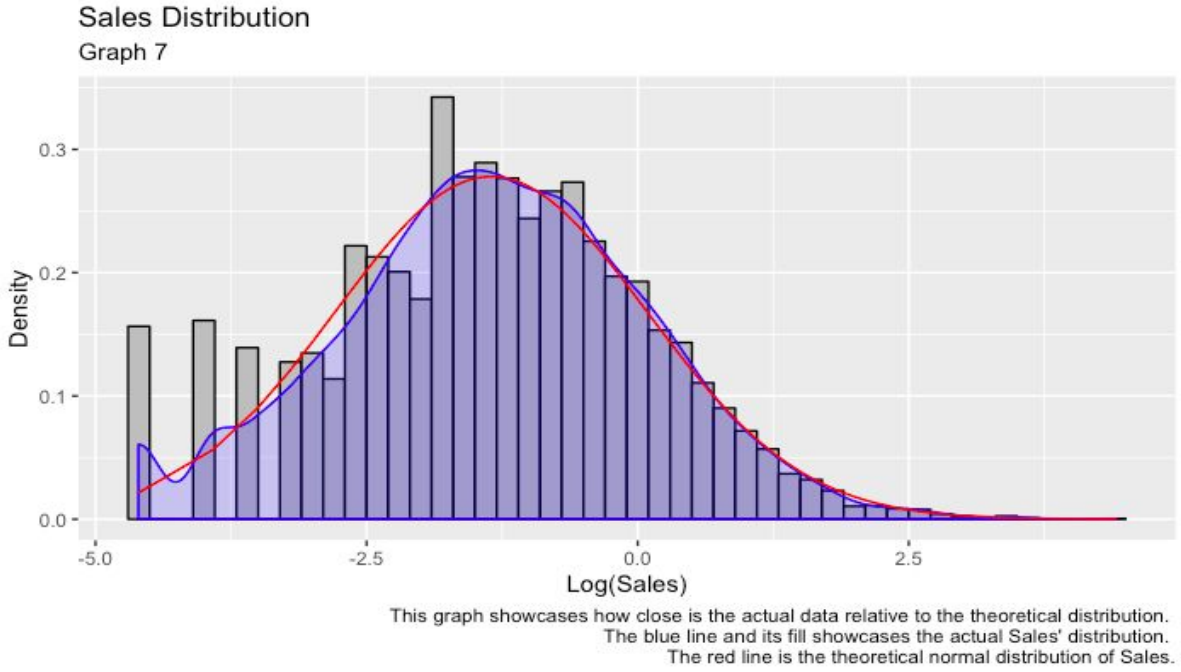
3. Modeling Results

a. Transformation of Variables

In order to achieve normality in our 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) has an exponential relationship with the variables Genre, Critic Score, and User Score. This allows having clearer and more aesthetic curves. Therefore, using the natural log for our dependent variable only linearizes the relationship curve and results in a better fitted model.

b. Check for Assumptions using Linear Regression

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.



According to the above curve, video game sales 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.

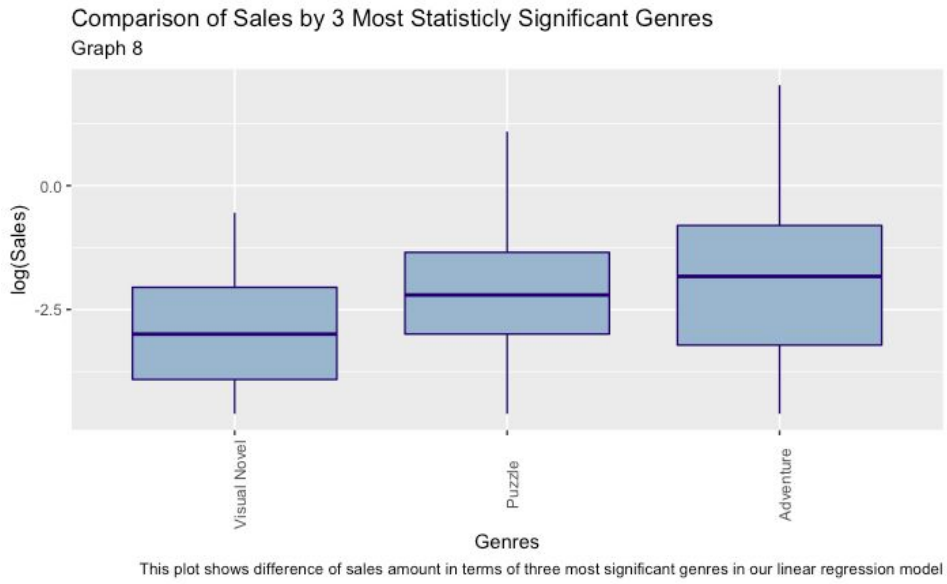
After plotting the linear regression model, the first thing we have done was checking the assumptions that must be held in order for the model to be accurate. Those graphs and descriptions for each can be observed in Appendix 3.

