Jake Evans

Background

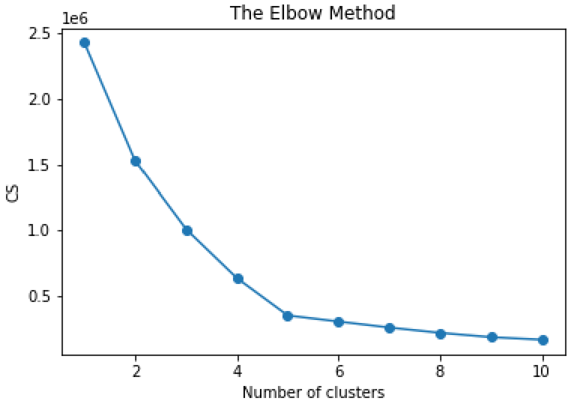
Turtle Games aims to improve their sales performance. This investigation tries to understand how customers acquire loyalty points, review the products on social media, and can be segmented to inform future marketing campaigns. Sales data will be analysed to determine relationships between regions, product impacts on sales and test the useability of the data in predicting future outcomes.

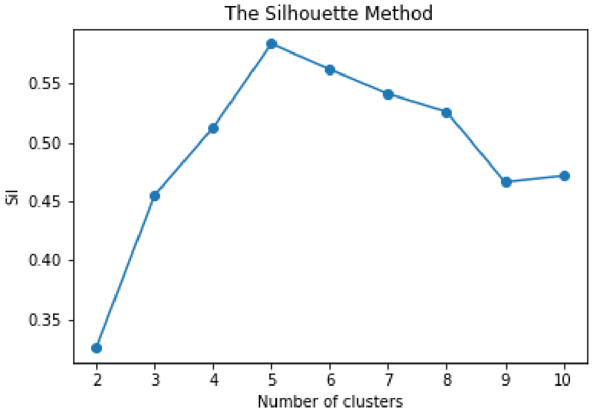
Analytical approach

The Customer Review dataset contains quantitative and qualitative data, with 11 columns, 2,000 rows and no missing values. The columns 'language' and 'platform' were removed, then 'renumeration' and 'spending_score' were renamed.

Linear regression is used to predict how loyalty points are accumulated with the OLS function from *Statsmodels*. Spending scores, income and age were tested as predicters, producing regression tables. Income and spending scores predict loyalty points similarly, having x-coefficients of 33.1 and 34.2 then r-squared results explaining 45% and 38% of the variation. These are statically significant (f-statistic < 0.05), meaning increasing spending score by 1 or income by £1,000 will likely increase loyalty points by 1. Age does not predict loyalty points with a r-squared value (0.002) that is almost 0 and not statically significant.

K-means clustering was used to determine the optimal number of segments based on spending and income with the KMeans and Silhouette_Score packages. The Elbow and Silhouette methods indicate an inflexion point at 5. Pairplots with 4,5 and 6 clusters confirm 5 is the optimal number of segments.





Customer reviews and summaries were converted to lower case with punctuation, duplicates, stop and alphanumeric words removed, the word tokenize function then separated them into words. Polarity scores help understand public sentiment of the reviews. They had an average of 0.2, indicating a positive sentiment. Lists were produced of the most positive and negative 20 strings.

After importing sales data into R, columns other than product and sales were removed. Grouping the data by product increased the minimum and mean global sales from £0.01M to £4.2M and £5.3M to £10.7M respectively.

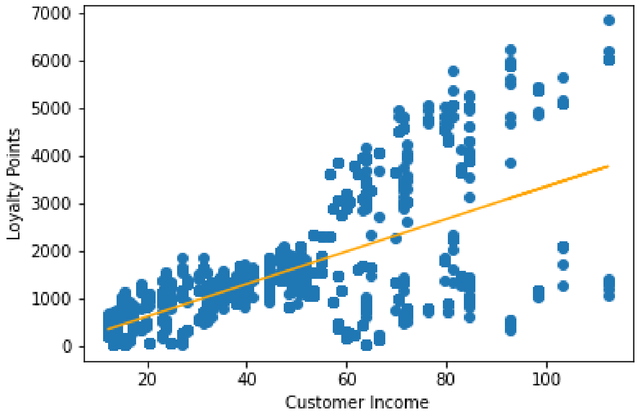
Shapiro test on sales were used to gauge reliability where they produced smaller p-values than 0.05. Kurtosis values for all regions are above 15 (>3) confirming leptokurtic distribution, meaning there are many outliers. Q-Q plots show that points begin to deviate from the diagonal line towards the tails for each region. These 3 tests all evidence the data is not normally distributed.

