Luke Ogilvie Thompson

Data Analytics Career Accelerator

Course 3: Technical Report

Background

This report forms part of a wider analytical project assisting Turtle Games' marketing and sales departments.

The marketing department were interested in:

- Accumulation of Loyalty Points.
- Looking at customer clusters.
- Analysing customer sentiments with reviews.

The sales department were interested in:

- The impact of sales per product.
- The reliability of the data.
- Identifying relationships between NA and EU sales with Global salesd.

Turtle Games provided two csv files:

1. turtle_reviews.csv
2. turtle_sales.csv

Metadata was also provided in text format:

3. metadata_turtle_games containing metadata relating to both turtle_reviews.csv and turtle_sales.csv explaining the makeup of the data.

Analytical approach

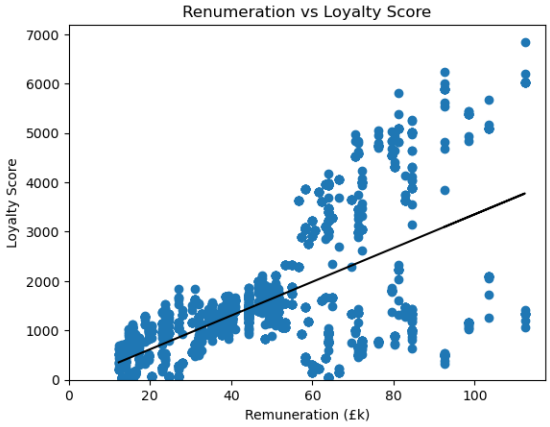
Reviews Data

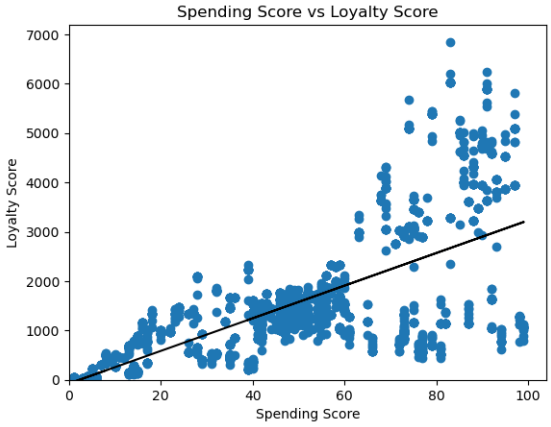
The reviews data was imported and cleaned in Python. It was checked for missing values, duplicates and explored to get an understanding of the nature of the data. The cleaned file was saved as a new csv file for future reference.

Accumulation of Loyalty Points

Loyalty Points, Spending Score, Remuneration, and Age's distributions were checked. Only Age was approximately normally distributed.

Loyalty Points was designated as the depended variable and plotted against each of the other variables with them as the independent variables. Age clearly showed no correlation with Loyalty Points. Neither Spending Score nor Renumeration appeared to have a linear relationship with Loyalty Points, but they did appear loosely correlated. Linear regression with OLS was used to investigate the relationships:





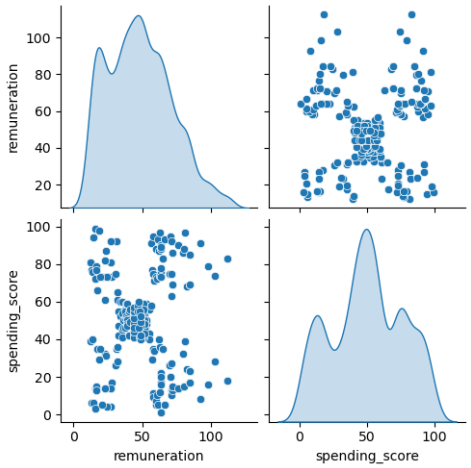
Spending score had a stronger ability to predict Loyalty Score than Renumeration did.

Using both Spending Score and Renumeration together showed an improvement in predicting Loyalty Points and the independent variables were shown to not be multicollinear.

