# Turtle Games Analysis

## Commission

This report aims to provide insights into customer trends to improve sales performance at Turtle Games.

## Key Results

Accumulation of **loyalty points is associated with age, remuneration, spending score, and gender**. In a multiple linear model, loyalty points increase as age, remuneration, and spending score increase, and if the customer is female.

**Five distinct groups have been identified in the customer base** using K-means clustering with remuneration and spending score. We suggest that grouped customers will respond similarly to marketing, and so different groups can be targeted with specific campaigns.

Results of a sentiment analysis on customer reviews can also inform our marketing campaigns – for example, **the most positive reviews have been identified**, and could be **displayed in marketing materials**. We could also focus on marketing the products that receive the most positive (on average) reviews, as these products appear likely to have a positive reception.

**Product 107 was the best-selling product** by global, North American, and European sales. We suggest that different product genres may be preferred in North America (platform) compared to Europe (sports), and this insight could be utilised in marketing.

The sales data are not normally distributed, containing positive skew and extreme values. It is suggested that this **could lead to non-normal residuals** in a linear model based on this data. This may make the **linear model less reliable**, as the errors are not random. Further analysis is recommended to investigate this further and possibly transform the sales data.

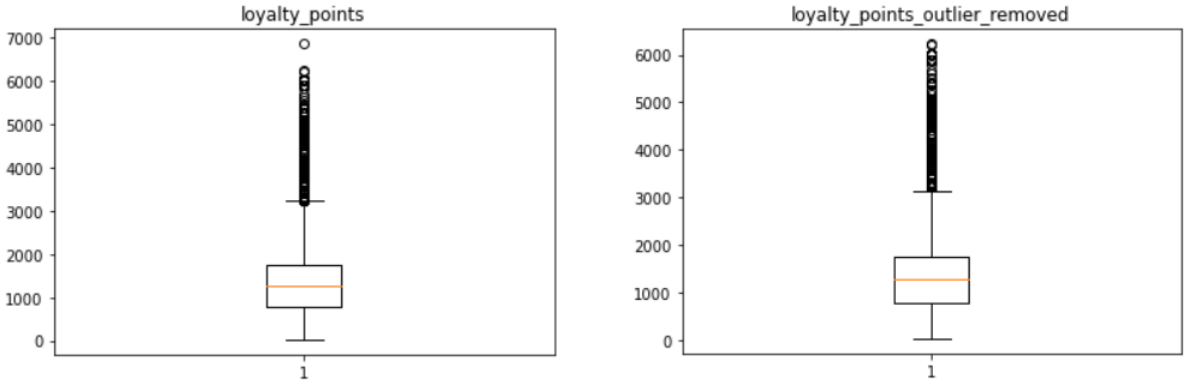
North American, European, and global sales are all positively correlated. A multiple linear model can **predict the global sales with some accuracy from North American and European sales**, which may be useful in forecasting total sales when games are released on different dates across the world.

## Approach

The customer data (*turtle_reviews.csv*), which included 2,000 data points, was analysed using Python. There were no missing values in the dataset.

A boxplot of customer loyalty points revealed some high outliers. One of these outliers was visually identified as extreme, suggesting a possible data entry error, and this data point was therefore removed from the analysis. No other outliers in loyalty points (*mean=1575.4*) were removed.

No outliers were identified in age (*mean=39.5*), remuneration (*mean=48.0*), or spending score (*mean=50.0*).

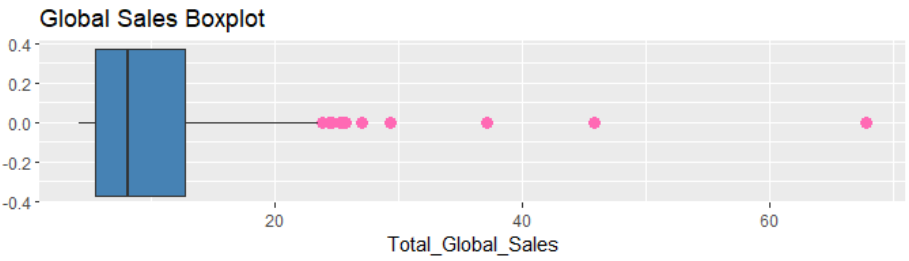


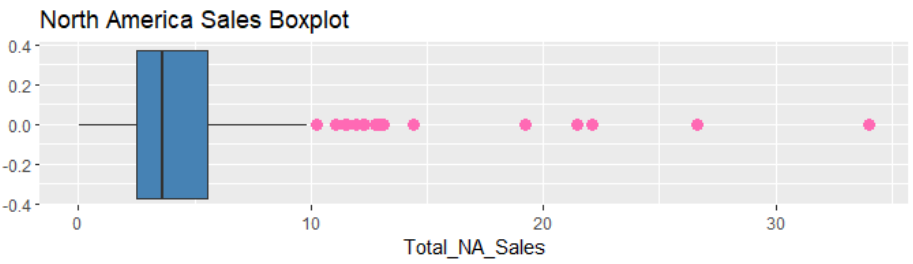

**Question for Turtle Games:** It may be useful to understand how the summary score has been calculated, so that we can more comprehensively highlight trends in the data. For instance, does a higher spending score indicate higher spending?

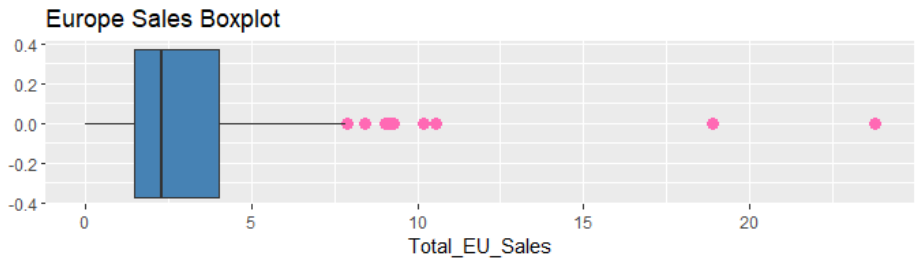
The sales data (*turtle_sales.csv*), which included data on global, North American (NA), and European (EU) sales, was analysed in R. Sales were in millions of GBP (£).

The sales for each product were grouped together, regardless of gaming platform, to give the total global sales (*mean=10.73*), total NA sales (*mean=5.06*), and total EU sales (*mean=3.31*) of 175 products.

Boxplots reveal that there are high outliers in each of the three sales categories. Outliers were not removed at this point in the analysis, as we planned to investigate these top-selling products.



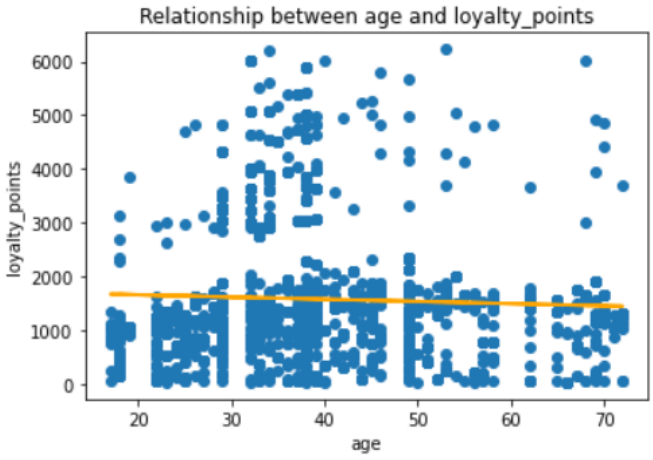

**Question for Turtle Games:** It may be useful to see the sales of each product by year. For example, do almost all sales occur within the first year of release, or do sales tend to gradually decline year on year? This information may allow us to predict not just the total sales, but the total sales per year.

