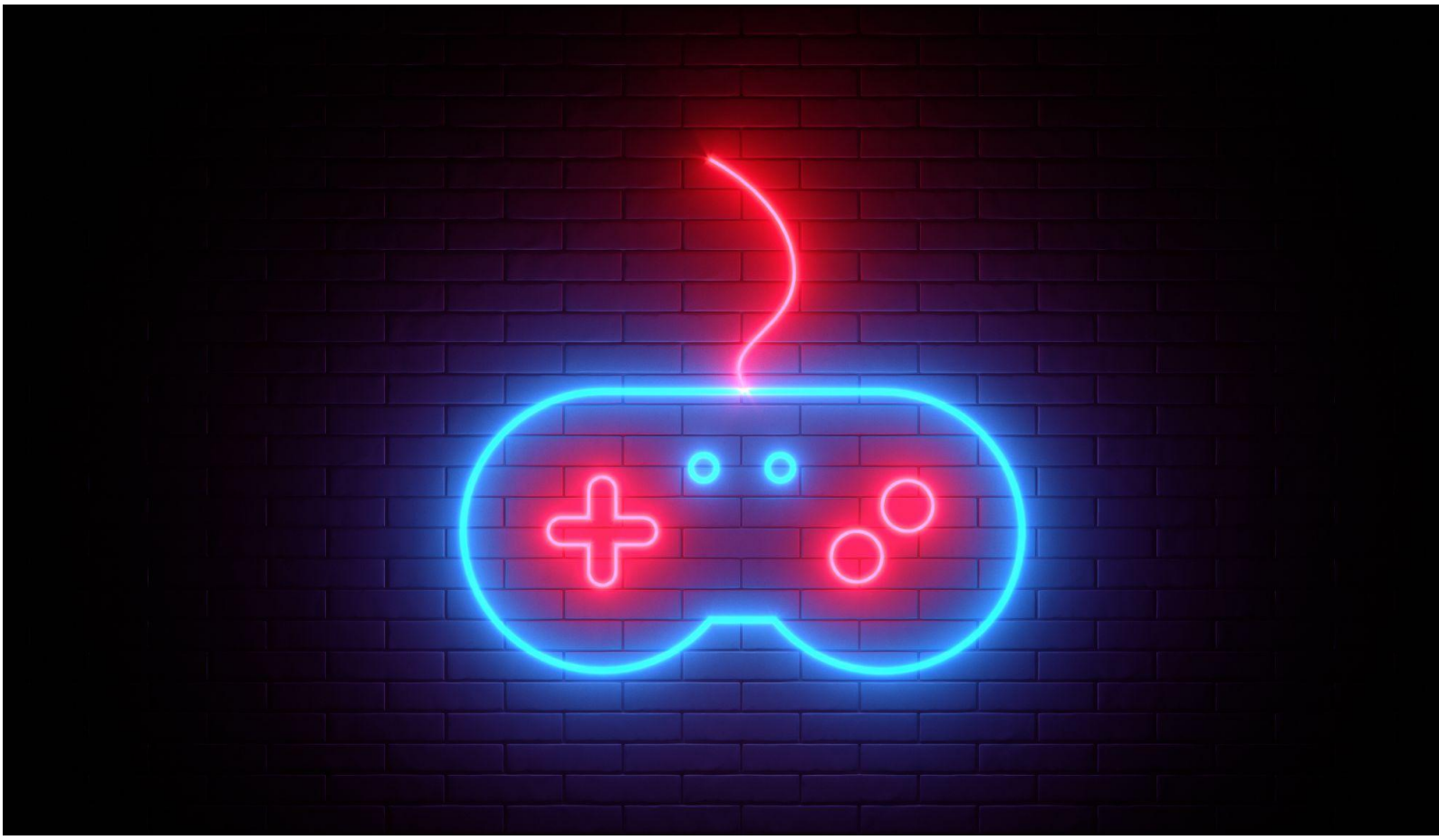


Turtle Games

Identifying customer trends and improving sales performance

Report by Nikki Stopford

Contents

Background and context

Turtle Games is an international game manufacturer and retailer. It manufactures and sells own-branded products and those from other companies. Its product range includes books, board games, video games and toys.

Problem statement

Turtle Games wants to analyse its sales and customer data to identify trends and insights that can be used to improve its overall sales performance.

Questions

Turtle Games would like this analysis to address the following critical business questions to help its decision-making:

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

Analytic approach

For analysis, Turtle Games provided two dataset:

1. **Customer reviews**: customer online reviews of products bought. This dataset includes customer demographics - including gender, age, income, education - and spending behaviour data.
2. **Customer sales**: sales data for video games sold globally, including game ranking, product information and volume of sales across North America, the EU and total global sales.

Data cleaning

I used Python to clean the 'Customer reviews' dataset:

- **Data completeness.** I checked whether there were any missing values. There were none.
- **Data consistency.** I checked the columns were formatted according to the correct data type.
- **Drop columns.** I dropped unnecessary columns.
- **Renaming columns.** For clarity, I renamed two columns in the reviews dataset:, 'remuneration (k£)' to 'income', and 'spending_score (1-100)' to 'spending_score'.

