Data Analysis carried out for Turtle Games, a game manufacturer and retailer with a global customer base, to improve overall sales performance by utilizing customer trends.

Predicting Future Outcomes

Turtle Games

Afshar Sanam

Friday, May 5, 2023

Table of Contents

1. Problem Statement 2

2. Analytical Approach 3

3. Visualization and Insights 4

3.1. How customers accumulate loyalty points 4

3.2 How groups within the customer base can be used to target specific market segments 6

3.3 How social data (e.g. customer reviews) can be used to inform marketing campaigns 11

4. Patterns and Predictions 18

4.1. QQ Plot for Sales data 18

4.2 Correlation plot between Sales data 19

4.3 Sales data clustering 20

4.10 Platform type grouping 20

4.4 Ring Plot for Genre 21

4.5 Analysis of Sale 22

4.6 Confidence level 25

4.7 Global Sales and Platform Type 26

4.8. Scatter Plots for Sales data 27

4.9 Histogram and Boxplots for Product and Platform 29

4.10 Global Sales against Product and Genre 30

4.11 Global sales, Product and Publisher 30

4.12. Global Sales density v/s NA Sales density and EU Sales density 31

4.12. Global Sales / Year + Residuals 32

4.13 EU Sales / Year + Residuals 33

4.14 NA Sales / Year + Residuals 34

4.15 Sales Comparison 35

Summary 36

1. Problem Statement

Data Analysis carried out for Turtle Games, a game manufacturer and retailer with a global customer base. The company manufactures and sells its own products, along with sourcing and selling products manufactured by other companies. Its product range includes books, board games, video games, and toys. The company collects data from sales as well as customer reviews. Turtle Games has a business objective of improving overall sales performance by utilizing customer trends.

The most important Questions to address are:

a. How customers accumulate loyalty points
b. How groups within the customer base can be used to target specific market segments
c. How social data (e.g. customer reviews) can be used to inform marketing campaigns
d. The impact that each product has on sales
e. How reliable the data is (e.g., normal distribution, skewness, or kurtosis)
f. What the relationship(s) is/are (if any) between Northern American, European, and Global Sales?

This analysis uses the

- `turtle_reviews.csv` - Details on customer gender, age, remuneration, spending score, loyalty points, education, language, platform, review and summary across products.

- `turtle_sales.csv` - Details of video games sold globally, such as the rank, product, platform, genre, publisher, and their sales across North America, Europe, and worldwide.

- `metadata_turtle_games.txt` – Details of the data set, data quality and reference.

The insights gained from the analysis will inform the Turtle Games for their decision-making so they can derive the next course of action(s)

2. Analytical Approach

- Applied Linear regression using Python for data analysis of social media data. For this, cleansed and sense-checked the data and checked for missing values (if any) in the dataframe and created a summary of descriptive statistics. Also, removed redundant columns and renamed the columns for ease of reference and finally saved/exported the cleaned dataframe into CSV file.
- Clustering using k-means: leveraged both elbow and silhouette methods to determine the optimal cluster number
- NLP using Python: Sense-checked the dataframe and determined the missing values. Prepare the data for NLP; applied tokenisation and created wordclouds. Next, checked for frequency distribution and polarity and removed alphanumeric characters and stopwords and identified 15 most common words. For sentiment analysis, extracted subjectivity score(s) apart from polarity score and derived the mean sentiment score. Performed Sentiment Analyzer using Vader sentiment and subsequently used Text Blob in extracting the percentage of positive, negative and neutral scores and built histograms for sentiment polarity scores
- Leveraged EDA using R in exploring sales data and built various plots to determine the insights into data set. Also, computed normal distribution of the data and performed Shapiro-Wilk test on all the sales data. Determined Skewness and Kurtosis of all the sales data and subsequently determined correlation between the sales data columns.
- Applied simple and multiple regression models in determining correlation between the sales columns. However, predictions were excluded

3. Visualization and Insights

3.1. How customers accumulate loyalty points

Out[397]:

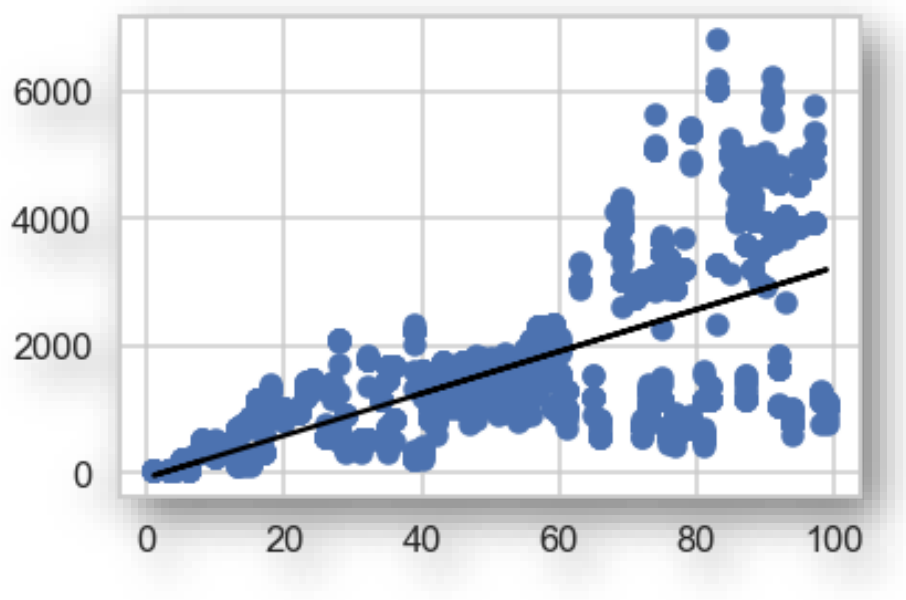
	age	remuneration (k£)	spending_score (1-100)	loyalty_points	product
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	39.495000	48.079060	50.000000	1578.032000	4320.521500
std	13.573212	23.123984	26.094702	1283.239705	3148.938839
min	17.000000	12.300000	1.000000	25.000000	107.000000
25%	29.000000	30.340000	32.000000	772.000000	1589.250000
50%	38.000000	47.150000	50.000000	1276.000000	3624.000000
75%	49.000000	63.960000	73.000000	1751.250000	6654.000000
max	72.000000	112.340000	99.000000	6847.000000	11086.000000

```
(352, 9)
```

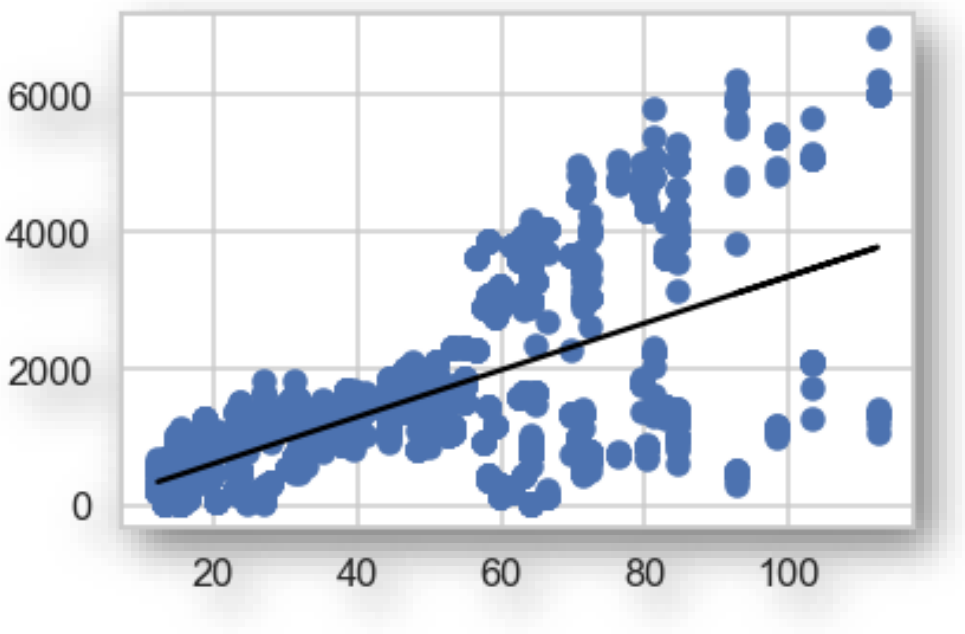
Higher the remuneration; higher the spending score and hence higher the loyalty points accumulated:

1. Average age: 39 years; remuneration: 48.07k£; spending score: 50; loyalty points: 1578
2. Min age: 17 yrs; remuneration: 12.3k£; spending score: 1; loyalty points: 25
3. Max age: 72 yrs; remuneration: 63.96k£; spending score: 73; loyalty points: 1751.25
4. Total sum of loyalty points: 3156064
5. The maximum value of Loyalty Points = 6847
6. The maximum value of Loyalty Points = 6847
7. Interquartile range of Loyalty Points = 979.25

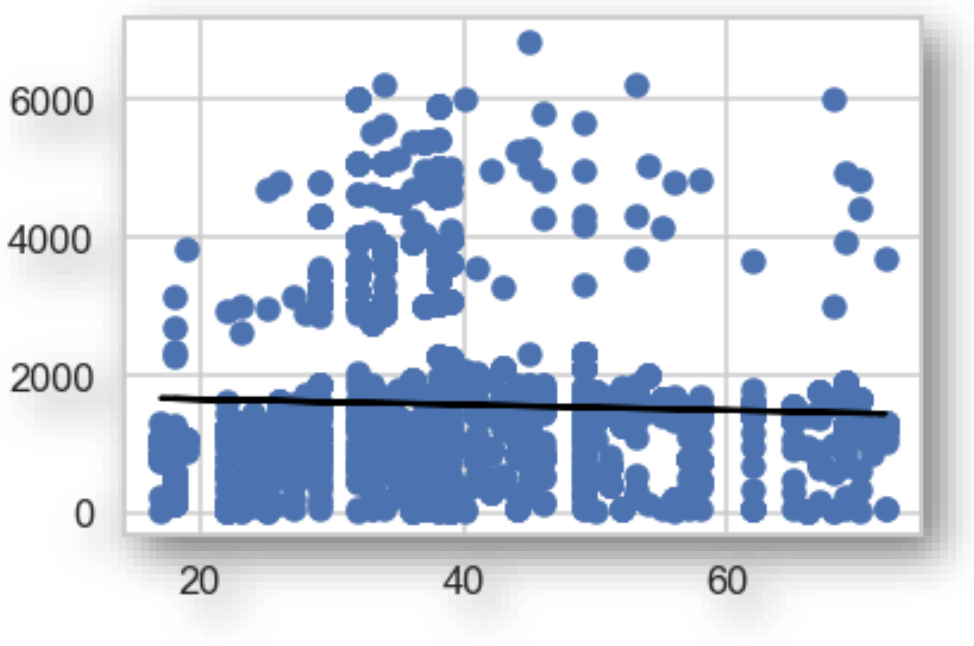
Spending v/s Loyalty (Linear regression): Higher the Spending score -> Higher the Loyalty Points



Remuneration v/s Loyalty (Linear regression): Higher the Remuneration -> Higher the Loyalty Points



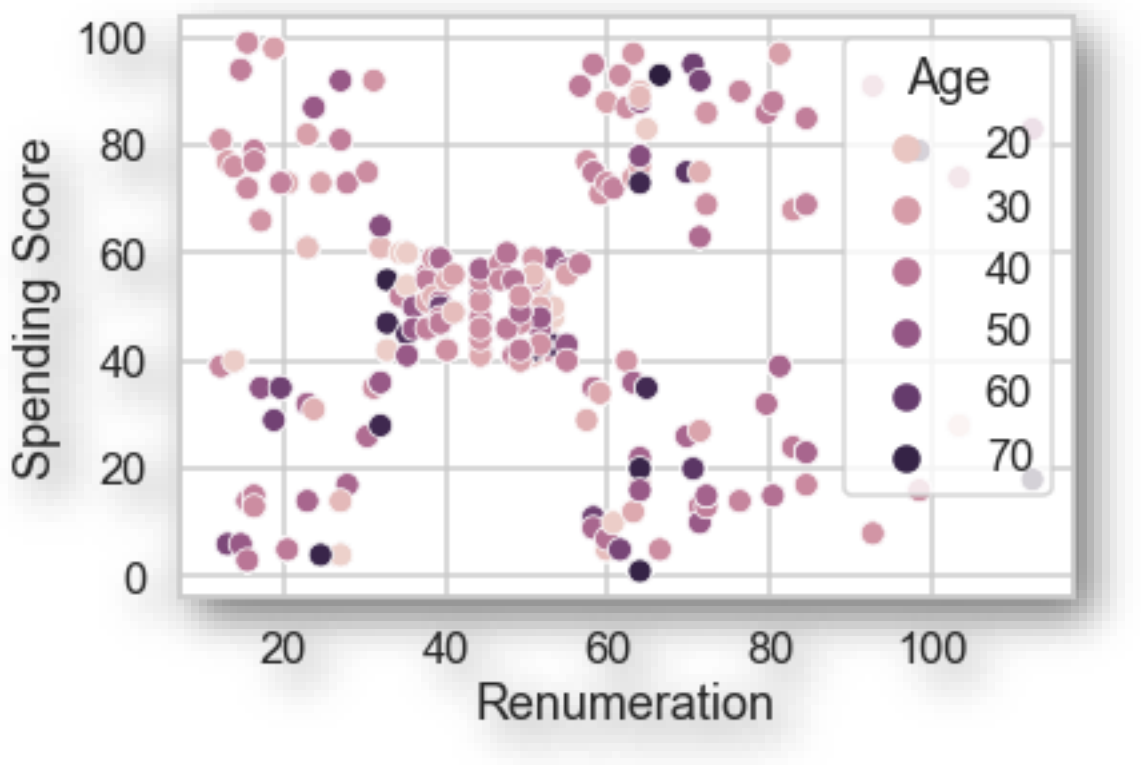
Age v/s Loyalty (Linear Regression): constant (not much of a difference): Need to further analyse