# Customer Insights

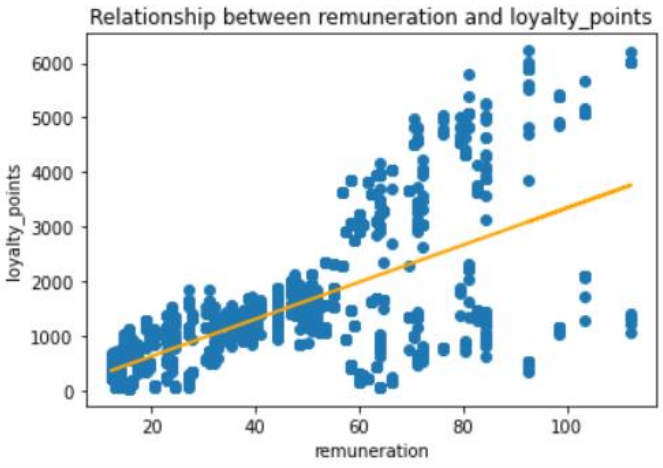
### *Accumulation of Loyalty Points*

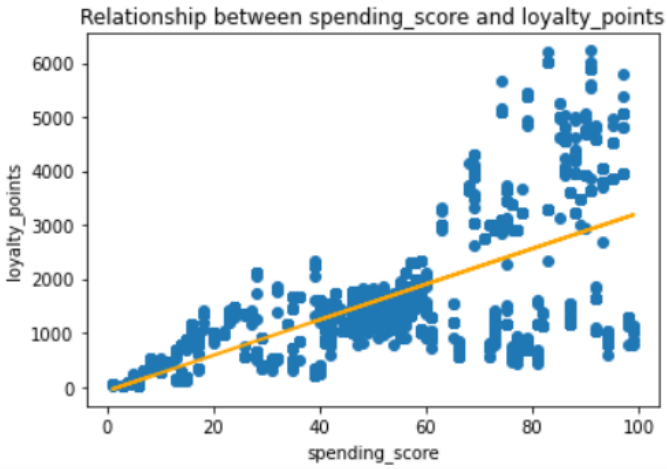
Simple linear regression was used to investigate the relationship between the dependent variable loyalty points, and each of the independent variables: age, remuneration, and spending score.

Viewing the scatter plot, there does not appear to be a linear relationship between age and loyalty points (*r = -0.04*).




Both remuneration (*r = 0.61*) and spending score (*r = 0.67*) appear to have a positive correlation with loyalty points, although we notice from the scatter plots that the scatter around the regression line appears wider at higher values.



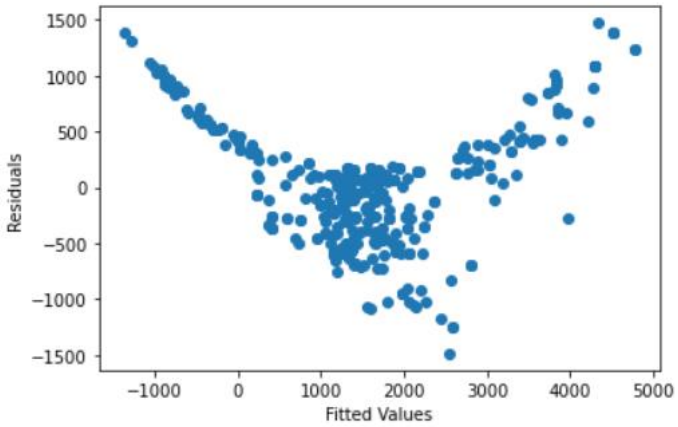




Multiple linear regression was used to investigate whether loyalty points are better predicted when taking into account multiple variables.

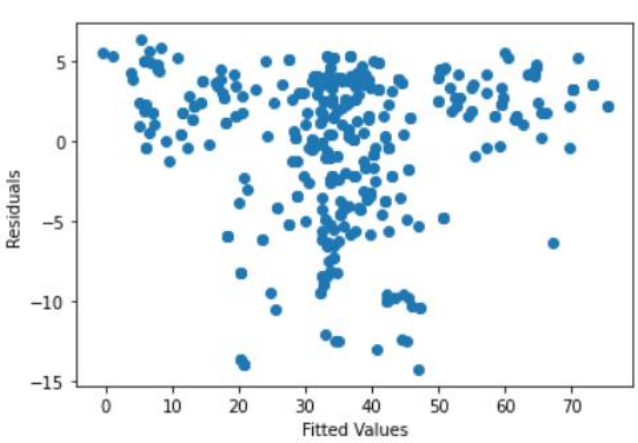
A model with variables age, remuneration, and spending score explained 84% of the variance in loyalty points ($R^2 = 0.840$, *adj* $R^2 = 0.840$). Despite not appearing correlated with loyalty points when assessed alone, age was a significant variable in this model ($p < 0.05$).

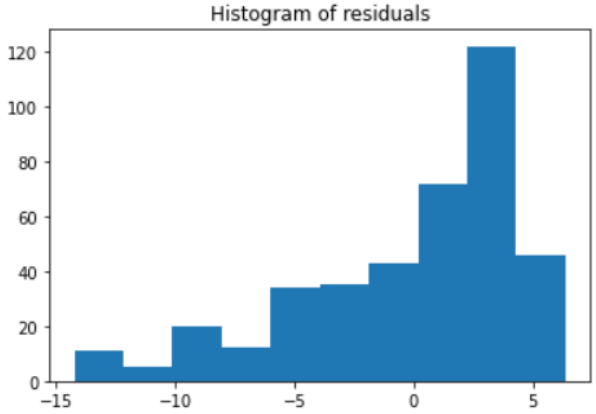
We attempted to add a gender indicator to this model, which improved the model very slightly ($R^2 = 0.841$, *adj* $R^2 = 0.841$). The gender indicator was a significant variable ($p < 0.05$).

When checking the assumptions of the model, however, the residual plot displayed a pattern rather than random scatter, suggesting heteroscedasticity in the model. This was confirmed by a Breusch–Pagan test (*LM p-value < 0.05*).




To resolve this, we created a new model, using age, remuneration, spending score and gender to predict the square root of the loyalty points. This resolved the issue (*LM p-value = 0.22*), and appears to be a strong predictive model ($R^2 = 0.907$, *adj* $R^2 = 0.906$). We take this as our final model.




The Durbin-Watson statistic for this model is 2.018, indicating that there is almost no autocorrelation. However, we also note that for this particular model, the assumption that residuals follow a normal distribution may not be met. The distribution of the residuals is left-skewed.

This final model can be used to predict the square root of loyalty points, and therefore loyalty points. For example, for a 40-year-old female customer with a remuneration value of 60 and a spending score of 50:

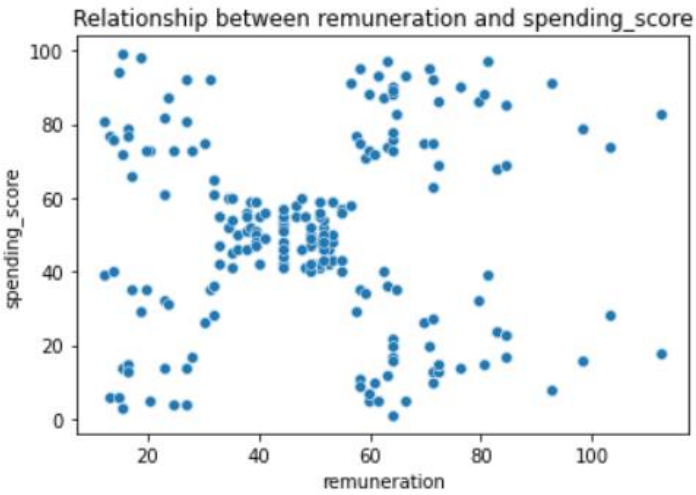
$$\sqrt{loyalty\ points} = 0.1527(40) + 0.4077(60) + 0.4393(50) + 1.3981(1) - 11.7737 = 42.1594$$

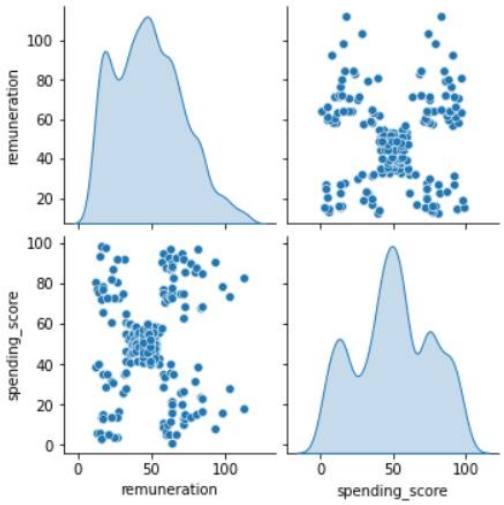
$$loyalty\ points = 42.1594^2 = 1777.4150 = 1777$$

**Summary:** We have found that customer loyalty points can be predicted using age, remuneration, spending score, and gender, providing some insight into the accumulation of loyalty points.