Python data analysis

I started by exploring the reviews data using a pairplot to visualise the strength of relationships between paired variables within the data. This helped identify patterns that may show a linear relationship and be suitable for linear regression analysis.

Business question 1: How do customers accumulate loyalty points?

Python was used to interrogate the customer reviews dataset to investigate how users accumulate loyalty points. Linear regression analysis was used to predict the likely impact that different variables have on the accumulation of loyalty points.

Simple linear regression modelling using the ordinary least squares (OLS) method was conducted to investigate the relationships between loyalty points, age, income and spending. Loyalty was set as the dependent variable (y) variable and spending, income and age as the independent (x) variables.

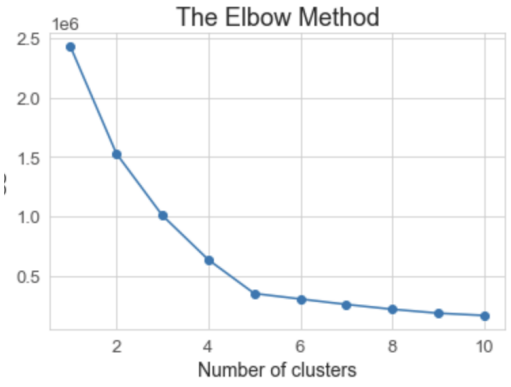
Further to this, I used the results of the simple linear regression modelling to identify the variables that would be most interesting to investigate how customers accumulate loyalty points using multiple linear regression modelling.

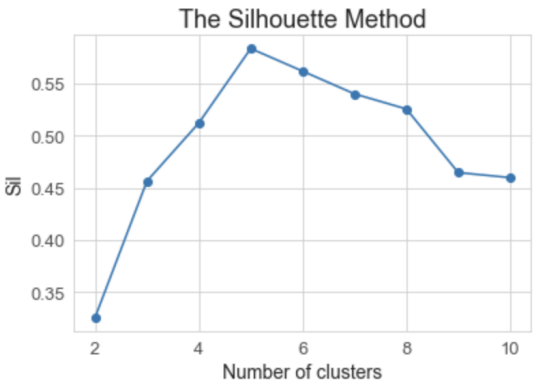
Business question 2: How can groups within the customer base be used to target specific market segments?

The Turtle Games' marketing team want to understand how useful the data on remuneration ('income') and spending ('spending_scores') is to help target specific market segments. This data is unlabelled so a clustering model has been used to help group and understand similarities in the data.

Specifically, the k-means clustering algorithm was used. This is a popular centroid model which provides an efficient way of clustering similar data points based on closeness to the centroid - or, in this case, by identifying the statistical mean of each cluster.

The initial scatterplot and pairplot highlighted five distinct clusters. The elbow and silhouette methods were used to validate this - see charts, below. Looking at both these charts, the optimal number of clusters is five but the model was also evaluated using four and six clusters before finalising on a k-value of five.





Business question 3: How can social data (e.g. customer reviews) be used to inform marketing campaigns?

Natural Language Processing (NLP) using sentiment analysis techniques was used to on the unstructured customer reviews to identify the 15 most common words used in online reviews, and the top 20 positive and negative reviews received from customers.

To prepare the data for analysis, the raw data (the unstructured review text) was first pre-processed. This included dropping unnecessary columns, checking for missing values, transforming the data to lowercase, removing punctuation and duplicates, applying tokenisation and visualisation techniques, and filtering out stopwords.

R data analysis

Business question 4: What impact does each product have on sales?

The sales department prefers for data analysis to be conducted using the R programming language. So, to address this question the Tidyverse package was used in R to analyse and explore the Turtle Sales data.

Business question 5: How reliable is the data (e.g. normal distribution, skewness or kurtosis)?

Statistical tests were conducted on the data to explore and determine the normality of the data set. These included:

- Shapiro-Wilk test to determine whether the continuous sales data follows a normal distribution, which is important to know when considering further statistical test which might require a normal distribution of data (e.g. *t*-tests).
- Skewness and kurtosis tests, using the Moments R package, to check the shape of the data

Business question 6: What are the relationships, if any, between North American, European and global sales?

Linear regression was used to model the relationships between the sales data variables.

Observations and insights

Business question 1: How do customers accumulate loyalty points?

Summary

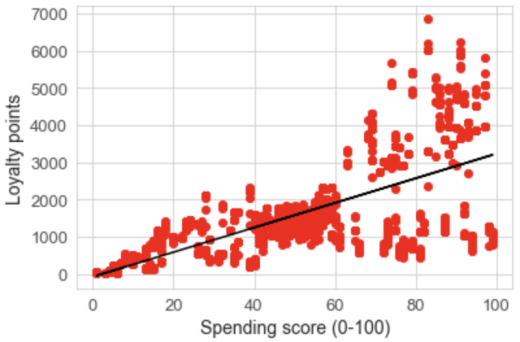
1. The 2000 Turtle Games customers have a mean average of 1578 loyalty points. This ranges from a minimum of 25 to a maximum of 6847. To explore drivers of loyalty points we have correlated this against customer spend, income and age.
2. Multiple linear regression analysis looking at accumulation of loyalty points shows that 83% of variation can be explained by customer spending and income. In other words, customers will have more loyalty points if they spend more and have higher incomes.
3. Breaking this down using simple linear regression analysis using the OLS method shows the amount spent by customers accounts for 45% of the variation in loyalty points and customer income accounts for a further 38%. Age has minimal impact on how customers accumulate loyalty points.