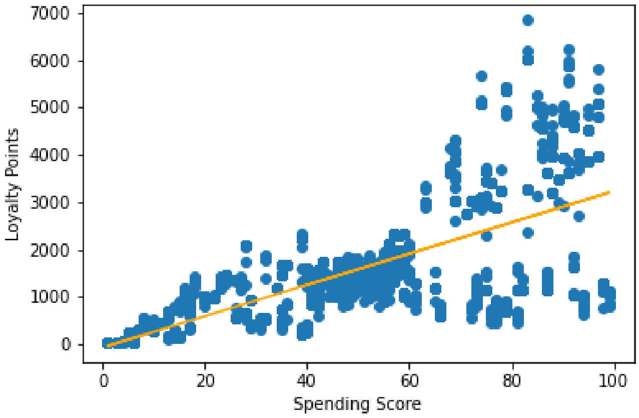
Skewness is above 2.8 for all regions, indicating the data is highly positively skewed. Global sales have a strong correlation (0.92 and 0.85) with NA and EU sales respectively. NA and EU sales have a positive correlation (0.62) indicating that success carries across regions.

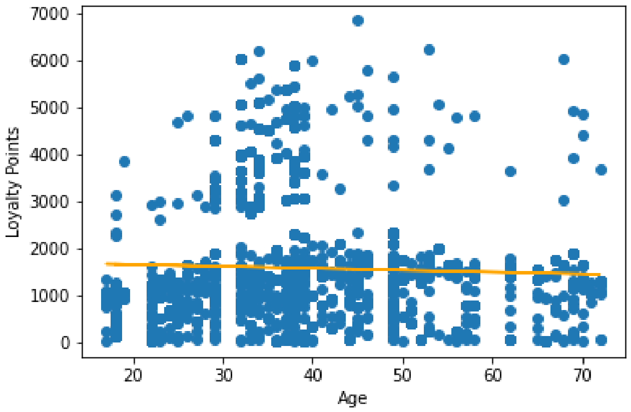
A MLR model was created to predict global sales. The adjusted R-squared value is larger, thus stronger, when product id is included. Testing the model with 5 examples from the data frame, 2 predicted values fell within the confidence interval, whereas only 1 was within when product id was excluded.

Visualisation and insights

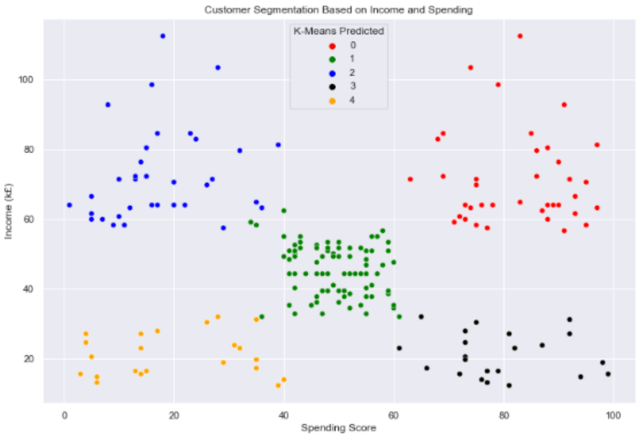
Scatterplots with a regression line best illustrate the relationship between loyalty points and the other variables. A positive correlation can be clearly observed with spend and income, supporting the regression table. No relationship with age is confirmed by the almost horizontal regression line.



