Turtle Games is a global game retailer aiming to improve sales performance. The business has commissioned a report exploring customer trends and sales activity. Turtle Games provided historical sales data and product reviews, which were analysed to capture insights, extrapolate trends and make predictions. The findings presented in this report will improve Turtle Games' understanding of their customers, their loyalty programme, and broader insights across sales and marketing. This report is a methodological review of the analytical process and findings, for detailed business recommendations refer to the stakeholder presentation.

Turtle Games posited six questions which have been expanded upon to generate an analytical framework, which is as follows:

Turtle Games Query	**Supporting Queries**	**Relevant File**	**Analytical Method**
How groups within the customer base can be used to target specific market segments.	Is there any relationship between the loyalty programme and other variables? Can customers be segmented?	Reviews	Linear Regression with OLS in Python *K*-Means Clustering in Python
How social data (e.g. customer reviews) can be used to inform marketing campaigns	What is the sentiment of product reviews? Of product summaries? Are there any trends in popularity versus unpopularity of products?	Reviews	NLP Sentiment Analysis in Python NLP Subjectivity & Polarity Analysis in Python
The impact that each product has on sales	Which products sell well? What is the overall spread of product sales?	Sales	Plots in R
How reliable is the data?	What is the distribution, skewness, and kurtosis? What does this mean for analysis?	Sales	Shapiro-Wilks test in R QQ Plots in R
Are there any relationships between North American, European, and global sales?	What is the nature of the relationship? Can this be used to reliably predict future sales?	Sales	Plots in R Simple and Multilinear Regression in R

The team developed supporting queries. Queries are tabulated in the order they were investigated.

With an analytical framework established**,** exploratory analysis was conducted to ascertain descriptive statistics and the distribution of the data, followed by wrangling. This began with exploring the 'Reviews' CSV, which contained the following values.

Reviews CSV	**Age**	**Remuneration**	**Spending Score**	**Loyalty Points**
Count	2000	2000	2000	2000
Mean	39.50	48.08	50	1578.03
STD	13.57	23.12	26.09	1283.24
Min	17	12.30	1	25
25%	29	30.34	32	772
50%	38	47.15	50	1276
75%	49	63.96	73	1751.25

Max	72	112.34	99	6847

The columns relevant for analysis were stored in an alternative data frame, then cleaned and standardised. Where necessary, additional cleaning was performed in other temporary data frames per the requirements of analysis. The 'Sales' CSV was cleaned and analysed using a different methodology to determine statistical reliability, detailed in the latter portion of this essay.

With the benchmark statistics for Reviews established, trends were analysed across customer spend, remuneration, and age respectively, versus loyalty. These variables were first plotted to determine signs of linearity, then a linear regression was conducted using the OLS model. The intercept 'b' and coefficient 'β' were input into equation $y = \beta X + b$ to build the model. The results were plotted to visualise predicted outcomes (Appendix A). The R-squared, or proportion of variance in each model, is summarised below.

Independent Variable (X)	**Dependent Variable (y)**	**R-Squared value**	**Observations**
Spending	Loyalty	0.45	Somewhat reliable
Remuneration	Loyalty	0.38	Less reliable than other models but still somewhat reliable
Age	Loyalty	0.002	Highly unlikely to be reliable, requires further investigation

No model was wholly conclusive in relationship, but spending and remuneration did show a significant enough R-squared value to indicate a linear relationship and some degree of reliability in predicting future outcomes.

Next, data was analysed using a *K*-Means Clustering method, to ascertain whether customers could be segmented. Customer spending scores and remuneration were plotted on a series of charts, which initially indicated clustering. To determine the number of clusters to model, the elbow and silhouette methods were employed, both suggesting 5 clusters were optimal. Two clustering tests were conducted where $k = 4$ and 5 respectively, and the model of best fit was $k = 5$ (Appendix B). The results of these models were more conclusive than the OLS models, producing 5 distinct and relatively predictable clusters in spending score and remuneration.

The next step was to determine sentiment in product reviews and summaries. To prepare data for NLP analysis, additional cleaning was employed and stored in a secondary data frame, which converted the textural data in the 'Review' and 'Summary' columns into strings. These strings were tokenised and stop words were removed to generate respective word clouds (Appendix C). The word clouds visualised an ostensibly positive overall sentiment, and additional frequency scores corroborate this. Subsequently, 'Review' and 'Summary' sentiment scores were plotted on histograms (Appendix C), with both showing a positive skew in score.

To counterbalance any bias in the reviews, subjectivity scores were also measured and plotted (Appendix C). The modal subjectivity score for reviews was .5, indicating some degree of subjectivity in the data. Subjectivity scores for summaries appeared more objective, with a modal score of 0. The top 20 positive and negative comments were cross examined against their respective subjectivity scores and tabulated.

Category	**Average Polarity (out of 20)**	**Average Subjectivity (out of 20)**

Positive Reviews	1.0	0.89
Negative Reviews	-0.49	0.66
Positive Summaries	1.0	0.86
Negative Summaries	-0.68	0.76

Overall, the scores reveal several trends in customer behaviour. Firstly, the most positive comments were higher in polarity than the most negative but also on average more subjective. This is not conclusive, but would suggest that those who feel positively about products do so because of sentimentality. Conversely, negative comments are less polarised but also less subjective on average. These scores, as well as the additional insights gleaned from the NLP analysis, provide myriad insights to support marketing, detailed in the recommendations.

Following this, the 'Sales' CSV was explored to generate insights pertaining to product sales and the relationship between different regions. To do so, it was first necessary to establish the reliability of the data, and whether this would impact modelling. To begin, a DataExplorer export was generated, which tested for missing values, outliers, and the overall descriptive statistics of the data set. The data was then cleaned by creating several temporary data frames to aid visualisation. It was noted during initial exploration that there were a large variety of gaming platforms recorded in the data, which limited the interpretability of visualisations. To understand macro sales trends, the 'Platforms' column was remapped using the recode R function. The additional column united platforms by brand and reduced the number of variables from 22 to 10 (Appendix D). A box-plot was generated to determine the range of data across global sales, which also highlighted outliers.

From exploratory analysis, several trends in product sales became apparent. Firstly, sales figures were grouped by product ID and plotted, showing a distinct negative correlation between product ID and sales, implying the ID is arranged in rank order. This also highlighted the significant range of data:

Sales CSV	**North America**	**Europe**	**Global**
Mean	2.51	1.64	5.33
Maximum	34.02	23.8	67.85
Minimum	0	0	0.01

As evidenced in the table, there is a wide berth between the most popular products (in sales) and the least. The most popular products were the same for all regional and global sales. Shown in the histograms in Appendix D, the highest selling item is Product 107 for the Wii platform.

Given the distribution of the data and the outliers, it was necessary to further establish the reliability of the data. Q-Q plots were generated to determine if the sales data followed a normal distribution (Appendix E). The resulting plots suggested an abnormal distribution, so a series of follow up tests were conducted, including Shapiro-Wilk, skewness, and kurtosis, producing the following results.