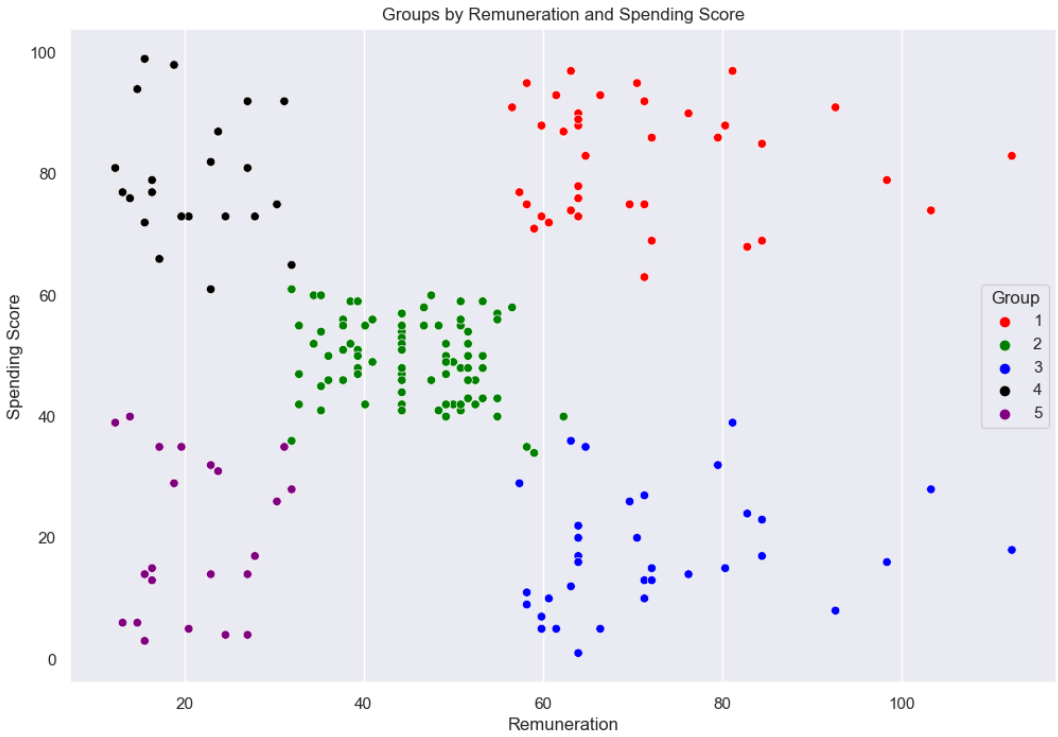
However, expected errors were quite considerable and one should use caution if trying to predict Loyalty Points with Spending Score and Renumeration.

Customer Clustering

Renumeration and Spending Score were compared using a pair plot. This initially suggested a possible five groups.



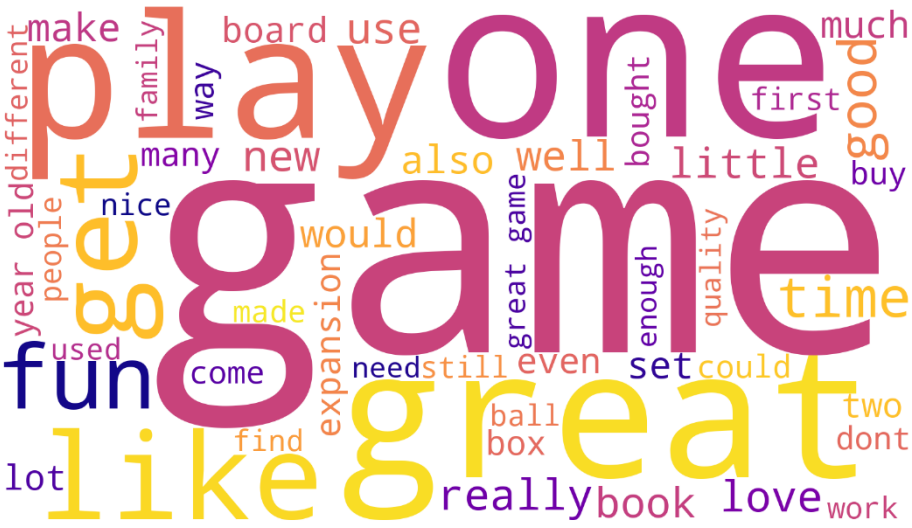
They were then compared using Kmeans clustering to identify and test the appropriate number of groups. The Elbow and Silhouette methods both indicated five clusters as well. To confirm this, three options were tested with four, five and six clusters. Five clusters was identified as the best option:

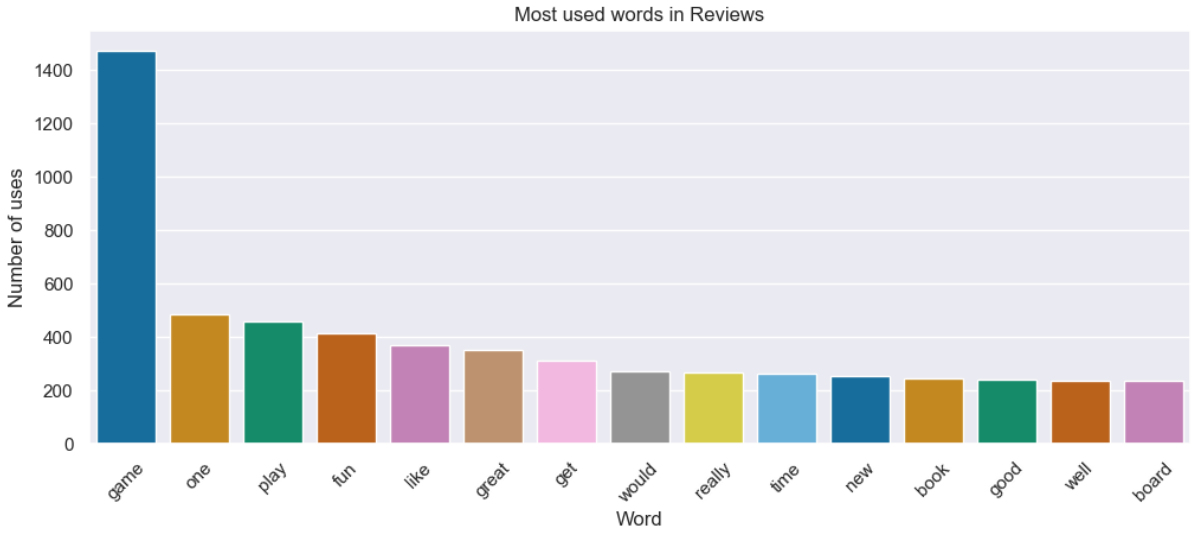


Analysing customer sentiments with reviews

WordClouds without stopwords were generated for both reviews and the summaries. For the words from the reviews, the frequency distribution of the most common words was identified and they were plotted in a bar chart before being analysed for their polarity.

