1. Background/context of the business

Turtle Games is a global game manufacturer and retailer, offering an array of products like books, video games, board games, and toys. The firm's main objective is to boost sales by utilising customer trends. Key areas of interest include understanding how customers gather loyalty points, identifying market segments within the customer base, leveraging social data for marketing, evaluating product-wise impact on sales, checking data reliability, and exploring relationships between North American, European, and global sales. The aim of this report is to provide actionable insights that can help inform sales and marketing strategies to improve Turtle Games' overall sales performance.

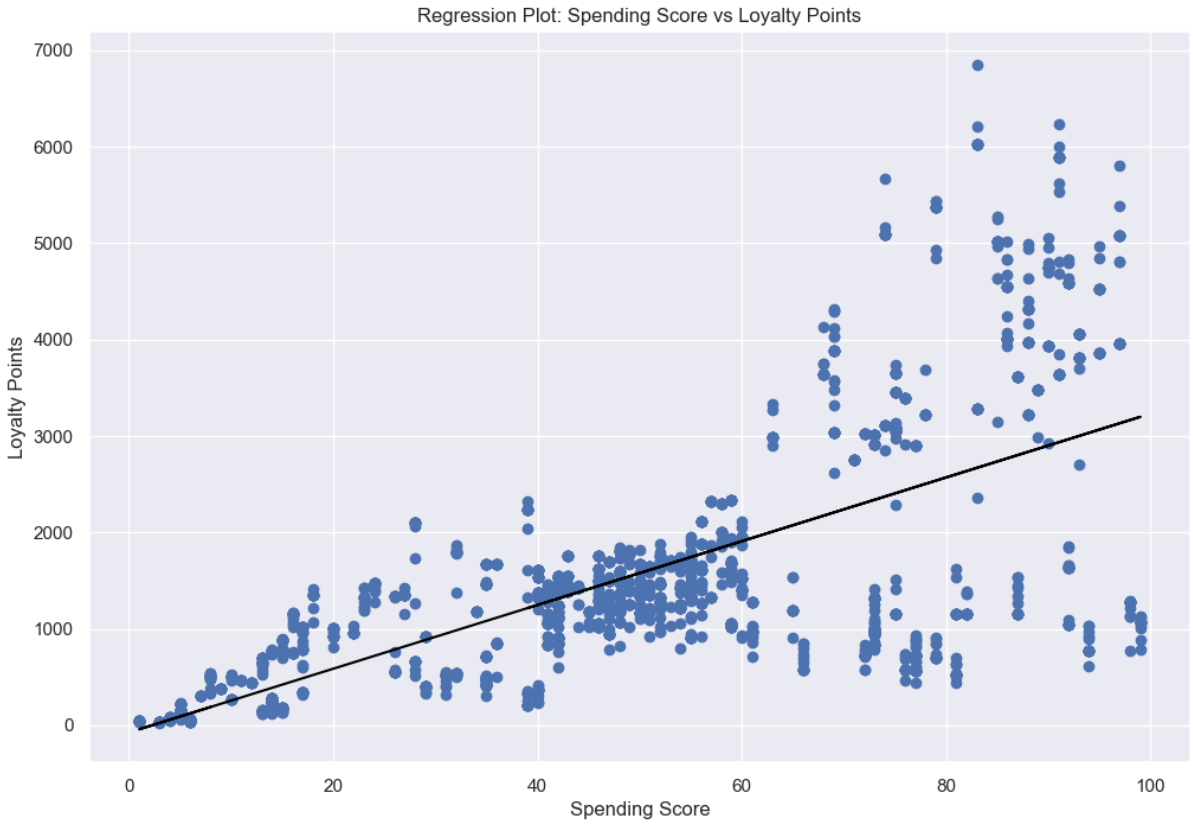
2. Analytical approach

- Linear and multilinear regression, via the statsmodels library, were used to evaluate relationships between loyalty points and factors such as age, remuneration, and spending scores. This was done because these techniques effectively identify relationships and inform predictive models. The Ordinary Least Squares (OLS) model was chosen for simplicity and efficiency, aiming to understand if these factors could predict loyalty points, to give Turtle Games a better understanding of how users accumulate said points.
- Using the K-means clustering method, five customer groups were determined based on remuneration and spending score. The optimal cluster number was established using the Elbow and Silhouette methods. These methods were chosen due to their ability to effectively determine the optimal number of clusters, avoiding overfitting and facilitating efficient data analysis, which will allow Turtle Games to accuratley target specific market segments.
- The sentiment of reviews and summaries was analysed using functions to determine polarity and sentiment. Polarity and sentiment scores were plotted using histograms for both columns. The top 20 positive and negative reviews and summaries were also identified. These steps were undertaken to gain insights for the marketing department, thereby enabling the creation of more informed future campaigns.
    - When cleaning the data for NLP analysis, the apply() function was used to convert 'review' and 'summary' text to lowercase and concatenate words. The str.replace() function removed punctuation, and drop_duplicates() eliminated duplicate entries. These preprocessing steps were performed to increase the accuracy of the NLP analysis and prevent errors.
    - Alphanumeric characters were removed from 'review' and 'summary' using a list comprehension. Then, the stopwords.words function was used to identify and filter out stopwords, this was done to further increase the NLP analysis.
- Scatterplots, histograms, and boxplots were created using the qplot function to gain insights into a sales dataset. The scatterplots highlighted relationships between North American, European, and Global sales. Histograms showcased the distribution of these sales, while boxplots displayed differences in sales by platform. This graphical exploration was undertaken to better understand the sales data prior to further analysis.
    - Using R's group_by(), summarise_all(), and as.data.frame() functions, sales data was grouped by 'Product'. This aggregation simplifies the dataset making it more useful for regression analysis.

- Influence of individual products on sales across North American, European, and Global markets was investigated using scatter plots, histograms, and box plots. The intention was to uncover individual performance of products across the different regions.
- Q-Q plots, Shapiro-Wilk tests, skewness and kurtosis tests were used to determine the trustworthiness of the data, the aim of this was to make insights drawn from said data more reliable, however ended up finding potential flaws instead.
- Simple linear regression and multilinear regression models were used to gain insights into the relationships between North American, European, and Global sales.