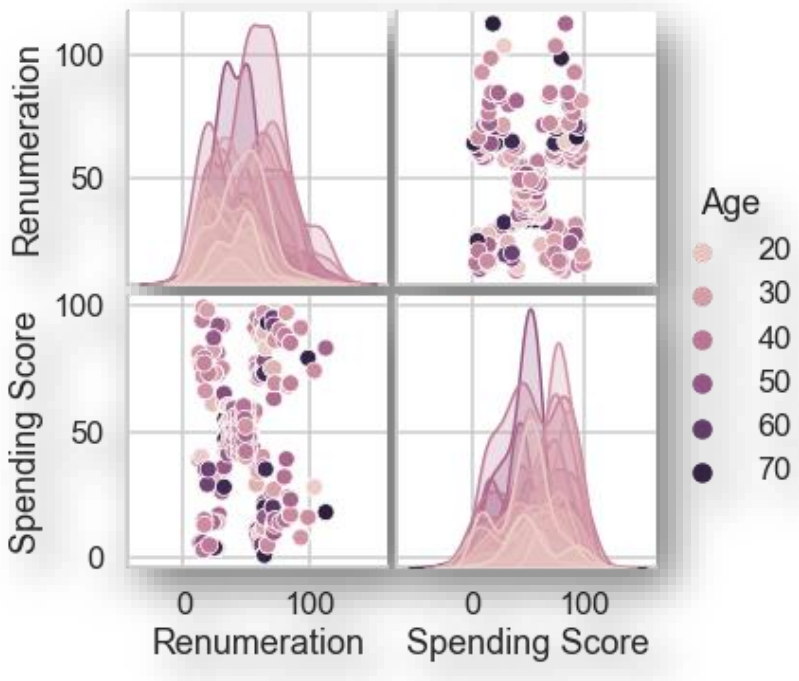
3.2 How groups within the customer base can be used to target specific market segments

To identify the proper clustering of the age group and how relevant remuneration and spending scores are in deriving the target market segments is the key objective of this exercise:

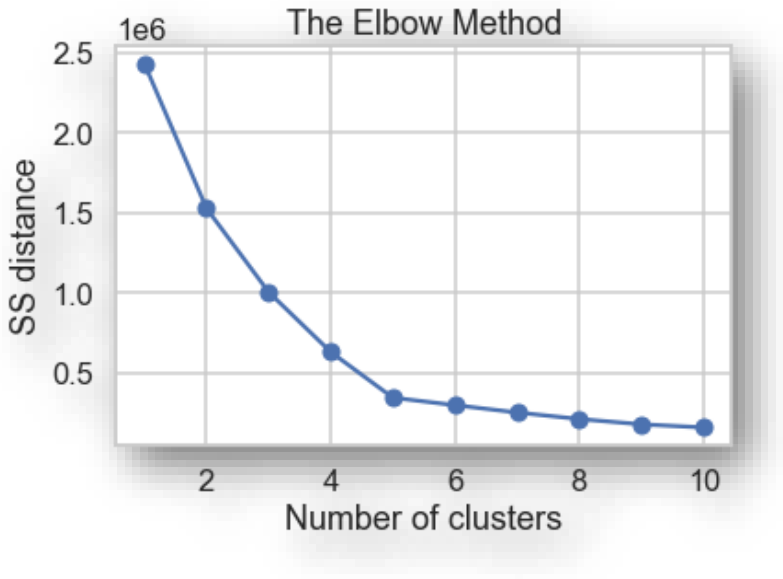
- Per the analysis, was able to identify the customer base as per the age, remuneration and spending scores as the key attributes to focus
    - the average age group is 39; average remuneration is 48 k£; average spending score is 50
    - the min age group is 17; min remuneration is 12.3 k£; min spending score is 1
    - the max age group is 72; max remuneration is 112.34 k£; max spending score is 99



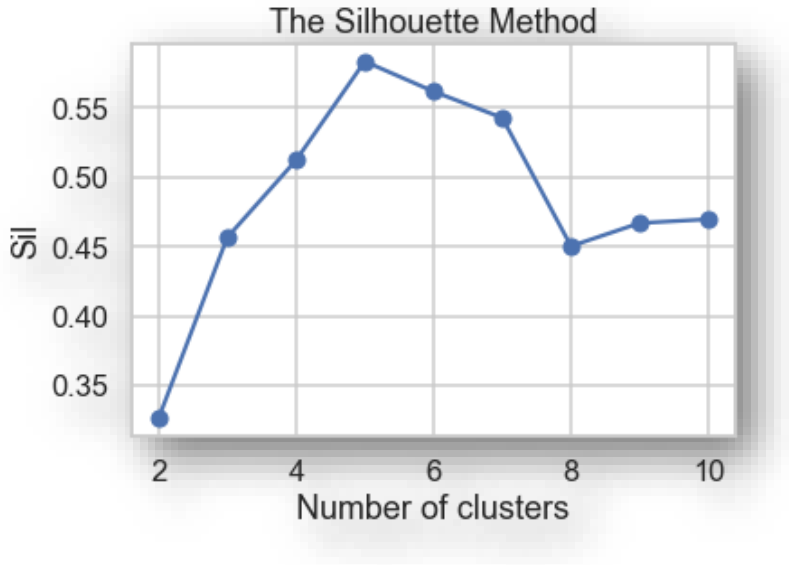
Normal scatterplot derived 6-age groups: 20, 30, 40, 50, 60 and 70 years



- Pair plot generated the combination of spending score + remuneration; remuneration + age; age + spending score and spending score + remuneration + age as 4 categories which needs to be further analysed to derive proper clustering
- K-means clustering was carried out in deriving number of optimal clusters:
    - Elbow method: 5 clusters (WSS value started to diminish)



    - Silhouette method: 5 cluster ranges (averages intra-cluster distance) as the peak in the plot



- After evaluating as per the scatter plot grouping of 6 clusters; the spending scores were skewed at an average age group is 39; average remuneration is 48 k£; average spending score is 50 which was the same.

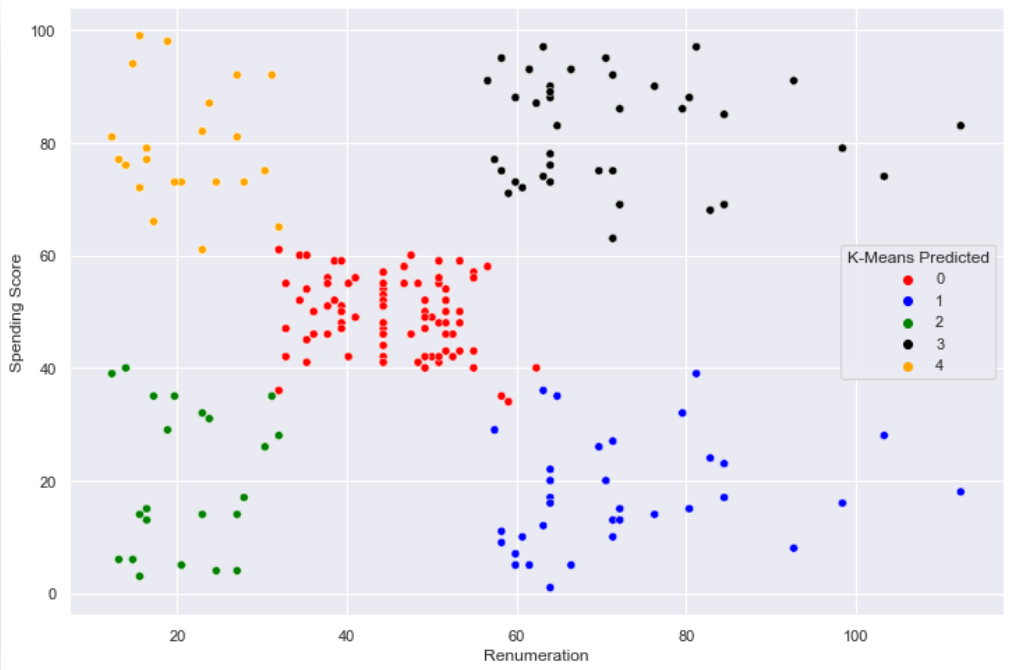
VALUE COUNTS: Clusters to target!¶

- Cluster 0 count: 774
- Cluster 3 count: 356
- Cluster 1 count: 330
- Cluster 2 count: 271
- Cluster 4 count: 269

	Renumeration	Spending Score	K-Means Predicted
0	12.30	39.0	2
1	12.30	81.0	4
2	13.12	6.0	2
3	13.12	77.0	4
4	13.94	40.0	2

Out[484]:

	Renumeration	Spending Score	K-Means Predicted
count	2000.000000	2000.000000	2000.000000
mean	48.079060	50.000000	1.508000
std	23.123984	26.094702	1.479199
min	12.300000	1.000000	0.000000
25%	30.340000	32.000000	0.000000
50%	47.150000	50.000000	1.000000
75%	63.960000	73.000000	3.000000
max	112.340000	99.000000	4.000000



3.3 How social data (e.g. customer reviews) can be used to inform marketing campaigns

Summary Frequency Distribution data...

TOP 15 words: FreqDist({'game': 268, 'great': 237, 'fun': 175, 'good': 84, 'love': 70, 'like': 54, 'kids': 48, 'book': 42, 'expansion': 42, 'cute': 40, ...})

Reviews Frequency Distribution data...

TOP 15 words: FreqDist({'game': 1360, 'one': 475, 'play': 442, 'fun': 407, 'great': 392, 'like': 373, 'get': 293, 'cards': 281, 'tiles': 280, 'really': 278, ...})

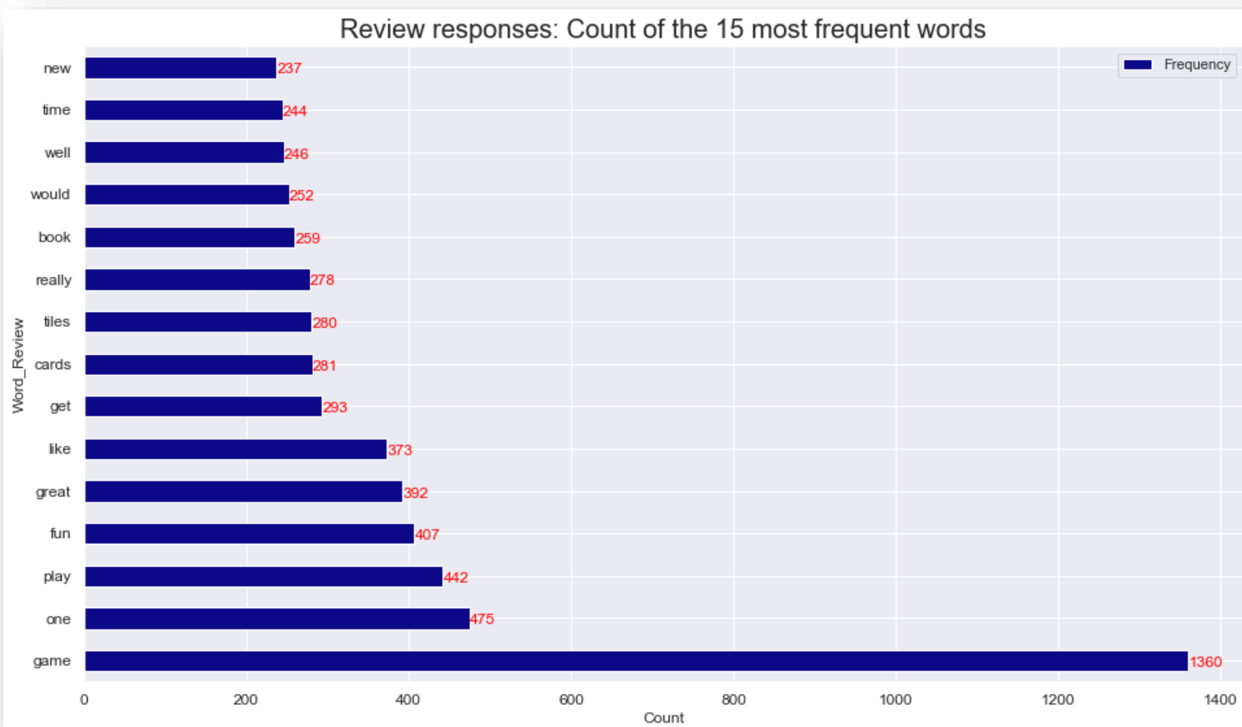
Out[457]:

Word_Review	Frequency
game	1360
one	475
play	442
fun	407
great	392
like	373
get	293
cards	281
tiles	280
really	278
book	259
would	252
well	246
time	244
new	237

