Turtle Games

Introduction

Turtle Games, a game manufacturer and retailer, would like to improve its overall sales performance by utilising customer trends. To improve sales performance, Turtle Games wants to understand:

- How customers accumulate loyalty points.
- How groups within the customer base can be used to target specific market segments.
- How social data can be used to inform marketing campaigns.
- The impact that each product has on sales.
- How reliable the data is.
- What the relationship(s) is/are (if any) between North American, European and Global Sales?

Analytical Approach

Matplotlib, pandas, seaborn and statsmodels libraries were imported into Python using aliases to keep code concise. Matplotlib and seaborn libraries enabled data visualisation. Numpy and pandas libraries assisted with data analysis. The statsmodels library has various statistical analyses and modelling techniques such as linear regression. The reviews csv was loaded into python using the pd.read_csv() pandas function and passed through to a variable. The resulting dataframe was sense-checked to ensure they've been correctly imported. The number of rows, column names, data types, missing values, metadata and descriptive statistics were checked in addition to the first and last five rows for any headers or footers.

Redundant columns were removed from the dataframe using the .drop() function and passed through to a new dataframe. The resulting dataframe had its columns renamed using the df.rename(columns=) function. This dataframe was exported and saved using the df_new.to_csv function. The dataframe was then imported and used with the statsmodel library to determine if there were any possible linear relationships between various variables.

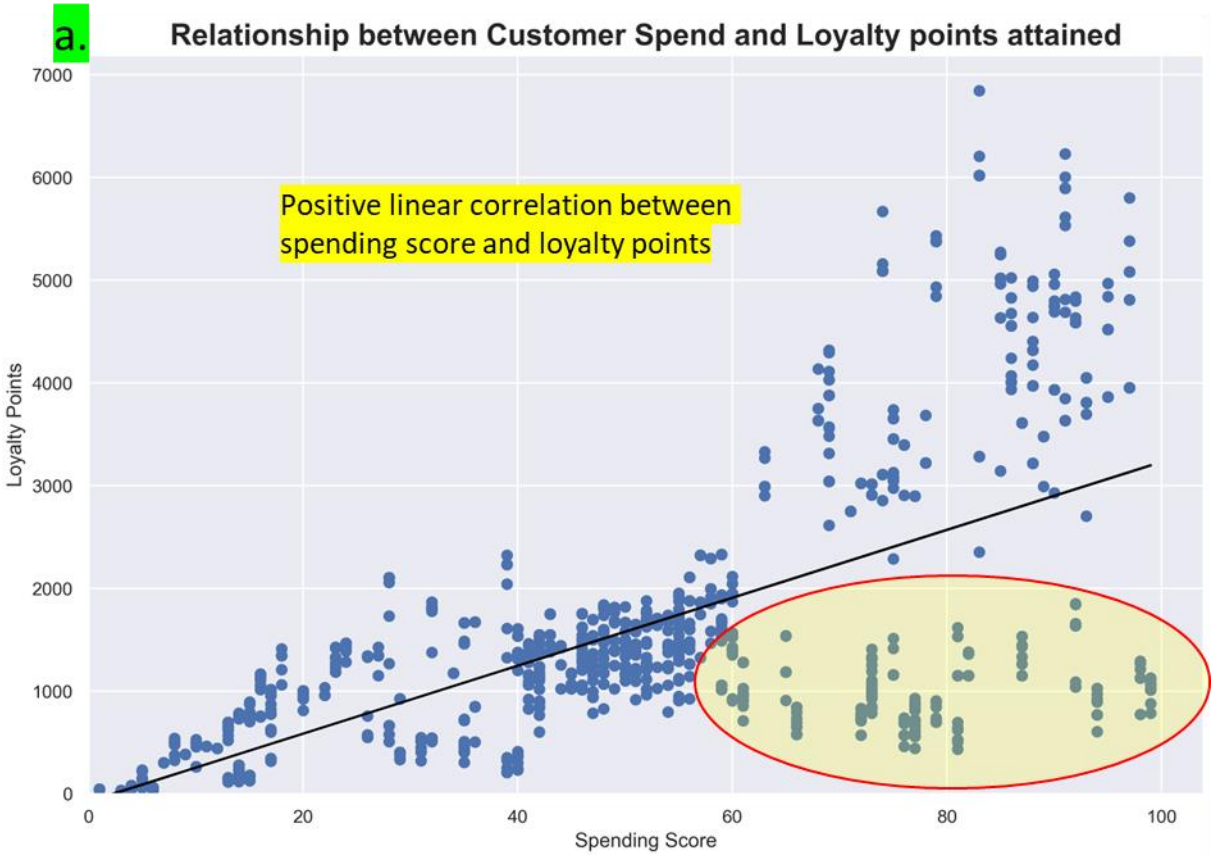
Sklearn library was imported to conduct k-means clustering. The cleaned dataframe was imported and sense-checked using aforementioned procedure. A scatterplot was created from two columns (renumeration and spending score) using seaborn. The silhouette and elbow methods were used to verify the number of clusters present in the scatterplot. Different values for k(number of clusters) were evaluated for suitability.

