CO4 LSE DA301

Advanced Analytics for Organisational Impact

Predicting future outcomes: Sales Prediction and customer analysis

Elaine Wong

April 2023

Section I - Background

The purpose of this report is to describe current market conditions and identify customer trends that may affect the marketing strategies of Turtle Games.

Turtle Games wish to increase customer loyalty and they have a loyalty program. The marketing team want to know what may affect the customer loyalty point.
On the other hand, video game market trend may also affect Turtle Games's business.

Section II - Analytical Approach

To analyze the customer reviews and global sales, exploratory data analysis and 3 predictive methods have been applied.

1. Exploratory Data Analysis

Explore the dataset by visualization:

- Histogram: categorical variables
- Box plot: numeric variables, or show relationships between numeric variable and categorical variable
- Scatter plot: show relationships between 2 variables
- Line chart: trend

2. Linear Regression

Simple linear regression and multiple linear regression are used for

- **predict loyalty points** in dataset 1 (Python)
- **predict the global sales** in dataset 2 (R)

2.1 Simple linear regression

- Step 1: Build model

  - identify the independent variable (x) and dependent variable (y)

  - fit linear regression model and run OLS in Python, lm in R

- Step 2: Evaluate model’s R-squared

In dataset 1, we used 5 independent variables for simple linear regression but the model R-squared for categorical variables are very low.

Independent variable (x)	Dependent variable (y)	R-squared
education^	Loyalty points	0.001
gender_Male*	Loyalty points	0.000
age	Loyalty points	0.002
renumeration	Loyalty points	0.380
spending_score	Loyalty points	0.452

^ Tranformed the education variable into label, 0=basic and diploma, 1=graduate, 2=postgraduate, 3=PhD
* gender_Male is the dummy variable of gender, 0=Female and 1=Male

In dataset 2, we used 7 independent variables for simple linear regression.
In order to improve the performance, I also take log transformation for the dependent variable.

Independent variable (x)	Dependent variable (y)	R-squared
NA_Sales	Global_Sales	0.8745
	log_ Global_Sales	0.3993
EU_Sales	Global_Sales	0.7705
	log_ Global_Sales	0.382
Other_Sales	Global_Sales	0.6732
	log_ Global_Sales	0.3958
num_of_platforms	Global_Sales	0.1514
	log_ Global_Sales	0.2175
Ranking	Global_Sales	0.1535
	log_ Global_Sales	0.7847
Year	Global_Sales	0.05953
	log_ Global_Sales	0.04965
console_cat*	Global_Sales	0.004149

2.2 Multiple linear regression

- Step 1: Estimate correlation coefficients
- Step 2: Build model

  - identify the multiple independent variables (x) and dependent variable (y)

  - fit linear regression model and run OLS in Python, lm in R

- Step 3: Evaluate model result

  - p-value of x: significance

  - Adjusted R-squared: percentage of data explanatory

- Standard Error: percentage of data explanatory

- Residual / Root Mean Squared Error

In dataset 1, we pick the model 4, the R-squared did not decrease a lot but it is a simple model.

	Independent variables (x)	Dependent variable (y)	Adjusted R-squared
1	gender, education, age, renumeraion & spending scores	Loyalty points	0.846
2	education, age, renumeraion & spending scores	Loyalty points	0.844
3	age, renumeraion & spending scores	Loyalty points	0.842
4	renumeraion & spending scores	Loyalty points	0.830

In dataset 2, we pick the model 2 with highest R-square.

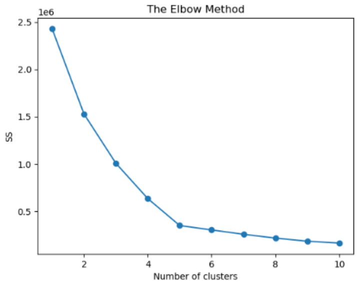
	Independent variables (x)	Dependent variable (y)	Adjusted R-squared
1	NA_Sales, EU_Sales Other_Sales	Global_Sales	1 (Over-fitting)
2	NA_Sales, EU_Sales	Global_Sales	0.9685
3	NA_Sales, Other_Sales	Global_Sales	0.9562
4	EU_Sales Other_Sales	Global_Sales	0.8669

3. K-means clustering

This is an unsupervised machine learning to cluster the data. As we found that renumeration and spending score are highly related with the loyalty points, we can have some customer segments with those 2 variables.

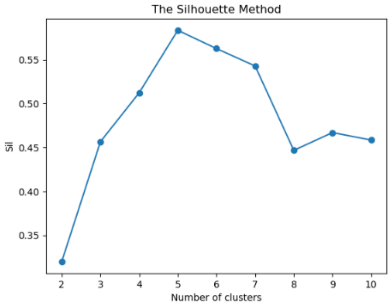
- Step 1: Create scatterplot to show the relationship between renumeration and spending score
- Step 2: Create pairplot to compare the distribution
- Step 3: Elbow method

  The plot suggests that the elbow is formed with K value around 5.
  After K=5, the SSE starts decreasing slowly.



- Step 4: Silhouette method

  The plot suggests the average Silhouette Index is high for K-value around 5-7.



- Step 5: Evaluate k-means model with different k values
- Step 6: Visualise the data with different number of cluster and compare the difference between clusters in the plots

4. Natural language processing

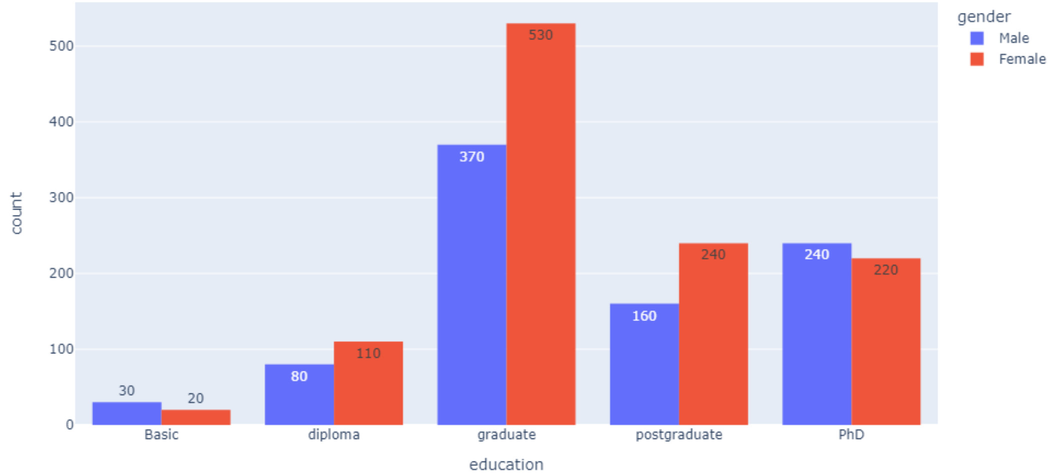
By using NLP, it helps to identify top common words from customer review & comment summary, and also helps to analyze the sentiment and polarity.