Spending v loyalty

The line of best fit in chart 1 indicates a positive correlation between the accumulation of loyalty points (the dependent variable) and customer spend (the independent variable).

The OLS regression results shows an R-squared value of 0.452. In other words, 45% of the variation in loyalty points can be explained by customer spending. That currently leave 55% of variation unexplained.

Chart 1: Linear regression - spending v loyalty

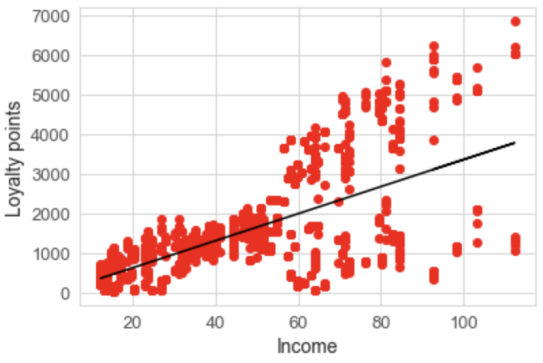


Spending v income

The line of best fit in chart 2 indicates a positive correlation between the accumulation of loyalty points (the dependent variable) and customer income (the independent variable).

The OLS regression results shows an R-squared value of 0.38. In other words, 38% of the variation in loyalty points can be explained by customer spending.

Chart 2: Linear regression - income v loyalty

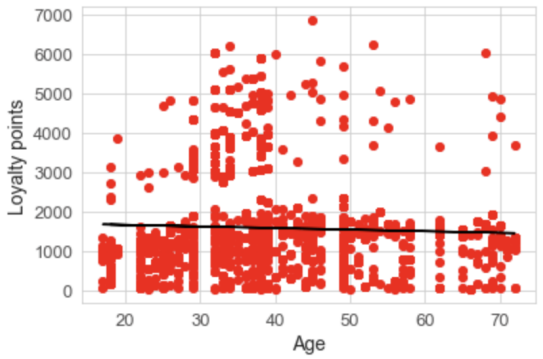


Age v loyalty

The line of best fit in chart 3 indicates no positive correlation between the accumulation of loyalty points (the dependent variable) and customer age (the independent variable).

The OLS regression results shows an R-squared value of 0.002. In other words, 0.2% of the variation in loyalty points can be explained by customer spending.

Chart 3: Linear regression - age v loyalty

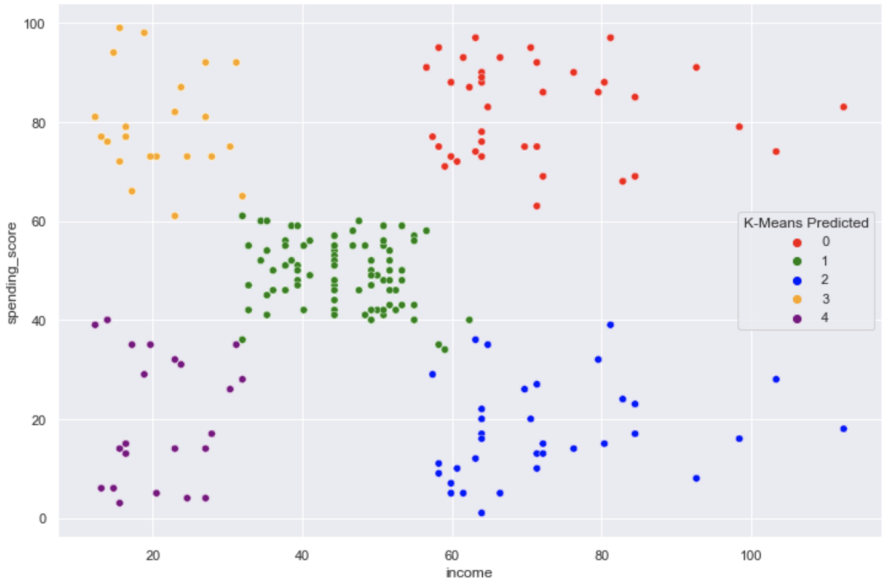


Business question 2: How can groups within the customer base be used to target specific market segments?

Using the the k-mean model with k=5, cluster 1 (middle income/middle spend) has most data points, followed by cluster 0 (high income/high spend) and then cluster 2 (high income/low spend).

These three customer segments could be used by the marketing team to target future advertising. A particularly interesting target market could be cluster 2, these customers are high earners but have lower spending scores so the ambition could be to move them into the high income/high spend segment.

Chart 4: K-Means visualisation



Business question 3: How can social data (e.g. customer reviews) be used to inform marketing campaigns?