Data Set	**P-value (Shapiro-Wilk)**	**Skewness**	**Kurtosis**
Global Sales	< 2.2e-16	4.04	32.64
NA Sales	< 2.2e-16	4.31	31.37

EU Sales	< 2.2e-16	4.82	44.69

As the P-value is < .05 in all three data sets, we reject the null hypothesis and assume the data is not normally distributed. The data is also highly positively skewed and is leptokurtic, containing several significant spikes in the data. The results do not significantly hinder the applicability of models, but it is important to factor this information into the findings of this study.

With the limitations of the dataset established, the next step was to assess the relationship between regional and global sales. The correlation plot (Appendix F) highlights several strong positive relationships which demonstrated the viability of regression models, as the variables implied a relationship between regional sales and their influence on global sales. A series of simple linear regression model were then fitted to the variables, which combined with the correlation tests, produced the following results:

Independent Variable (x)	**Dependent Variable (y)**	**Correlation**	**Significance**	**Adj. R-Squared**
North America Sales	Global Sales	0.93	< 2e16 ***	0.87
Europe Sales	Global Sales	0.88	< 2e16 ***	0.76
North America Sales	Europe Sales	0.71	< 2e16 ***	0.50

The results produced statistically significant findings, suggesting the strongest relationship between North American Sales and Global Sales. The regression values were then plotted to predict future values (Appendix F). To validate the findings of the simple linear regression model, a multilinear regression was fitted with the aforementioned variables and produced an adjusted R-Squared value of .96, which implies that almost all of global sales behaviour can be explained and predicted using the regional sales variables. The strong statistical validity of the multilinear regression model makes it a powerful predictive tool. Using the formula $y = \beta X + b$ and corresponding values (Appendix F), future performance can be modelled.

The insights from this study can be used to support the business in multiple ways by highlighting consumer trends and attitudes, and predicting future sales performance. There are several key recommendations derived from these insights, summarised in this report and further detailed in the stakeholder presentation. Firstly, given the distinct customer segments identified in the *K*-means model, tailoring marketing and sales activities, such as the loyalty programme, to these distinct groups could be a source of additional revenue. For example, targeting the cluster with the highest spending score with promotional offers and the cluster with the lowest spending scores with reduced-price items, such as in a sale. The clustering and OLS models provide the basis for further investigation into the loyalty programme and how this can be tailored to consumer types.

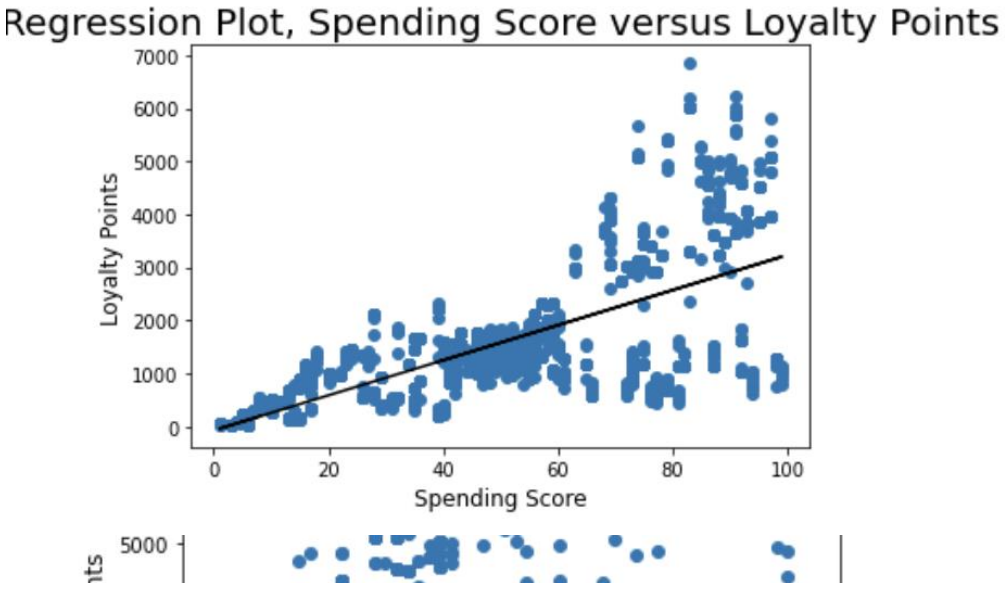
The NLP analysis also generated actionable insights. Firstly, there is valuable data in positive reviews, with the overall skew towards positive sentiment suggesting that Turtle Games have more favourable products than unfavourable. There is also useful insight within the negative reviews and summaries given their lower subjectivity scores, implying higher objectivity such as product functionality. These negative reviews can be parsed for information pertaining to defective products to manage inventory. An area for further exploration could be grouping negative reviews to investigate specific concerns, such as around a faulty product.

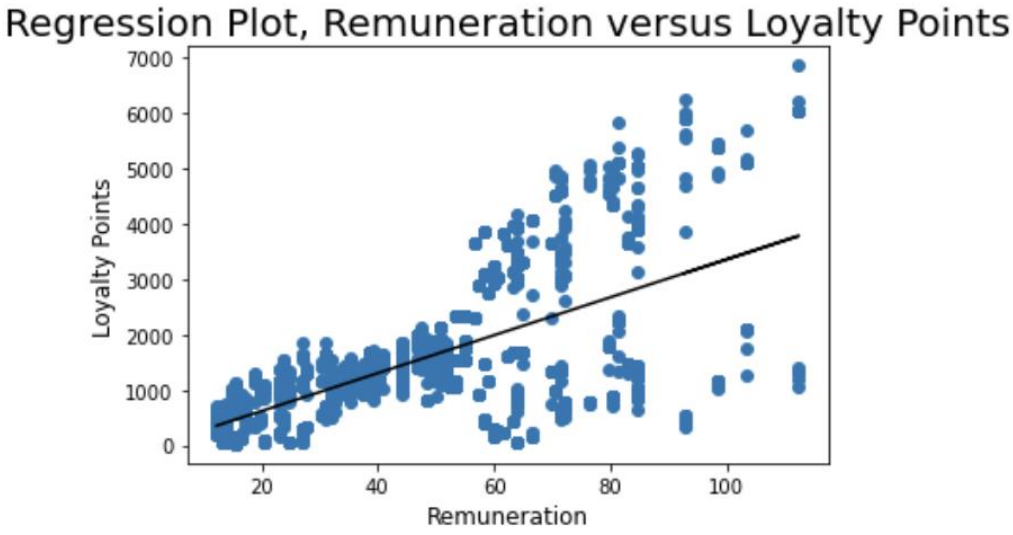
Despite the limitations of the sales data in terms of distribution and skew, there are still useful insights presented. Firstly, given the spread between the most and least popular products, with several products not generating any revenue, in order to improve sales performance Turtle Games should investigate inventory and potentially reduce stock of less popular items. Furthermore, there is a strong relationship between regional and global sales, which implies that regional sales can be

leveraged to predict future performance. For example, product launches can be trialed regionally with their sales performance as a strong indicator of their success on the global scale. While additional research would be required to ascertain precisely how well a product would sell globally, this model presents a strong foundational basis for predictions.

