DA301

Turtle Games: Analysis of Customers and Sales Data

Kei WAI

2023

DA301

Background/context of the business (100 words):

Turtle Games is a global game manufacturer and retailer that sells books, board games, video games, and toys. The company collects data from sales and customer reviews to improve overall sales performance by utilizing customer trends. The business problem that the team of data analysts has been contracted to solve is to provide insights on how to improve sales performance by understanding how customers accumulate loyalty points, how different market segments can be targeted, how social data can be used to inform marketing campaigns, the impact of each product on sales, the reliability of the data, and the relationships between North American, European, and global sales. The team will analyze the data and present actionable insights that can help Turtle Games achieve its business objective.

Analytical approach (350 words):

The analytical approach for data importation, cleaning, and analysis involved several steps using Python and R.

In Python, the steps included importing libraries (pandas, nltk, TextBlob), importing data using pandas' read_csv() function, data cleaning by removing unnecessary columns/rows, handling missing values, and duplicates. Text preprocessing involved converting text to lowercase, removing punctuation/special characters, and stop words using the nltk library. Sentiment analysis was performed on preprocessed text using TextBlob to calculate polarity and subjectivity scores. Data analysis provided insights into reviews and summaries, identifying common themes and sentiments. Visualizations were created using matplotlib to further understand the data.

The choice of libraries, functions, and variables was based on their suitability and relevance to Turtle Games' objectives. The approach was organized and pertinent to the company's goals.

In R, the data was imported using read.csv() and choose.files() functions. Initial data exploration was conducted using functions like str(), summary(), head(), and dim(). Data cleaning involved checking for missing values using is.na() and sum() functions, dropping irrelevant columns, and verifying duplicate rows. The dplyr and ggplot2 libraries were used for data analysis and visualizations.

To understand how customers accumulate loyalty points, the data was grouped by "Platform" to calculate average "User_Score" and "Critic_Score". The PS4 platform exhibited the highest scores and user ratings. For targeting specific market segments, the cut() function created bins for "User_Score" to calculate average "Global_Sales". Higher user scores correlated with higher global sales. Scatter plots of "User_Score" vs. "Global_Sales" indicated that games with higher scores and user ratings had higher sales. The impact of each product on sales was analyzed by calculating total sales for each platform and genre, with PS4 and Action genre exhibiting the highest sales. Reliability of data was assessed by calculating skewness and kurtosis values for sales columns using the moments library.

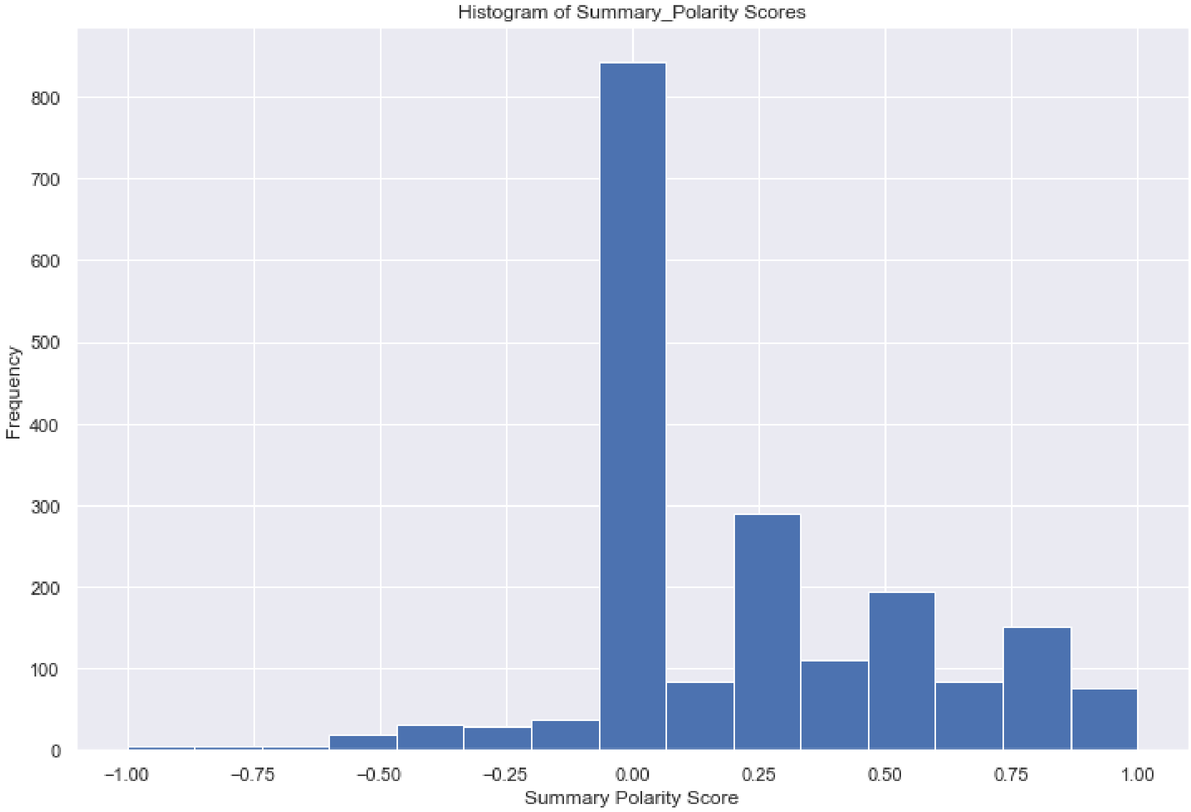
In summary, the analytical approach using Python and R involved importing data, cleaning, preprocessing, analysis, and visualization. The approach provided insights into customer sentiments and common themes, aiding in Turtle Games' decision-making process. The analysis identified the PS4 platform and Action genre as significant contributors to sales, and established a positive correlation between higher user scores and global sales.

Visualisation and insights (350 words)

A variety of visualizations were used to identify patterns and trends in the data, including histograms, scatter plots, correlation matrices, wordclouds and bar plots.

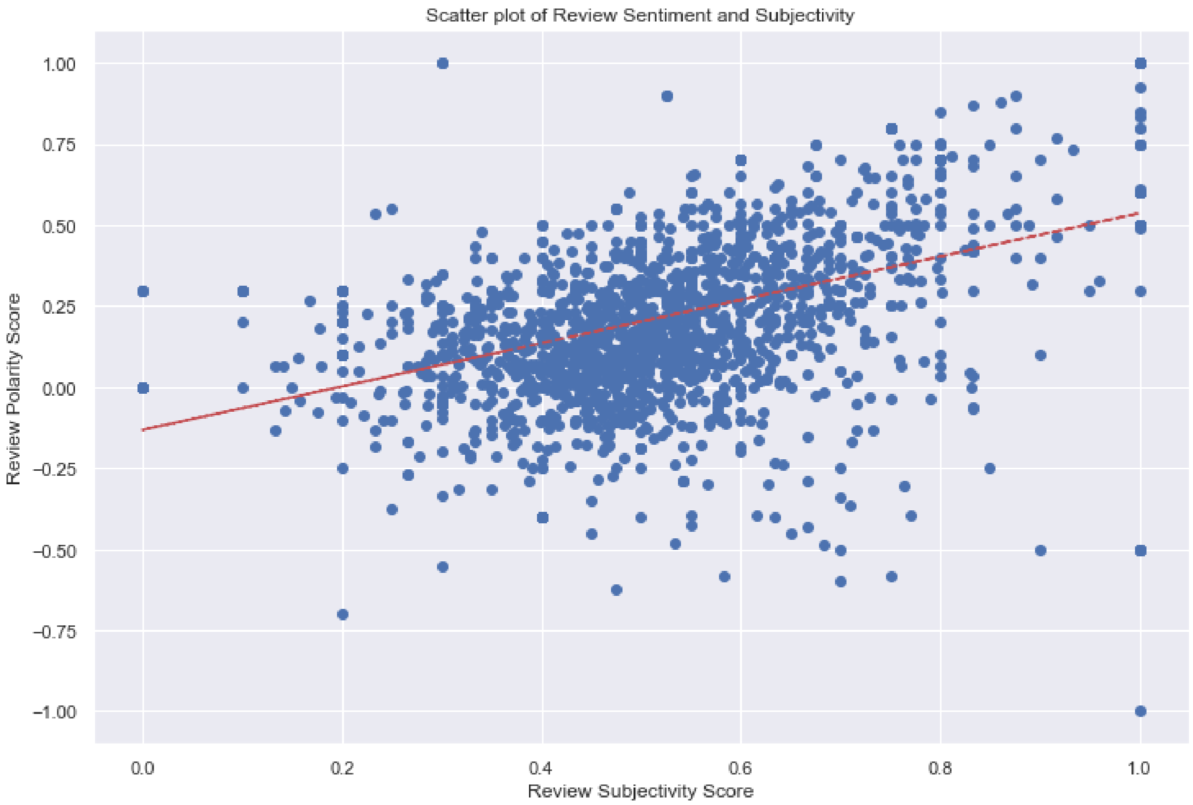
Histograms

1. The histogram of review polarity scores showed that the sentiment of the reviews was mostly neutral, with the majority of reviews having a polarity score between -0.2 and 0.2.


