Turtle Sales business objective: improving overall sales performance by utilising customer trends.

1. Customer analysis into loyalty points accumulation

The marketing department of Turtle Games wants to better understand how users accumulate loyalty points.

Data was imported, sense-checked, cleaned and saved in a new file. A data dictionary was created to describe the features in the datasets.

Linear regression analysis was conducted to predict each set of feature(s) against loyalty points. The results of the r-squared values were as follows:-

- spending_score: 0.452
- renumeration: 0.380
- age: 0.002
- age,renumeration, spending_score: 0.829 (From predictions test set; multicollinearity VIF factor: (1.1,1.0,1.1))

Scatter plots were used for the 4 linear regression analysis.

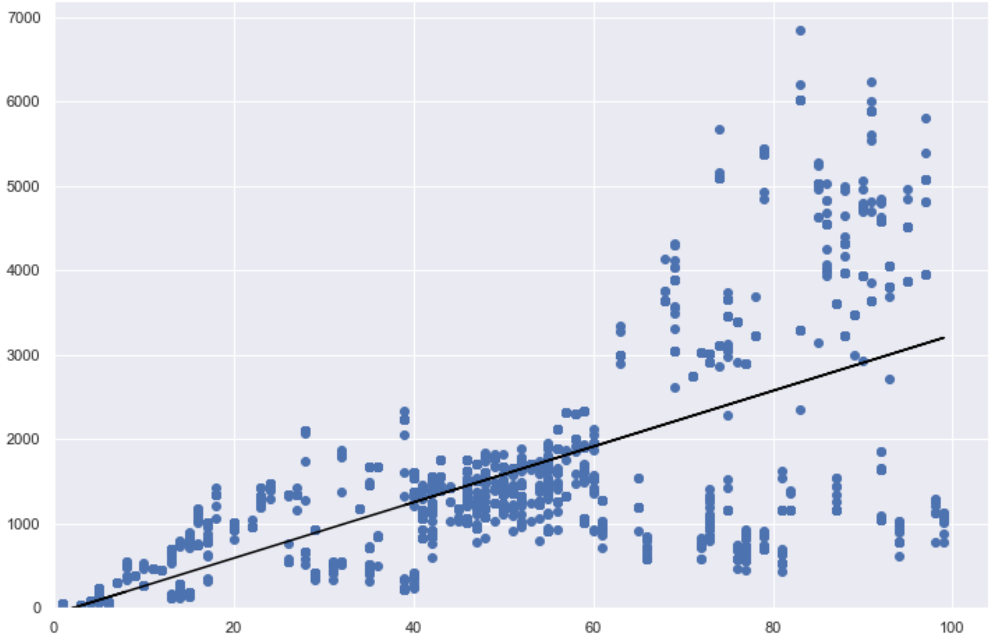


Fig 1.1 (spending vs loyalty)

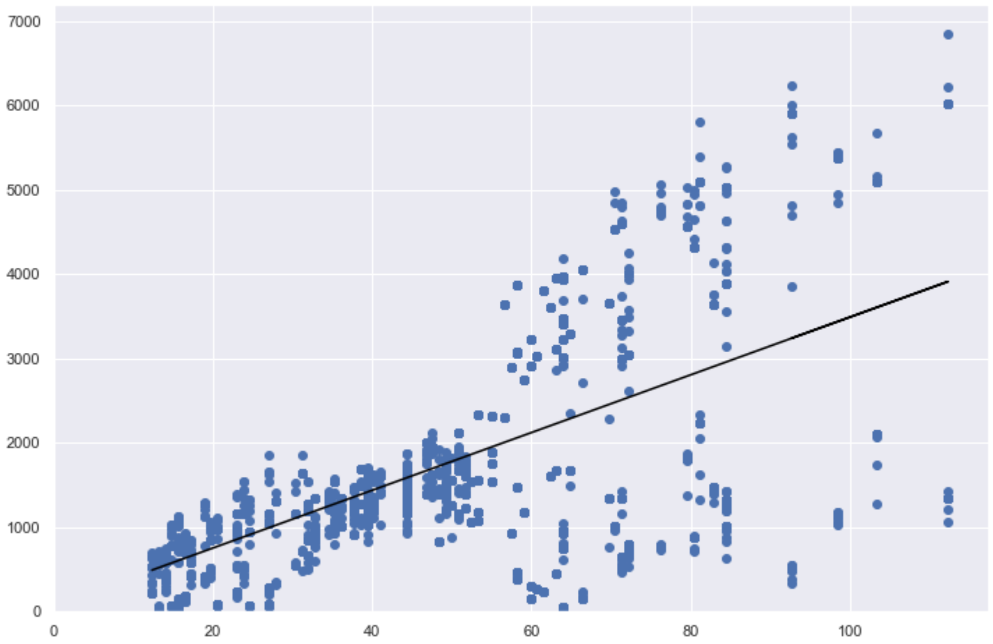


Fig 1.2 (renumeration vs loyalty)

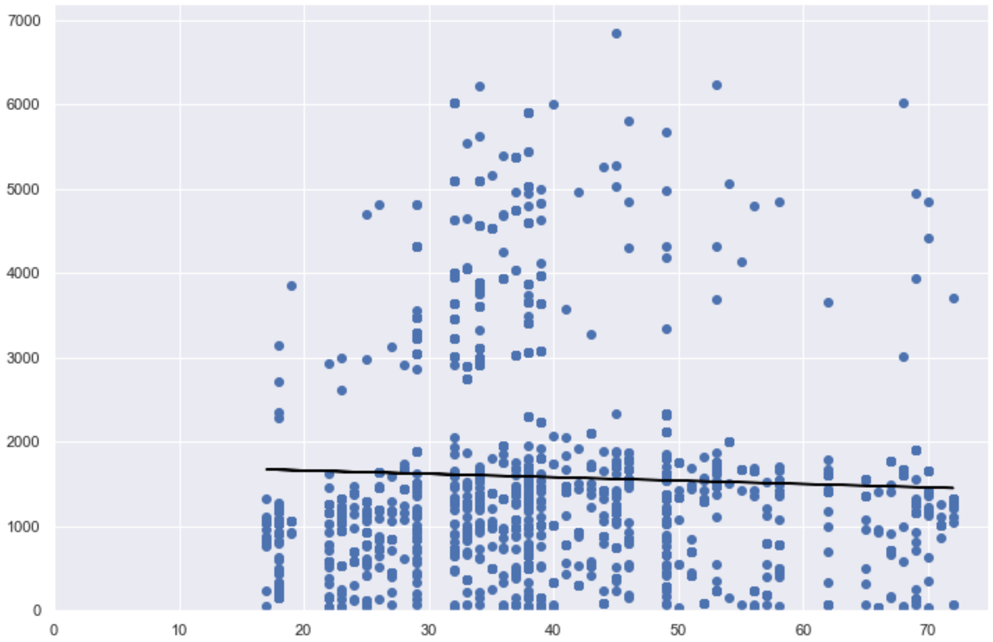


Fig 1.3 (age vs loyalty)

Based on the scatter plots and relevant statistical analysis, the MLR regression analysis produced the highest R-squared value of 0.842, accounting for more than 84% of the variation in the dependent variable that can be explained by the independent variables. Based on the VIF factor (~1.0), there are limited to no correlation between the 3 independent variables. Hence, it is likely that age, renumeration and spending score contributed to the customer accumulating higher loyalty points and spending more.

2. Renumeration and spending scores analysis

Following the previous customer analysis, we now examine the groups of customers based on renumeration and spending score that the team can focus on to boost overall sales performance; tailoring a different strategy to each group.

A Centroid clustering model, k-means clustering was deployed to organised data into non-hierarchical clusters. Scatterplot was used to visually inspect the plausible cluster relationship.

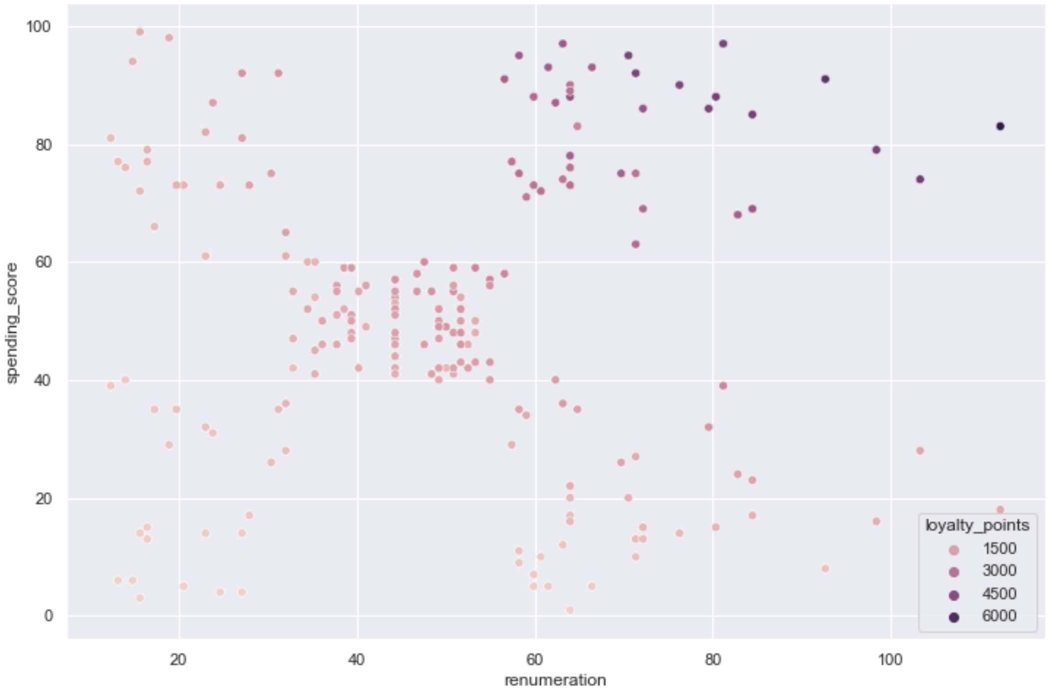


Fig 2.1 (renumeration-spending-loyalty plot)

The plot above was used in conjunction with the elbow and silhouette methods in determining the number of clusters.