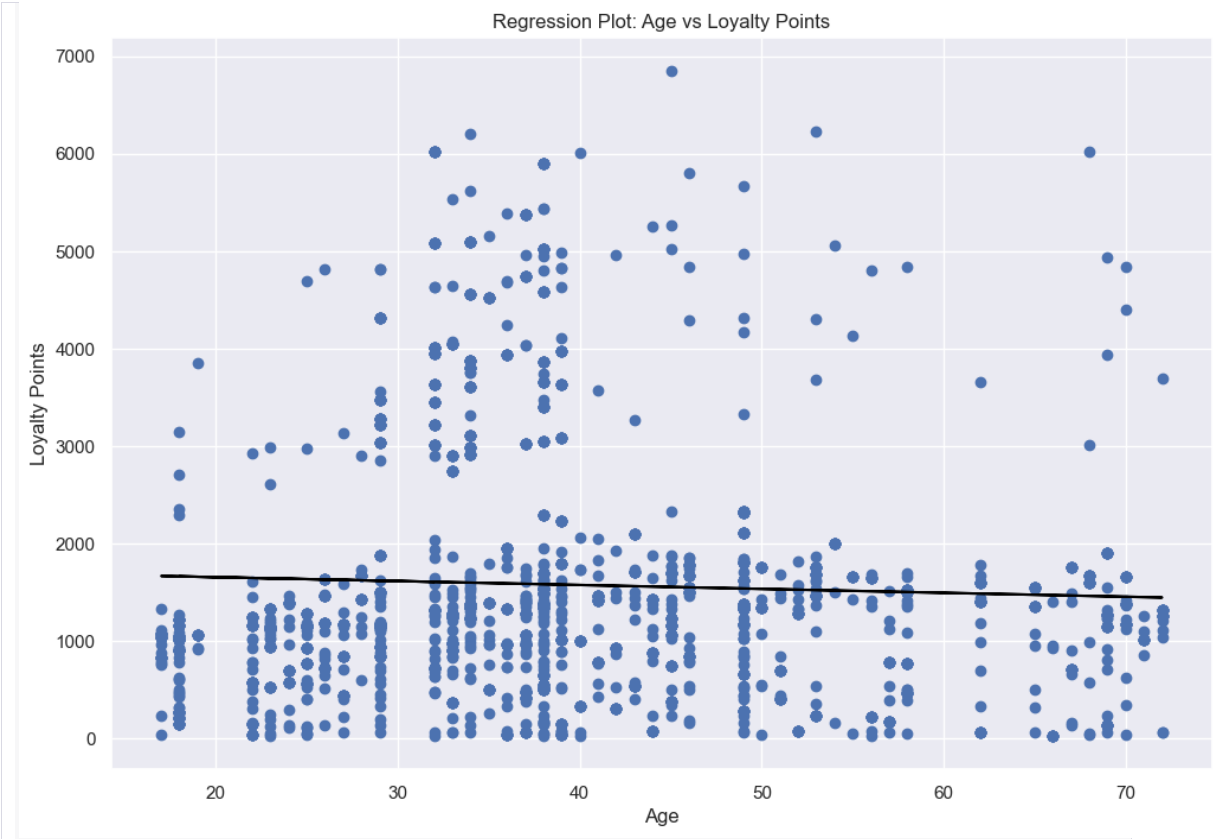
3. Visualisation and insights

- **how customers accumulate loyalty points**



- There is a positive correlation between spending score and loyalty points, indicating that more spending leads to more points.

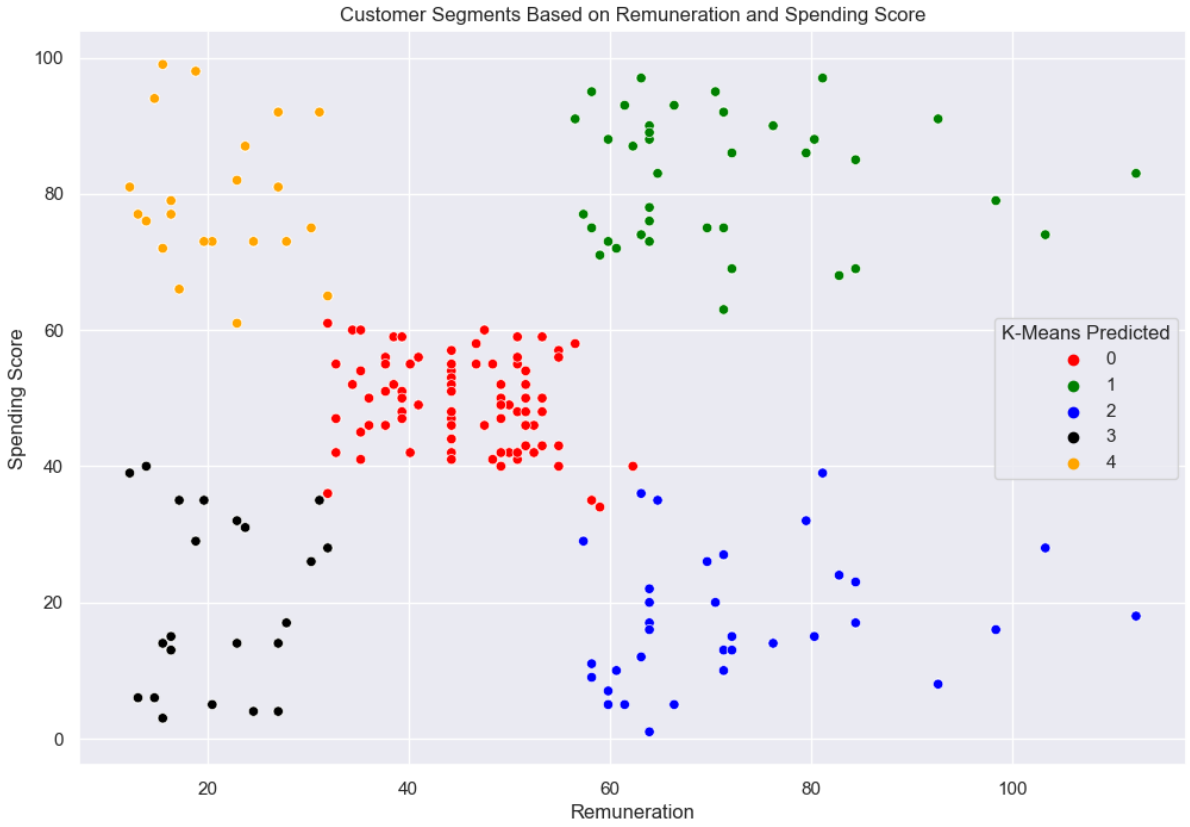
- Similarly, a positive relationship is evident between remuneration and loyalty points, suggesting higher income leads to more points.



- Age doesn't significantly influence loyalty points, showing equal engagement across age groups.

