Assignment: Predicting Future Outcomes

Omer Khan

Word count: 1008

Overview:

Turtle Games, a game manufacturer with a global customer base, has a business objective of improving overall sales by analysing customer trends to extract insights. The existing data collected from sales and customer reviews has been analysed to gain a better understanding to the following questions posed by Turtle Games:

1. How customers accumulate loyalty points
2. How groups within the customer base can be used to target specific market segments
3. How social data (e.g. customer reviews) can be used to inform marketing campaigns
4. The impact that each product has on sales
5. How reliable the data is (e.g. normal distribution, skewness, or kurtosis)?
6. What the relationship(s) is/are (if any) between North American, European and global sales?

Approach

Python and R were utilised to clean, explore, manipulate and then visualise the data in order to answer the questions posed.

The initial sense check for both Python and R involved the following steps:

- Viewing the data frame
- Checking dimensions of the data frame
- Checking and removing any errors
- Viewing descriptive statistics to the see the spread, max values, min values and mean of the data set
- Checking missing values and removing or replacing based on the analysis required
- Removing irrelevant variables

In order the answer questions 1, 2 and 3 above, the data containing customer reviews was manipulated using Python. Linear regression with OLS, K-Means Clustering, NLP Sentiment Analysis and NLP Subjectivity and Polarity Analysis were conducted on this data set.

Questions 4, 5 and 6 were explored in R by creating various plots, conducting Shapiro-Wilks test, QQ plots and simple and multilinear regression.

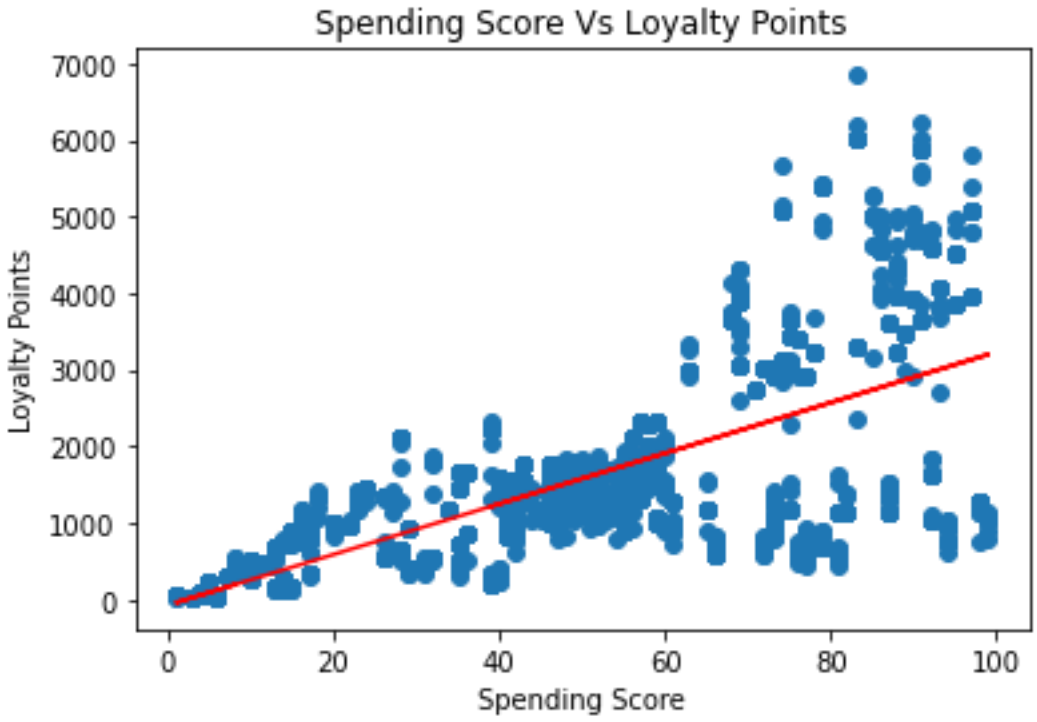
Insights

Reviews Data:

After performing sense check, cleaning the data and removing unnecessary columns Linear regression was performed. The following R-Squared values were observed:

Independent Variable (x)	Dependent variable (y)	R-Squared Value
Spending	Loyalty	0.45
Remuneration	Loyalty	0.38
Age	Loyalty	0.002

The above R-squared values suggest that there is a slight positive correlation and linear relationship between Spending vs Loyalty and Remuneration vs Loyalty. On the other hand, no correlation between age and loyalty points can be suggested based on the almost 0 R-squared value. This might, however, require to be further investigated.