Histogram of Summary_Polarity Scores
Frequency
0
100
200
300
400
500
600
700
800
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Summary Polarity Score

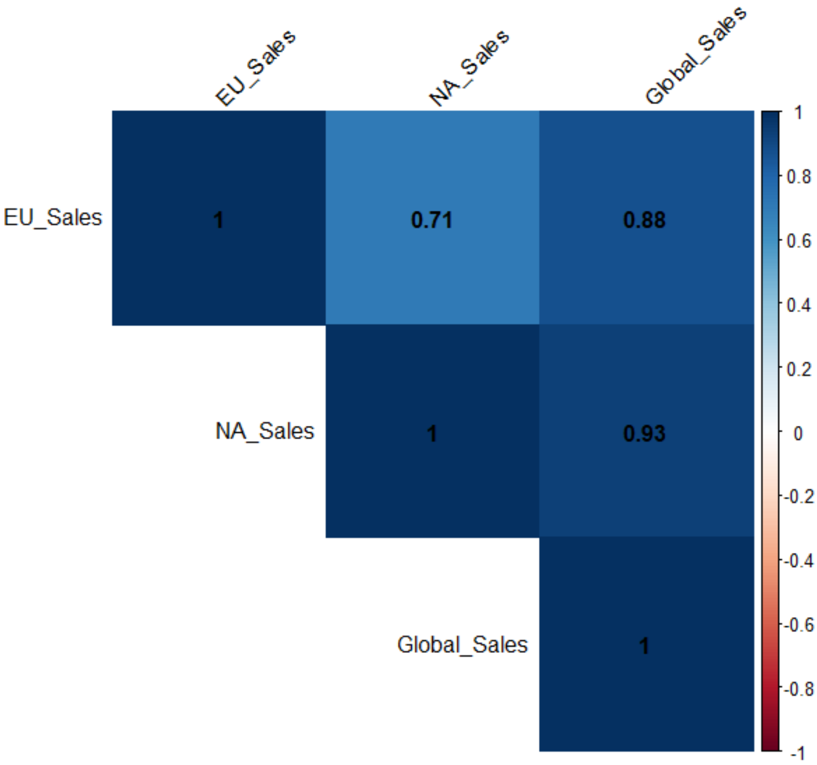
Scatter Plots

1. A scatter plot was used to show the relationship between the sentiment and subjectivity of the reviews, The scatter plot of the relationship between sentiment and subjectivity of the reviews showed a weak positive correlation between the two variables, indicating that highly subjective reviews could be either positive or negative.



Correlation Matrix

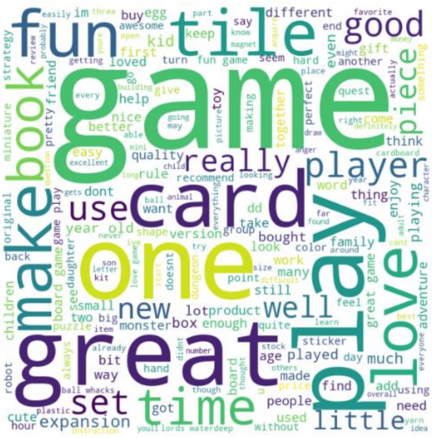
1. A correlation matrix was used to visualize the pairwise correlations between sales data columns. This visualization showed that all pairs of sales data columns were strongly positively correlated, with the highest correlation between NA_Sales and Global_Sales. This visualization was relevant to Turtle Games' objective of understanding the relationship between North American, European, and global sales and identifying any groups within the customer base that could be used to target specific market segments.



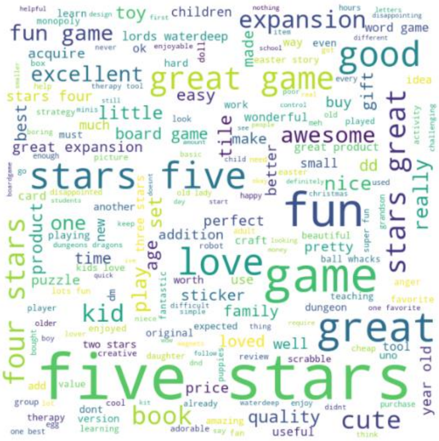
Wordclouds

1. Two wordclouds based on the review and summaries of customers are generated. These visualizations help identify the most frequent words used in the reviews and summaries, allowing for the identification of key themes and sentiments.