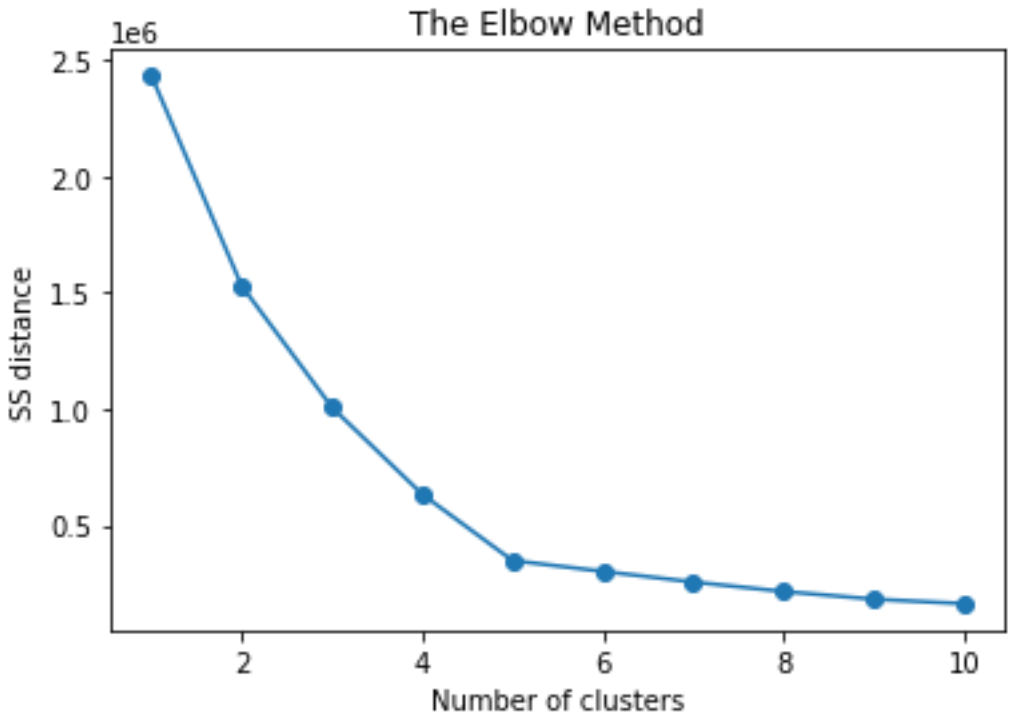
These visualisations were chosen as scatter plots are effective in depicting the relationship between two variables. The line of best fit was used as it provides a clear summary of the trend of the relationship between the variables.

- ***how groups within the customer base can be used to target specific market segments***



Utilising k-means clustering on customer remuneration and spending scores, five distinct market segments were identified. These include average earners with average spend (774 customers), low earners with low spend (356 customers), high earners with low spend (330 customers), high earners with high spend (271 customers), and low earners with high spend (269 customers).

The optimal cluster number was confirmed by the Elbow and Silhouette methods

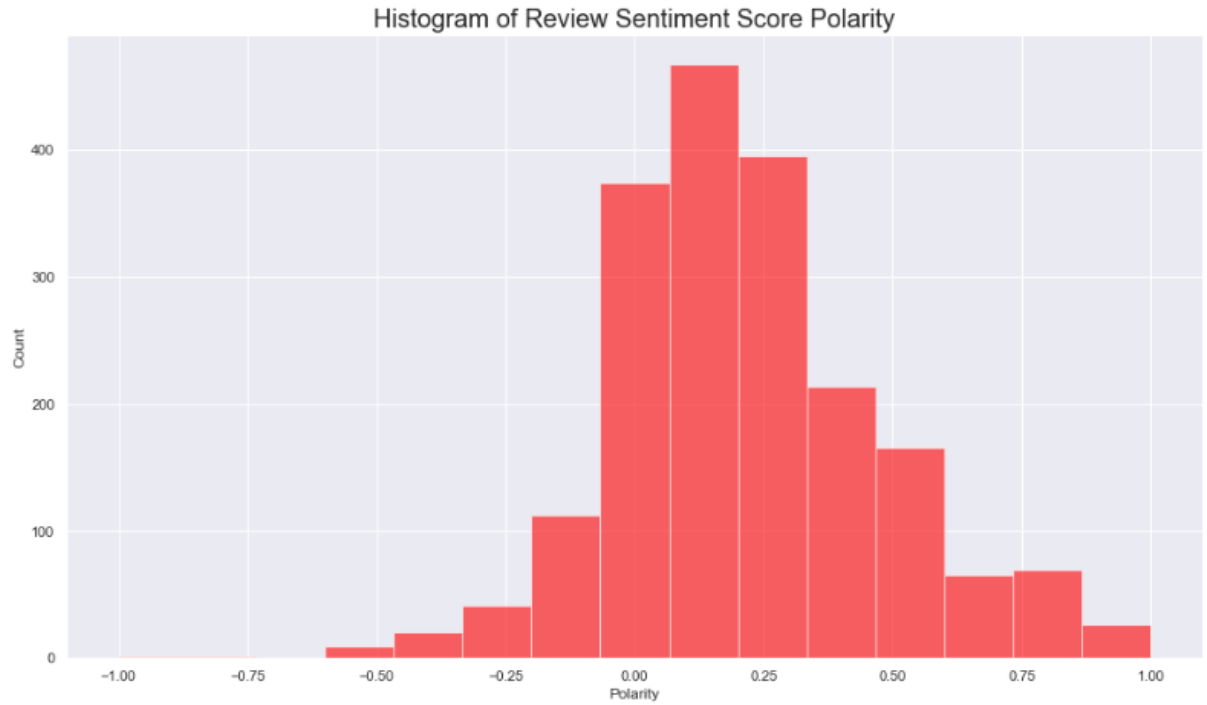


- ***how social data (e.g. customer reviews) can be used to inform marketing campaigns***

