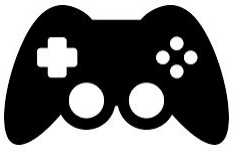


MSBA CONSULTING

VISUAL ANALYSES AND RECOMMENDATIONS FOR VIDEO GAMES DEVELOPERS

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY
ISOM 5620 Visual Analytics for Business Decisions
Group 11 Final Project

Pan Huiyi, 20660002
Liu Xu, 20632237
Yuke Xie, 20667878
Dian Yu, 20662165
Tang Viola Jing Xu, 20579899
Zhong Weicheng, 20673736

TABLE OF CONTENTS

Introduction

In 2018, the global gaming market is estimated to have generated US$138.7 billion in revenue, and is predicted to grow by 9.6% to US$152.1 billion in 2019[1]. Sales performance is similar to the movie and other entertainment industries where games are characterised as “hit” or “miss” with potential exorbitant sales for high performing “hits”. Yet there are over 2,000 game developers in the United States alone, yielding a highly competitive market.

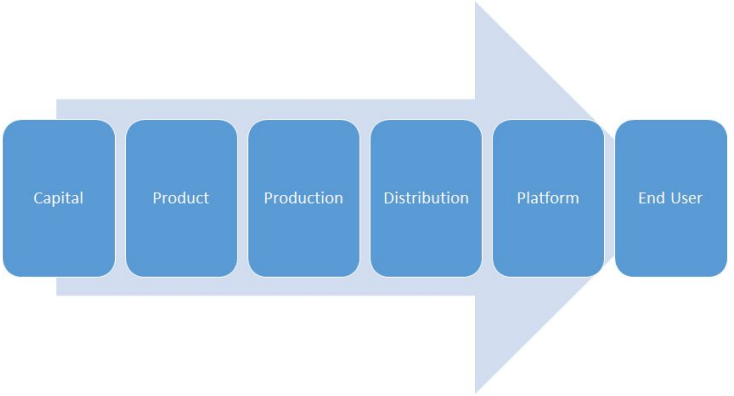
To support developers of new video games, we conduct a thorough analysis of the console and PC video games market to enable a better understanding of the industry and determinants of global sales and provide recommendations to increase the likelihood of developing and publishing a high performing video game. This will help future game developers to better meet their business objectives of achieving product/market fit, determining the best marketing strategy and optimising sales.

In this report, we will provide a brief overview of the industry, our data sources and treatment methods, then conduct detailed analyses on the main variables and key aspects of our dataset. We introduce a model for preliminary exploration of the predictive power of big data in the video games industry. Finally we conclude with a development proposal for developers, a summary of our findings and suggest topics for further study.

Industry Overview

The history of video games can be traced back to the 1970s when the first arcade game, Computer Space, was released in 1971. One year later, the first commercially successful video game, Pong, was released and sold over 19,000 machines. In the same year, Magnavox released the first video game console to the home market. Since then, there have been waves of innovation within the industry that have propelled the development of video games as well as the development of modern computing. There are eight generations of video games, primarily defined by the consoles or platforms by which games are played. The eighth generation began in 2012 with the launch of Wii U by Nintendo.

The video game industry value chain consists of six main layers: capital, product, production, distribution, platform and end users (gamers). Capital provides financing for the development of games, which is often conducted by publishers. Product includes developers, designers and artists who build the game. Production consists of software and tools used to generate content and support the development process. Distribution entails the marketing and selling of the game, which is conducted by publishers. Platform consists of the hardware, software and network infrastructure required to play the game[2].



[1] https://newzoo.com/insights/articles/the-global-games-market-will-generate-152-1-billion-in-2019-as-the-u-s-overtakes-china-as-the-biggest-market/

[2] Flew, Terry; Humphreys, Sal (2005). "Games: Technology, Industry, Culture". New Media: an Introduction (Second Edition). Oxford University Press. pp. 101–114. ISBN 0-19-555149-4

Data Source and Data Treatment

Data Source

The two main sources for our data are VGChartz and Metacritic. Video game sales volume data (global and regional) is obtained from VGChartz via their proprietary methodology, which includes surveys to video games users on their purchases, polling of sellers on their sales and consultations with publishers and manufacturers on distribution. This polling data is then overlaid with an analytical layer (including statistical trend fitting of historical and comparative data and analysis of resell prices to obtain demand and inventory levels) to determine the final sales values. Other data (platform, publisher, year of release, genre, critic score, critic review count, user score and user review count) are obtained from Metacritic and sourced from Rush Kirubi's[3] scrape of the Metacritic website. Critic and user scores and review counts are based on voluntary submission of feedback to the Metacritic site by individuals.

The raw dataset of video games sales contains 15 variables: Name, Platform, Publisher, Year_of_Release, Genre, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, Critic_Score, Critic_Count, User_Score, User_Count and Rating. There are a total of 16,719 observations primarily from 1980 to 2016. A description of each variable before treatment, as well as of additional variables that we introduce are provided below.

Variable	Type	Description
Name	Categorical	Name of the video game
Platform	Categorical	The hardware and software on which the game is launched and played. There are 17 unique categories
Publisher	Categorical	The party that is responsible for the manufacturing, financing, marketing and distribution of the game. 583 unique categories are involved
Year_of_Release	Numerical - Ordinal (Discrete)	Year of release of the game, ranging from 1980 to 2016
Genre	Categorical	Genre of the game, including 12 types: Action, Adventure, Fighting, Misc, Platform, Puzzle, Racing, Role-Playing, Shooter, Simulation, Sports and Strategy
NAS	Numerical - Count (Discrete)	Cumulative sales in North America (in million units). Previously named NA_Sales
EUS	Numerical - Count (Discrete)	Cumulative sales in Europe (in million units). Previously named EU_Sales
JPS	Numerical - Count (Discrete)	Cumulative sales in Japan (in million units). Previously named JP_Sales
OtherS	Numerical - Count (Discrete)	Cumulative sales in Other regions (in million units). Previously named Other_Sales
GlobalS	Numerical - Count (Discrete)	Cumulative sales worldwide (in million units). Previously named Global_Sales

[3] https://www.kaggle.com/rush4ratio

Critic_Score	Numerical - Ordinal (Discrete)	Score compiled by Metacritic based on scores submitted by critics
Critic_Count	Numerical - Count (Discrete)	The number of critics that submitted scores
User_Score	Numerical - Ordinal (Discrete)	Score compiled by Metacritic based on scores submitted by users
User_Count	Numerical - Count (Discrete)	The number of users that submitted scores
Rating	Categorical	The rating given to each game. There are 5 unique categories: E (everyone), E10+ (everyone over the age of 10), T (teenager), M (mature), AO (adults only). Prior to 1998, the E rating was labeled K-A (Kids to Adults)
Age	Numerical - Ordinal (Discrete)	The age of each game (i.e. the number of years since its release as of 2019)
Is_popular	Categorical - Binary	If the publisher is on the top publisher ranking from year 2010-2016 as provided by Metacritic. '1' if Yes, '0' if No
newPlatform	Categorical	Redefined parent platform name. A value from the following 5 options: Sony Playstation, Microsoft Xbox, Nintendo, Sega, PC
Is_highsale	Categorical - Binary	If the global sales of the game is larger than 0.75 million units. '1' if Yes, '0' if No

Data Treatment

For the purpose of conducting further analysis, the following data cleaning steps are taken:

- Excluded all entries with Null values
- Limited the observation period to between 1998 and 2016 in order to improve the relevance and reliability of our analysis to the current times. We set the start of the period to be the start of the sixth generation of video games, which was in 1998
- Revised a few Chinese characters that appeared in game names to avoid error reporting
- Modified the K-A rating to E in order to maintain consistency. The K-A (Kids to Adults) rating was changed to E (Everyone) in 1998
- Assumed that all games sales occurred in the year of release. Although some video games may continue to generate sales following the year of the release, we are not able to determine the percentage of sales in the year of release and the following years. Moreover, different types of video games have different lifespans. Some games can generate significant sales for many years while other video games may generate most of its sales in the year of release such as FIFA 2019 or 2K19

In addition, four new variables are added for analytical purposes.

- ***newPlatform:*** In order to compare platforms more directly, we regrouped 17 platforms into 5 major categories, based on their parent companies. More specifically, 'PS', 'PS2', 'PS3', 'PS4', 'PSP', and 'PSV' are a series of 'Sony PlayStation'; 'X360', 'XB', 'XOne' are under 'Microsoft Xbox'; '3DS','DS','GBA','GC','Wii','WiiU' belongs to 'Nintendo'; 'DC'· refers to 'Sega'. 'PC' is not changed in this new column.

- ***Age:*** Age represents the number of years since the release of the game. For example, if the game was released in 2006, then its 'Age' is 13 (calculated as the numerical difference between 2019 and 2006). This new numerical variable is used in our analysis to determine its impact on sales.
- ***is_popular:*** As there are 583 different publishers in the dataset, we create a binary variable based on the ranking provided by Matecritic from 2010 to 2016. We create a dataset with the top 8 publishers from each year. If the publisher is on the list, it would take on a '1' value and '0' if not.
- ***is_highsale:*** If global sales is larger than 0.75 million units, the value is '1'. Otherwise, the value is '0'. This binary variable is used in the logistic regression model.

Following the data treatment and addition of variables, we have 6,810 observations remaining with 19 variables. Summary statistics of the numerical variables after treatment are provided below.

