TURTLE GAMES ANALYSIS

Predicting Future Outcomes

23/07/2023

SANA SALEEM
[Company address]

Contents

Turtle Games

Problem Statement

Turtle Games, a global game manufacturer and retailer has an objective to improve its sales performance based on customer trends. It seeks to analyse and predict data that aims to garner increased sales and hence customer loyalty.

The following processes have been carried out in order to analyse and predict data:

- Linear Regression to understand customer loyalty
- Clustering to identify potential target market segments
- Sentiment Analysis of online customer reviews to devise marketing campaigns
- Visualisation to understand the relationship between products and sales
- Determining the data reliability through Regression Analysis
- Determining the relationship between North American, European and Global Sales

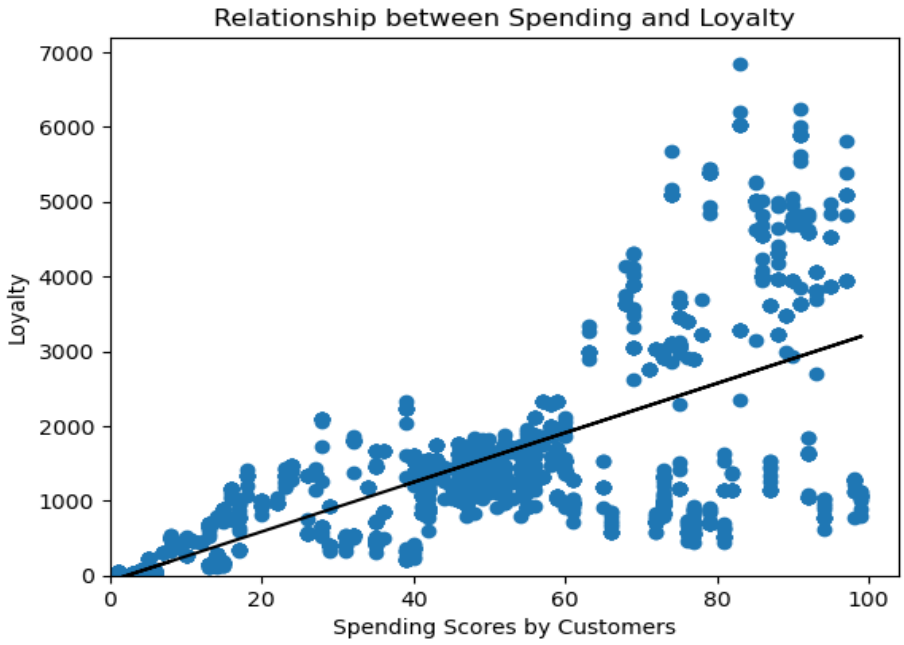
Customer Loyalty towards Turtle Games

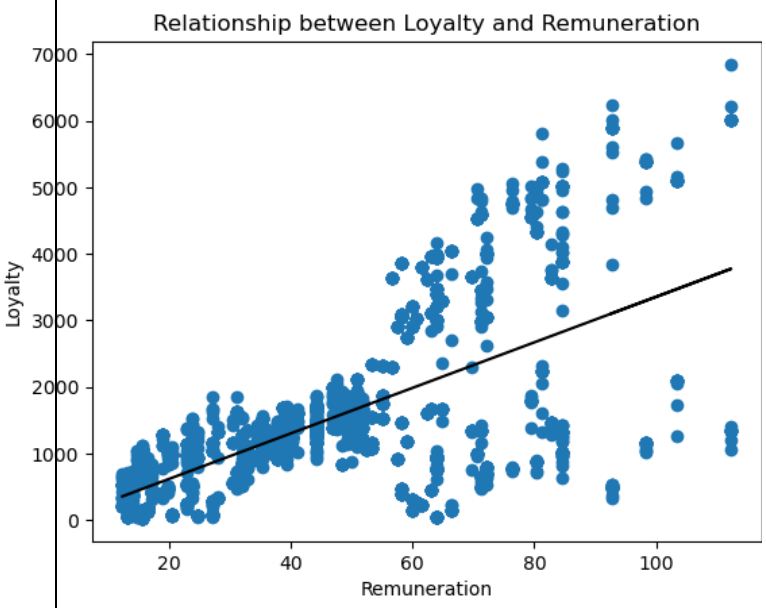
Process of determining Customer Loyalty

In order to understand customer loyalty, the data was loaded and explored for any missing values along with removing irrelevant columns and renaming existing ones. The new DataFrame was saved as a csv file and explored further to identify that it has 2000 rows and 10 columns.

Linear Regression was performed to understand if there is a relationship between customer spending, remuneration and age on the loyalty of the customers. Customer loyalty was treated as the dependant variable (y) while spending, remuneration and age were treated as independent variables. The OLS function was used to fit the model and other useful values such as estimated parameters, standard errors and predicted values were adopted to identify the coefficient, intercept and standard errors of the loyalty. Lastly, a simple linear regression model was used to predict loyalty based on the independent variables and a line of best fit was plotted.

Following are the graphs displaying the effect spending, age and remuneration have on the loyalty of the cus





Observations and Recommendations

The R squared value of 45.2% for spending is the greatest followed by 38% and 0.2% meaning 45.2%, 38% and 0.2% variability of loyalty is explained by spending, remuneration and age respectively. The p-value is below 0.05 for all the variables indicating that the data is highly significant. It can be observed that the variation of all three dependant variables is quite low as 45.2% which is the highest R squared value is towards the lower end. There is however a trend indicating that as spending and remuneration increases, customer loyalty also increases.