Nltk library was imported for tokenisation, TextBlob was imported for sentiment analysis. Wordcloud was imported to visualise the most common words. The review and summary columns were prepared for natural language processing. The text was changed to lower case and punctuation removed. Any duplicates that appeared in the columns were removed using the df.drop_duplicates() function. The columns underwent tokenisation and wordclouds were generated. Alphanumeric characters and stopwords were removed and new wordclouds were made. The FreqDist function was used to count the most common words in both columns. Sentiment analysis was undertaken on the words in these columns using TextBlob through a user-defined function.

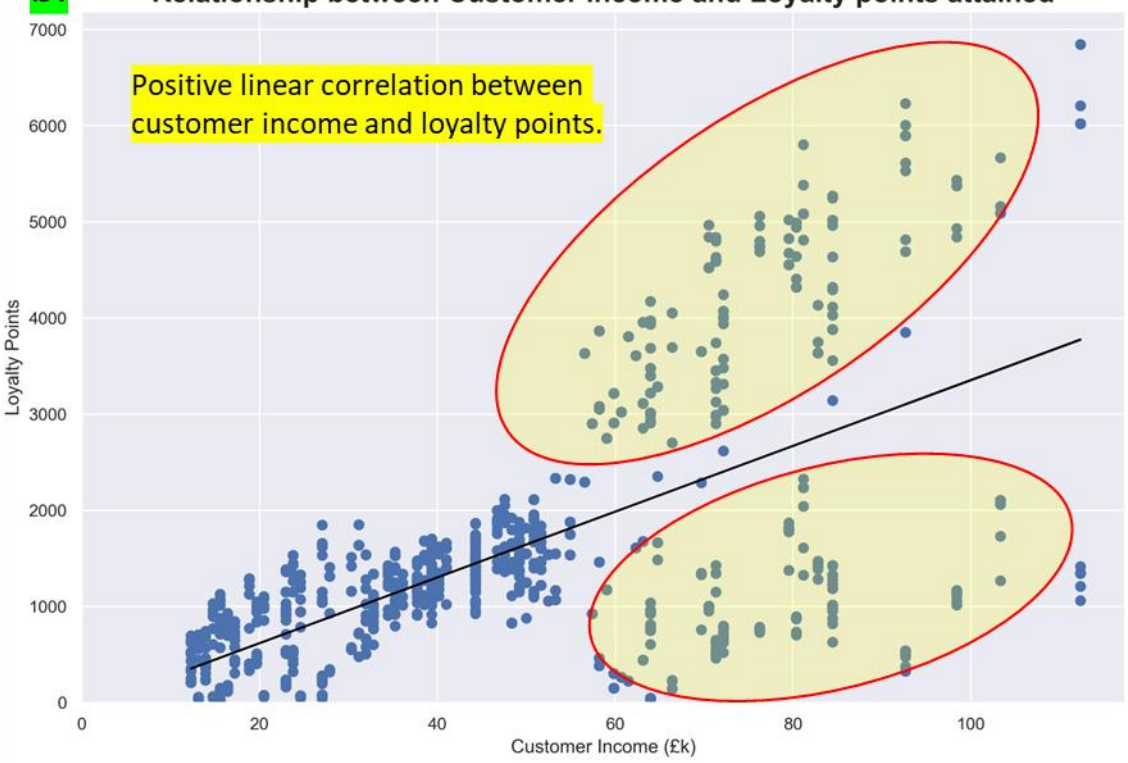
R was used to analyse sales data. The tidyverse, plotly, moments and psych packages were imported. The tidyverse package has a host of data analysis functions. Plotly makes interactive plots, moments was used for statistical analysis. Psych provides a correlation matrix between different variables. The csv file was imported into R and sense checked in a similar way to Python as described earlier. The

qplot() function was used to generate scatterplots, histograms and boxplots for exploratory analysis(EDA). The ggplot() function was later used to enhance the EDA visualisations. The moments library was utilised to conduct a Shapiro-Wilk test to determine normality of sales data. Skewness, kurtosis and correlation were determined on the sales columns. A multiple linear regression model was formed using the North American and European sales as regressors and the Global sales as the explained variable.