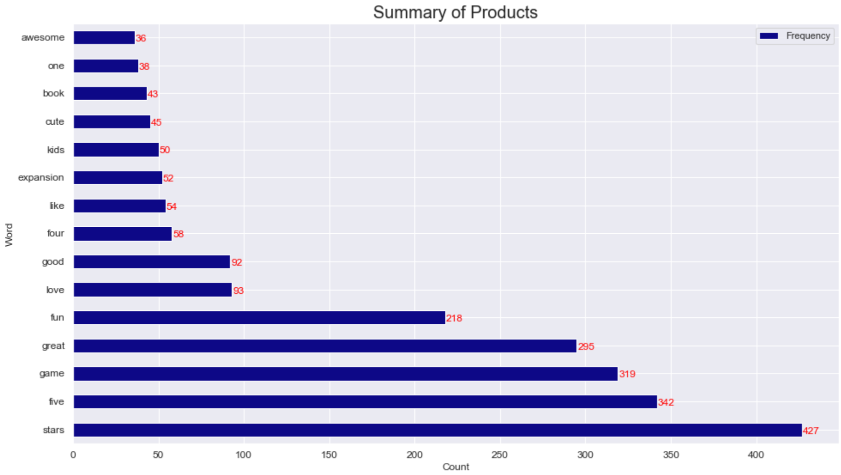


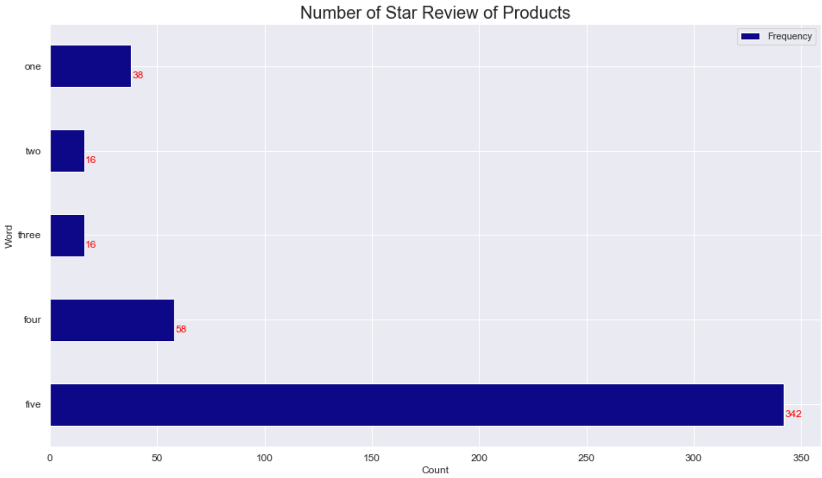


A colour coded scatter plot identifies segments based on income from K-means clustering. The largest segment by far is the green cluster with average income and spending, with the next two largest clusters being those with higher incomes, blue and red.

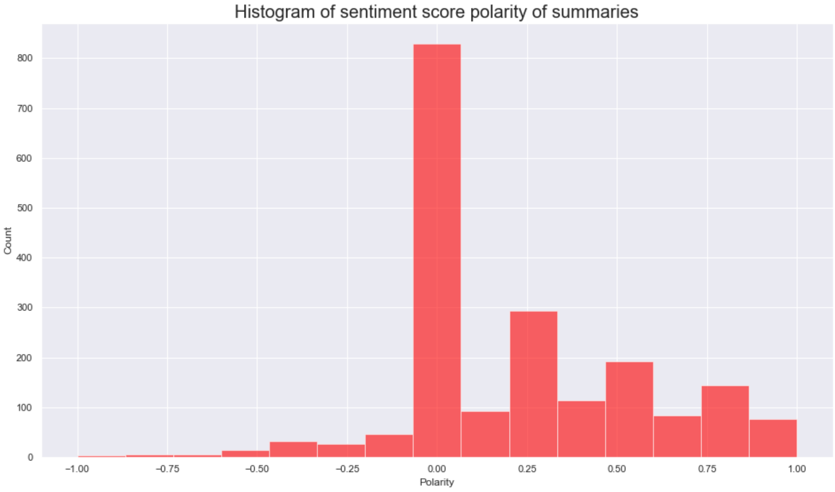


Bar charts best illustrate the frequency distribution of words from reviews, with the top 15 selected allowing for easier comparison. The most common words tend to be positive such as ‘great’, ‘love’ and ‘fun’. Summaries also contained some star ratings, filtering and plotting a frequency distribution of those, 73% gave ‘five stars ’and had an average of 4.4 stars.