- Step 1: Data preparation
  - change to lower case, replace punctuation and remove duplicated records
- Step 2: Tokenise the words
- Step 3: Remove alphanumeric characters and stopwords
- Step 4: Identify the most frequent word and visualize by create wordcloud
- Step 5: use TextBlob package to analyze the polarity sentiment scores of summary and reviews, visualize th distribution of polarity sentiment scores by histrogram
- Step 6: identify the top 20 top positive/negative review/summary by extract the record with highest/lowest polarity sentiment scores respectively

Section III – Visualisation and insights

1: Education

- **Use a plotly grouped histogram plot to show the distribution (can filter the gender easily)**

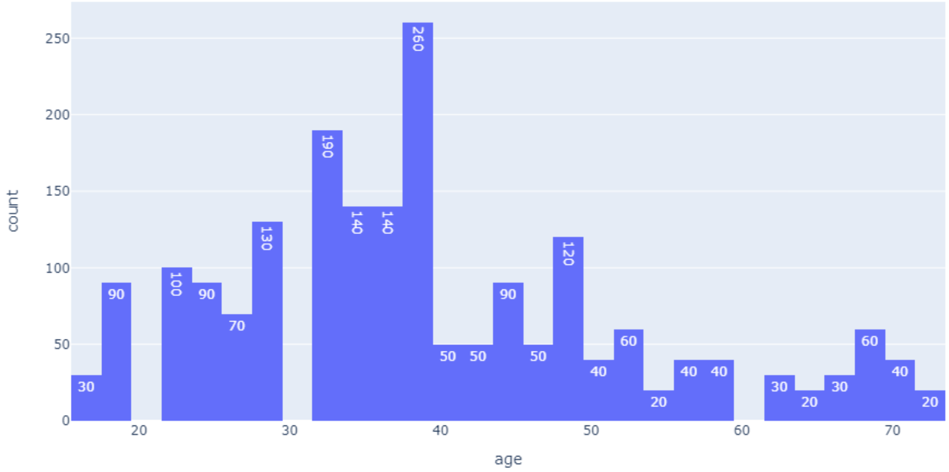


Over 40% of customers are postgraduate or PhD, it is a high portion.

The no. of males in the dataset is less than no. of female, but we found that the male PhD are more than female PhD in our data.

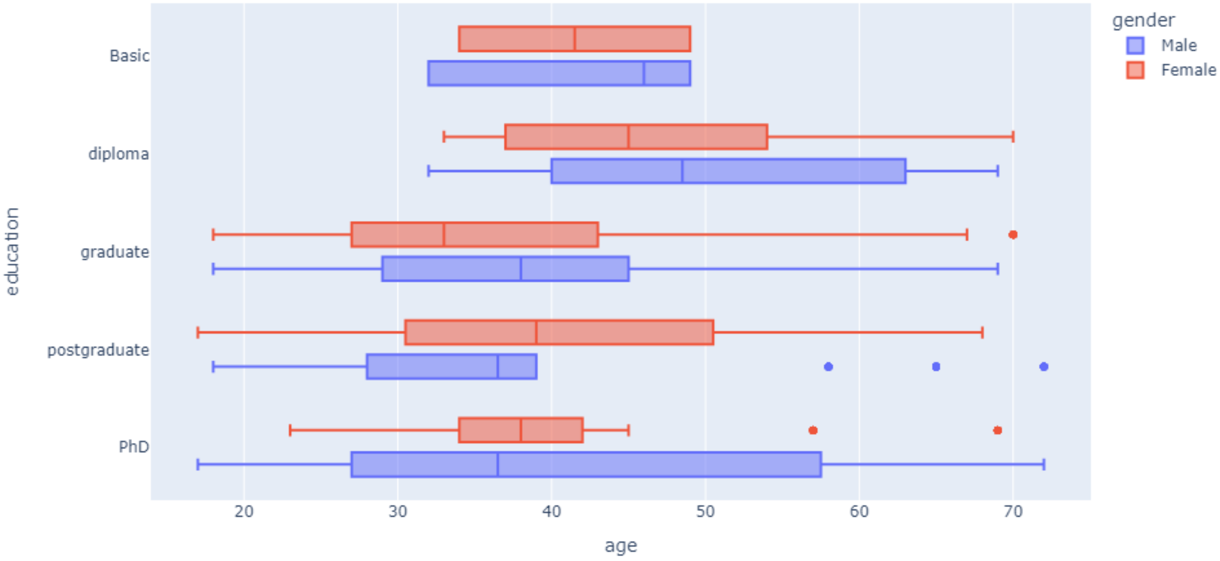
2: Age

- **Use a histogram to show the age distribution**



The histogram is right-tailed.
The majority customer age group is 30s. We saw that the average age was 39.5 as well, so this is not surprising.