The above scatter plot and data being around the regression line suggest the possible linear relationship between spending and loyalty points.

Following this, K-Means Clustering was conducted. Elbow and silhouette methods were used to determine the ideal number of clusters. Since the max value for silhouette was 5 it

was concluded that 5 is the optimal number. The following scatterplot was obtained:

K-Means Clusters Scatter plot, k=5

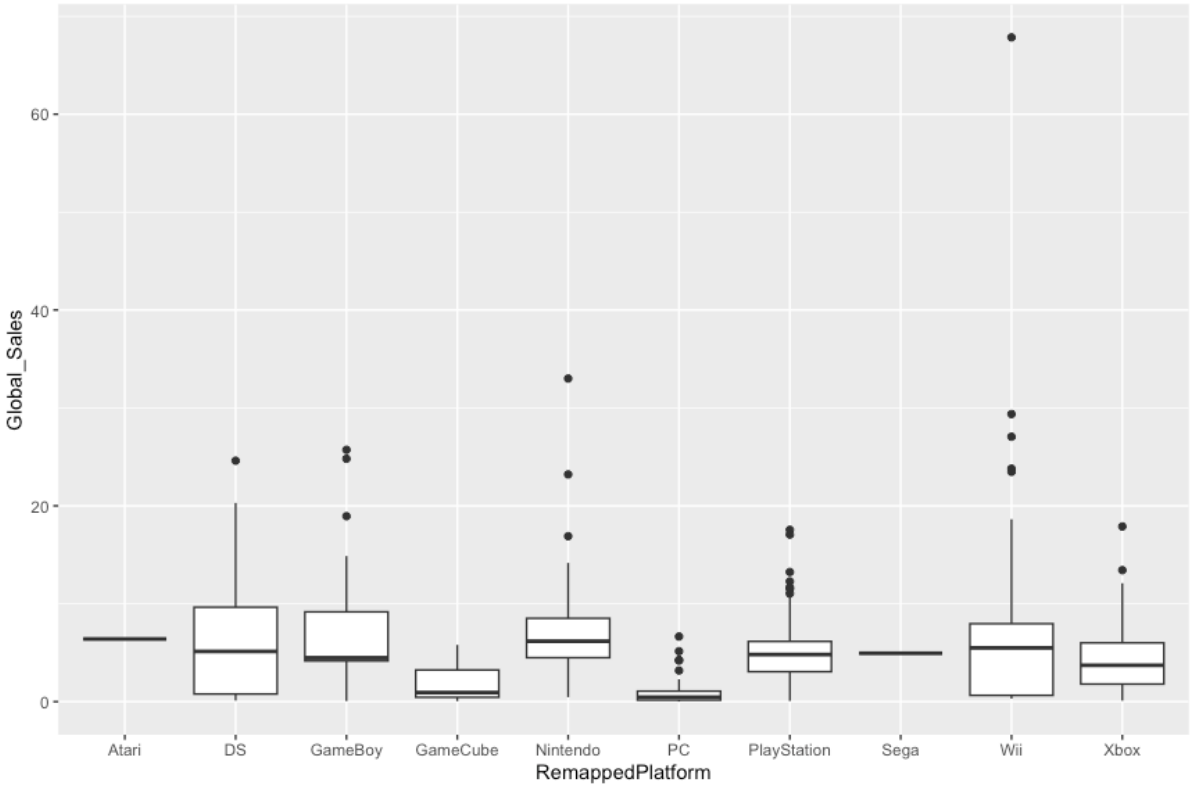
From the above scatter plot it can be inferred that there is a strong relationship between remuneration and spending score.

Following this NLP sentiment analysis was conducted. The Review and Summary columns from the data set were converted to strings which were tokenised and stop words were removed in order to product word clouds for both review and summary. Both these word clouds reflected generally positive sentiments. In order to minimise any bias, subjectivity scores were also measured. 4 Histograms were plotted which have been included in Appendix A. From these histograms a strong relationship between sentiment score and subjectivity score can clearly be observed suggesting that customers reviews were less objective. It is also evident that most reviews written are positive.