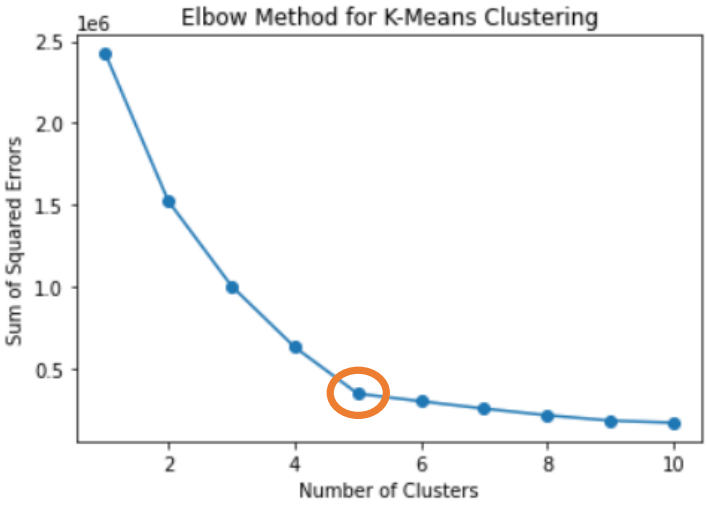
### *Groups within the Customer Base*

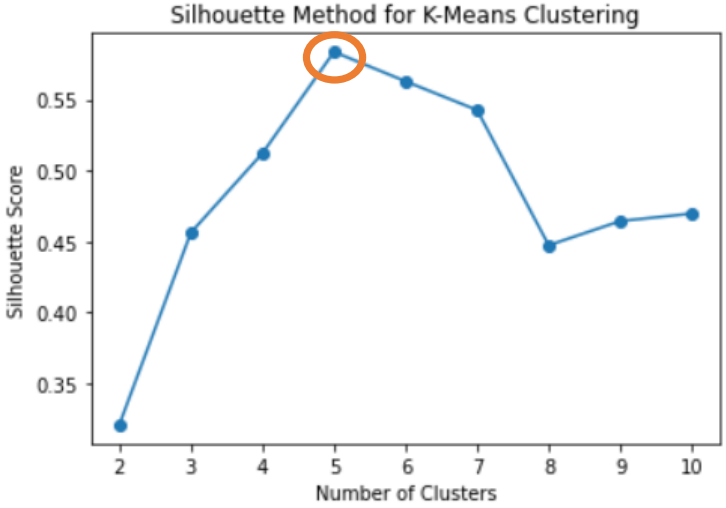
K-means clustering was used to identify different customer groups, using remuneration and spending score. Plotting these variables against each other, we can see that there do appear to be some clusters.




The Elbow method and the Silhouette method were both used to determine the optimal number of clusters (*K*). As highlighted below, both methods indicated *K*=5 (identified by the elbow in the curve with the elbow method, and the highest point on the curve using the silhouette method.)

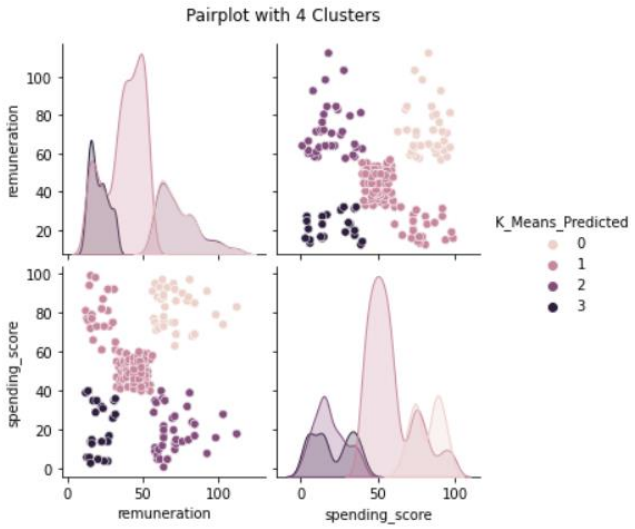


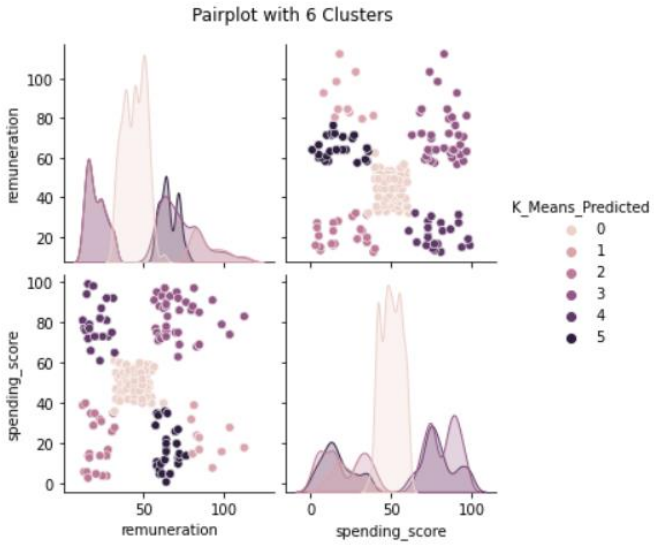



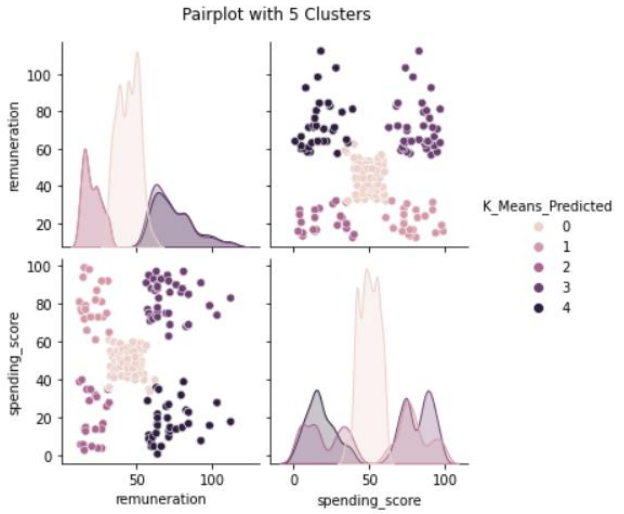

However, we also decided to investigate the clusters with *K*=4 and *K*=6, for comparison.

Looking at the distribution across clusters, it appears that the clusters in *K*=4 are quite unbalanced, with a large proportion of customers (1013 of 1999) falling into one cluster. With *K*=6, there is a concern of overcomplexity. In practical terms, it may be that a lower number of groups is easier to manage when it comes to marketing.