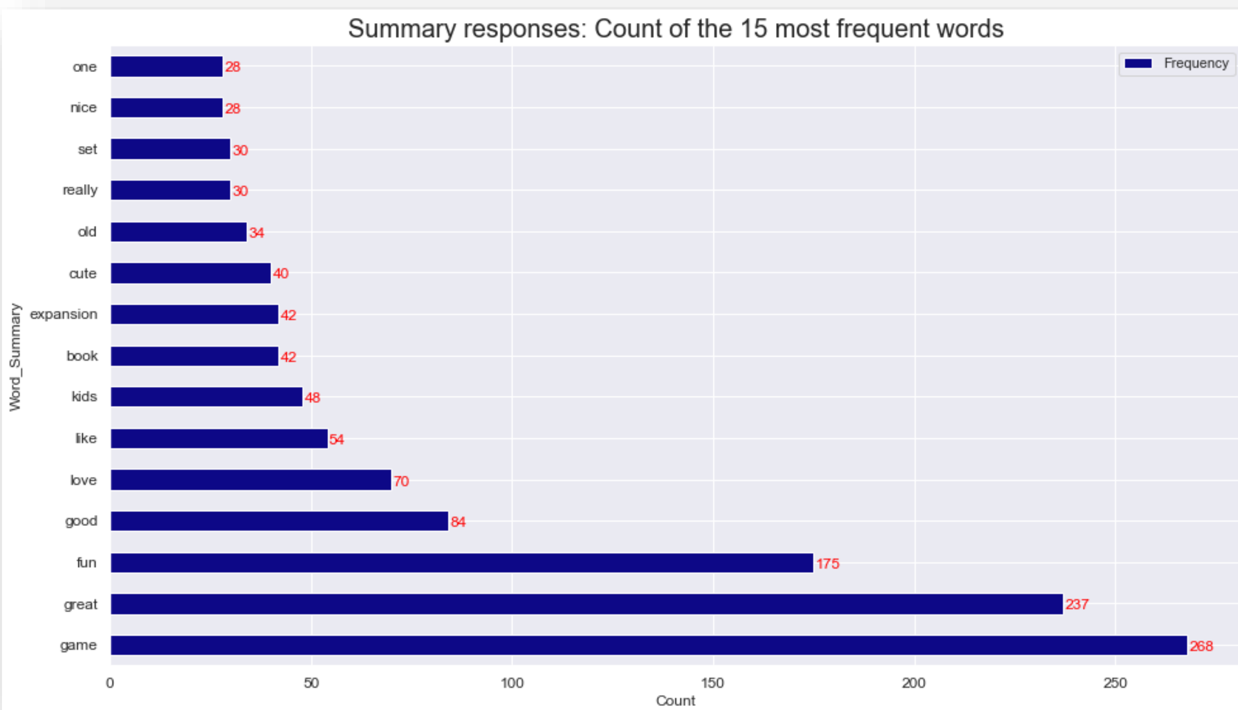
Out[459]:

Word_Summary	Frequency
game	268
great	237
fun	175
good	84
love	70
like	54
kids	48
book	42
expansion	42
cute	40
old	34
really	30
set	30
nice	28
one	28


game
one
play
card
great
fun
tile
book
make
time
love
piece
player
new
set
well
use
really
little
good
expansion


Review responses: Count of the 15 most frequent words
Frequency
Word_Review
Count
new 237
time 244
well 246
would 252
book 259
really 278
tiles 280
cards 281
get 293
like 373
great 392
fun 407
play 442
one 475
game 1360
0
200
400
600
800
1000
1200
1400


game
great
fun
good
expansion
love
really
little
awesome
nice
playbook
product
quality
excellent
price
kid
one
make
giftset
cute
tile
best
year old
time
sticker
puzzle
family
better
acquire
made
toy
easy
much
age
board game
perfect
loved
pretty
lords waterdeep
kids love
addition
enjoyed
scrabble
fantastic
old lady
new
craft
use
adorable
useful
beautiful
playing
expected
strategy
card
cheap
buy
picture
worth
dungeon
small
amazing
original
well
word game
monopoly
therapy
happy
children
wonderful
creative
another
version

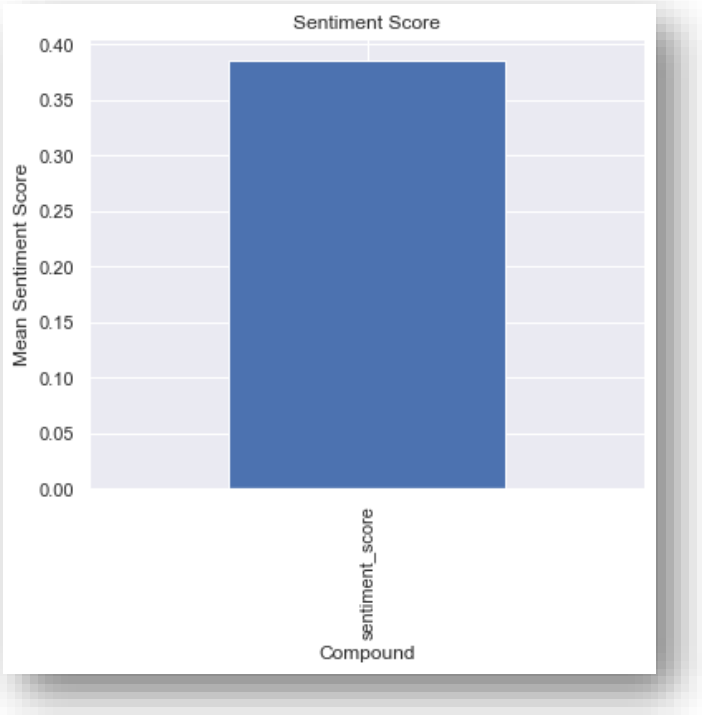

Summary responses: Count of the 15 most frequent words
Frequency
Word_Summary
Count
one 28
nice 28
set 30
really 30
old 34
cute 40
expansion 42
book 42
kids 48
like 54
love 70
good 84
fun 175
great 237
game 268
0
50
100
150
200
250

- **Polarity and Sentiment analysis:**

Out[497]:

	sentiment_score
count	1351.000000
mean	0.385885
std	0.347644
min	-0.905200
25%	0.000000
50%	0.510600
75%	0.624900
max	0.952400

  - The mean sentiment score is 0.385
  - 75% sentiment score is 0.62
  - Max sentiment score if 0.95

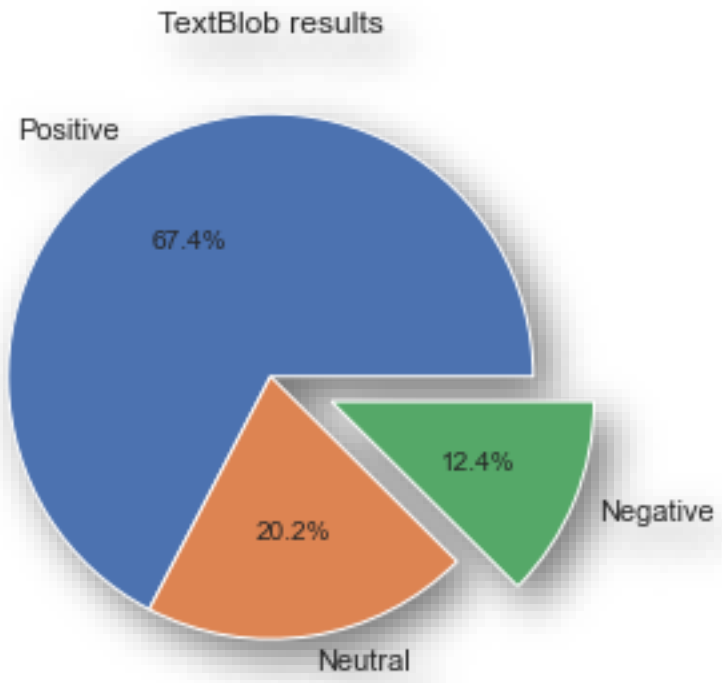


Sentiment Polarity Score of Reviews: Its mostly positive trend

- Polarity lies between -1 and +1 (-1 defines negative sentiment and 1 defines positive sentiment)

- Subjectivity lies between [0, 1]: (0 – objective; 1 – subjective)Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher the subjectivity means that the text contains personal opinion rather than factual information.

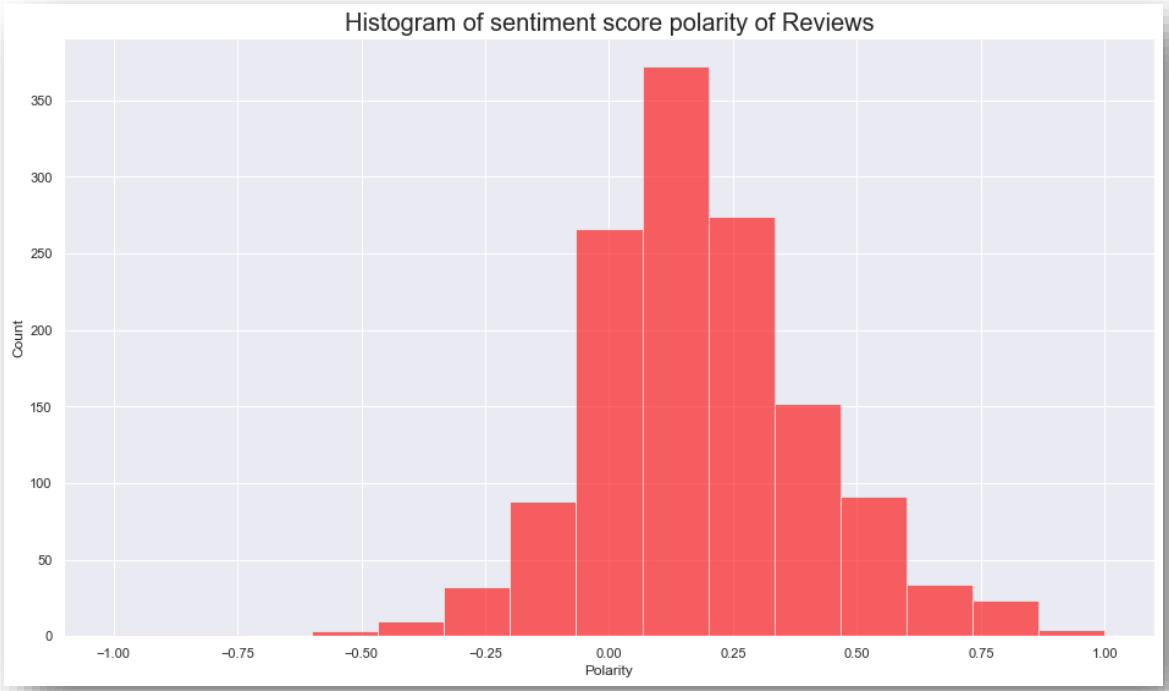
	review	summary	Polarity	Analysis
0	when it comes to a dms screen the space on the...	the fact that 50 of this space is wasted on ar...	0.15	Positive
1	an open letter to galeforce9 your unpainted mi...	another worthless dungeon masters screen from ...	-0.80	Negative
2	nice art nice printing why two panels are fill...	pretty but also pretty useless	0.00	Neutral
3	amazing buy bought it as a gift for our new dm...	five stars	0.00	Neutral
4	as my review of gf9s previous screens these we...	money trap	0.00	Neutral

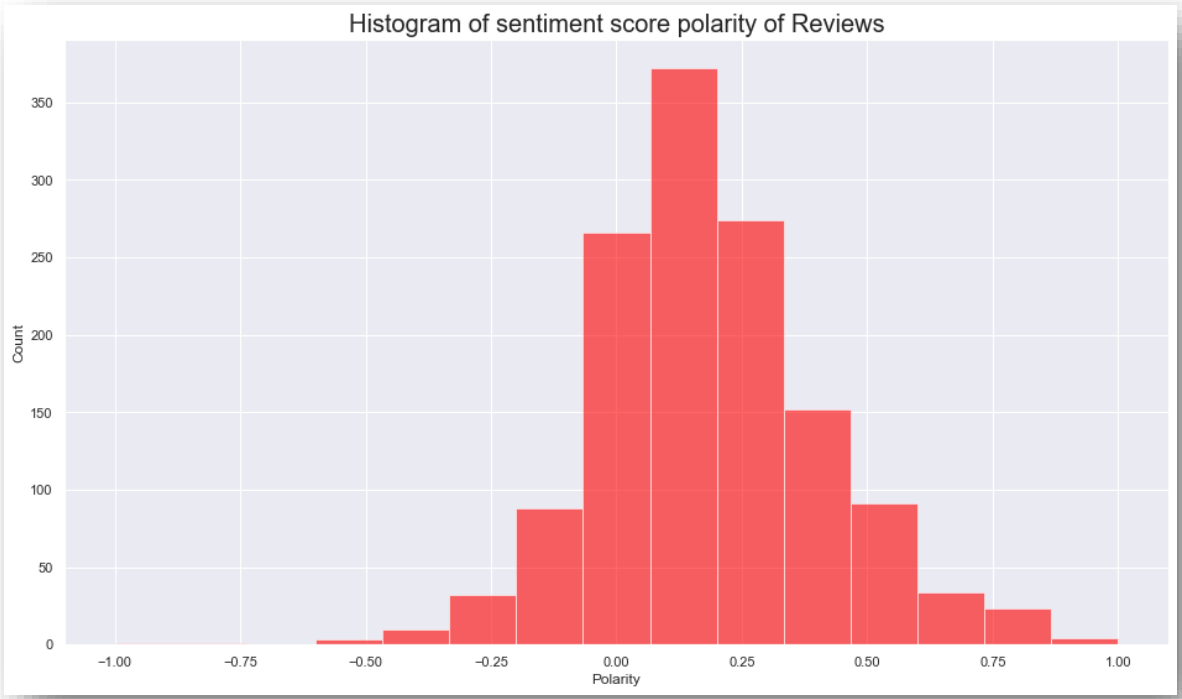


- The Text Blob pie chart shows **67.4%** Positive Polarity Analysis for both Summary & Reviews
- 20.2% Neutral Polarity Analysis and
- 12.4% Negative Polarity Analysis