The sentiment analysis of textual social data comments shows a leaning towards positive comments but there is no extreme sentiment in either direction - see the histogram of polarity plots for Reviews (chart 5) and Summary (chart 6) comments.

Chart 5: Histogram of polarity for reviews comments

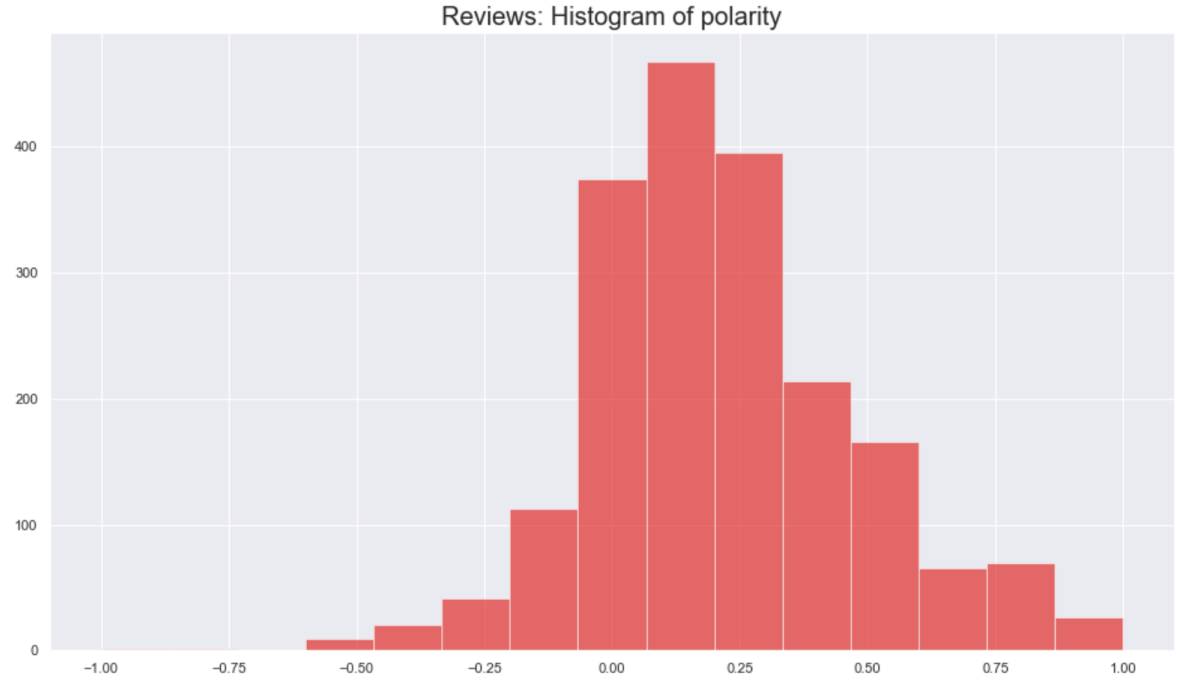
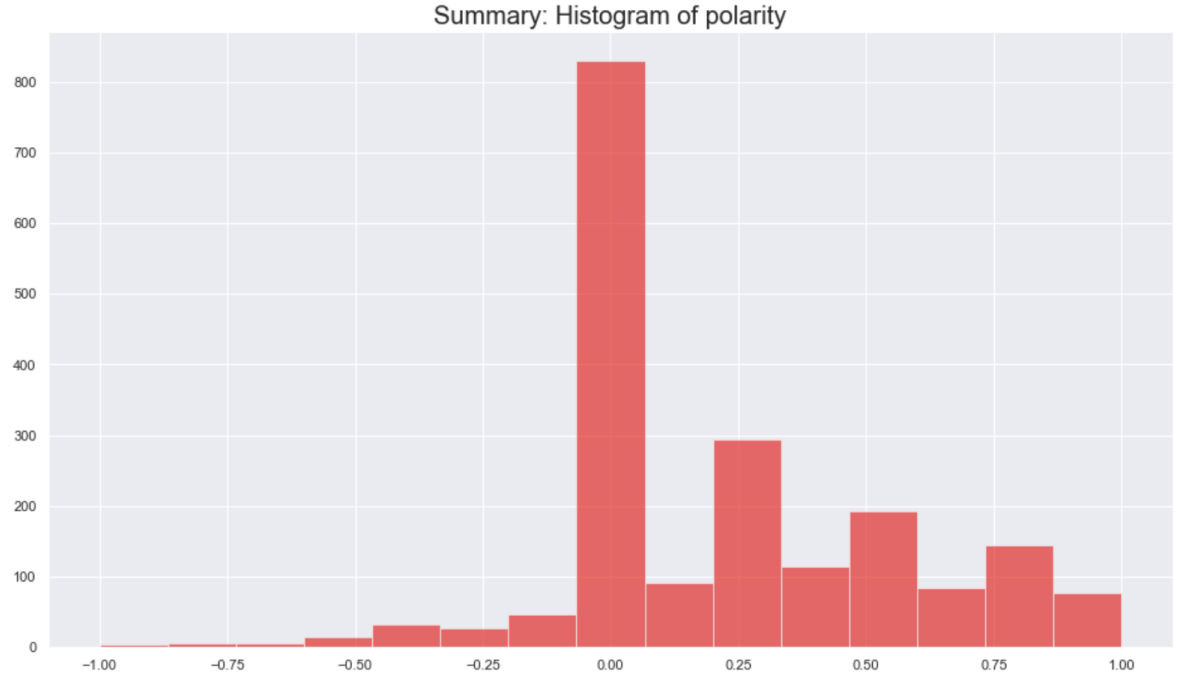