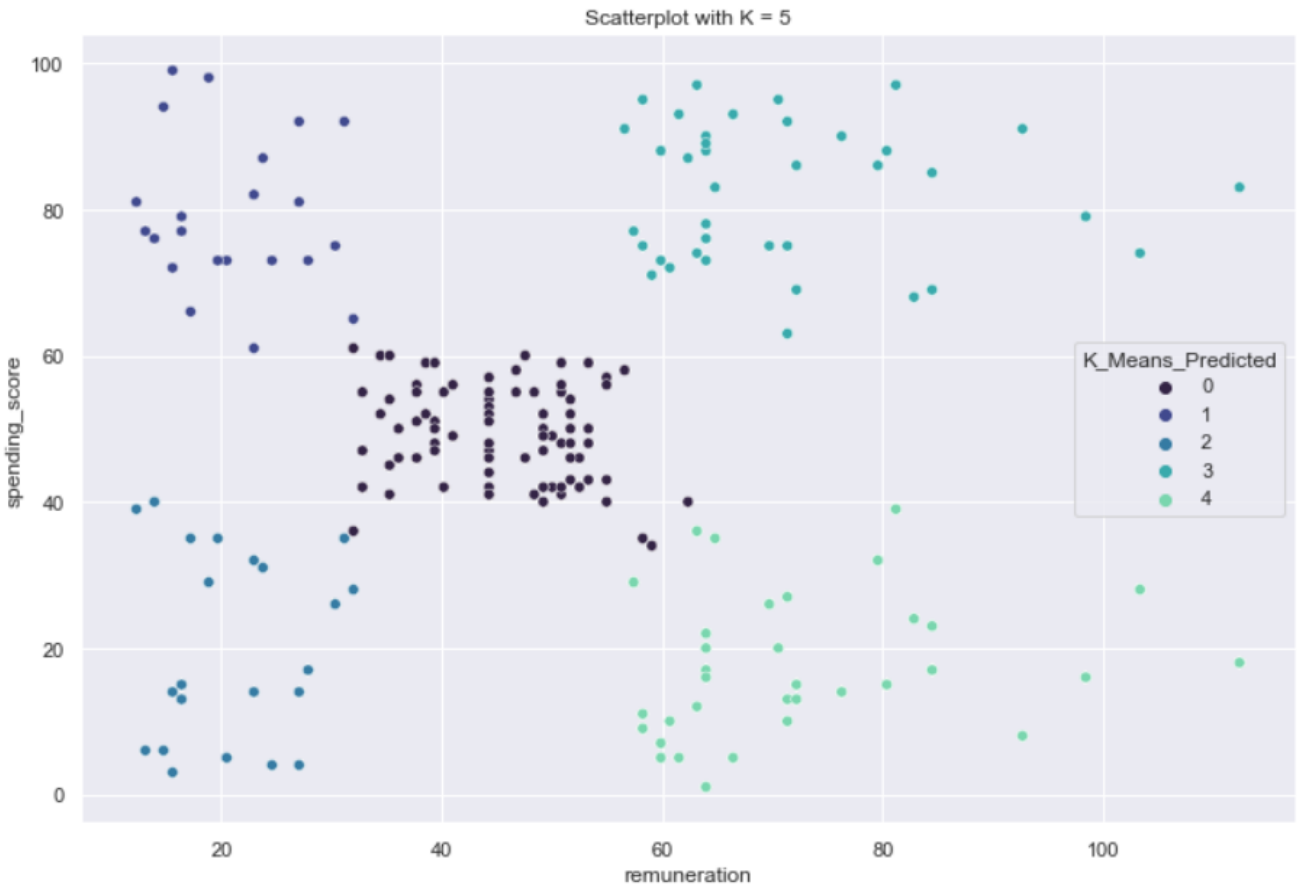
Viewing the figures for *K*=5, we confirm that five appears to be the optimal number of customer groups.




**Summary:** We have identified five groups of similar customers using remuneration and spending score. We suggest that customers within these groups may respond similarly to marketing, and therefore, customers can be targeted with specific campaigns based on their group.



***Customer Reviews***

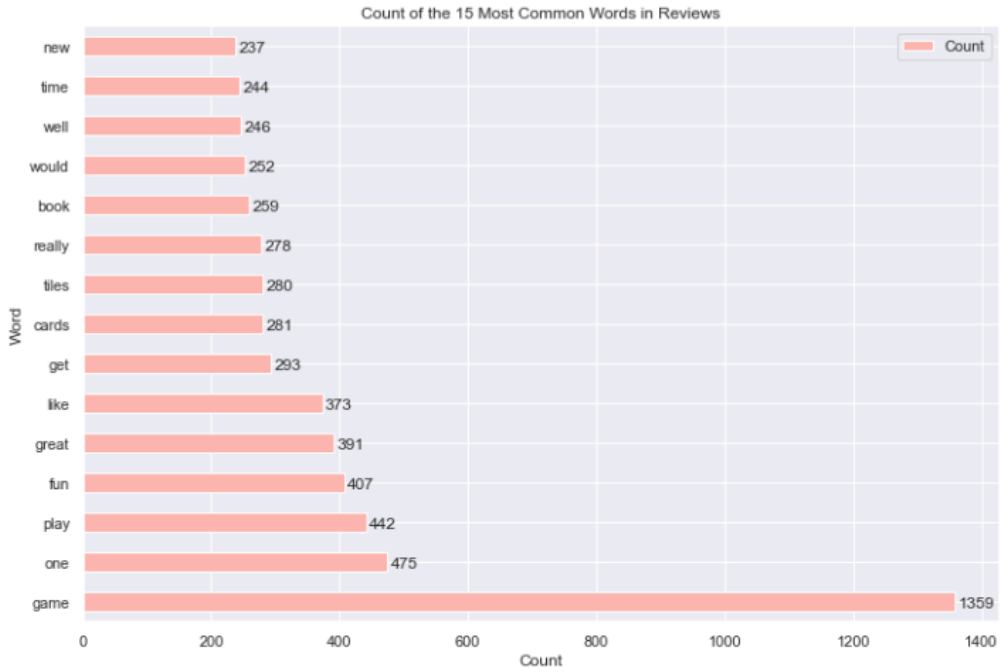
We used sentiment analysis methods to investigate the sentiments expressed in customer product reviews.

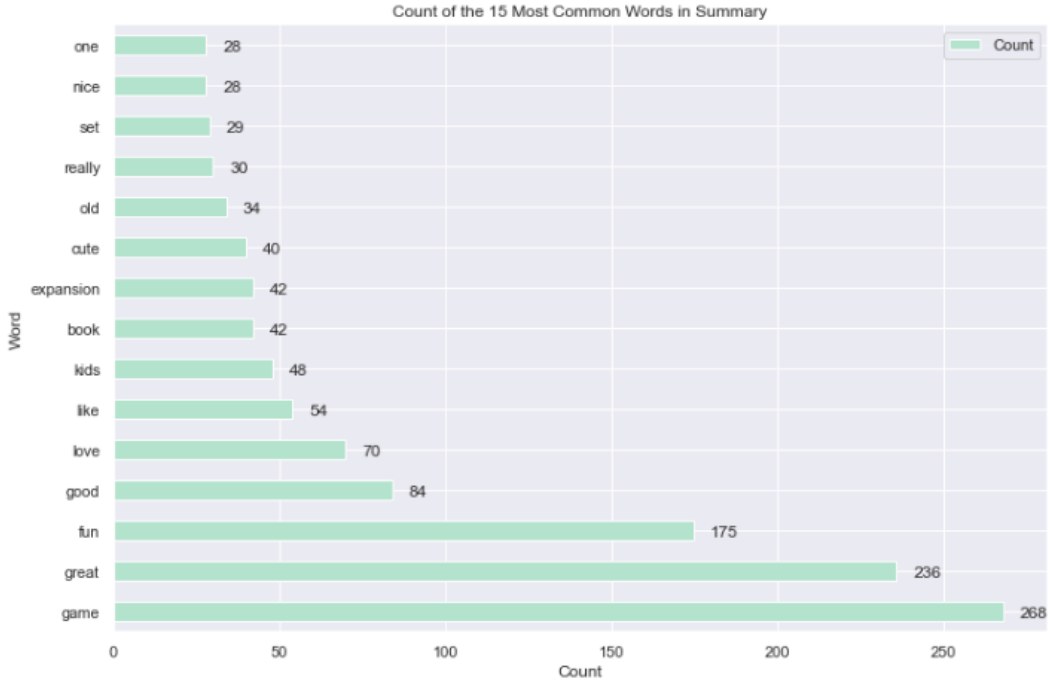
The text data from the reviews and review summaries was pre-processed, with the text being converted to lowercase, and punctuation removed. Any duplicated reviews or summaries were removed from the data, leaving 1,349 data points for analysis. Non-alphanumeric words, and stopwords (words such as ‘and’, or ‘the’), were removed from the text.

