Pixels to Profit: An Analysis of the Video Game Sales and Trends

Ha Doan
DA/MATH 220
Dr. Alice Miller

Spring 2023

1. Summary/ Abstract

This project analyzes data of video game sales globally from 1980-2017 to explore leading genres, platforms, and publishers in sales globally and also in North America, Europe, Japan, and Other regions. The project also aims to identify trends or relationships of video game sales with other variables to make future predictions and inform game development decisions. The statistical methods used in the project include descriptive statistics, (multivariate) linear regression, ANOVA, and Chi Square tests.

2. Introduction

In 2021, the video game industry is estimated to be worth roughly **$178.73 Billion** (an increase of 14.4% from 2020) with around 3.2 billion players worldwide. Gaming has been a popular and mainstream hobby (especially since the COVID-19), so it should come as no surprise that the industry will keep growing with a forecasted value of $268 Billion by 2025.

This project then will look at the data of video game sales globally over the past 40 years to understand more about the industry trends as well as its growth potentials. The data is from Kaggle (and sourced/ recorded from vgchartz.com). The data was published in 2022 (but mostly includes data from 1980-2017), and offers insights about video games' platforms, genres, publishers as well as their sales across the globe and in specific regions like North America, Europe, Japan, and Others. The aim of this project is to explore about:

1. What genres - platforms - publishers are leading global sales and compare sales in each region (mainly North America, Europe, Japan, and Other).
2. Trends/ relationships between video games sales with their genres - platforms, and make predictions about future trends.

These insights will not only give a better understanding of consumer behaviors but also potential/ targeted markets or genres-publishers that could inform future game development decisions.

Last but not least, to achieve the project's goals, the following statistics methods were used in the project: Descriptive Statistics Summary, Graphs, Linear Regression (with Residuals, Inference), Multivariate Linear Regression, ANOVA, Chi Square Tests.

3. Methods

3.1. Market Share of Each Region

a. Graphical and Descriptive Statistics

First, I calculate the percentage of each region's sales over global sales, which is shown in the table below.

North America	Europe	Japan	Others	Global
49.25%	27.29%	14.47%	8.99%	100%

Figure 1. Table of the overall sales proportion in each region from 1980s-2010s.

The following graphs also depict trends of sales percentage over time in each region.

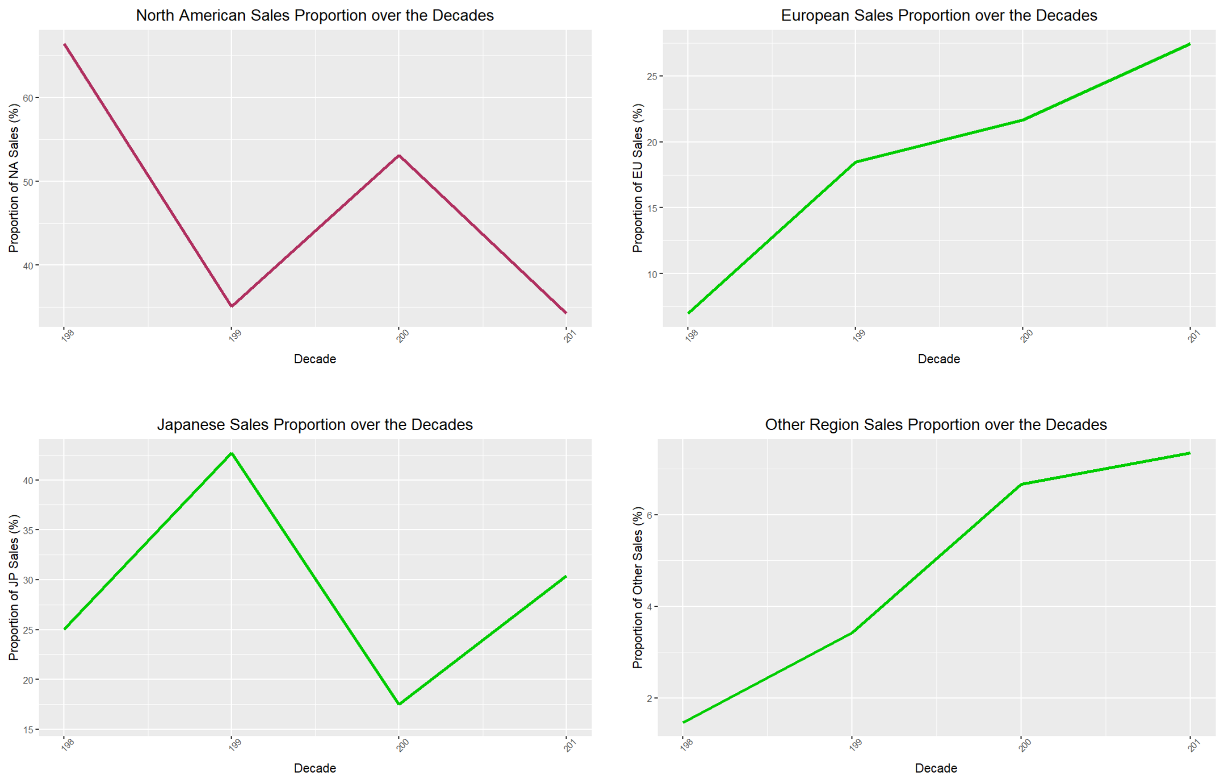


Figure 2. Sales proportion changes in NA, EU, JP and Other from 1980s-2010s.

Figure 2 suggests that the percentage of North America over global sales decreases generally (from more than 60% to around 40%), while European, Japanese and Others sales start to account more for global sales. Thus, while indeed North America still occupies the majority of global video game sales share (as suggested in Figure 1), producers might want to think about their growth potentials in other markets now.

b. Inference Test and Linear Regression Model

I am particularly interested in Other Sales, as it seems to be growing and yet remains quite ambiguous in this dataset. More particularly, I want to explore the relationship between Other Sales and NA/ EU/ JP, as well as if we can predict Other Sales. As a result, I opt for the linear regression model for this case. But first, I want to check the correlation between sales in these continents.