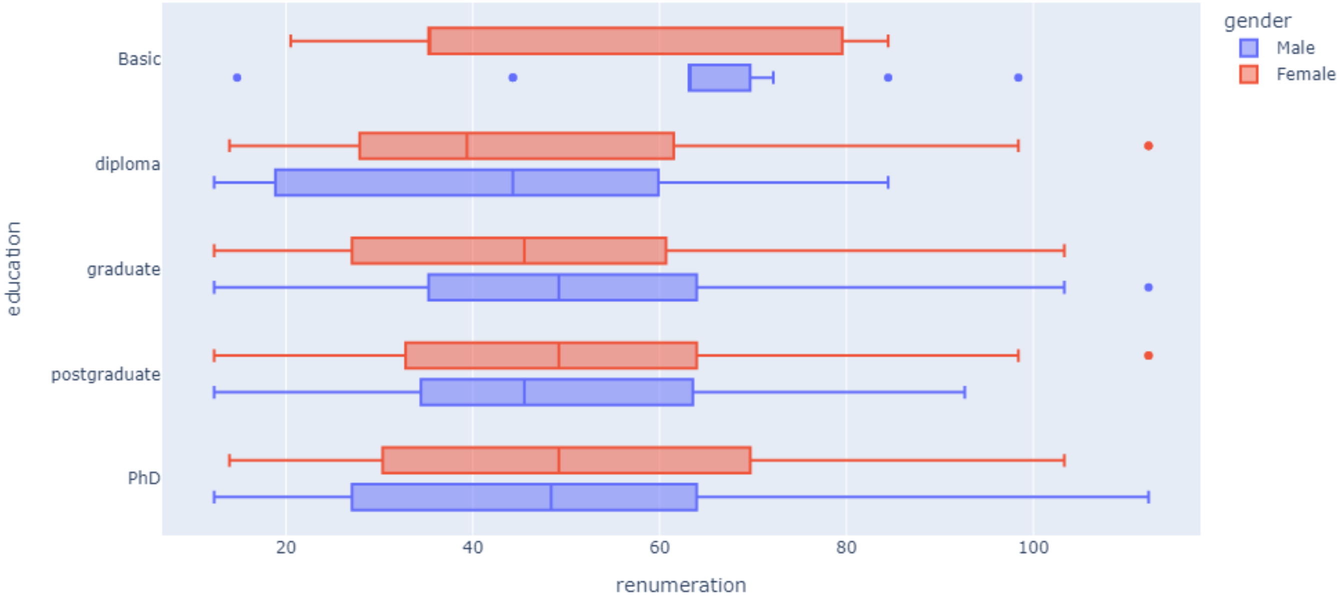
- **Use a plotly box plot to show the age range by education (can filter the gender easily)**



The young generation maybe received higher education.

3: Remuneration

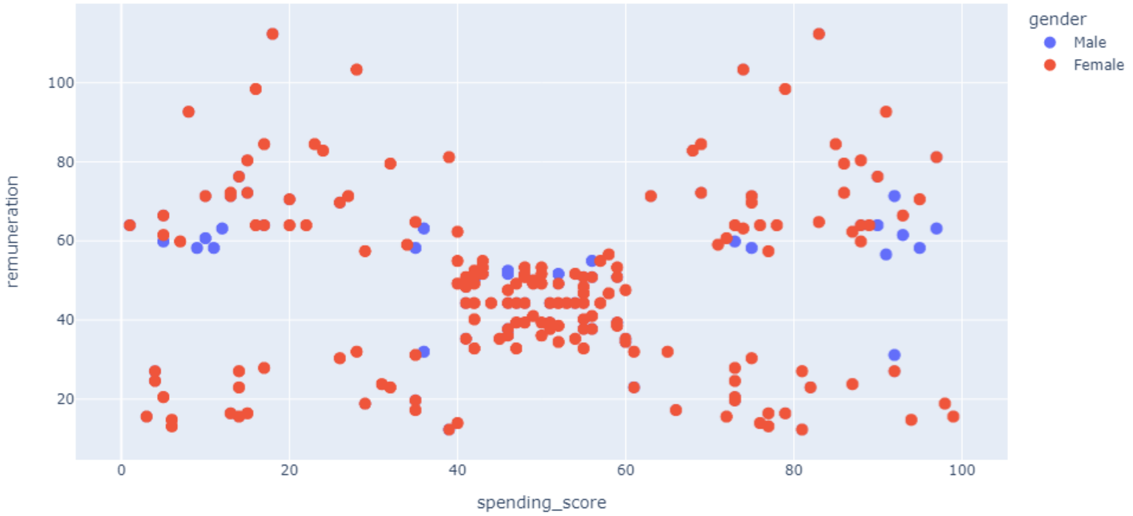
- **Use a plotly box plot to show the remuneration range by education (can filter the gender easily)**



The remuneration difference among education is not very big.

4: Spending score & Renumeration

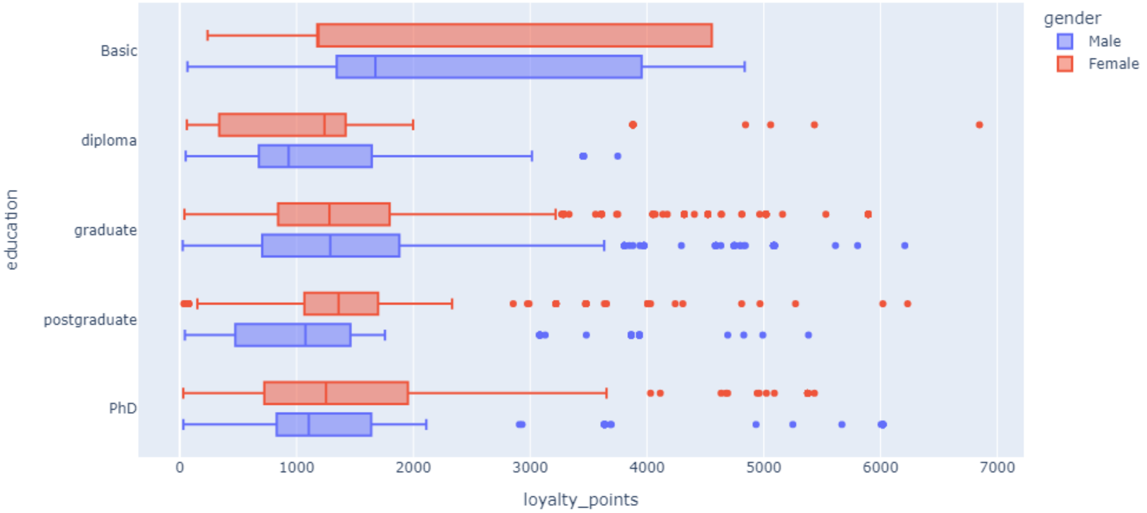
- **Use a plotly scatter plot to show the remuneration and spending score (can filter the gender easily)**



Interestingly, there are 2 customer groups with high spending scores: high remuneration group and low renumeration group. Meanwhile, the middle high remuneration group get middle level of spending scores.