Review Wordcloud

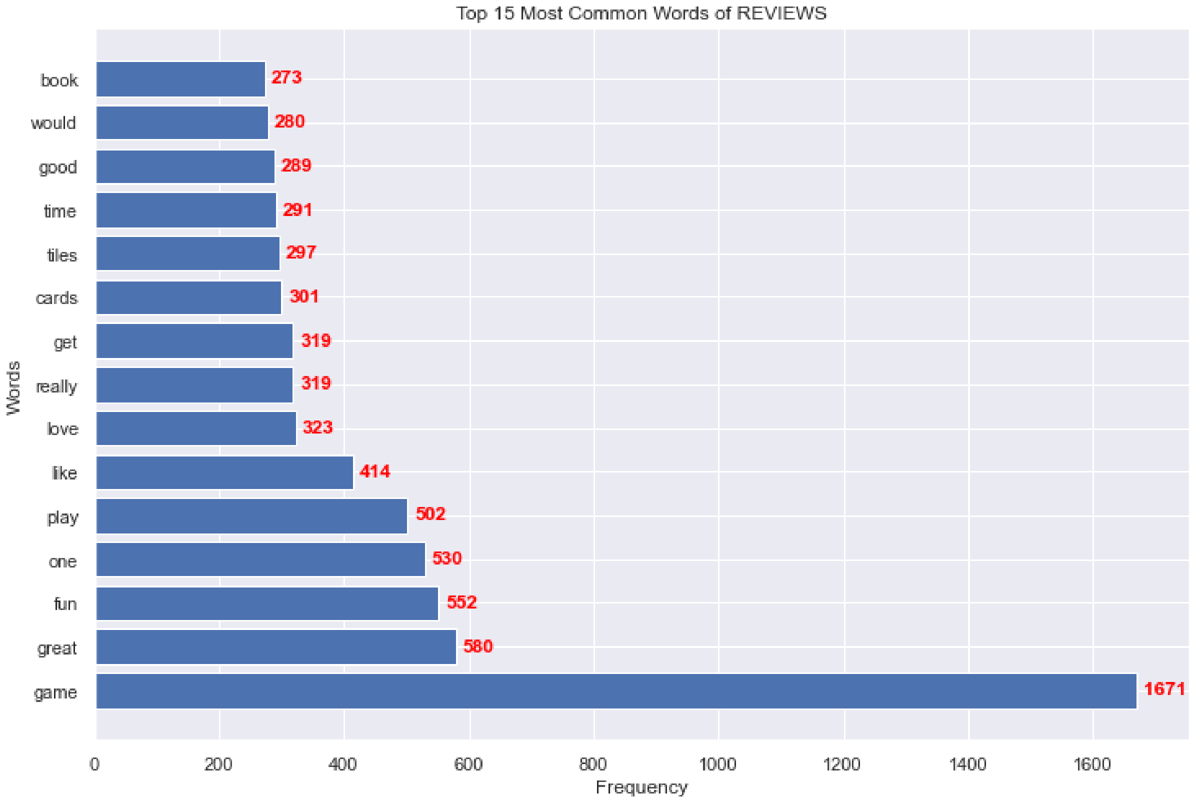


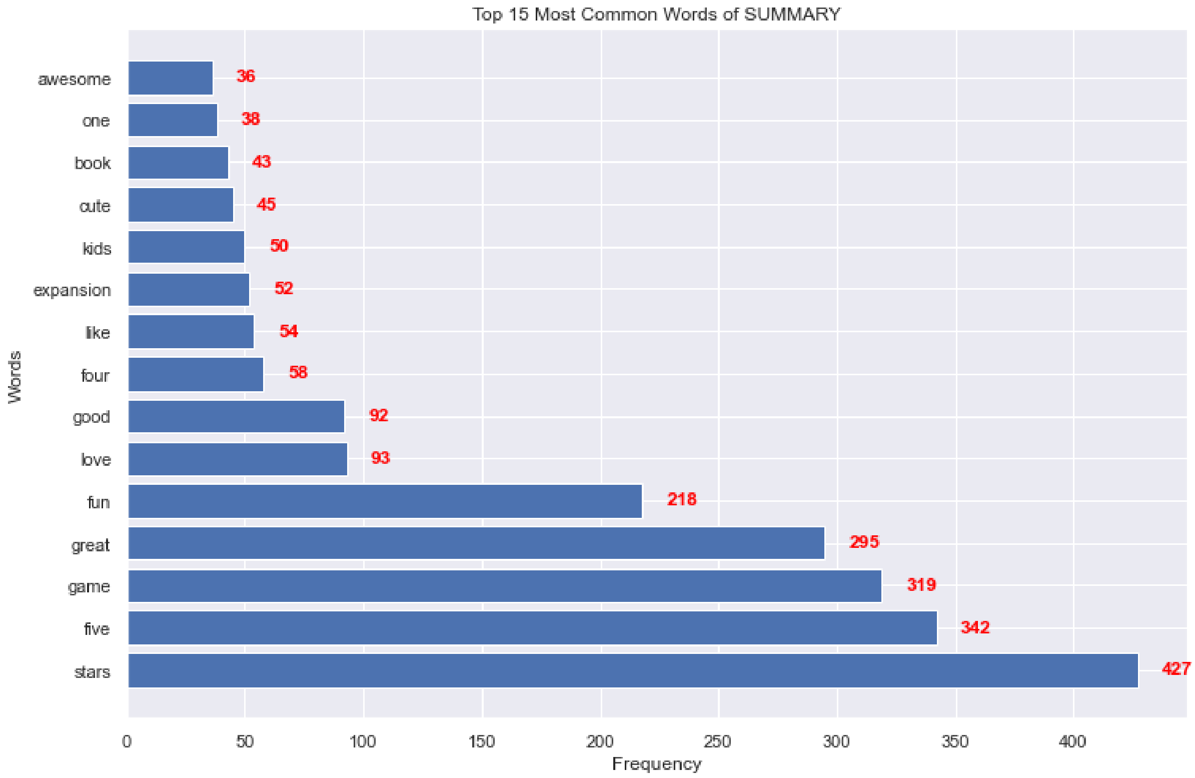
Summary Wordcloud



Barplots

1. Two horizontal barplots based on the review and summaries are generated. These visualizations help identify the most frequent words used in the reviews and summaries with numerical frequency presentations are most useful for data deep dive and quantitative analysis.





Overall, the selected visualizations provided clear and intuitive ways to understand the data and identify patterns and trends. The interpretations of the visualizations were detailed and insightful, providing Turtle Games with actionable insights to improve their sales performance. By using a variety of visualizations and methods, the team was able to paint a comprehensive picture of the data and identify actionable insights to improve sales performance.

Patterns and predictions (200 words):

1. Customer reviews indicate that Turtle Games' products are enjoyable and engaging based on the most frequent words in reviews and summaries, such as "game," "great," "fun," and "play."

2. The average summary polarity score of 0.224 suggests that customers generally have a positive opinion of Turtle Games' products. However, the average review polarity score of 0.210 indicates that customers may be more critical in their reviews than in their summaries. Turtle Games should pay attention to customer feedback to address concerns and criticisms.

3. The sales data analysis indicates a high degree of positive correlation between sales data columns. This means that successful marketing and sales strategies applied in one region can work in other regions as well. This can lead to increased sales and revenue.

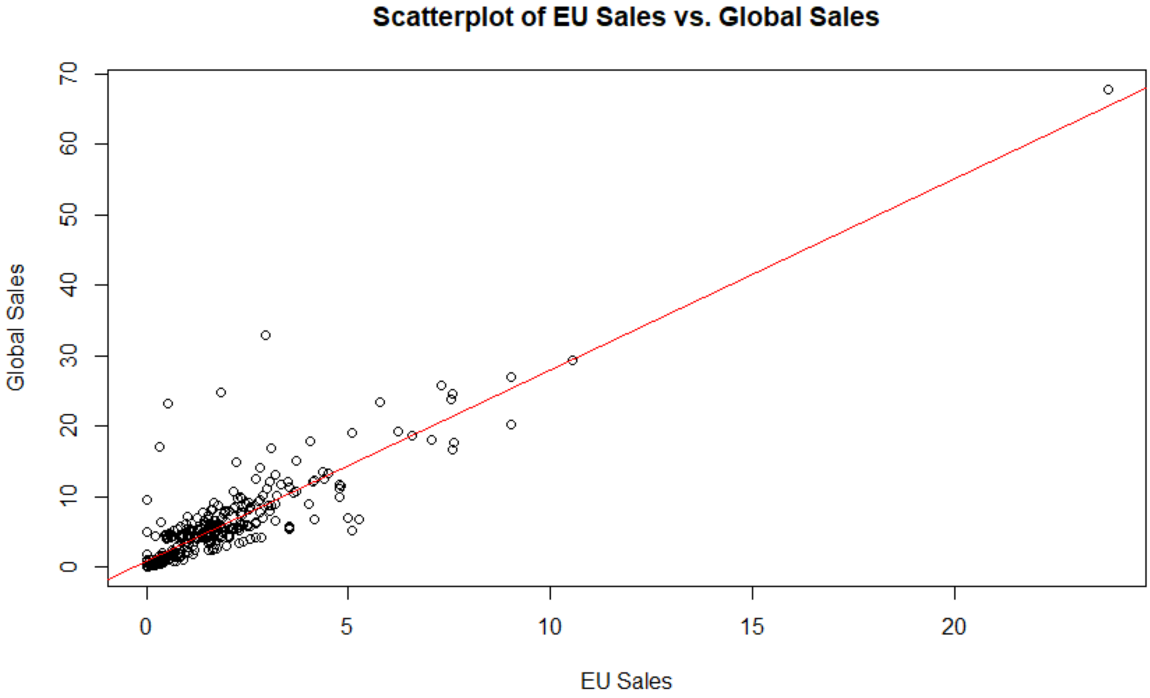
4. However, Turtle Games should exercise caution when making decisions based on the highly non-normal distribution of sales data with highly skewed distributions and heavy tails. The

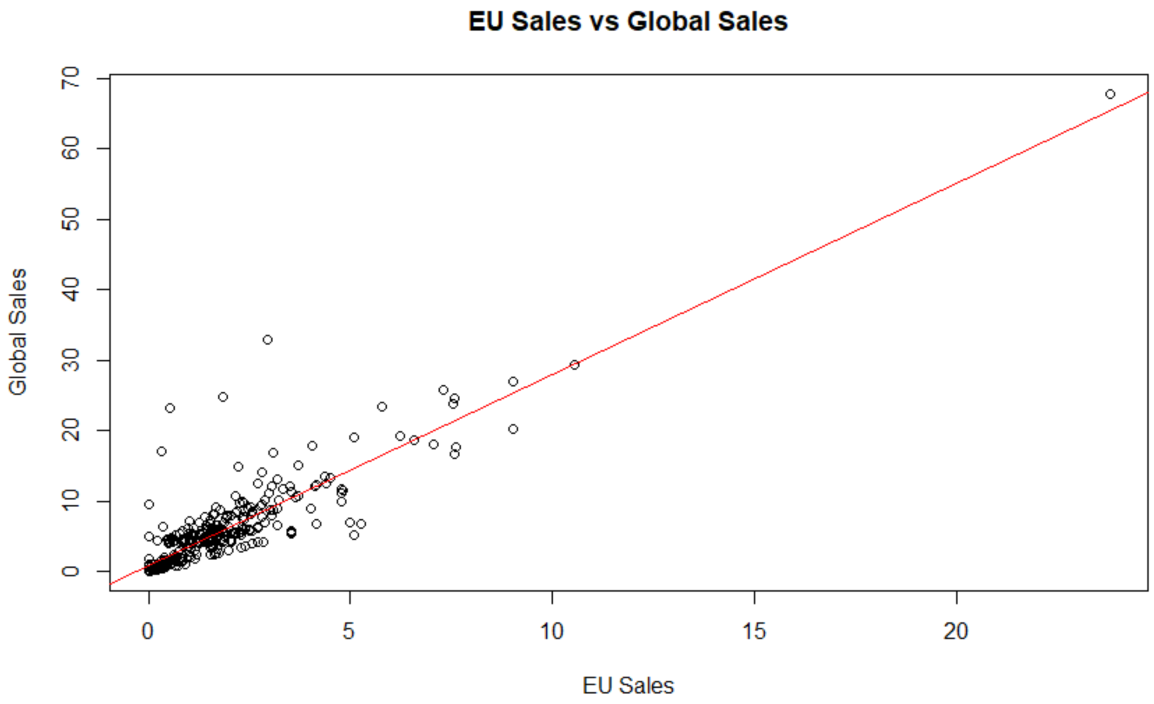
sales data may not be representative of the overall population, and Turtle Games should ensure decisions are based on reliable data.