c. Linear Regression Results

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. We have also plotted a model including interaction between Genres and User Score. What we observed is that almost no interaction terms showed significance, and our R^2 increasing to 11.2%. That being said, it does show a clear correlation between genre and score, and sales, something that game publishers might not have thought about.

d. Tests for Significance Difference

After fitting a linear regression model and getting p-values for each variable, we selected the genres that were 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.



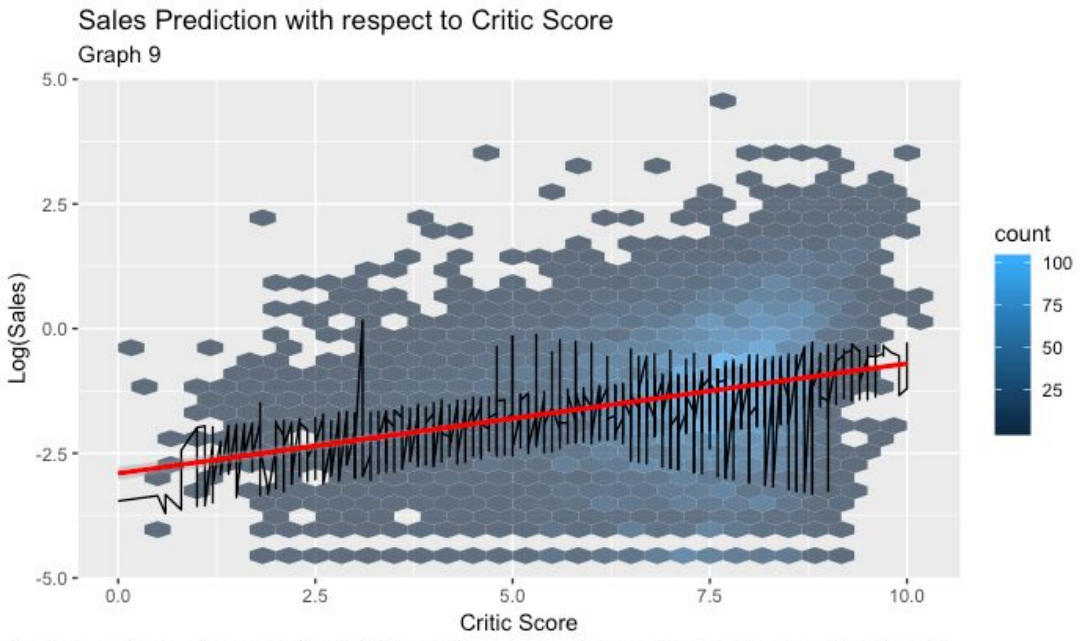
In the plot, we observe the three most statistically significant genres in terms of our regression model, Visual Novel, Puzzle and Adventure. The box plots represent the natural logarithm of sales. As we can see from the black horizontal lines, the medians for those genres are approximately the same. 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.

Our null hypothesis for Tukey's test is that the medians for the two genres are the same, in our alternative hypothesis, we say that they are not. The only statistically significant difference