Chart 6: Histogram of polarity for summary comments

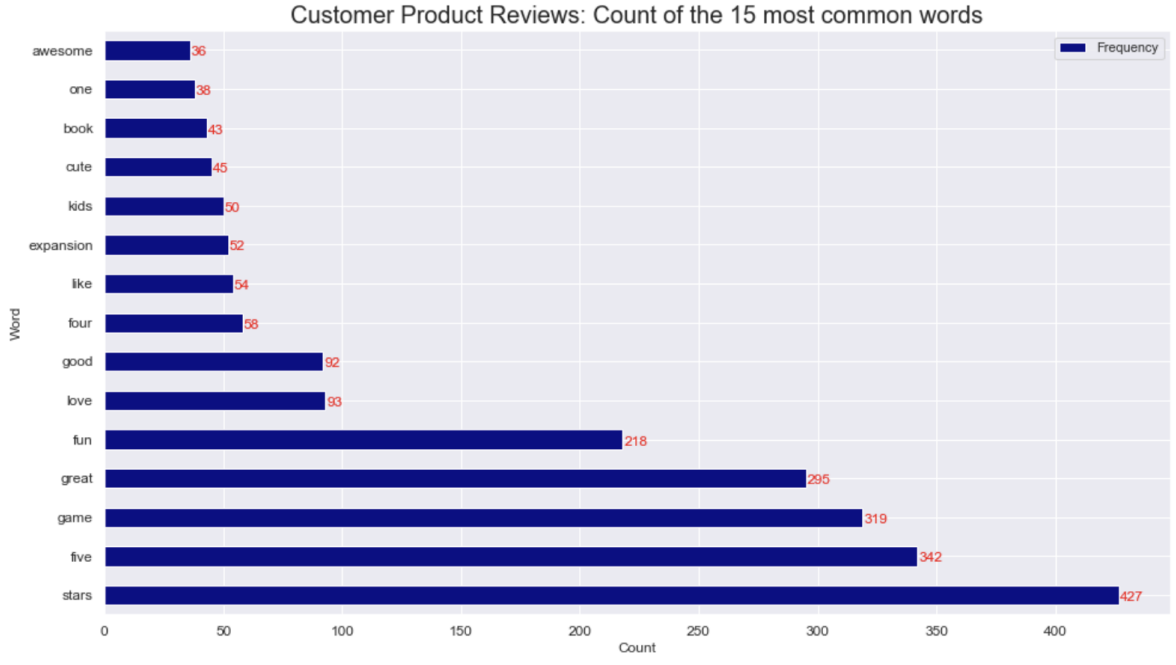


The most common words left in reviews tend to be positive - as can be seen in the

wordcloud visualisation (chart 7) and the plot of the 15 most common words used in customer product reviews (chart 8). Keywords like ‘five stars’, ‘great’, ‘fun’, ‘love’, ‘good’ feature strongly.

Chart 7: Customer product reviews wordcloud

Chart 8: 15 most commons words in customer product reviews



Taking a closer look at the top 20 positive reviews and top 20 negative reviews received from customers gives a strong indication of what customers did and didn’t like.

A theme coming from the more negative reviews showed that customers found products bought difficult or complicated to use. For example, one dissatisfied customer

wrote:

"Booo, unless you are patient, know how to measure, I didn't have the patience neither did my daughter. Boring unless you are a craft person, which i am not".

Conversely, more positive customer experiences revealed much higher satisfaction with the products bought. As one very satisfied customer wrote:

"Excellent activity for teaching self-management skills."

As an additional piece of analysis it would be useful for the team to interrogate whether there are any trends in the products bought - so, are some products more likely to elicit negative feedback than others.

Business question 4: What impact does each product have on sales?

The sales data has a total of 352 rows of data for products sold globally. Product lines are separated by the different types of gaming platforms bought for. Product data has been aggregated to explore trends in the data by product id. This aggregated data shows a total of 175 product types sold.

Looking at the sales revenue by product, Product 107 has generated the highest revenue in sales for Turtle - both across North America and Europe - generating a total of £67.85m in sales globally.

Exploratory visuals indicate there is a positive correlation between product sales across North America and in Europe. The scatterplot in chart 9 illustrates the tendency for products to generate similar revenue levels across regions.

There are some outliers where products have higher sales revenues in one region. An example is product 123 which generated £26.64m in revenue in North America but only £4.01m in Europe.