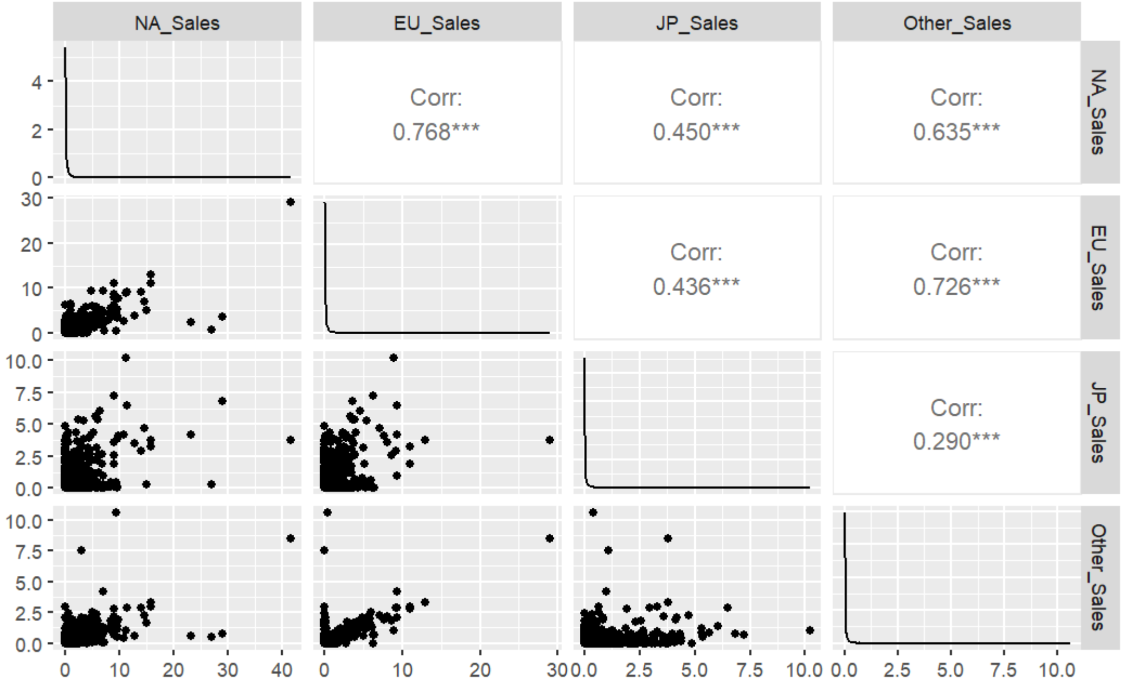


Figure 3. Graphs of correlation between NA, EU, JP and Other Sales.

Looking at this figure, it is clear that *Other_Sales* has the strongest correlation with *EU_Sales*. Yet, we can do a hypothesis test for Inference of Linear Regression to see if there is an actual linear relationship between video games sales of Europe and Other.

i. Hypothesis

H_0: ρ = 0 (no linear relationship)

H_a: ρ ≠ 0 (some linear relationship)

Where ρ is the actual correlation between Other Sales and European Sales.

ii. Test statistics

$$df = 16597 - 2 = 16595$$

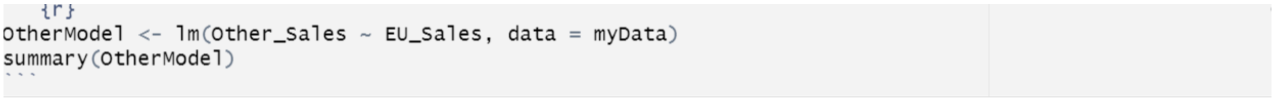
$$t - value = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.72\sqrt{16597-2}}{\sqrt{1-(0.72)^2}} = 133.7$$

iii. Finding p-value

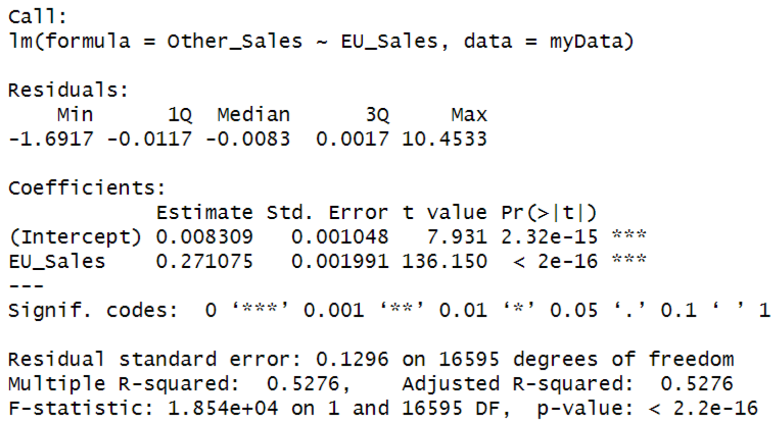
$$p - value = 1 - pt(133.70,\ 16595) \sim 0$$

iv. Conclusion

Since the p-value is small (<0.05), we could reject the null hypothesis and conclude that there is an actual linear relationship between sales in Europe and Others. Then, the linear regression is used to see if we can predict Other Sales based on the EU sales.



~~~
  {r}
OtherModel <- lm(Other_Sales ~ EU_Sales, data = myData)
summary(OtherModel)
```
~~~