From the graph, we can clearly observe that there are many residual data points above the line of best fit which occur after a spending score of 60 and a remuneration amount of £55k. The data points are dense around spending scores between 40 and 60 and remuneration between £20k and £50k. There is however a negative impact on customer loyalty based on the age of the customer as most points lie below loyalty points of 2000. The R value is 0.2% which is extremely low and hence there is no proper line of best fit.

Factors which can be explored further:

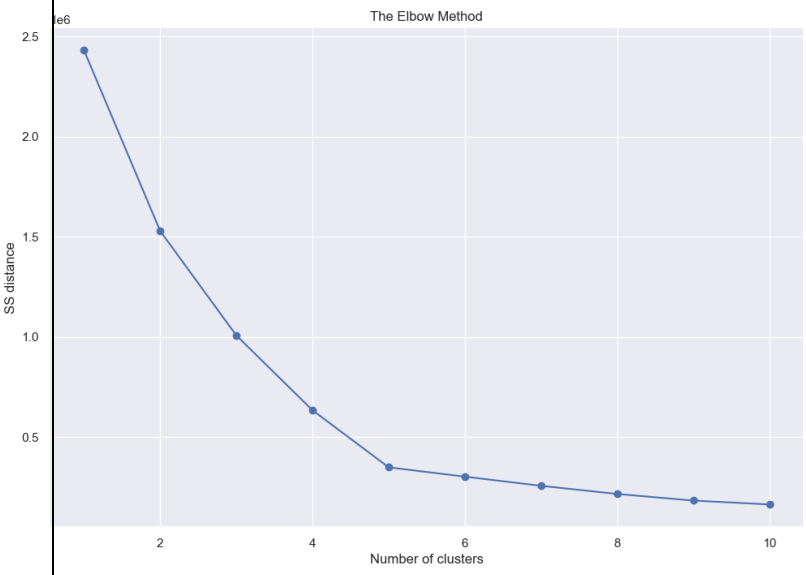
- Impact of profession on loyalty.
- Why some customers with a spending score of 100 may have such low loyalty.
- Focus on customer's loyalty towards other competitor online games.
- How much time spent on social media has an impact on loyalty.
- Any loyal users without a social media presence.
- Products with a greater loyalty.
- Demographics of customers.

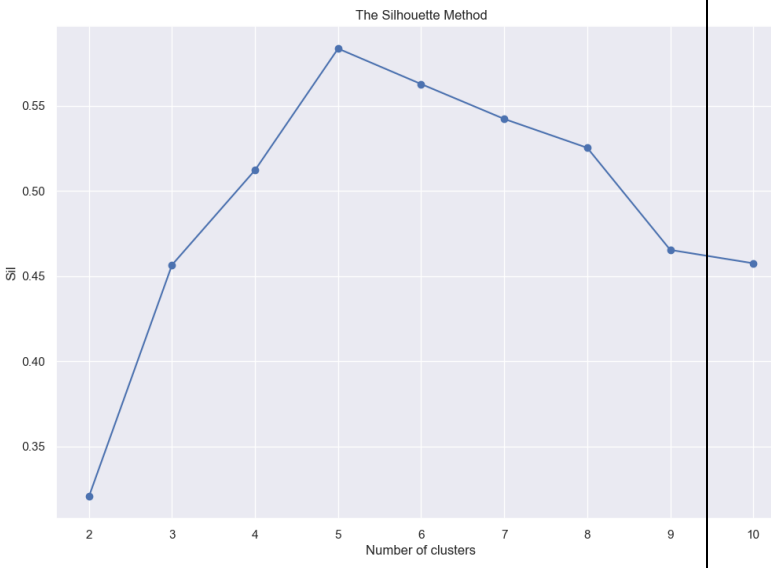
Identifying and targeting market segments based on K-means Clustering

K means clustering technique was employed in order to segment the market based on homogeneity by determining the optimal number of clusters through the Elbow and Silhouette methods focusing on gender wise spending and remuneration.

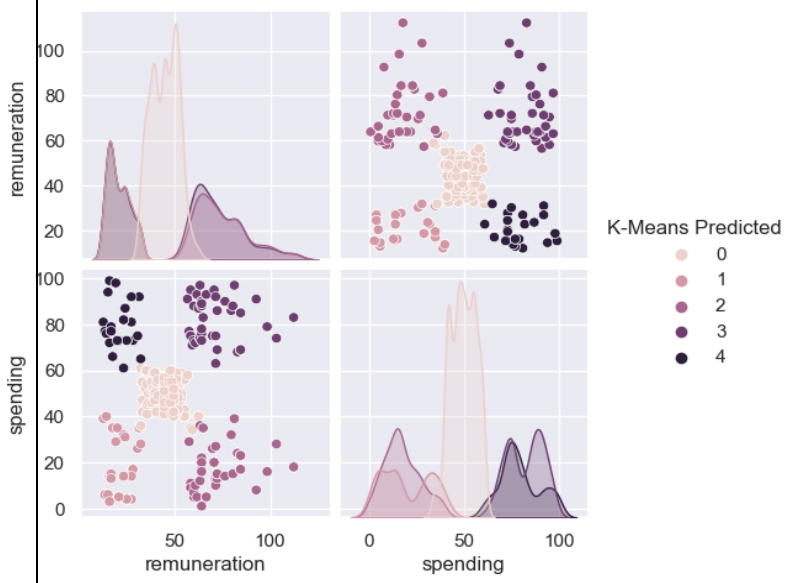
The data was loaded and explored and irrelevant columns were dropped. A scatterplot was then created using remuneration on the x axis and spending on the y axis. This was followed by creating a pairplot keeping the heu as gender. Elbow chart and Silhouette methods were used to determine the no of the clusters which indicated a no of 5 along with predicted values. A test was carried out using 3 and 4 clusters however 5 clusters were eventually used to fit the model.