Variable	**Mean**	**Standard Deviation**	**Minimum**	**Maximum**
NAS	0.3924055	0.9656525	0	41.36
EUS	0.2346008	0.6856241	0	28.96
JPS	0.06238494	0.2819887	0	6.5
OtherS	0.08247317	0.2699405	0	10.57
GlobalS	0.7720497	1.957049	0.01	82.53
Critic_Score	70.22026	13.85786	13	98
Critic_Count	28.97912	19.22243	3	113
User_Score	7.18215	1.439822	0.5	9.6
User_Count	172.9257	582.9021	4	10665
Age	11.5211	4.154679	3	21

Limitations of the Data

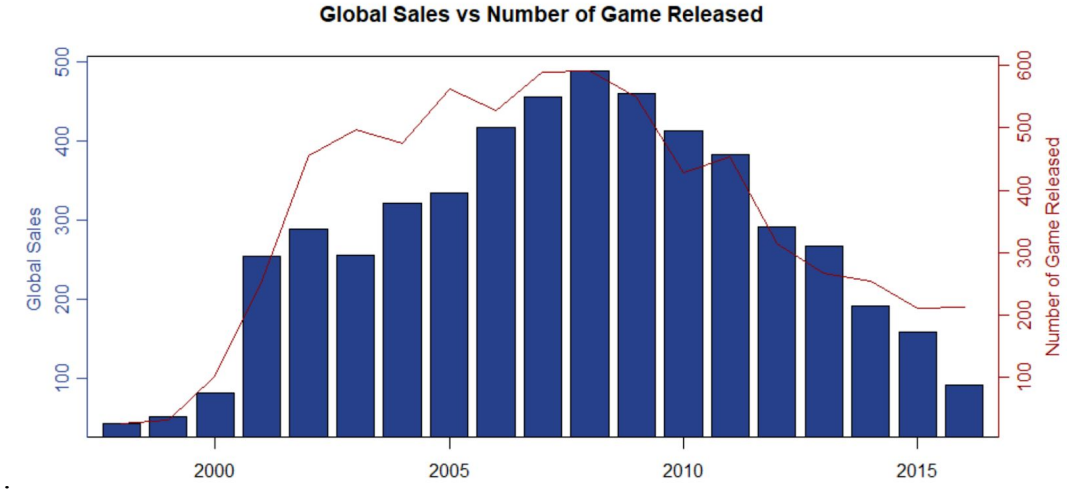
Although this dataset provides an interesting foundation for our study on the industry and yields a number of insights for business strategies, there are many limitations to the data, which include:

- Data is only updated to 2016, thereby limiting insight on a fast-moving industry based on the most recent developments
- Data limited to console and PC games only and excludes mobile games, which have grown significantly
- For analytical purposes, we assume that all games sales occurred in the year of release
- There are likely inaccuracies in the data due to errors or biases in the collection methodology (polling by VGChartz, voluntary submission of scores by critics and users on Metacritic, data entry by Metacritic staff and web scraping of Metacritic by Rush Kirubi). This may have a particularly significant impact across geographies as the systems and stakeholders in different regions may differ substantially
- The dataset from Metacritic is not complete as it only covers a limited number of platforms
- Our data treatment methods, such as removing observations with 'NA' values, may bias the results
- The industry is characterised by the existence of many outliers, which affects our analyses

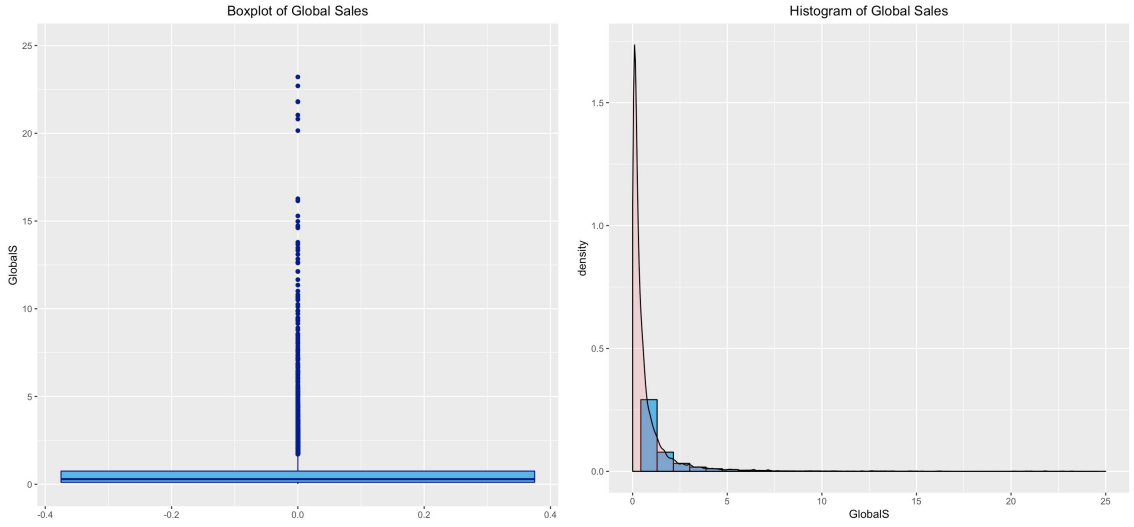
Visual Analysis

Global Sales and Game Releases

Global sales and the number of games released each year display a similar trend over time, thus indicating a positive correlation between the two. Both the number of games released and global sales for games released grew rapidly in 2001. An explanation for this could be the release of Xbox. From 2001, number of games released and sales continued to increase, peaking in 2008 when the 8th generation of consoles were developed, before decreasing year by year afterwards. This gradual decline may be attributed to the impact of the global financial crisis coupled with competition resulting from the rise of smartphone usage and mobile games in recent years

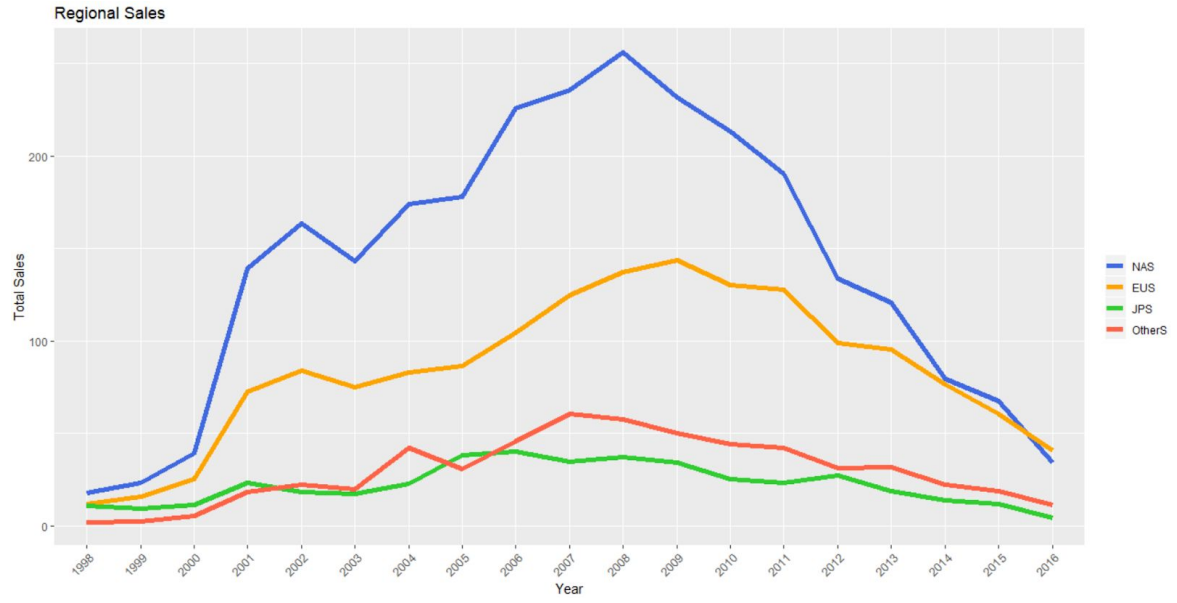


To better understand the nature and distribution of global sales, we examine its boxplot and histogram.



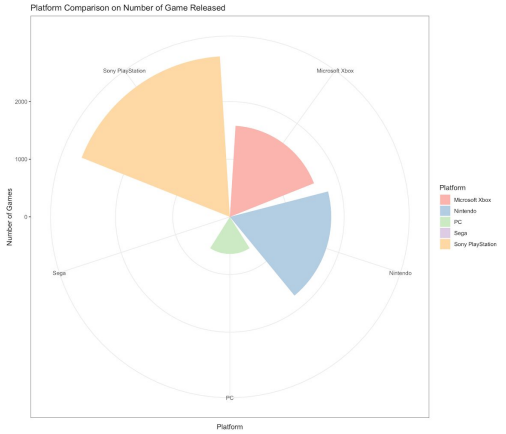
Since there are some extreme values in the dataset such as 82.53 for the global sales of the Will sports game, to better visualize the distribution of the Global Sales (GlobalS) variable, we adjusted the range to 0 to 25, which omits 6 outliers from the boxplot. However, we can see that there is still a large number of outliers in the boxplot. The median of global sales is 0.29 and the third quartile of global sales is 0.75. From the histogram, Global Sales has a very small mean of 0.7721 and a large standard deviation of 1.957049. The distribution is strongly skewed to the right and does not follow the normal distribution. Hence, we created a binary variable to classify global sales into high global sales and low global sales. Games with global sales larger than 0.75 million units is considered as high since it exceeds the 75th percentile of the data.

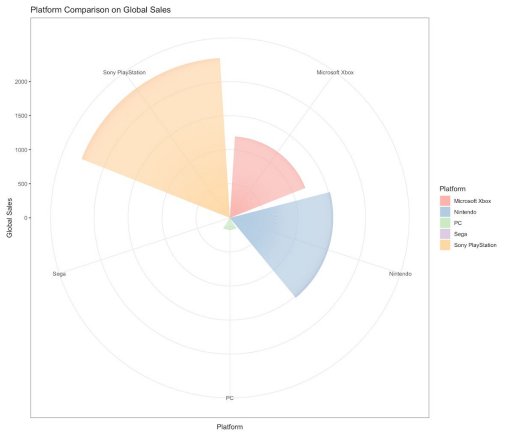
If we examine sales by region, we see that historically North America has been the major video game market, followed by Europe. Japan has made comparatively smaller sales. This regional difference could be due to the nature of the market, where people in North America and Europe simply purchase more video games than people in other regions. Although China as a huge gaming market, it does not contribute much to the sales value of 'Other Regions' as the sale of game consoles was banned until 2015. Sales in North America have been more volatile than in other regions, including a greater percentage increase in 2001 and a greater decline since the great financial crisis. For games released in 2016, Europe registered the highest sales, surpassing North America. Compared to the US, Europe, Japan and Other Regions have shown more stable sales over time. Japan in particular appears to have more resilient demand.