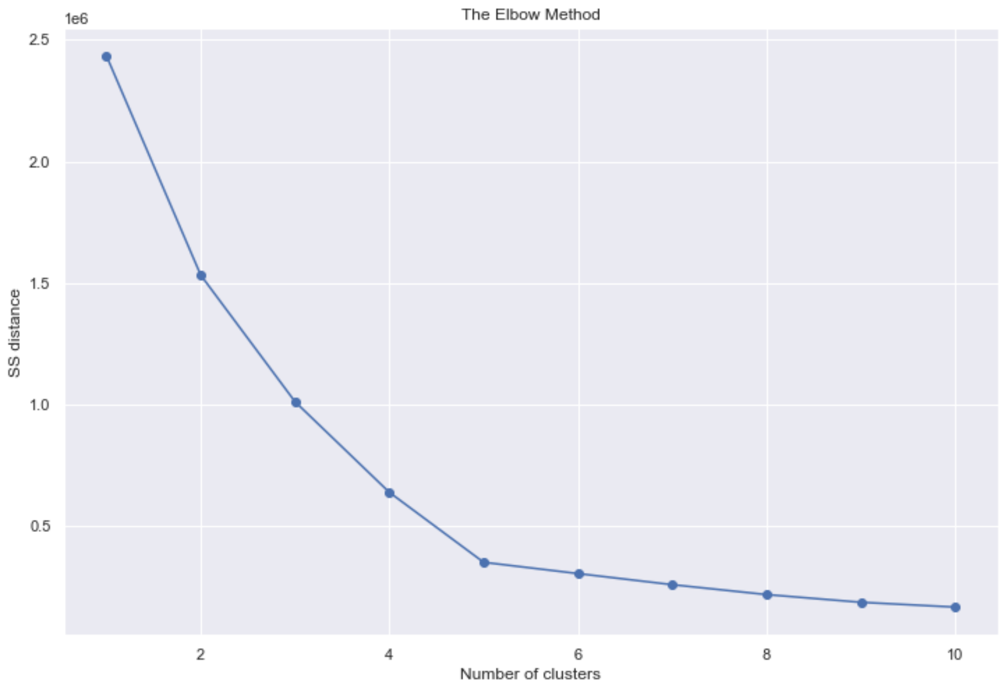


Fig 2.2 (Elbow Plot)

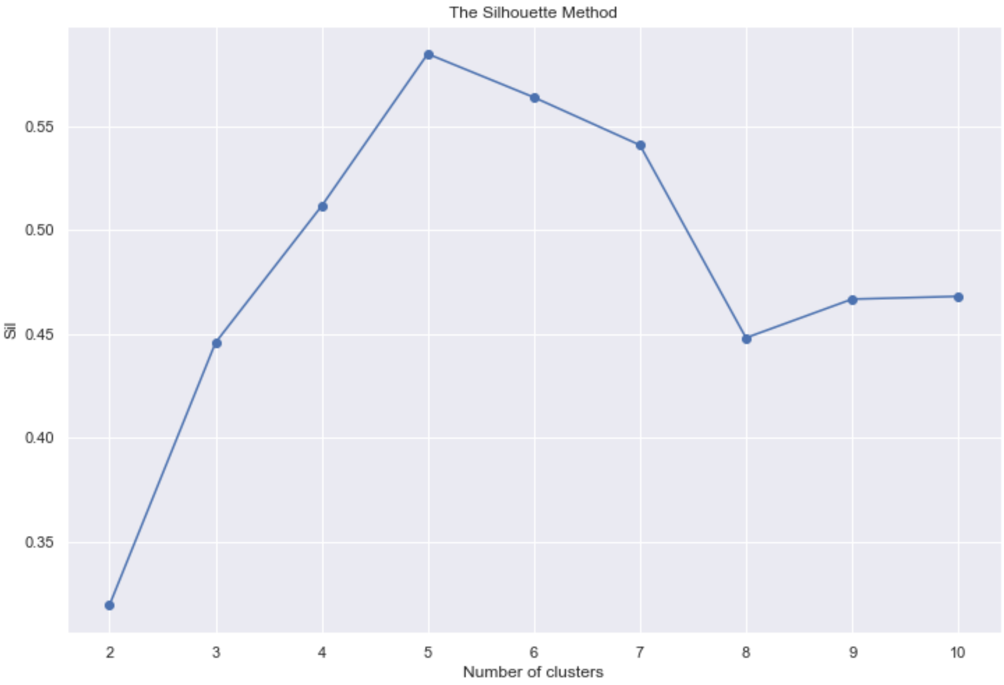


Fig 2.3 (Silhouette plot)

Based on the cross-validation of various plots above, we noted that the optimal number of clusters to be 5.

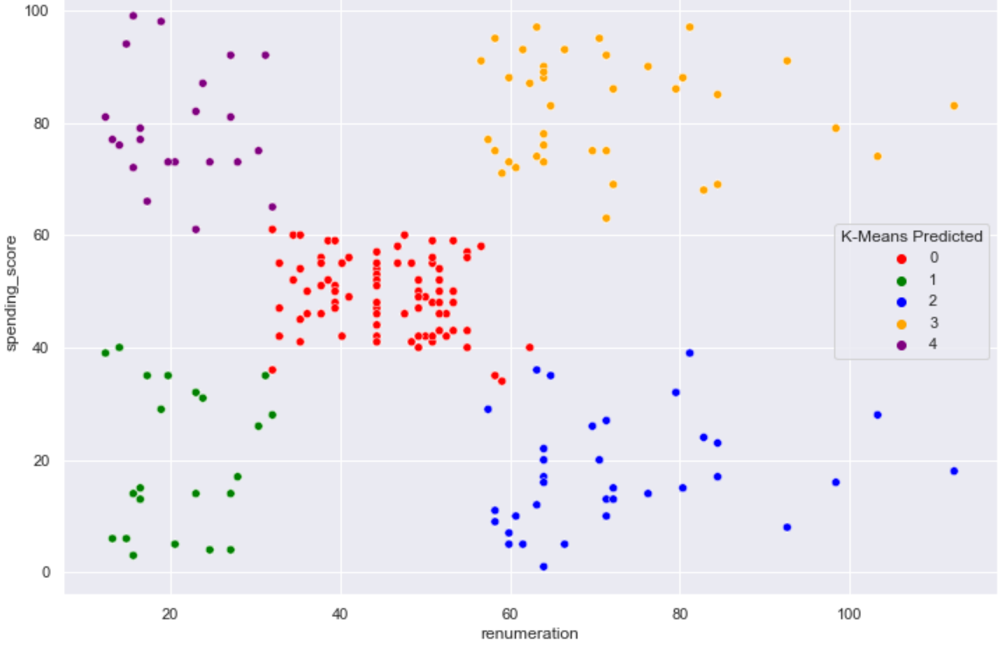


Fig 2.4 (k-means clusters plot)

The elbow method seems to be following a linear pattern after k=5. Also, the silhouette score is slightly higher with 5 clusters. Visualising the scatter plots above after evaluating the model, we also noted that the 5 cluster groups are generally distributed fairly as compared to 4 cluster groups; with 1 cluster having a higher number of observation. The number of predicted values per class indicates a better distribution for `k=5` than `k=4`. Hence, we will use 5 clusters for our k means model.

The clusters can be labelled as such (renumeration,spending) :-

- High spending low renumeration (0-36,60-100)
- High spending high renumeration (58-110, 60-100)
- Low spending high renumeration (58-110,0-40)
- Low spending low renumeration(0-35,0-40)
- Average spending average renumeration (25-55, 40-60)

3. Sentiment analysis on customers review

Customer reviews were downloaded from the website of Turtle Games. This data will be useful for the marketing department because it can inform future campaigns.

We will deploy the natural language processing on the dataset to conduct sentiment analysis on the customers reviews.

The customer review dataset was cleaned for the following:

- Lower case
- Punctuations
- Joining elements of the same row using whitespace
- Duplicates

Thereafter, the dataset was tokenised and a word cloud was created for both the summary and review column.

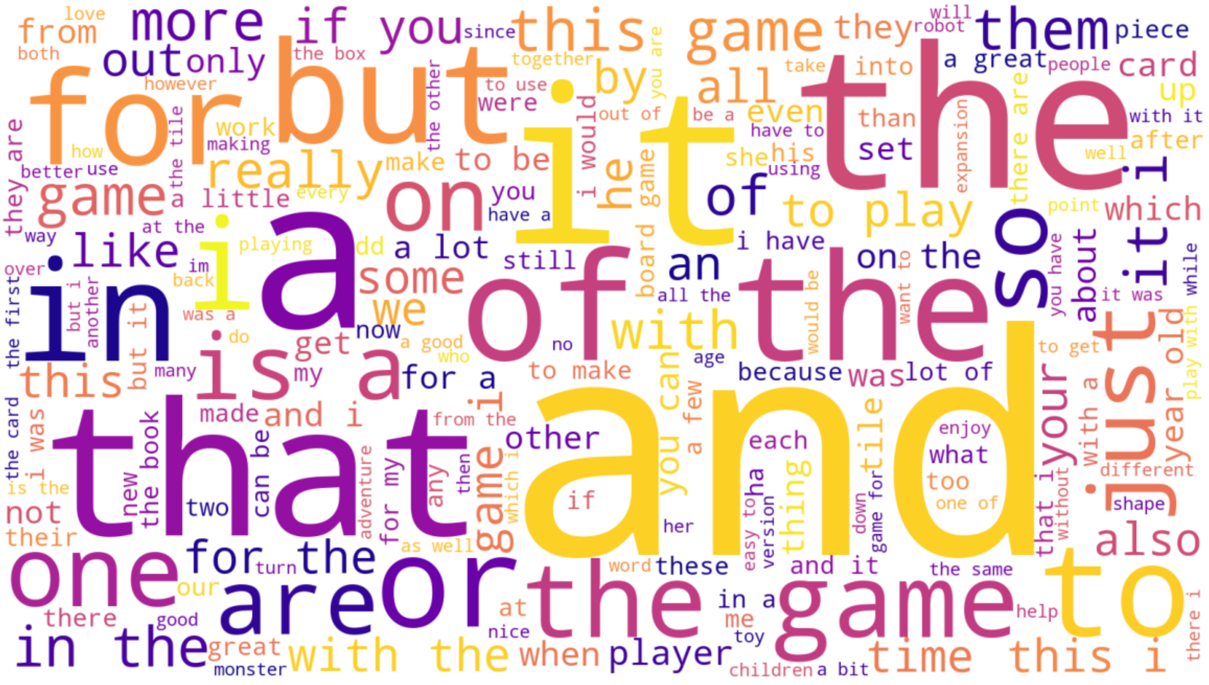


Fig 3.1 (wordcloud-review)

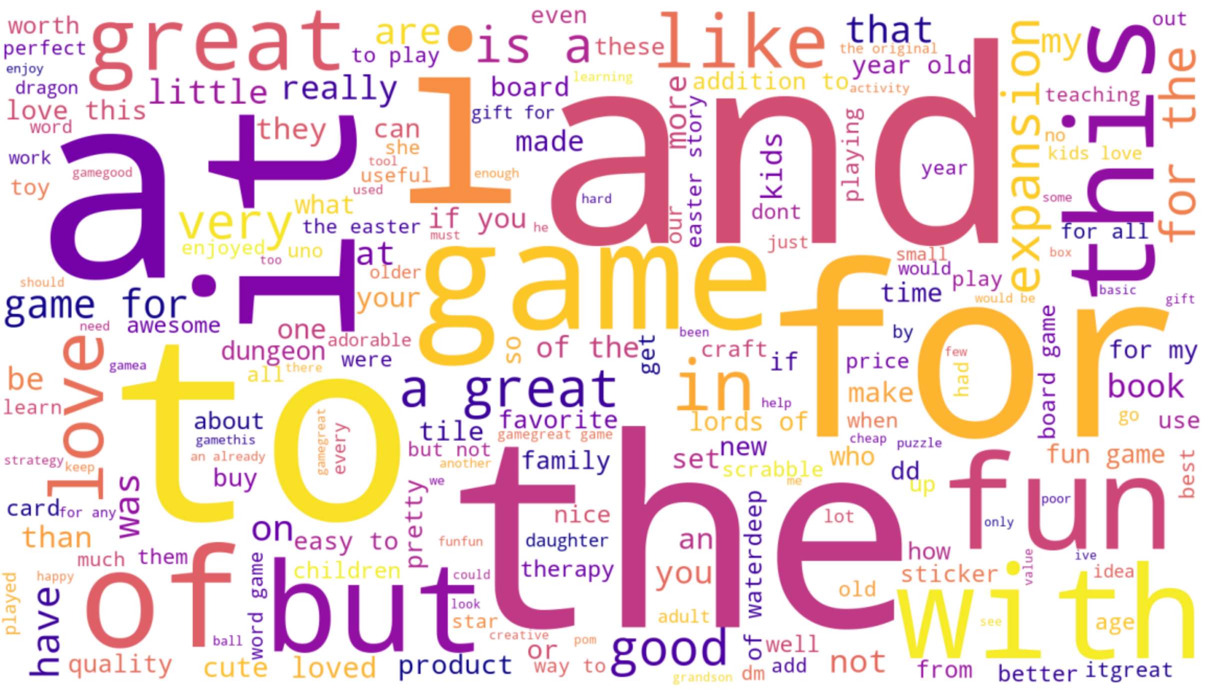


Fig 3.2 (wordcloud-summary)

Another set of wordclouds was created without stopwords from the NLTK library.