Sentiment Score Polarity of Review and Summaries: Its positive trend

`





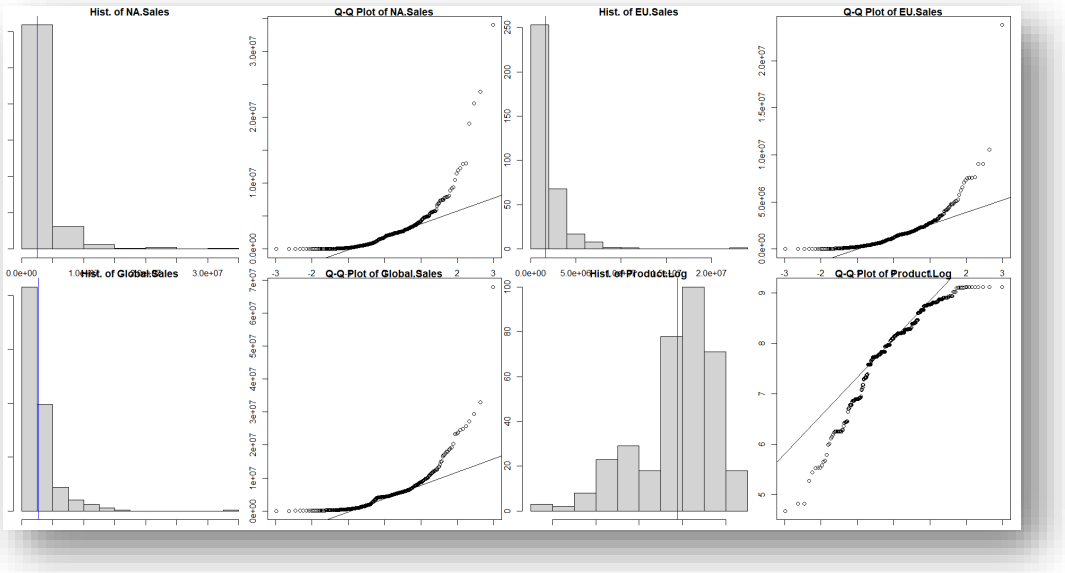
- Extract negative summaries (Top 20)
    - 8 out of 20 had personalized subjectivity scores (>0.5 or >50%)
- Extract negative reviews
    - 18 out of 20 had personalized subjectivity scores (>0.5 or >50%)
- Extract positive summaries
    - 9 out of 10 scores (polarity and subjectivity) were >0.75 to 1 or 75% to 100%
- Extract positive reviews
    - 10 out of 10 scores (polarity and subjectivity) were >0.8 to 1 or 80% to 100%

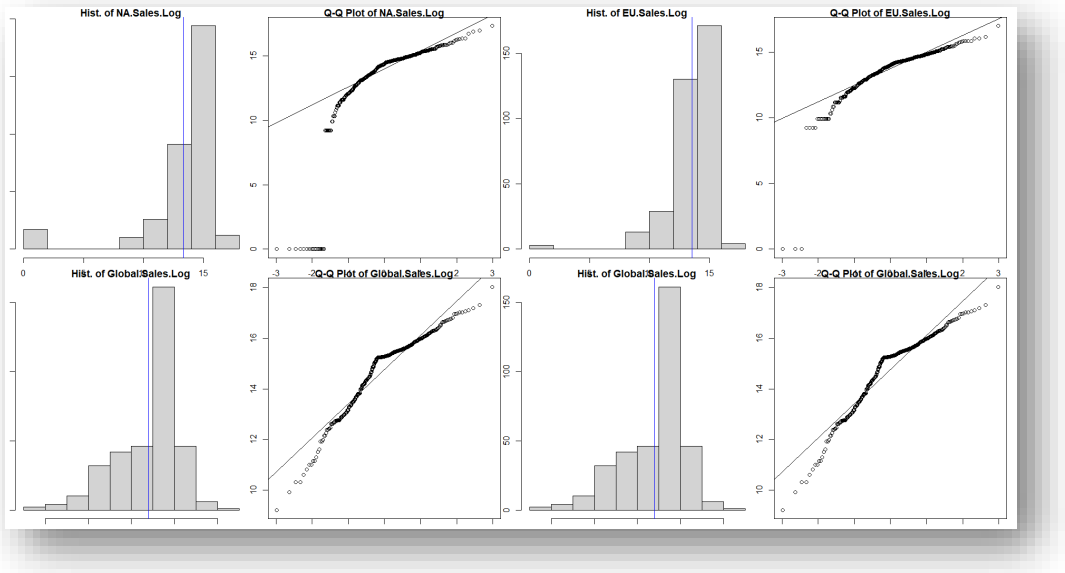
However, NPS scores are equally important to be derived as there are 20.2% neutral and 12.4% negative (detractors) and need to convert more promoters to improve the marketing strategy!

4. Patterns and Predictions

4.1. QQ Plot for Sales data

Pre-analysis shows that the variables (Global.Sales, NA.Sales, EU.Sales) are not normally distributed -> logs to transform and combine with the original variables. The above diagram is the histogram and QQ plot for the transformed data set. The original Histograms and qq plots of sales show abnormal distributed, but the log values of these sales are much close to normal distribution, especially the log value of global sales. The data set shows the distribution of global and NA and EU sales, their relationship. The other factors relating to genre, platform and publisher is being derived subsequently.

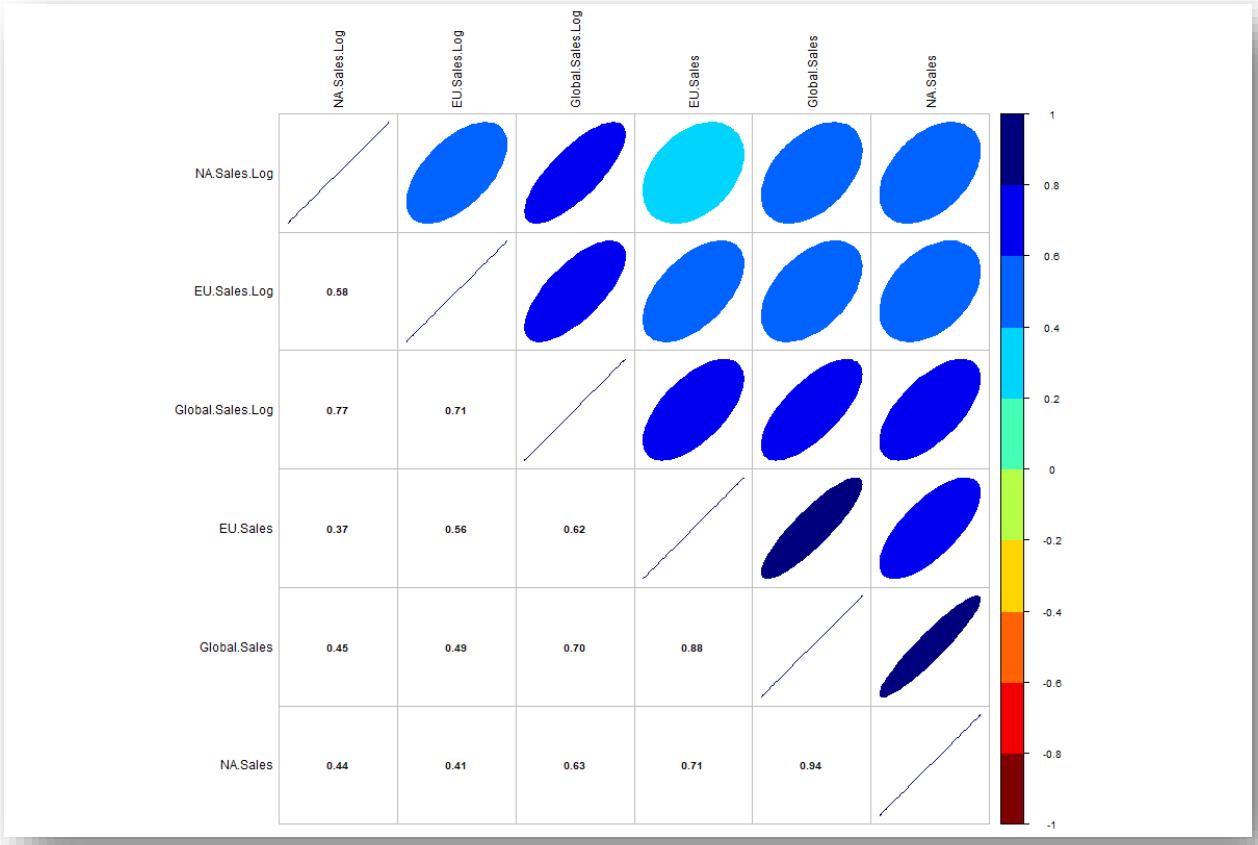




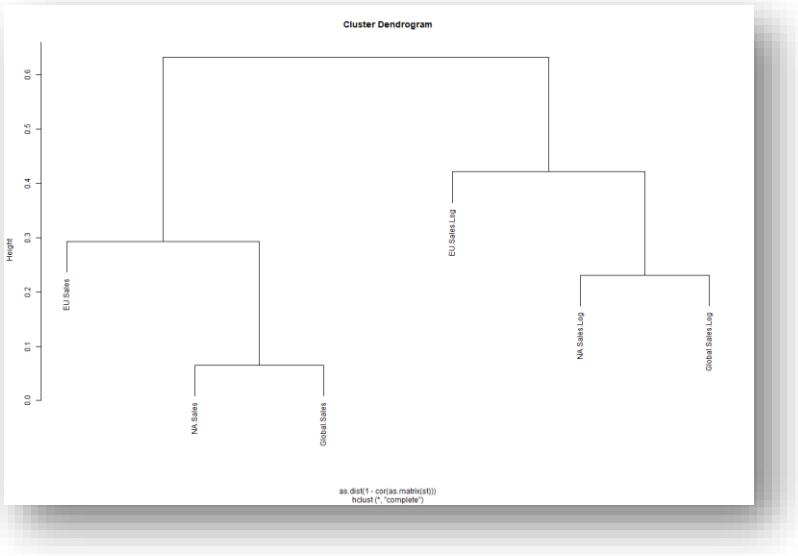
4.2 Correlation plot between Sales data

The correlation plot between Global Sales & Regional Sales Logs. There are high r values of 0.94, 0.88, 0.77, 0.71, 0.70 between the values of:

- Global Sales and NA.Sales: 0.94
- Global Sales and EU Sales: 0.88
- Global Sales Log and NA Sales Log: 0.77
- NA Sales and EU Sales (Global Sales Log and EU Sales Log): 0.71
- Global Sales and Global Sales Log: 0.7



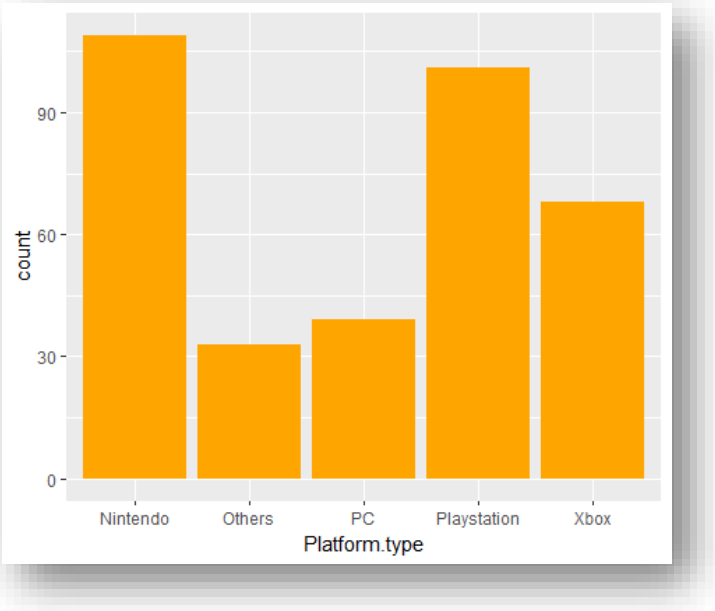
4.3 Sales data clustering



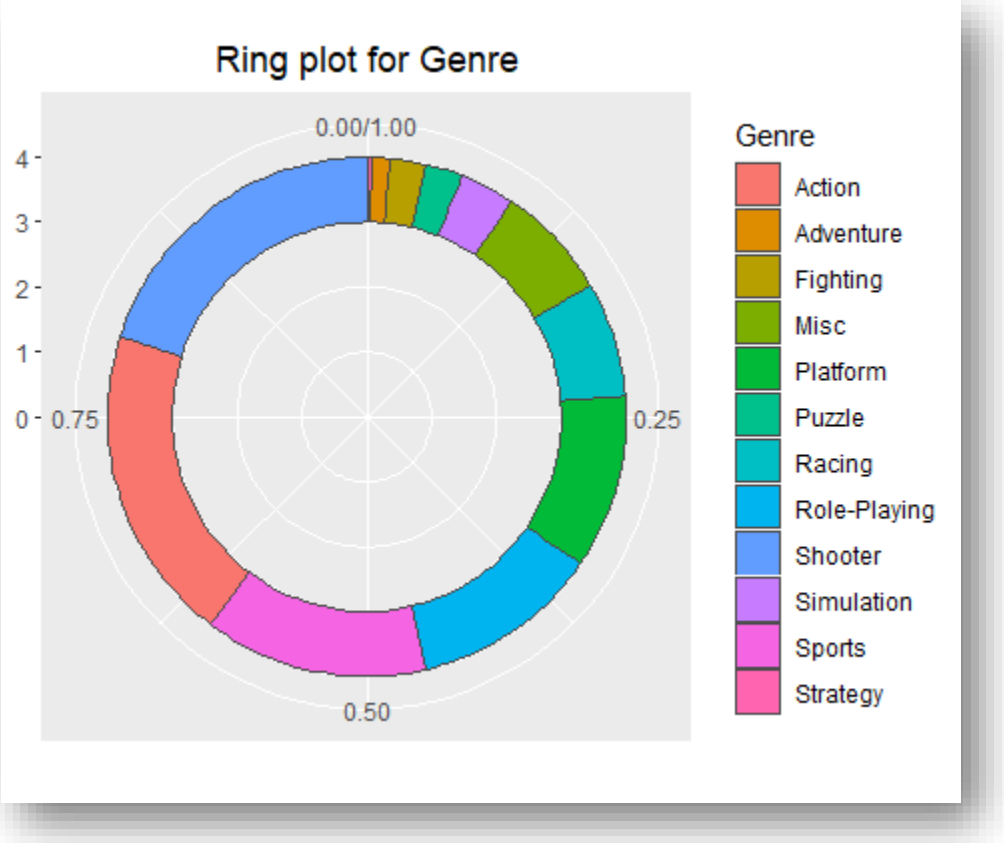
- NA Sales Cluster is closest to Global Sales Cluster and then to EU Sales Cluster
- NA Sales Log Cluster is closest to Global Sales Log Cluster and then to EU Sales Log Cluster

4.10 Platform type grouping

The Platform type has been regrouped for simplification. Per the graph, Nintendo is the biggest group, next is Playstation and the third is Xbox. The smallest is the “Others” category.