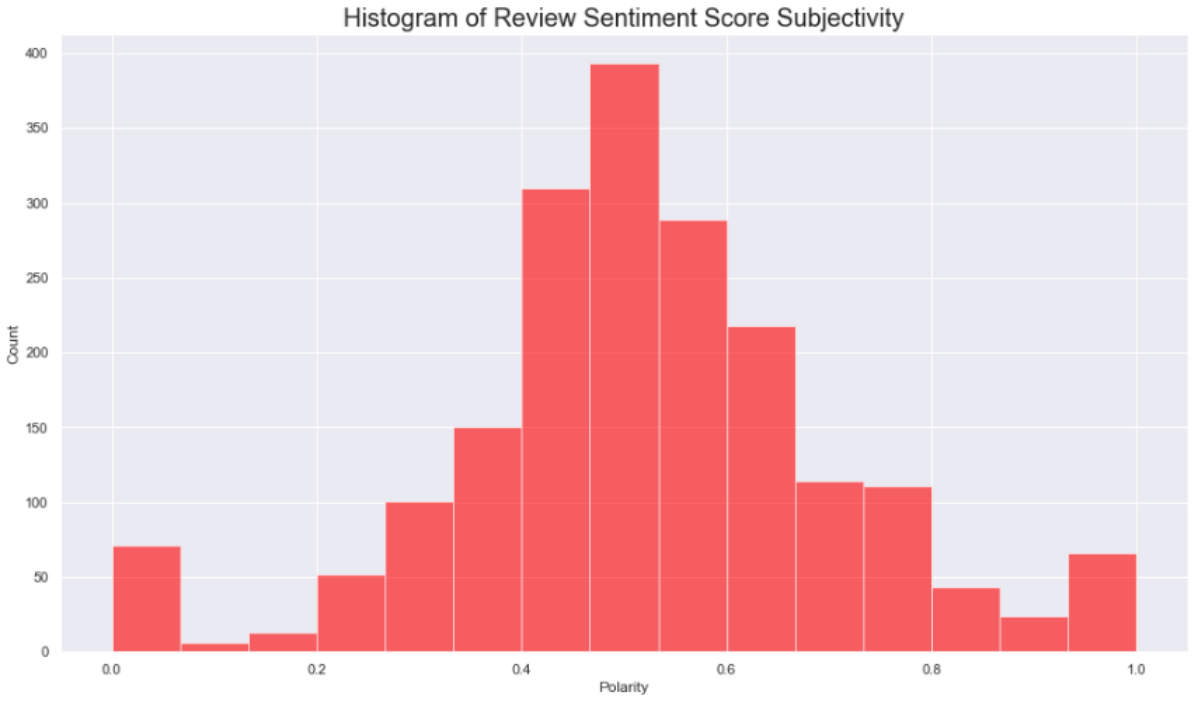
The most common words used in reviews:

Word	Frequency
game	1990
great	875
fun	770
one	568
play	528
like	468
stars	464
love	416
good	381
five	362
really	349
get	333
tiles	317
book	316
time	309

The Wordcloud was used to visualise the frequency of the most used words in Turtle Game’s customer reviews.

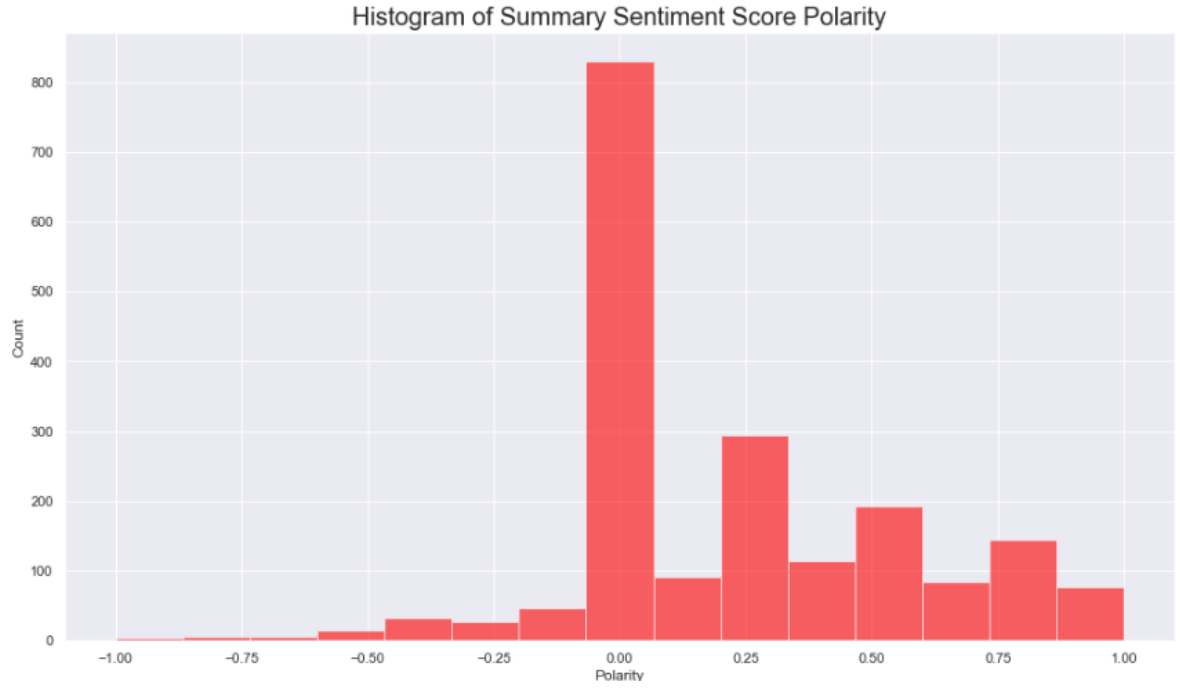


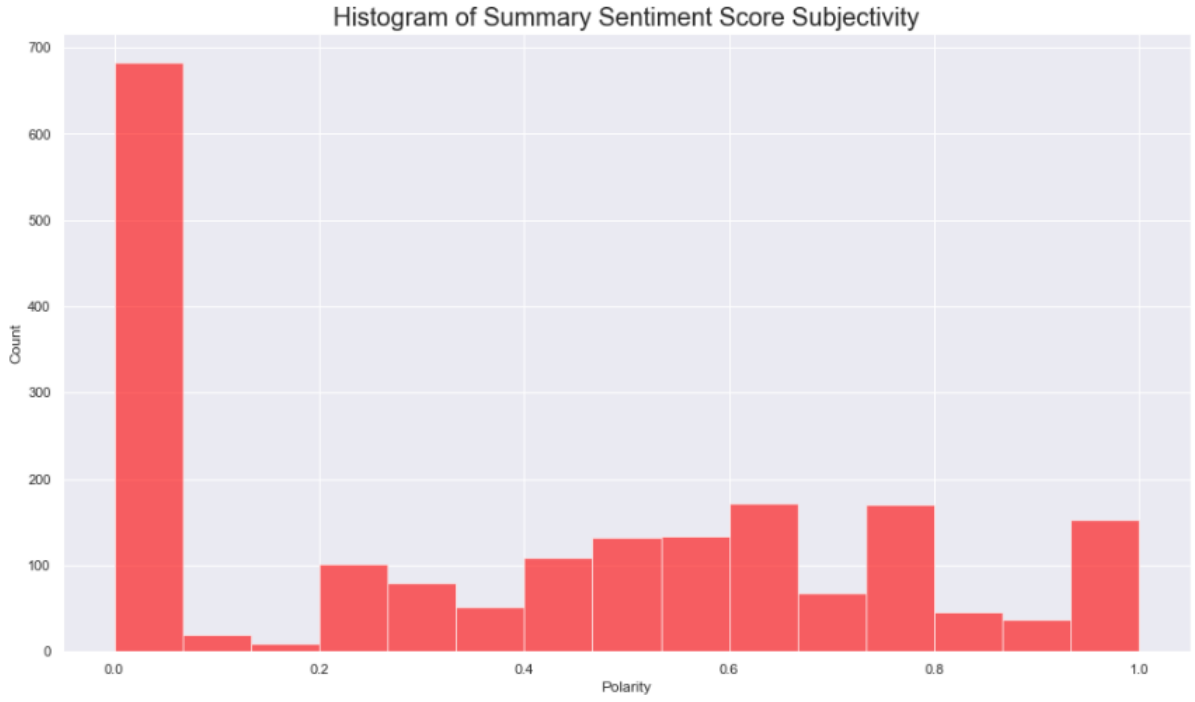
The sentiment score for reviews was normally distributed, with most reviews having neutral to slightly positive sentiment scores (-0.07 to 0.2). Strongly negative or extremely positive sentiments are infrequent, suggesting minimal polarisation.