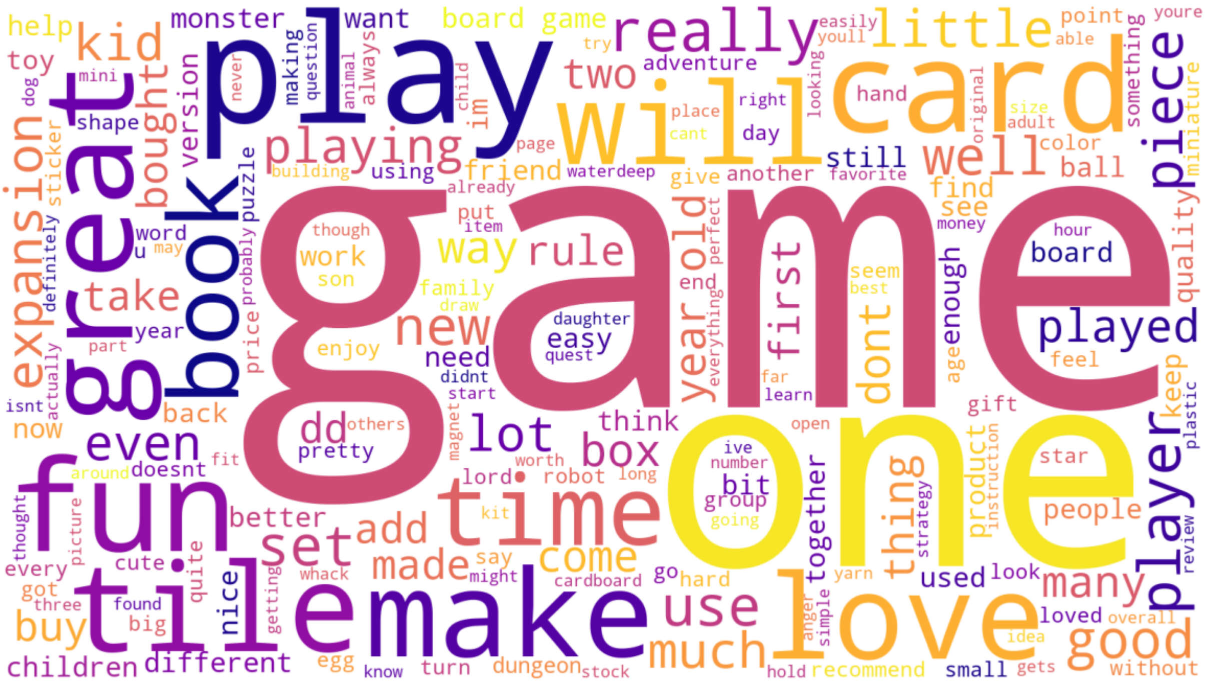


Fig 3.3 (wordcloud2-review)

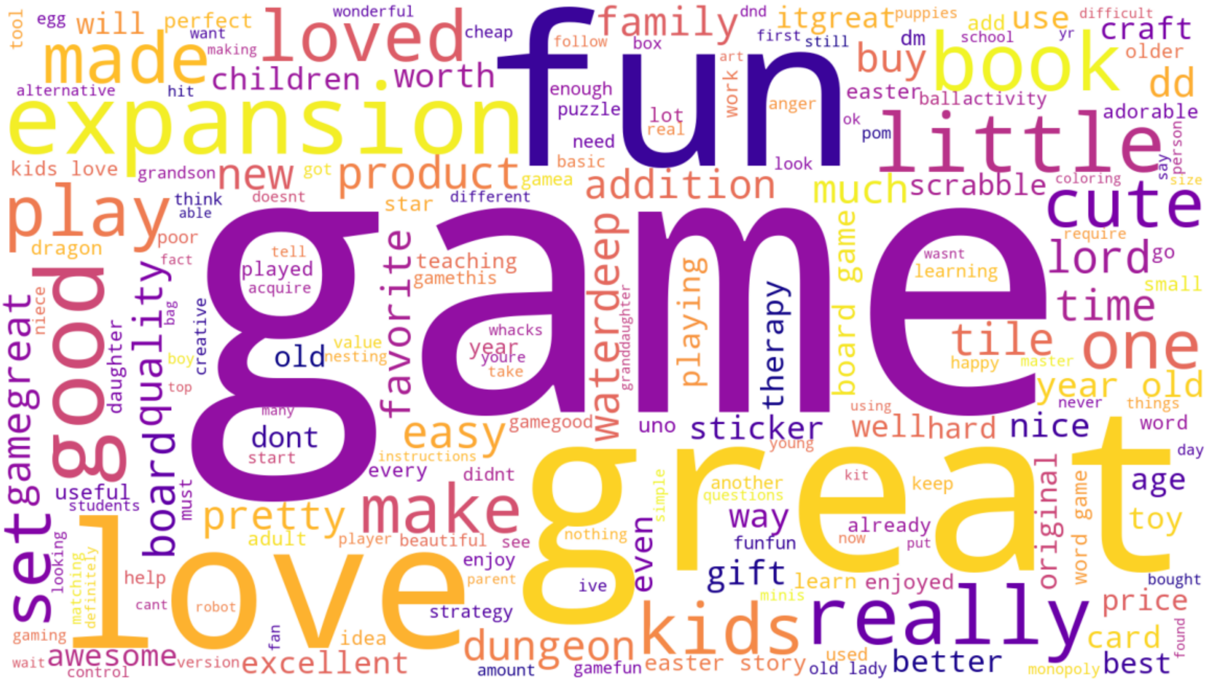


Fig 3.4 (wordcloud2-summary)

Comparing both sets of wordclouds, we start to see a fair bit of prominent words surfacing. Words such as game, love, great, card, tile, family, fun and player. From these words, it is likely that the the reviews and summaries were generally targeted at games that were enjoyable, foster a sense of closeness in terms family. It predominantly tells us that the games that the customer likes to play were cooperative in nature and also related to logic thinking puzzles.

Moving forward with the analysis, we will focus on finding the top 15 most common words and sentence polarity using textblob library. A horizontal barplot was used for the most frequent words.

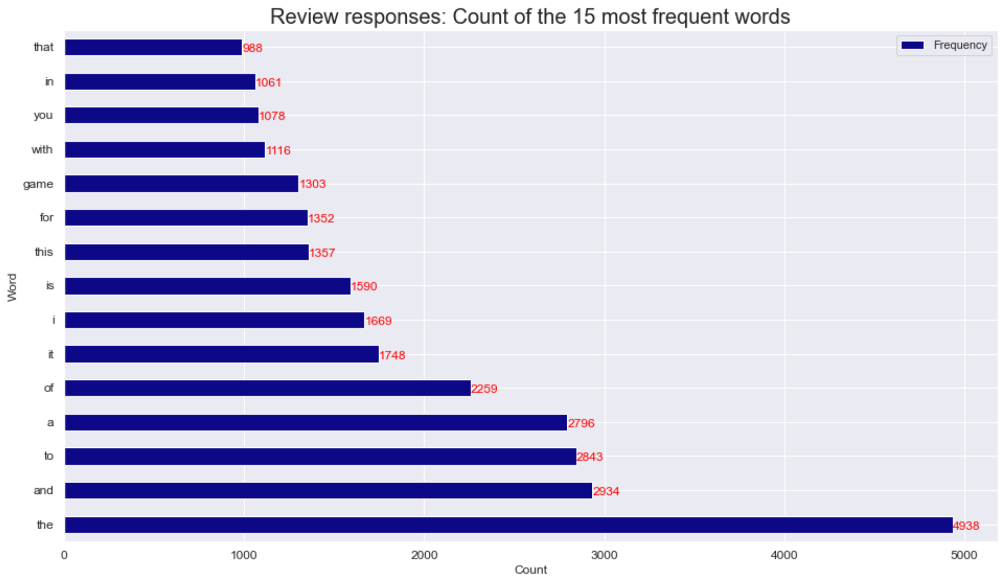


Fig 3.5 (top 15 words for review column)

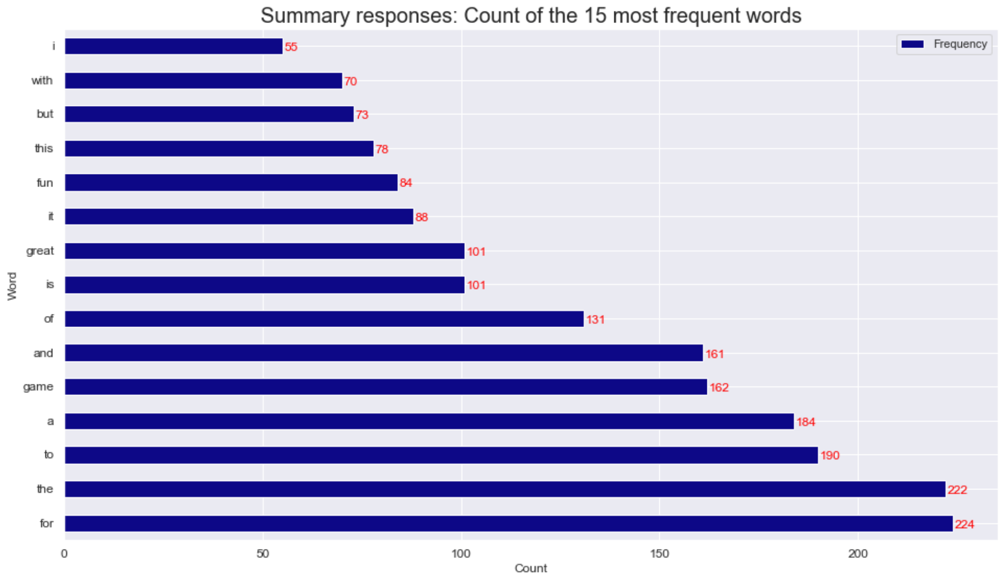


Fig 3.6 (top 15 words for summary column)

The barplot generally agrees with the wordcloud with a few similar words used. However, noted that most of the words on the bar plots were generally neutral in nature.

A histogram plot was used to show the sentiment analysis of the review and summary column.