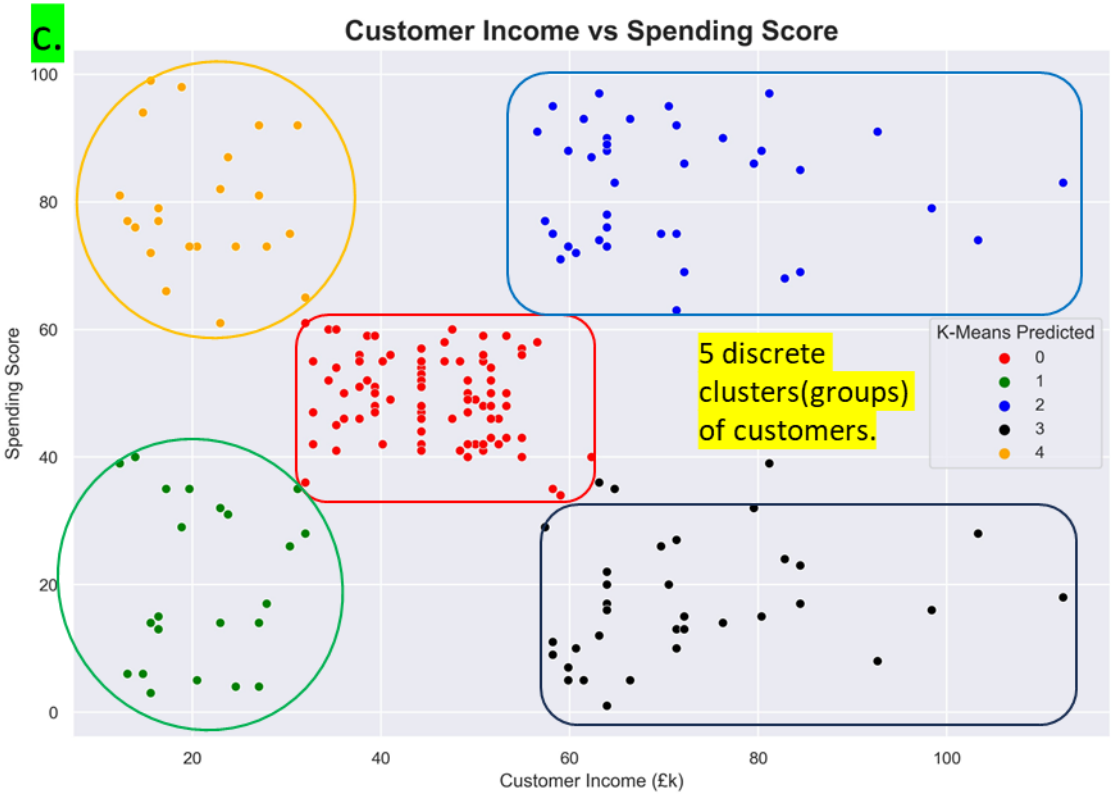
Visualisations and Insights



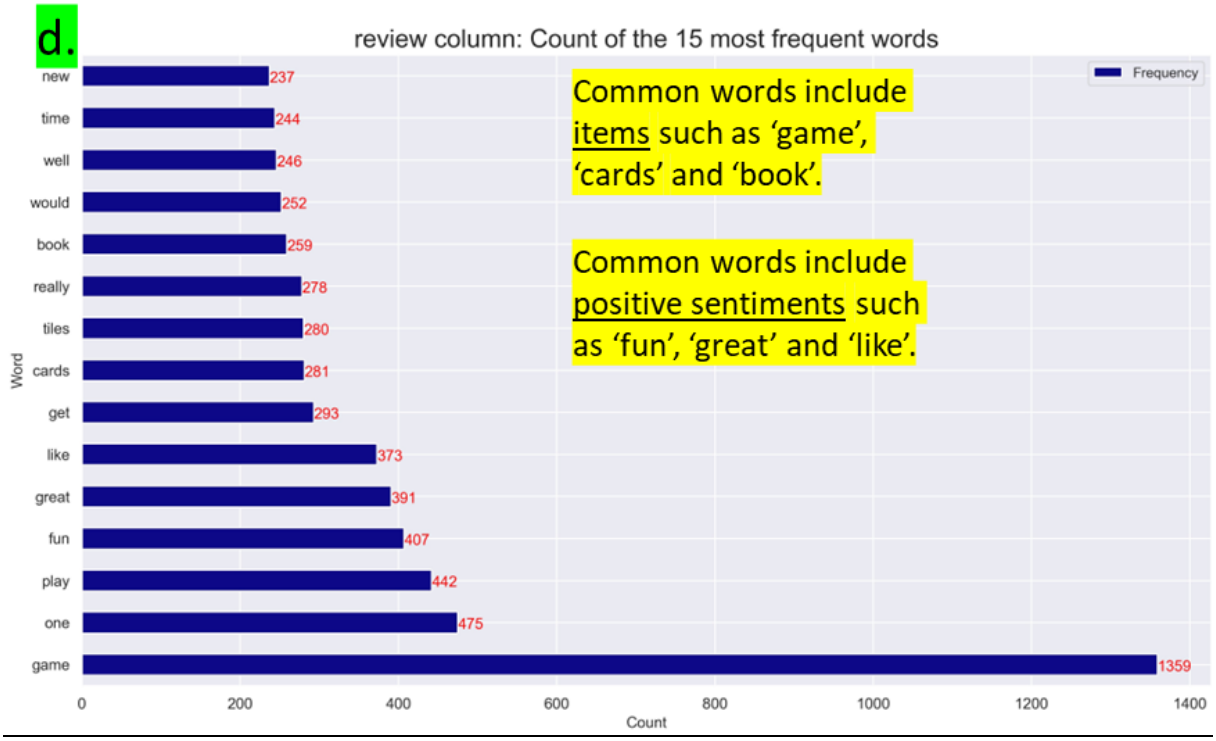
b.


Relationship between Customer income and Loyalty points attained
Positive linear correlation between customer income and loyalty points.
Loyalty Points
Customer Income (£k)