America and Europe

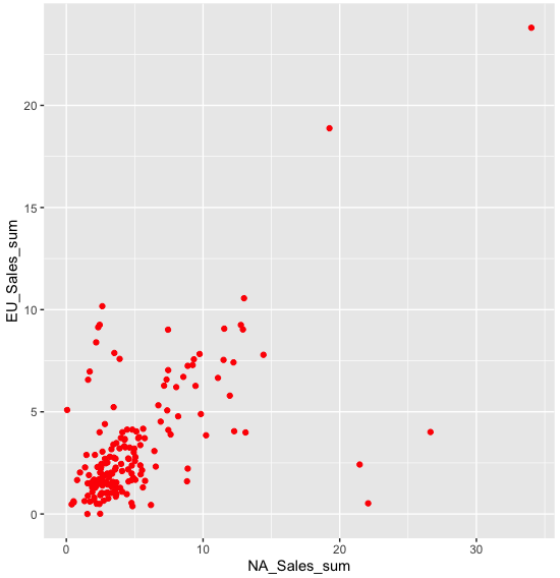
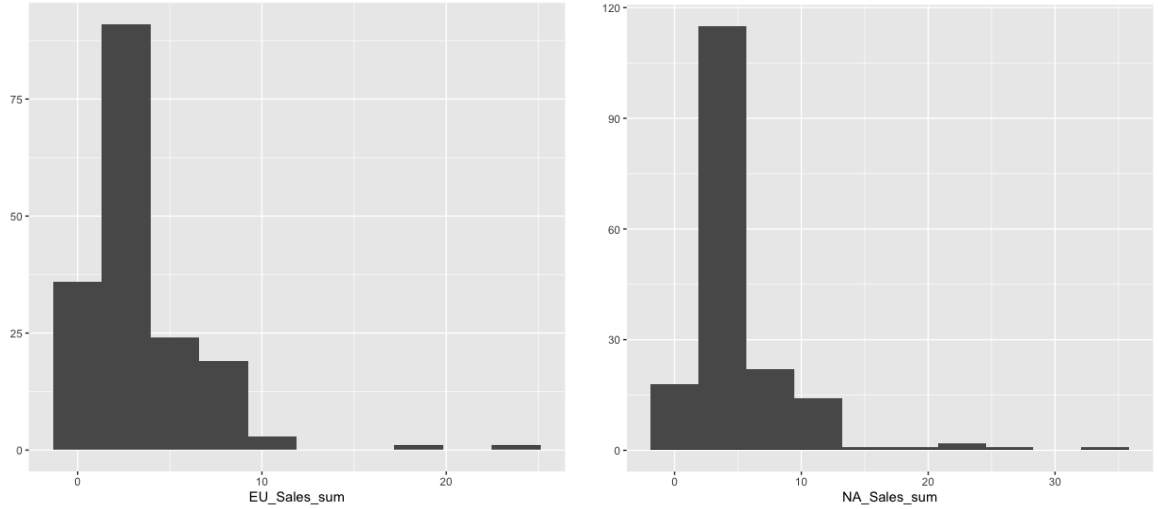


Chart 9: Scatterplot of product sales in North

Chart 10 illustrates histogram plots in 'bins' of 10 the product revenues in North America and EU. Both plots show a similar shape to the data across regions - with

products sales across North America and Europe being positively skewed. This shows that most products generate sales of less than £10m.

Chart 10: Histograms for product sales in North America and Europe

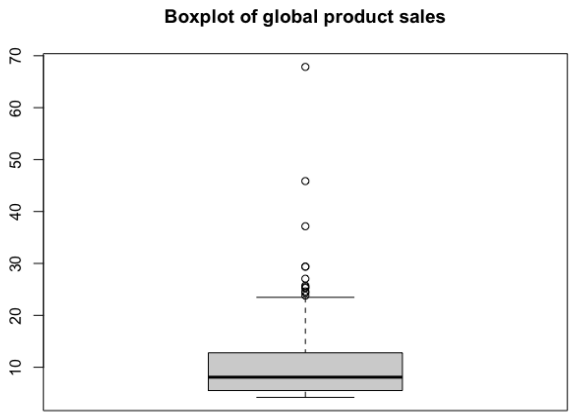


Mean and median calculations of aggregated sales data across North America, Europe and globally result in different values with higher mean values than median. This suggests the mean is being biassed by high values. We will confirm this with kustosis and skewness tests for the next business question.

Examining the descriptive statistics more closely for global sales grouped by product we can see that the median average of sales across the 175 products is £8.09m.

The lowest global sales achieved for a product was £4.2m. The middle 50% of product sales (the interquartile range) fall between £5.52m and £12.79m. This is illustrated in the boxplot in chart 11.

Chart 11: Boxplot of global product sales



Business question 5: How reliable is the data (e.g. normal distribution, skewness or kurtosis)?

The histogram for global sales (see chart 12) shows the right skewedness of the data. This means the data does not follow a normal distribution. The boxplot for global sales, chart above, shows visible outliers in the data.