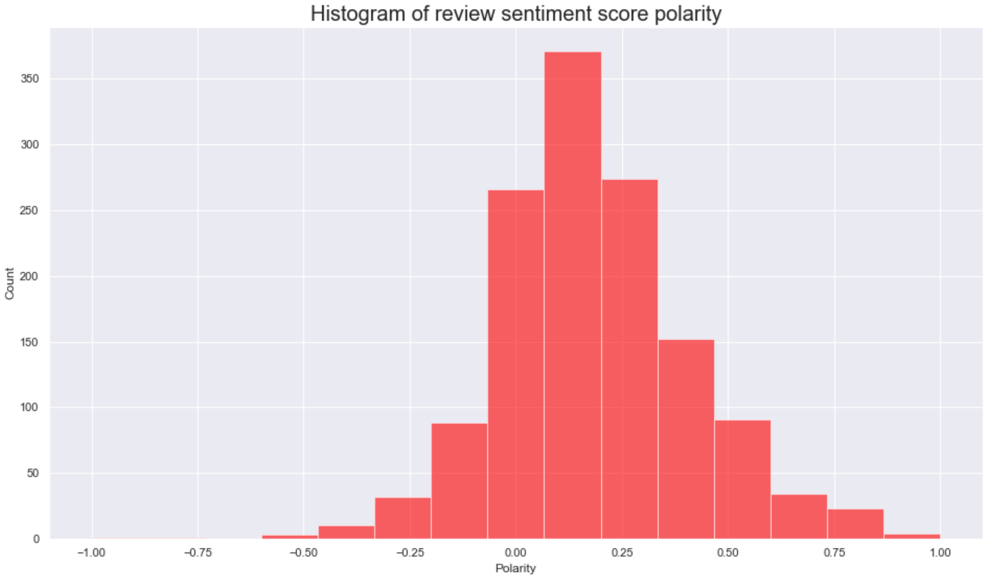


Fig 3.7 (sentiment analysis for review column)

This plot shows us that most comments sit closest to neutral, not expressing a particularly strong sentiment in either direction.

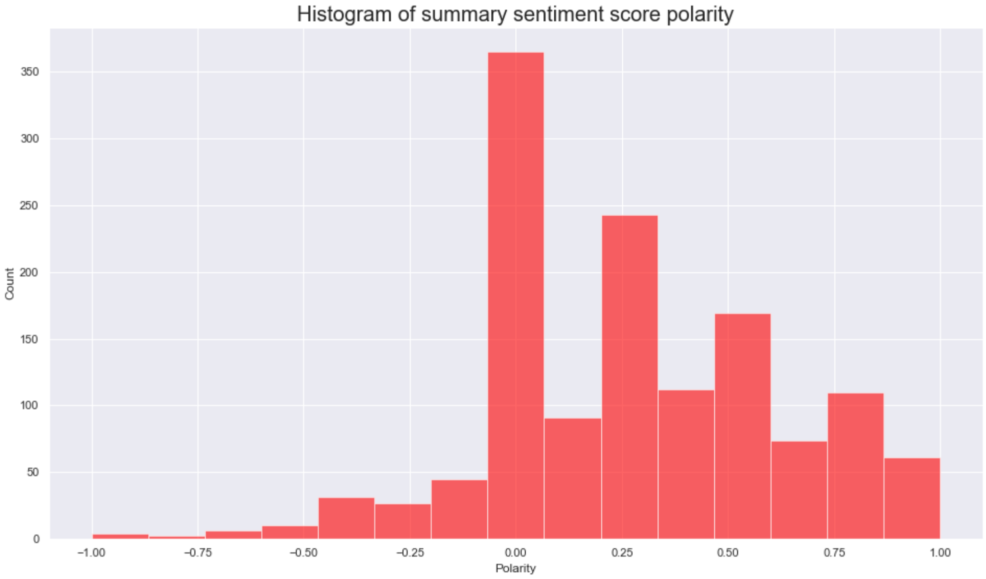
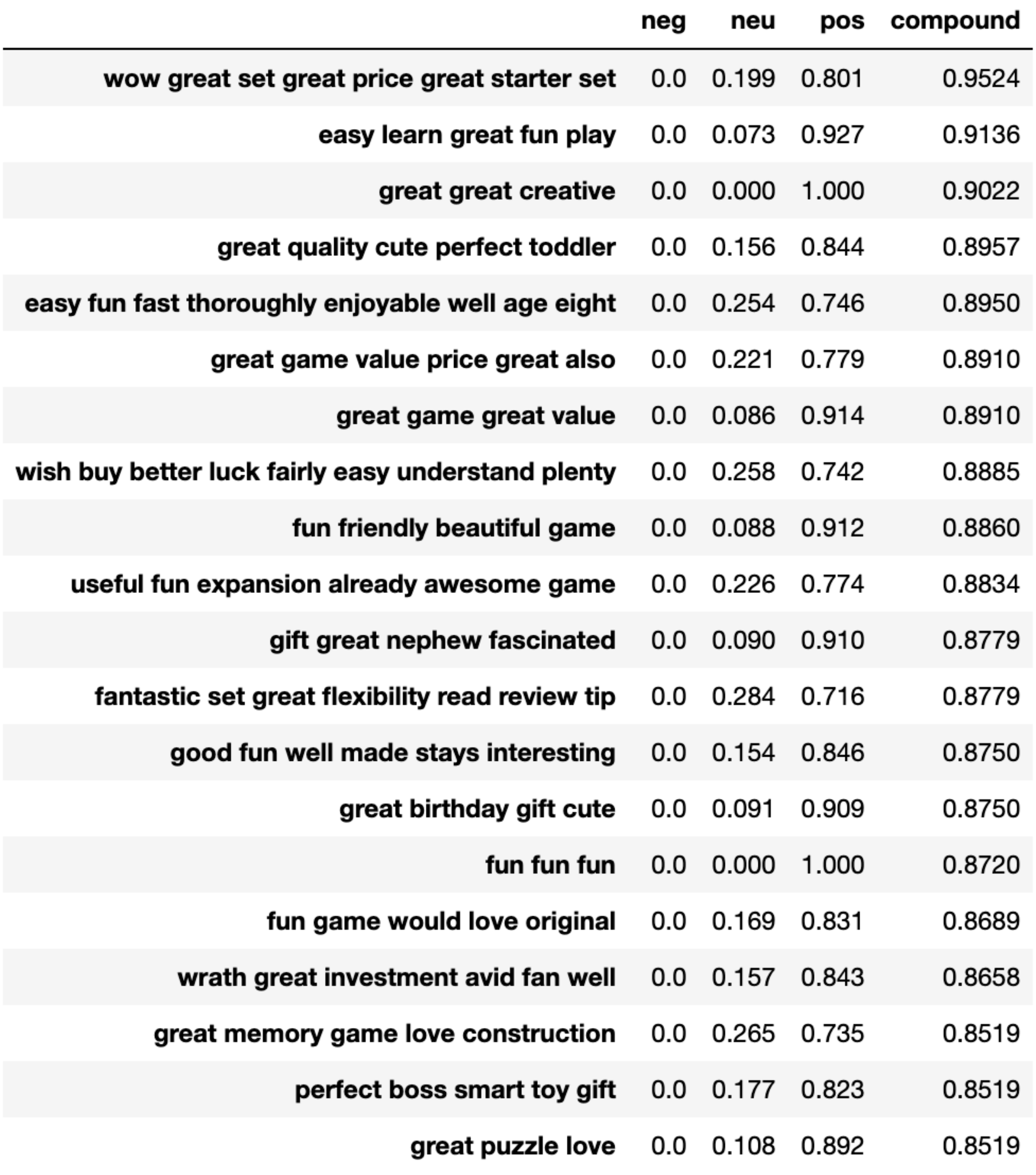


Fig 3.8 (sentiment analysis for summary column)

This plot shows us that most summary sit closest to neutral, expressing a relatively stronger sentiment in the positive direction.

We now cross reference our findings with the sentence polarity; investigating the top 20 positive and negative sentences from the review and summary columns. The same dataset was re-tokenised after processing the raw data by removing stopwords and non-alphanumeric words. The VaderSentiment library was used for this analysis.

The results from the summary column were particularly helpful in describing the products.



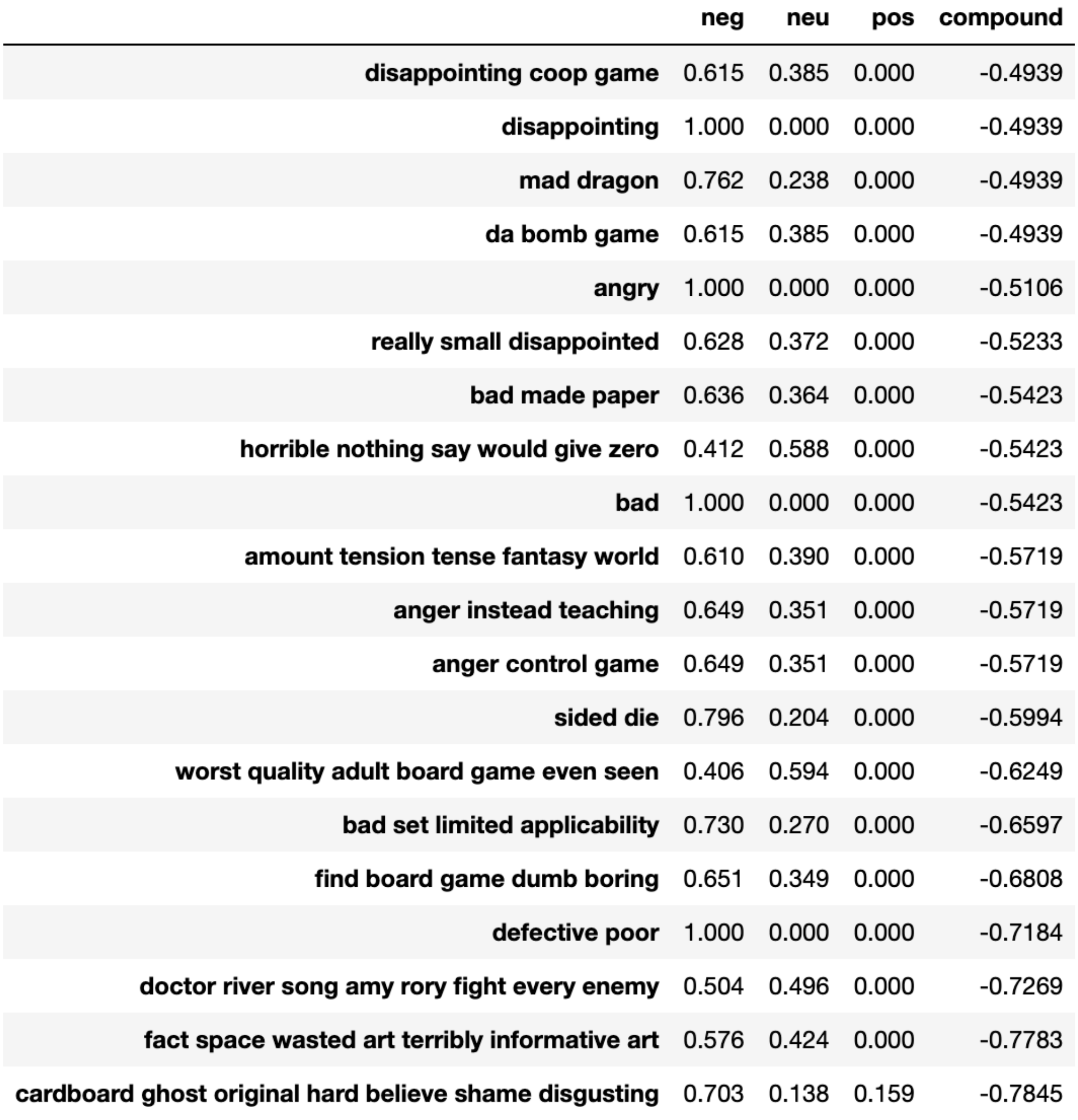
	neg	neu	pos	compound
wow great set great price great starter set	0.0	0.199	0.801	0.9524
easy learn great fun play	0.0	0.073	0.927	0.9136
great great creative	0.0	0.000	1.000	0.9022
great quality cute perfect toddler	0.0	0.156	0.844	0.8957
easy fun fast thoroughly enjoyable well age eight	0.0	0.254	0.746	0.8950
great game value price great also	0.0	0.221	0.779	0.8910
great game great value	0.0	0.086	0.914	0.8910
wish buy better luck fairly easy understand plenty	0.0	0.258	0.742	0.8885
fun friendly beautiful game	0.0	0.088	0.912	0.8860
useful fun expansion already awesome game	0.0	0.226	0.774	0.8834
gift great nephew fascinated	0.0	0.090	0.910	0.8779
fantastic set great flexibility read review tip	0.0	0.284	0.716	0.8779
good fun well made stays interesting	0.0	0.154	0.846	0.8750
great birthday gift cute	0.0	0.091	0.909	0.8750
fun fun fun	0.0	0.000	1.000	0.8720
fun game would love original	0.0	0.169	0.831	0.8689
wrath great investment avid fan well	0.0	0.157	0.843	0.8658
great memory game love construction	0.0	0.265	0.735	0.8519
perfect boss smart toy gift	0.0	0.177	0.823	0.8519
great puzzle love	0.0	0.108	0.892	0.8519

Fig 3.9 (Positive sentiment analysis for summary column sentences)

From the positive sentences, It generally agrees and validate the initial findings from the word cloud. (cooperative and logic thinking puzzles)

The marketing team could use this information to move more of these game products through active promotions as it tends to be the current interest of the customers.