```
Call:
lm(formula = Other_Sales ~ EU_Sales, data = myData)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6917 -0.0117 -0.0083  0.0017 10.4533 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.008309   0.001048   7.931 2.32e-15 ***
EU_Sales    0.271075   0.001991 136.150  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1296 on 16595 degrees of freedom
Multiple R-squared:  0.5276,	Adjusted R-squared:  0.5276 
F-statistic: 1.854e+04 on 1 and 16595 DF,  p-value: < 2.2e-16
```

Figure 4. Linear Regression Model for Other and EU Sales.

The model above tells us that if our EU sales increase by 1 million, then we expect/ predict the Other Sales to increase by 0.27 million. This also means that if EU sales is 0, then we expect/ predict the Other Sales to be just 0.008 millions.

The p-value is very small and less than 0.05, which means that the model is statistically significant. However, The R-squared tells us that of all variation in Other Sales, about 52.76% is due to the linear relationship between Other Sales and EU Sales. But before any final conclusion, I want to double-check if linear regression is the best model for this case.

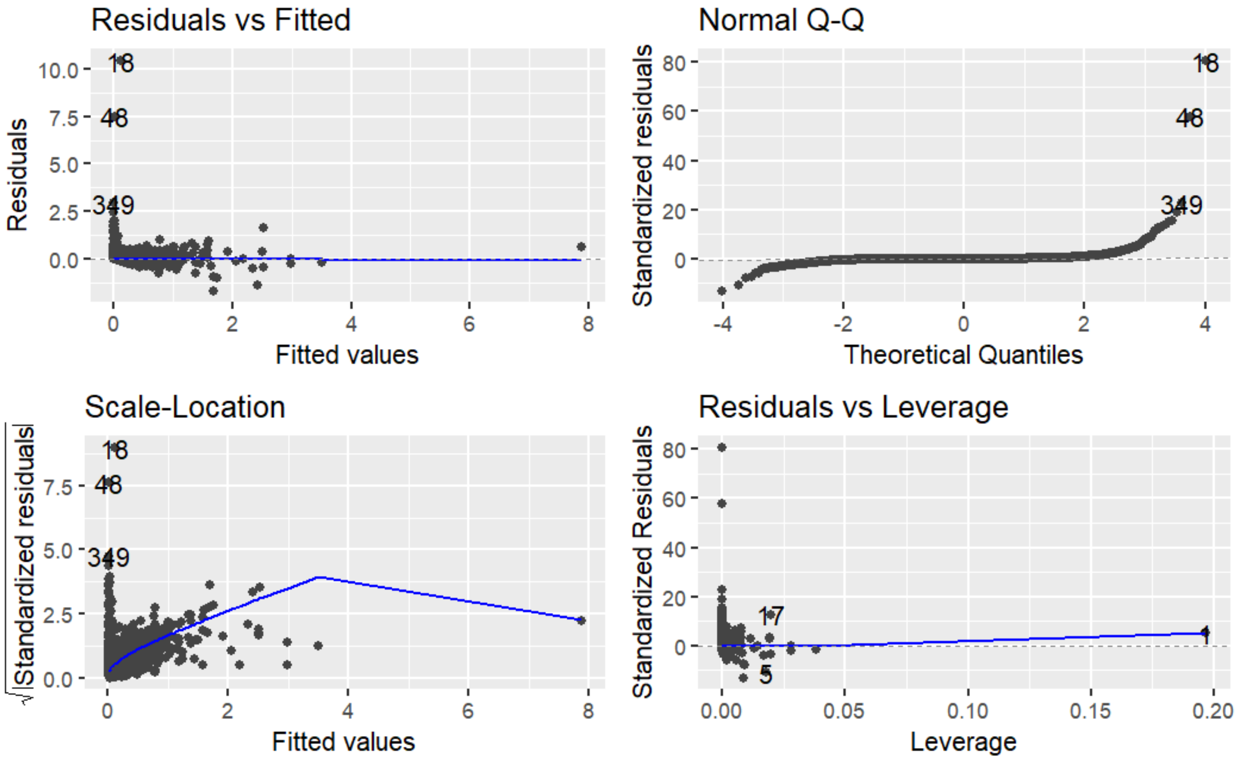


Figure 5. Linear Regression Model for Other and EU Sales.

Residuals vs Fitted and *Scale-Location* show that all residuals are clustering together and very heteroskedastic, while the *Normal Q-Q* plot suggests similarly, since our residuals deviate from the reference line a lot and thus do not follow normal distribution. All of these evidences mean that the linear regression model might not be the best indicator of relationship between EU Sales and Other Sales in this case, and that the prediction might not be very well-established.

3.2. Video Games Sales and Platforms

As there are 31 total platforms in this dataset, I decide to only focus on the Top 5 platforms. This filter is based on their charted frequency as the dataset is a combination of top 100 video games in sales each year (from 1980-2017), and so the more a video game is charted the more likely it is to have high sales. Anyhow, the top 5 platforms are DS, PS2, PS3, Wii, and X360.

a. Graphical and Descriptive Statistics

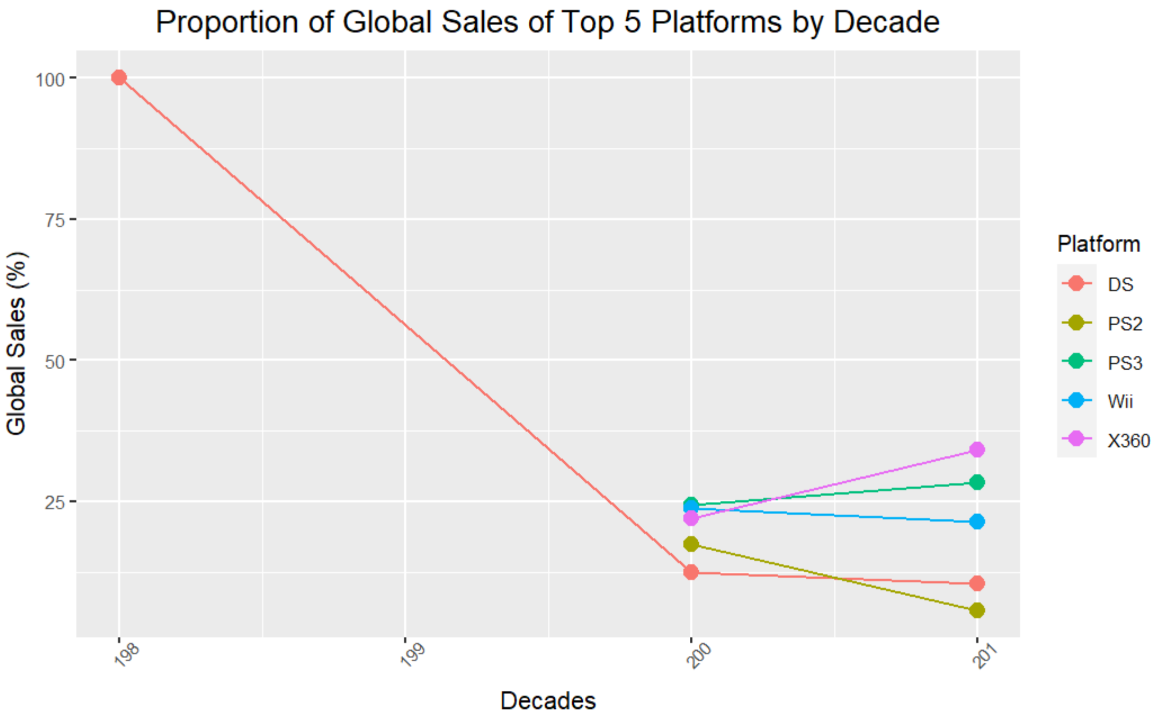
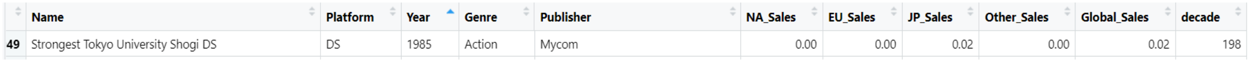


Figure 6. Trend of Top 5 Platforms" Global Sales from 1980s-2010s.

It is worth noting that the DS was in fact released in 2004, and so I have to re-examine that dataset and find out that there is one video game named Strongest Tokyo University Shogi marked as DS (which is correct) and release year of 1985 (which is incorrect since it is actually [released in 2007](#)). Hence, this wrong information might have misrepresented the trend here a little bit. But ignoring that, we could still see that the top 5 platforms have quite similar market share to each other, and all of them were released quite recently (around the 2000s).



	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	decade
49	Strongest Tokyo University Shogi DS	DS	1985	Action	Mycom	0.00	0.00	0.02	0.00	0.02	198

Figure 7. Capture of the incorrectly recorded data point.

Again, to understand the trends in each region, I also look at the sales percentage of these platforms in NA, EU, JP and Other specifically.

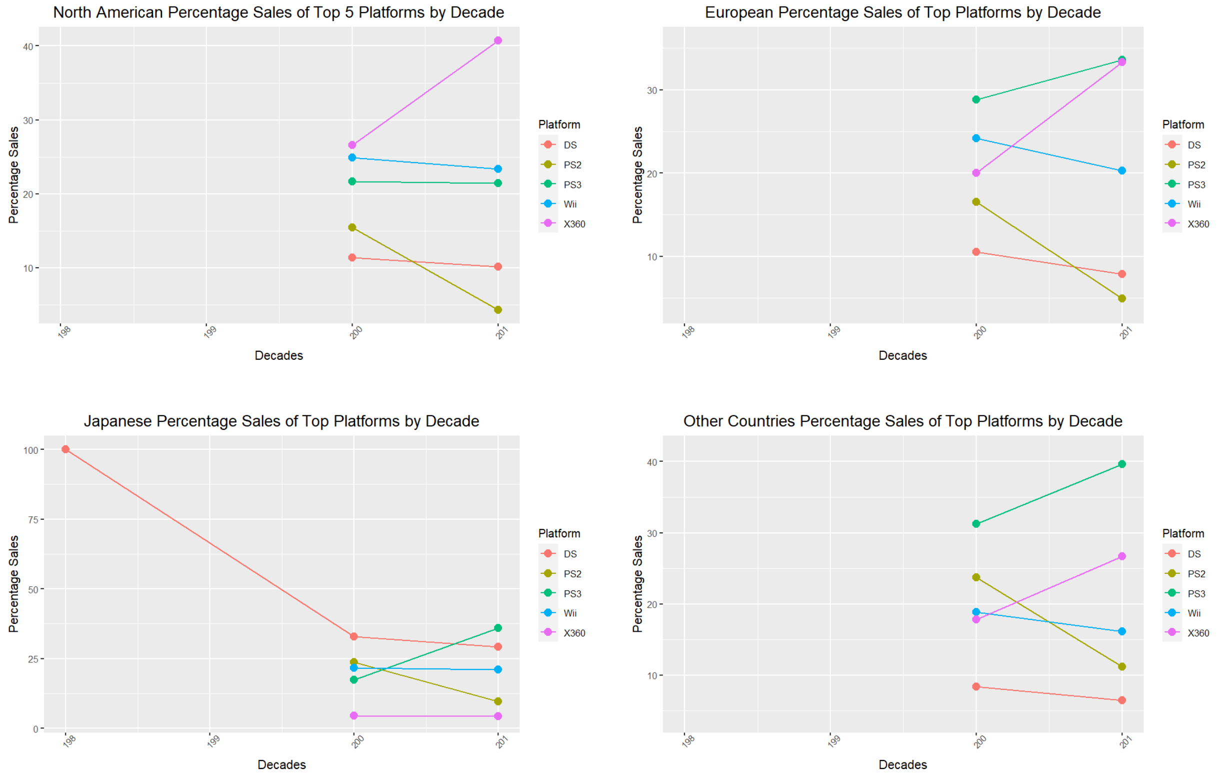
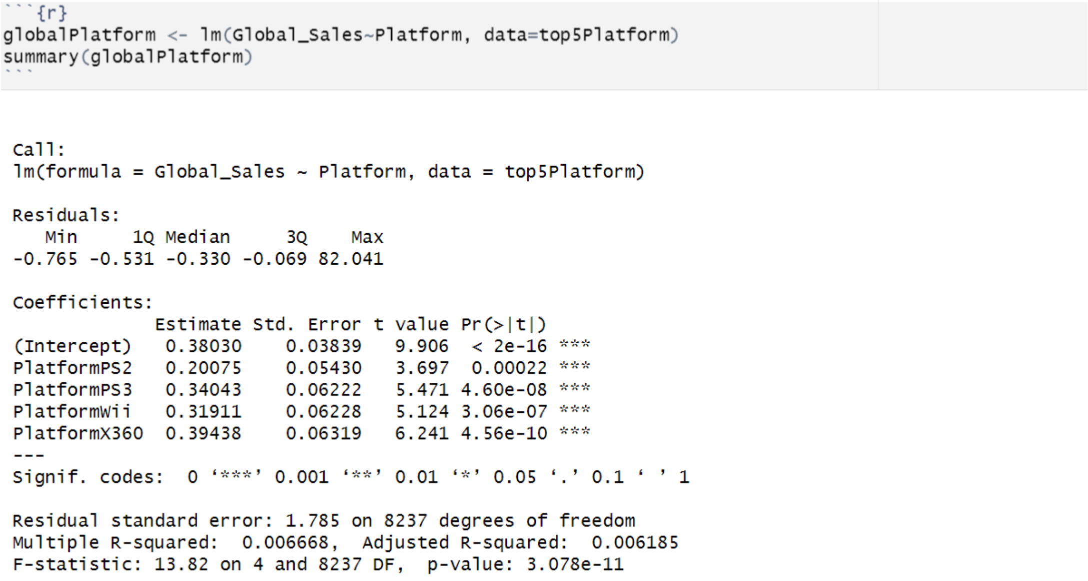


Figure 8. Sales proportion changes in NA, EU, JP and Other from 1980s-2010s (by Platforms).

Each region seems to have a different preference for platforms, but overall it seems that PS3 and X360 account for large share in North America, Europe, and Other, while PS2 share generally declines across these four continents.

b. Multivariate Linear Regression Model

To examine whether there is a relationship between platforms and global sales, or if platforms could dictate or predict global sales, I decide to run a multivariate linear regression model. It is worth noting that since *Platform* is a categorical variable, our prediction is just the mean value for the category that gets “absorbed” in the model, or in this case DS.