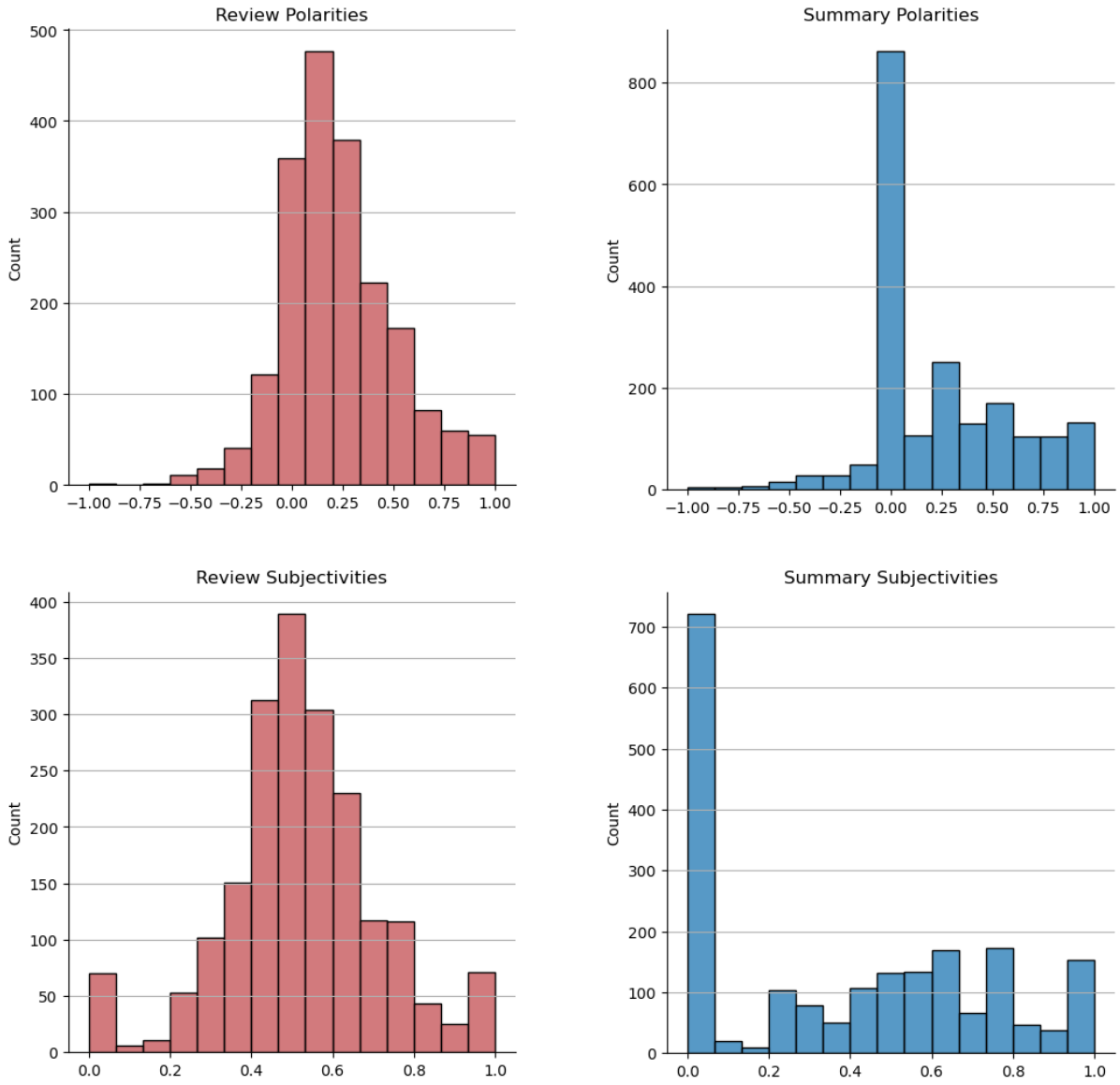
Reviews WordCloud (50 words)





Given the lack of context when analysing individual words, the polarity of the most frequent words was of little assistance.

Reviews and summaries were analysed for their polarity and subjectivity. Both were predominately positive, however, the reviews tended to be more subjective than the summaries.



The most positive/negative reviews and summaries were gathered. The sentiment analysis used had been good at identifying when reviews were positive due to words such as “Amazing”, some more nuanced comments were designated as negative when the overall sentiment was positive but some words appeared negative. One comment was “Boring unless you are a craft person which I am ... “, where clearly the user had enjoyed the product but the analysis had flagged it as negative due to “Boring”.