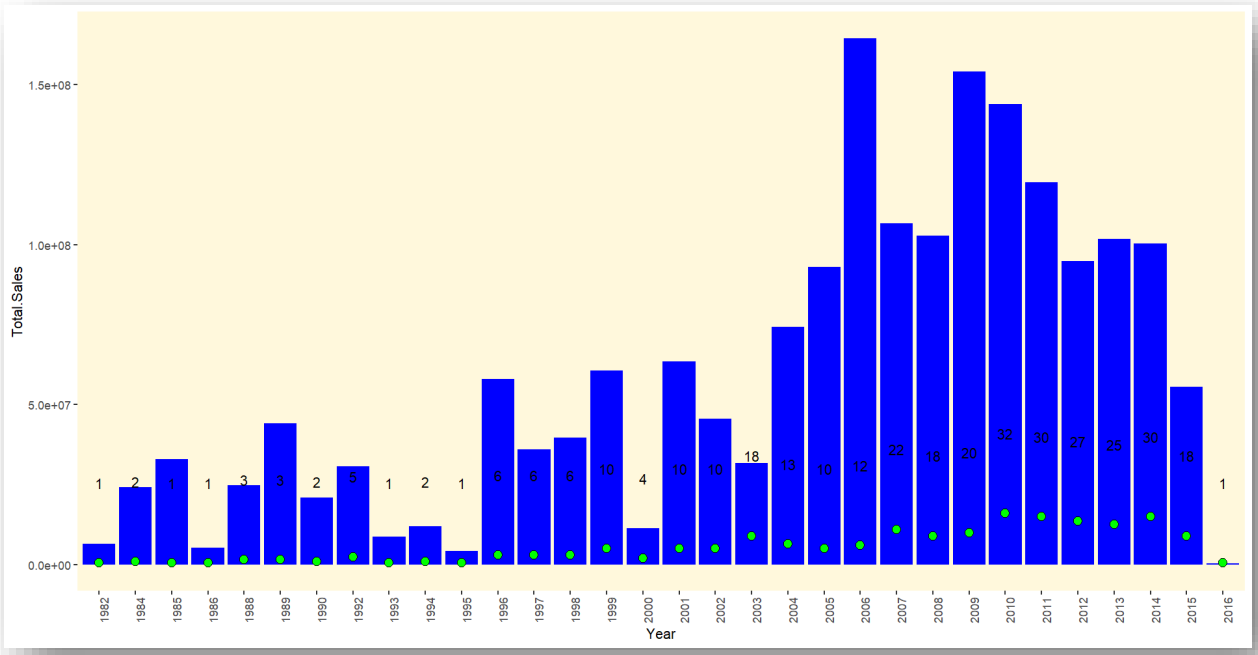
4.4 Ring Plot for Genre



Shooter, Action, Sports are the top 3 biggest genres. Shooter occupies close to 25% of genre. The top 3 genres contribute to over half of the genre count. The bottom 3 are Adventure, Fighting and Puzzle

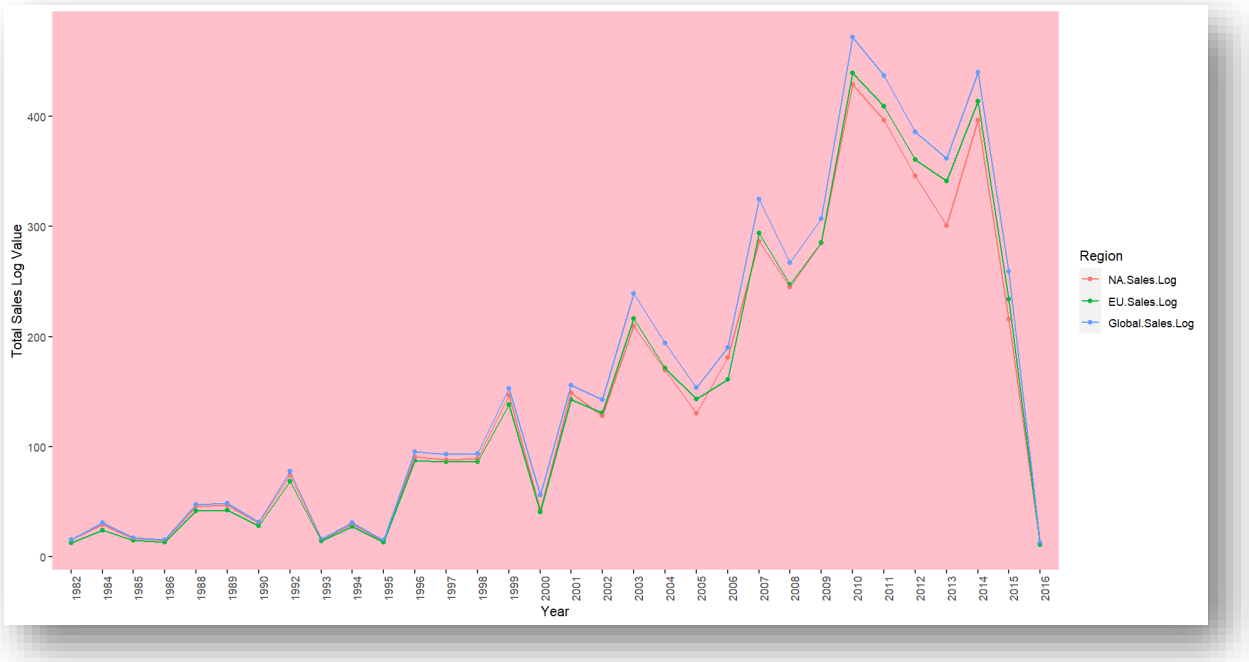
4.5 Analysis of Sale

1. By Year:



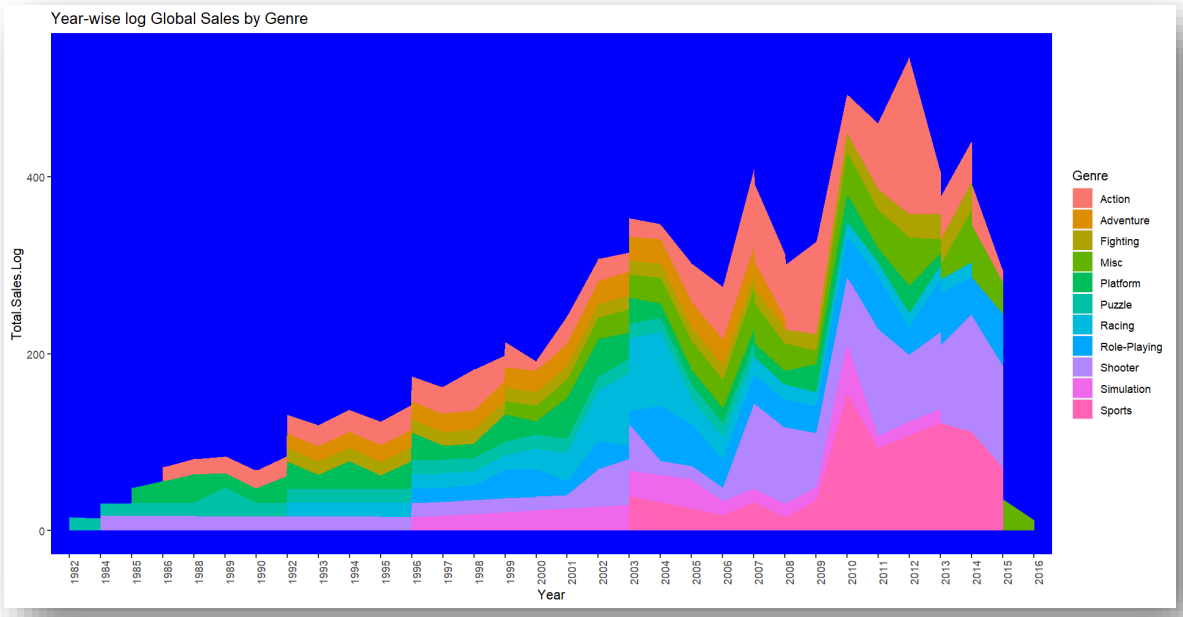
The global sales has seen a slow pickup since 1992 with less than 10 games until 1999 and there was a slump in the sales in 2000 and then got picked-up back from 2001 onwards. **2006** saw a huge spike on the Total sales globally despite having just 12 games released to the market and then the sales went down by adding +10 more products and then 4 got retired and there was some change on the gaming strategy and the sales picked up again in 2009. However, 2016 has been a huge cut down on the no. of games by almost 50% and 2016 was just 1 !?!?! It's like gone back to 1992!!!!

2. By Region:



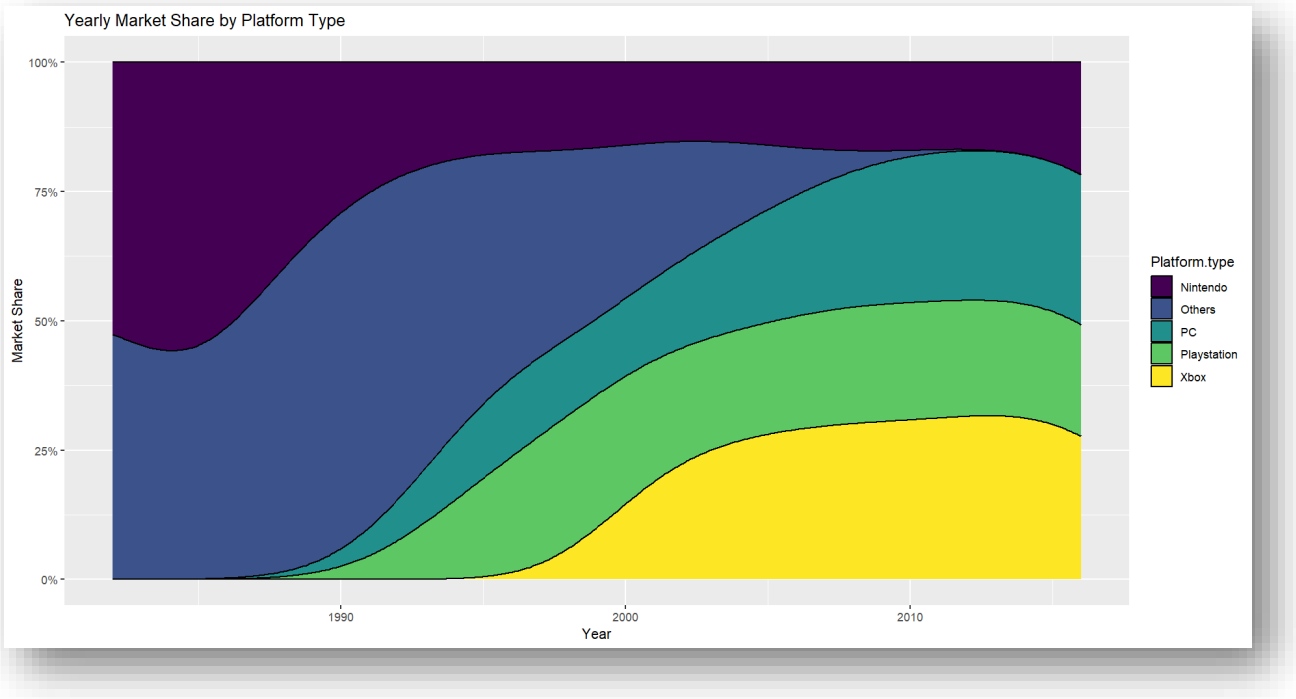
The pattern of log value for the regional sales are similar for Global, North America, Europe which matches the cluster analysis.

3. By Genre:



2010-2014 seems like a very good year for games. The games produced above 400 Total.Sales.Log in each of those years. Action games are on the top sale during this time, which contributed the biggest portion of the total global sales log. Adventure, Fighting and Puzzle are at the bottom.

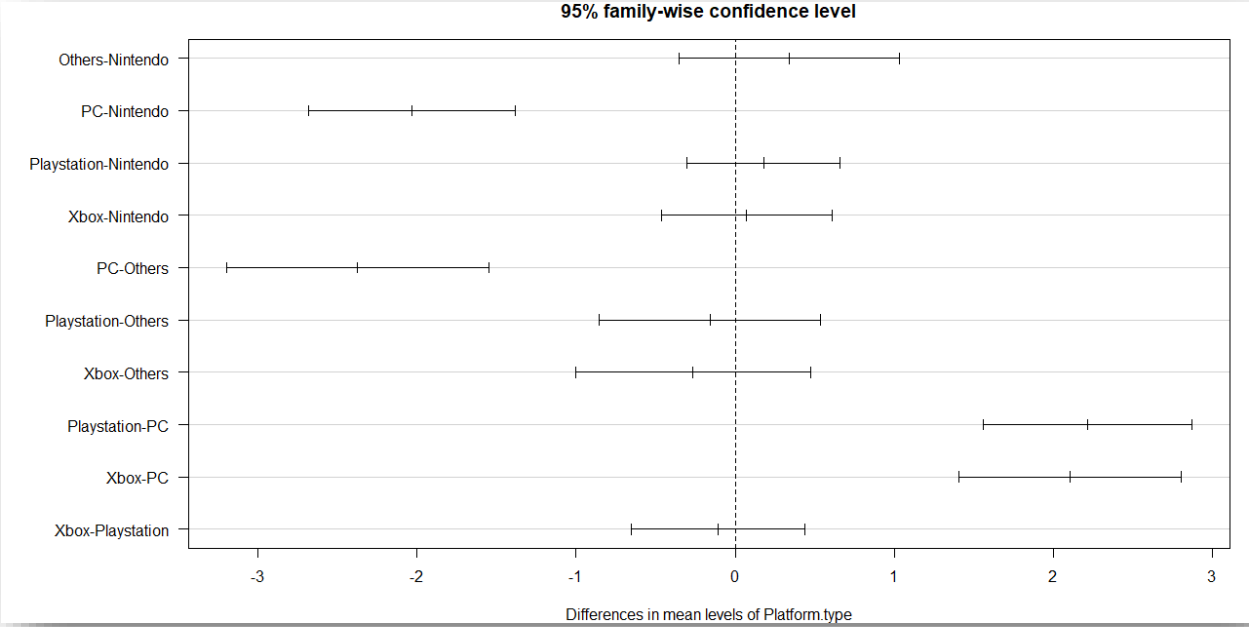
4. By Platform:



Xbox came after 1990 almost nearing 2000. However, Nintendo and Others were the main market players with PC being the key platform for gaming and got continued. However, the Others category got reduced during 2010 but Nintendo, PC, Playstation and Xbox continued with almost equal market share.