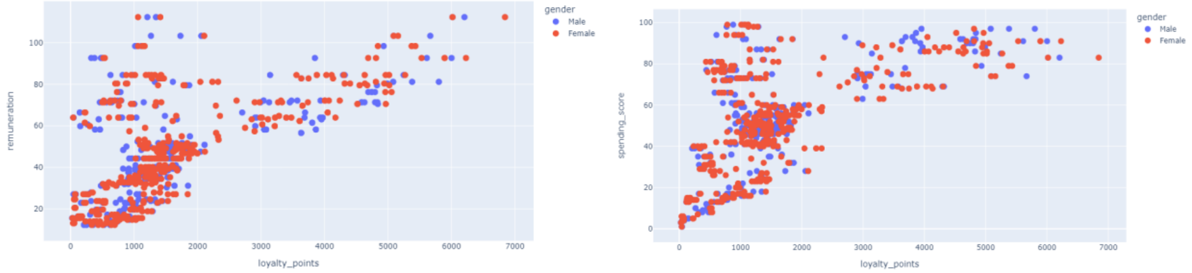
5: Loyalty point

- **Use a plotly box plot to show the loyalty point range by education (can filter the gender easily)**



There are wide range of loyalty points and we can see median between 1000-2000 scores.

- **Use a plotly scatter plot to show the relationship between loyalty points and remuneration (left diagram) ,and between loyalty points and spending score (right diagram)**

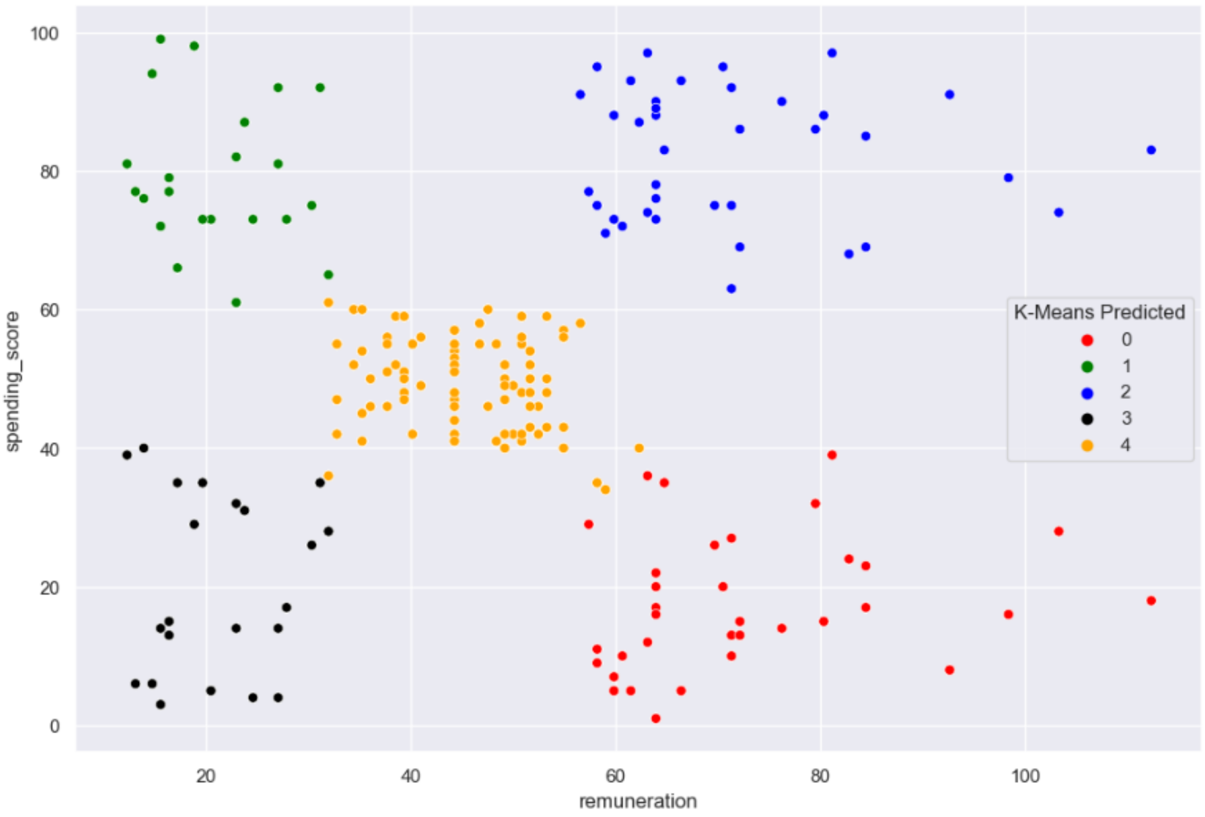


Both diagrams shown linear relationship between loyalty points and remuneration (left diagram), and between loyalty points and spending score.

For predict the loyalty point with multiple linear regression in section 2, remuneration and spending scores are the most important variables.

6: Customer clusters

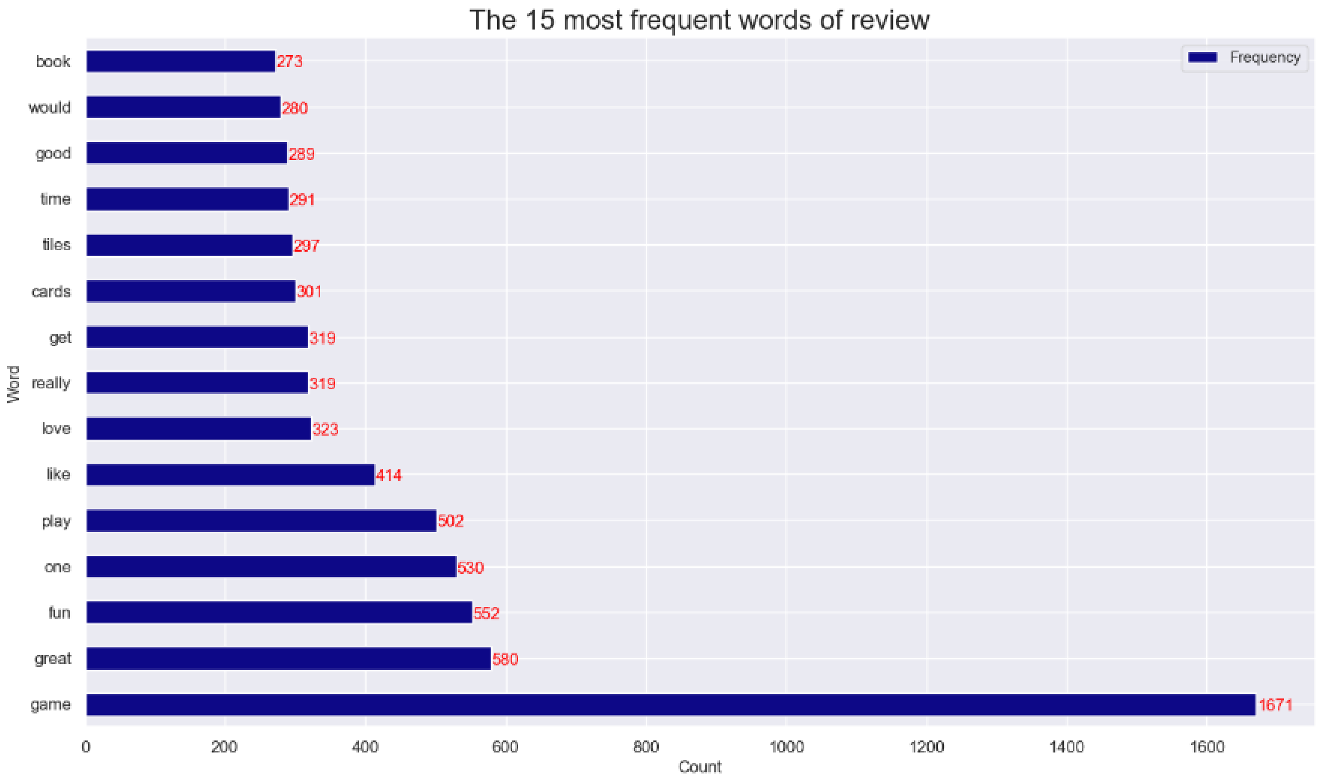
- **Use a scatter plot to show the clustering by remuneration and spending score**