Chart 12: Histogram of global sales

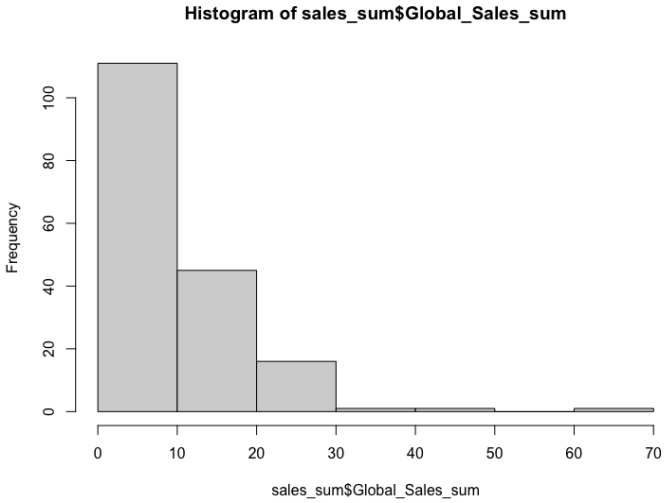
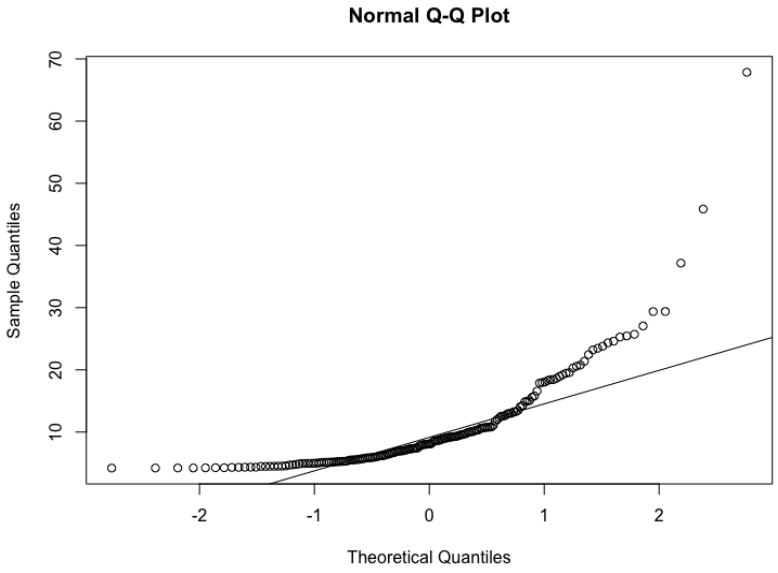


Chart 13 also indicates that the data isn't following a normal distribution because the data points on the Q-Q Plot are not following the straight reference line.

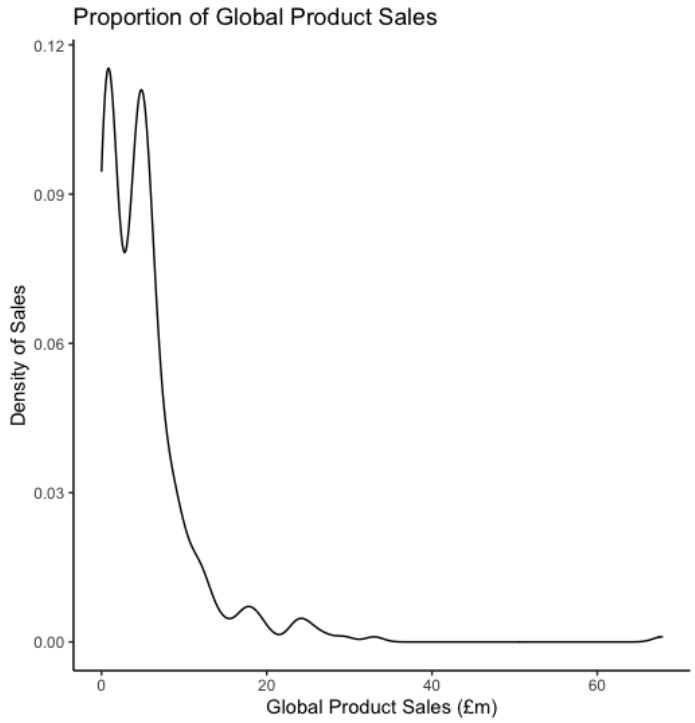
Chart 13: Quantile-Quantile Plot of product global sales



A Shapiro-Wilk normality test was conducted to confirm that the results do not follow a normal distribution. The results across the sales data - NA, EU and global - show a p-value that is very close to zero (for example, for global sales the p-value < 2.2e-16), which confirms the data is not normally distributed.

The skewness test confirms as we've seen visually that the data is positively (or right) skewed - see Chart 14. We saw evidence of this from the results of the mean and the median, where the mean is larger than the median. Data is considered to be normal if the skewness is less than -1 and greater than 1. Skewness results for all sales data is above 2.