A word cloud, which visualises words with the most frequently occurring words appearing proportionally larger, was generated to display the words appearing in the reviews.


And, again, for the words appearing in the review summaries.

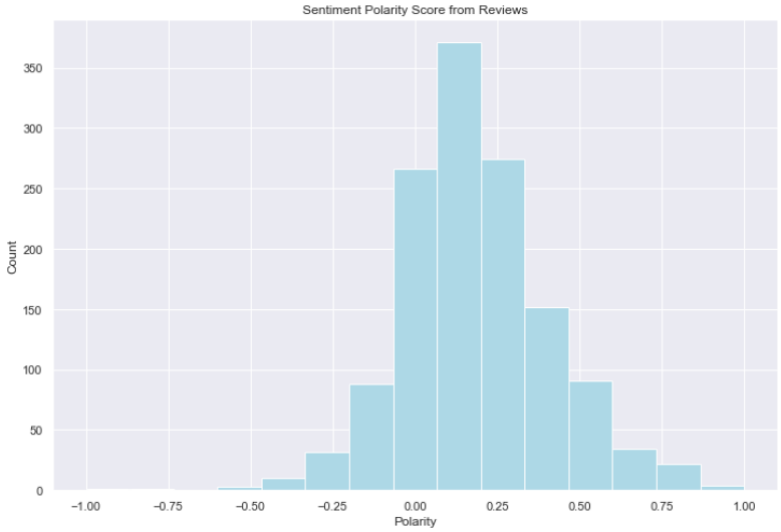
Of the most commonly occurring words, some are neutral and expected, such as ‘game’, or ‘book’. However, there are also some frequent positive words, such as ‘fun’ or ‘great’, which appear in the top 5 words from reviews, and the top 3 from summaries.

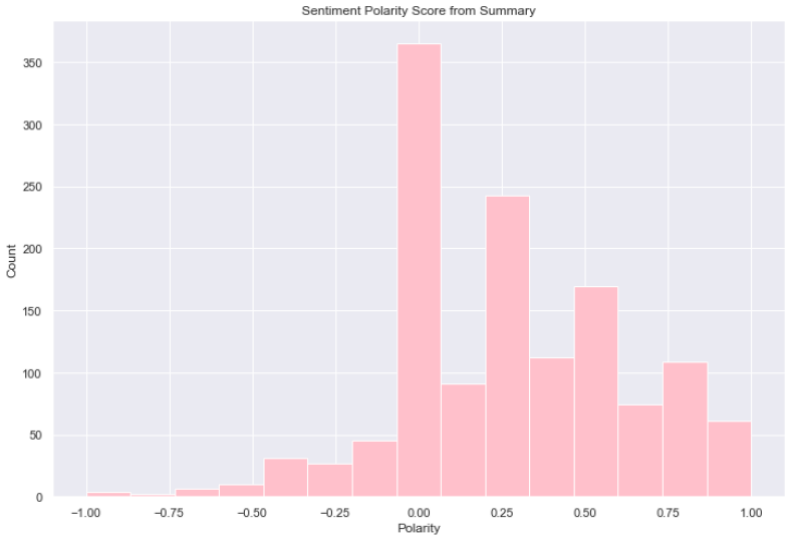



The sentiment polarity, which scores text data between -1 and 1, where -1 is highly negative sentiment and 1 is highly positive sentiment, was then calculated for each review (*mean=0.18*) and review summary (*mean=0.27*).

Histograms of the sentiment polarity indicate that the sentiment is generally slightly positive, with more strong positive scores than strong negative scores.




The top 20 most positive and most negative reviews and review summaries have been identified. Further analysis is recommended to investigate common themes across these top scoring reviews. For example, at a glance we notice that the word 'tool' appears multiple times in the top 20 positive reviews.

| review_polarity | review |
|---|---|
| 1.000000 | perfect |
| 1.000000 | my daughter loves her stickers awesome seller thank you |
| 1.000000 | perfect for tutoring my grandson in spelling |
| 0.880000 | the best part i see is the box what a wonderfully diverse and rounded set for the cost i am so happy and as the dm you know that if i am happy my players are happy |
| 0.816667 | great quality very cute and perfect for my toddler |
| 0.800000 | this is a great tool to have at hand when playing quiddler |
| 0.800000 | great seller happy with my purchase 5 starrrr |
| 0.800000 | prompt service and a great product |
| 0.800000 | a great creation tool it helps me concentrate |
| 0.800000 | husband seems happy with it |
| 0.800000 | great price arrived on time with no damage will be a great addition to my collection |
| 0.800000 | the pictures are great ive done one and gave it to a friend of mine who likes dragons |
| 0.800000 | bought this because i wanted it all these dd games are great |
| 0.800000 | great easter gift for kids |
| 0.800000 | my granddaughter loves these so happy to find peppa pig items for her |
| 0.800000 | great doll to go with the book animals cant wait to read book with the doll to the grandkids |
| 0.800000 | great accessory to use with the playing mat |
| 0.800000 | this is a great accessory to the starter set i would recommend this to anyone who owns the starter set |
| 0.800000 | these are great |
| 0.790000 | this is a great product i use it as a therapeutic tool and it has been very effective |

Note that sentiment polarity analysis does have some limitations. For example, the polarity score is not always accurate, particularly where 'negative' words are used in a positive context (for example, 'this game is the bomb!').

We have also identified the top 10 products by average review polarity. The highest-scoring product on average was product 11004 (*mean=0.44*).

A limitation of this approach to identifying the top products is that duplicate reviews and summaries have been excluded. Therefore, repeated positive reviews will be overlooked.

| | product | mean_review_polarity | number_of_reviews |
|---|---|---|---|
| **0** | 11004 | 0.437552 | 4 |
| **1** | 9119 | 0.397751 | 6 |
| **2** | 2139 | 0.385115 | 7 |
| **3** | 3158 | 0.383138 | 6 |
| **4** | 326 | 0.359691 | 7 |
| **5** | 515 | 0.353342 | 9 |
| **6** | 6310 | 0.350492 | 8 |
| **7** | 2457 | 0.347050 | 6 |
| **8** | 466 | 0.346270 | 3 |
| **9** | 11086 | 0.341155 | 6 |

**Summary:**

These insights from customer reviews can inform our marketing campaigns. For example, the most positive reviews for Turtle Games could be used in marketing materials, such as marketing emails.

We could also utilise common words, or words that appear frequently in highly positive reviews, in marketing materials (such as 'fun', 'great', or 'tool', which are words that appear to resonate for current customers).

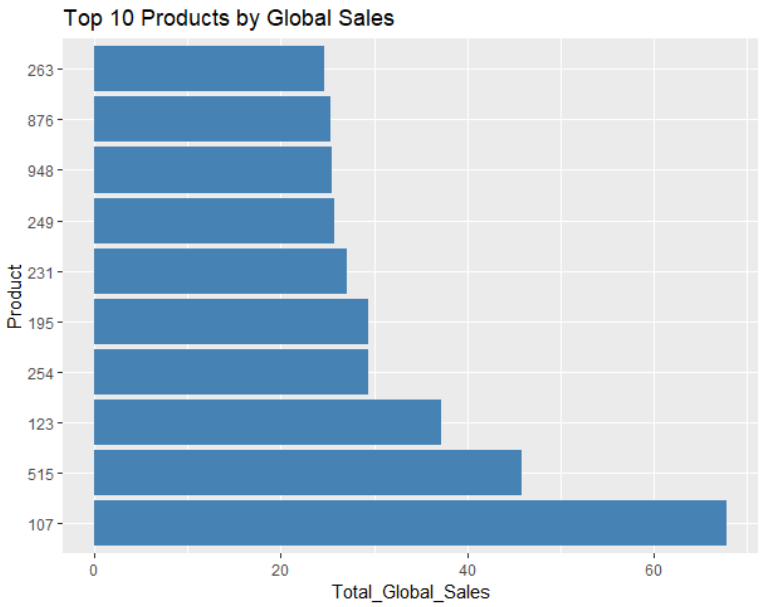
In addition, we could focus on marketing the products that have a higher average review polarity (such as 11004, 9119, 2139, 3158), as we suggest that this indicates that these products are likely to have a positive reception.