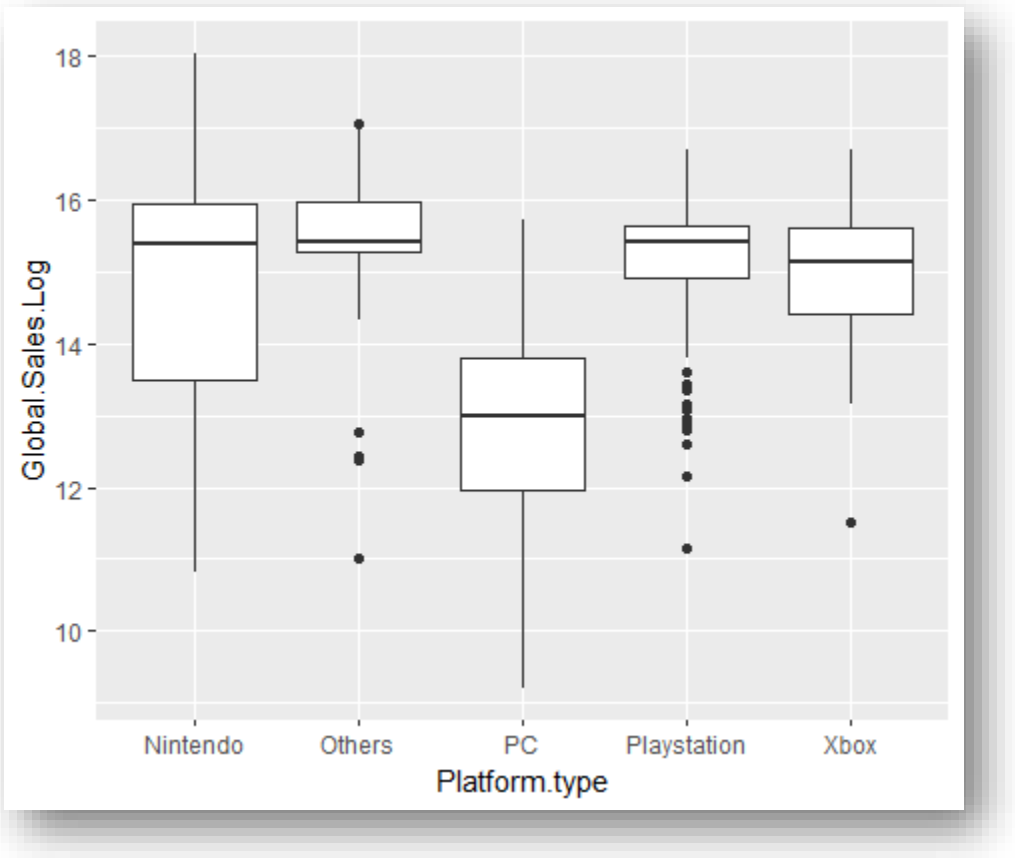
ANOVA test shows that at least one of the mean values of Global.Sales.Log for those platform types is significantly different from the others like PC-Nintendo; PC-Others; Platstation-PC and Xbox-PC. However, Others-Nintendo; Playstation-Nintendo; Xbox-Nintendo; Playstation-Others; Xbos-Others and Xbox-Playstation has the 95% confidence level

4.6 Confidence level



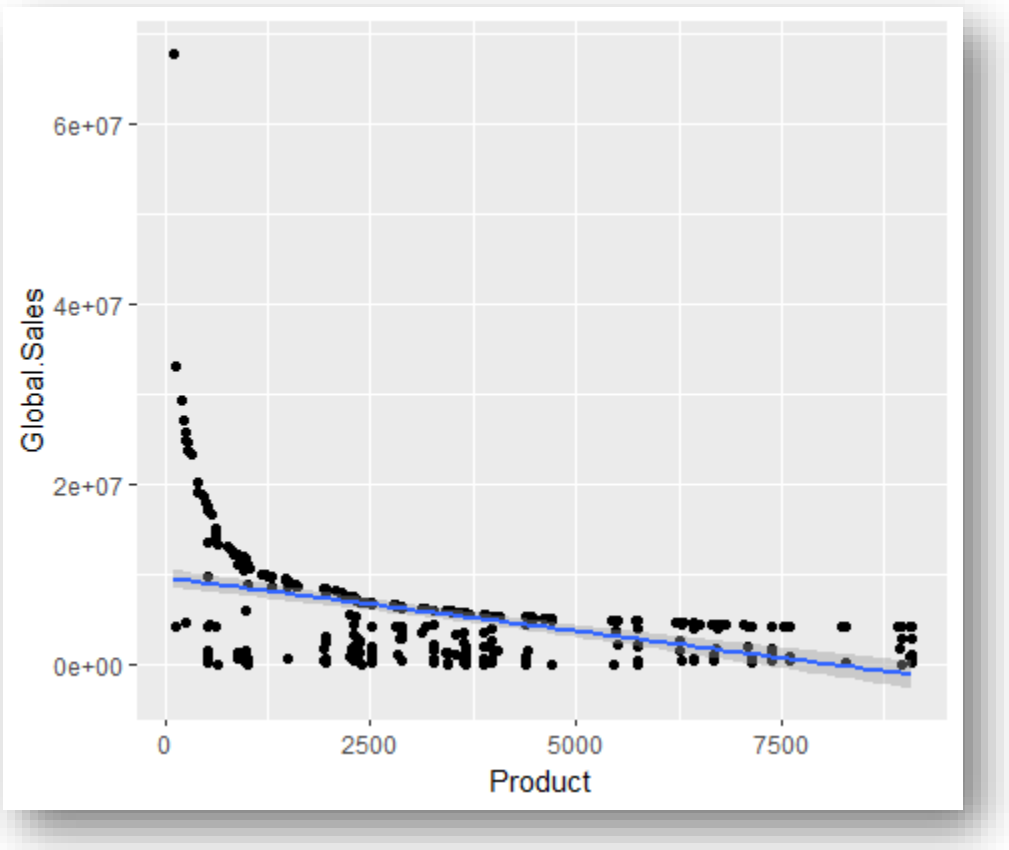
ANOVA test shows that at least one of the mean values of Global.Sales.Log for those platform types is significantly different from the others like PC-Nintendo; PC-Others; Platstation-PC and Xbox-PC. However, Others-Nintendo; Playstation-Nintendo; Xbox-Nintendo; Playstation-Others; Xbos-Others and Xbox-Playstation has the 95% confidence level

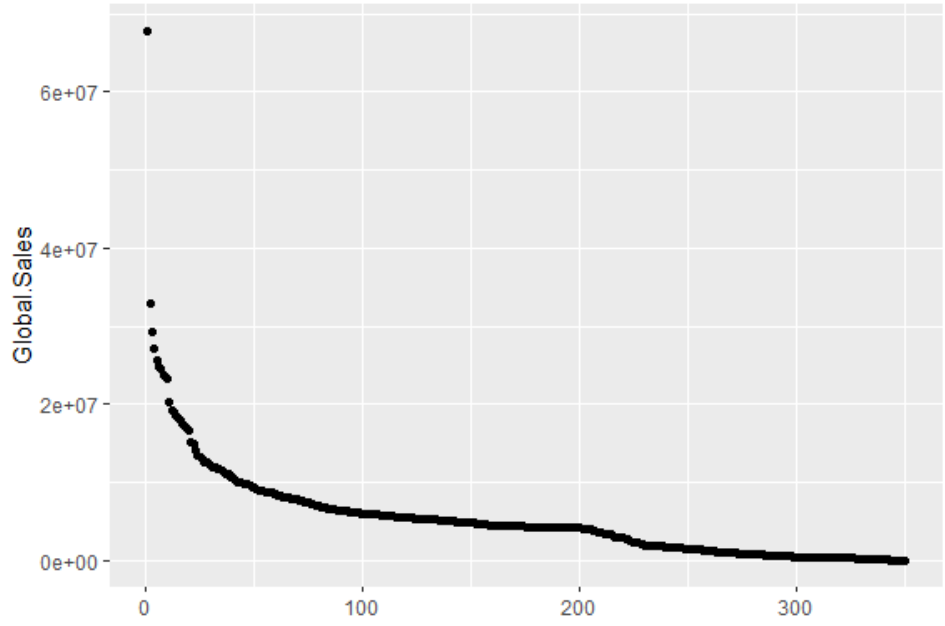
4.7 Global Sales and Platform Type

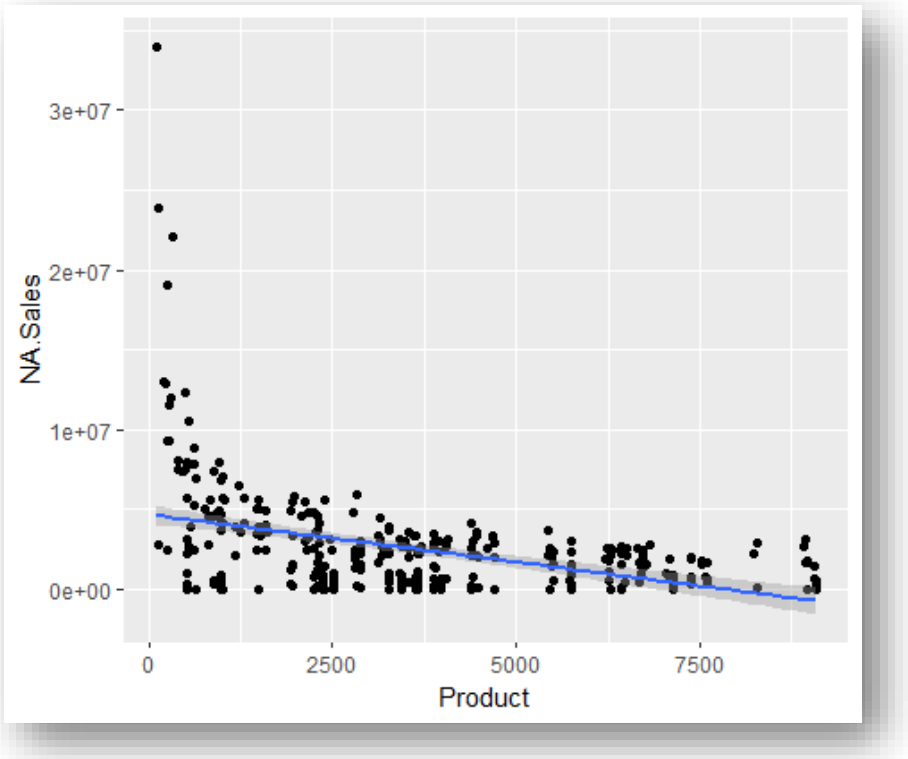


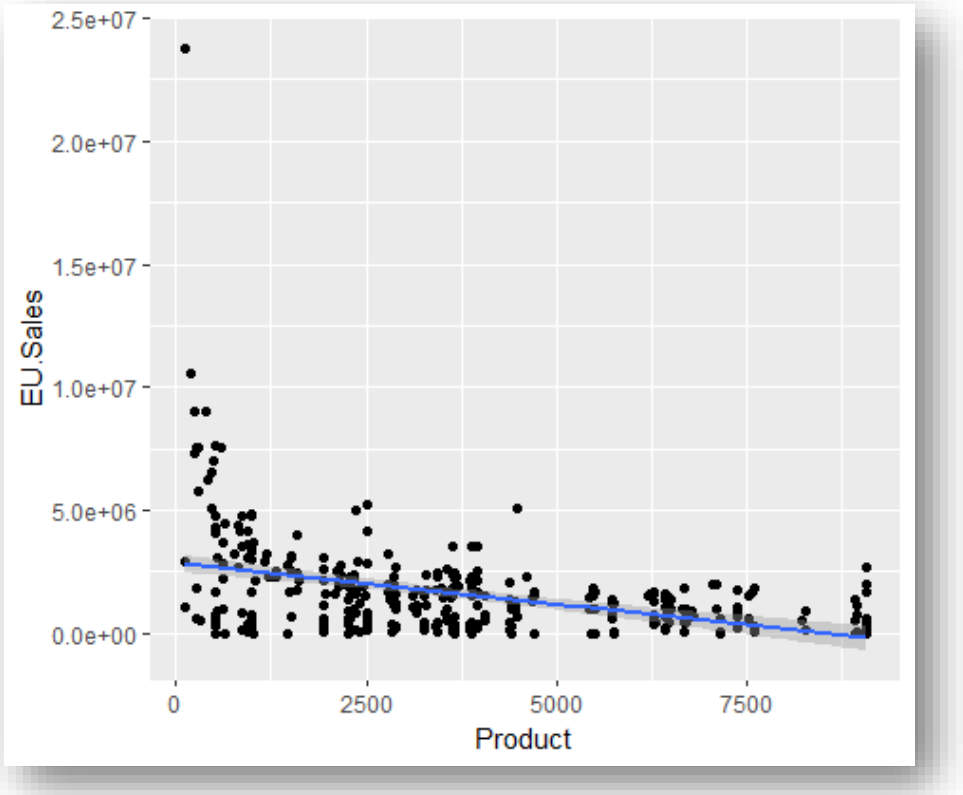
1. PC has lowest global sales
2. Nintendo has the highest global sales with equivalent from Playstation, Xbox and Others