0
1000
2000
3000
4000
5000
6000
7000
0
20
40
60
80
100

c.


Customer Income vs Spending Score
5 discrete clusters(groups) of customers.
K-Means Predicted
0
1
2
3
4
Spending Score
Customer Income (£k)
0
20
40
60
80
100
20
40
60
80
100

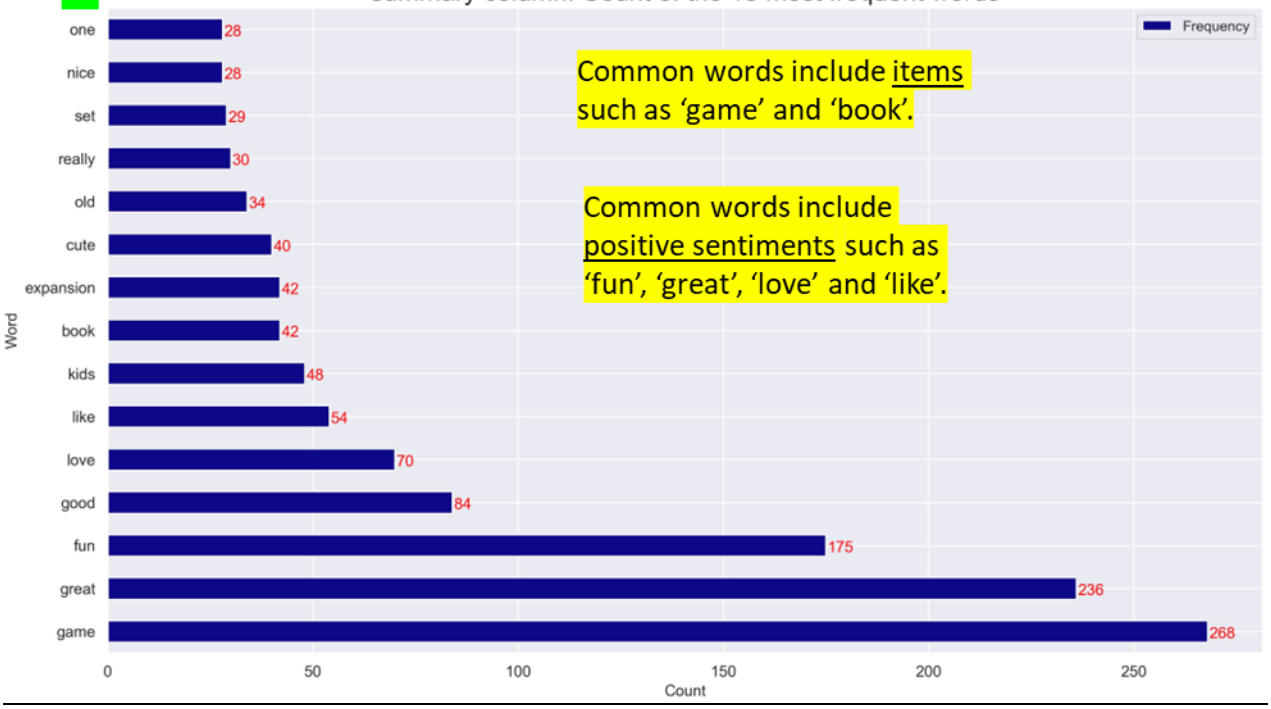
d.