~~~
```{r}
globalPlatform <- lm(Global_Sales~Platform, data=top5Platform)
summary(globalPlatform)
```
~~~

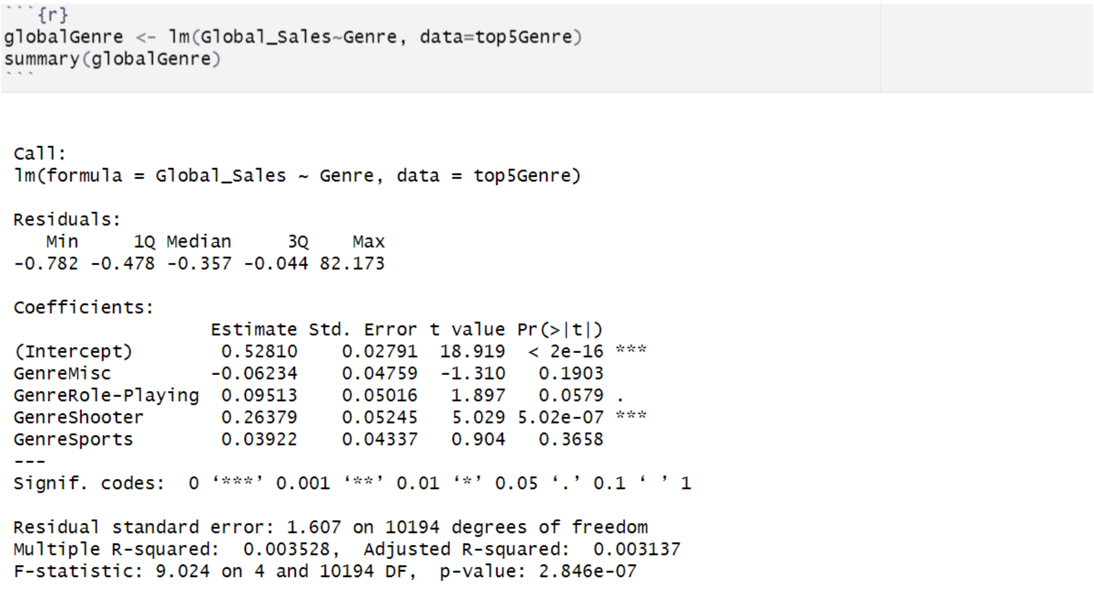
```
Call:
lm(formula = Global_Sales ~ Platform, data = top5Platform)