Chart 14: Proportion of Global Product Sales by £m

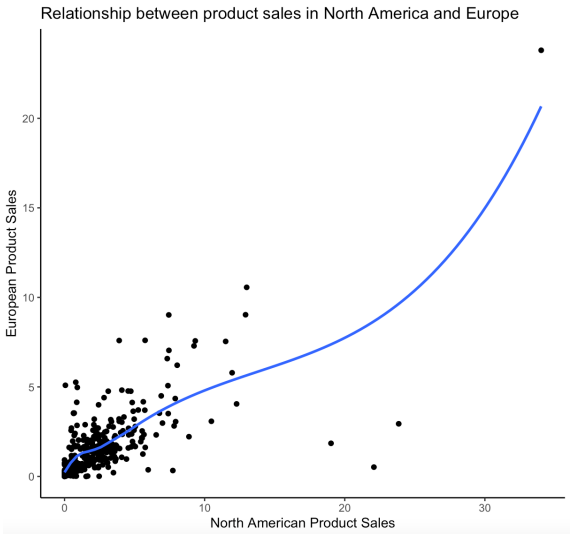


A normal distribution will have a kurtosis coefficient of 3. The values for the sales data are much higher than this (for example, 17.8 for global sales), which indicates that a leptokurtic (or heavy-tailed) distribution. This confirms that we are seeing more extreme outliers than you would in a normal distribution.

The Pearson's correlation coefficient (*r*) test was run to determine whether there is a correlation between the sales data columns. This test is typically only used when on data with a normal distribution. However, the correlation coefficients of the sales data columns suggest a strong positive correlation - so, for instance, higher sales of a

product in Europe typically correlates with higher sales of a product in North America. We can see this clearly in chart 14.

Chart 15: Correlation between North American and European product sales



Business question 6: What are the relationships, if any, between North American, European and global sales?

There is a strong positive and highly significant correlation between North American sales and global sales, with North American sales explaining 83.85% variability in global sales (see Chart 16). Similarly, the correlation between European sales and global sales is a highly significant value, explaining 71.85% variability in global sales (see Chart 17). The relationship between European and North American sales is weaker, explaining only 38.2% of variability. Together, European and North American sales account for 96.64% of variation in global sales.

Chart 16: Correlation between North American and Global product sales

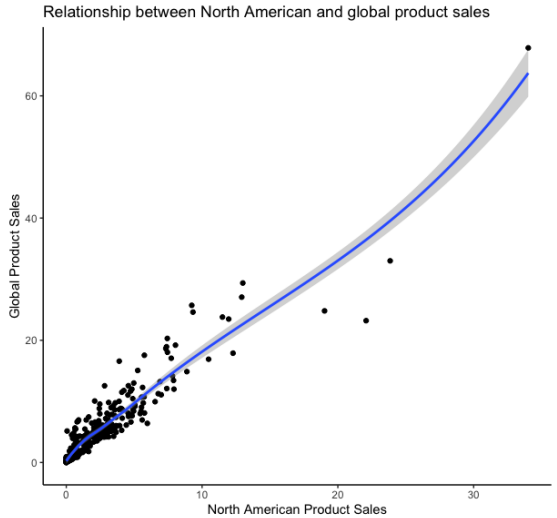
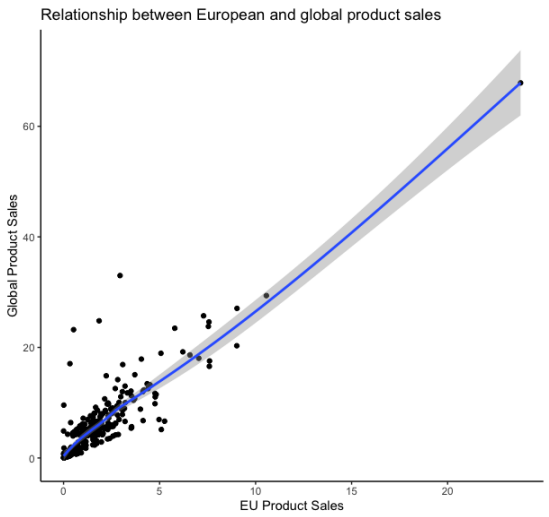


Chart 17: Correlation between European and Global product sales



Conclusions and recommendations

Action 1 Understand how loyalty rewards are being spent to inform marketing.

Analyse how high-spend, high-income customers are spending loyalty rewards to better target advertising and marketing to our high-income low spend customers with the intention of moving them into the high-income/high-spend segment.
This action is based on the insight that customers who spend more and are on higher incomes drive the accumulation of loyalty points (business question 1) and that high-income/low-spend is our third biggest customer segment (business question 2).

Action 2 Analyse negative reviews based on products bought to identify any trends.

Carry out further analysis on negative reviews to identify whether there are any trends with specific products bought being more likely to elicit negative reviews. This insight can be used to eliminate products from the range or improve product information.
This action is based on the sentiment analysis that showed a trend in negative reviews of customers finding the products they bought difficult or complicated to use (business question 3).

Action 3 Identify more big hitting product lines.

Learn from the success of produce 107, which generated high sales revenues in Europe and North America, and identify similar products that might be as successful.
This action is based on the middle 50% of product sales (the interquartile range) generate between £5.52m and £12.79m in revenue (business question 4). But we're seeing a small number of best-selling products (business question 5). European and North American sales account for 96.64% of variation in global sales (business question 6).