Common words include items such as 'game', 'cards' and 'book'.

Common words include positive sentiments such as 'fun', 'great' and 'like'.

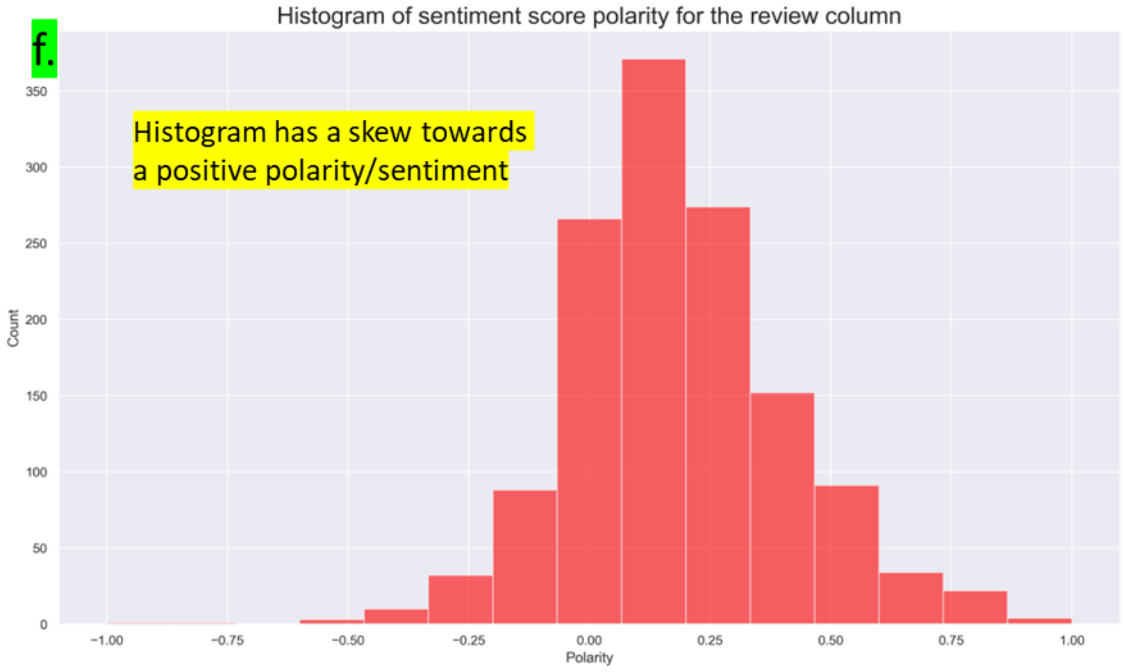
e.