Residuals:
   Min     1Q Median     3Q    Max 
-0.765 -0.531 -0.330 -0.069 82.041 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.38030    0.03839   9.906  < 2e-16 ***
PlatformPS2   0.20075    0.05430   3.697  0.00022 ***
PlatformPS3   0.34043    0.06222   5.471 4.60e-08 ***
PlatformWii   0.31911    0.06228   5.124 3.06e-07 ***
PlatformX360  0.39438    0.06319   6.241 4.56e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.785 on 8237 degrees of freedom
Multiple R-squared:  0.006668,	Adjusted R-squared:  0.006185 
F-statistic: 13.82 on 4 and 8237 DF,  p-value: 3.078e-11
```

Figure 9. Multivariate Linear Regression Model for Global Sales and Top 5 Platforms.

The model indicates that the PS2 platform will increase global sales by 0.2 millions on average compared to if that video game is DS Platform, or PS3 will have increased global sales by 0.34 millions on average compared to sales of DS (and so on). The p-value is very small, which means that our model is statistically significant. However, adjusted R-squared is extremely small (0.62%).

A low adjusted R-squared value suggests that our independent variables are not explaining much in the variation of our dependent variable, regardless of the variable's significance. It shows that the identified independent variables, even though significant, are not accounting for much of the mean of the DS sales, let alone the global sales. There are several reasons why it happened, one of which might be our focus on just 5 categories across 31 different platforms in the original dataset, or that because we just choose only 1 variable among others (like Genres, Publishers). During the process of omitting the other variables, we might remove some categories/ variables that are closely related to our dependent variables.

All in all, the conclusion is that we can not predict global sales based solely on just one variable like Platforms, and there are many other factors that influence the global sales of video games. Moreover, even multivariate linear regression might not be the most suitable model to examine relationships of any variables to video games sales, because there are so many variables and categories to account for and each only has a small impact on the sales. And these conclusions are further solidified when I attempt to run another multivariate linear regression model for Global Sales and Top 5 Genres alone - which also shows very small adjusted R-Squared (0.31%) and many coefficients are not even statistically significant enough (p-value > 0.05).



~~~
```{r}
globalGenre <- lm(Global_Sales~Genre, data=top5Genre)
summary(globalGenre)
```
~~~

```
Call:
lm(formula = Global_Sales ~ Genre, data = top5Genre)