From a glimpse, it can be observed that remuneration and spending is greater for females.

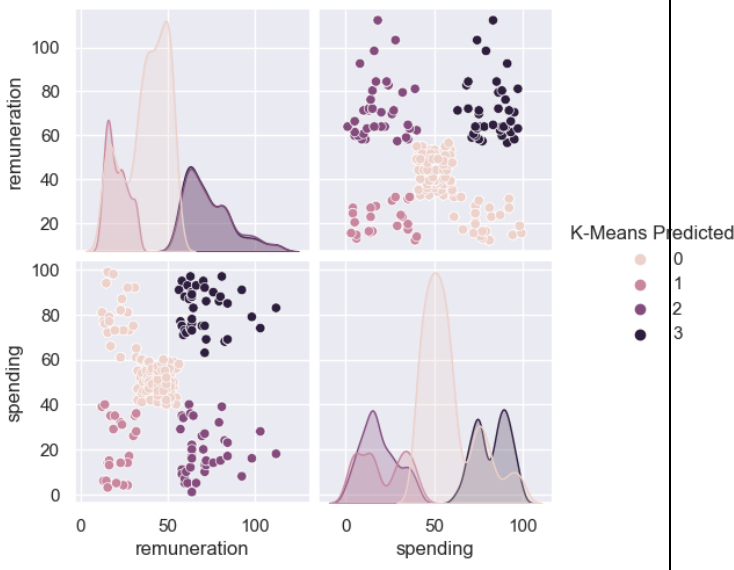




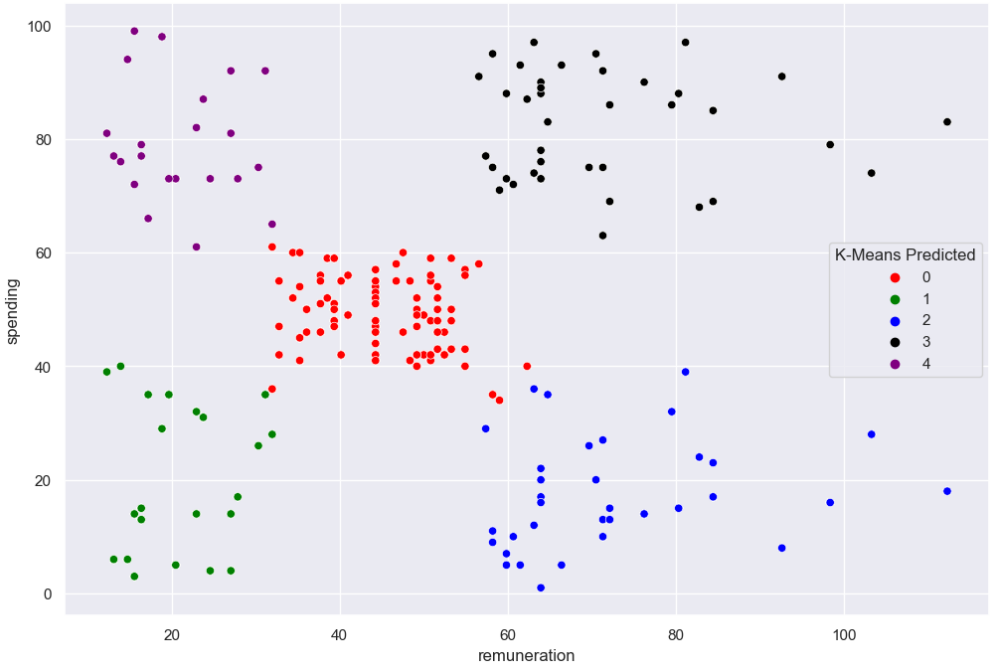
5 Clusters



4 Clusters



Cluster 5 shows a remuneration and spending of £50k and 50 pts respectively. A model of 4 and 3 clusters was also observed but 5 seemed to be the best fit as the clusters do not seem to be overlapping unlike in 3 or 4 clusters.



There can be seen Five prominent clusters of the following:

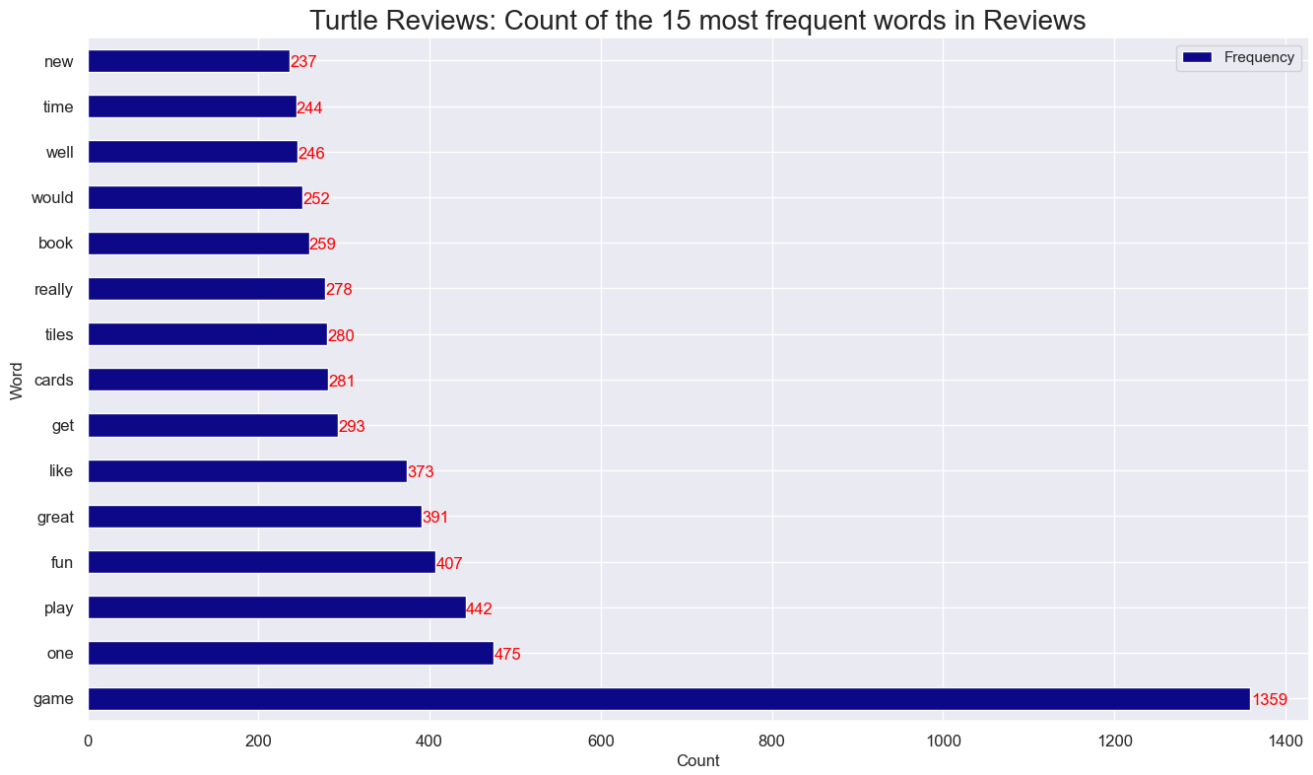
Low Remuneration	Low Spend
Low Remuneration	High spend
High remuneration	High spend
High remuneration	Low spend
Average Remuneration	Average Spend

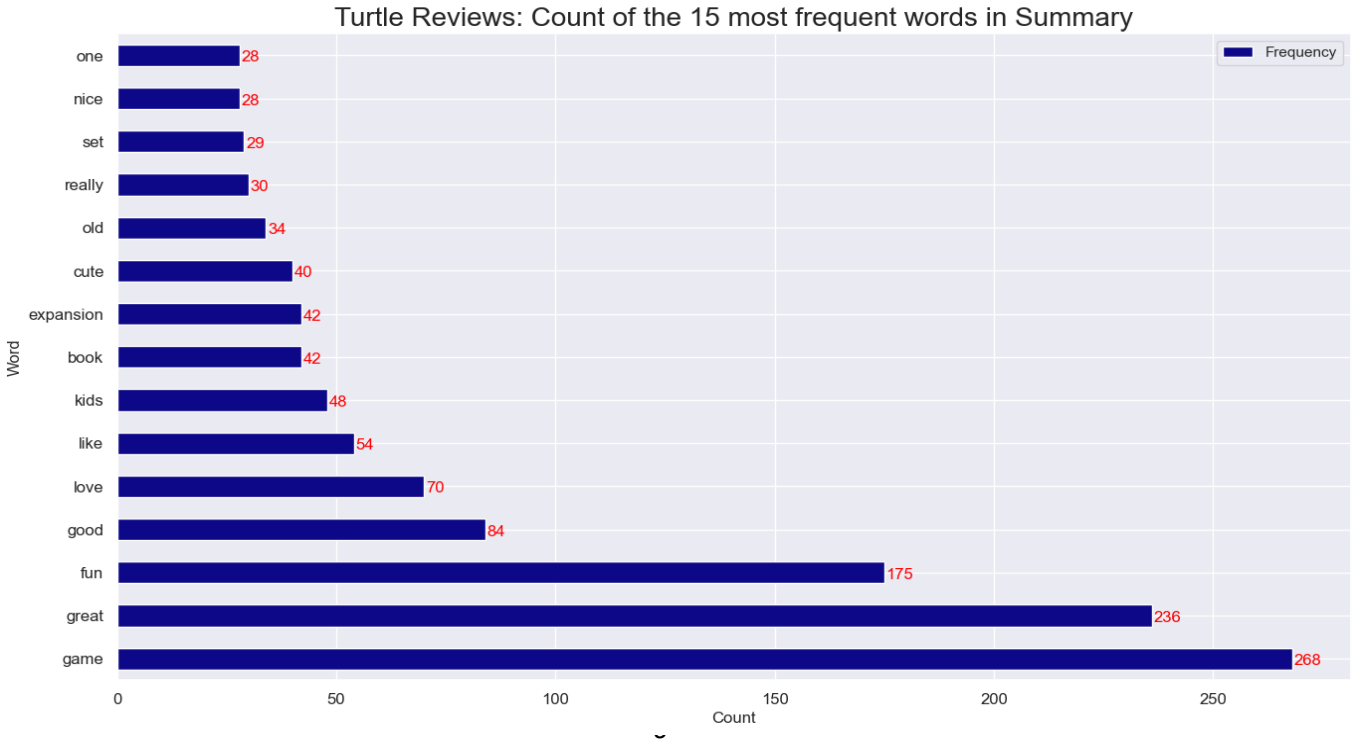
Observations and Recommendations

There are 5 prominent clusters which should be targeted by Turtle Games as these are homogenous group of individuals with ones having a remuneration between £40k-£60k and spending between 40 and 50 are the most widespread. There should be a more personalized approach, targeting this group of individuals with ads on social media. The ones with high remuneration which is above £60k and a low spend which is below 40 should also be targeted through social media and other marketing campaigns as these individuals may not have enough time to play.