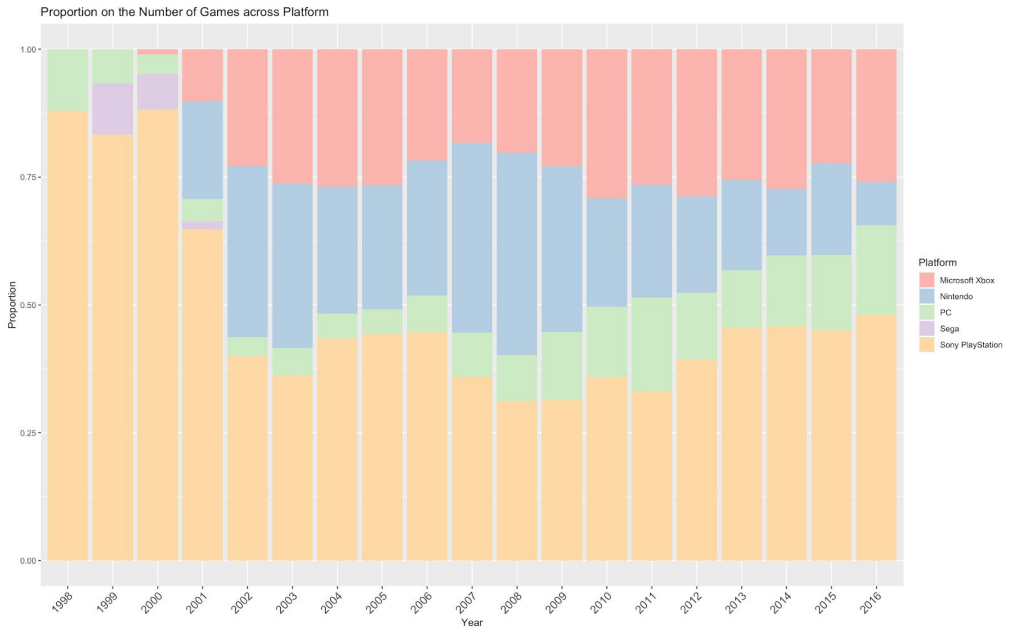
Platform

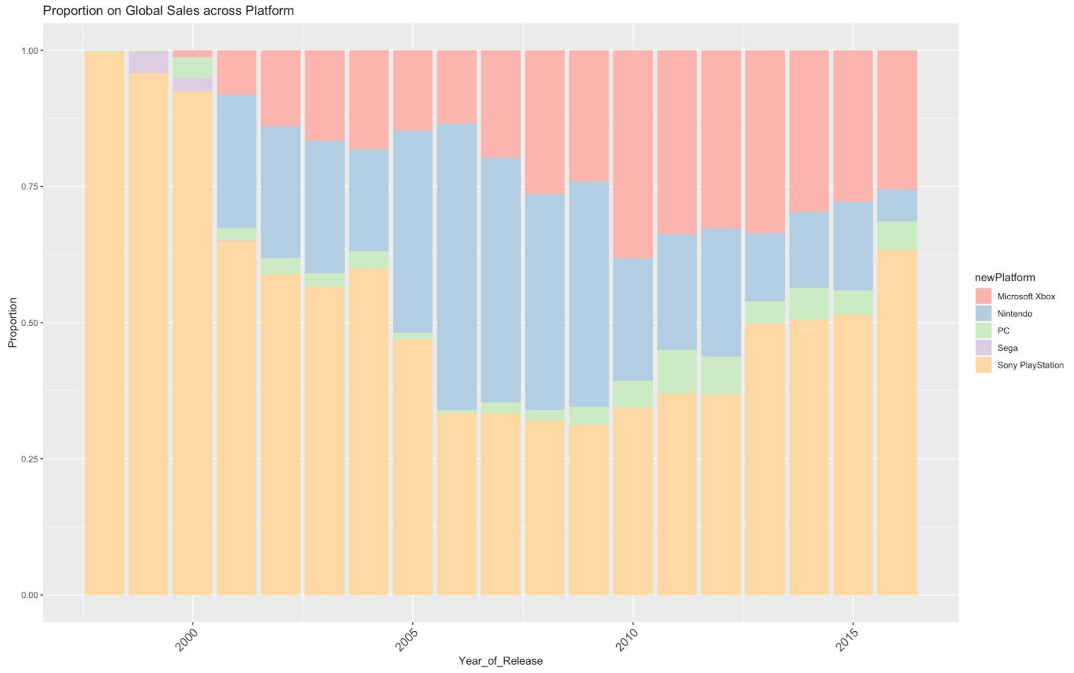
Polar plot is used to present global sales and number of games across the five newly defined platforms. Sony PlayStation published the largest number of video games and generated the highest sales volume, followed by Nintendo, Microsoft Xbox, PC and Sega. Sega only published 11 games and sold 2.51 million units.

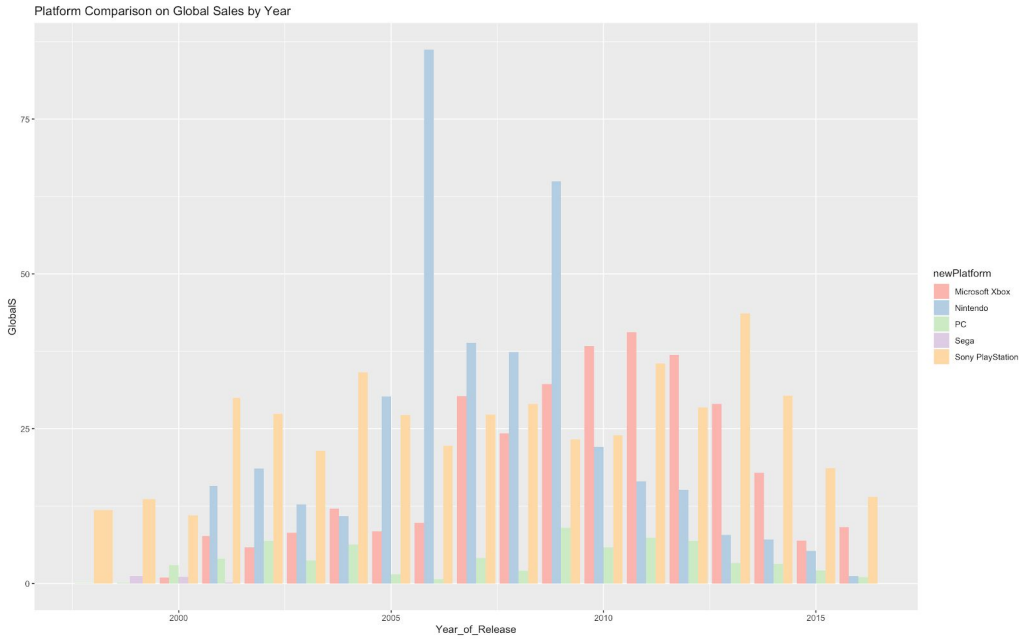




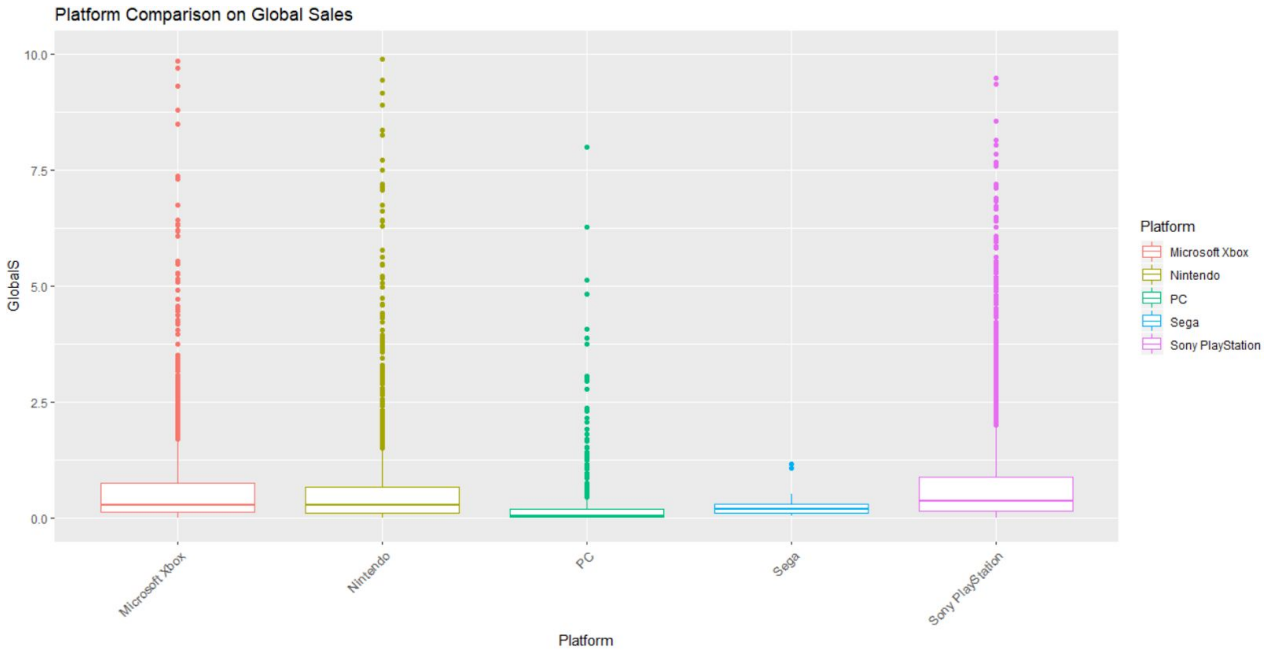
We further our market share analysis using stacked column charts. Sony PlayStation released more than 87.5% of total games in 2000, and generated the highest global sales in the following 6 years. Following the release of Xbox in 2001, Microsoft Xbox quickly increased in market share, ranking in the top 3 by number of games released and global sales. Nintendo saw outstanding sales in 2006 due to the release of Wii Sports. However, we can see that its market share gradually declined afterwards both in terms of number of games released and global sales. Following the turn of the decade, Sega was no longer a competitive platform.


Proportion on the Number of Games across Platform
Proportion
1.00
0.75
0.50
0.25
0.00
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Year
Platform
Microsoft Xbox
Nintendo
PC
Sega
Sony PlayStation


Proportion on Global Sales across Platform
Proportion
1.00
0.75
0.50
0.25
0.00
2000
2005
2010
2015
Year_of_Release
newPlatform
Microsoft Xbox
Nintendo
PC
Sega
Sony PlayStation

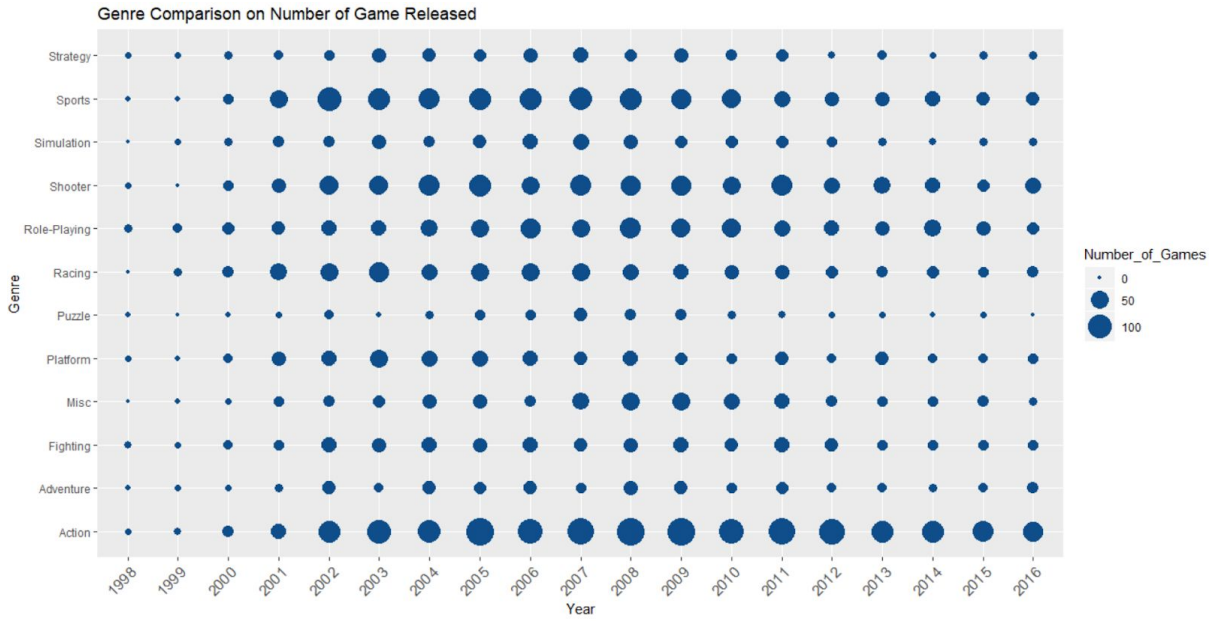

Platform Comparison on Global Sales by Year
GlobalS
75
50
25
0
2000
2005
2010
2015
Year_of_Release
newPlatform
Microsoft Xbox
Nintendo
PC
Sega
Sony PlayStation

We further analyse the relationship between Global Sales and Platform by comparing boxplots of Global Sales by Platform. From the boxplots, we can see that Sony, Microsoft, and Nintendo have dominated the video game market because of their high median global sales. In addition to similar median global sales, these three giants also have similar interquartile ranges. We also notice that the median global sales for these three platforms are all closer to the first quartile than the third quartile and a significant amount of outliers exists for all three, which means that global sales for these platforms were most likely driven by a number of popular and good selling games that were identified as outliers in the boxplots.