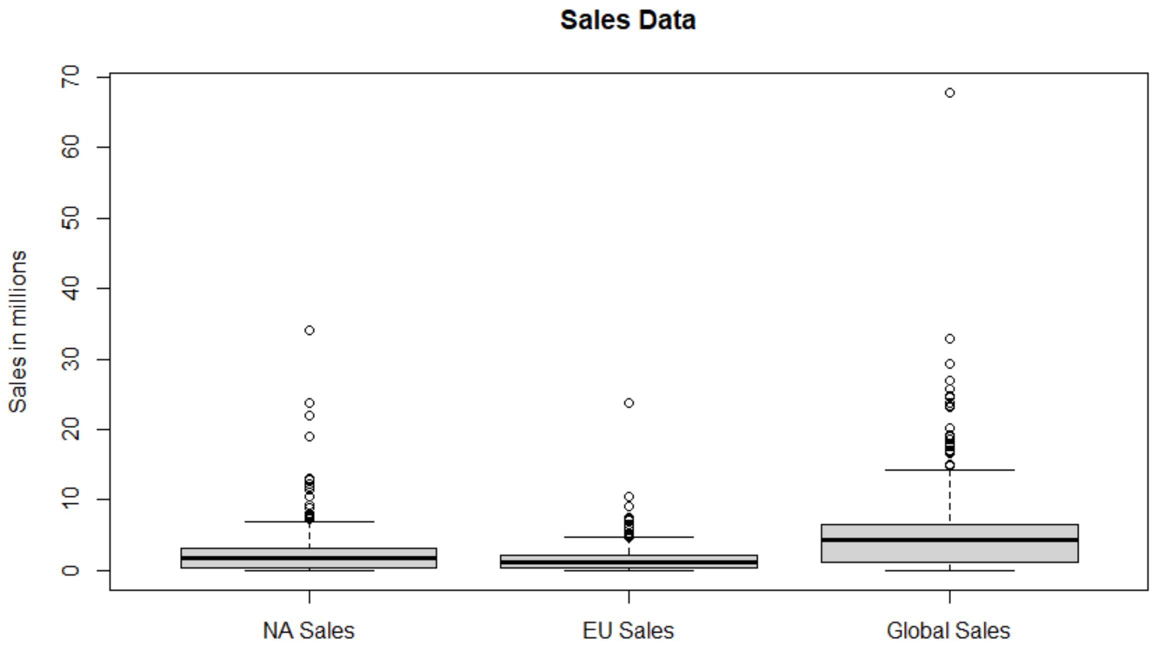
5. The team used the present sales data to create Linear Regression Model for Sales prediction. However, Root mean squared error (RMSE) of the linear regression model for predicting global sales is evaluated and resulting RMSE of 19.25 indicates that, on average, the model's predictions are off by approximately 19.25 units of global sales. Moreover, residual plot to check for linearity and homoscedasticity has a cone-like pattern, it suggests that the assumptions of linearity and homoscedasticity may not be met.

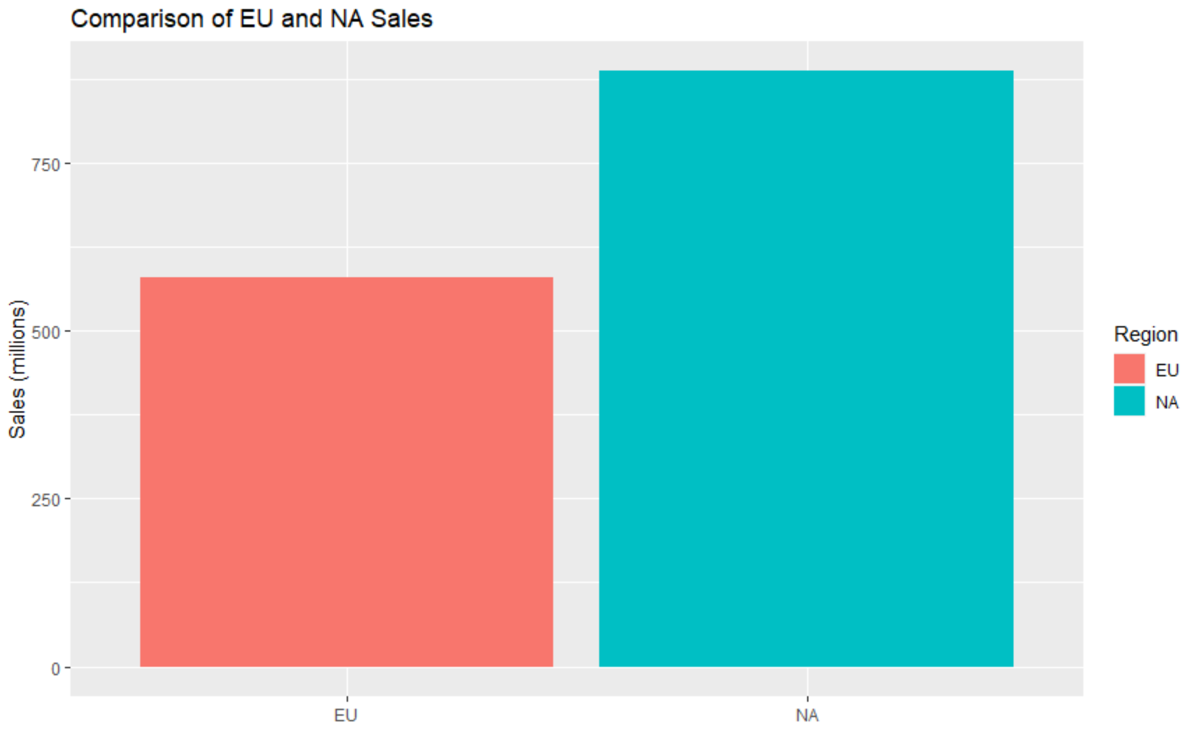
Overall, the analysis provides valuable insights for Turtle Games. By leveraging positive customer sentiment, identifying successful sales strategies, and monitoring customer feedback, Turtle Games can continue to grow its customer base and improve brand reputation. The company should be cautious when making decisions/ predictions based on sales data and address customer concerns to improve satisfaction and retention. Shall indeed consider to checking for outliers, transforming the data, considering non-parametric methods, or exploring different distributions prior to any further analysis and decision making.