This is the K-mean clustering result. As spending score & remuneration are highly related with the loyalty point and Turtle Game should focus the key customer group 2 (blue point).

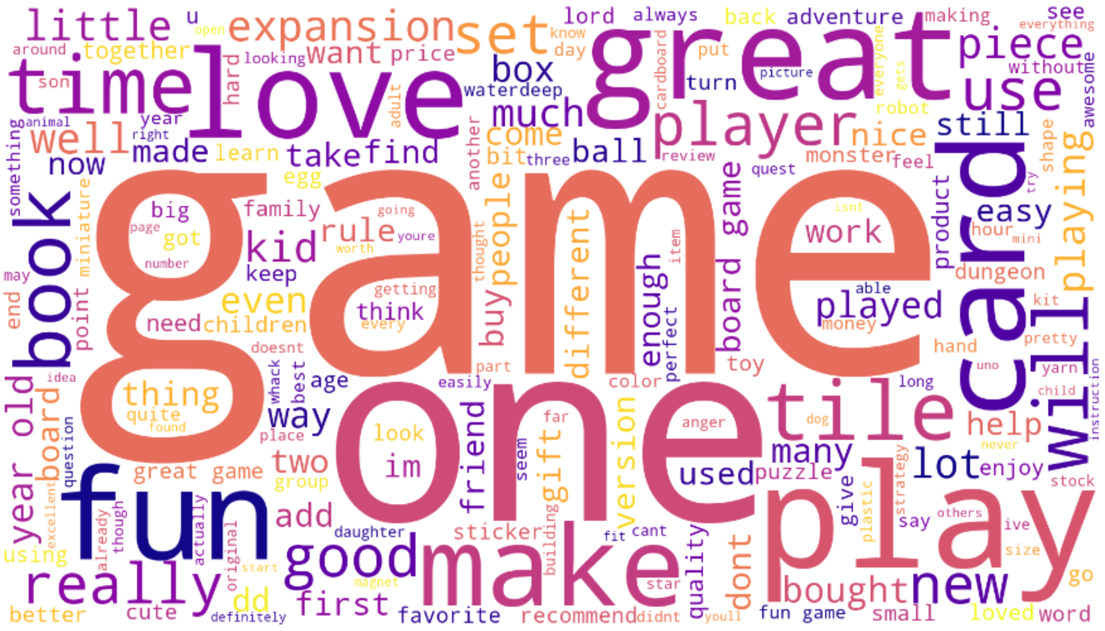
7: 15 most common words

- **Use a bar plot show the word frequency in ascending order, easily to compare the count**



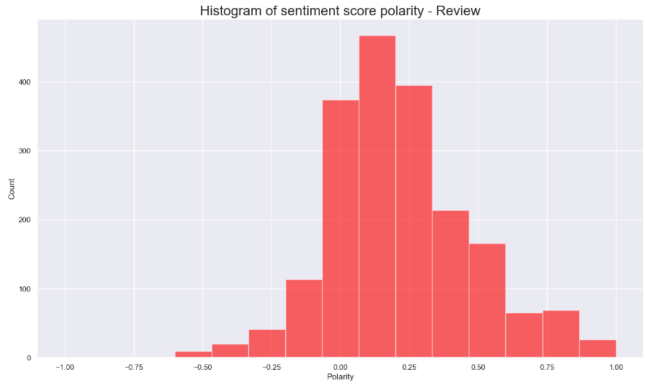
The popular product types are game, card and book. Many words related positive sentiment.

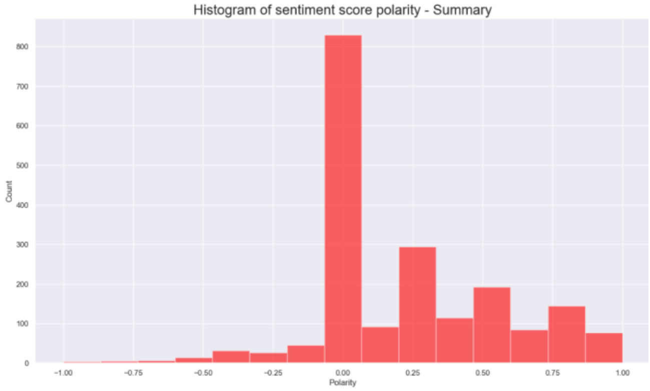
- **Use a word cloud show the word frequency by fancy and eye-catching style**



8: customer review & summary

- **Use histrograms to show the sentiment score popularity of review and summary**