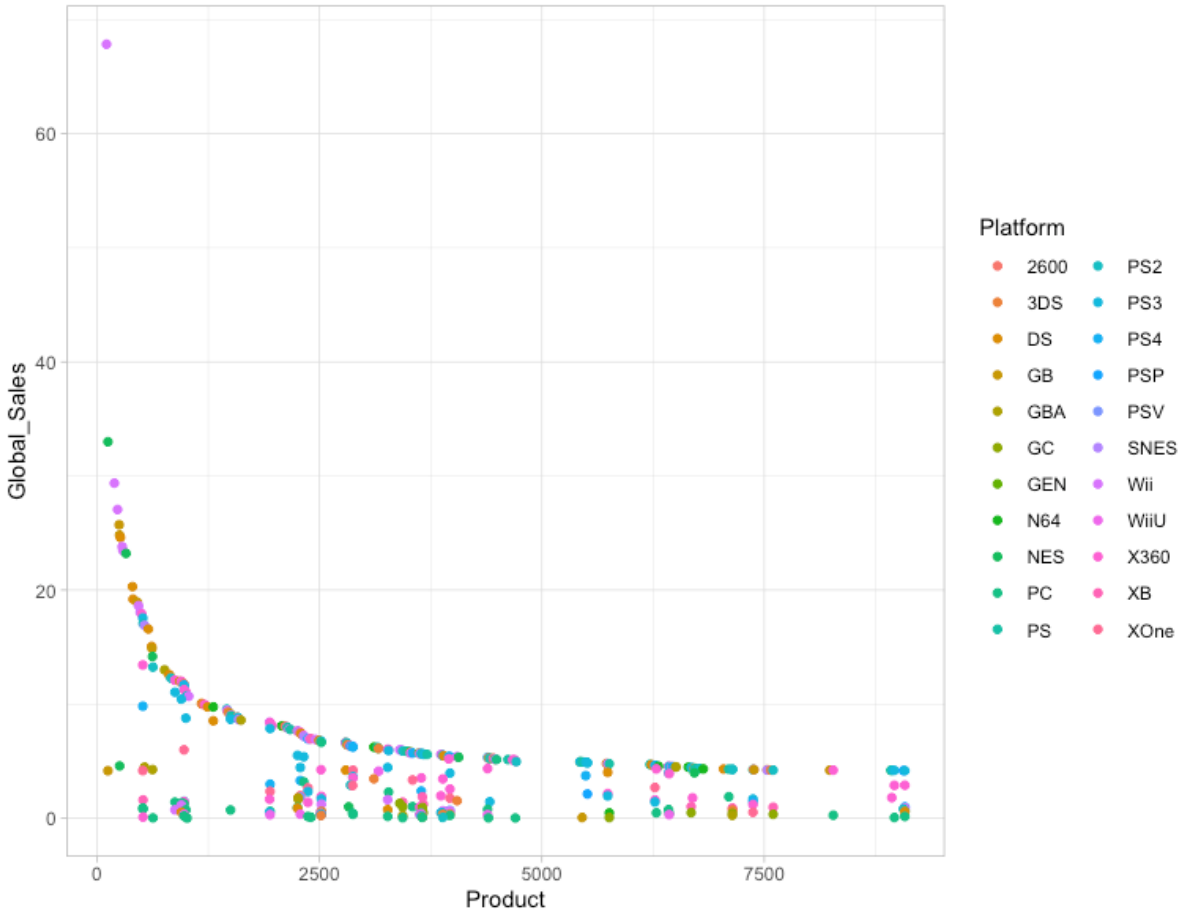
Sales Data:

After the initial sense check of the data, DataExplorer was generated using R, in order to determine reliability, testing for missing values and the descriptive statistics of this data set. The following box-plot identifying outliers was plotted:

Box and whisker plot of gaming platforms



The following graph was also plotted showing sales and product ID:



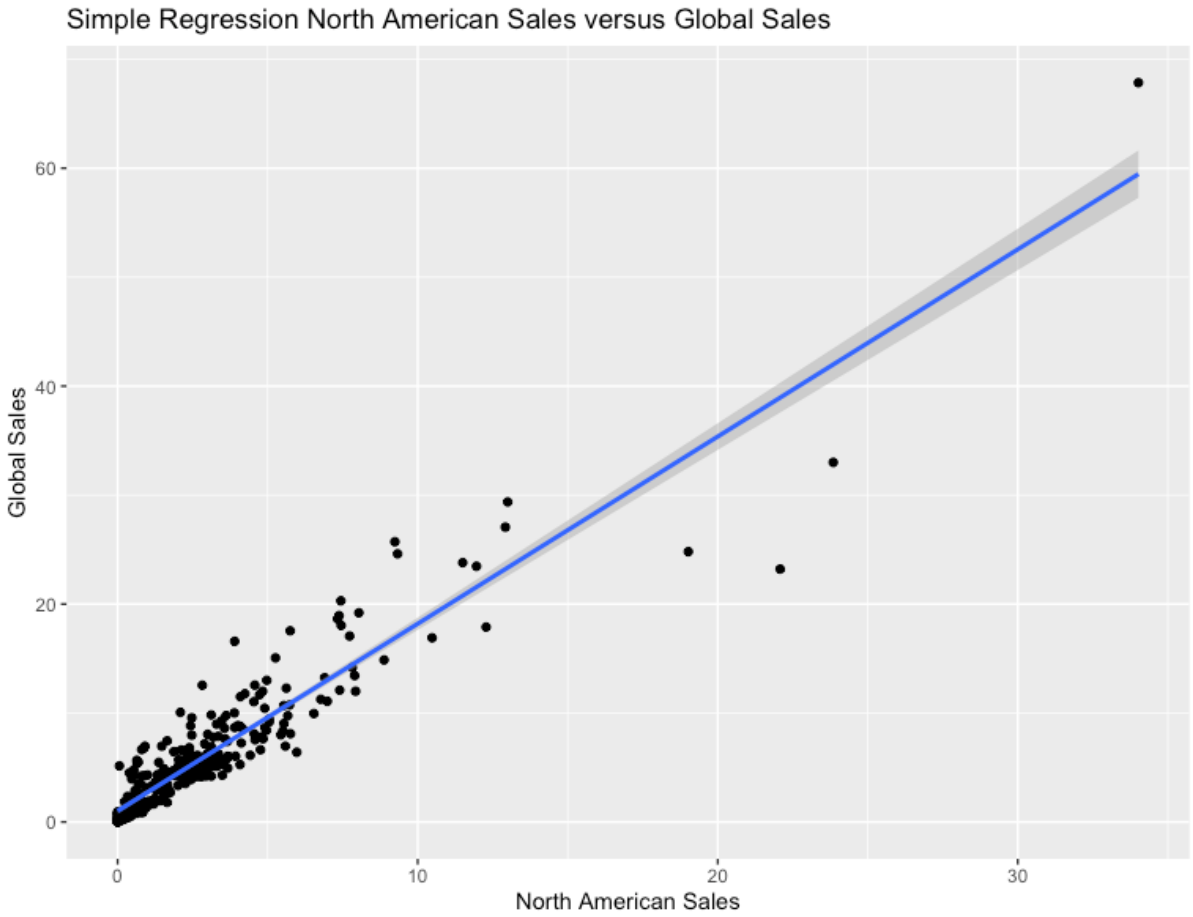
The above graph shows an obvious negative correlation between Product ID and sales. As product ID increases sales decreases. The likely explanation of this being that Turtle Games has given product ID based on number of sales.