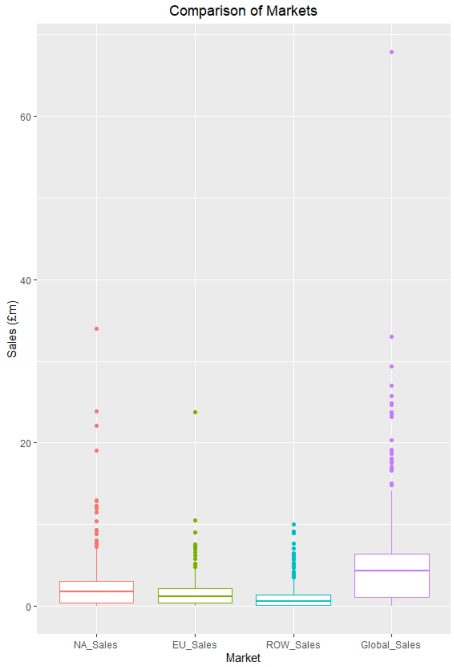
Sales Data

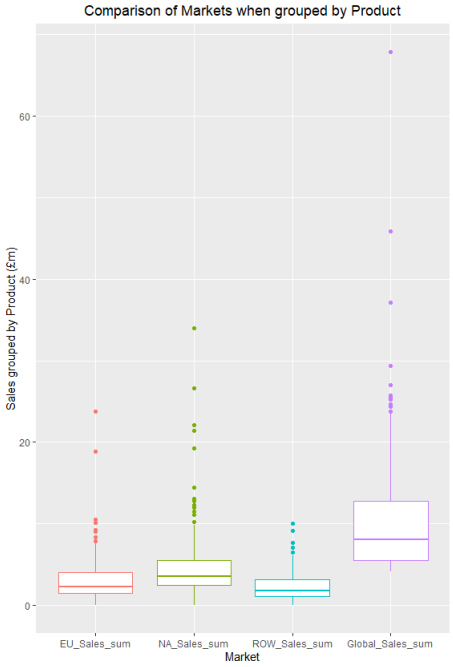
The sales data was imported and cleaned in RStudio and validated in a similar fashion to the review data.

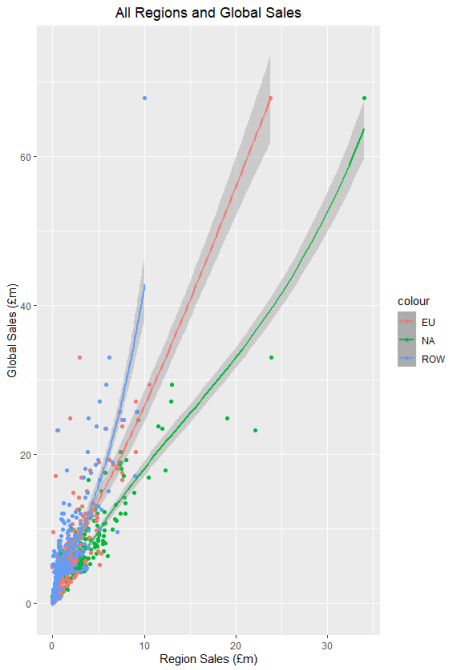
The impact of sales by product

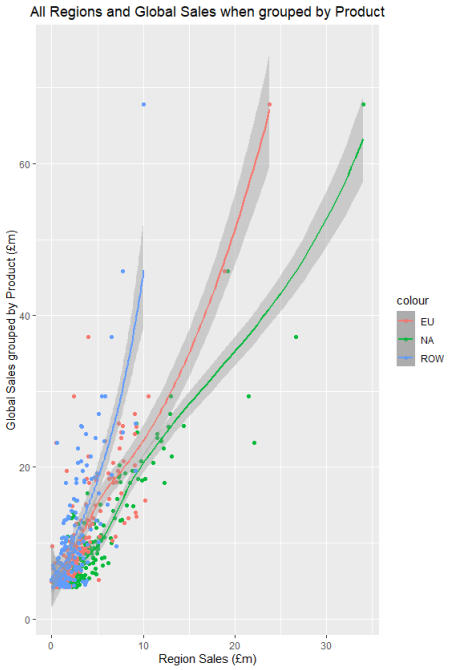
Scatterplots, histograms, boxplots, and violin plots were used to explore the data and compare the NA, EU and Global sales. Before and after being grouped by Product, the NA, EU and Global sales data demonstrated a significant level of skewness with a large number of outliers at the upper end of the sales data. These outliers were considered relevant to the analysis as they represent the best selling games of all time.

Through grouping by Product it reduced the number of outliers and the total skewness of the data.