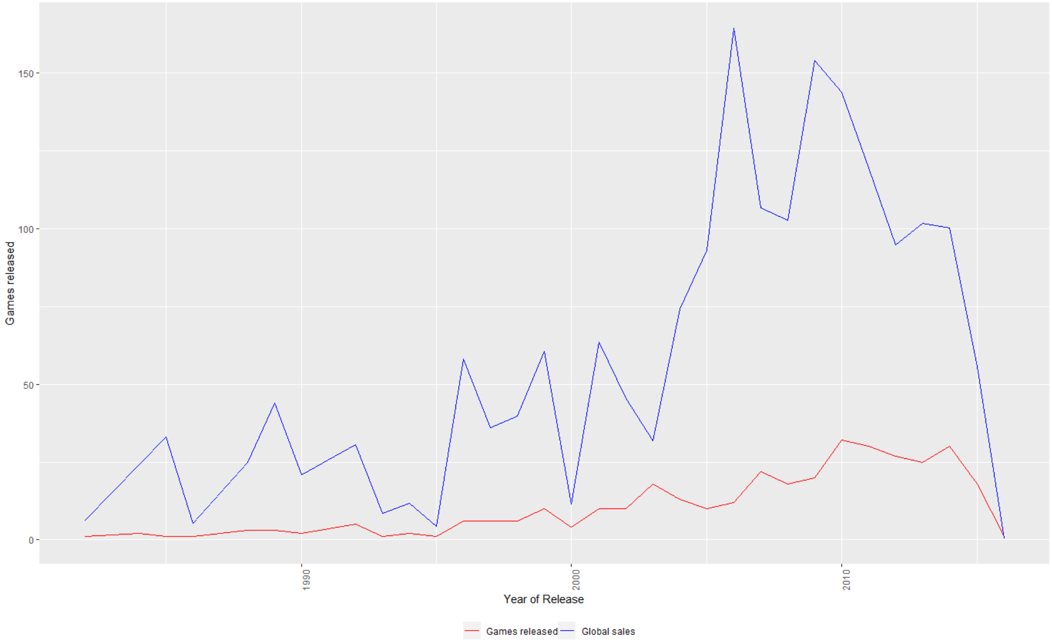


More positive reviews than negative reviews, and the review tends to positive (not very high score)
For summary, we got many neutral summaries, and other summaries are tends to positive (not very high score)

■ **Top 20 positive reviews** - awesome	■ **Top 20 positive summaries** - Gift - Teaching tool - Excellent therapy tool
■ **Top 20 negative reviews** - Boring - Complicated and difficult, unclear instruction - Recommended by therapists and counselor	■ **Top 20 negative summaries** - Disappointed - Bad quality - expensive than other seller

10. Game release trend

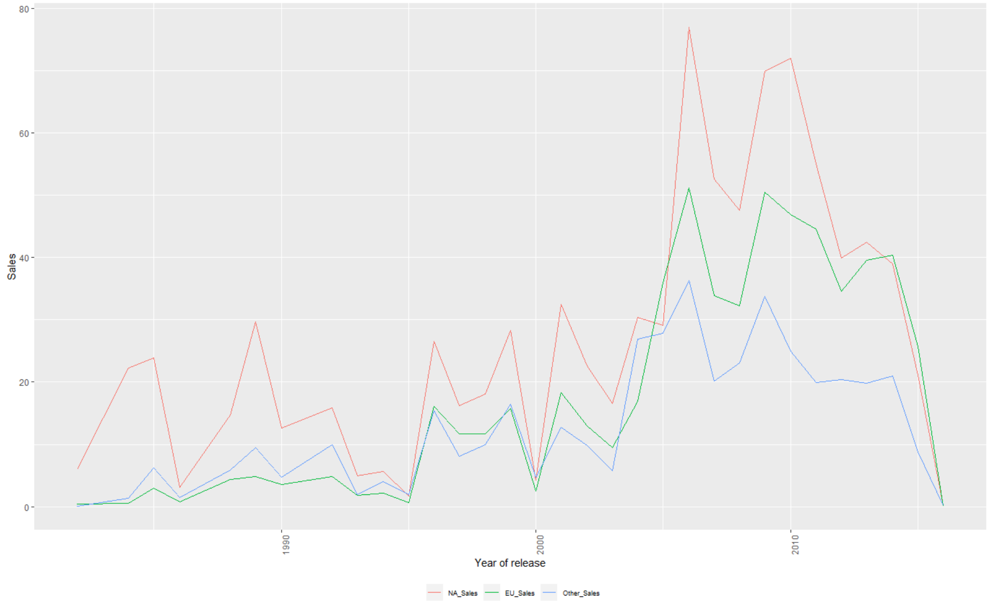
- **Use line chart to show the global sales trend and game released trend**



The number of game released is in an increasing trend (red line) and the global sales also increasing (blue). The market expanded.

11. Global sales trend per region

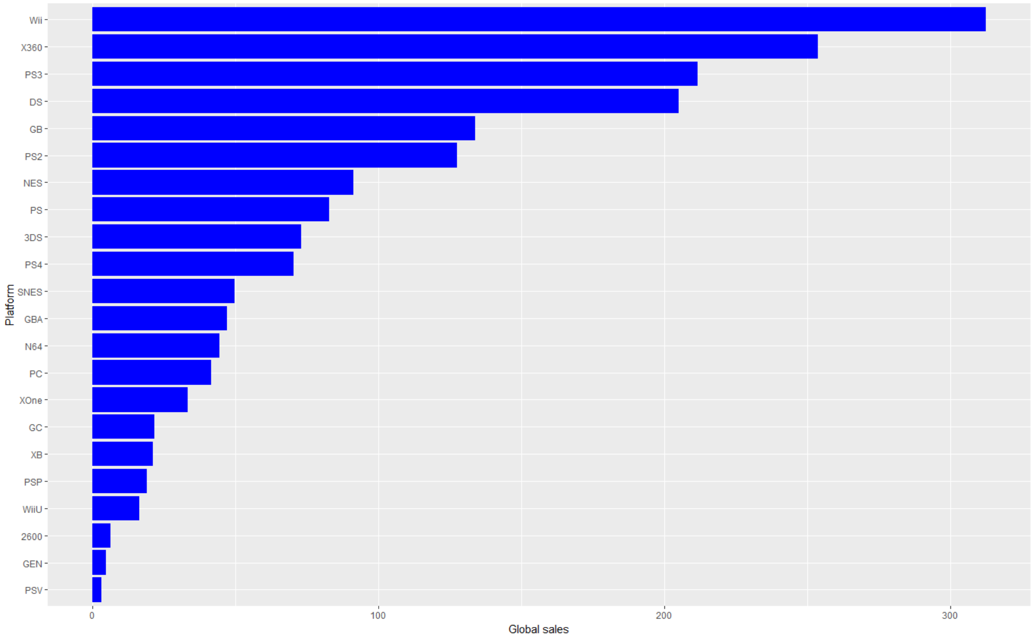
- **Use line chart to show the global sales trend per region**



North America market is the top sales region.