Residuals:
   Min     1Q Median     3Q    Max 
-0.782 -0.478 -0.357 -0.044 82.173 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        0.52810    0.02791  18.919  < 2e-16 ***
GenreMisc         -0.06234    0.04759  -1.310   0.1903    
GenreRole-Playing  0.09513    0.05016   1.897   0.0579 .  
GenreShooter       0.26379    0.05245   5.029 5.02e-07 ***
GenreSports        0.03922    0.04337   0.904   0.3658    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.607 on 10194 degrees of freedom
Multiple R-squared:  0.003528,	Adjusted R-squared:  0.003137 
F-statistic: 9.024 on 4 and 10194 DF,  p-value: 2.846e-07
```

Figure 10. Multivariate Linear Regression Model for Global Sales and Top 5 Genres.

3.3. Video Games Sales and Genres

Now we can not predict the sales of video games, I decide to look at their trends and make conclusions based on them alone. Similar to *Platform*, I filter out the top 5 genres to examine their trends, including Action, Sports, Misc, Role-Playing, Shooter.

a. Graphical and Descriptive Statistics

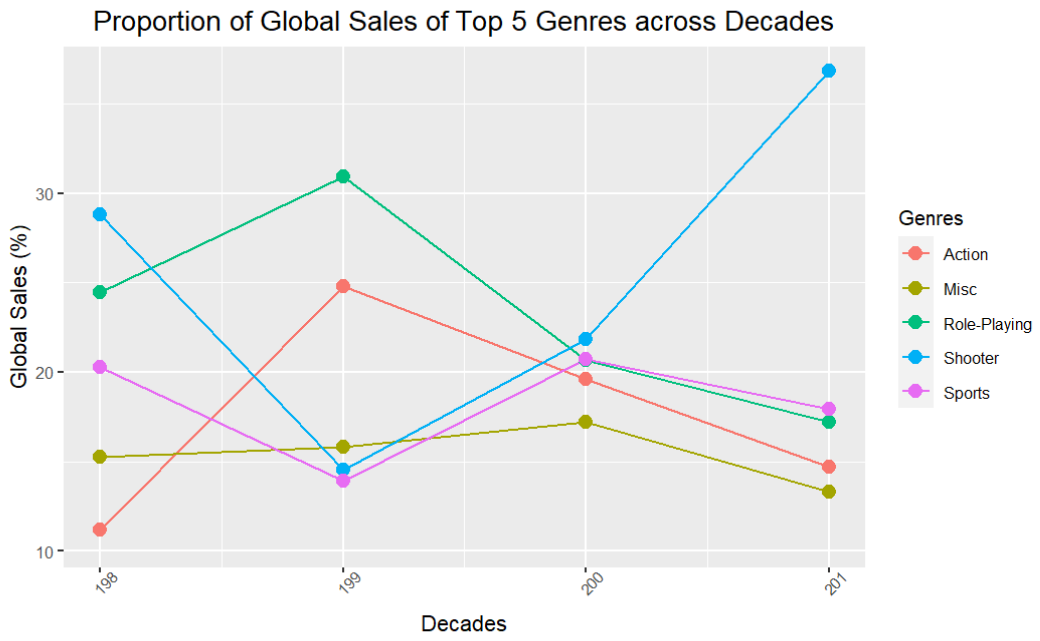
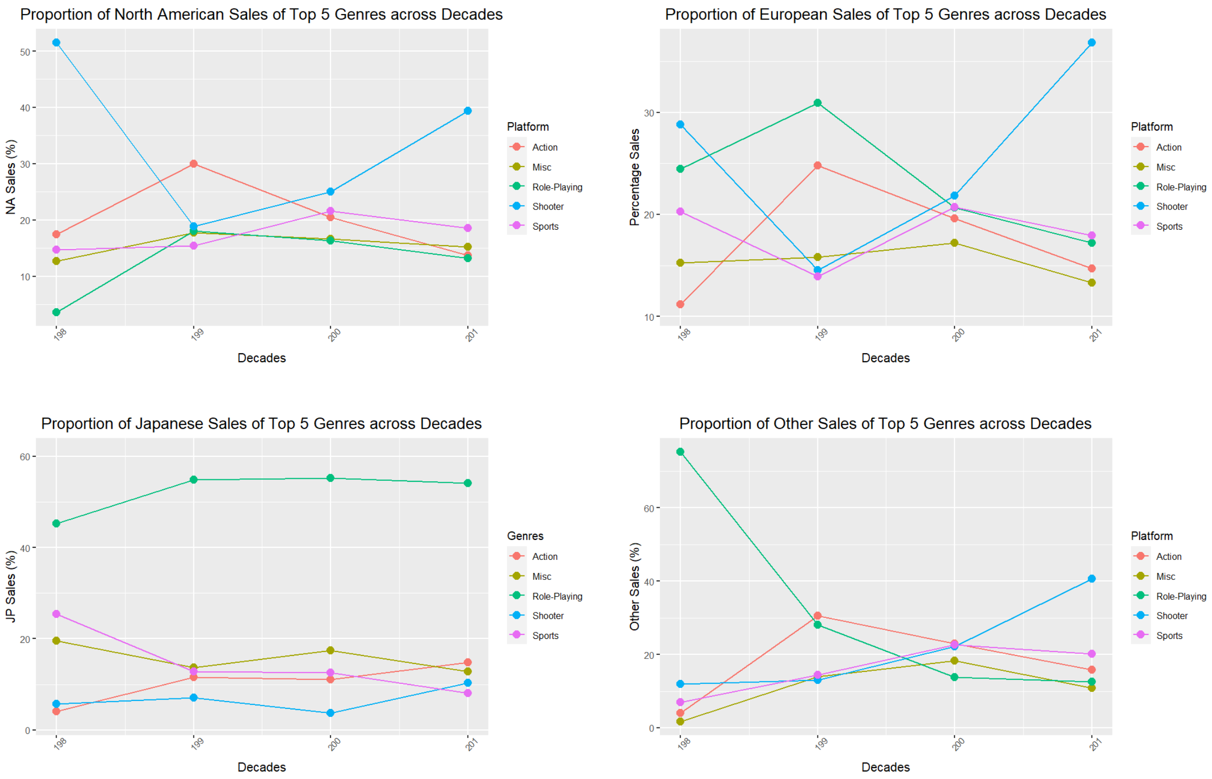


Figure 11. Trend of Top 5 Genres' Global Sales from 1980s-2010s.

The graph suggests that Shooter video games sales are rising lately, while the rest are shrinking down. Look at specific region:


Proportion of North American Sales of Top 5 Genres across Decades

Proportion of European Sales of Top 5 Genres across Decades

Proportion of Japanese Sales of Top 5 Genres across Decades

Proportion of Other Sales of Top 5 Genres across Decades

Figure 12. Sales proportion changes in NA, EU, JP and Other from 1980s-2010s (by Genres).

North America, Europe and even Other sales share a similar trend, while Japan is quite different. In Japan, Role-Playing games are the highest-selling ones, and still remain popular throughout the years.

b. ANOVA

To find out which genres might be driving sales, I decide to first illustrate a box plot to compare global sales between them.

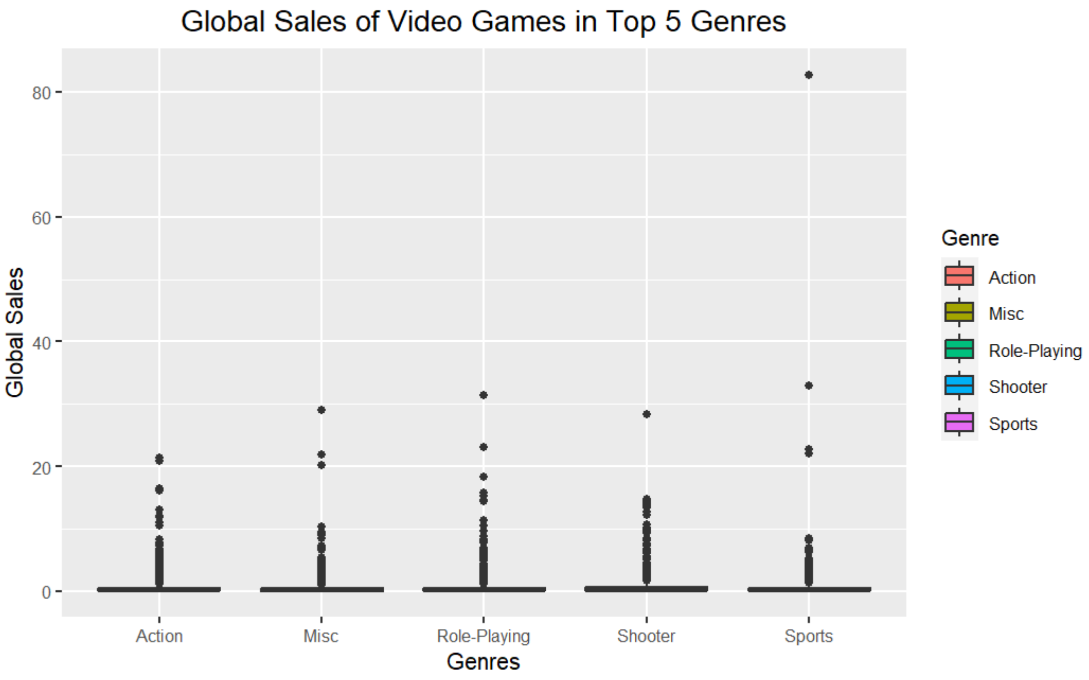


Figure 12. Graph of Video Game Global Sales by Genres.

It is perhaps quite impossible to deduct anything from this box plot, and a quick summary statistics on mean sales of each genre has been done and arranged in descending order.

Genres	Mean Global Sales (millions)
Shooter	0.79
Role-Playing	0.62
Sports	0.56
Action	0.52
Misc	0.46

Figure 13. Table of Global Sales on Average by Genres.

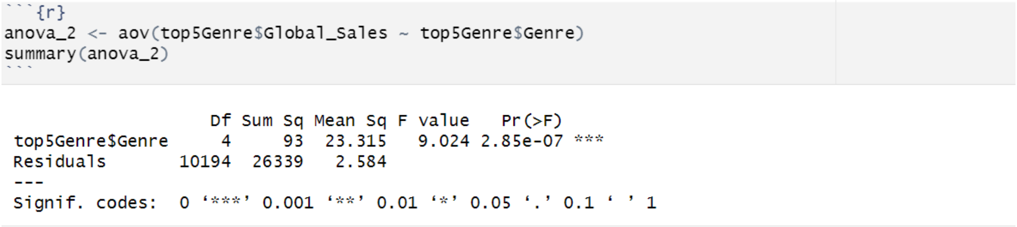
Shooter game is leading the chart - which seems to align with what the trend in Figure 12 suggests. However, these means are quite close to each other and so it is better to check if their differences are statistically significant enough by conducting an ANOVA test.

i. Hypotheses

H_0: all means are equal

H_a: at least one differs

ii. Test statistics