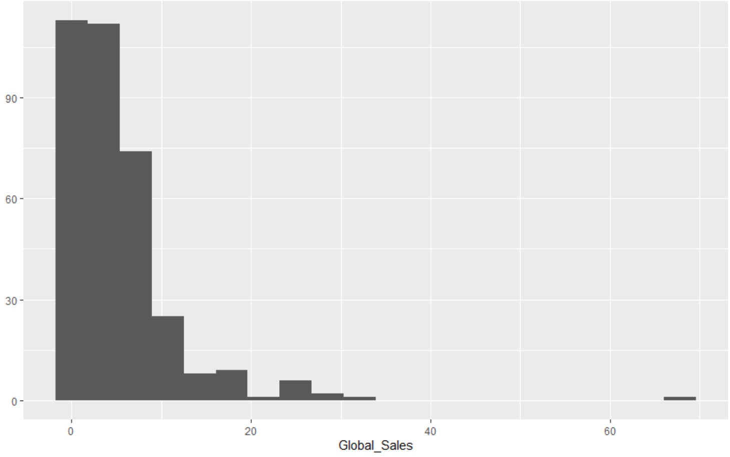


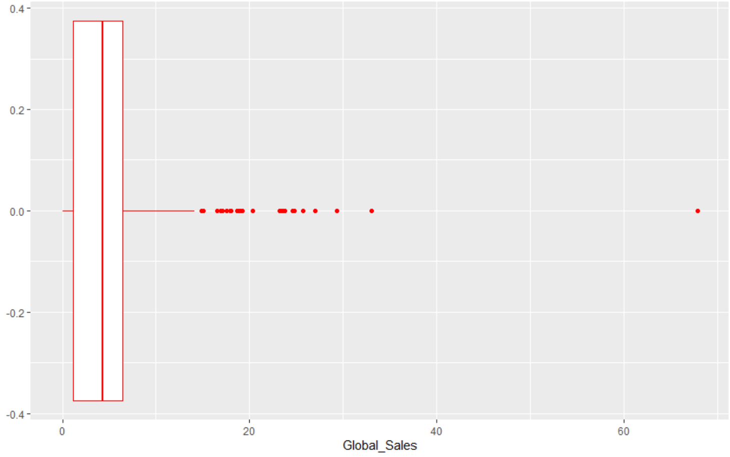


Histograms best highlight the distribution of polarity scores. They indicate a normal distribution which is expected as 74% of review summaries are neutral (between -0.05 and 0.05). However, there are overwhelmingly more positive review summaries (25%) than negative (1%).



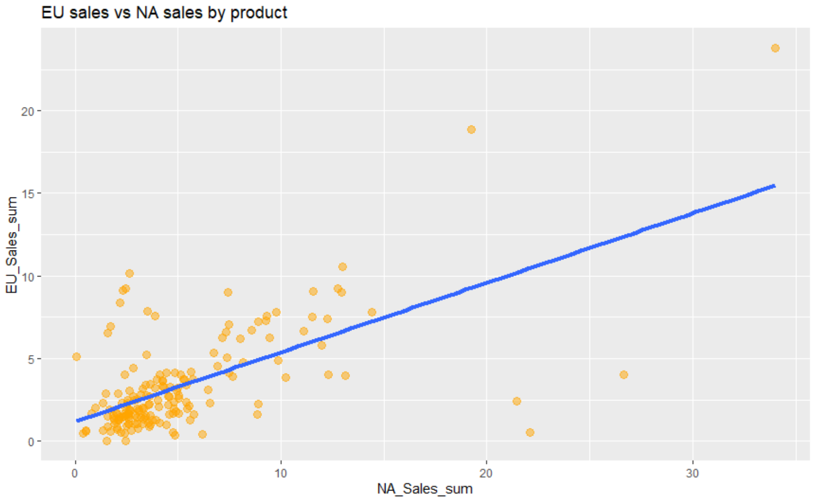
To support the sales descriptive statistics, various plots were produced. Histograms shows there is a positive skew, with there being larger quantities of products with fewer sales than those that produced larger volumes. Boxplots further this, showing there are several outliers above the upper limit for all regions.