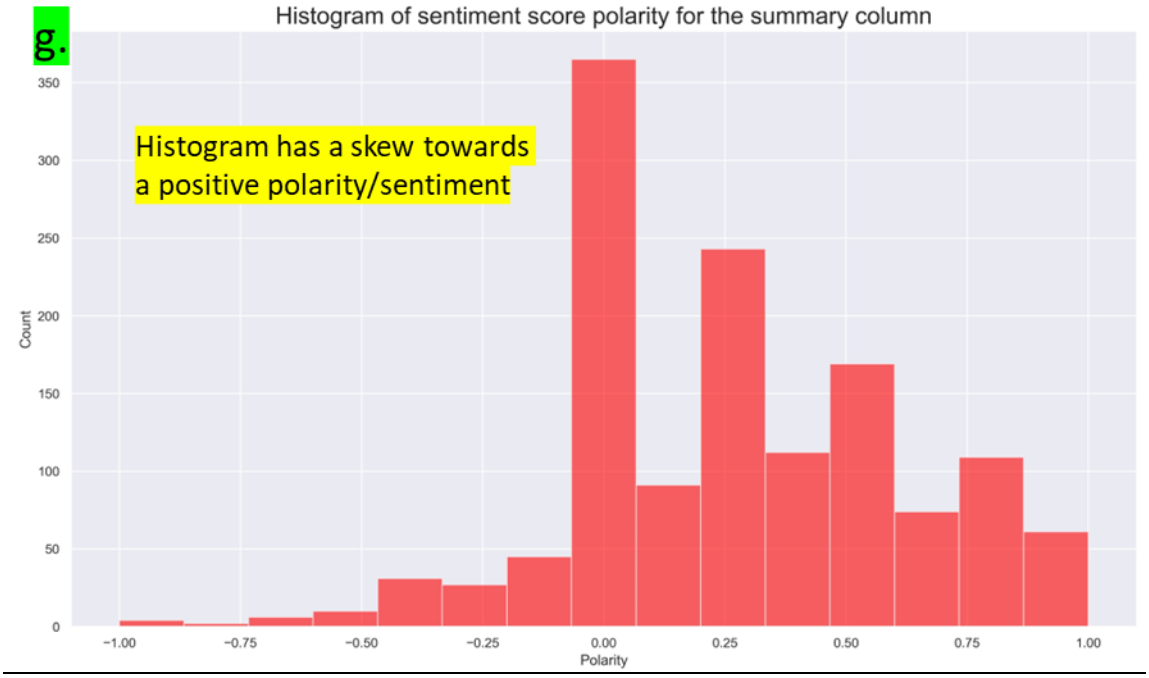
Common words include items such as 'game' and 'book'.

Common words include positive sentiments such as 'fun', 'great', 'love' and 'like'.

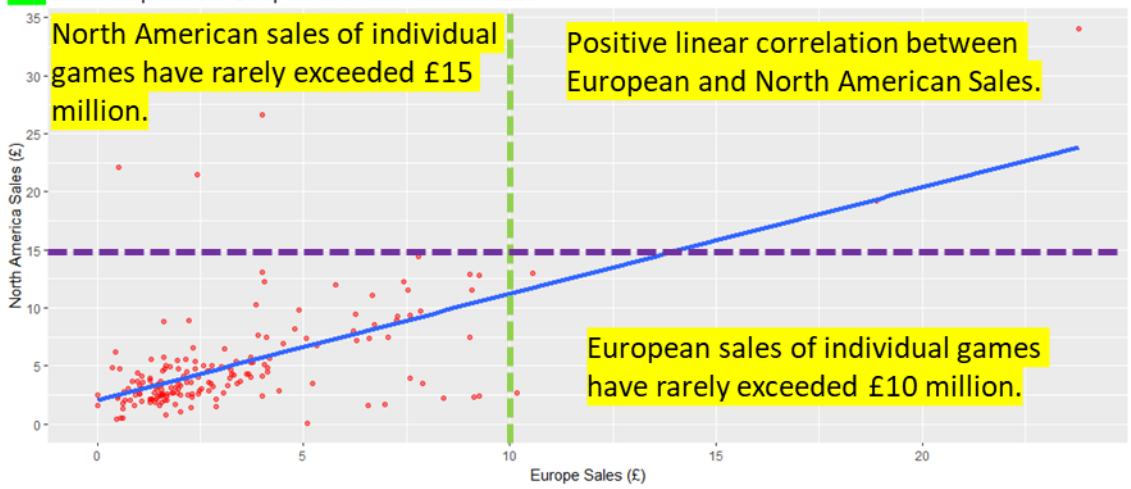
f.