With an ambition to support Turtle Games' sales activities, a rigorous and multi-platform analysis was conducted. This report captures consumer trends regarding Turtle Games' customers, their preferences, and how to tailor sales activities towards them both at an individual and regional level. To ensure the findings were relevant and actionable, strategic recommendations were implemented and advice on suggested additional investigation that could bolster the findings of this analysis was included.

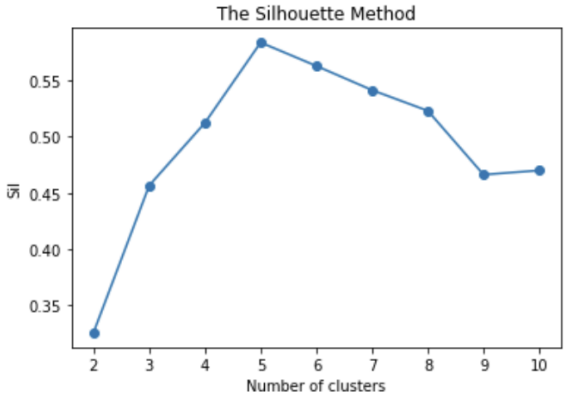
Appendix A: OLS Models

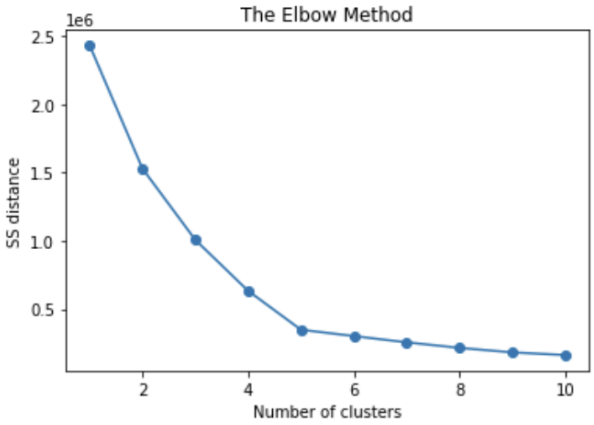