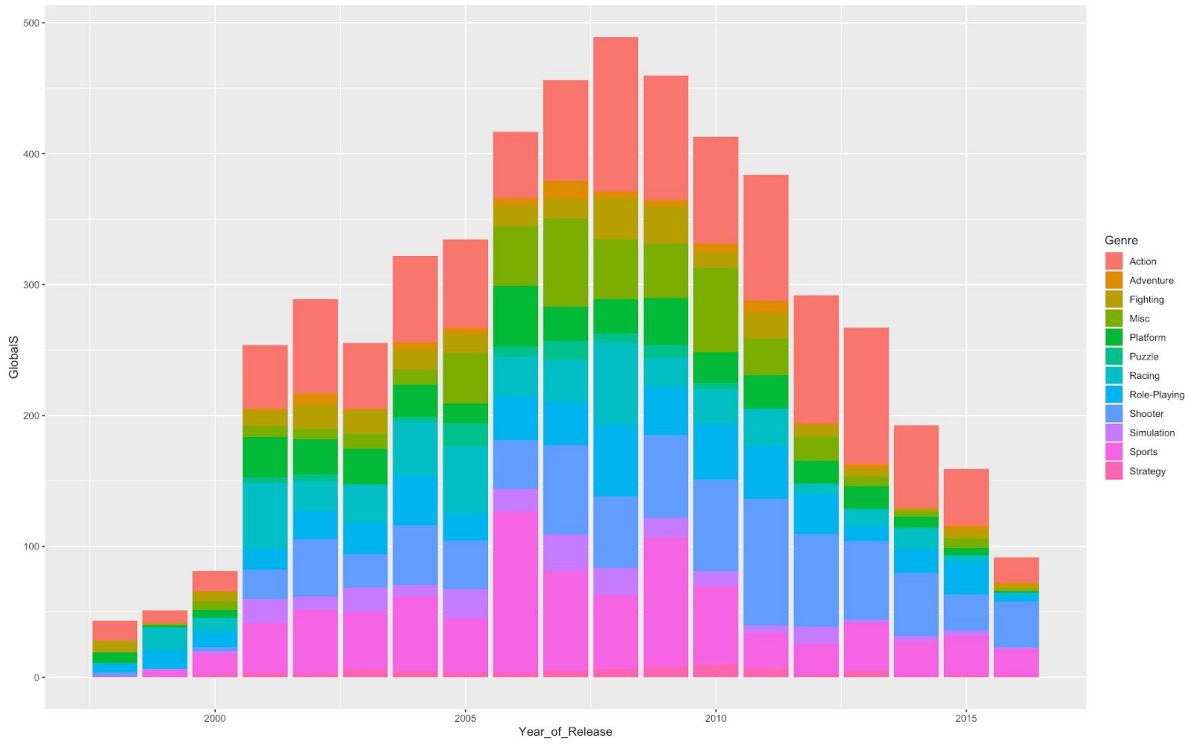


Genre

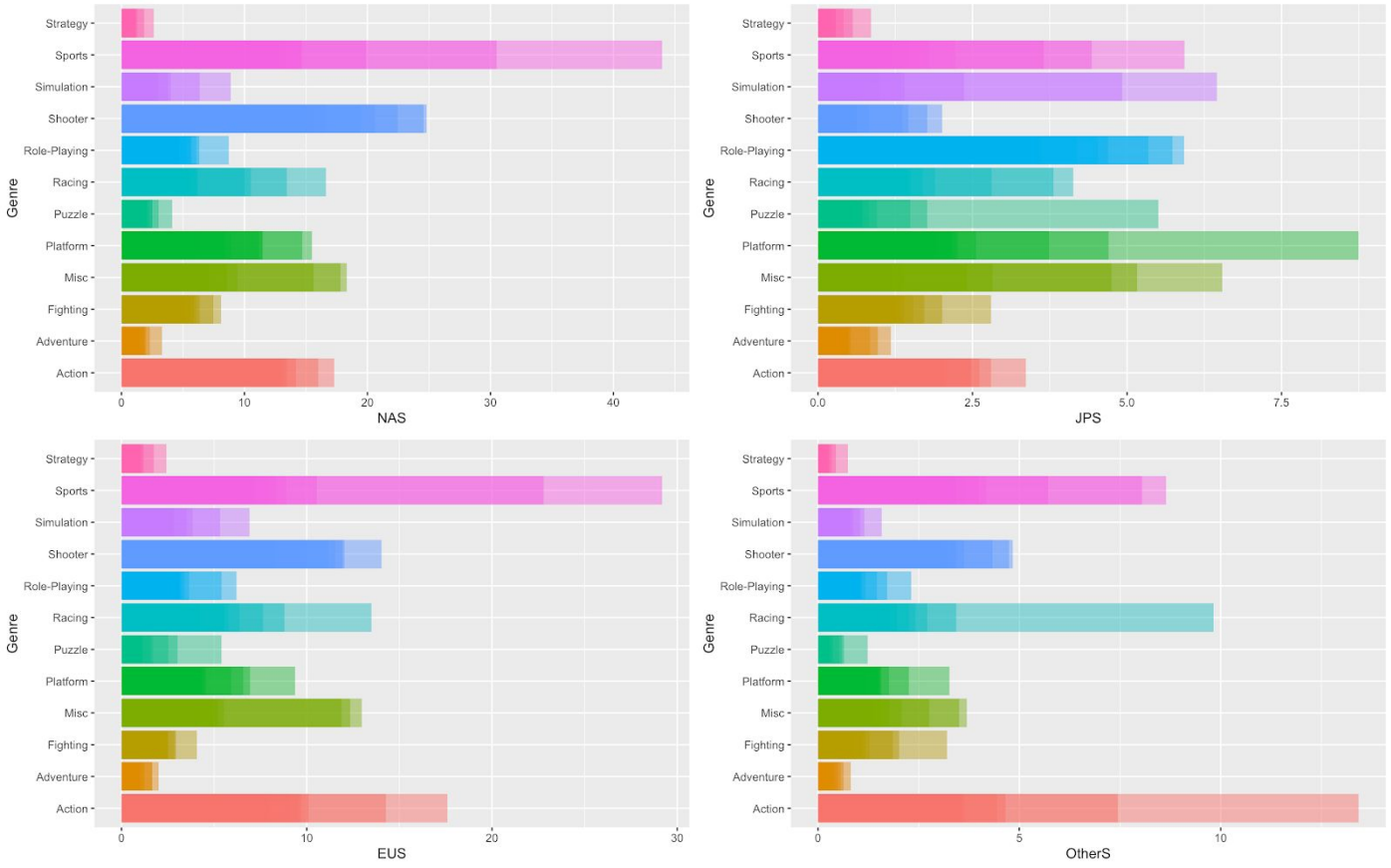
From a bubble plot of genre against the number of games released, we can see that developers and publishers clearly prefer releasing Action games. The number of Action games released over time is significantly higher compared to other genres. Additionally, Sports and Shooter games also consistently have a high number of games released each year.



The below stacked barchart shows global sales by genre. Action and Sports games take up the greatest proportion of global sales. Overall sales for Shooter games is not the highest, but it was very popular in 2011. Demand for Action and Shooter games appear more defensive, with sales generally stable over time.



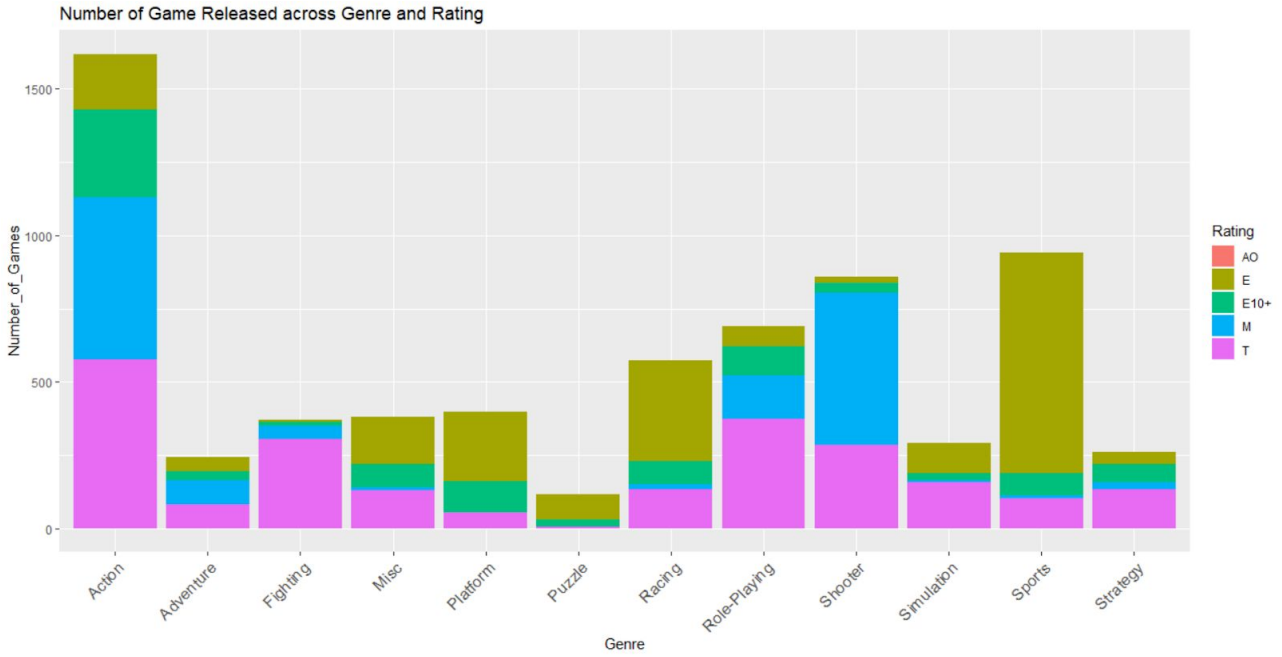
When we investigate sales by genre in each region, we find that Japanese gamers have a different taste in games while the other regions are quite similar to each other in terms of preference. Compared to gamers in other regions, Japanese gamers purchase more Platform, Simulation, Role-Playing and Puzzle games. Outside of Japan, Action and Sports games are popular.