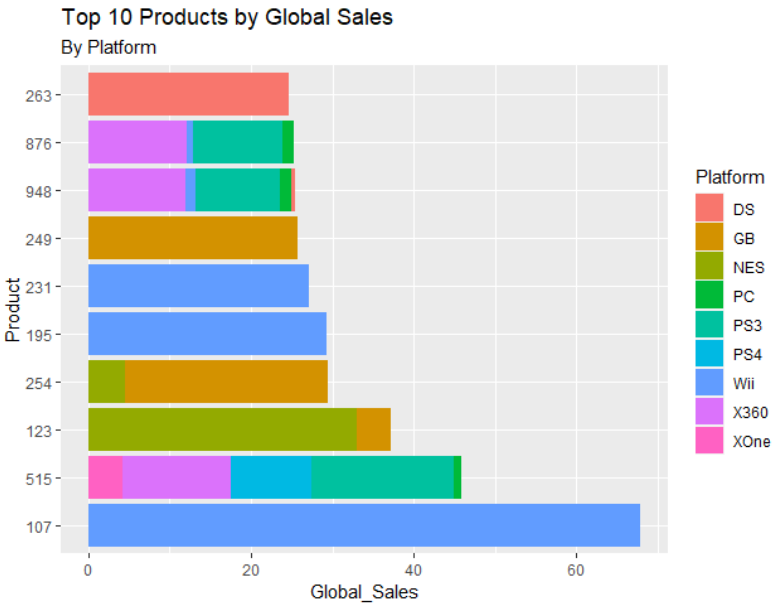
# Sales Insights

### *Product Sales*

The top three products by global sales are 107, 515, and 123 (of which, 107 and 123 appear in the top three products by NA sales, and 107 and 515 appear in the top three products by EU sales).

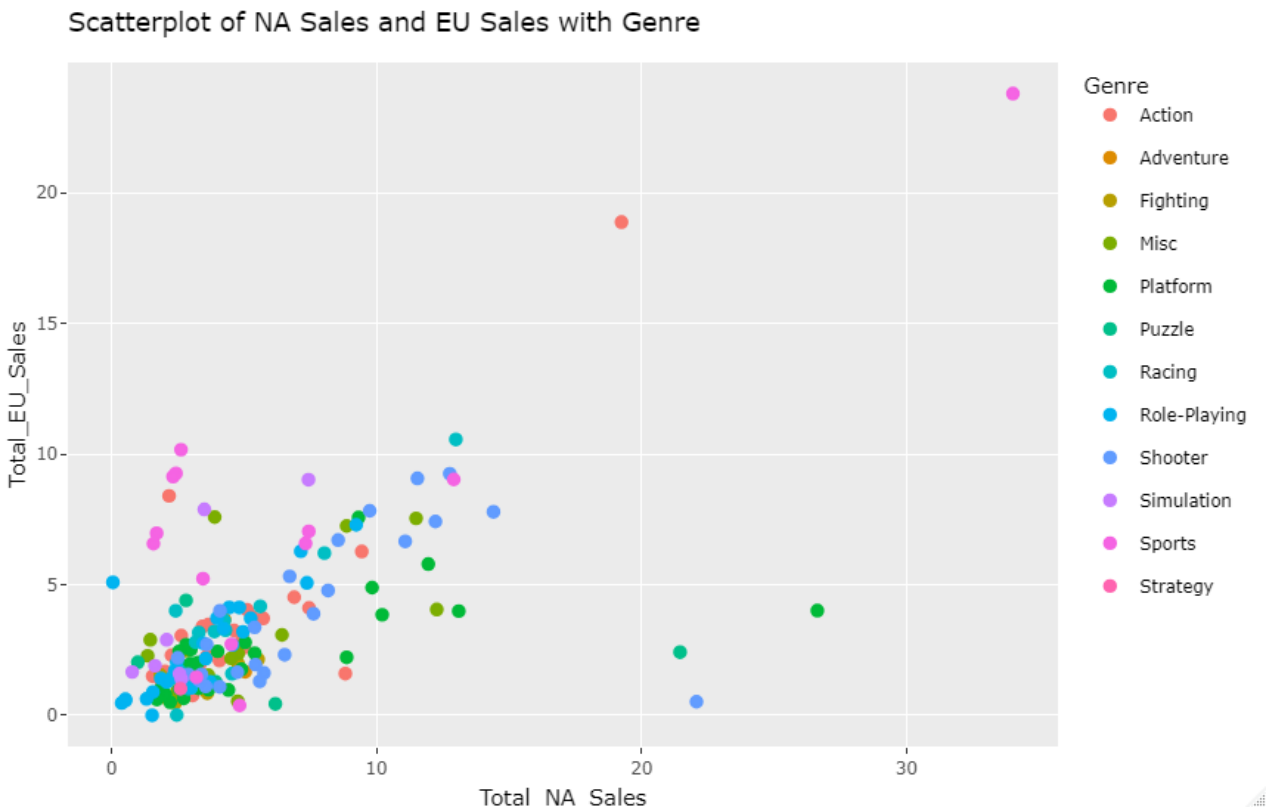
We note that most of the top 10 products by global sales appear to be on Nintendo platforms (Wii, NES, GB, DS).




The largest customer base is NA, making up 885.62 out of 1877.81 global sales.

We suggest that different genres may be preferred in NA and EU. From the plot below, it looks as though NA may prefer platform games, while EU may prefer sports games.



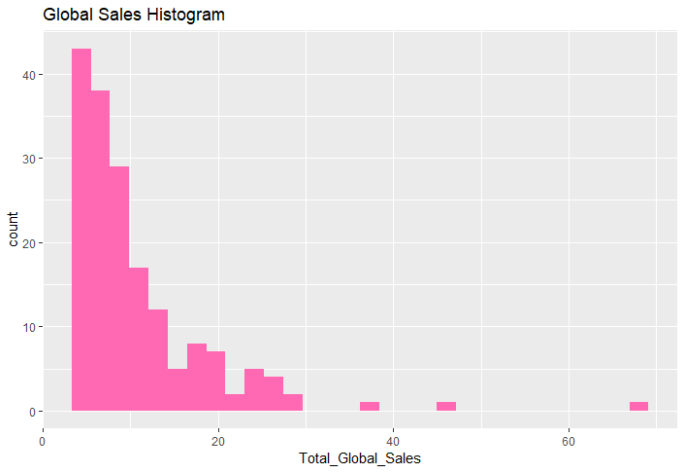
**Summary:** The best-selling product across all three areas (global, North America, and Europe) is product 107, which appears to be a Wii Sports game. For future product development, we may wish to consider the preferences (such as platform games) of our largest customer base, North America.