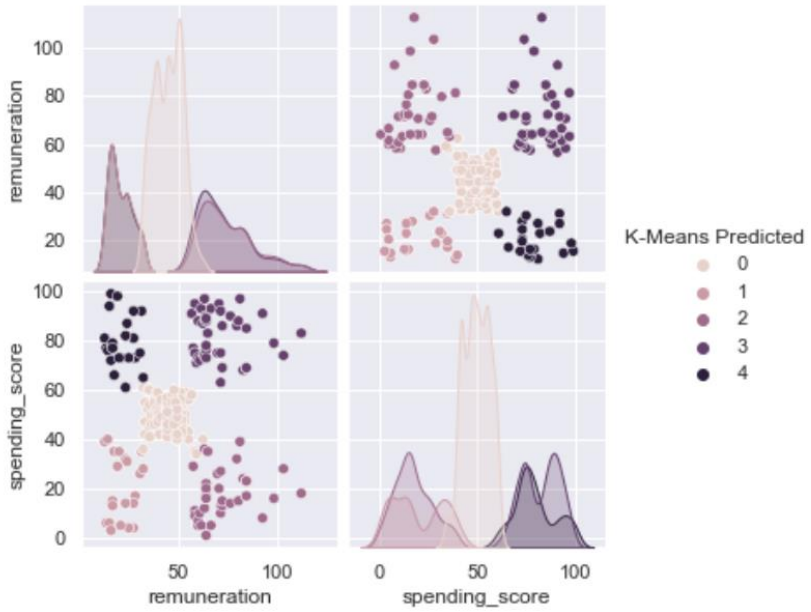
Appendix B: Clustering Models

Appendix B1: The Silhouette & Clustering Methods

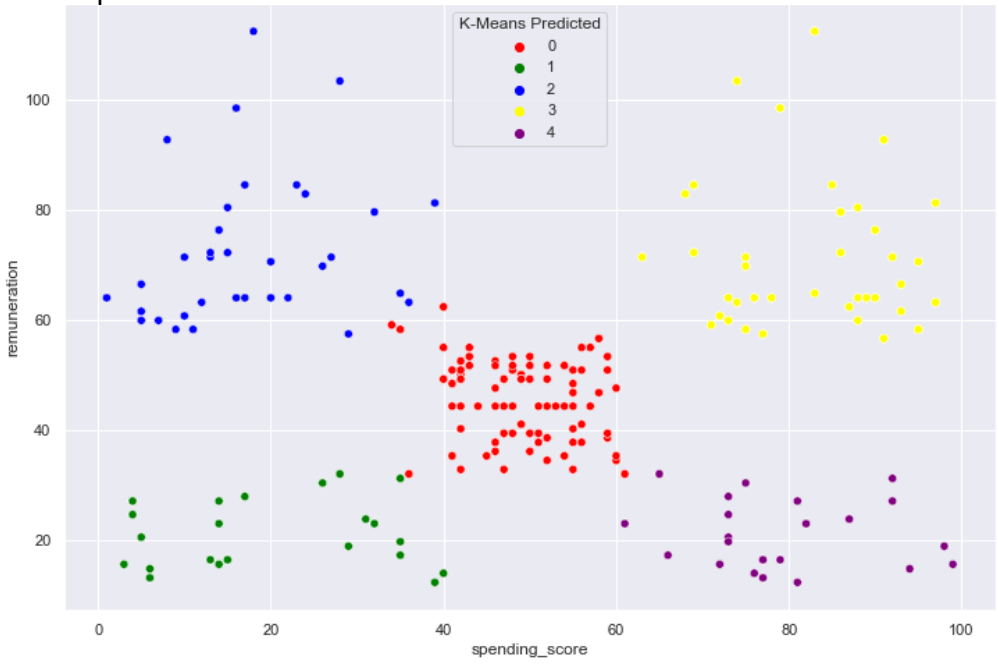




Appendix B2: Pair plot of data where *K*=5

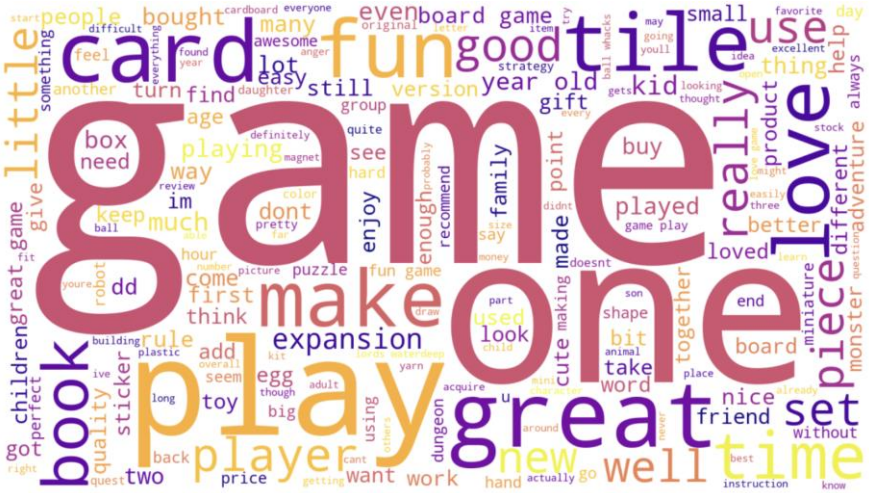


Appendix B3: Scatterplot of *K*-Means Clusters where *K*=5



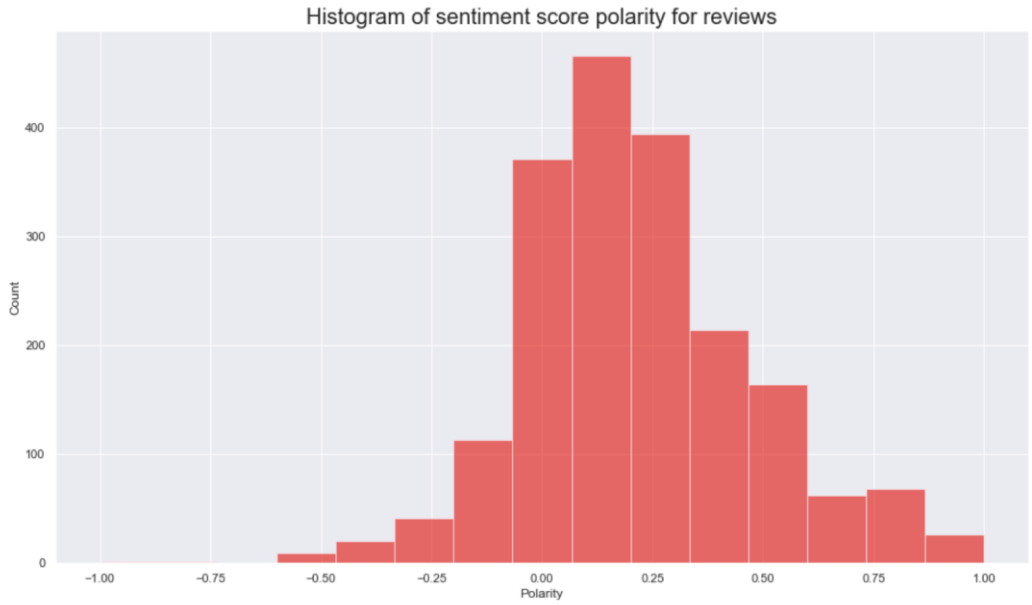
Appendix C: NLP Models

Appendix C1: World Clouds for Review

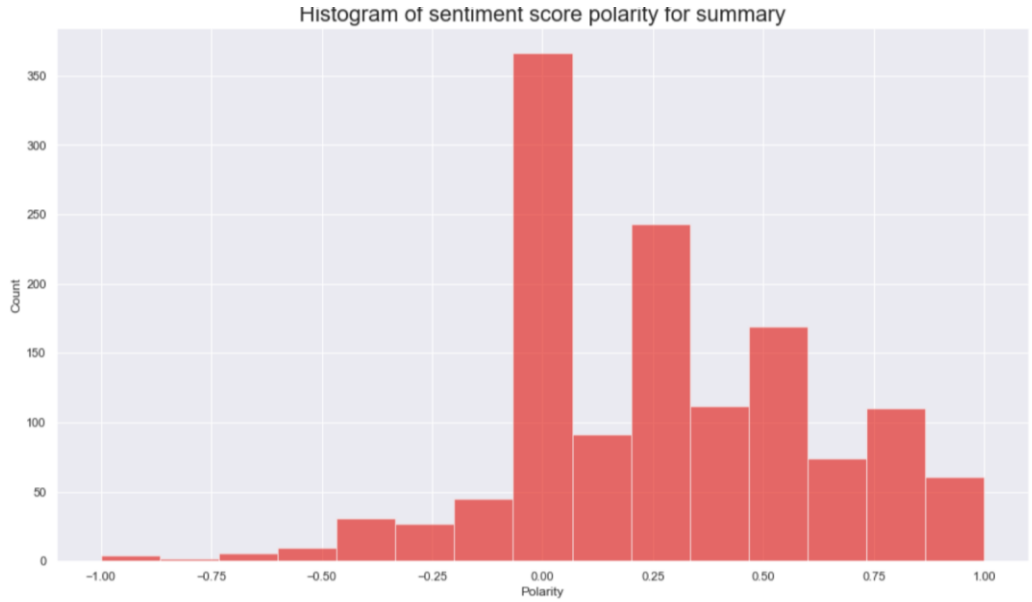


Appendix C2: Word Cloud for Summary

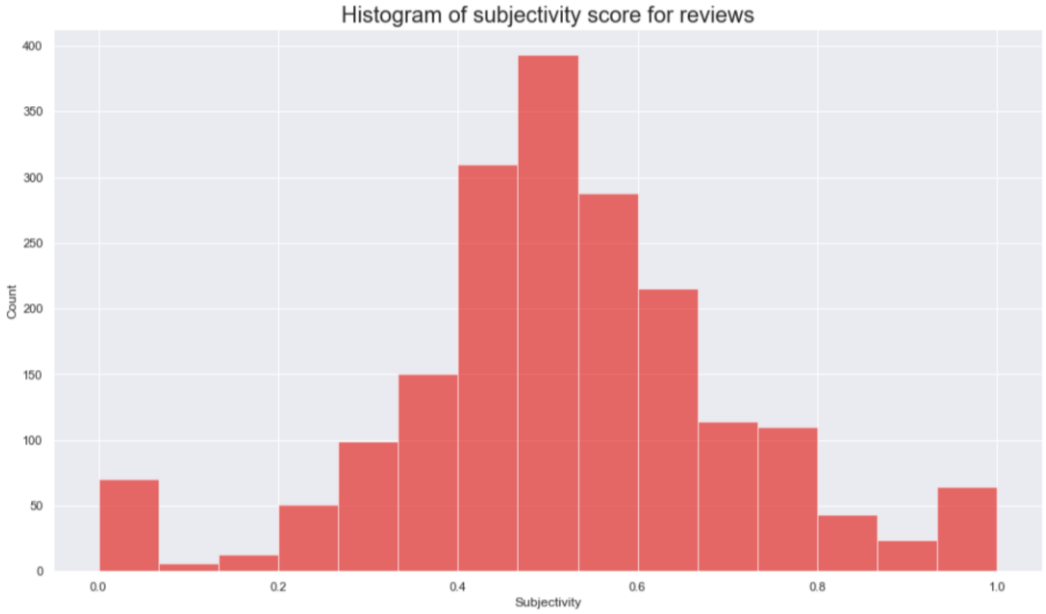
Appendix C3: Histogram of Polarity score for Reviews