From the negative sentence we realise that customer would also prefer games that are instructive, cooperative and logic-thinking puzzles with great art designs. Hence, it reinforced the particular games that the customers prefers.



	neg	neu	pos	compound
disappointing coop game	0.615	0.385	0.000	-0.4939
disappointing	1.000	0.000	0.000	-0.4939
mad dragon	0.762	0.238	0.000	-0.4939
da bomb game	0.615	0.385	0.000	-0.4939
angry	1.000	0.000	0.000	-0.5106
really small disappointed	0.628	0.372	0.000	-0.5233
bad made paper	0.636	0.364	0.000	-0.5423
horrible nothing say would give zero	0.412	0.588	0.000	-0.5423
bad	1.000	0.000	0.000	-0.5423
amount tension tense fantasy world	0.610	0.390	0.000	-0.5719
anger instead teaching	0.649	0.351	0.000	-0.5719
anger control game	0.649	0.351	0.000	-0.5719
sided die	0.796	0.204	0.000	-0.5994
worst quality adult board game even seen	0.406	0.594	0.000	-0.6249
bad set limited applicability	0.730	0.270	0.000	-0.6597
find board game dumb boring	0.651	0.349	0.000	-0.6808
defective poor	1.000	0.000	0.000	-0.7184
doctor river song amy rory fight every enemy	0.504	0.496	0.000	-0.7269
fact space wasted art terribly informative art	0.576	0.424	0.000	-0.7783
cardboard ghost original hard believe shame disgusting	0.703	0.138	0.159	-0.7845

Fig 3.10 (Negative sentiment analysis for summary column sentences)

4. Impact on sales per product

Using R, we look into the turtle_sales dataset. Data was imported and cleaned to include only relevant data. (Platform, sales, product) Scatterplots, boxplots and histograms were used to gain insights on the data.

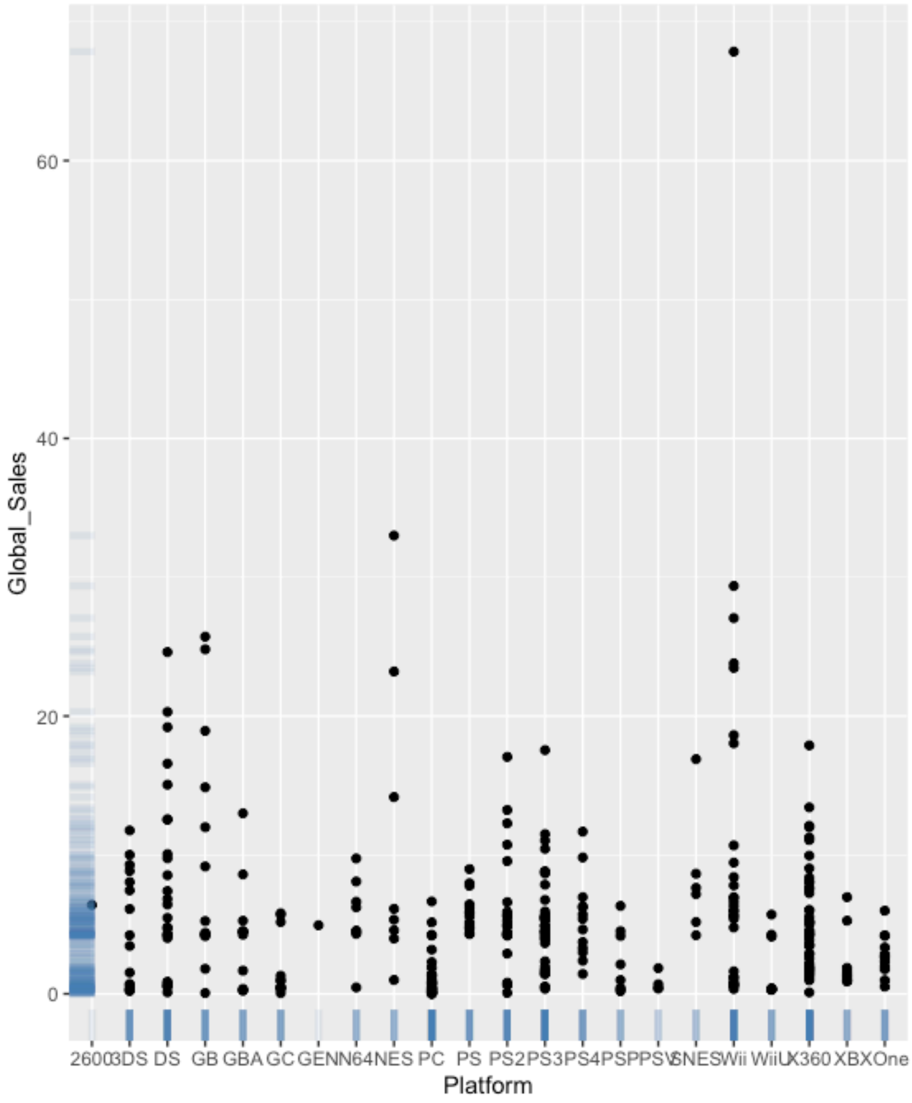


Fig 4.1 (Platform vs Global sales scatterplot)

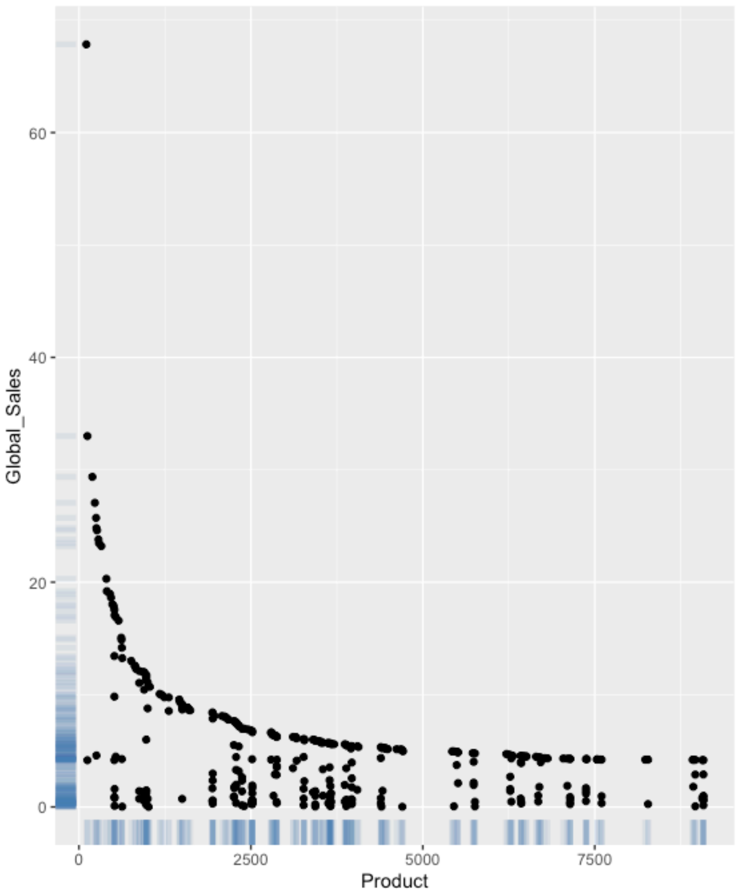


Fig 4.2 (Product vs Global sales scatterplot)

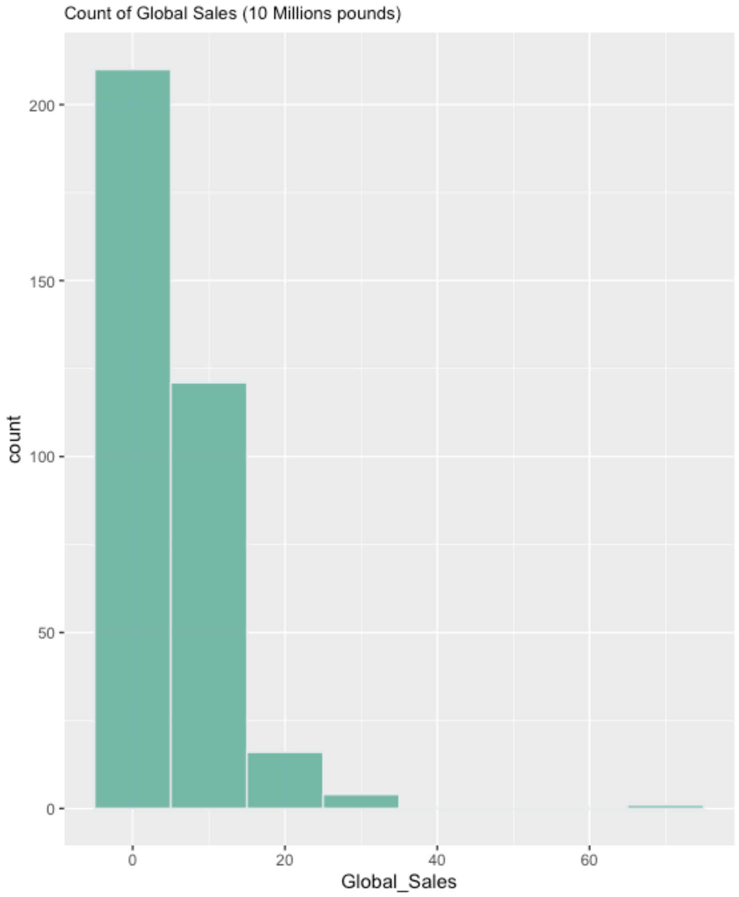


Fig 4.3 (Global sales histogram)