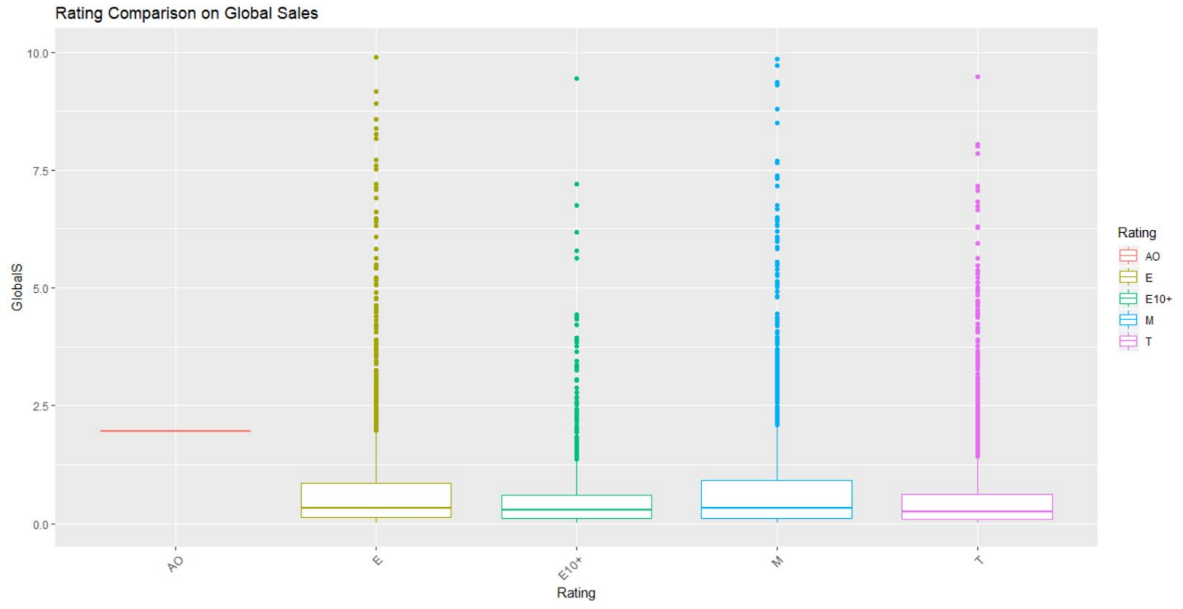
Rating

If we examine the number of video games by rating, we see that T (teenager) and E (everyone) are the most commonly occurring. Games for mature audiences usually appears in the Action and Shooter genres. The

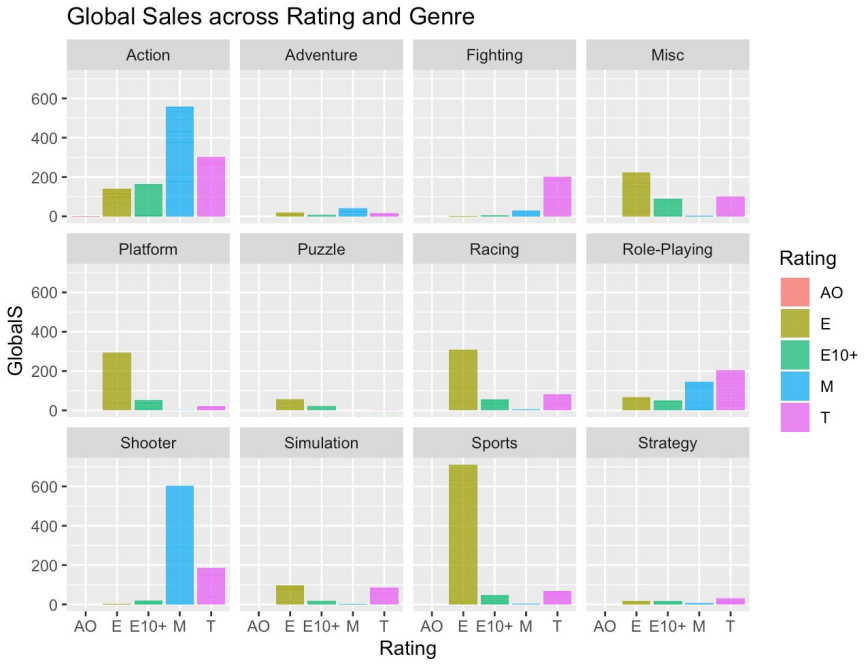
teenager audience is a significant customer group for Action, Fighting, Role-playing, Simulation and Strategy games. Additionally, Sports games tend to target everyone.



The boxplot of global sales against the games' ratings indicates that the games rated as Adult Only have the highest median global sales. However, because the number of Adult Only games considered in the analysis is significantly lower than other types of games, we do not have sufficient evidence to confirm that games rated as Adults Only tend to have better sales than other types of games. If we omit Adult Only games, E (Everyone) and M (Mature) games have the highest global sales. The IQR of E rated games is smaller than the IQR of M rated games, which means that the distribution of global sales of E rated games is less dispersed than the distribution of the global sales of M rated games. Overall, games rated as E or M tend to have better sales than other types of games.

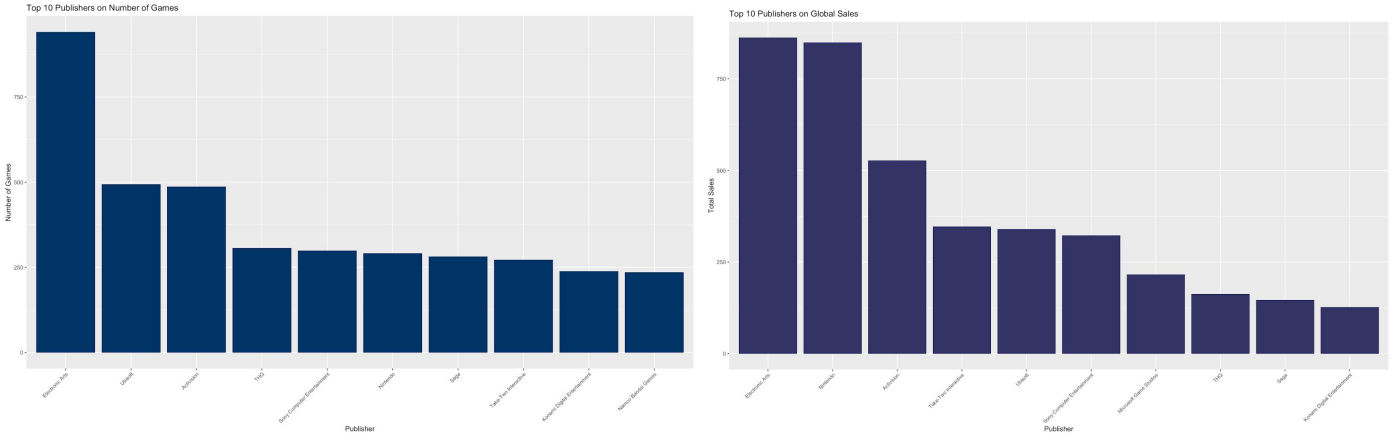


If we look at global sales by genre and by rating, we see that Action, Shooter and Sports are the best selling genres. Action and Shooter games target teenagers and mature audiences, while everyone is targeted for Sports games.

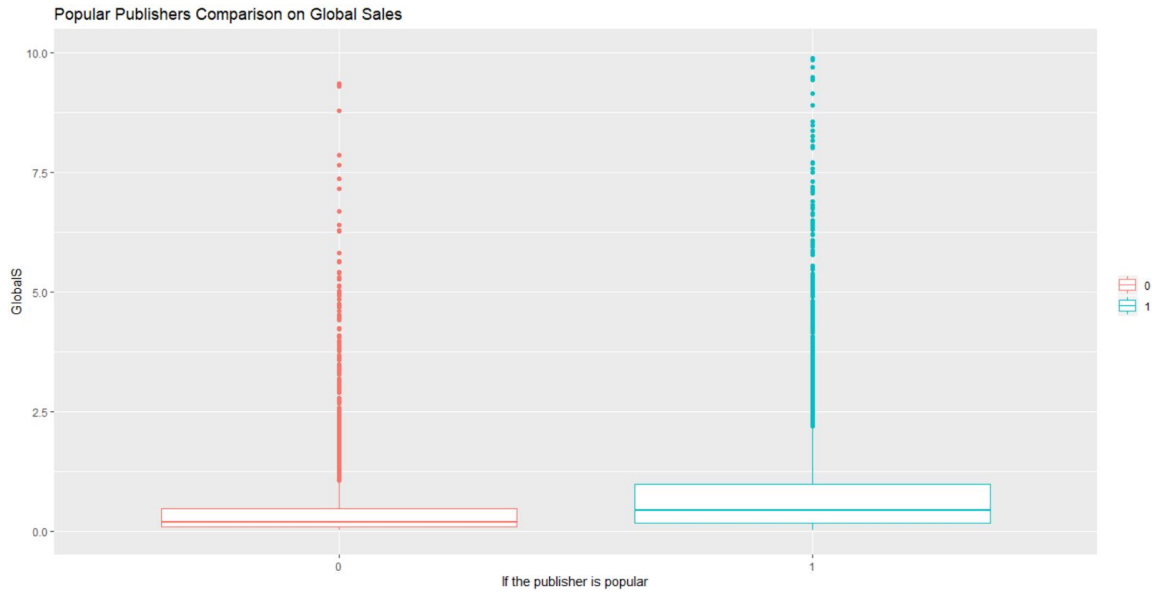


Publisher

EA, Ubisoft and Activision are the most active publishers by number of games, which may be attributed to economies of scale at these firms. Looking more broadly, the top ten publishers by number of games and global sales is almost the same, except Namco Bandai Games, which ranked 10th by number of games, and Microsoft Game Studios, which ranked 7th by global sales. Microsoft Game Studios, an internal Games Group for the development and publishing of video games for Microsoft Windows, published 141 games and sold 216.49 million units of games from Year 2000 to 2016. Its relatively stronger monetisation capabilities may be attributed to the success of "Kinect Adventures!" and "Halo 3", which were released to the market in 2010 and 2007 respectively. Similarly, the release of "Wii sports" may be the major reason why Nintendo can be the second largest by global sales despite a relatively smaller number of games released.