Most reviews had a subjectivity score of 0.33 to 0.6 implying that Turtle Game's reviews are mostly subjective rather than factual or objective statements.

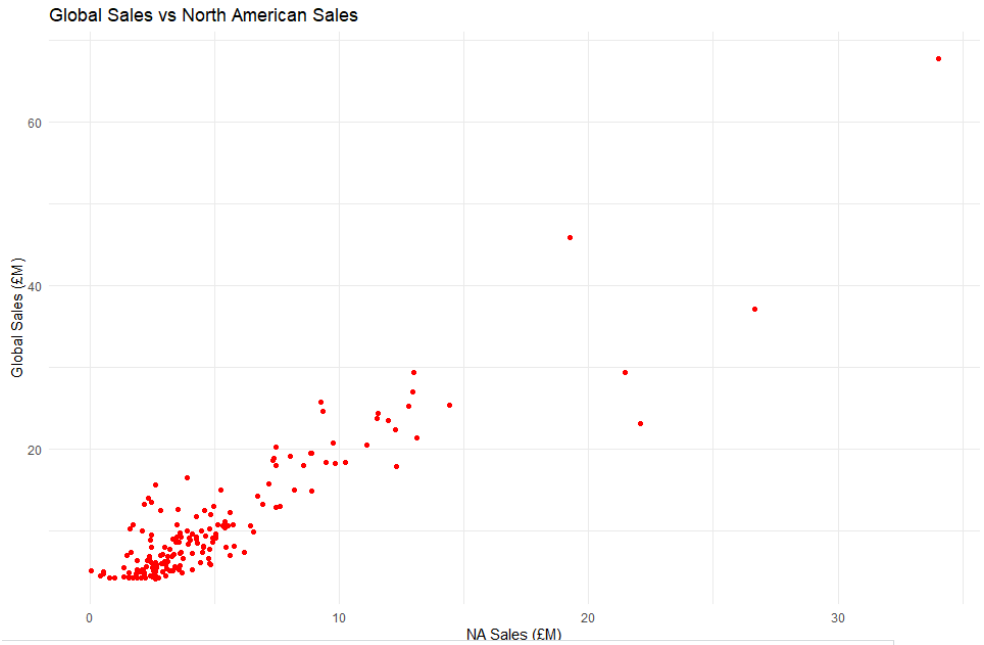
Most review summaries are neutral and objective, showing considerable uniformity. This lack of distinctiveness makes them less informative than the full reviews for directing marketing decisions.

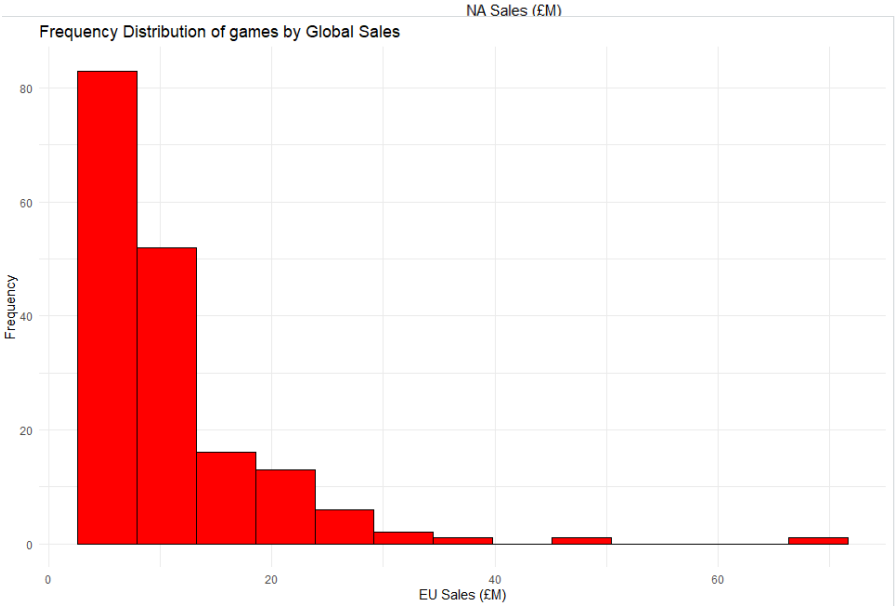




- ***the impact that each product has on sales***

Turtle Games offers 175 different products, with each product averaging £10.73M in global sales. The top seller is product 107, with sales of £67.85M, accounting for 3.6% of global sales. This figure, alongside the charts below show that Turtle Game's income is extremely diversified with no individual product/game having an overwhelming contribution to the total sales.