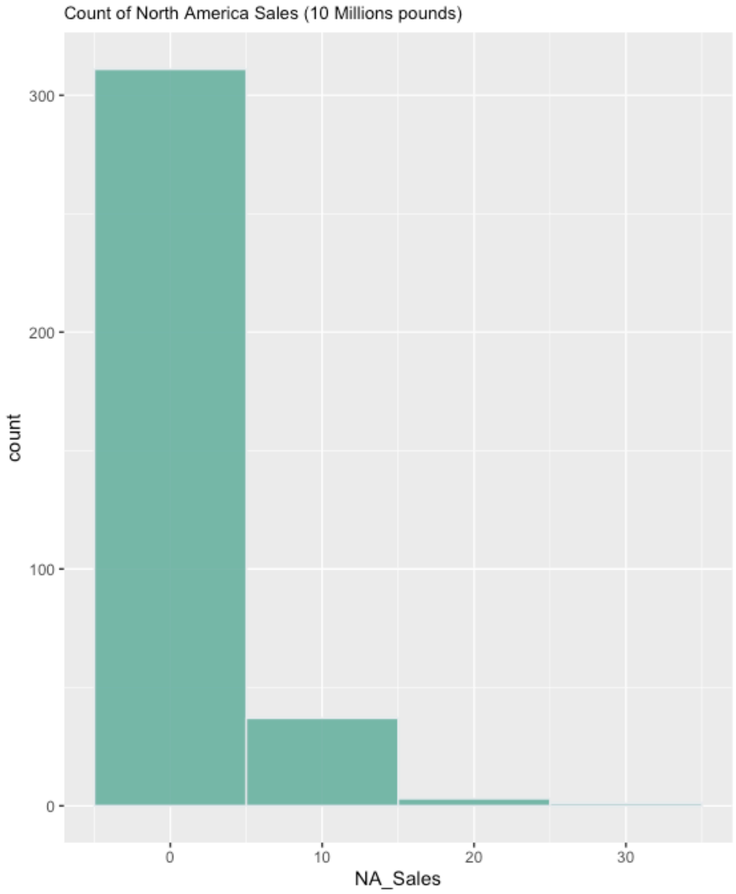


Fig 4.4 (North America sales histogram)

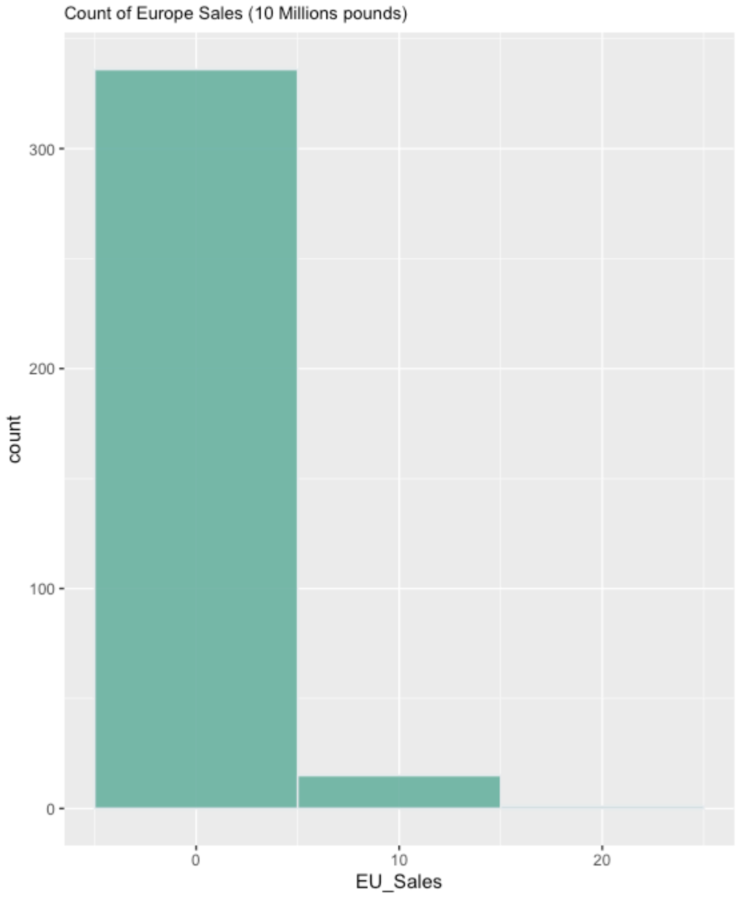


Fig 4.5 (Europe sales histogram)

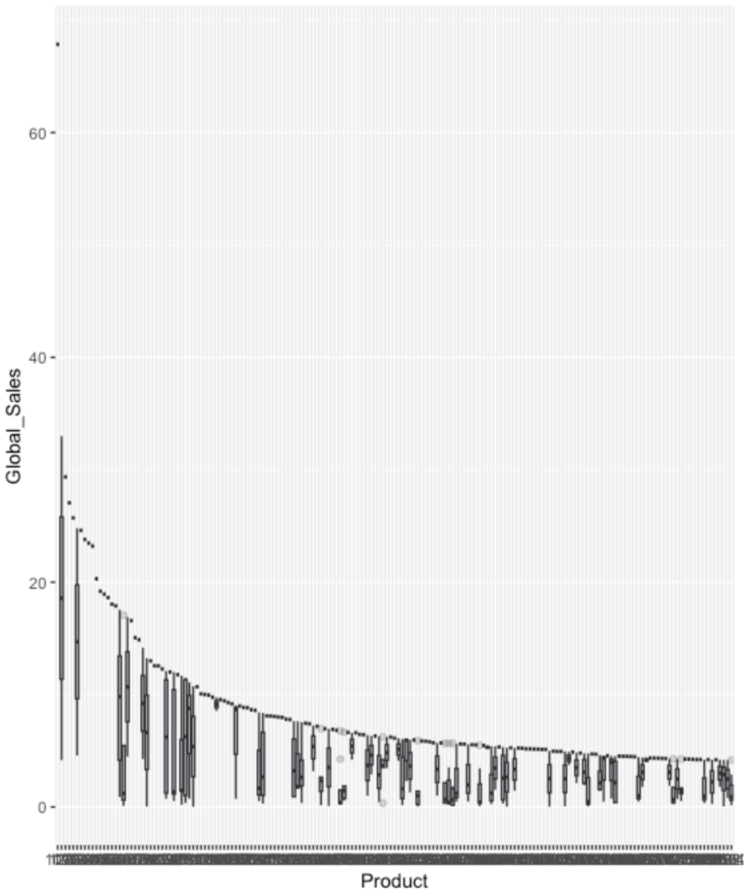


Fig 4.6 (Product vs Global sales boxplot)

Based on the Histogram on sales, Majority of the global sales were less than 10 million pounds. Breaking down the sales further into different regions, North America tend to sell more of the products as compared to Europe given that there are more sales numbering over 10 million pounds.

One particularly interesting trend we can see from the scatter and boxplot of the product sales is that, the number of sales decreases as new product are released across the years. (Starting from #1) It is likely that the customers are playing a lot lesser platform games throughout the years.

5. <u>Reliability of data</u>

The same dataset was reused and cleaned. The dataset was grouped by product and it’s total sales (EU,NA and global) via groupby function.

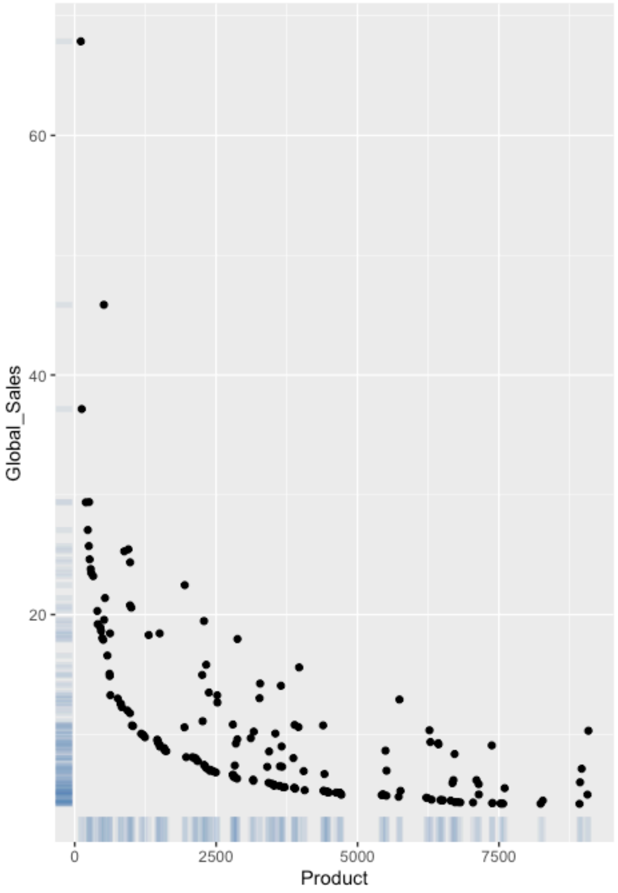


Fig 5.1 (Product vs total Global sales boxplot)

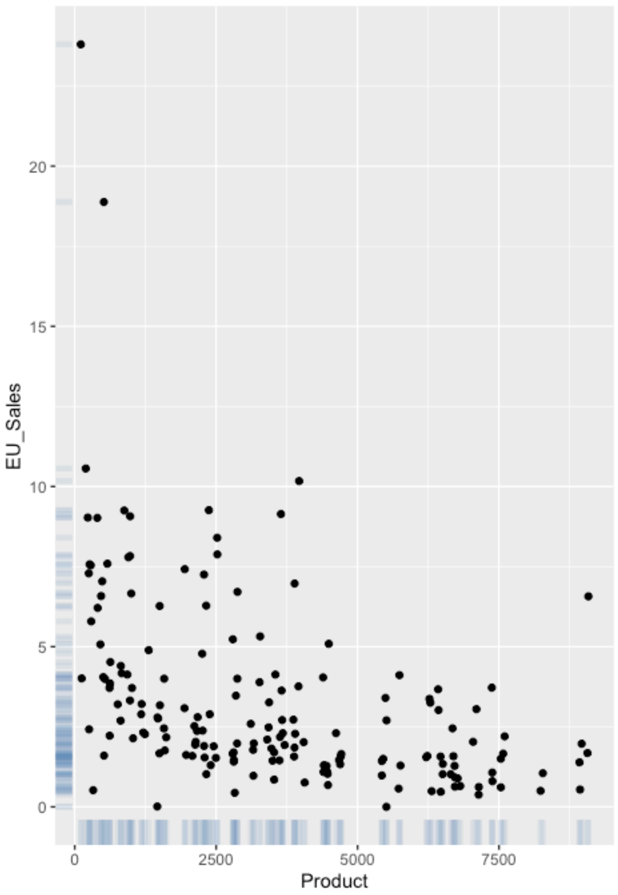


Fig 5.2 (Product vs total Europe sales boxplot)

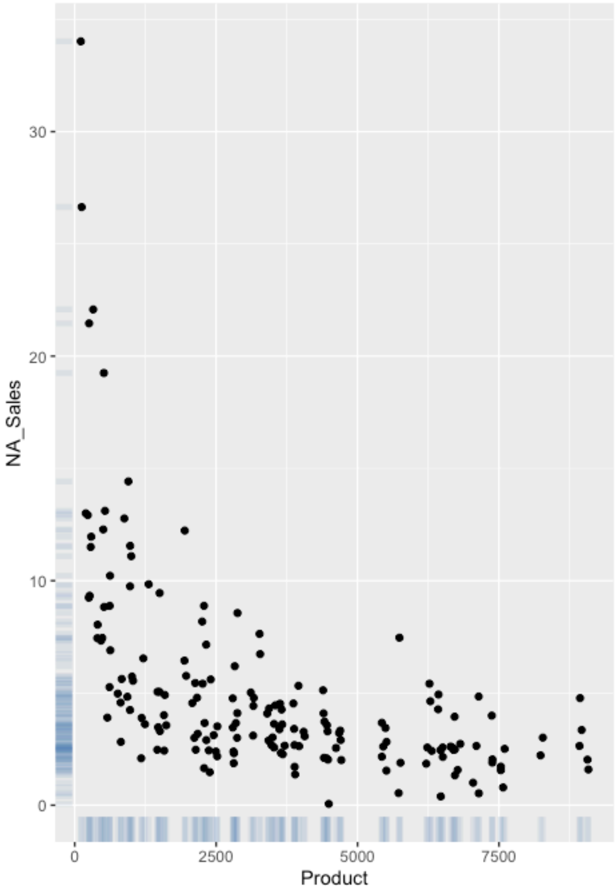


Fig 5.3 (Product vs total North America sales boxplot)

The scatter plot above shows the same trend where most of the products were selling within the achieving about 10,000,000 pounds in total global sales. This trend is the same for both North America and Europe. As the product were sequentially numbered, we also see that the total sales decreases as new product were released.