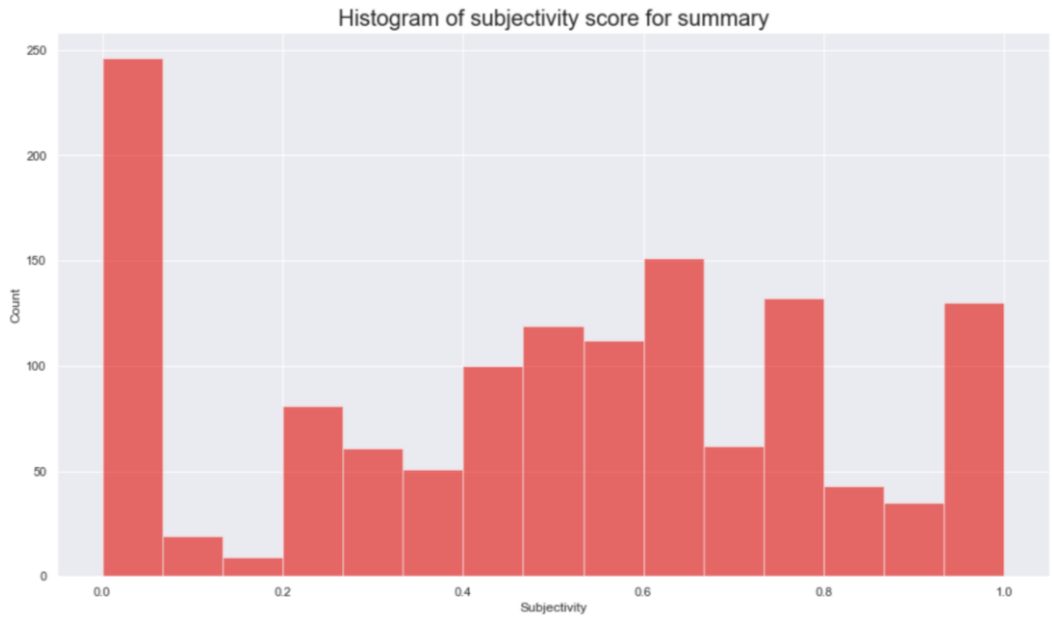
Appendix C4: Histogram of Polarity score for Summaries