Sentiment Analysis of Customer Reviews

A sentiment analysis was performed on the review and summary column of the data by first using apply to change all the words to lower case and using replace function for punctuation this was followed by dropping any duplicates then using tokenisation to create word clouds. Frequency distribution was carried out in order to remove stop words and a polarity and subjectivity score was observed.

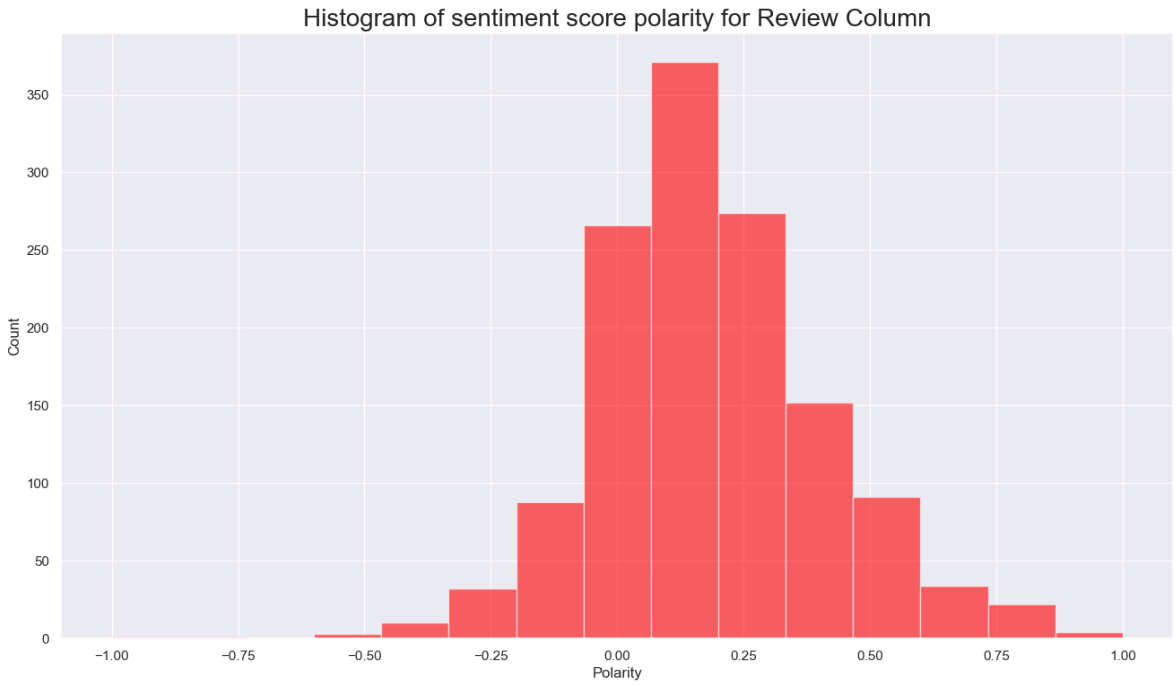




Sentiment Polarity Scores for Review and Summary

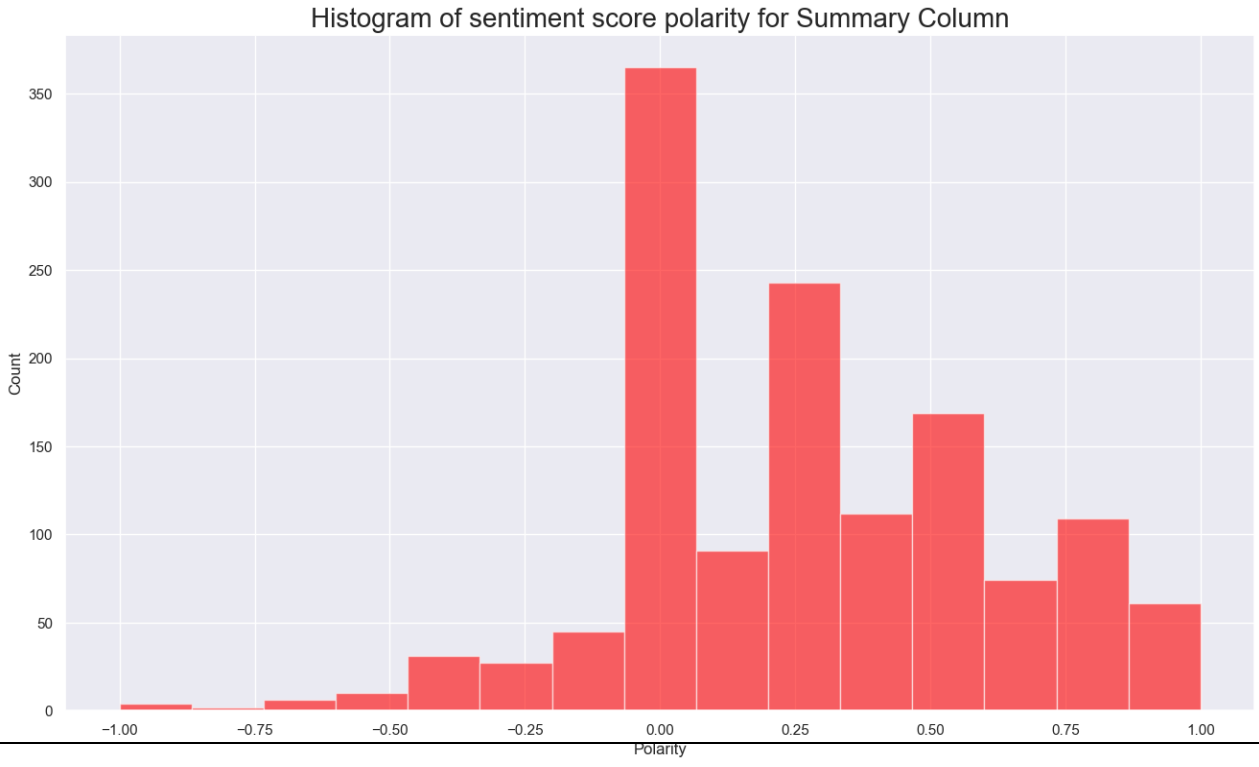
Review Column

The sentiment polarity scores for the Review column suggest a large amount of comments between the range of 0 and 0.5 indicating a trend towards the positive sentiment while most of the comments were fairly neutral suggesting that customers are overall liking the game and associating positive feelings towards it.



Summary Column