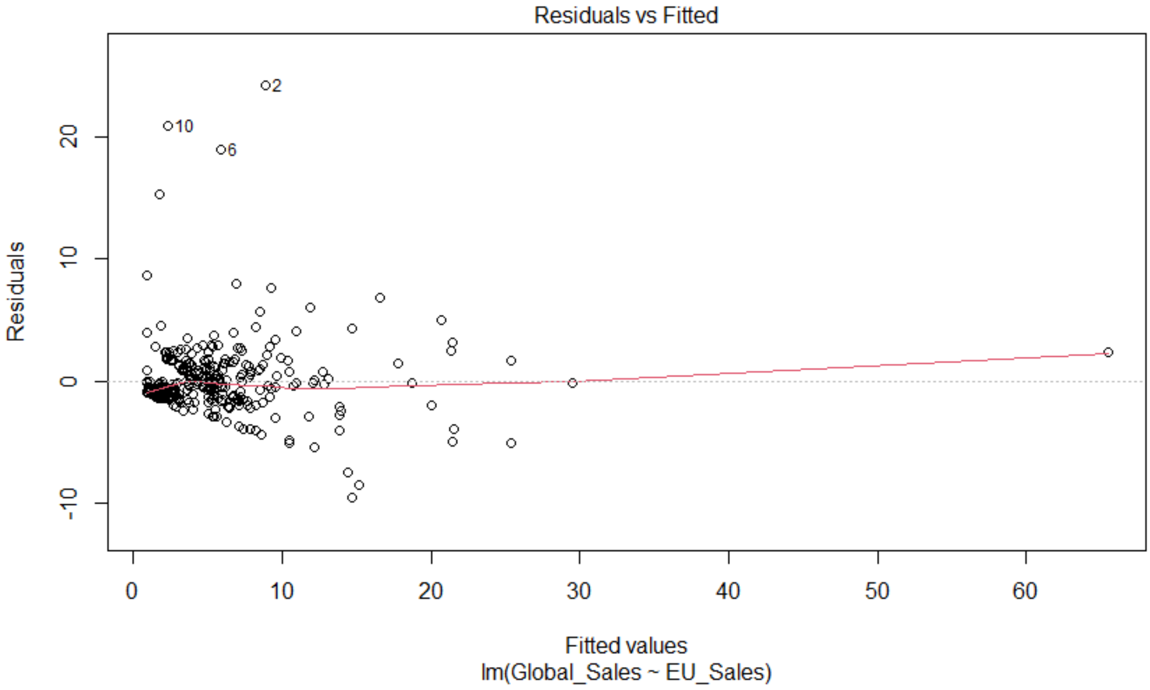
Appendix:

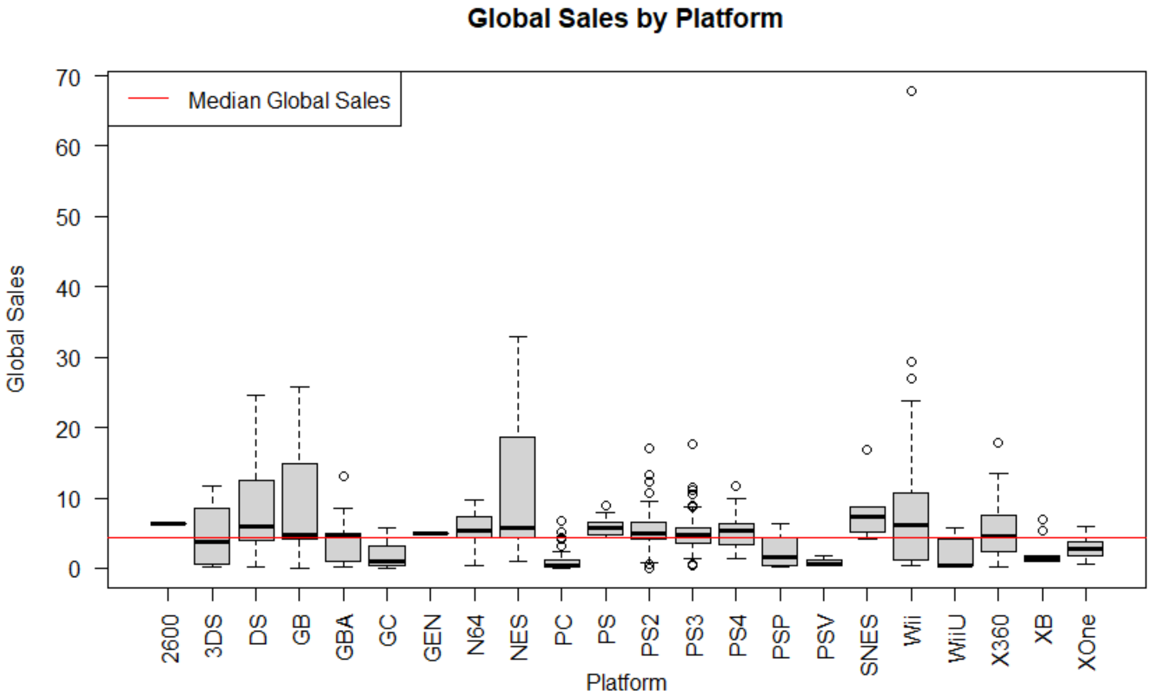

Scatterplot of EU Sales vs. Global Sales
Global Sales
0
10
20
30
40
50
60
70
EU Sales
0
5
10
15
20

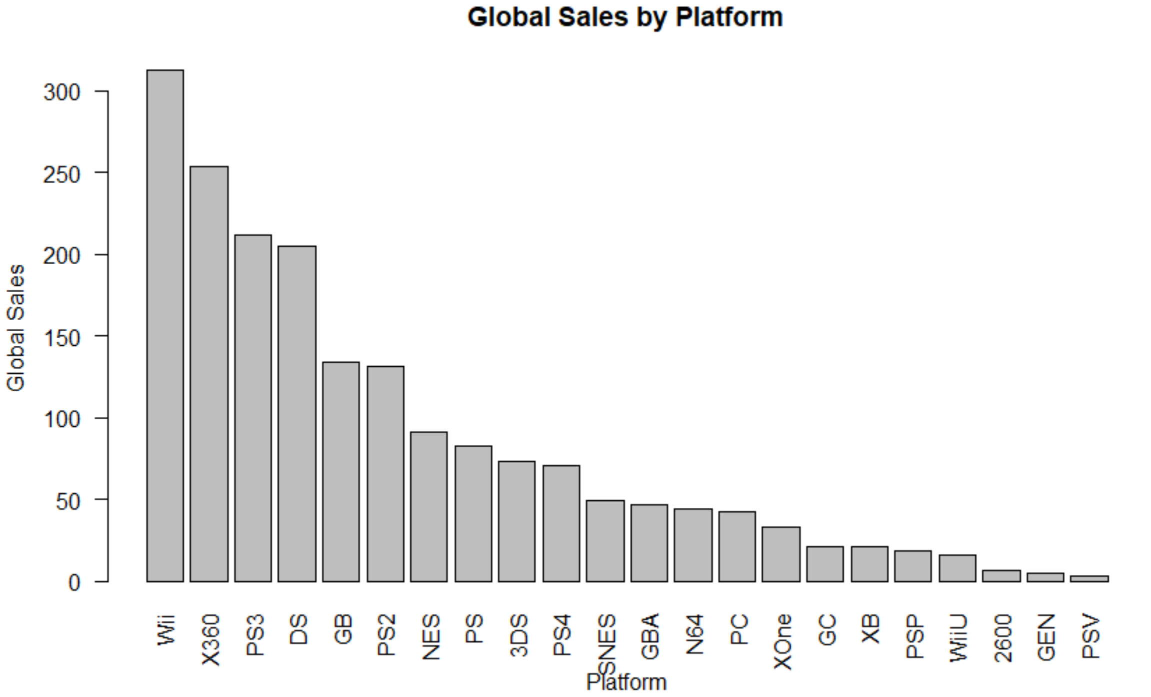

EU Sales vs Global Sales
Global Sales
0
10
20
30
40
50
60
70
0
5
10
15
20
EU Sales