Appendix C5: Histogram of Subjectivity score for Reviews

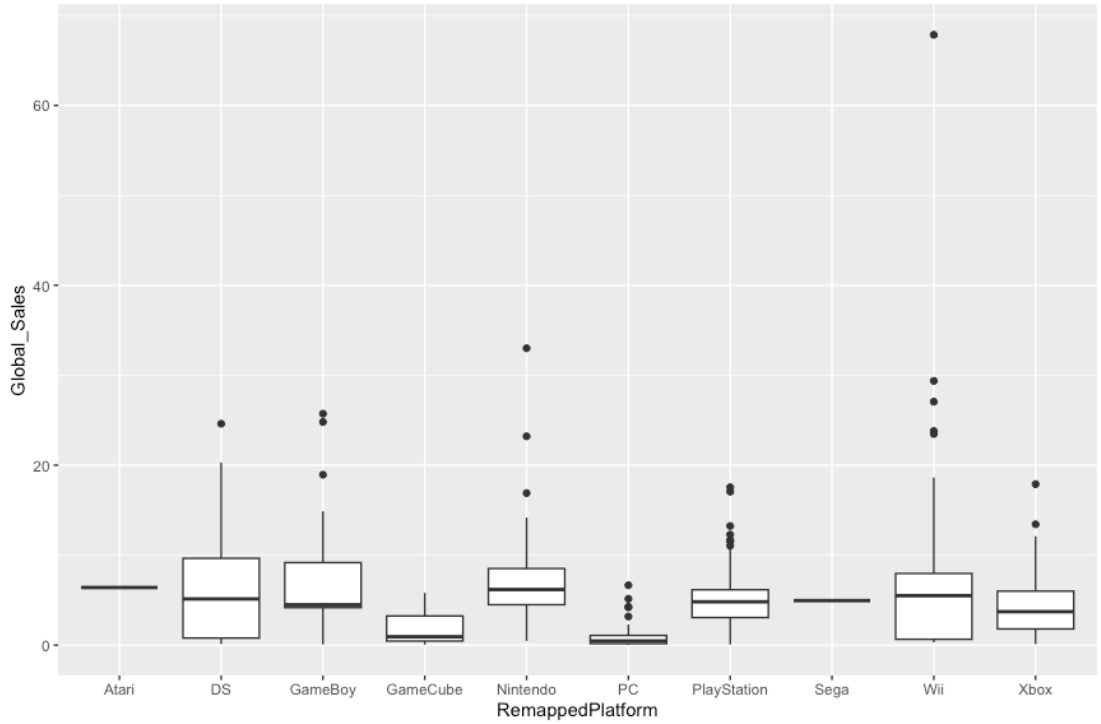


Appendix C6: Subjectivity score for Summaries

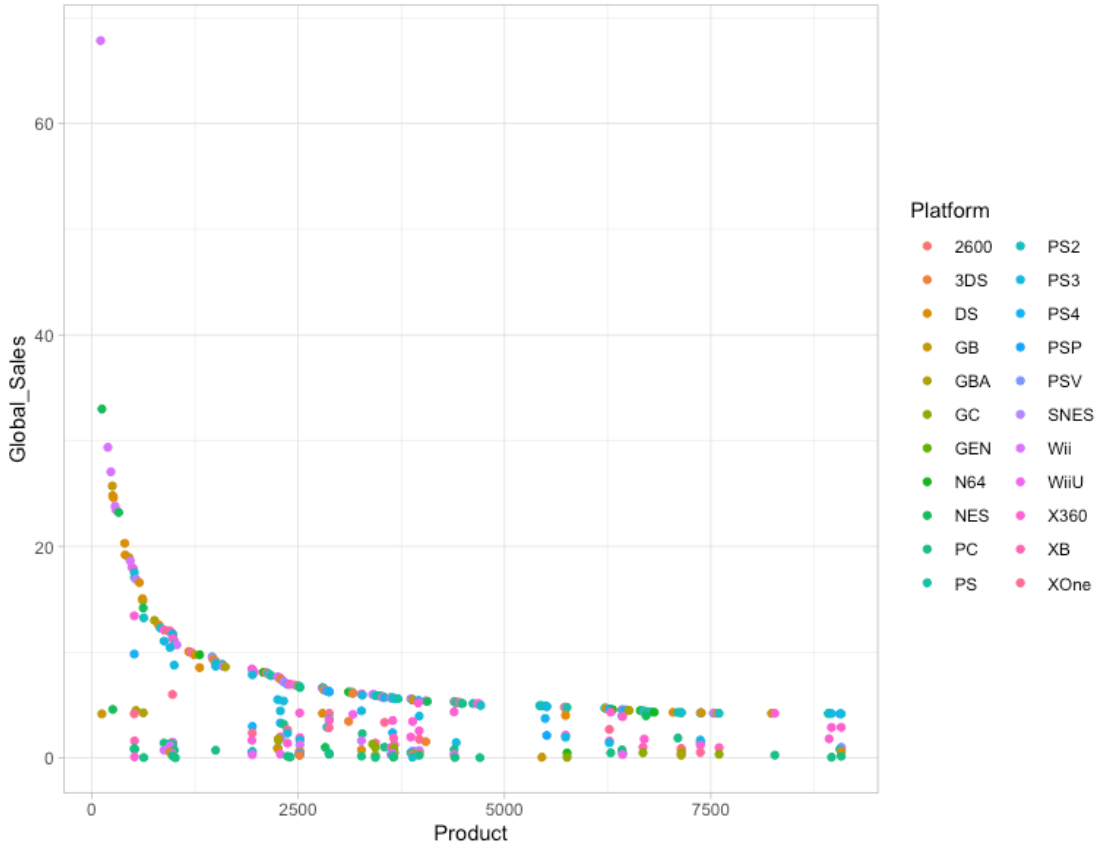


Appendix D: Plots in R

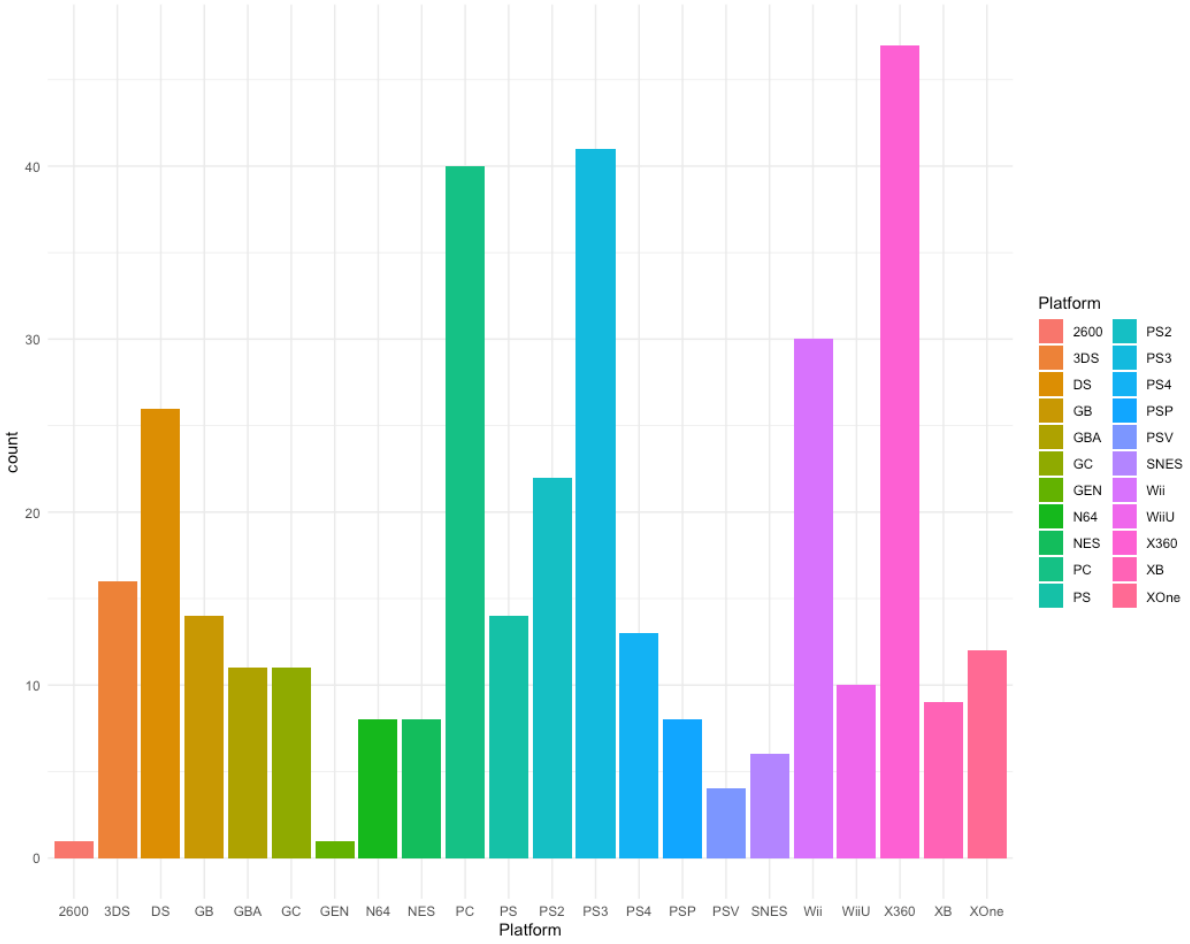
Appendix D1: Box-and-Whisker Plot of Distribution Across Platforms



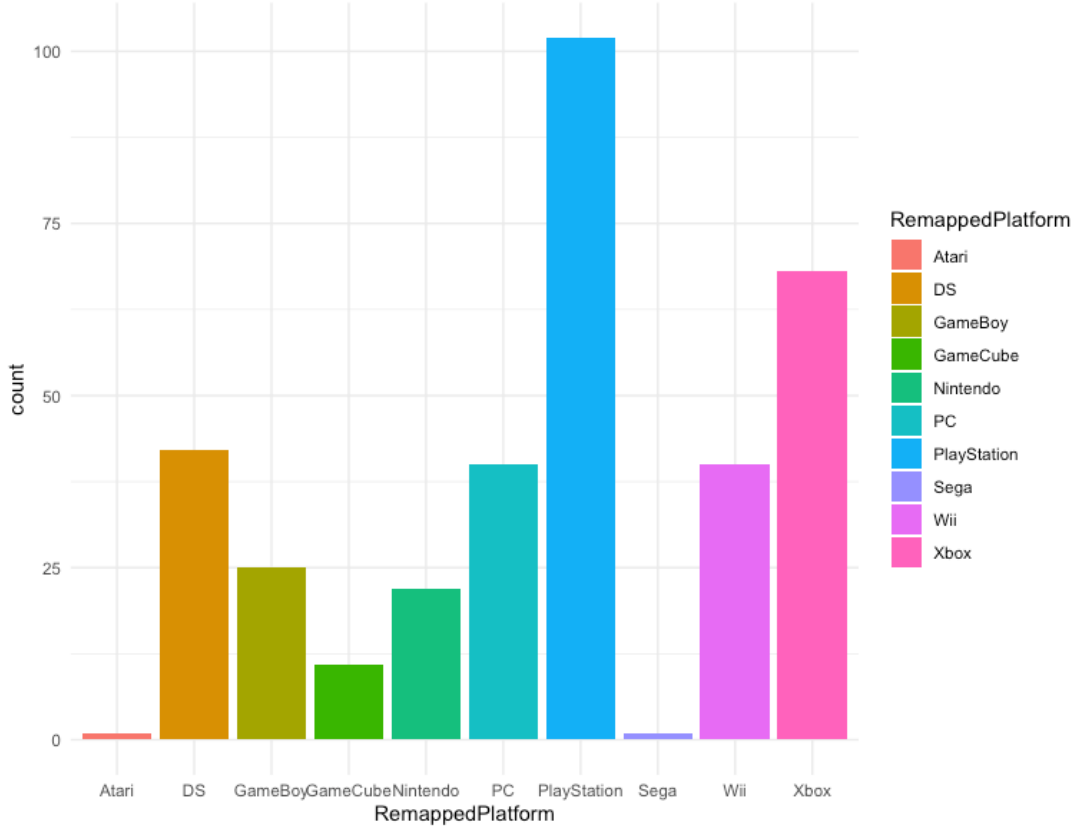
Appendix D2: Product Number versus Global Sales, Categorised by Platform