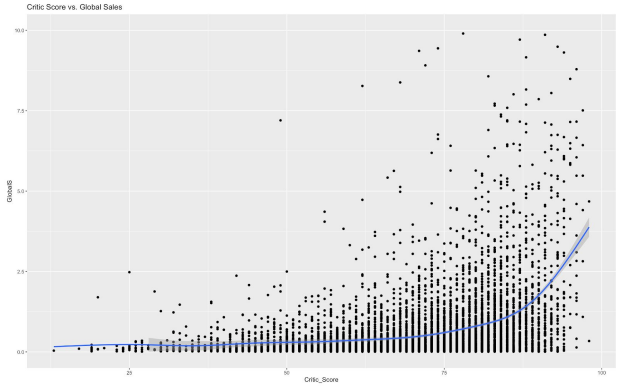
We define a new variable, popular publishers, as publishers listed in the top publisher ranking from 2010-2016 provided by Metacritic. After dividing the data by this new variable, we plot boxplots of global sales by popular/not-popular publishers. From the boxplots below, we can see that popular publishers tend to have better sales than the less popular publishers due to the significant difference in median global sales. We can also observe that the interquartile range for popular publishers is much bigger than the interquartile range for less popular publishers, which means games published by popular publishers tend to have more variation in the amount of global sales. We may conclude that popular publishers are better at manufacturing, marketing and distributing games which subsequently result in higher game sales.



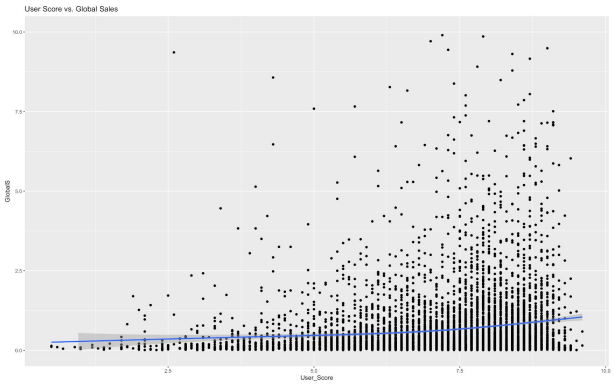
Critic Score and User Score

The upward tendency of the line in a plot of critic score against global sales and user score against global sales allows us to conclude that the score of a game increases with an increase in global sales. The direction of causality is unclear as it could be that higher ratings lead to higher sales, or that higher ratings is a reflection of higher sales. However, the rate of increase between critic scores and user scores is slightly different. For critic scores over 87, sales increases significantly with a small increase in critic score. Therefore, we think that the relationship between critic score and global sales is exponential whereas the relationship between user score and global sales is more linear. For games developers, it would thus be important to market to game critics effectively to boost ratings obtained from them.

Critic Score vs. Global Sales

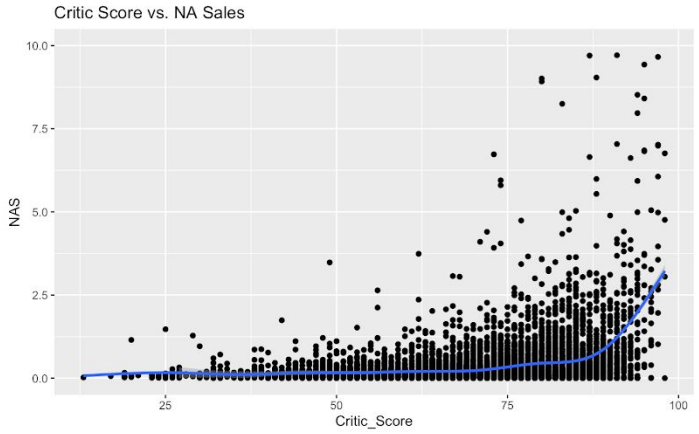


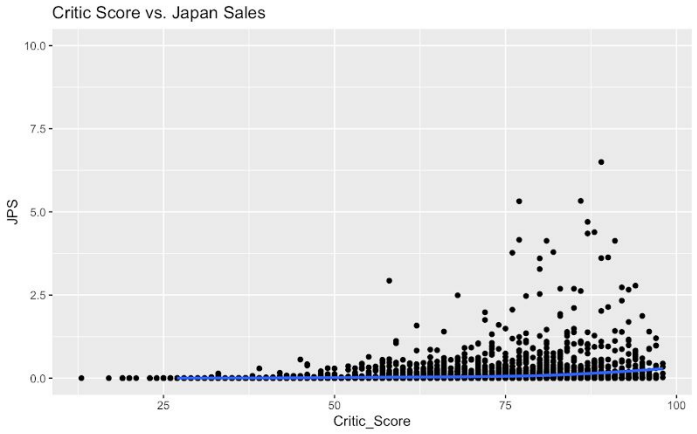
User Score vs. Global Sales

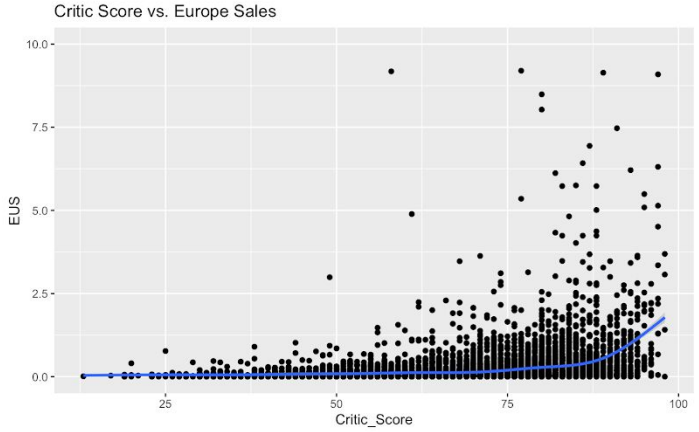


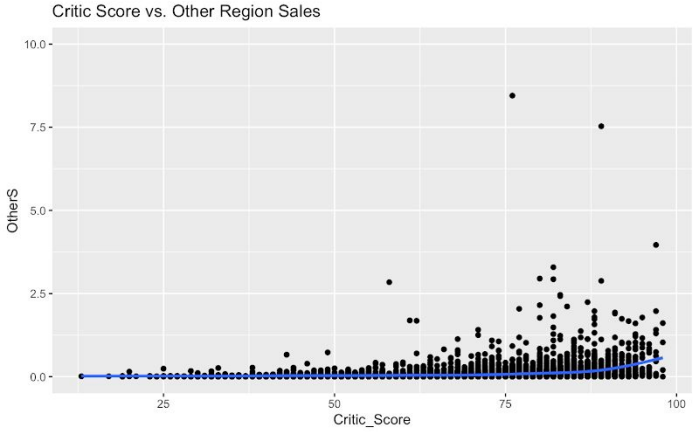
Critic Score vs. Regional Sales

By further checking scatter plots of critic scores over regional sales, we can see that the exponential relationship between critic score and sales is more obvious in North America and Europe. In other regions, more linear patterns are found to be more common.



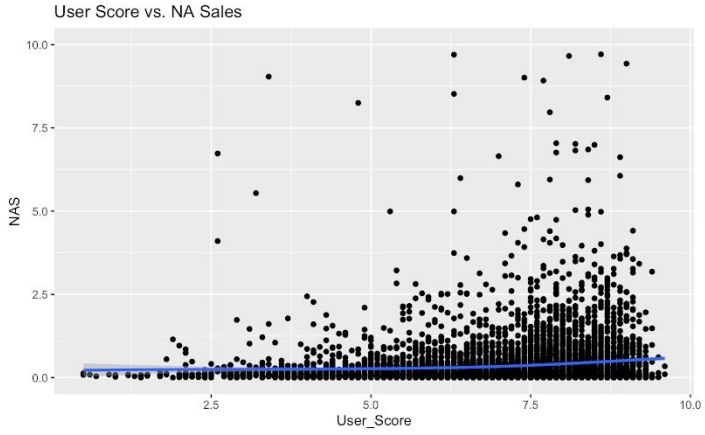


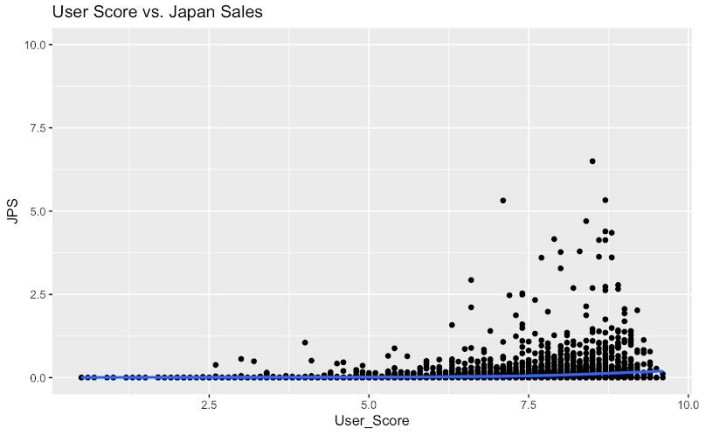


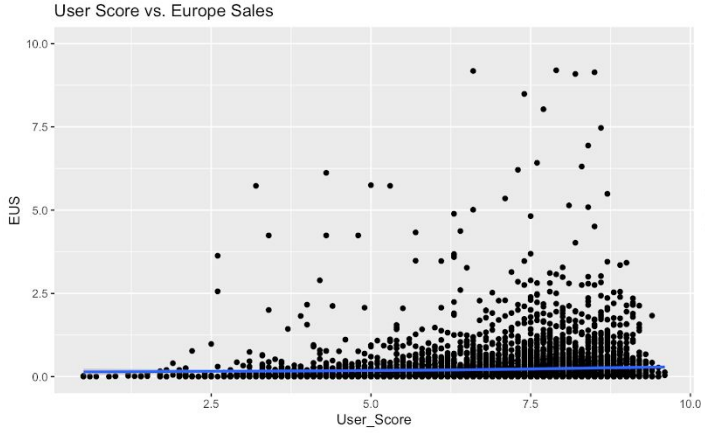


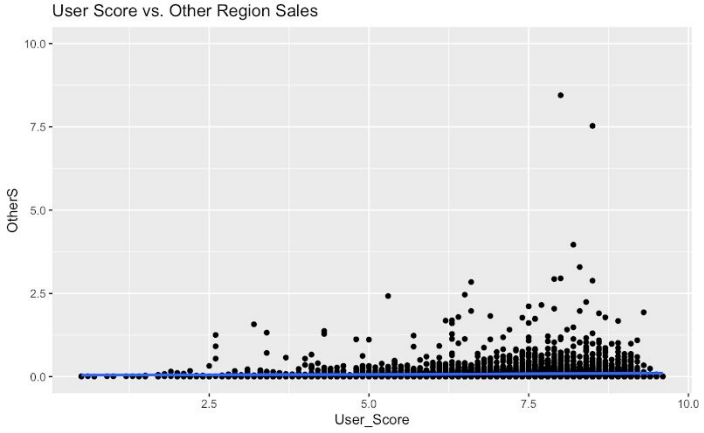
User Score vs. Regional Sales

Regarding user scores and global sales in North America and Europe, there is a smooth linear trend. However, it seems that user scores have less effect on sales in Japan and Other Regions.