Histogram of sentiment score polarity for the review column
Histogram has a skew towards a positive polarity/sentiment
Count
Polarity
0
50
100
150
200
250
300
350
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

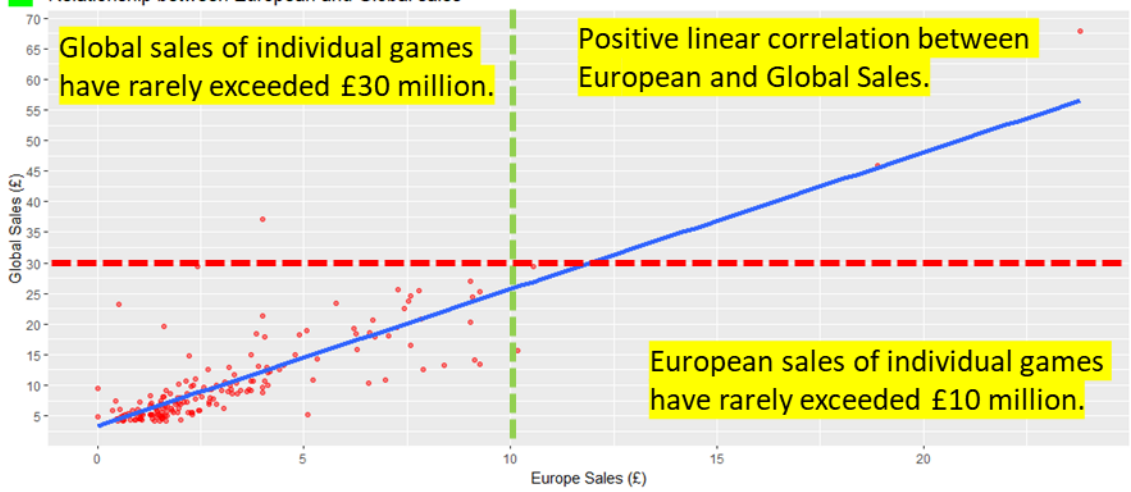
g.


Histogram of sentiment score polarity for the summary column
Histogram has a skew towards a positive polarity/sentiment
Count
Polarity
0
50
100
150
200
250
300
350
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

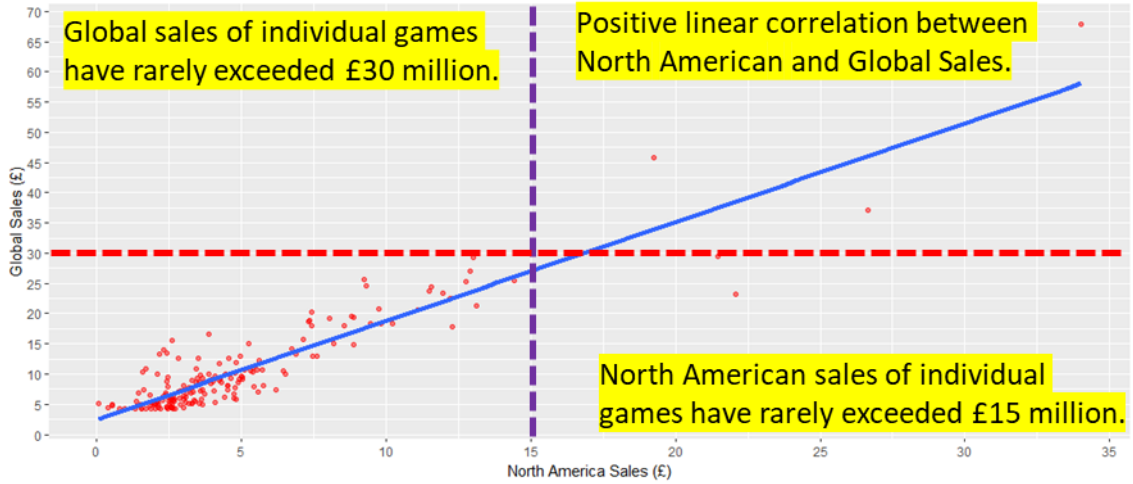
h. Relationship between European and North American sales