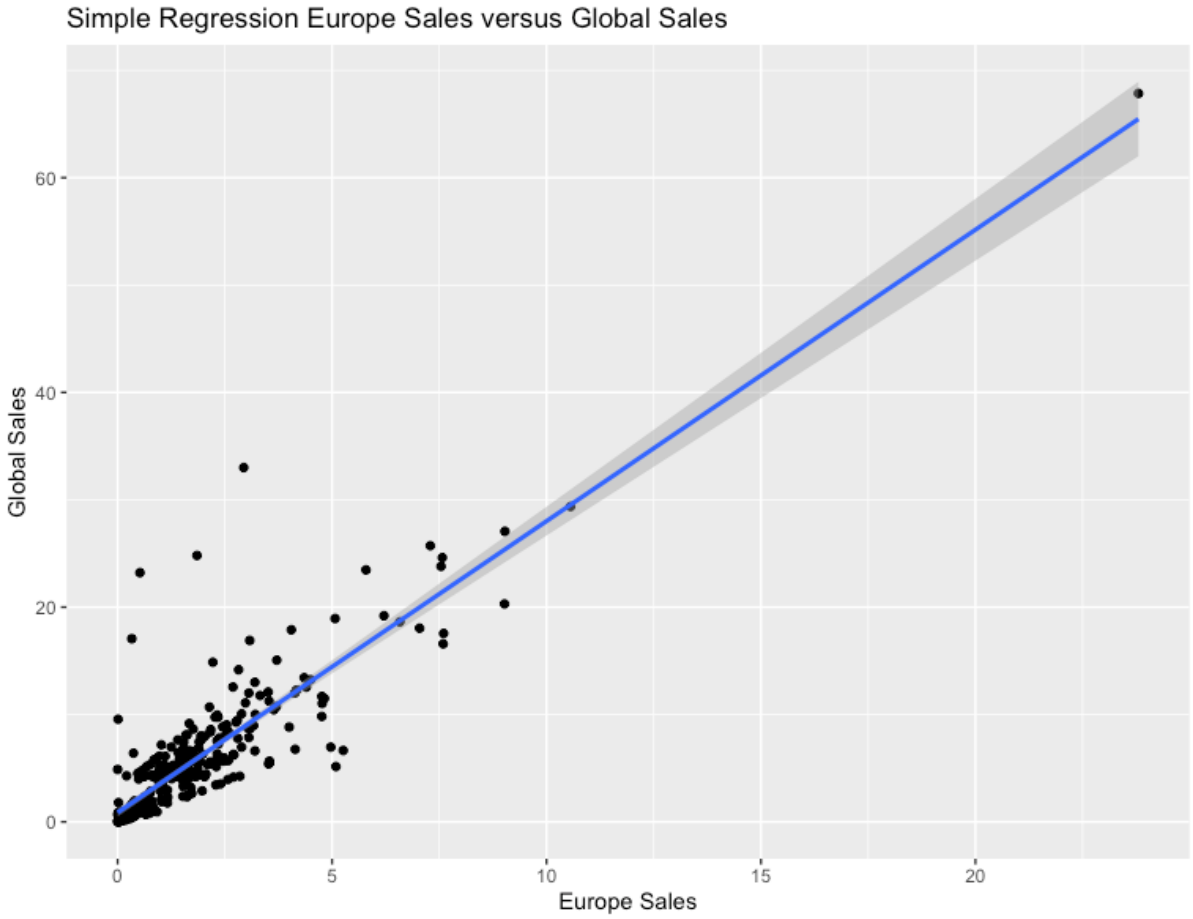
To further determine the reliability of the data set, due to the spread and outliers identified in the boxplot, QQ plots were plotted to determine if the data is normally distributed. Shapiro Wilk, kurtosis and skewness tests were also performed and it was concluded that the data is very positively skewed and likely not normally distributed.

In order to determine the relationship between global and regional sales the following correlation plot was created:

The above plots clearly highlights the strong correlation between regional and global sales with the correlation between NA and Global Sales being stronger.

Finally, simple linear regression model and a multilinear regression was conducted to plot future predicted values. The follower linear regression models were obtained:


Simple Regression North American Sales versus Global Sales
Global Sales
North American Sales
0
20
40
60
0
10
20
30


Simple Regression Europe Sales versus Global Sales
Global Sales
Europe Sales
0
20
40
60
0
5
10
15
20