The sentiment polarity scores of the Summary Column suggest that large amount of comments are neutral with a polarity score of 0. A great number of comments further have a score of 0.25, 0.5 and 0.75.



Subjectivity Scores for Review and Polarity

The Subjectivity Score for Review column is greatly between 0.4 and 0.6 hence suggesting opinions which are a mix of fact based and opinion based.

The Subjectivity Score for Summary column however is largely focused on 0 thereby indicating mostly fact based summaries.

Observations

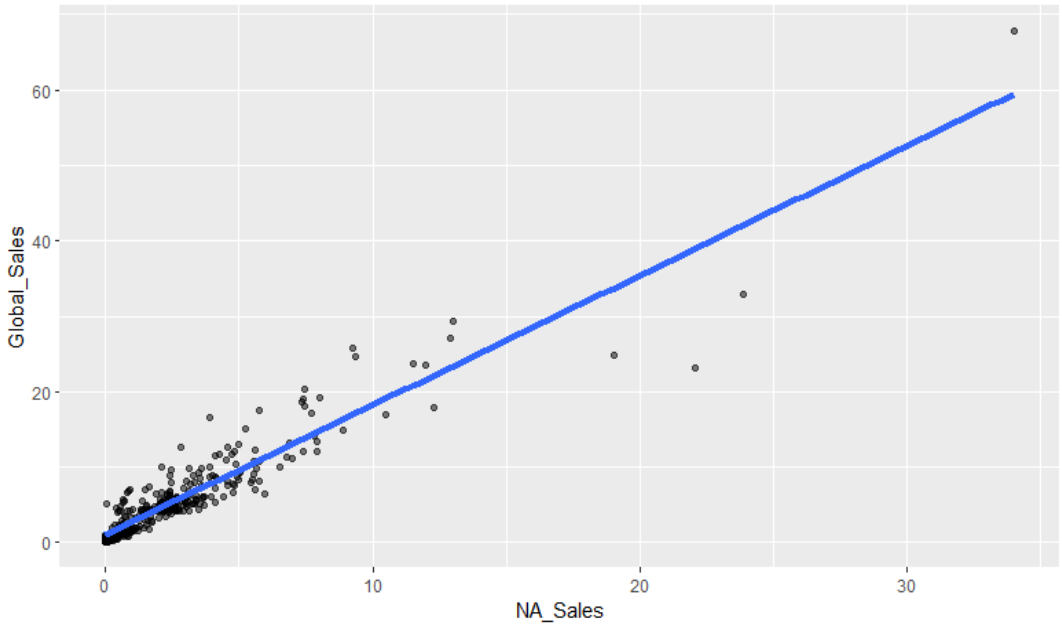
Top 20 positive reviews and summaries also have a subjectivity score close to 1 hence indicating that they are mostly opinion based.

A TextBlob was carried out on a comment “wonderful design” indicating a polarity and subjectivity score of 1 which is highly positive and opinion based.

Moreover a Document Term Matrix was generated for review and summary columns which suggested the most used document words of screen, book, adventure, pretty and art which overall suggest a positive sentiment towards the game.