4.8. Scatter Plots for Sales data

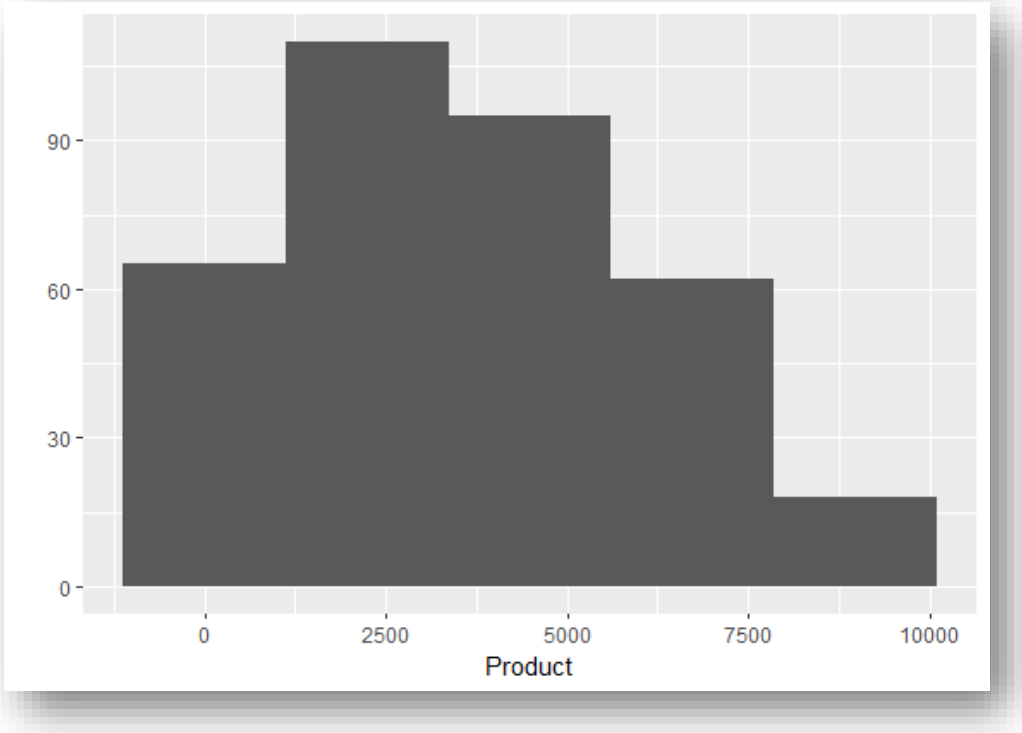


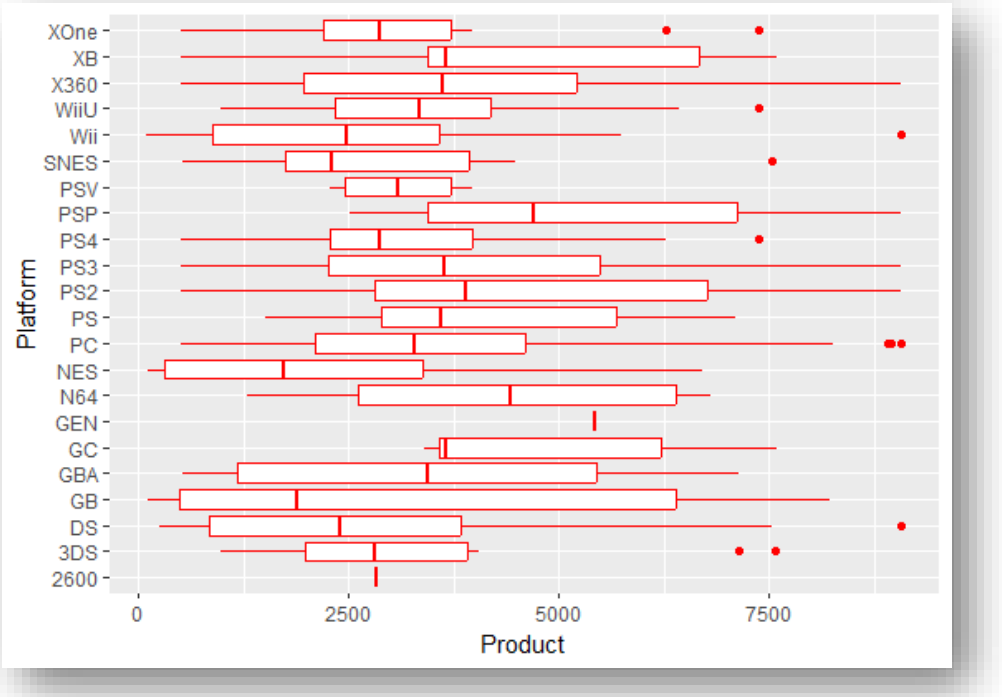



3e+07
2e+07
1e+07
0e+00
NA.Sales
0
2500
5000
7500
Product

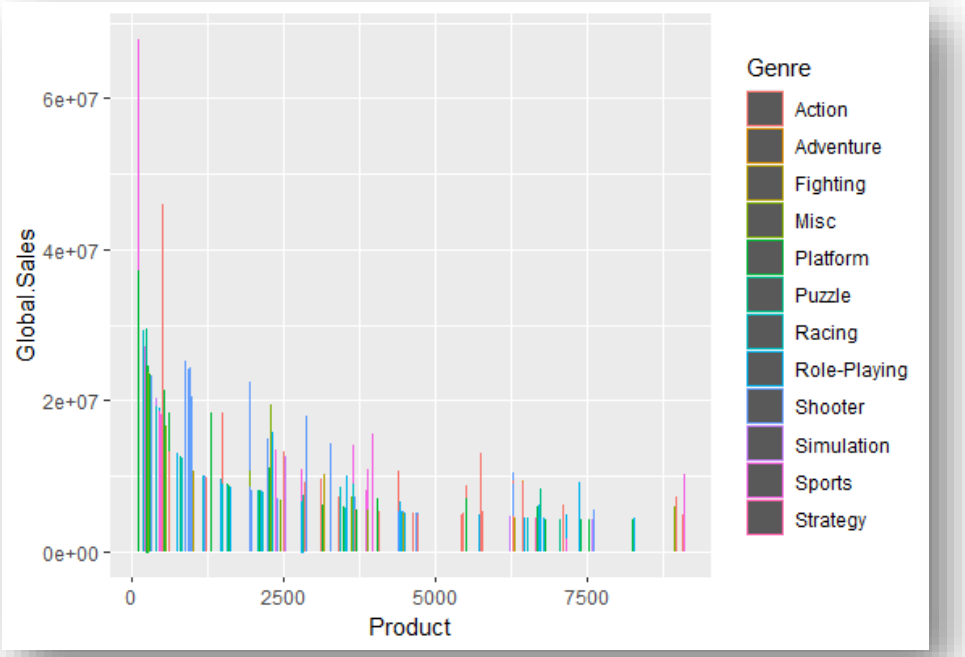

2.5e+07
2.0e+07
1.5e+07
1.0e+07
5.0e+06
0.0e+00
EU.Sales
0
2500
5000
7500
Product

4.9 Histogram and Boxplots for Product and Platform

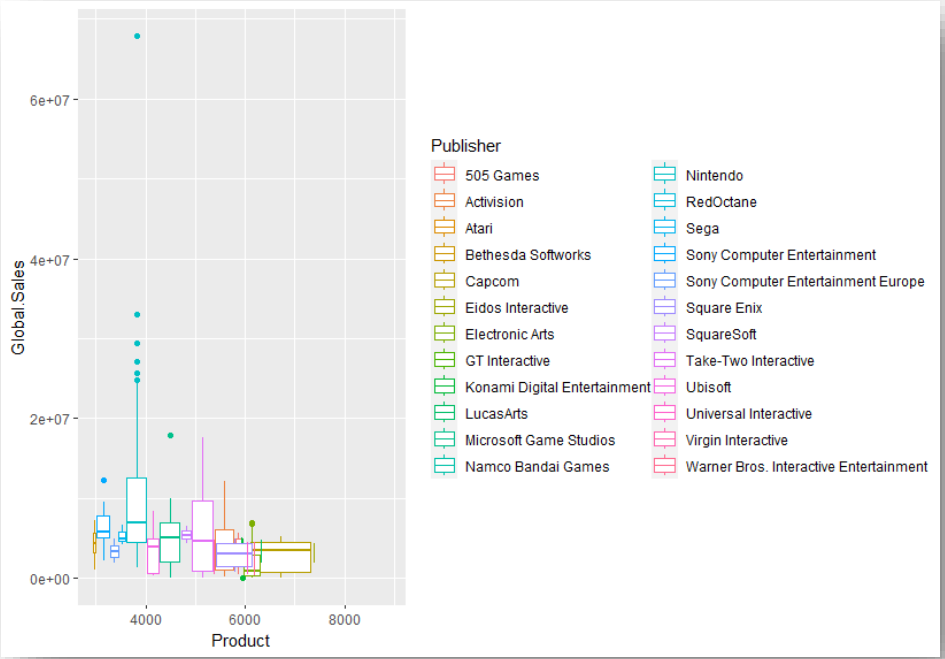




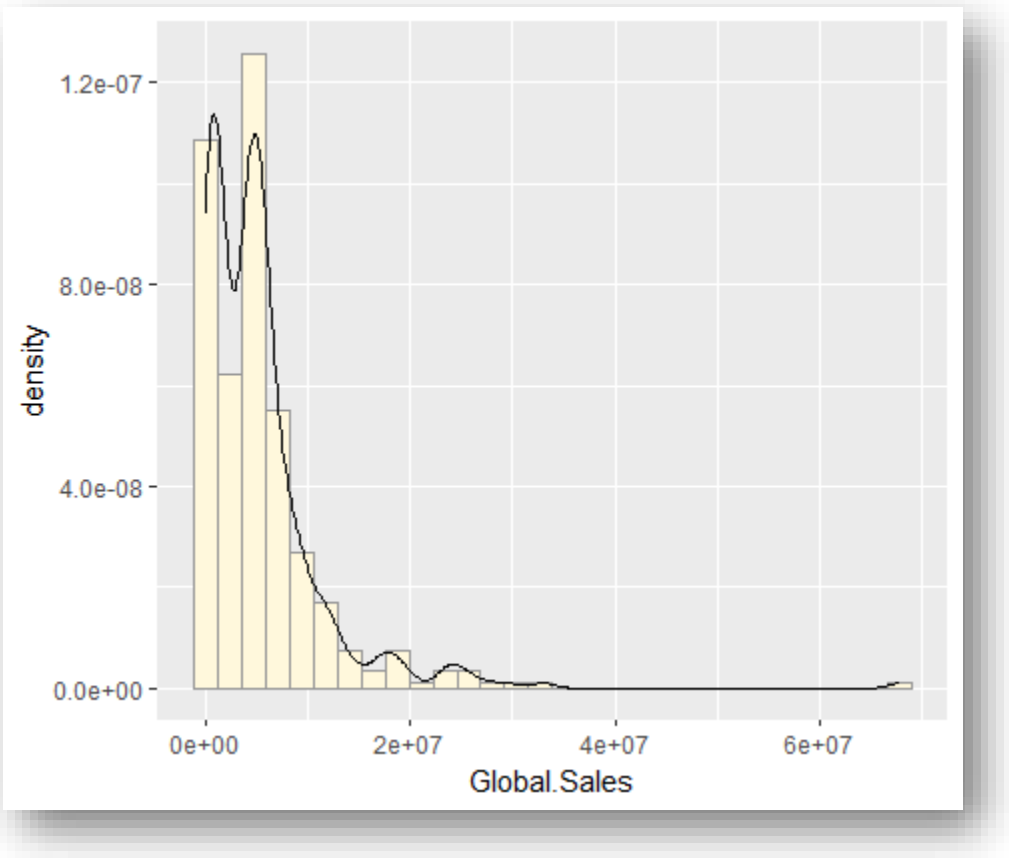
4.10 Global Sales against Product and Genre

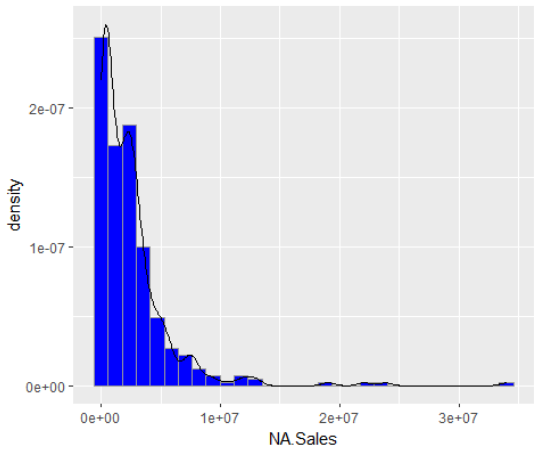


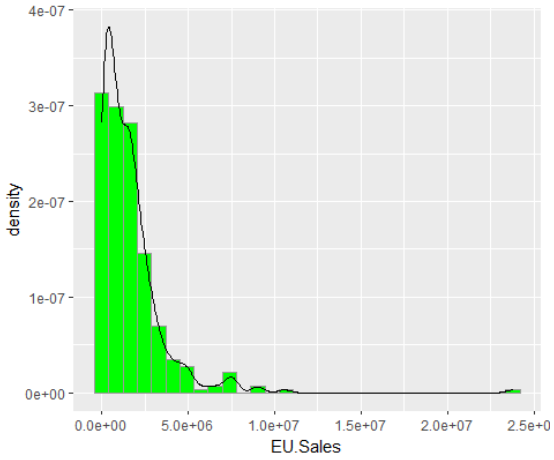
4.11 Global sales, Product and Publisher