Critic Score vs. Global Sales by Genre

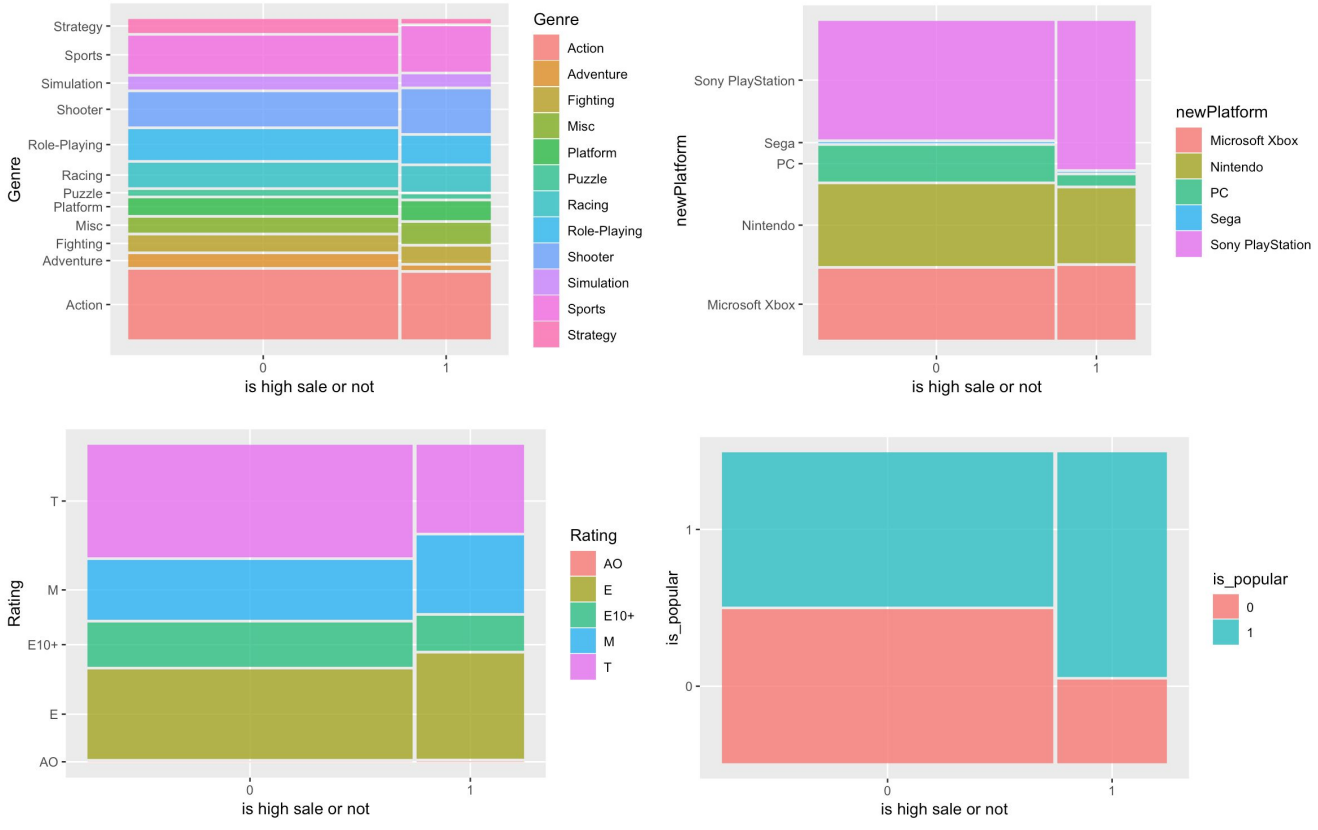
Overall, the effect of critic score on sales is consistent among different genres except for Puzzle and Simulation. If the critic score is higher than 87, sales of games in these two genres actually decrease. One relevant example is “World of Goo” published by “2D boy”, a small publisher company. Even though the game has a score of 90, the publisher may not have enough resources to advertise it worldwide and hence the game has relatively low sales despite the high critic rating.

User Score vs. Global Sales by Genre

The relationship between user score and global sales is linear among most genres. However, user scores have less effect on Adventures, Simulation, and Strategy games. This may be because people who buy these games are less likely to be affected by user scores.

High Sales and Categorical Variables

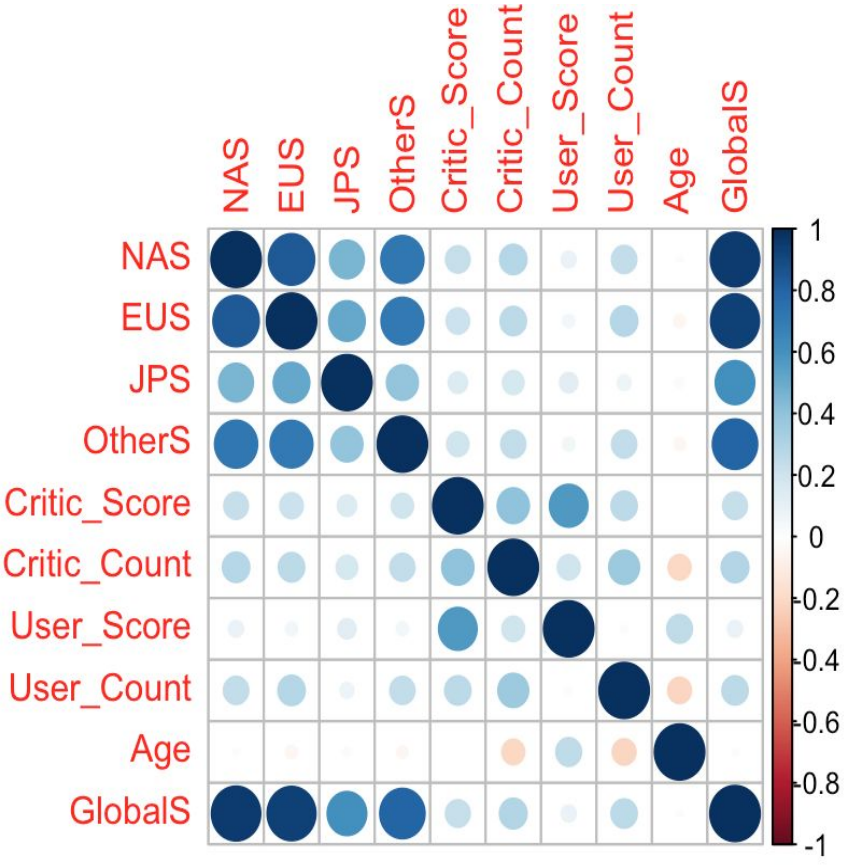
Due to the high dispersion and prevalence of outliers in the global sales data, we define a new variable - whether a global sales value is high or not (1-yes, 0-no) to analyse its correlation with categorical variables Genre, Platform, Rating and Publisher (via the new variable is_popular) using mosaic plots.



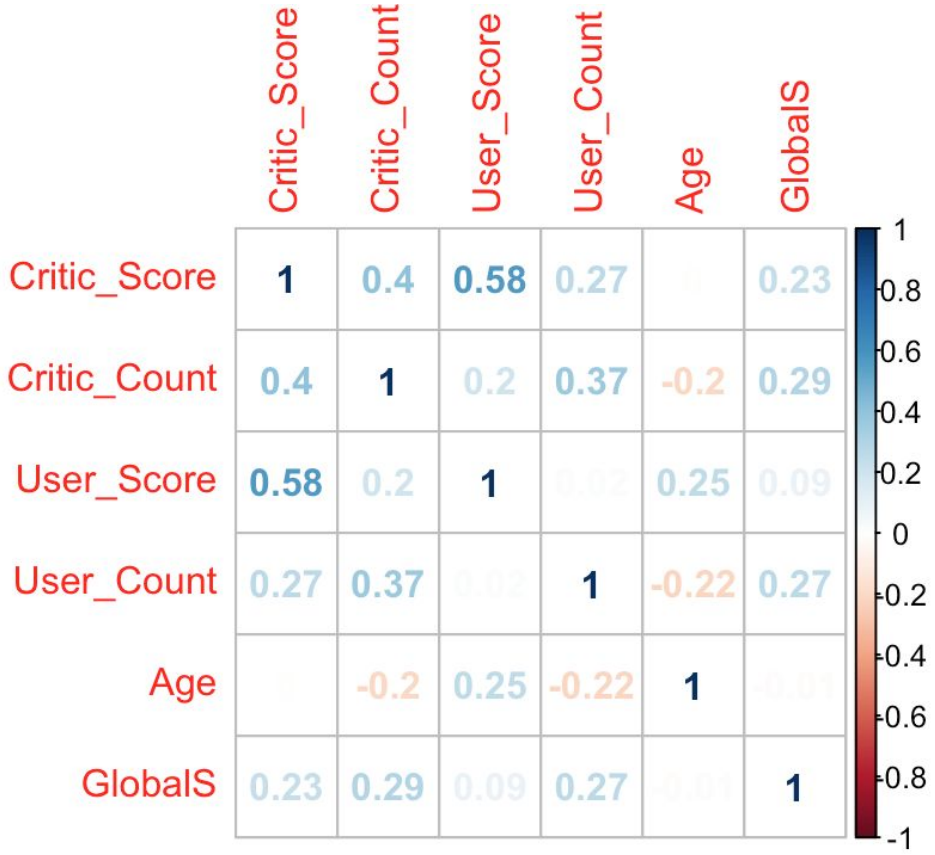
From the first mosaic plot above, we can conclude that global sales amount is related to genre because the width of 12 game types on the left are slightly different from the right. In particular, Sports and Shooter games are more likely to have high sales. From the second plot, we can see that in the category of having a high sale, there are more games from the Sony platform and less from the PC platform, which means that there is a relationship between the platform and global sales. From the third plot, we can see that games with high sales tend to have more rating E, M and less E10+ and T. From the last mosaic plot, it can be seen that the more popular the game's publisher is, the more global sales it will make.

Correlation Analysis

To evaluate the correlation between variables, we conduct a correlation plot on our dataset, as shown below:



In the correlation plot, the size and color of the bubble imply the significance of the relationship between variables. Regional sales and Global Sales have the strongest positive relationships with each other, followed by User Score and Critic Score. As Global Sales is simply a summation of regional sales, we remove regional sales and re-examine the correlation plot.