Analysis of Global, North American and EU Sales

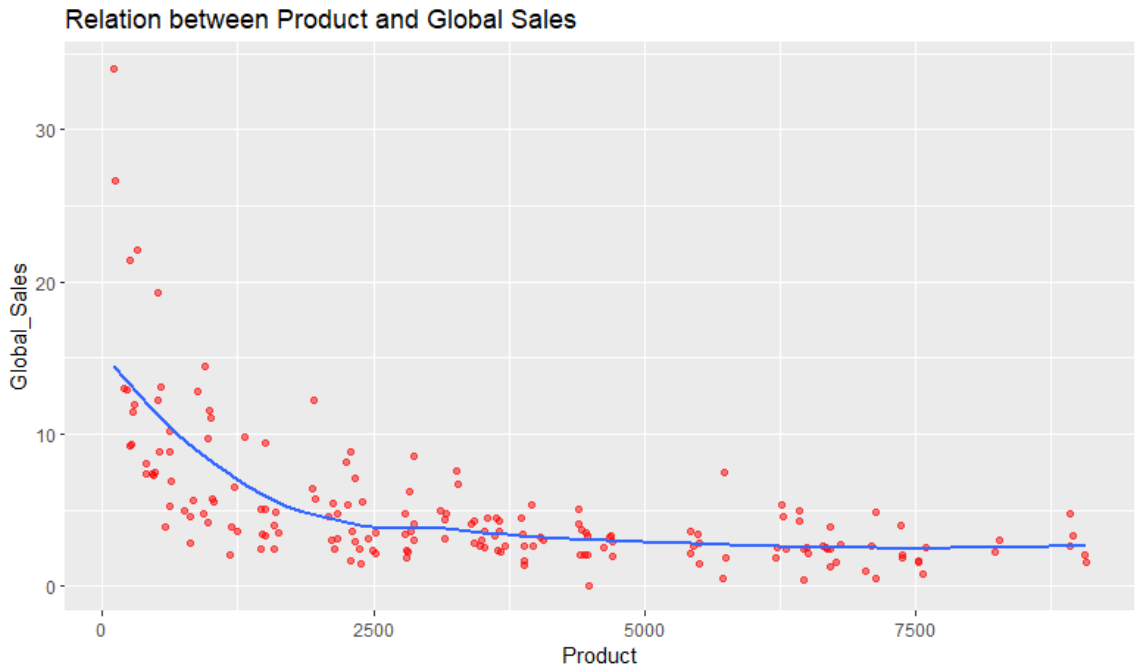
The Scatterplot displaying the relationship between NA_Sales and EU _Sales is heavily dense at the tail indicating that the data is positively skewed. There is an extreme outlier which can be observed in all the plots. The relation between Global and NA Sales however is more linear and follows a positive direction as shown in the diagrams below. The boxplot and histogram for NA and Global Sales also indicate a very skewed relation with a few outliers above 10M which should be observed.



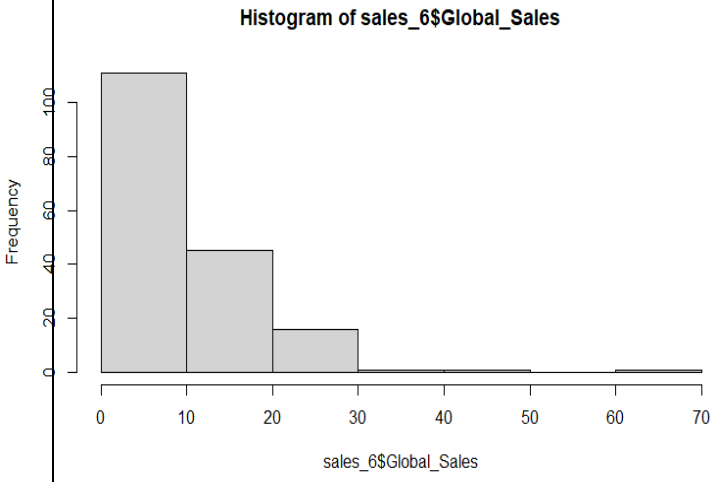
Testing the reliability of the Data

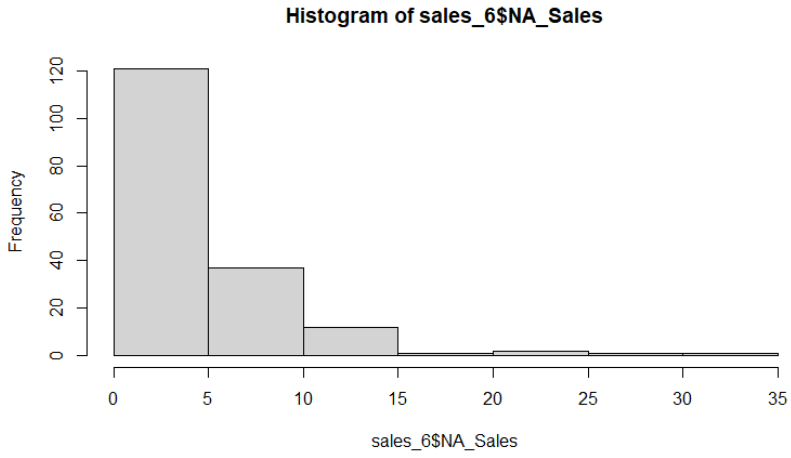
It can be observed that the mean (2.5) and maximum amount of sales (34) obtained in North America is higher compared to the EU indicating that on average **2.5 million sales** are made in NA. In addition to this, the InterQuartile Range for NA is also greater than EU with a value of 2.65. There also seems to be a greater variance of 11.6 for NA. This however indicates that **North America is a more lucrative market for Turtle Games**.

A scatterplot representing the relation between Product and Global Sales suggests that as Product ID increases, the amount of global sales decreases while most of the data points are concentrated around the smoothing line. There are however a few outliers ranging from 20M to 60M.

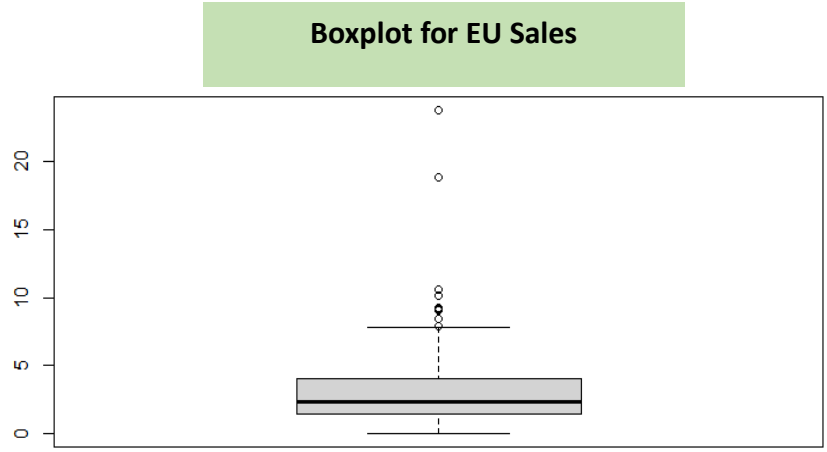


A Histogram of Global Sales represents most of the sales are concentrated around 10M followed by 20M where NA sales contribute significantly as most of the NA Sales are around 5M. The tail is concentrated at one end.