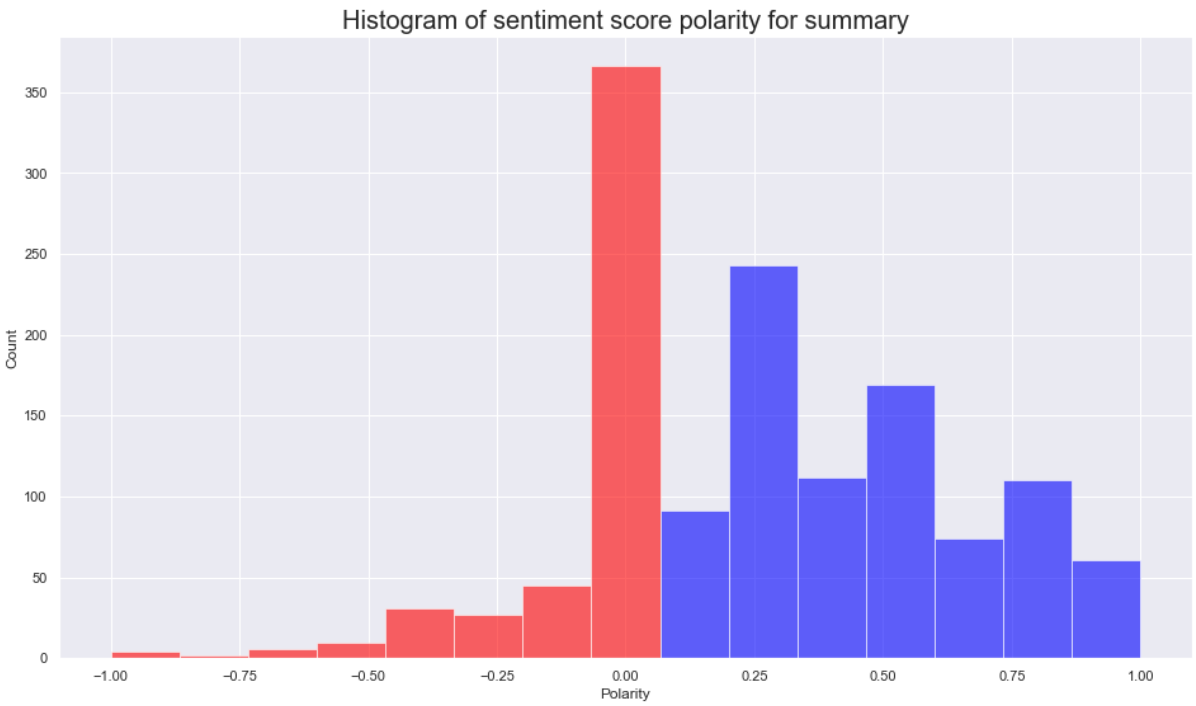
The above plots clearly highlight the strong correlation suggesting that global sales can be predicted using regional sales for both NA and EU.

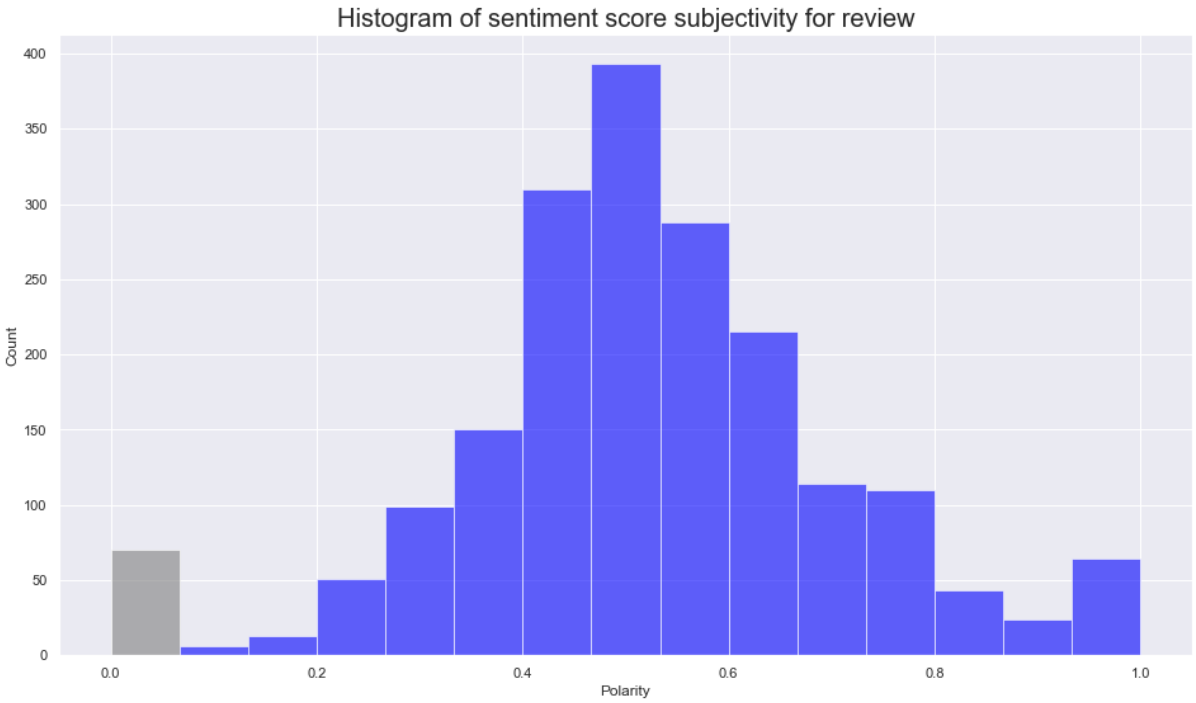
Patterns and Predictions:

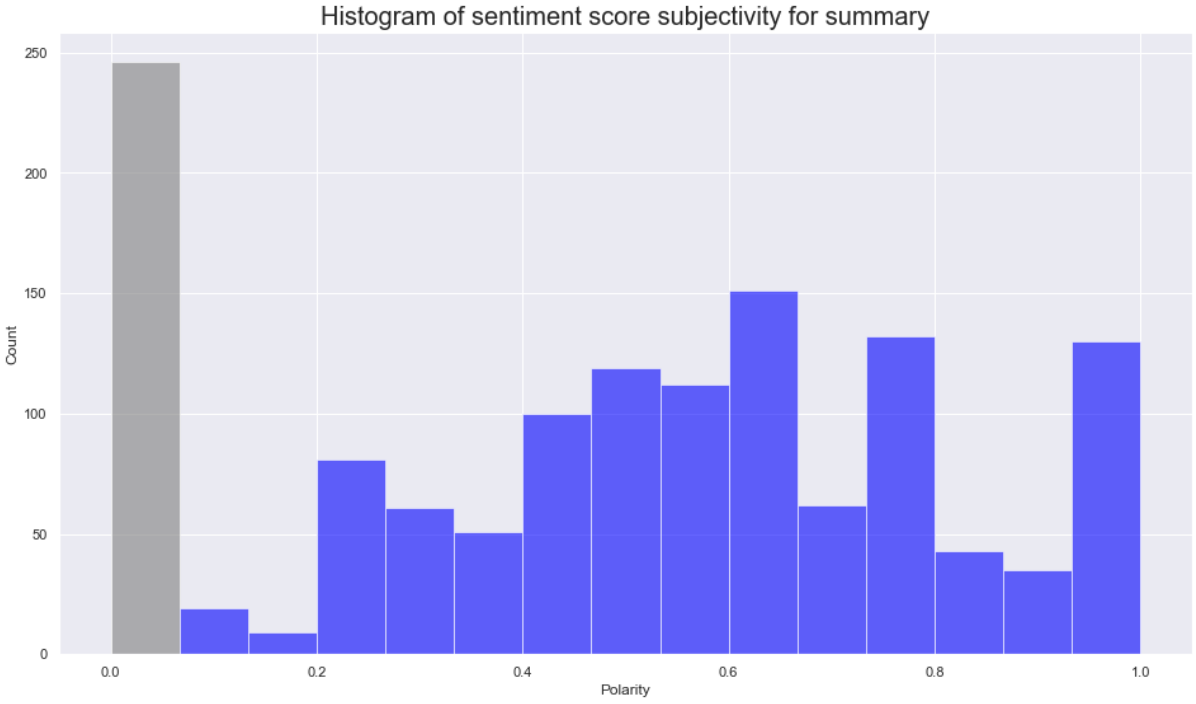
These insights can help the business in the following ways:

- targeting the clusters with the highest spending score should be targeted by marketing and sales.
- NLP Sentiment identifies that majority of the products offered are liked by customers.
- Negative reviews should be investigated further to identify any concerns
- The strong relationship between regional and global sales can be used to predict future product performance

Appendix A: Polarity and Subjectivity Histograms




Histogram of sentiment score subjectivity for review
Count
Polarity
0
50
100
150
200
250
300
350
400
0.0
0.2
0.4
0.6
0.8
1.0


Histogram of sentiment score subjectivity for summary
Count
Polarity
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0