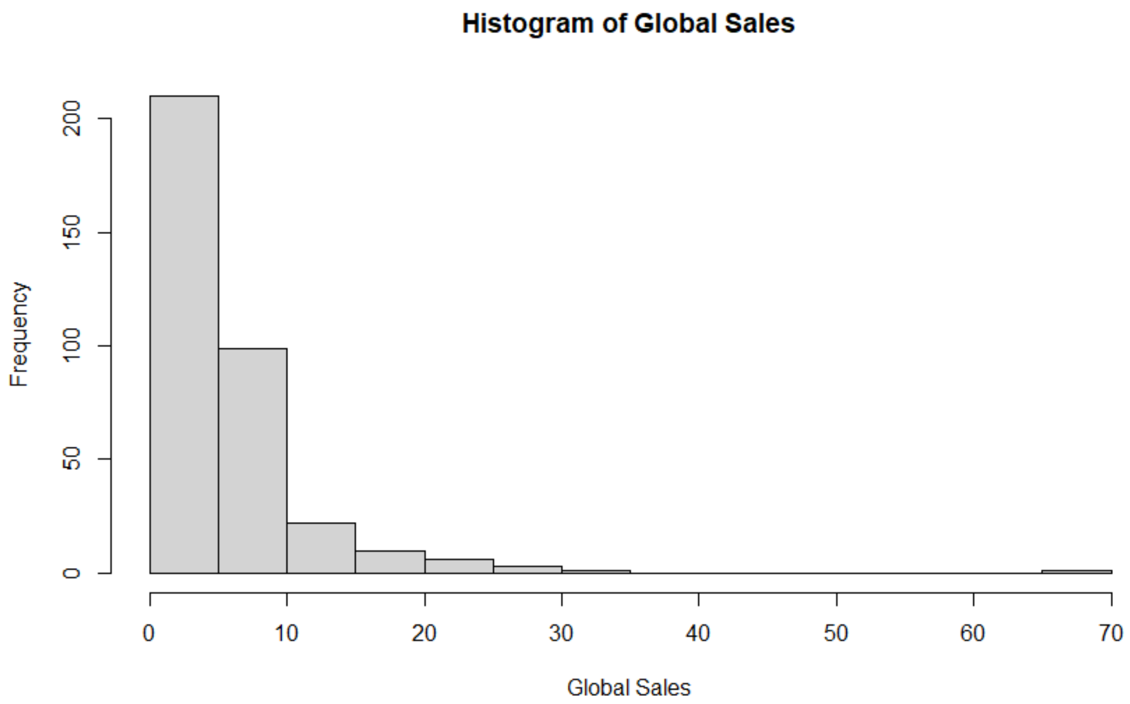

Sales Data
Sales in millions
0
10
20
30
40
50
60
70
NA Sales
EU Sales
Global Sales

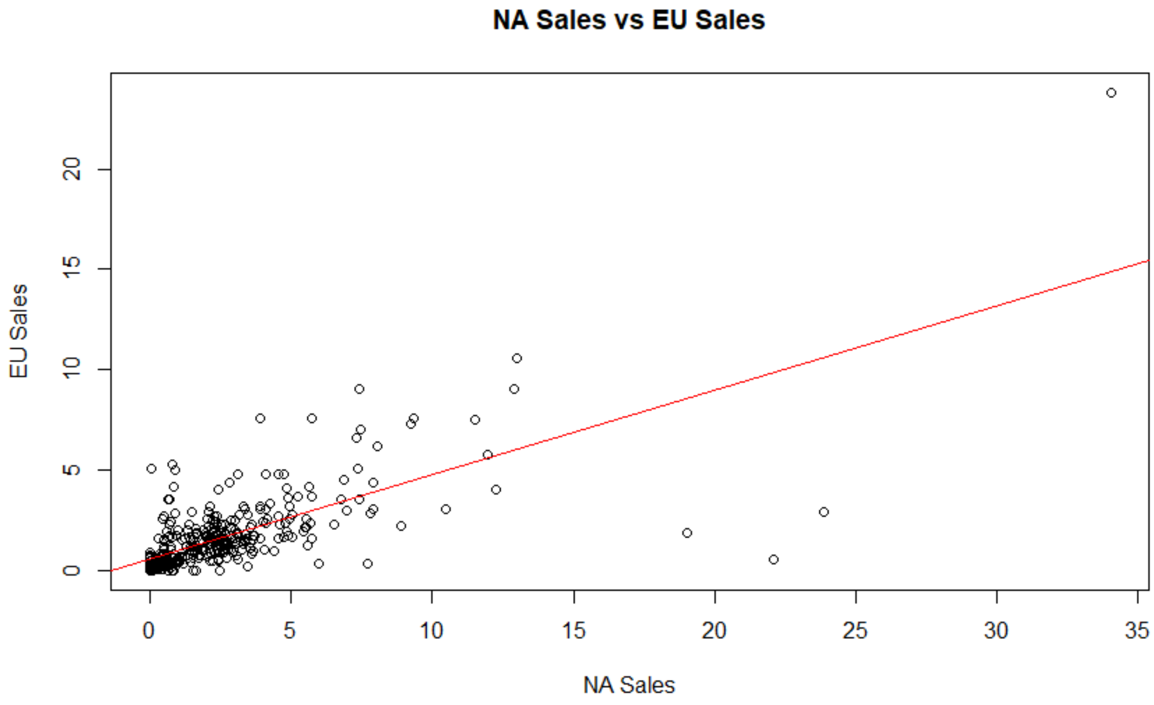

Comparison of EU and NA Sales
Sales (millions)
750
500
250
0
EU
NA
Region
🟥 EU
🟦 NA

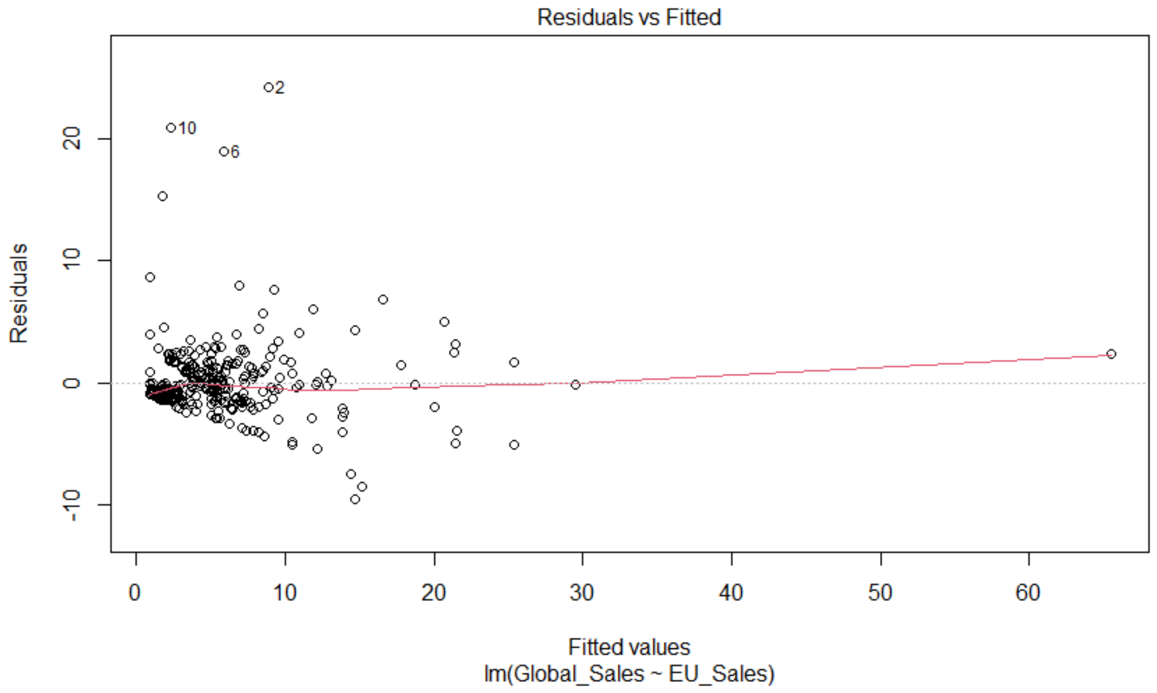

Residuals vs Fitted
Residuals
20
10
0
-10
0
10
20
30
40
50
60
Fitted values
lm(Global_Sales ~ EU_Sales)
2
10
6