12: global sales per platform

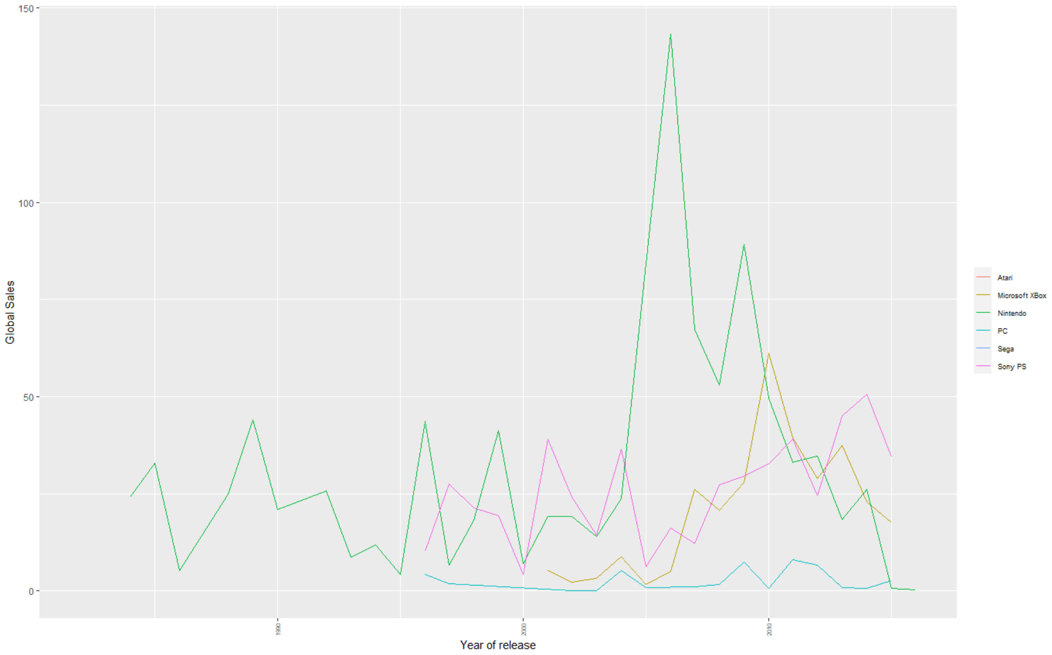
- **Use histrograms to show the global sales per platform**



The top 4 platforms contribute the most of the global sales.

13. Sales per platform categories

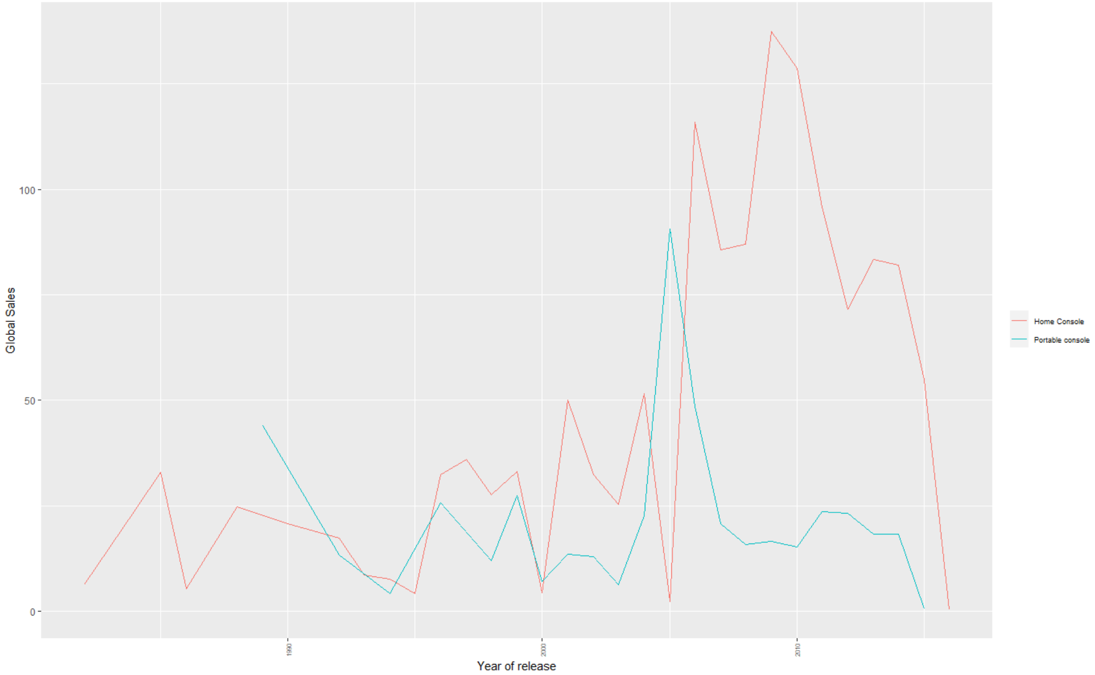
- **Use line chart to show the global sales trend per platform category**



As there are too many platforms in the dataset, let group them as 6 categories and see the sales trend. Xbox series is the newest category but it grows rapidly and become the top 2 category.

14. Sales per console category

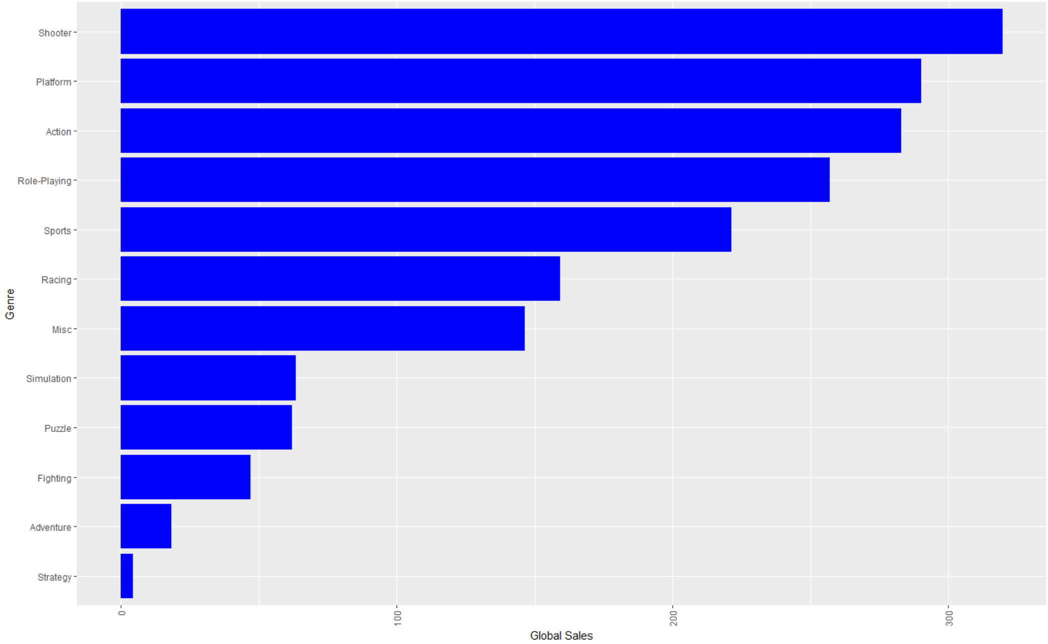
- **Use line chart to show the global sales trend per console category**