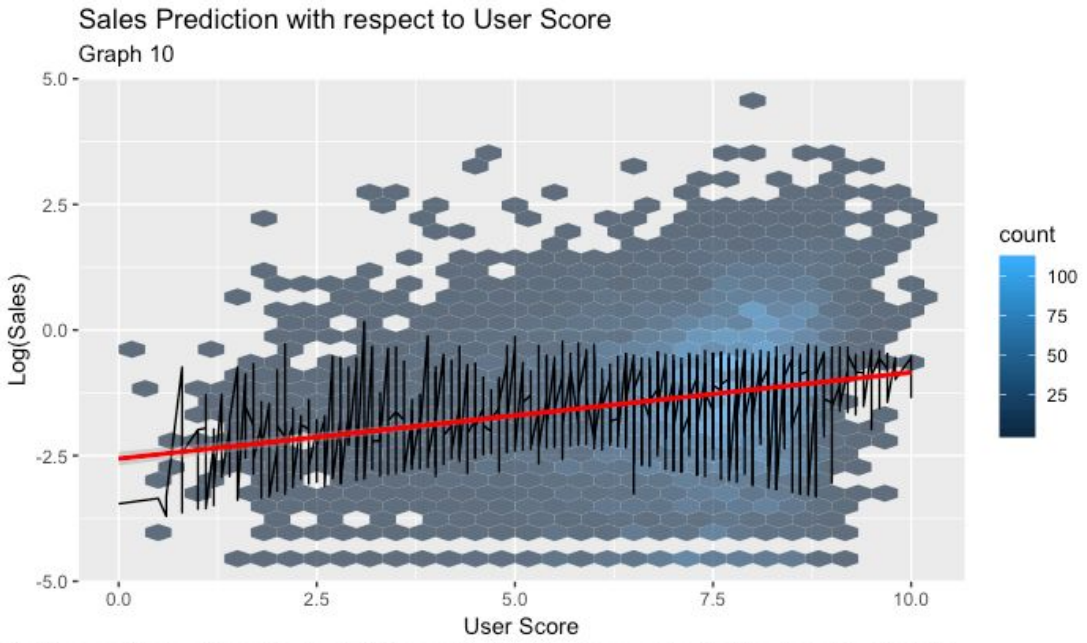
is observed between puzzle and adventure, as the p-value is bigger than 0.05. Another reason to conclude that would be 0 that is included in the confidence interval, which means that the potential difference between those groups can be 0. Other than that, we conclude that there are significant differences between Visual Novel and Adventure, and Visual Novel and Puzzle. Therefore, sales will be affected differently by Visual Novel and other genres. Whereas, there won't be a huge difference in the effect on sales between puzzle and adventure.

e. 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.



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 9 and 10 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 make sense as we have established earlier that critic scores have more effect on sales than user scores. Graph 9's curve also has more residuals for higher critic scores and graph 10'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 11 and 12 portray the residuals with respect to critic scores and user scores, respectively, and can be found in Appendix 4. 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.

4. Discussion

The main goal of this research was to identify to what extent “less important” factors 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. Even though those results provide us with an interesting understanding of game genre trends, it is important to keep in mind that the overall effect of genres is not very significant.

Finally, Another interesting finding is that even though “Action” games have been the most popular over the years, they do not affect global sales positively. This observation proves again that genre does not affect global sales as it is based on taste of customer. While the model therefore does not work for this specific category, it could be used to evaluate how scores affect sales.

5. References

1. Alqunber, Abdulshaheed. “Video Games Sales 2019.” *Kaggle*, 13 Apr. 2019, www.kaggle.com/ashaheedq/video-games-sales-2019.
2. “NPD: US Video Games Sales Reached $43.4 Billion Last Year.” *GamesIndustry.biz*, www.gamesindustry.biz/articles/2019-01-23-npd-us-video-games-sales-reached-USD43-4-billion-last-year.
3. “PlayStation.” *Sony's First Video Game Console Established the PlayStation Brand. It Dominated the 32/64-Bit Era and Was the Best-Selling Home Console up until the PlayStation 2.* https://www.giantbomb.com/playstation/3045-22/

Appendix 1

Table 1: Variables used in the study and their meaning

Variable	Meaning
Name	Name of game
Genre	Classification of type of game
Platform	Gaming system used
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
Total_Shipped	Number of games sold by Nintendo

Appendix 2

Table 2: Summary of the linear model results

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	Significance Level
Intercept	-2.69662	0.07622	-35.378	< 2e-16	***
Critic_Score	0.25772	0.01789	14.408	< 2e-16	***
User_Score	-0.4760	0.01795	-2.652	0.00802	**
Action - Adventure	0.21220	0.09809	2.163	0.03054	*
Adventure	-0.73042	0.06094	-11.986	< 2e-16	***
Education	-1.37765	1.36033	-1.013	0.31121	
Fighting	0.02117	0.07197	0.294	0.76863	
Misc	-0.46826	0.05822	-8.043	9.79e-16	***
MMO	0.70457	0.32230	2.186	0.02883	*
Music	-0.10565	0.14896	-0.709	0.47819	
Party	0.75403	0.33151	2.264	0.02362	*
Platform	0.12814	0.06653	1.926	0.05412	.
Puzzle	-0.88597	0.09137	-9.697	< 2e-16	***
Racing	-0.11020	0.06204	-1.776	0.07571	.
Role-Playing	-0.15262	0.05657	-2.698	0.00699	**
Sandbox	0.14799	0.78564	0.188	0.85059	
Shooter	0.22905	0.05600	4.090	4.35e-05	***
Simulation	-0.29440	0.06985	-4.215	2.52e-5	***
Sports	0.10093	0.05152	1.959	0.05013	.
Strategy	-0.58090	0.08200	-7.084	1.50e-12	***
Visual Novel	-1.81732	0.23241	-7.819	5.88e-15	***