Appendix D3: Number of Products in Inventory per Platform

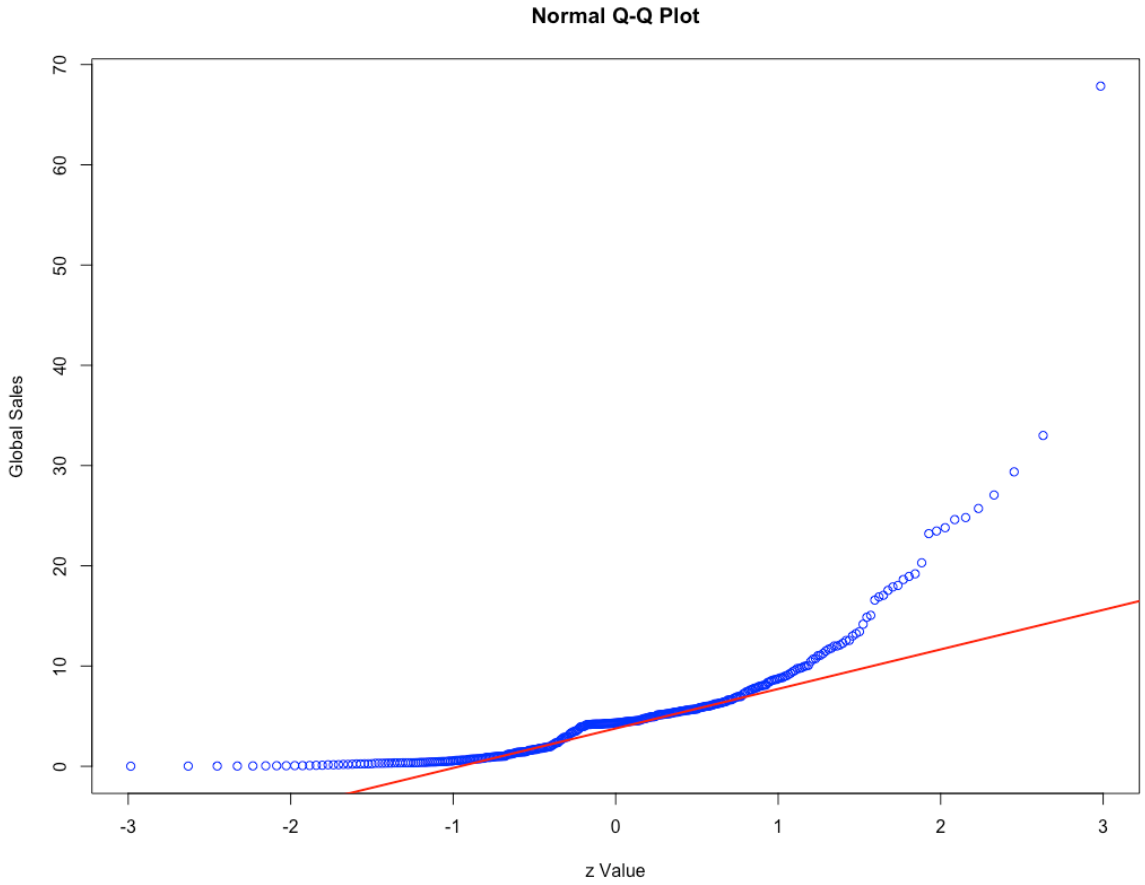


Appendix D4: Number of Products in Inventory per Remapped Platform

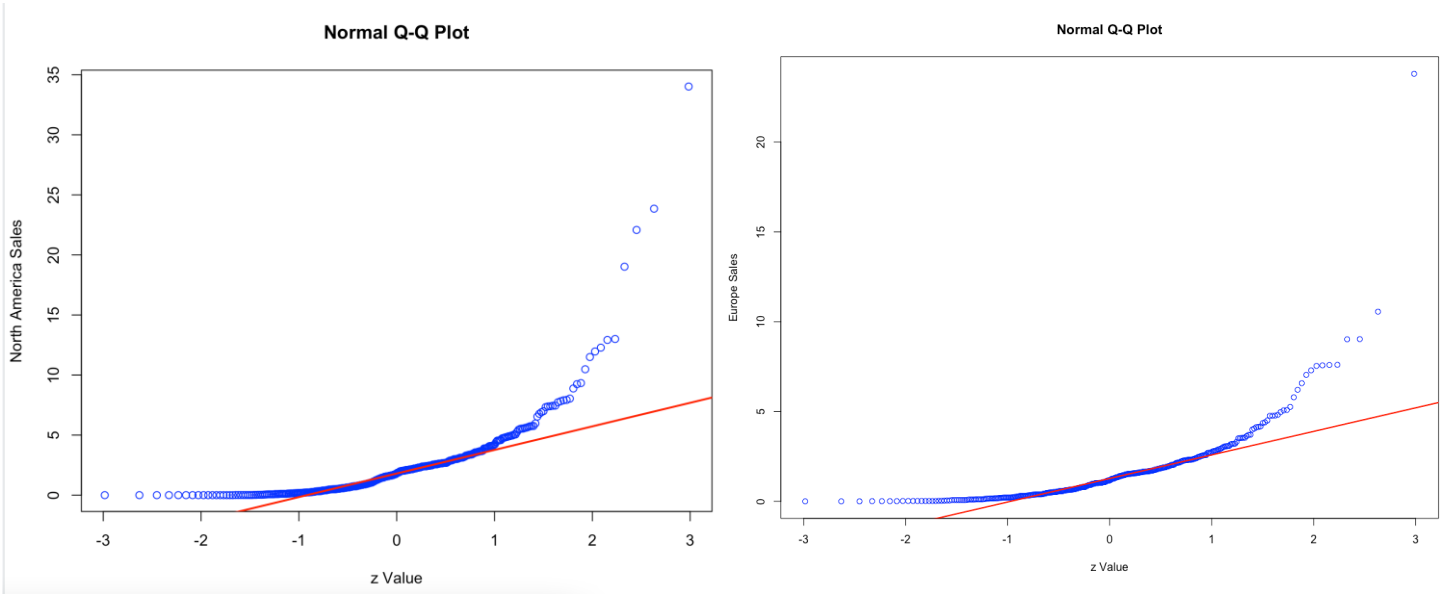


Appendix E: Q-Q Plots

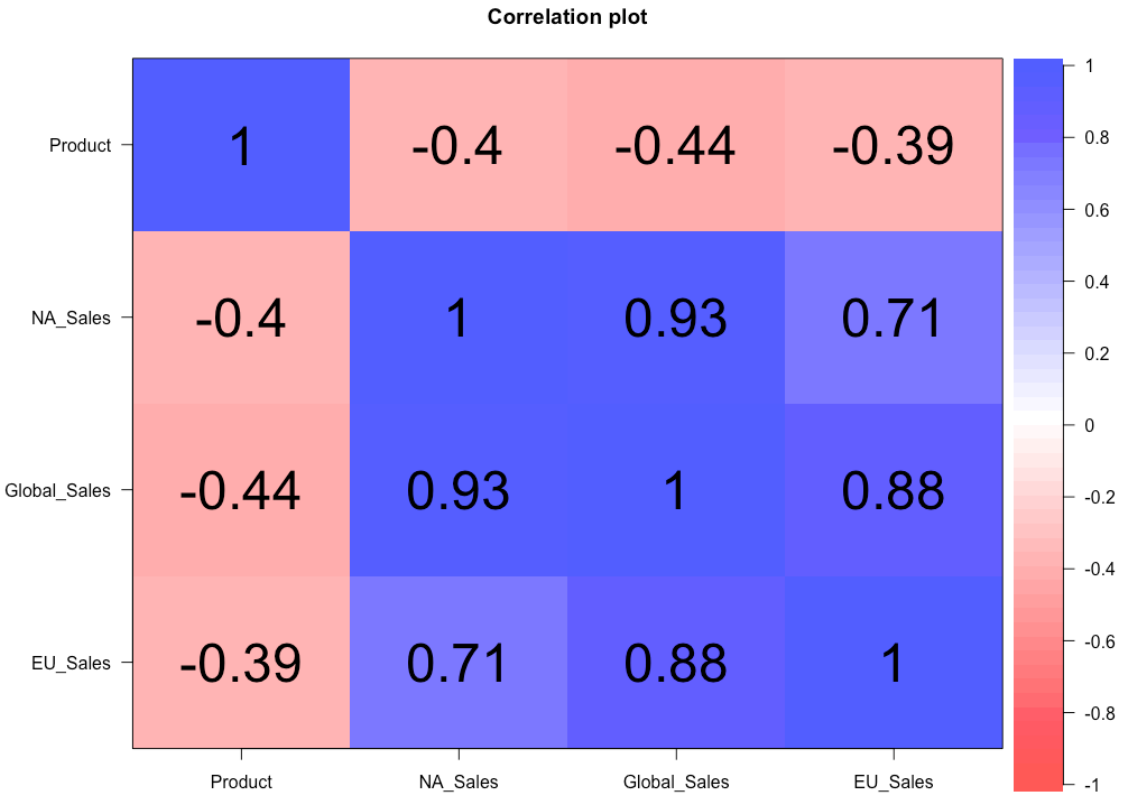
Appendix E1: Q-Q Plot of Global Sales