The home console is mainstream, the sales of portable console games in decreasing trend.

15: global sales per game genre

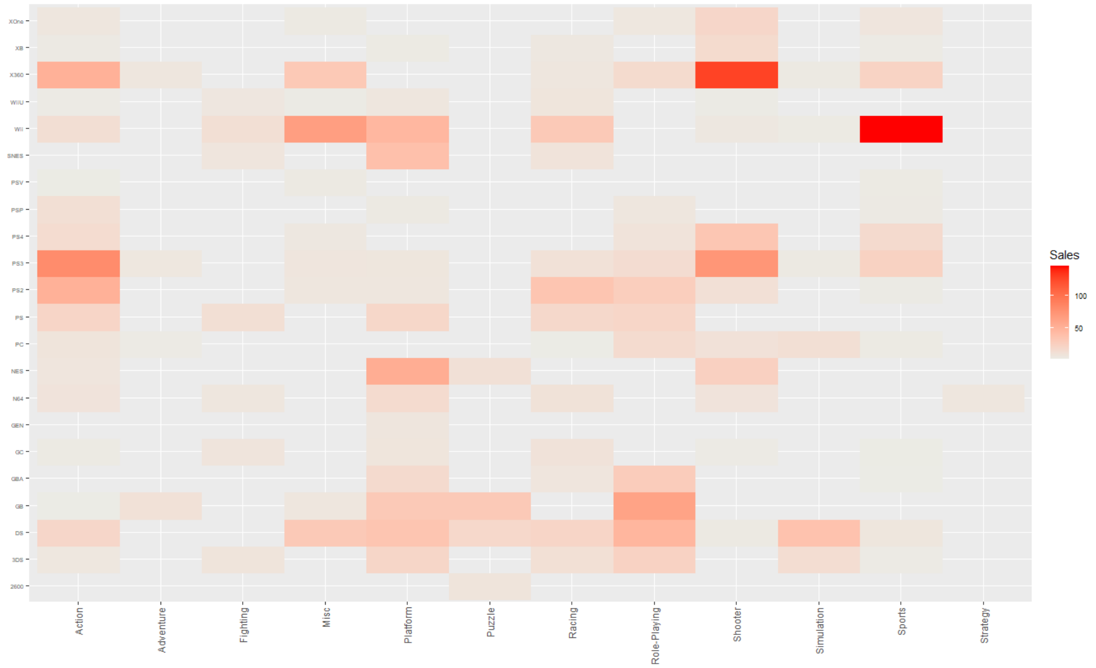
- **Use histrograms to show the global sales per game genre**



Top 5 genres are shooter, platform, action, role-playing and sports. Since the quantity sold information is not provided, one of possibilities for higher sales is higher selling price.

16: global sales per game genre

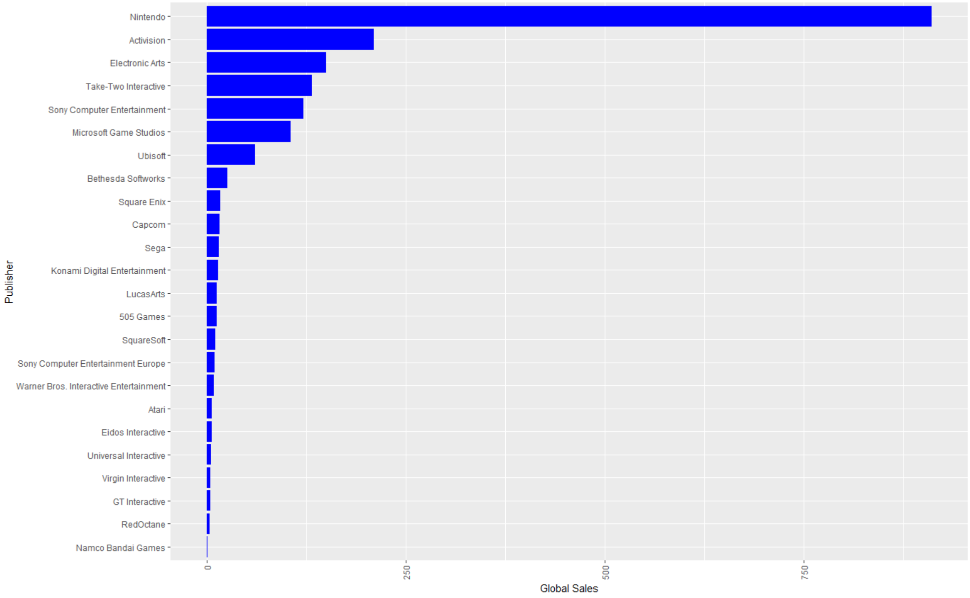
- **Use heatmap to show the global sales for each platform in each genre**



Top 2 are Wii-Sports and X360-Shooter. It may reflected that game experience of certain genres are much better than using other platform

17: global sales per publisher

- **Use histrogram to show the global sales per publisher**



Nintendo is the top publisher.

Section III – Recommendation

Increase customer loyalty by increase spending frequency, amount and willingness.

- **Focus on DIY market and improve quality**

  Target mass market and promote make DIY as gift to beloved
- **Develop product line of therapy and counselling usage**

  They concerns functions and less price-sensitive, high opportunity to re-purchase
- **Focus on selling video games of Nintendo, Xbox, PlayStation. Also selling console**

  Video-gamers are loyal with platform