Signif. Codes:

0 `***`

0.001 `**`

0.01 `*`

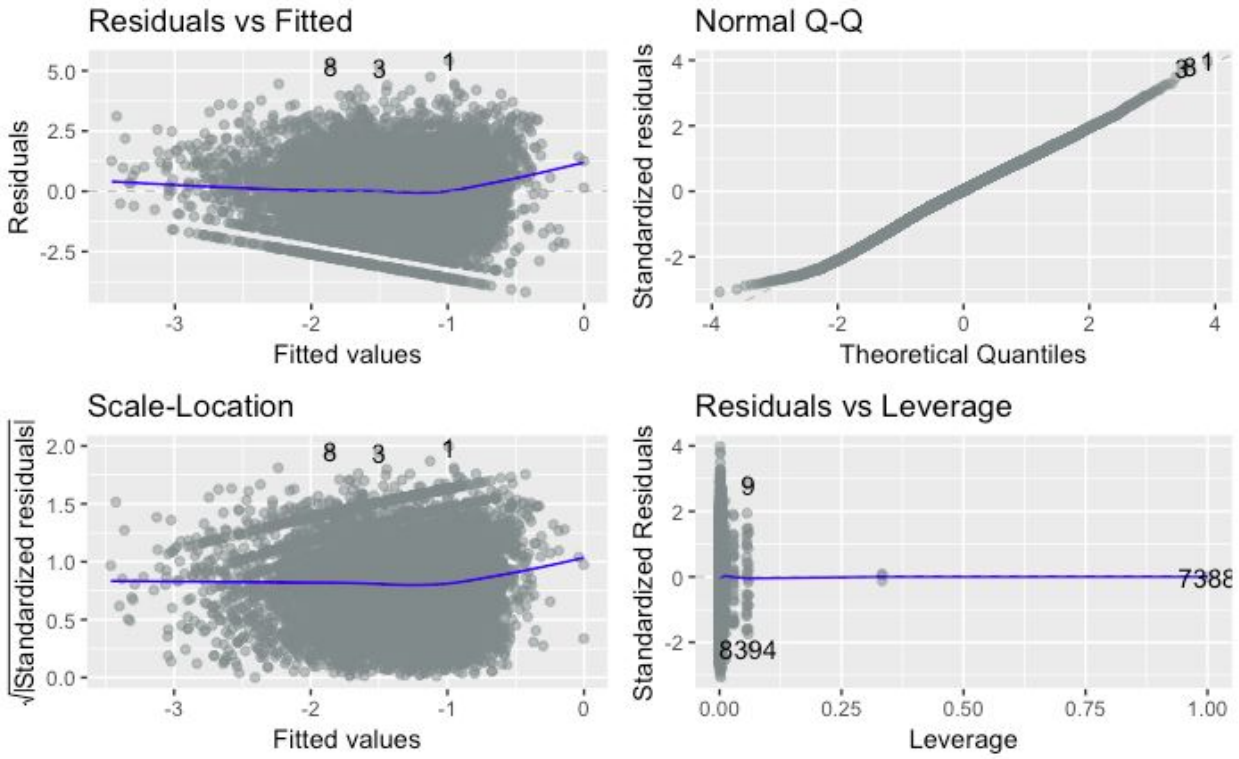
0.05 `.`

1 ``

Multiple R-squared: 0.1052

Appendix 3

Residual Analysis to Check Model Assumptions



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 a 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.