Q-Q plots were deployed to verify the normality of the datasets.

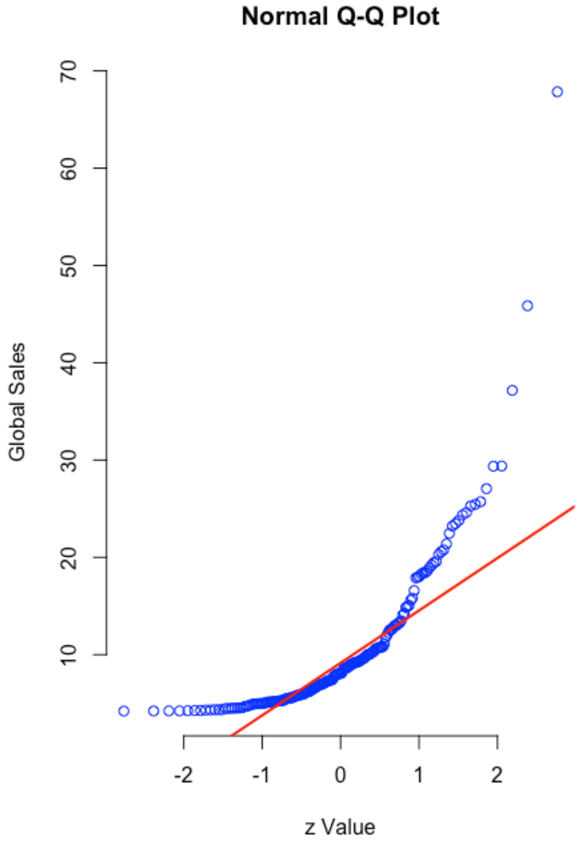


Fig 5.4 (Q-Q plot for total global sales)

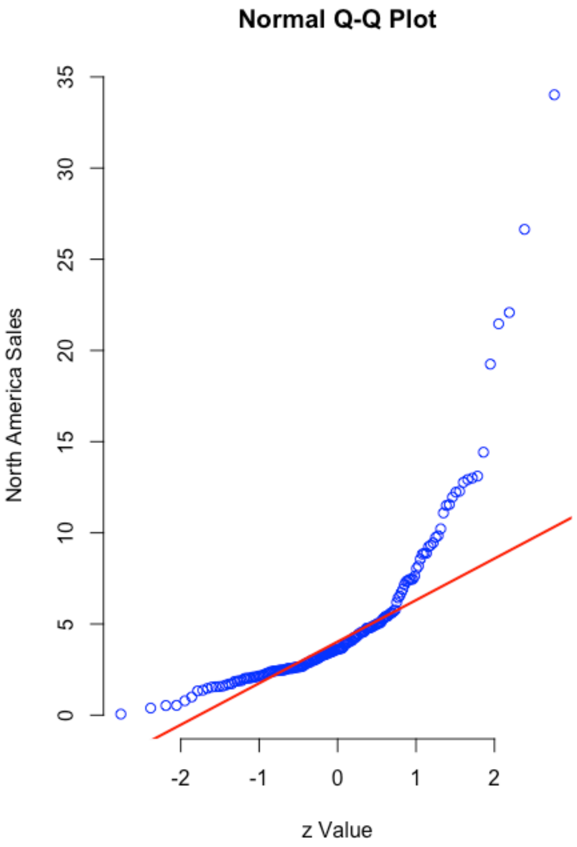


Fig 5.5 (Q-Q plot for total North America sales)

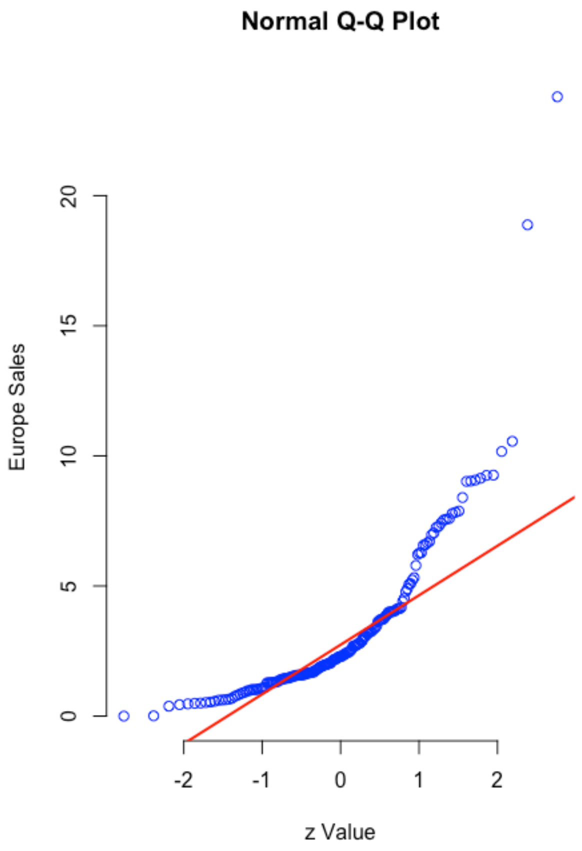
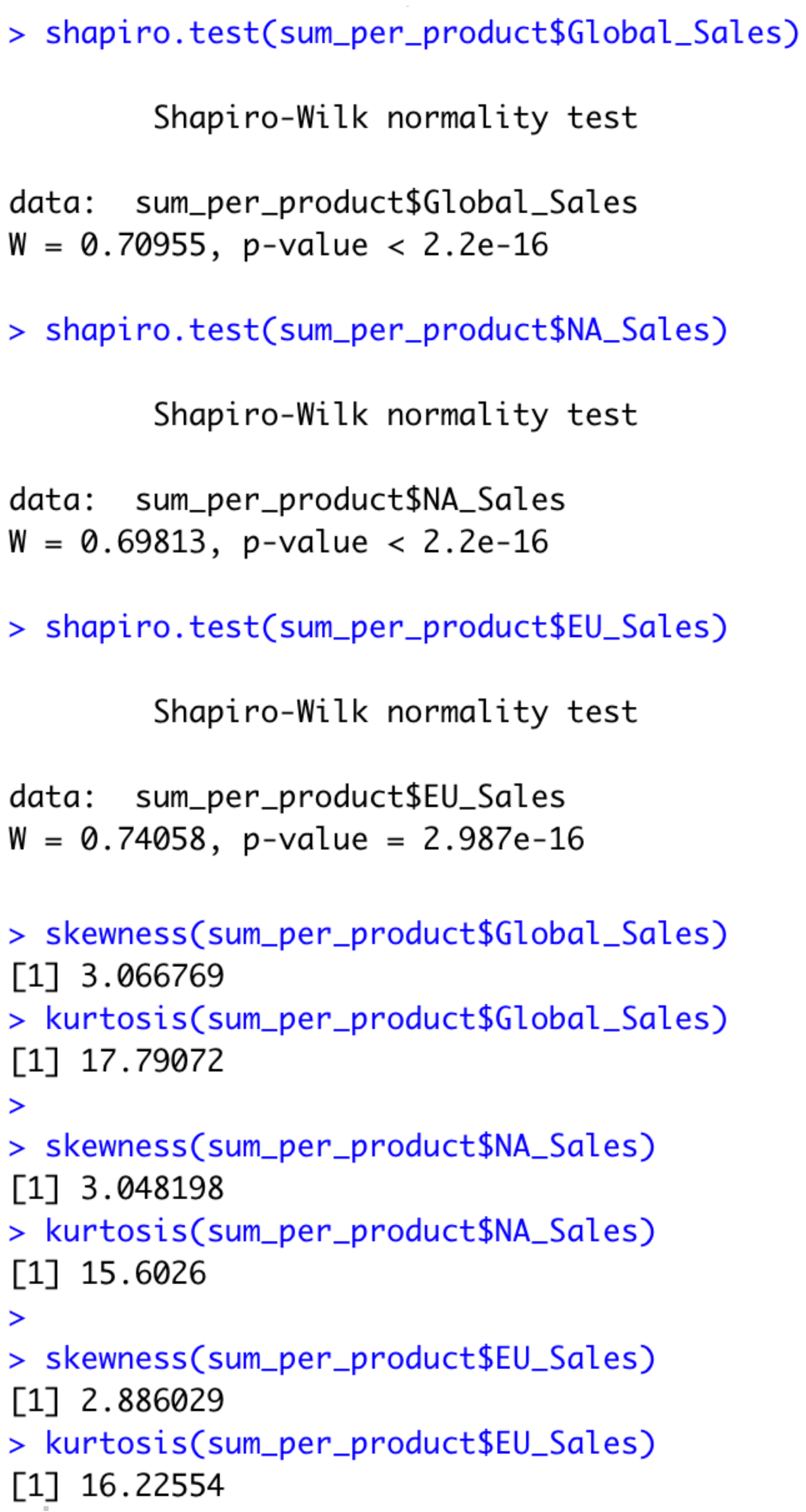


Fig 5.6 (Q-Q plot for total Europe sales)

Performing the Shapiro-Wilk test and determining skewness and kurtosis, the results were as follows:-



```
> shapiro.test(sum_per_product$Global_Sales)

	Shapiro-Wilk normality test

data:  sum_per_product$Global_Sales
W = 0.70955, p-value < 2.2e-16

> shapiro.test(sum_per_product$NA_Sales)

	Shapiro-Wilk normality test

data:  sum_per_product$NA_Sales
W = 0.69813, p-value < 2.2e-16

> shapiro.test(sum_per_product$EU_Sales)

	Shapiro-Wilk normality test

data:  sum_per_product$EU_Sales
W = 0.74058, p-value = 2.987e-16

> skewness(sum_per_product$Global_Sales)
[1] 3.066769
> kurtosis(sum_per_product$Global_Sales)
[1] 17.79072
>
> skewness(sum_per_product$NA_Sales)
[1] 3.048198
> kurtosis(sum_per_product$NA_Sales)
[1] 15.6026
>
> skewness(sum_per_product$EU_Sales)
[1] 2.886029
> kurtosis(sum_per_product$EU_Sales)
[1] 16.22554
```

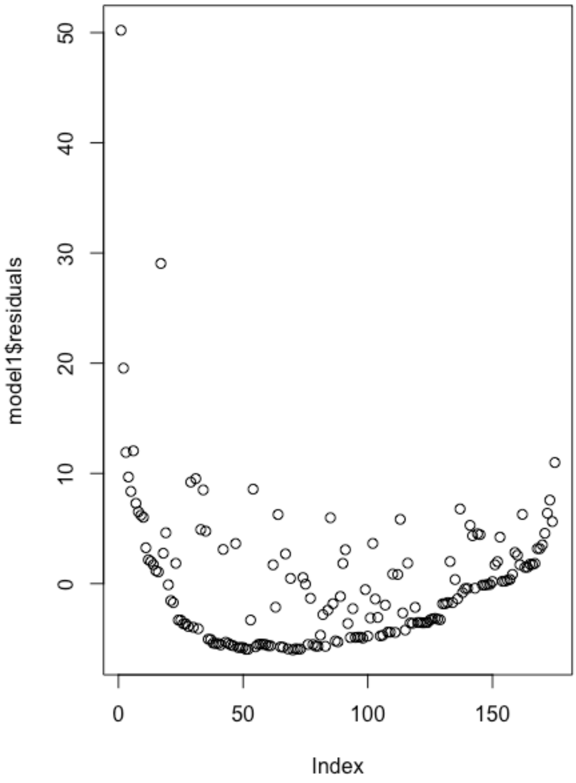
The Q-Q plot indicates a positive correlation. However, for an alpha level of 0.05, all the p values tested were less than 0.05 and rejects the null hypothesis that the data are from a normally distributed population. This is supported by the fact that the skewness of all the sum of sales (global,NA,EU) were not close to zero. The kurtosis values (>3) also supported the fact that the data does not has a peak.