The result of the new plot is consistent with the former, where only the User Score and Critic Score variables have strong positive correlation, where the correlation coefficients is 0.58.

Modeling

To explore the applications of our analysis and dataset, we build a preliminary version of a predictive regression model. Due to the heavy skewness of the distribution of global sales and large number of categorical variables, we decided to build a logistic model on the training dataset and then made use of this model to classify whether the video game belongs to high global sales or low global sales.

First, using R we split global sales data into two groups: is_highsale or not. The response variables are 0 and 1. A value of 1 indicates that the global sales of the video game is larger than 0.75 million units and a value of 1 indicates that the global sales of the video game is smaller than or equal to 0.75 million units. After the data adjustment, the number of entries belonging to is_highsale is 1,687. Based on the descriptive analysis above, we decided to include both numerical and categorical variables into the logistic regression model. The numerical variables are Critic_Score, Critic_Count, User_Score, User_Count and Age while the categorical variables include Genre, Rating, newPlatform and is_popular.

Generally, a categorical variable with n levels will be transformed into n-1 variables each with two levels. These n-1 new variables contain the same information as the single variable. For example, there are 5 levels in the Rating variable and thus Rating transforms into 4 variables in our model. In this way, we can describe the 3 categorical variables with 3 series of dummy sub-categorical variables. Then we split the full data into a training set and a test set with a weighting of 0.7 and 0.3 respectively. We use the training set to build the model and the test set to validate the performance of the model.

To verify whether the predictions for a given classification model are accurate or not, a baseline prediction model is introduced. It provides a set of predictions that can be used to compute a baseline error rate which becomes a value for comparison when evaluating the logistic regression model. Here we use the Zero Rule algorithm to obtain the baseline. It is a benchmark procedure that uses more information about a given problem to generate a rule and then make a prediction. For our problem, this rule predicts the value of is_highsale which is most common in the training set. Given a list of class values observed in the training set, it suggests that there is a higher probability that the outcome is 0 rather than 1. Thus outcome equal to 0 is selected as a predictive class value for all the predictions in the baseline prediction model. Using the prediction outcomes compared to the class values in the test set, we computed the error rate equal to 23.8% (round to 1 decimal places).

	Number of 0 in dataset	Number of 1 in dataset	Error Rate
Training Set	3559	1201	1201/4760 * 100% = 25.2%
Test set	1555	486	486/2041 * 100% = 23.8%

Furthermore, we fit the model as the function below and obtain the result in R:

```
glm(formula = is_highsale ~ Critic_Score + Critic_Count + User_Score +
    User_Count + Age + Genre + Rating + newPlatform + is_popular,
    family = "binomial", data = sample_train)
```

Now we can analyze the fitting and interpret the implications of the model.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              3.286e+00  1.970e+02   0.017 0.986691
Critic_Score             6.791e-02  4.846e-03  14.013  < 2e-16 ***
Critic_Count             3.484e-02  2.623e-03  13.280  < 2e-16 ***
User_Score              -2.109e-01  4.110e-02  -5.131 2.89e-07 ***
User_Count               1.165e-03  1.625e-04   7.167 7.68e-13 ***
Age                      4.865e-02  1.152e-02   4.223 2.41e-05 ***
GenreAdventure          -1.375e+00  3.663e-01  -3.755 0.000174 ***
GenreFighting           -3.117e-02  1.893e-01  -0.165 0.869204
GenreMisc                7.238e-01  1.781e-01   4.064 4.82e-05 ***
GenrePlatform           -1.935e-01  2.013e-01  -0.961 0.336432
GenrePuzzle             -7.466e-01  3.120e-01  -2.393 0.016700 *
GenreRacing             -7.317e-02  1.749e-01  -0.418 0.675750
GenreRole-Playing       -4.190e-01  1.587e-01  -2.640 0.008301 **
GenreShooter             2.218e-02  1.442e-01   0.154 0.877767
GenreSimulation          5.004e-01  2.075e-01   2.412 0.015883 *
GenreSports             -1.331e-01  1.593e-01  -0.835 0.403465
GenreStrategy           -9.117e-01  3.058e-01  -2.981 0.002874 **
RatingE                 -1.046e+01  1.970e+02  -0.053 0.957635
RatingE10+              -1.071e+01  1.970e+02  -0.054 0.956634
RatingM                 -1.094e+01  1.970e+02  -0.056 0.955705
RatingT                 -1.083e+01  1.970e+02  -0.055 0.956133
newPlatformNintendo      6.076e-01  1.268e-01   4.791 1.66e-06 ***
newPlatformPC           -2.448e+00  3.108e-01  -7.876 3.37e-15 ***
newPlatformSega         -1.368e+00  1.119e+00  -1.222 0.221566
newPlatformSony PlayStation  7.911e-01  1.062e-01   7.450 9.30e-14 ***
is_popular               4.697e-01  8.884e-02   5.287 1.25e-07 ***
```

As shown in the result table above, all of the numerical variables (Critic_Score, Critic_Count, User_Score, User_Count and age) are significant in predicting the global sales of video games. Additionally, whether the publisher is popular is also useful in predicting global sales. As for the 3 categorical variables (Genre, Rating and Platform), since both Genre and Platform have significant dummy sub-categorical variables, they

also contribute to the prediction of global sales. Although Rating does not appear to be significant, we cannot conclude that it is insignificant at this junction and will do further tests to determine its impact on global sales. For Genre, the reference level is GenreAction and compared with GenreAction, the negative coefficient of GenreAdventure for this predictor suggests that all other variables being equal, the global sales of a GenreAdventure game is less likely to have a high value compared to GenreAction. For Platform, the reference level is Platform-Microsoft Xbox. The positive coefficient of Platform-Nintendo for this predictor suggests that all other variables being equal, the global sales of a Platform-Nintendo game is more likely to have a higher global sales value compared to a Platform-Microsoft Xbox game.

To further test the significance of the Rating variable, we perform an ANOVA Chi-square test to check the overall effect of variables on the dependent variable.

The difference between the null deviance and the residual deviance shows how our model is doing against the null model (a model with only the intercept). The wider this gap, the better. Analyzing the table we can see the drop in deviance when each variable is added. From the low p-value, we can conclude that all the variables we use in the model are significant in predicting global sales under the 5% significance level. In particular, using the Critic_Score, Critic_Count, Genre and platform variables can significantly reduce the residual deviance.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			4759	5377.5		
Critic_Score	1	670.06	4758	4707.4	< 2.2e-16	***
Critic_Count	1	289.02	4757	4418.4	< 2.2e-16	***
User_Score	1	20.70	4756	4397.7	5.383e-06	***
User_Count	1	6.50	4755	4391.2	0.0108	*
Age	1	19.67	4754	4371.5	9.215e-06	***
Genre	11	127.37	4743	4244.2	< 2.2e-16	***
Rating	4	30.44	4739	4213.7	3.989e-06	***
newPlatform	4	243.98	4735	3969.7	< 2.2e-16	***
is_popular	1	28.31	4734	3941.4	1.033e-07	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

In addition to the earlier correlation analysis, we also calculate VIF values for all variables to determine if there is a problem of multicollinearity. We find that all VIF values are less than 10 and hence there is no multicollinearity problem that needs to be addressed in the model.

	GVIF	Df	GVIF^(1/(2*Df))
Critic_Score	1.730199	1	1.315370
Critic_Count	1.521165	1	1.233355
User_Score	1.696089	1	1.302340
User_Count	1.982549	1	1.408030
Age	1.419006	1	1.191220
Genre	3.123809	11	1.053139
Rating	3.093679	4	1.151621
newPlatform	2.130138	4	1.099135
is_popular	1.109175	1	1.053174

Training Set		Actual 0	Actual 1
	Predict as 0	3351	656
	Predict as 1	208	545
Error Rate = (656+208) / 4760 * 100% = 18.2%			
Test Set		Actual 0	Actual 1
	Predict as 0	1475	258
	Predict as 1	80	228
Error Rate = (258 + 80) / 2041 * 100% = 16.6%			

The error rate of the logistic regression model calculated from the test set is 16.6%, which is 7.2% lower than the baseline model, which shows that our logistic regression model is effective in classifying the level of global sales of video games given the related variables. However, there exists some limitations in the model that can be further improved in the future. For example, due to the large number of categorical variable classes, we did not conduct additional analysis on the effect of each interaction term on the model.

Recommendation and Conclusion

From our analyses, we would recommend the developers to develop a Shooter game, rated M, on the Sony Playstation platform, distributed by Electronic Arts with a focus on North American and European markets.

We recommend the Shooter genre, because it is one of the best selling and most resilient genres and has a higher correlation to high global sales than under independence. From the logistic regression model, all else being equal, it is also more likely to have higher sales compared to an Action game. However, if the developer prefers to focus on the Japanese market, we would instead recommend simulation, role-playing or puzzle genres due to the unique taste of gamers in that market. For rating, we would recommend the highly monetizable Mature segment that is both appropriate for a Shooter genre and has been shown to be more positively correlated to high sales. For platform, we would recommend the Sony Playstation platform due to the greater number of players already on this platform (which represents a higher addressable market) and its effective market expansion strategy, as evidenced by the increase in its market share over recent years. The logistic regression model also shows that developing for the Sony Playstation platform is more likely to generate higher sales compared to the Microsoft Xbox platform, all else being equal. Regarding distribution, we would recommend that the developer select a popular distributor, which is more likely to achieve higher sales, and that has economies of scale to leverage. EA satisfies both and is the top publisher by both number of games and total sales and hence is our recommended publisher. Finally, we recommend that particular resources be allocated to distributing the game in North America and European markets due to the significantly larger absolute size of these two markets, which would increase the likelihood of high sales.

These recommendations have been possible to synthesize due to the power of visual analytics and the wide variety of tools that can be used to analyse the relationship between variables. In this report we have endeavoured to use a multitude of visual methods and correlation analyses to explore the nature of our dataset. In particular, we have used stacked and clustered barcharts, scatter plots, polar plots, bubble charts, histograms, boxplots, correlation plots, mosaic plots, tables and more.

However, there are significant limitations to our data, methodology and modelling that can serve as interesting questions for future exploration. We would like to highlight the potential for examining underexplored opportunities in the industry. This can be done by searching for genres with high sales per game but less games released and/or low ratings from critics and users. This subset of video games would have high potential to perform, as evidenced by gamers' willingness to pay, and less competition (in terms of absolute number of video games and relatively poor performance by existing competitors), which should increase the likelihood of higher sales. Similar questions can be developed for further studies.

After detailed visual analyses on the main variables and key aspects of our dataset, we have gained many insights on the nature of the video games industry and the relationship between the various variables and global sales. These insights can serve to reduce some of the uncertainty in the industry and better guide developers in their efforts to build high performing games in the future.