~~~
```{r}
anova_2 <- aov(top5Genre$Global_Sales ~ top5Genre$Genre)
summary(anova_2)
```
~~~

```
                   Df Sum Sq Mean Sq F value   Pr(>F)    
top5Genre$Genre     4     93  23.315   9.024 2.85e-07 ***
Residuals       10194  26339   2.584                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 14. Table of Global Sales on Average by Genres.

iii. Conclusion

Since the p-value is less than 0.05, we reject the null hypothesis. There is strong evidence that the mean global sales of the 5 platforms (Shooter, Role-Playing, Sports, Action and Mics) are statistically different. Combining the evidence from Figure 12 and Figure 14, it could be likely that Shooter games have the highest sales among all.

3.4. Video Games Sales and Publishers

Last but not least, I look up into the top 5 publishers of video games from 1980-2017, namely Activision, Electronic Arts, Konami, Namco and Ubisoft.

a. Graphical and Descriptive Statistics

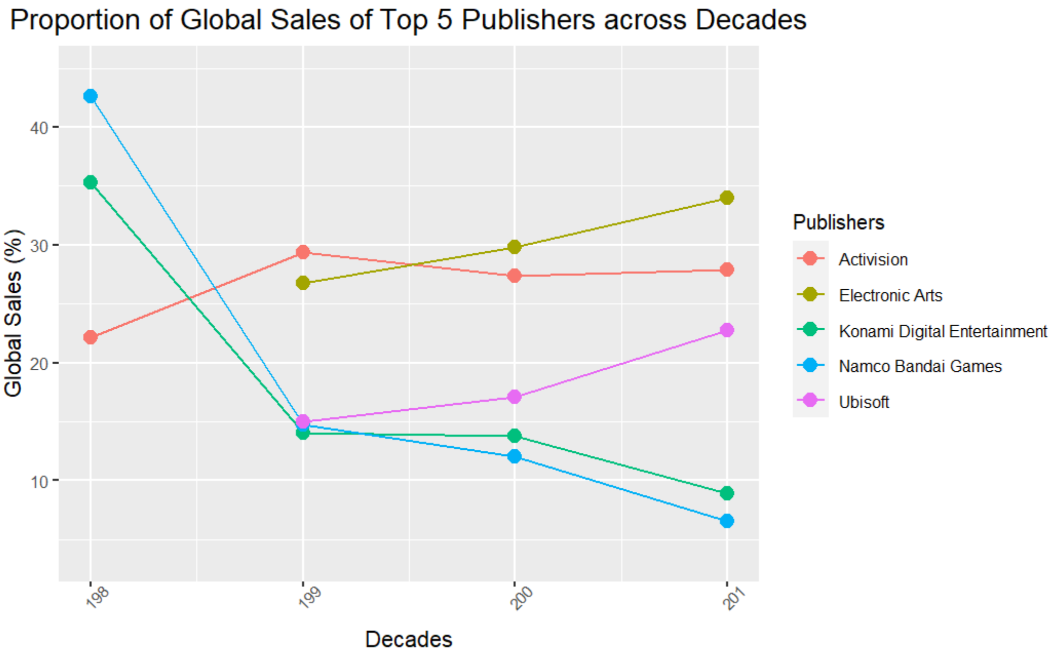


Figure 15. Trend of Top 5 Publishers’ Global Sales from 1980s-2010s.

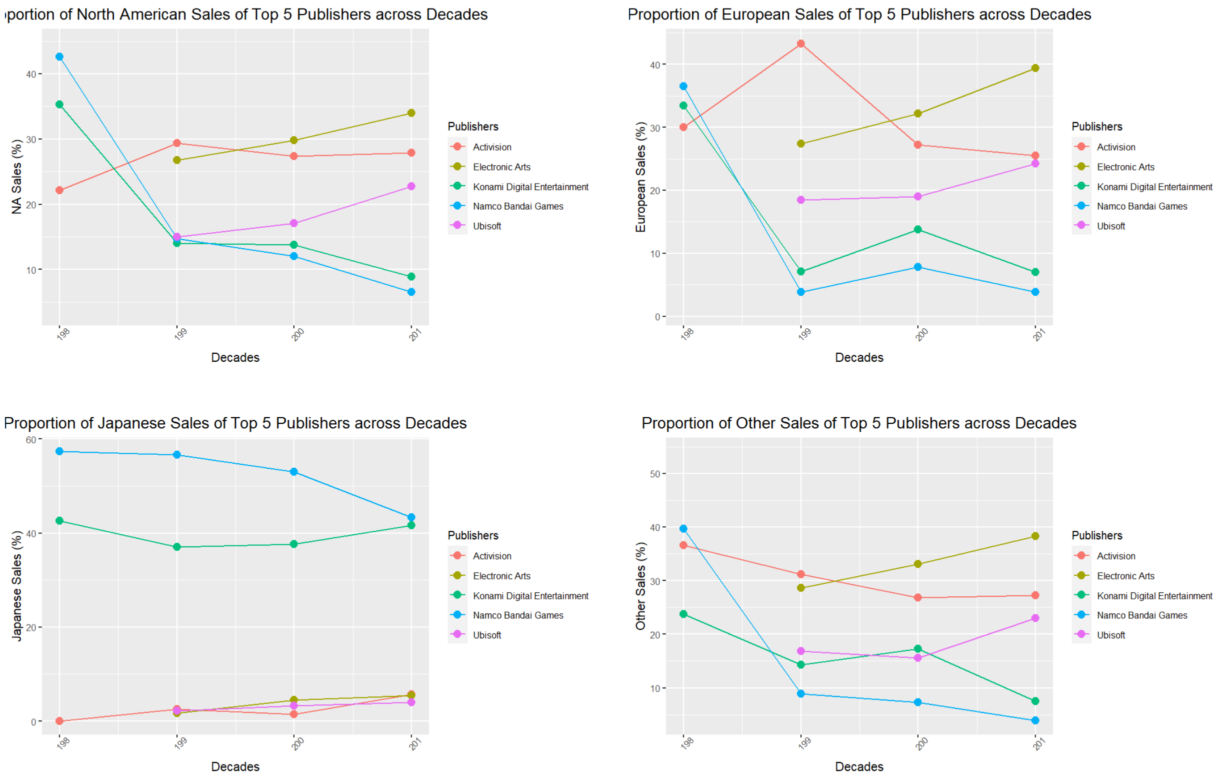


Figure 16. Sales proportion changes in NA, EU, JP and Other from 1980s-2010s (by Publishers).