- ***how reliable the data is (e.g. normal distribution, skewness, or kurtosis)***

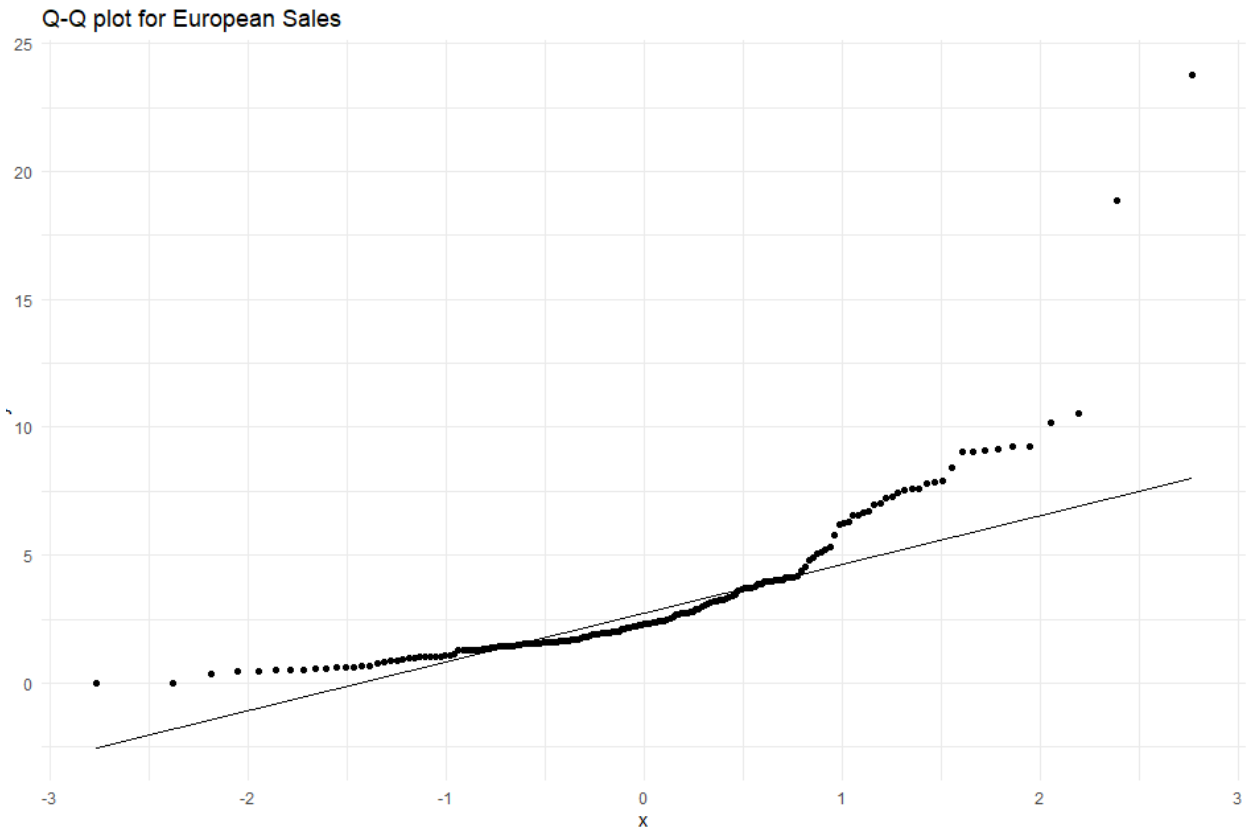
The sales data reliability is questionable as it doesn't follow a normal distribution. Significant right-skewness and high kurtosis values suggest the presence of outliers or extreme high values, potentially impacting the statistical analyses that presume normality (simple linear and multilinear regression. The Q-Q plots, Shapiro-Wilk tests, and skewness and kurtosis measures consistently confirm these findings.