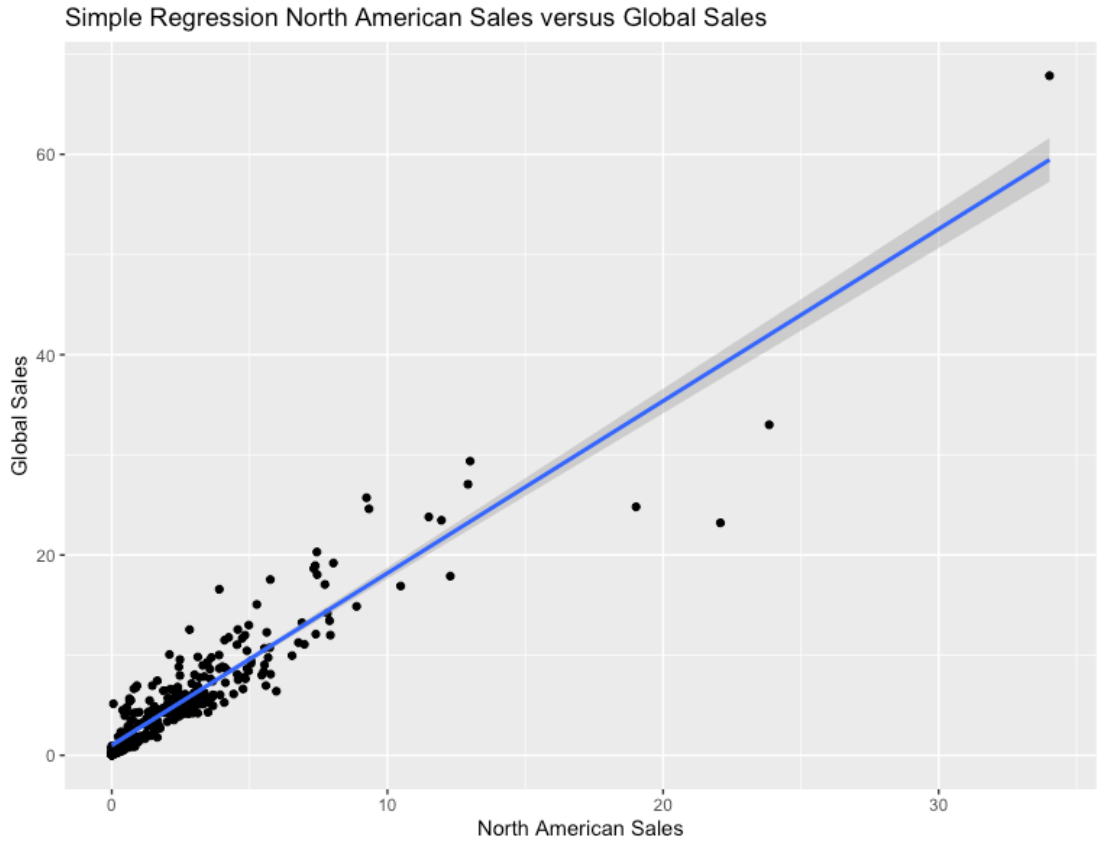
Appendix E2: Q-Q Plots of NA and EU Sales



Appendix F: Simple & Multilinear Regression Models
Appendix F1: Correlation Plot



Appendix F2: Simple Linear Regression North America & Global Sales



Appendix F3: Simple Linear Regression Europe & Global Sales