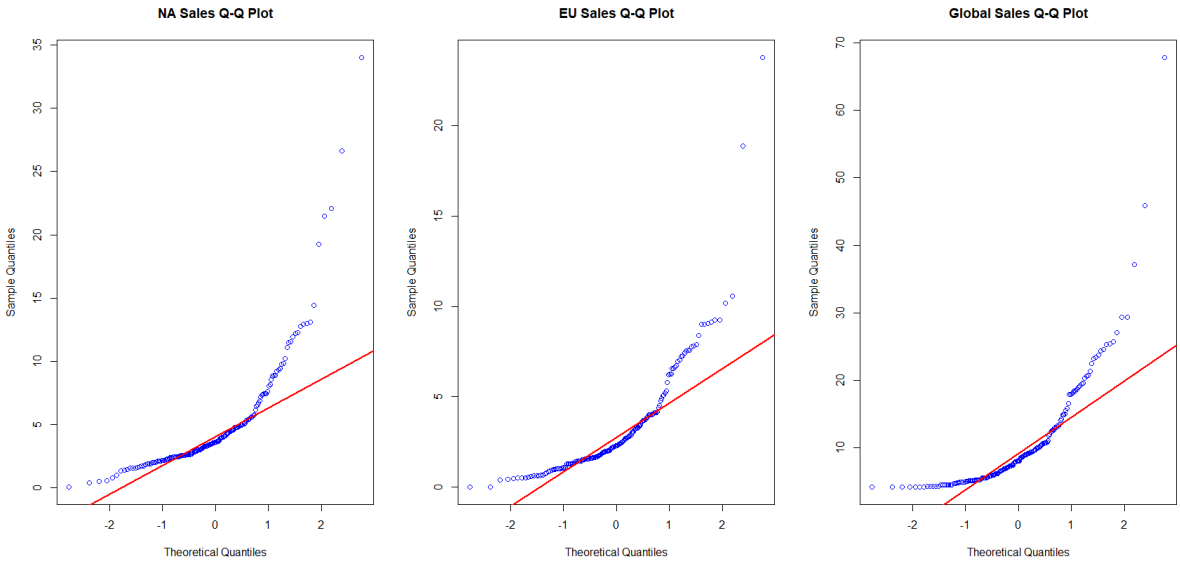






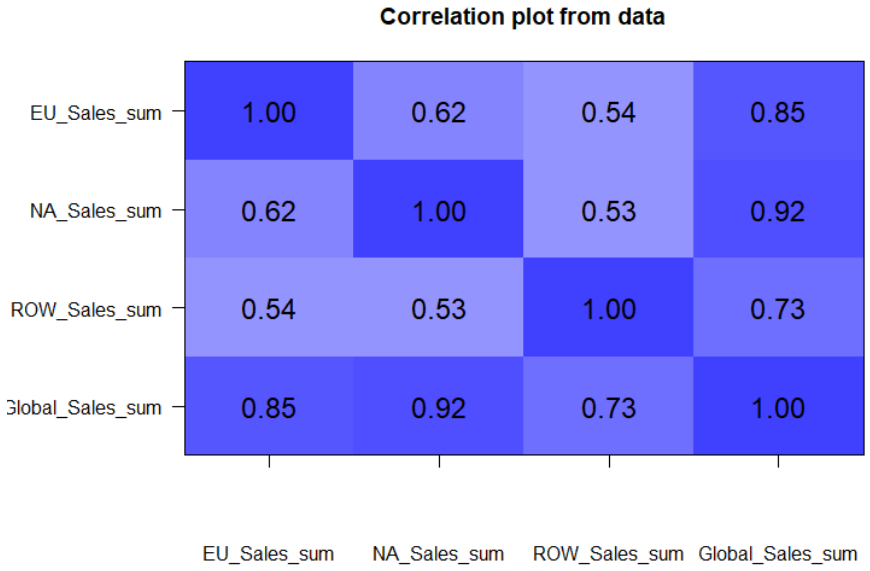
The reliability of the data

Q-Q plots were generated for the NA, EU and Global sales. They all showed the data was not following a normal distribution, and suggested that they follow exponential distributions. However, these possible distributions were not investigated further.



Shapiro-wilk, skewness and kurtosis tests all supported the finding that the sales figures were not normally distributed.