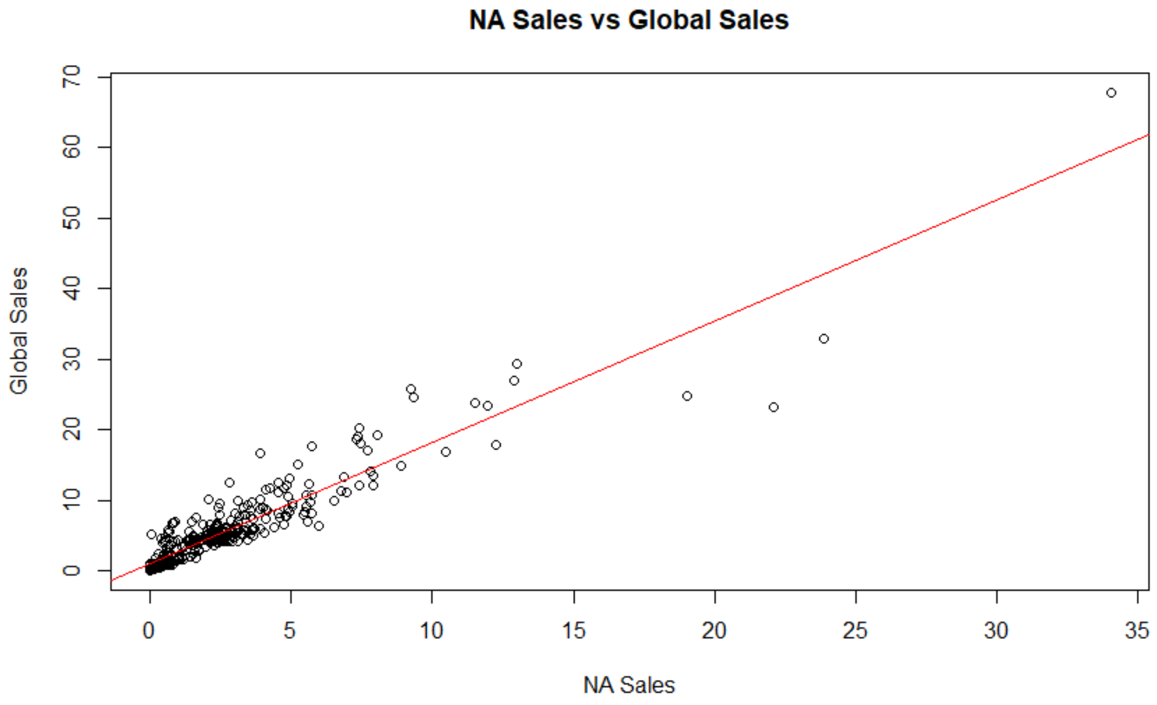

Global Sales by Platform
Median Global Sales
Global Sales
70
60
50
40
30
20
10
0
2600
3DS
DS
GB
GBA
GC
GEN
N64
NES
PC
PS
PS2
PS3
PS4
PSP
PSV
SNES
Wii
WiiU
X360
XB
XOne
Platform

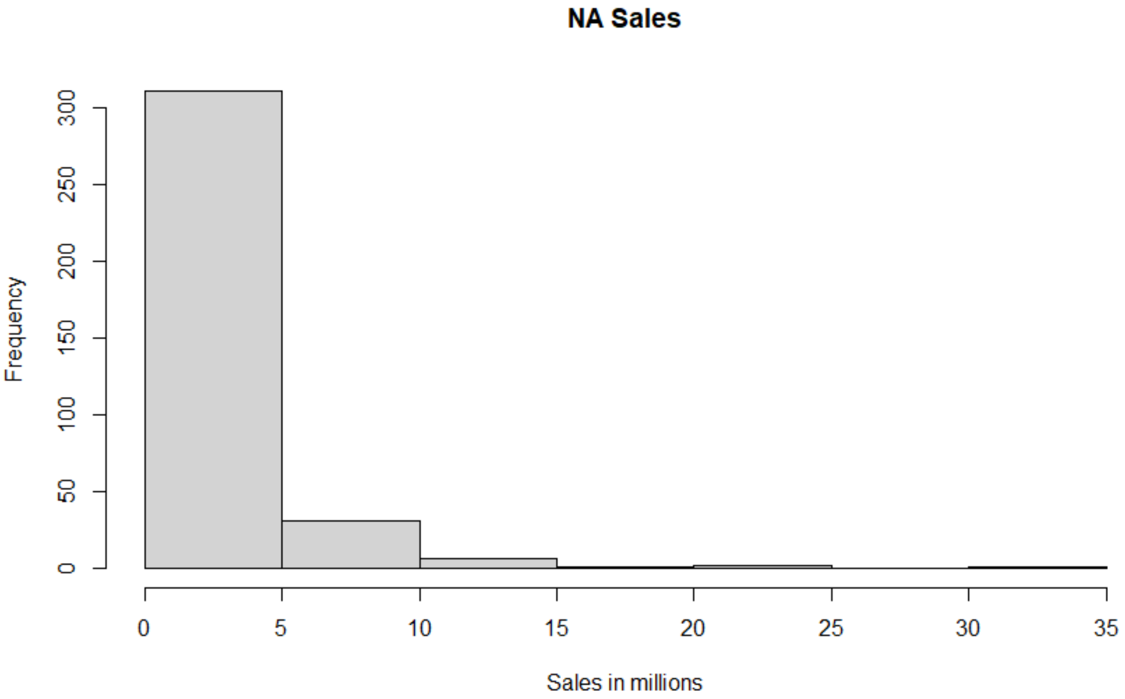

Global Sales by Platform
Global Sales
300
250
200
150
100
50
0
Wii
X360
PS3
DS
GB
PS2
NES
PS
3DS
PS4
SNES
GBA
N64
PC
XOne
GC
XB
PSP
WiiU
2600
GEN
PSV
Platform

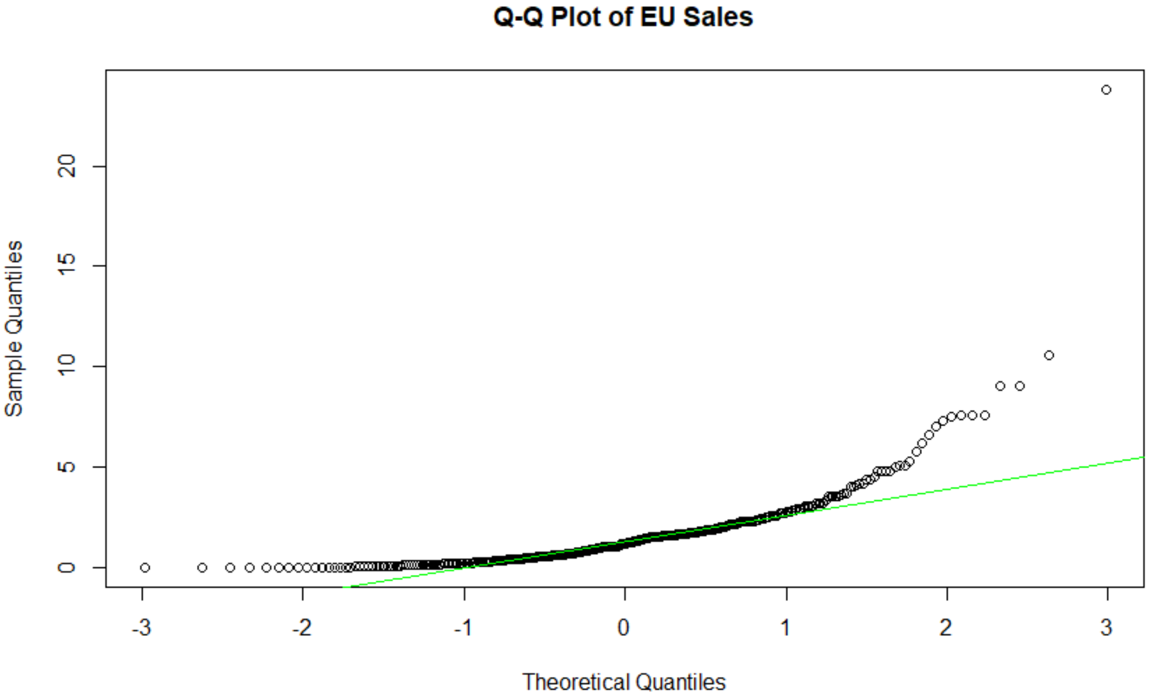

Histogram of Global Sales
Frequency
0
50
100
150
200
0
10
20
30
40
50
60
70
Global Sales