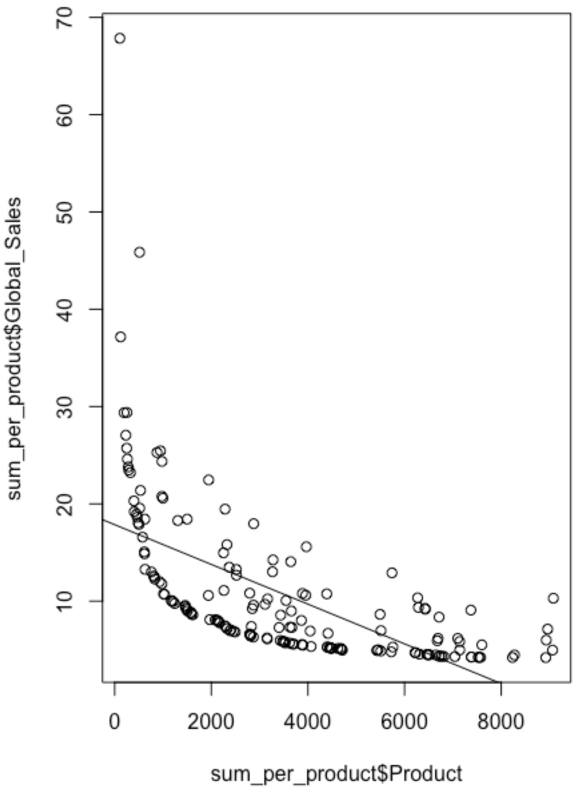
6. Sales analysis between global, Europe and North America

The analysis aims to find out possible relationship between the sales in regions and global and product. A simple Linear regression model was deployed. The results were as follows:-

Product vs Global sales
# p-value: < 2.2e-16; Adjusted R-squared: 0.3637
# F-statistic: 100.5 on 1 and 173 DF

Residual plot and correlation between global sales and product.





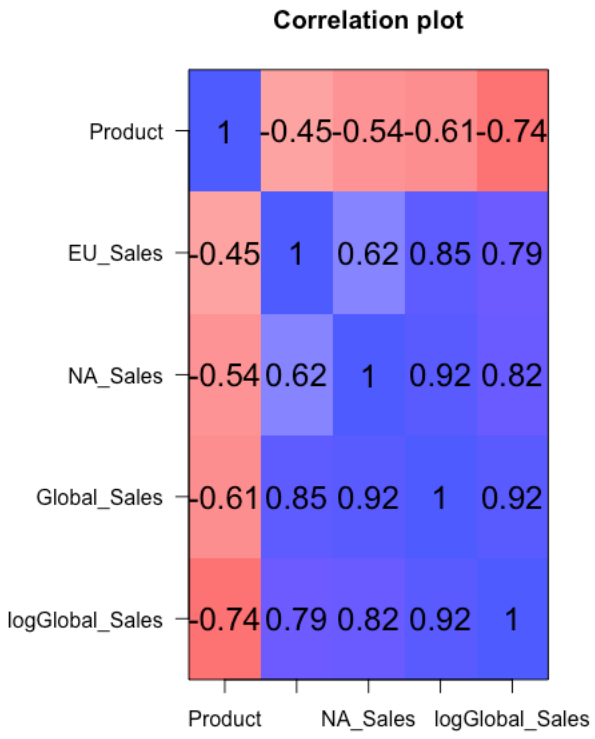
Noted that the linear regression model made a very poor fit as the relationship between Product and Global Sales is not a linear one. To try a log transformation on Global Sales on a 2nd linear model.

The second linear model results were as follows:-

Product vs log(Global sales)
Adjusted R-squared: 0.5459 , p-value: < 2.2e-16
F-statistic: 210.2 on 1 and 173 DF.

Model 2 seems to fit better than model 1 with a higher R-squared value and F-statistic value. The overall shape of the residual and correlation plot remains similar. However, noted that it still does not fit well based on the R-squared value.

A Multiple linear regression model (MLR) was deployed. The correlation plot below also indicated fairly strong correlations between the features.

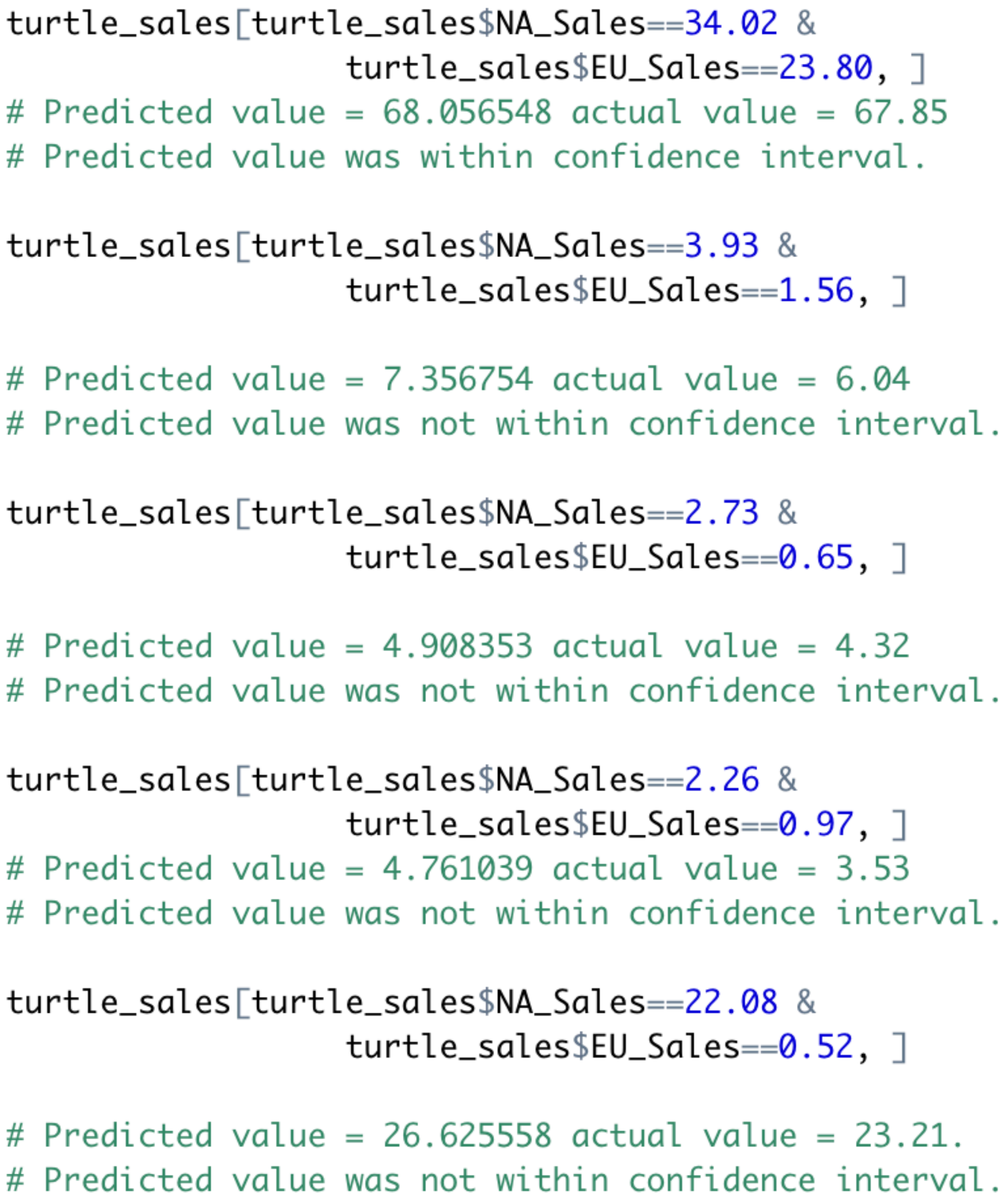


The results from the MLR were as follows:-

<u>Product vs global + EU + NA</u>
Adjusted R-squared: 0.9664
F-statistic: 2504 on 2 and 172 DF, p-value: < 2.2e-16

MLR is a stronger model than the LR model earlier with a higher R-squared value.

Using the MLR model, we predict the global sales of 5 data points of EU and NA sales.
The results were as follows:-



```
turtle_sales[turtle_sales$NA_Sales==34.02 &
                 turtle_sales$EU_Sales==23.80, ]
# Predicted value = 68.056548 actual value = 67.85
# Predicted value was within confidence interval.

turtle_sales[turtle_sales$NA_Sales==3.93 &
                 turtle_sales$EU_Sales==1.56, ]

# Predicted value = 7.356754 actual value = 6.04
# Predicted value was not within confidence interval.

turtle_sales[turtle_sales$NA_Sales==2.73 &
                 turtle_sales$EU_Sales==0.65, ]

# Predicted value = 4.908353 actual value = 4.32
# Predicted value was not within confidence interval.

turtle_sales[turtle_sales$NA_Sales==2.26 &
                 turtle_sales$EU_Sales==0.97, ]
# Predicted value = 4.761039 actual value = 3.53
# Predicted value was not within confidence interval.

turtle_sales[turtle_sales$NA_Sales==22.08 &
                 turtle_sales$EU_Sales==0.52, ]

# Predicted value = 26.625558 actual value = 23.21.
# Predicted value was not within confidence interval.
```

MLR is a stronger model than the LR model earlier with a higher R-squared value. However, when it came to predictions, only 1 predicted value was within the 95% confidence interval of the observed value. There is a 20% chance of accurately predicting the Global Sales. Based on the sample predictions test. Hence, it is likely that the dataset does not fit properly with the linear regression models (simple or multiple). Future tests may be conducted to find other models that fit the dataset more accurately. Further noted that the historic dataset analysis were conducted only for sales and did not account for the product genre, publisher and platform. Hence, if we use more dimensions , the accuracy of the model might change and give a better prediction.