The correlation between the different regional sales was also looked at:

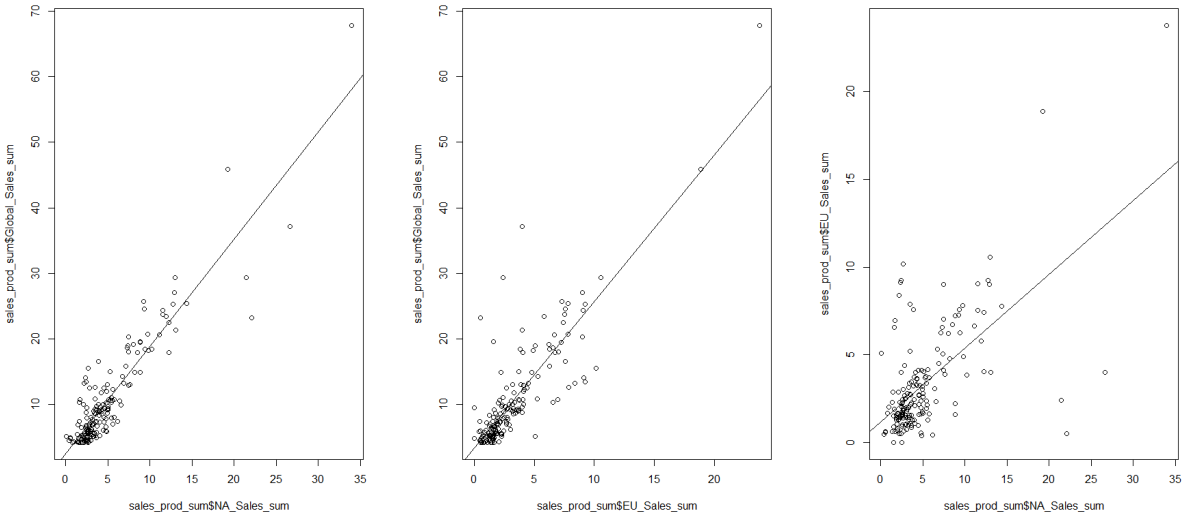


This showed NA sales had the highest correlation with Global sales

Identifying relationships between sales in various regions around the world.

To identify relationships between sales in different regions, the following simple linear regressions were investigated:

Independent Variable	Dependent Variable	Adjusted R-squared	P-value (0.05 level of sig)
NA sales	Global sales	0.8385	< 2.2e-16
EU sales	Global sales	0.7185	< 2.2e-16
NA sales	EU sales	0.382	< 2.2e-16



This showed that NA and EU sales could individually be used to predict Global sales to some degree.

A multiple linear regression model using NA and EU sales to predict Global sales was made. The adjusted R-squared value of this was 0.9664 with both p-values being < 2e-16, indicating the model could predict Global sales with a high degree of accuracy. NA and EU sales did not exhibit multicollinearity.

Sample NA and EU sales were used to predict Global sales:

Input NA Sales (£m)	Input EU Sales (£m)	Predicted Global Sales (£m)	Predicted Global Sales lower-bound (£m)	Predicted Global Sales upper-bound (£m)	Predicted Global Sales range (£m)
34.02	23.8	68.06	66.43	69.68	3.25
3.93	1.56	7.36	7.1	7.61	0.51
2.73	0.65	4.91	4.61	5.2	0.59
2.26	0.97	4.76	4.48	5.04	0.56

22.08	0.52	26.63	25.37	27.88	2.51

However, it should be noted that NA and EU sales directly contribute the majority to Global sales so it is not surprising that they can be used to predict their approximate sum. This predictive ability is of very limited use.

Recommendations

1. Be cautious using the sentiment analysis results of the reviews/summaries given the possibilities of reviews being considered overly negative without appreciating nuance.
2. Adding a quantitative value to reviews will help differentiate between negative and positive reviews/summaries.
3. Use predictive options cautiously:
    a. Loyalty Points based on Spending Score and Remuneration is prone to large errors
    b. Global sales based on NA and EU sales is little more than a sum of the two.
4. Future analysis could consider genres within sales data.