Overall, Activision and Electronic Arts have the highest market share, especially Electronic Arts - which appears later than the other four companies yet has increased quite consistently in terms of sales proportion. It is also worth pointing out that Japan also shares a totally different trend from every other continent, as it is dominated by either Konami and Namco (which are Japanese video game companies). On the other hand, Ubisoft share of the market stays pretty stable over the past 40 years.

b. Chi-Square Test

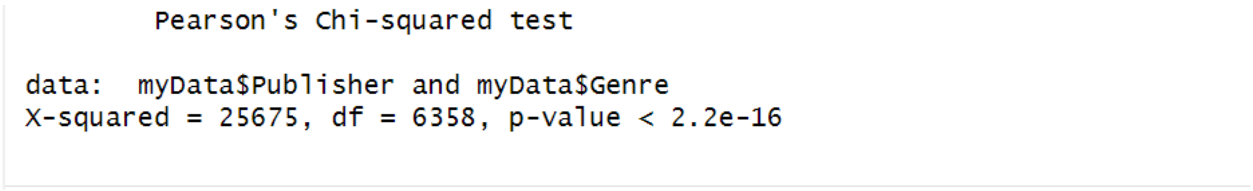
Since the mean sales of each genre differs and some are more popular than others, I want to check if there is a correlation between publishers and the kind of video games they sell by using a Chi-square test with two categorical variables.

i. Hypotheses

H_0: There is no relationship between publishers and video game genres

H_a: There is relationship between publishers and video game genres

ii. Test statistics



```
        Pearson's Chi-squared test

data:  myData$Publisher and myData$Genre
X-squared = 25675, df = 6358, p-value < 2.2e-16
```

Figure 17. Chi-squared Test Statistics (between Publishers and Genres).

iii. Conclusion

Since p-value is very small, we can reject the null hypothesis. In other words, there is a relationship between publishers and the genres of video games.

Further evidence about this relationship could be found in the plot below. Looking at the top 5 publishers, we could see that they tend to produce certain genres more to gain more profits, like Action (which is also present in the top genres).

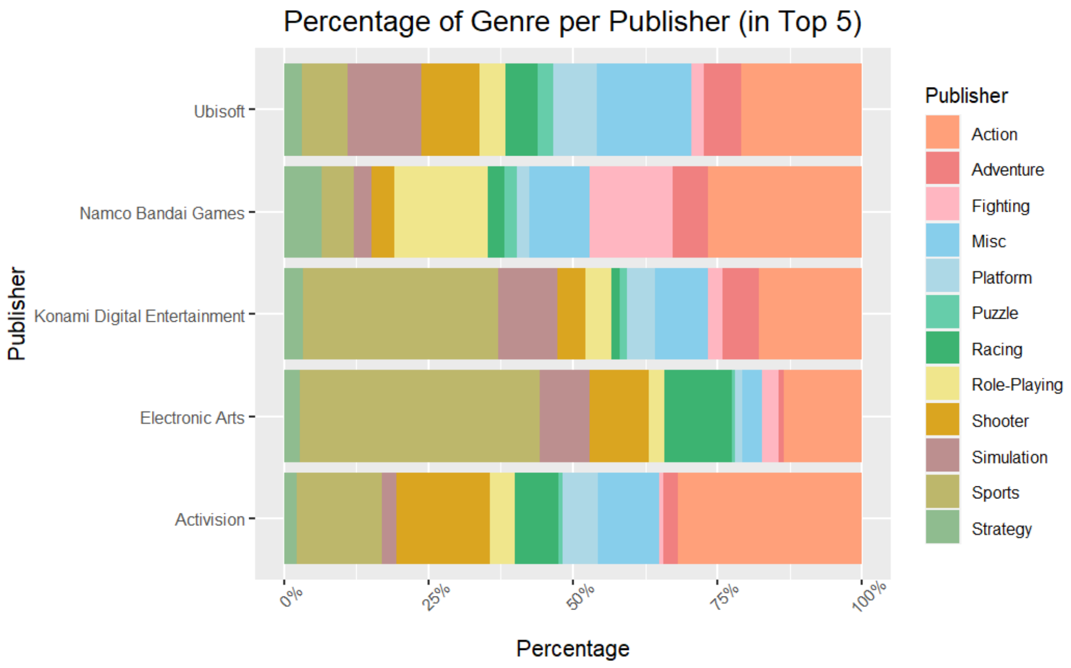


Figure 18. Proportion of Genres from each Top 5 Publishers.

4. Results and Conclusion

- Overall, the leading video game genre is Shooter while sales of PS3, Wii and X360 account for the majority. In terms of publisher, Electronic Arts seems to be growing and prevailing more in the industry. Publishers and Genres are also related to each other, which means that some Genres are more likely to make high sales and be invested by publishers than others.

- Japan, however, often shares a different trend compared to the rest of the world, with Role-Playing games sales remaining consistently high and with Konami and Namco dominating the market share.
- Sales can not be predicted by platforms nor genres alone. Thus, while there could be some relationship between them, there are many factors that contribute to the sales of video games.
- While North America still accounts for a large portion of global sales, it has a downward trend currently while Europe, Japan, and Other regions show more positive trends. And so, producers might want to shift their focus and investment to these places more. More specifically, I suggest looking at Asia Pacific region since it has been reported to growth considerably in the past few years, especially with emergence of China as a major gaming hub (Grand View Research)

5. General Discussion and Acknowledgements

- There are some incorrect data points and also 200+ missing data information (especially in terms of *Year*), which could affect the data quality and trends analysis.
- Data is before COVID-19, and so many trends and new growth could have emerged since COVID-19 has led to many changes in users behaviors. A few changes could include increased demand as people spent more time indoors due to lockdowns and social distancing measures, or the popularity of online gaming because of social distancing measures.
- "Other" region is unclear, and because the data is from a Western source/ chart, it tends to be more biased towards NA and EU games.

6. References

Devore, J. L. (2015*). Probability and Statistics for Engineering and the Sciences.* Cengage Learning.

Global Video Game Sales. Kaggle. (2023, February 6). Retrieved May 5, 2023, from https://www.kaggle.com/datasets/thedevastator/global-video-game-sales

Grand View Research. (n.d.). *Video game market size & share growth report*. Video Game Market Size & Share Growth Report. Retrieved May 5, 2023, from https://www.grandviewresearch.com/industry-analysis/video-game-market

Kirkcaldy, A. (2023, January 12). *Video game industry statistics, trends and data in 2023.* WePC. Retrieved May 5, 2023, from https://www.wepc.com/news/video-game-statistics/