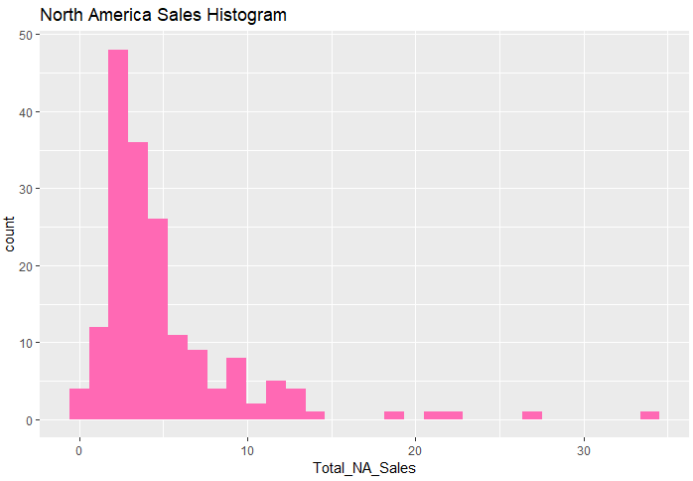
### *Reliability of Data*

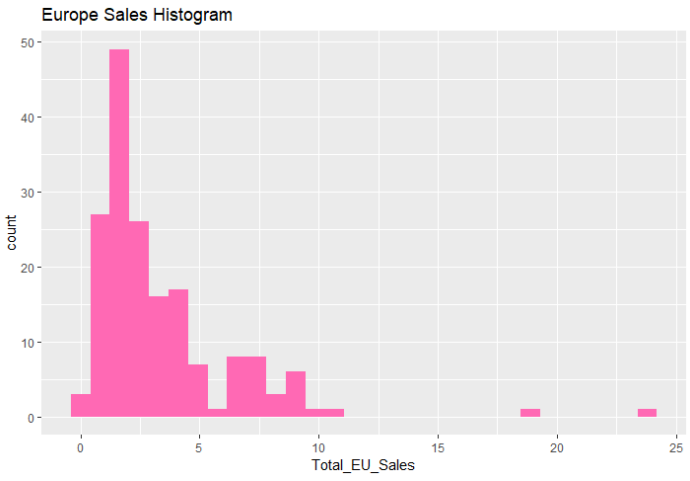
We tested for skewness in the sales data. The skewness values of our sales data were greater than zero, indicating a positive skew (*global sales = 3.07, NA sales = 3.05, EU sales = 2.89*).

We also tested for kurtosis in the sales data, a method of comparison to the normal distribution. Our kurtosis values were much higher than three, indicating distributions with more extreme outliers than the normal distribution (*global sales = 17.79, NA sales = 15.60, EU sales = 16.23*).

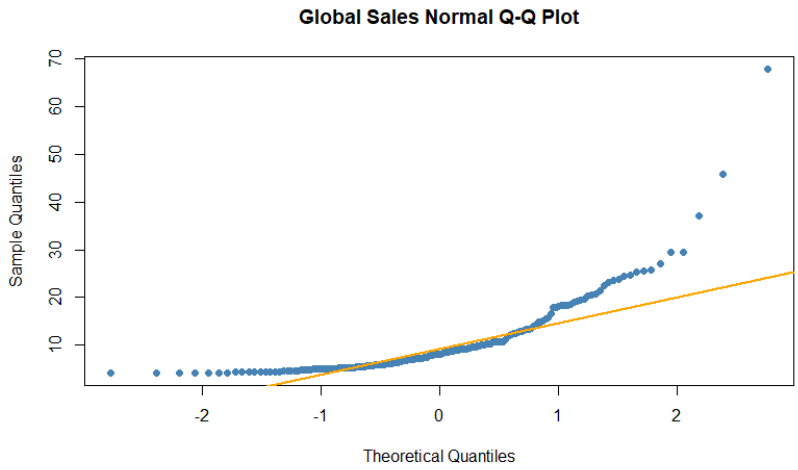
A Shapiro-Wilk test confirms that the data is not normally distributed ($p < 0.01$ *for global sales, NA sales, and EU sales*). This can be observed in the histograms below.

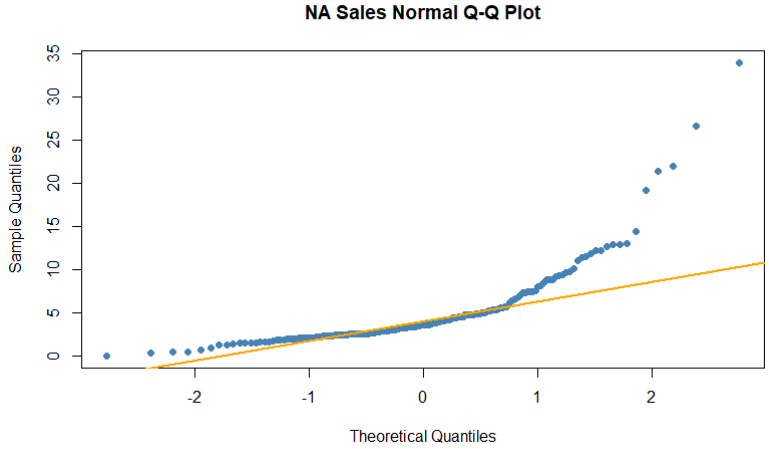


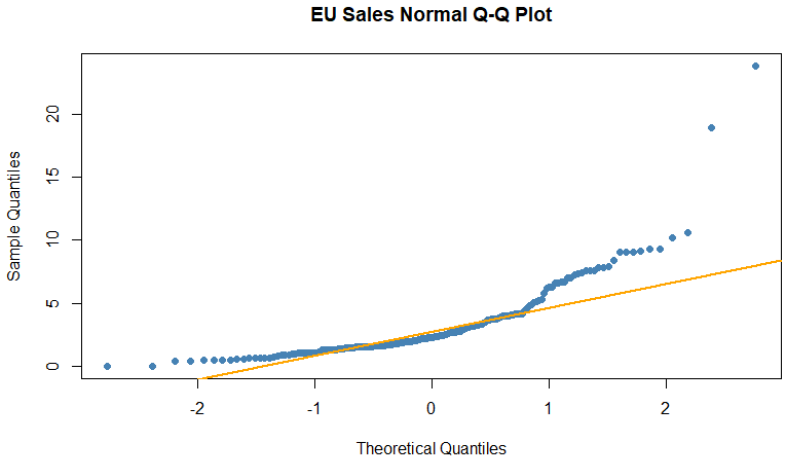




The Q-Q plots also display a severe departure from the normal distribution. While this alone does not break any assumptions for analysis, highly skewed data may be more likely to see non-normal distributions in residuals when used for linear regression modelling.



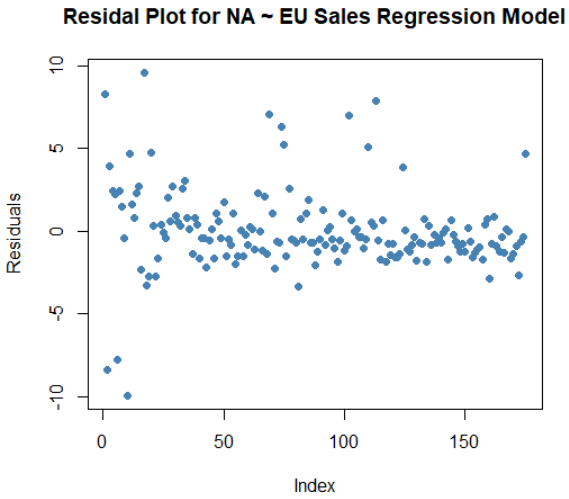



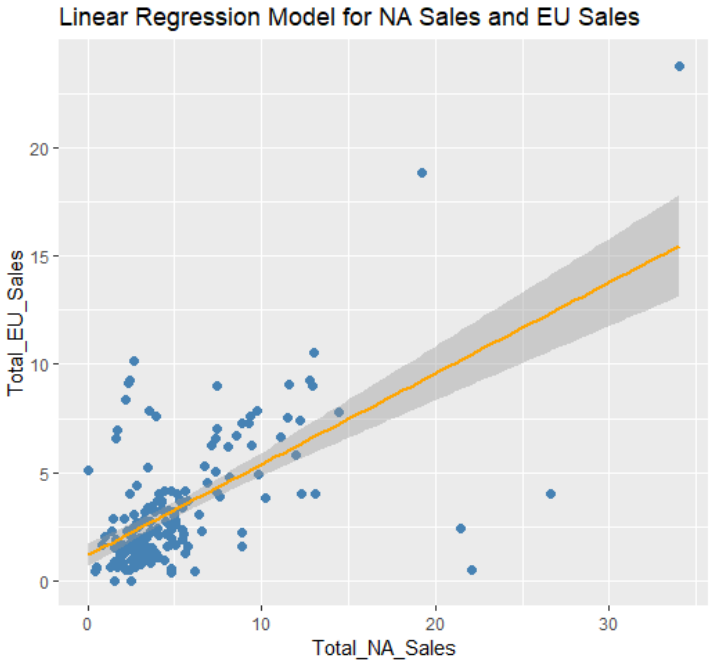

**Summary:** The presence of extreme outliers, along with the level of skew and non-normality displayed in the data, may mean that the residuals in a linear model based on this sales data would also not follow a normal distribution. This would break an assumption of linear modelling, and may mean that the resulting linear model is not reliable. Further analysis is recommended to consider transforming the sales data into a more normal format, where resulting residuals may be less likely to deviate from the normal distribution.