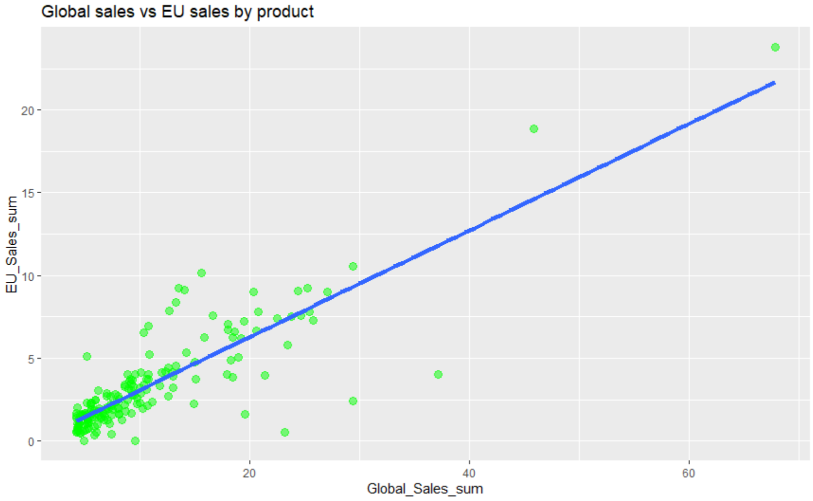


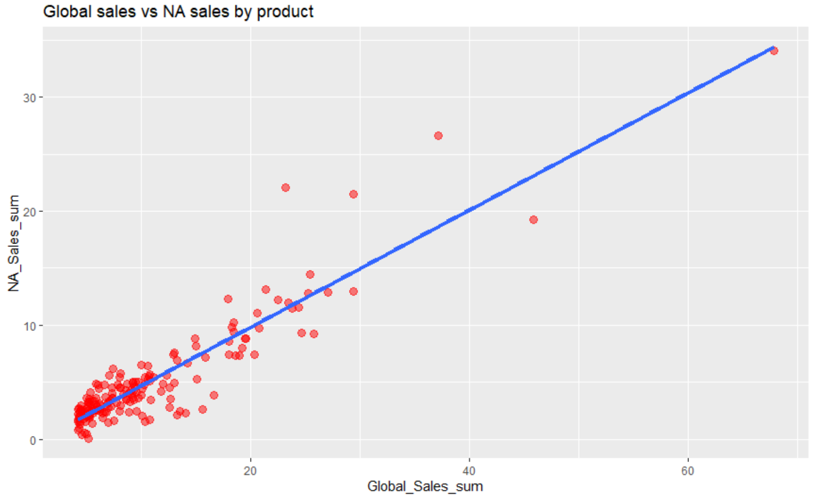


Jake Evans

Scatter plots show the positive correlation between NA and EU sales, indicating success translates from region to region with a few exceptions. Global sales correlate positively stronger with both NA and EU expectantly due to including both values. These are most useful in helping understand relationships between the different sales regions, thus a line of best fit is added to help Turtle Games visualise the relationship, with different colours being added to differentiate the plots.


EU sales vs NA sales by product


Global sales vs EU sales by product


Global sales vs NA sales by product

Patterns and predictions

Promoting loyalty points can improve sales, thus segmenting by income and spending scores can be utilised and tailored towards within marketing. A loyalty points campaign should be prioritised on the high-income, high-spending cluster, as they are most likely to acquire loyalty points. The high-income, low-spending cluster could be the most efficient segment to market to as spending behaviour can change with an advertisement unlike income.

The customer reviews are mostly positive, from the analysis there are frequently used positive words or perfect polarity scored reviews that could be included as customer testimonials. The statistic, 85% of star-based summaries being 4+ could be utilised. Overall, there are lots of positive materials to support trust in the quality and credibility of the products.

The sales database is not too reliable as high performing outliers distorts the data. However, the correlations between NA and EU games could be useful in future decision making. A MLR model to predict global sales is useable should be improved upon by collecting additional data to test and refine the model.

The investigation could be furthered by diving deeper into game genre and to determine any insight into what type of games should be prioritised by Turtle Games.