4.12. Global Sales density v/s NA Sales density and EU Sales density

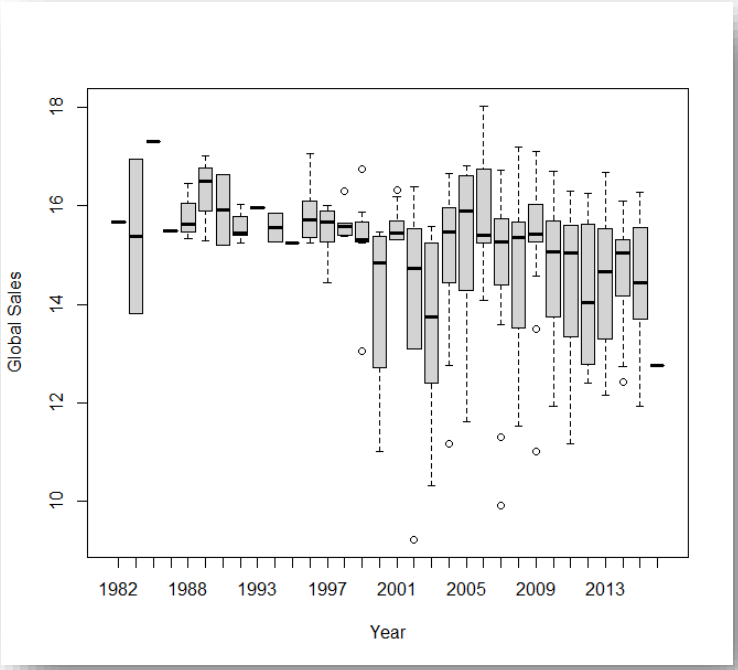


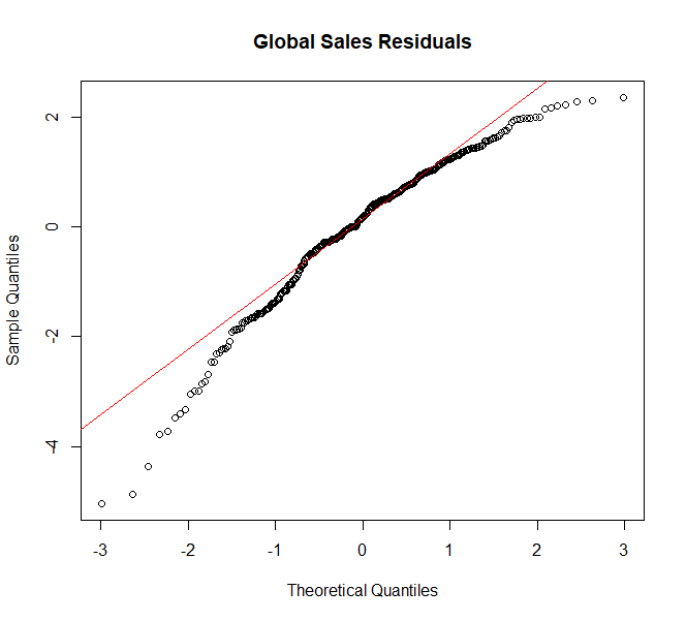




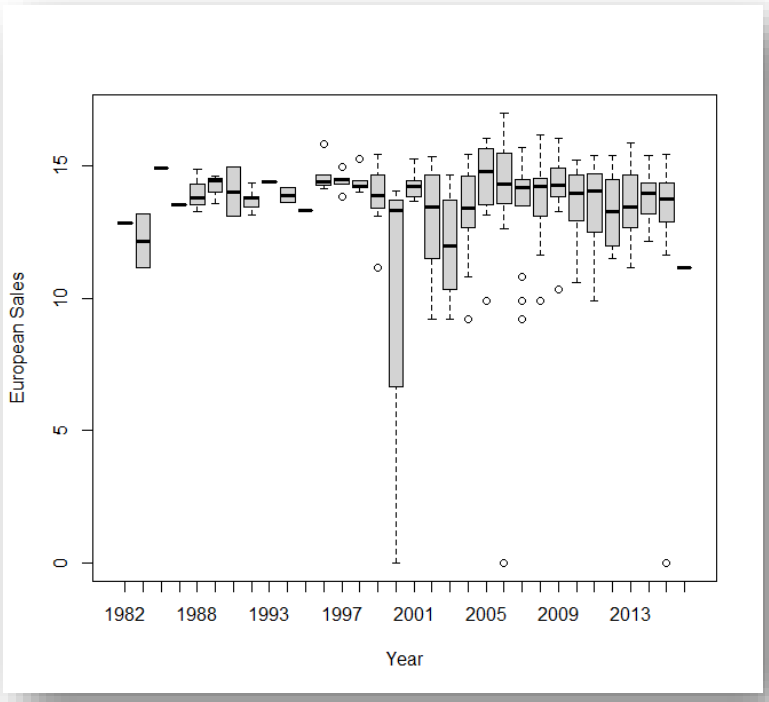
The pattern of density are similar for Global, North America, Europe which matches the cluster analysis.

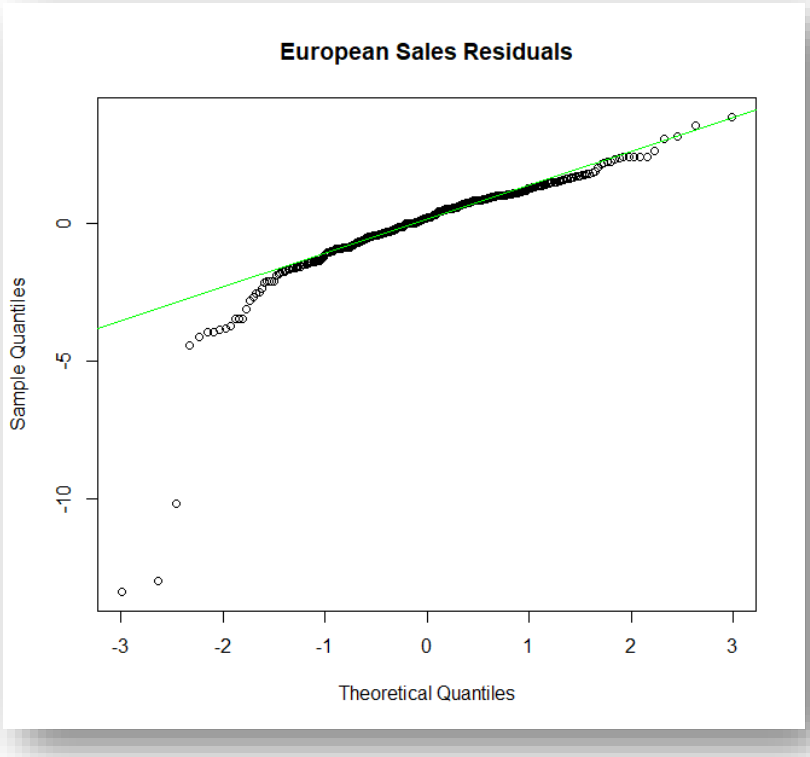
4.12. Global Sales / Year + Residuals



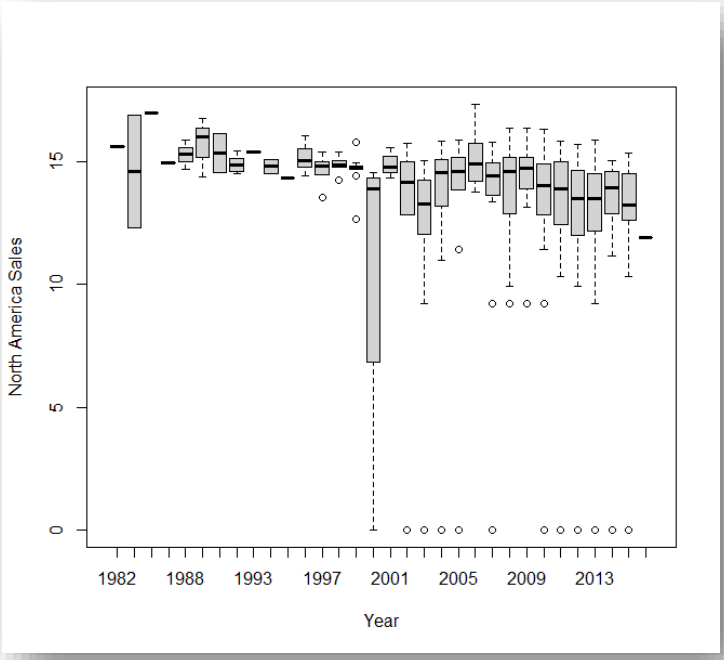


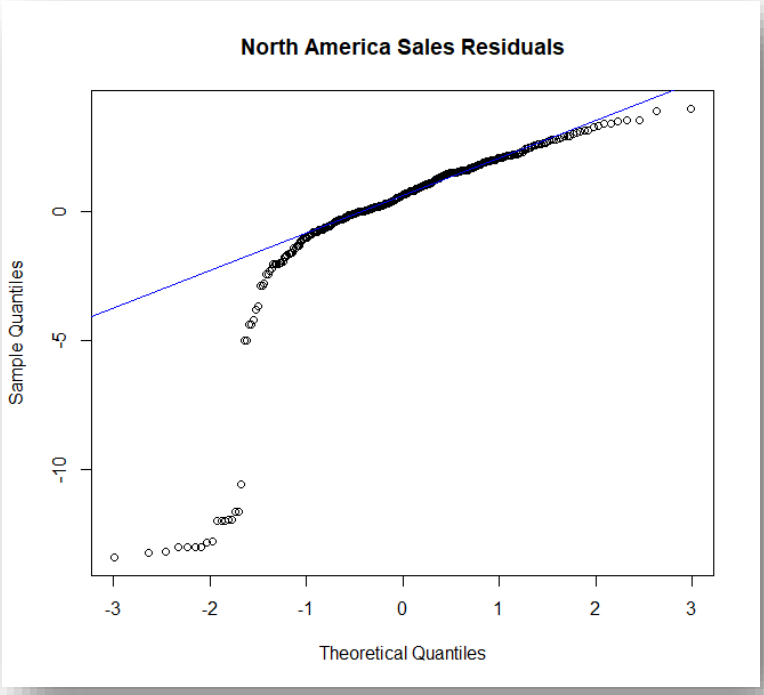
4.13 EU Sales / Year + Residuals



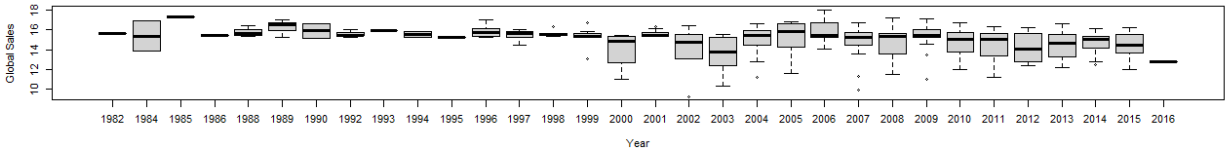


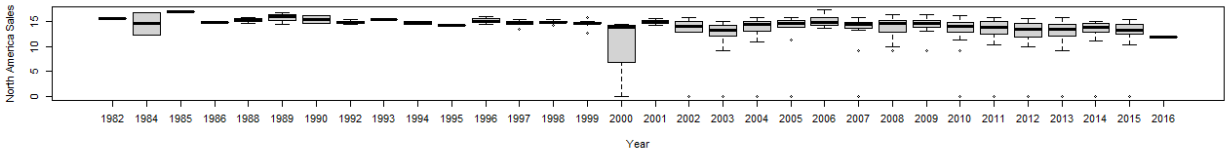
4.14 NA Sales / Year + Residuals

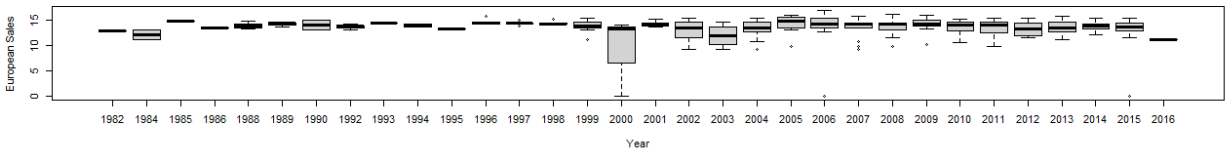




4.15 Sales Comparison







Summary

a. How customers accumulate loyalty points
    a. Higher the remuneration; higher the spending score and hence higher the loyalty points accumulated
    b. Spending v/s Loyalty (Linear regression): Higher the Spending score -> Higher the Loyalty
    c. Points Remuneration v/s Loyalty (Linear regression): Higher the Remuneration -> Higher the Loyalty Points
    d. Age v/s Loyalty (Linear Regression): constant (not much of a difference): Need to further analyze
b. How groups within the customer base can be used to target specific market segments
    a. VALUE COUNTS: Clusters to target!
        i. Cluster 0 count: 774
        ii. Cluster 3 count: 356
        iii. Cluster 1 count: 330
        iv. Cluster 2 count: 271
        v. Cluster 4 count: 269
c. How social data (e.g. customer reviews) can be used to inform marketing campaigns
    a. Frequency distribution, Word clouds, Polarity scores and Sensitivity analysis would determine the areas to improve marketing campaigns
    b. NPS scores are equally important to be derived as there are 20.2% neutral and 12.4% negative (detractors) and need to convert more promoters to improve the marketing strategy!
d. The impact that each product has on sales
    a. Products had an impact on the regional and global sales. More the number of products introduced to the market reduced the total sales and vice-versa.
e. How reliable the data is (e.g., normal distribution, skewness, or kurtosis)
    a. Had to log the data to transform the data set
f. What the relationship(s) is/are (if any) between Northern American, European, and Global Sales?
    a. Global and regional sales are not distributed normally, while their log values are close to normal distribution. Most regional sales have similar pattern as global sales.