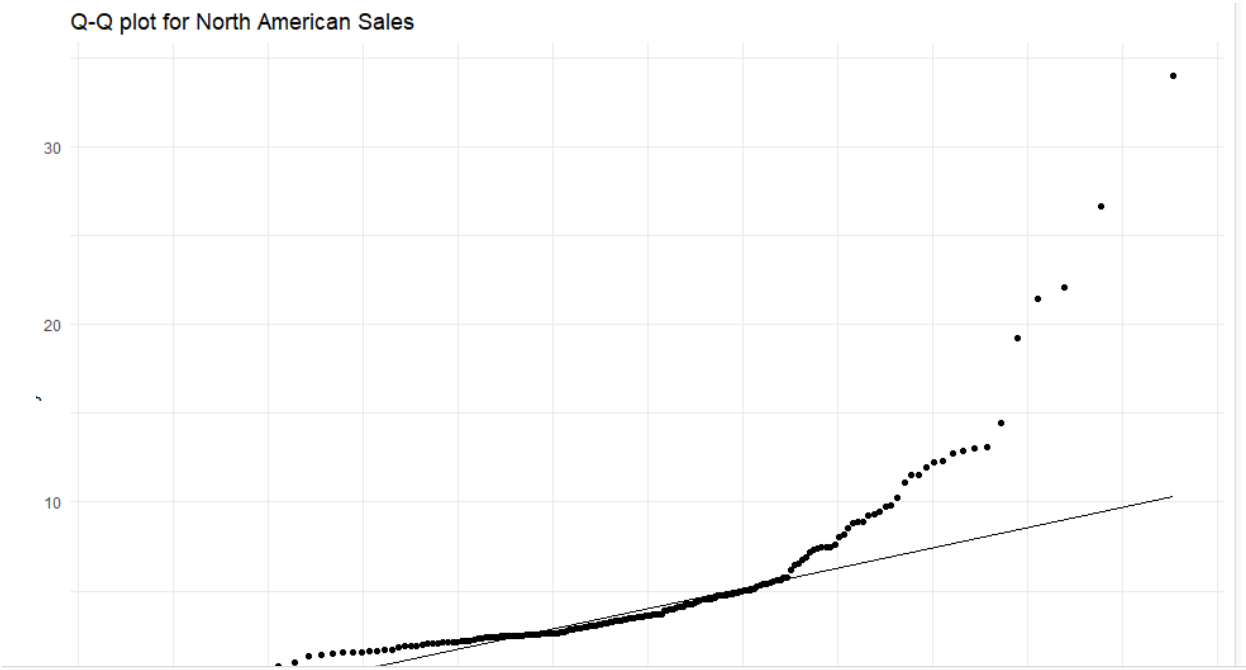
Skewness + Kurtosis

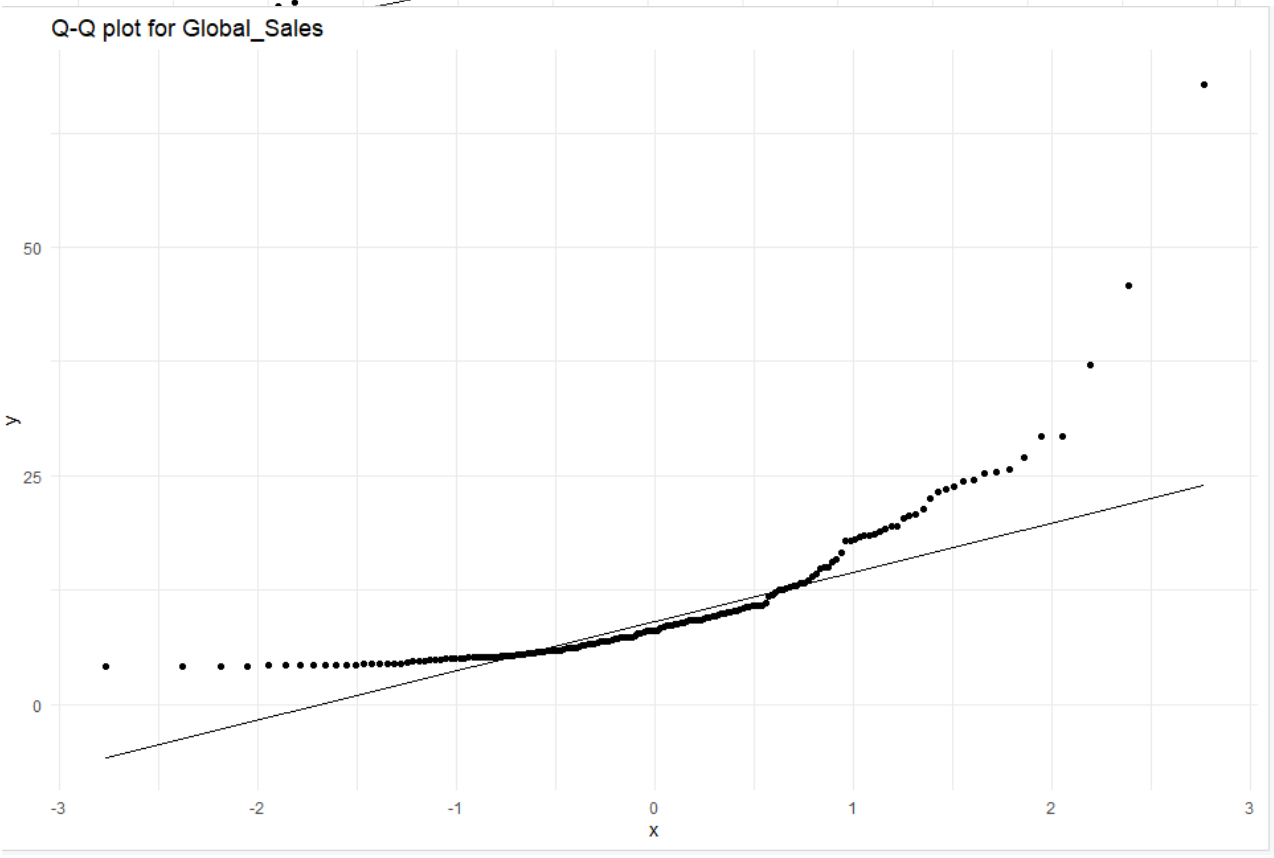
	Skewness	Kurtosis
Global Sales	3.066769	17.79072
European Sales	2.886029	16.22554
North American Sales	3.048198	15.6026

Shapiro-Wilk test

	W Value	P Value
Global Sales	0.70955	< 2.2e-16
European Sales	0.74058	2.987e-16
North American Sales	0.69813	< 2.2e-16