### *Relationship Between Sales*

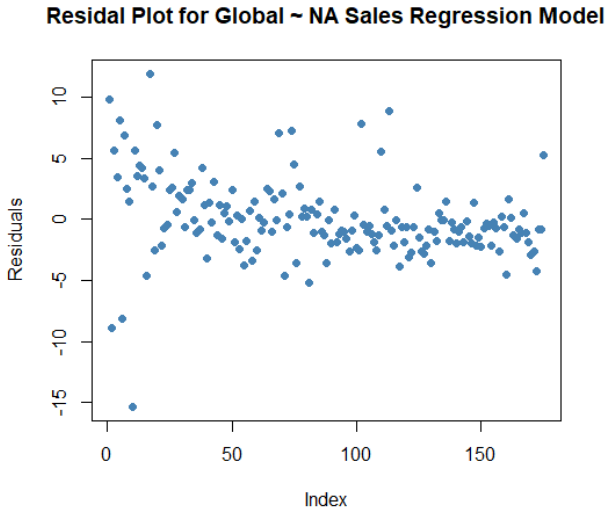
We used linear regression methods to investigate relationships between the sales variables.

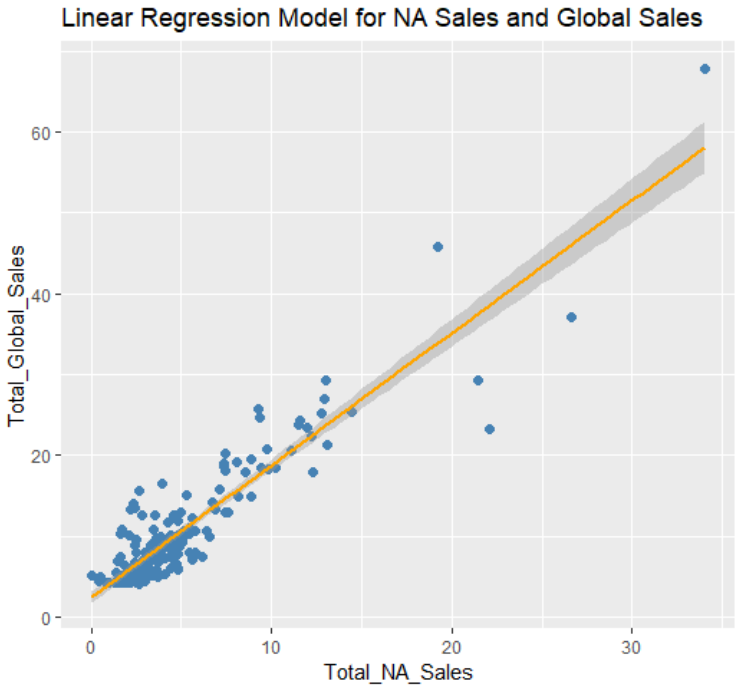
We found that NA sales and EU sales were positively correlated (*adj $R^2$ = 0.382*), though there is some scatter visible in the plot. The residual plot appears to show random scatter, indicating that the linear model is an appropriate fit.



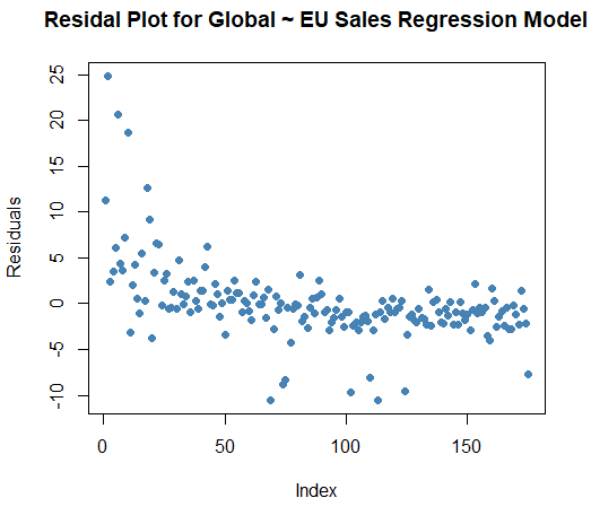



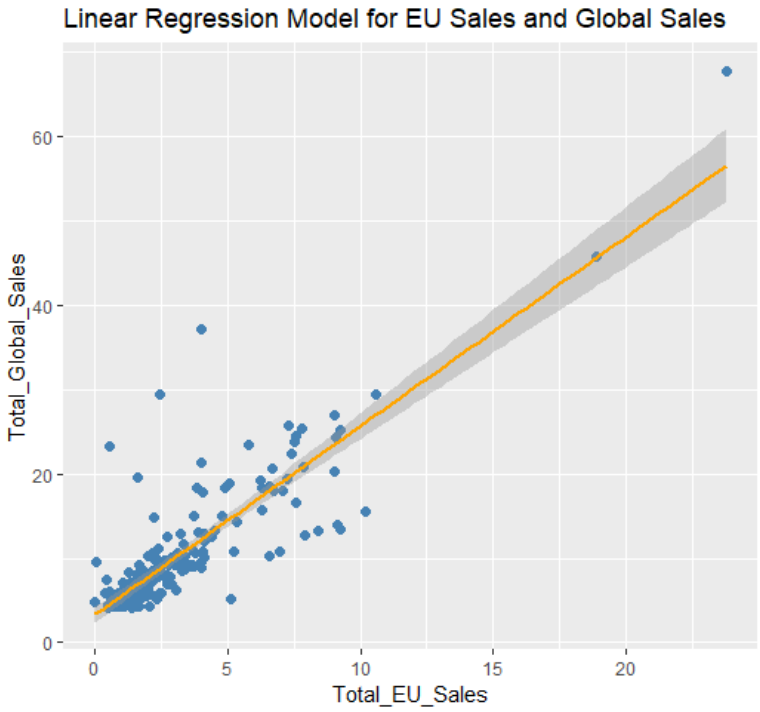

NA sales and global sales had a strong positive correlation (*adj $R^2$ = 0.839*). Once again, the residual plot mostly appears to display random scatter.



EU sales and global sales are also positively correlated (*adj* $R^2 = 0.719$). Here, however, we suspect that the residual plot may not show random scatter, but a slight curve. This could indicate that the linear model is not a good fit for this data.




We then tried a multiple linear regression model, using NA sales and EU sales to predict global sales (*adj* $R^2 = 0.966$).

To check the model accuracy, we provided the model with the values of the NA and EU sales for products 107, 326, 3267, 6815, 2877, and then compared the predicted values with the observed values.

| Total_Global_Sales | Predicted_Global_Sales |
|---|---|
| 67.85 | 68.033940 |
| 23.21 | 26.648166 |
| 17.96 | 18.770062 |
| 13.02 | 14.335029 |
| 4.32 | 4.908353 |

The model appears to be fairly accurate, though could be improved.

**Summary:** NA sales and EU sales are positively correlated with global sales and with each other. We can predict global sales with some accuracy from the NA and EU sales. This may be useful for calculating projected sales when a product is released at different times around the world.

## Summary

Key results and recommendations for stakeholders are presented in full at the start of this report.

Further analysis is recommended, in particular an analysis of any common themes in the most positive and most negative customer reviews, and a transformation of the sales data to reduce potential non-normality in the residuals when this data is used for linear modelling.