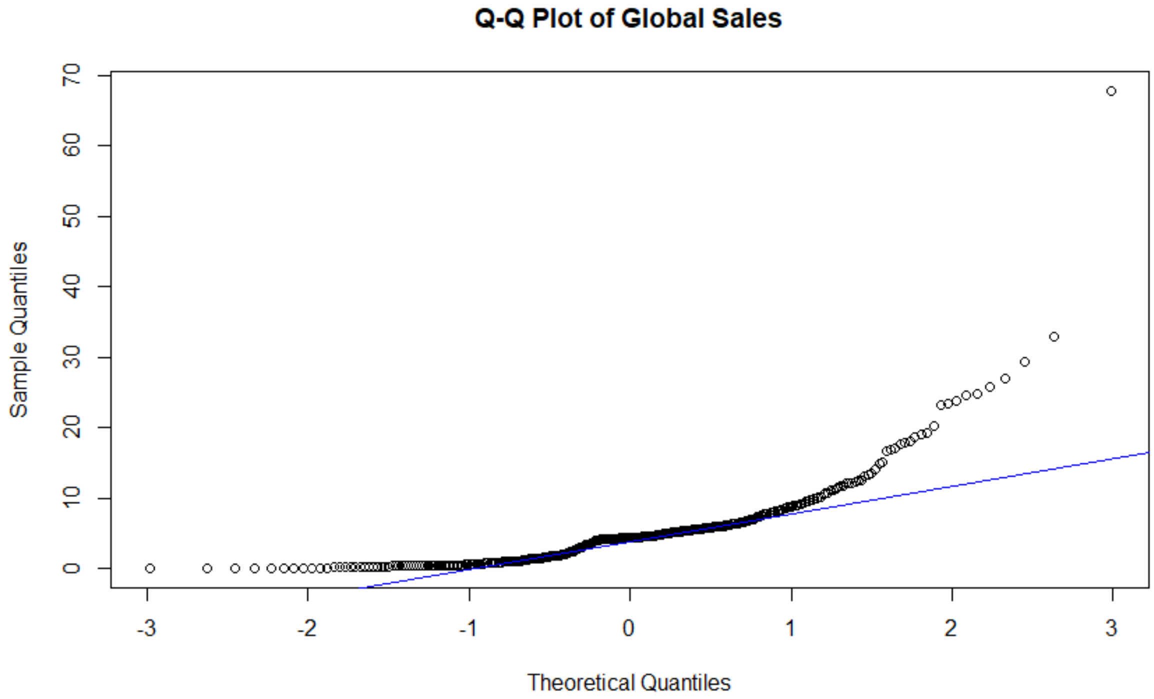

NA Sales vs EU Sales
EU Sales
0
5
10
15
20
0
5
10
15
20
25
30
35
NA Sales

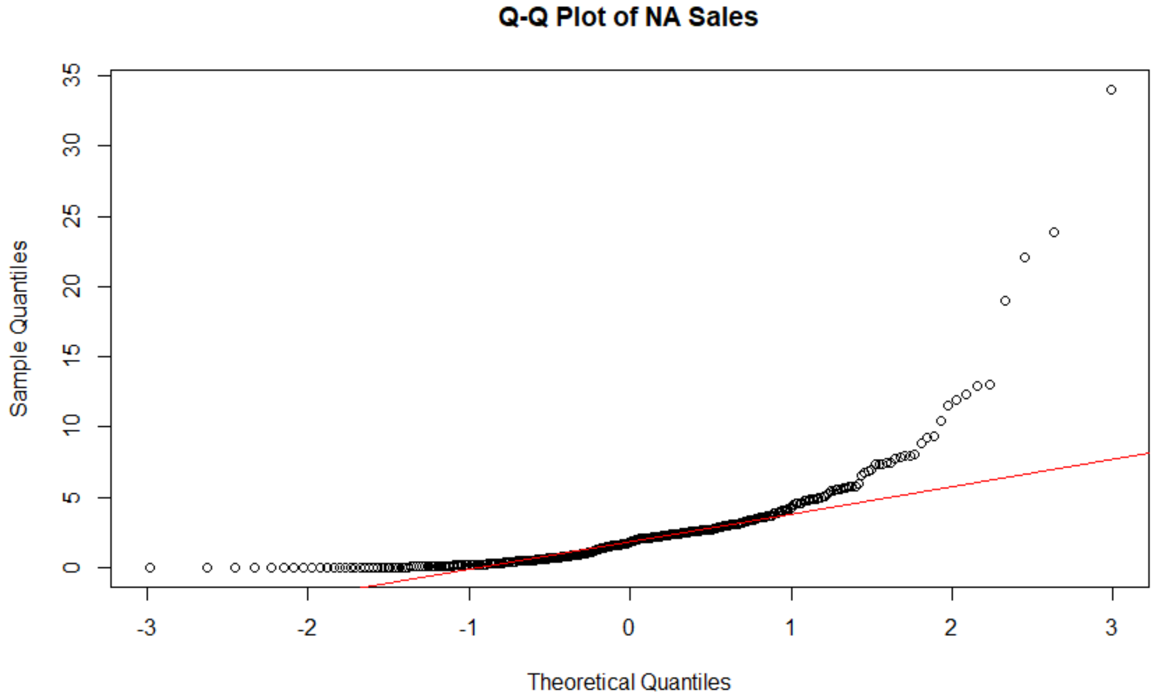

Residuals vs Fitted
Residuals
-10
0
10
20
2
10
6
0
10
20
30
40
50
60
Fitted values
lm(Global_Sales ~ EU_Sales)

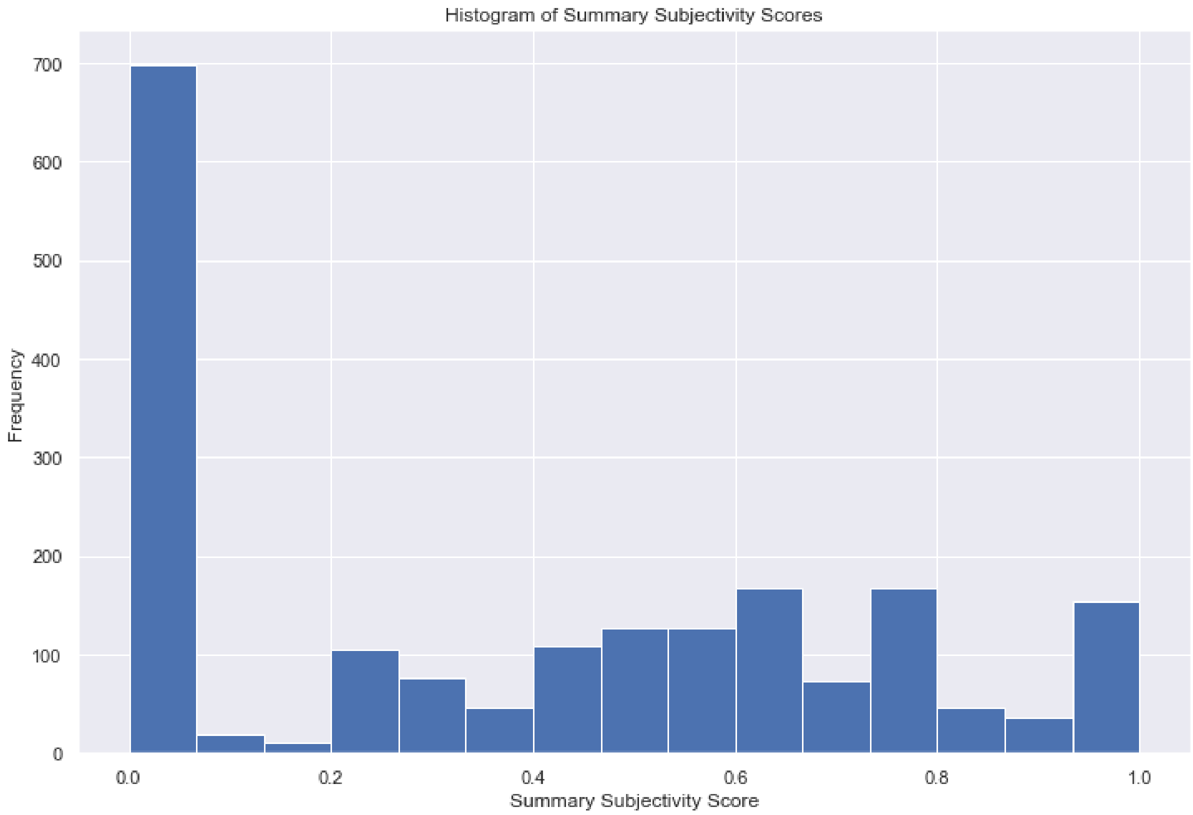

NA Sales vs Global Sales
Global Sales
0
10
20
30
40
50
60
70
0
5
10
15
20
25
30
35
NA Sales


NA Sales
Frequency
0
50
100
150
200
250
300
0
5
10
15
20
25
30
35
Sales in millions


Q-Q Plot of EU Sales
Sample Quantiles
0
5
10
15
20
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Q-Q Plot of Global Sales
Sample Quantiles
0
10
20
30
40
50
60
70
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Q-Q Plot of NA Sales
Sample Quantiles
0
5
10
15
20
25
30
35
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Histogram of Summary Subjectivity Scores
Frequency
0
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0
Summary Subjectivity Score