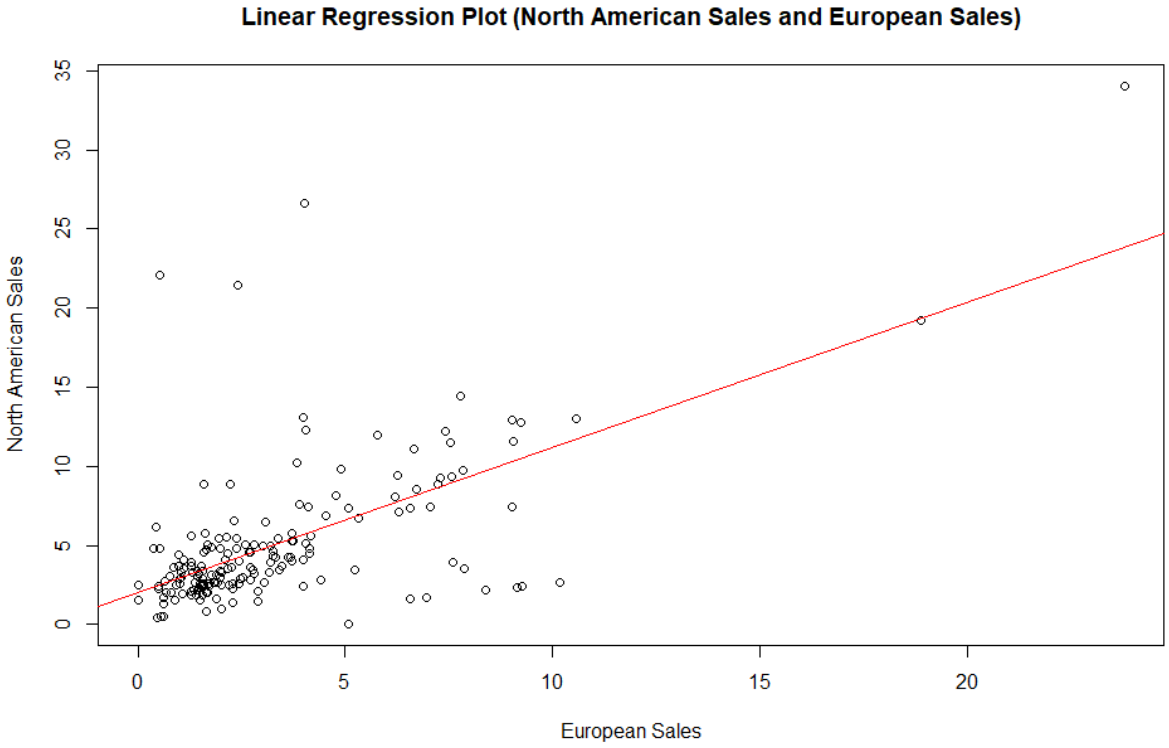


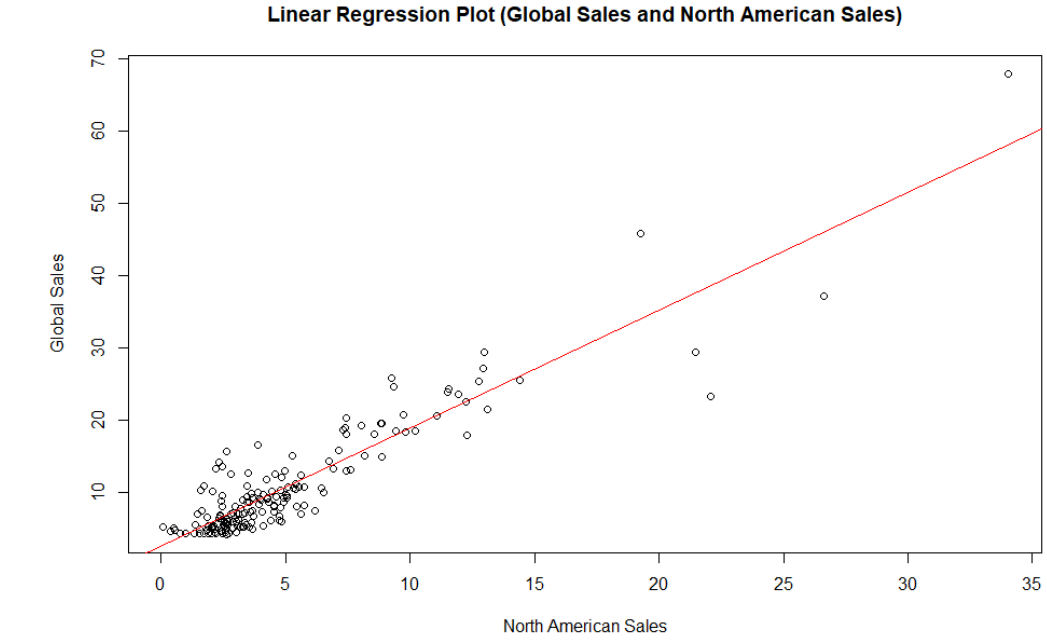

Q-Q plot for North American Sales
30
20
10

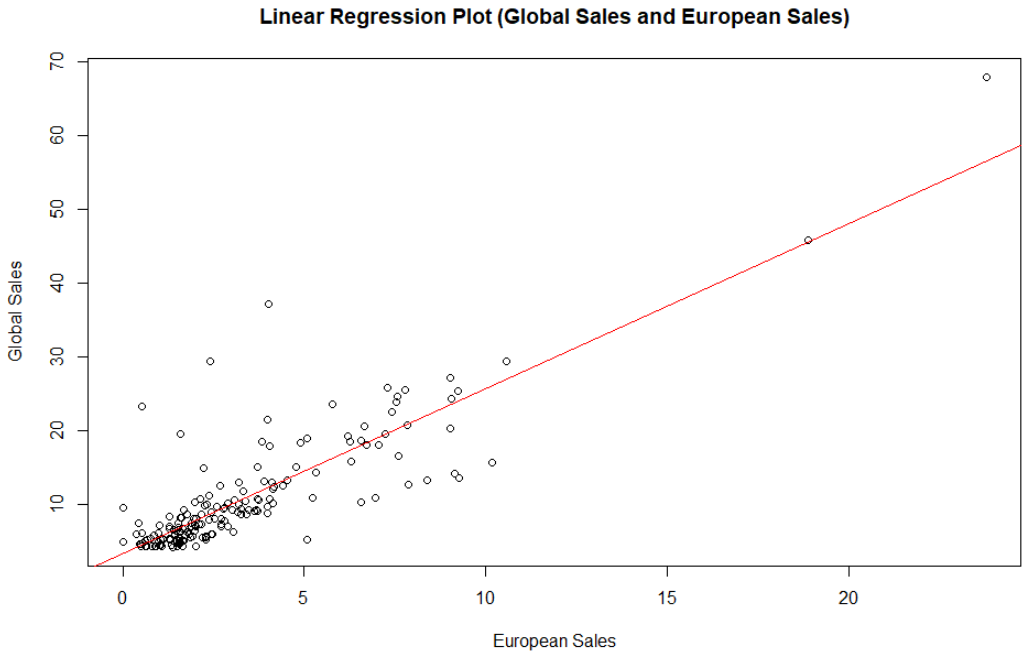

Q-Q plot for Global_Sales
50
25
0
y
-3
-2
-1
0
1
2
3
x

- ***what the relationship(s) is/are (if any) between North American, European, and global sales?***

There is a strong positive correlation between North American, European, and Global game sales. Using simple linear regression we can assume that an increase in sales by 1 million pounds in Europe and North America corresponds to respective increases in Global sales by roughly 2.237 and 1.635 million pounds. Furthermore, the multilinear regression model suggests that about 96.68% of Global sales variability is explained by North American and European sales. Both these markets are significant predictors of Global sales.




Linear Regression Plot (Global Sales and North American Sales)
Global Sales
North American Sales
0
5
10
15
20
25
30
35
10
20
30
40
50
60
70


Linear Regression Plot (Global Sales and European Sales)
Global Sales
European Sales
0
5
10
15
20
10
20
30
40
50
60
70

4. Patterns and predictions

In the multilinear regression analysis of Turtle Games' sales data, North American and European sales emerged as statistically significant predictors of Global sales. The regression coefficients for North American and European sales were 1.13040 and 1.19992 respectively. This implies that for every one million pound increase in North American sales, we'd expect an increase of about 1.13 million pounds in Global sales, assuming European sales are constant. Similarly, for each million pound increase in European sales, we'd anticipate an increase of 1.19 million pounds in Global sales, assuming North American sales are constant.

This MLR model was used to predict global sales for several scenarios, which can be found in the table below. The predictions align closely with the observed values, demonstrating the model's accuracy.

North American Sales	European Sales	Global Sales (Predicted)	Global Sales (Observed)
34.02	23.80	68.06	67.85
3.93	1.56	7.36	6.04
2.73	0.65	4.91	4.32
2.26	0.97	4.76	3.53
22.08	0.52	26.63	23.21

The model's R-squared value of 0.9668 indicates that it explains about 96.68% of the variability in global sales, demonstrating its effectiveness.

```
Call:
lm(formula = Global_Sales ~ NA_Sales + EU_Sales, data = sales_grouped2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4156 -1.0112 -0.3344  0.6516  6.6163

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.04242    0.17736   5.877 2.11e-08 ***
NA_Sales     1.13040    0.03162  35.745  < 2e-16 ***
EU_Sales     1.19992    0.04672  25.682  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.49 on 172 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.9664
F-statistic:  2504 on 2 and 172 DF,  p-value: < 2.2e-16
```

Stakeholders may use this model and predictions to inform marketing and go to market strategies, financial plans, inventory management, and sales target setting.

Further analysis should be undertaken to fix the skewness and normality issues with the sales dataset before Turtle Games actions any changes using these insights.