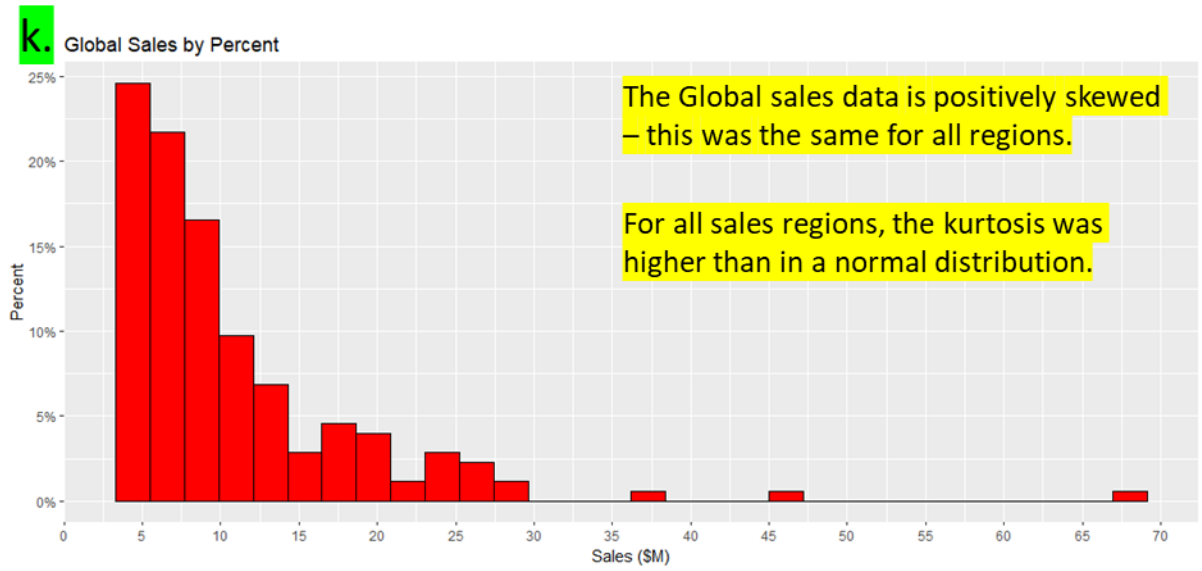


i. Relationship between European and Global sales



j. Relationship between North American Sales and Global sales





Patterns and Predictions

Plots a) and b) show that the customer income and spending score were both positively correlated with the loyalty points attained. However, plot a) shows customers with a high spending score and relatively low loyalty points(circled) – this needs to be investigated further to avoid unnecessary customer churn of customers with high purchasing power. In plot b), there is a dispersion of data so that it's further from the regression line once customer income exceeds ~£50k (two groups circled). The group containing high income and relatively low loyalty points may need to be incentivised to spend more.

Plot c shows five discrete customer groups:

1. Moderate income, moderate spending score (39% of sample).
2. High income, high spending score (18% of sample).
3. High income, low spending score (16.5%).
4. Low income, low spending score (13.5).
5. Low income, high spending score (13%).

Assuming that the sample is representative of the entire population, the focus ought to be on incentivising customers with a high income and low spending score to generate more revenue. As the most prevalent group has a moderate income and spending score, wholesale purchasers should consider whether the item will be affordable for this group of customers at Turtle Games.

Plots d) and e) show the most common words used in the review and summary columns – there was a positive sentiment and mention of items such as game, books and cards. The word 'kids' was also common which identifies a target audience for products – perhaps the range of books and cards can be expanded. Plots f) and g) support the notion of positive sentiment, but caution must be taken using machine learning tools to conduct sentiment analysis – it highlighted the word game as a negative but in the context of this scenario, it is neutral or positive.

Plots h), i), and j) show the relationship between European vs. North American sales, European vs. Global sales and North American vs. Global sales respectively. There is a positive linear correlation for all trends. It's noted that a game is unlikely to exceed sales of £10 million in Europe, £15 million in North America and £30 million globally.

Plot k) shows the distribution of sales of products globally, the same trend is seen for all sales regions. There is a positive skew. The kurtosis is higher than in a normal distribution. Caution must be taken when trying to predict sales using the supplied sales data. For example, in plots h), i) and j), the positive skew means fewer data points on the right-hand side of the scatterplot where sales are higher. This means there is a higher degree of uncertainty when predicting how well a game will sell in one region if it has sold well in another. There are data points on the right-hand side of the scatterplot which are significantly deviated from the regression line.

Recommendations

Give personalised product suggestions to customers and give incentives for referrals.

Become the exclusive retailer for pre-orders of highly anticipated releases.

Ensure all products give the same number of loyalty points per pound spent.

Encourage loyalty program engagement.

Incentivise high purchasing power customers with bonus loyalty points or promotions to encourage spending.

Expand the quantity and variety of books, cards and PS5 games sold.

Increase the marketing budget for PlayStation games in Europe and Xbox games in North America.

Use surveys to gain further insight into customer spending and product preferences.

Analysis could be enhanced by using twitter and social media data as well as YouTube sentiment analysis for products. More granular sales data by country rather than continent can help to extract more accurate insights. Recent sales data (e.g., weekly) as opposed to total sales can identify current best-sellers. It would also be useful to have sales data on all products sold by Turtle Games, not just video games.