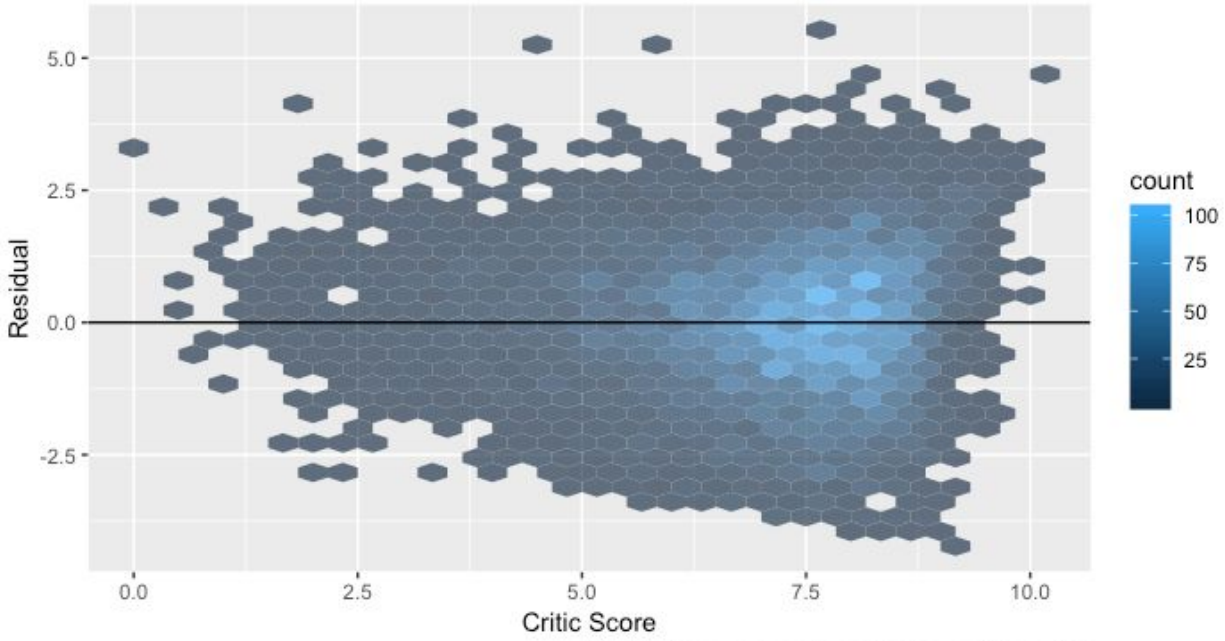
The third plot portrays a Scale-Location relationship, which provides us with information on whether residuals are spread equally along with the ranges of predictors. The horizontal blue line coupled with an equal spread of data points means that all of the independent variables have the same variance (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.

Appendix 4

Residual with respect to Critic Score

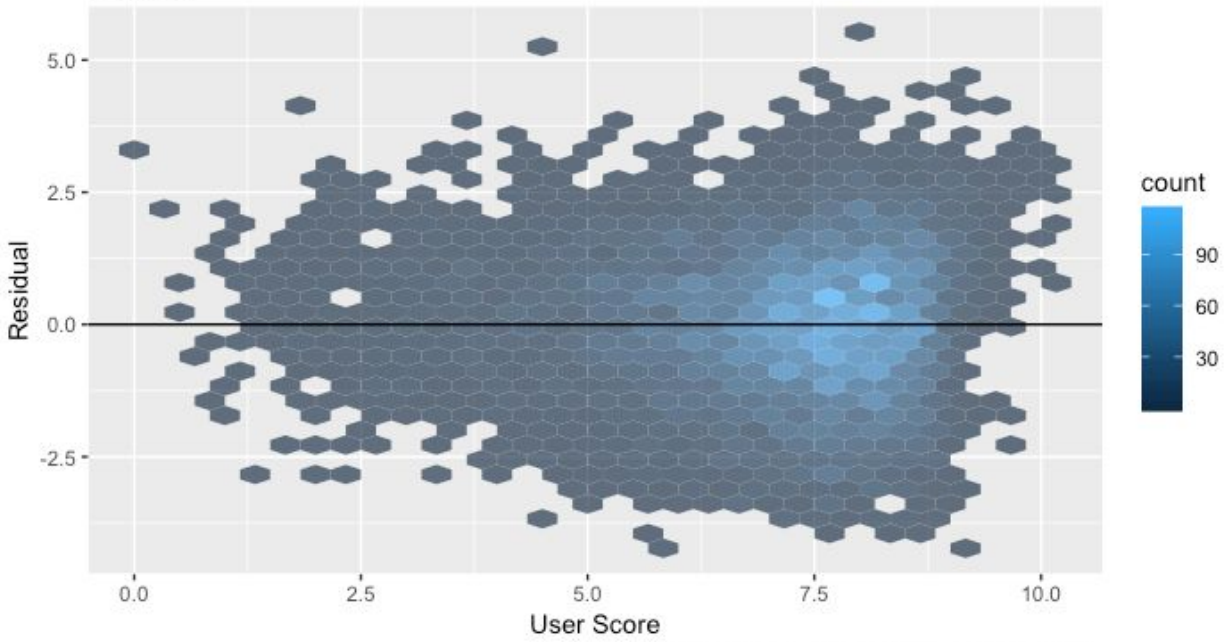
Graph 11



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.