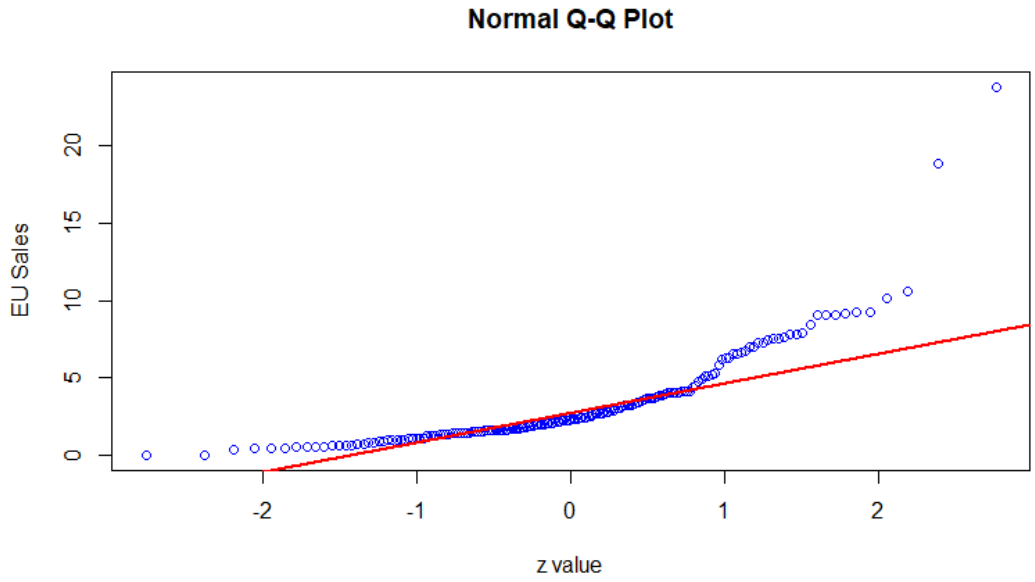
A boxplot of the sales strongly suggests that the data seems to be positively skewed since the median value lies quite below and there can be seen outliers especially in the boxplot for EU sales.



The three plots suggest that the data is positively skewed. We may need data regarding the sales being made in other regions.

The QQ Plot for Global Sales indicates that the data is normally distributed as the points are not too far from the normal line. The values in the tails of the distribution are not quite extreme.

The QQ Plot for EU Sales however indicates a number of values towards the tail end above the distribution line.

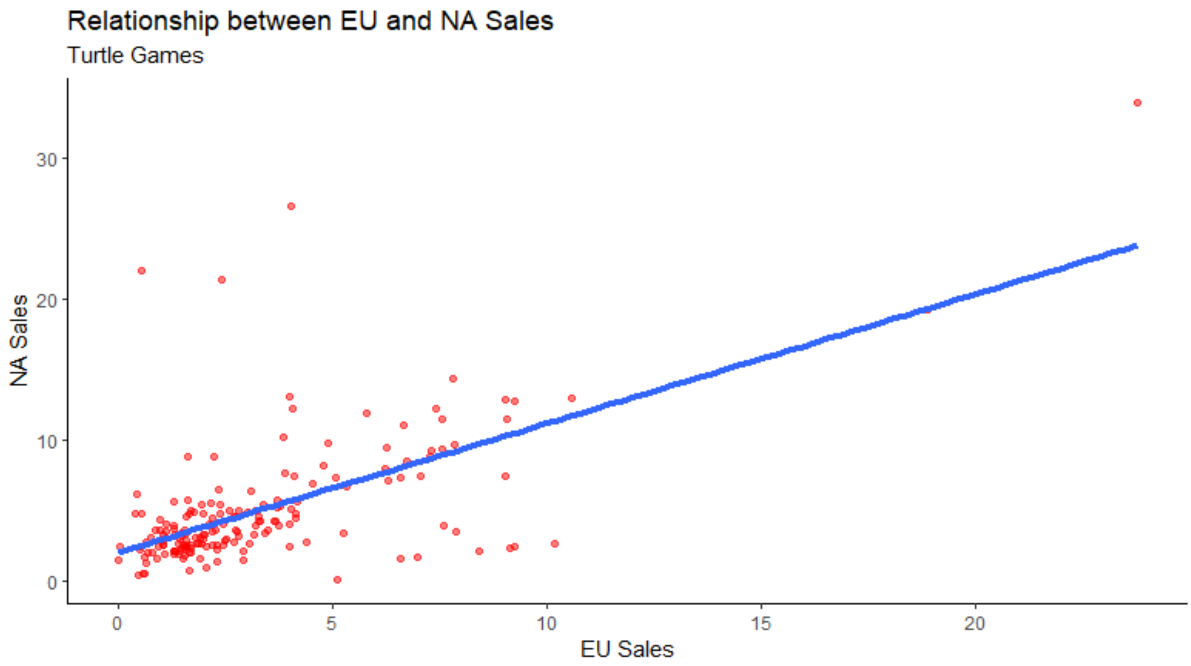


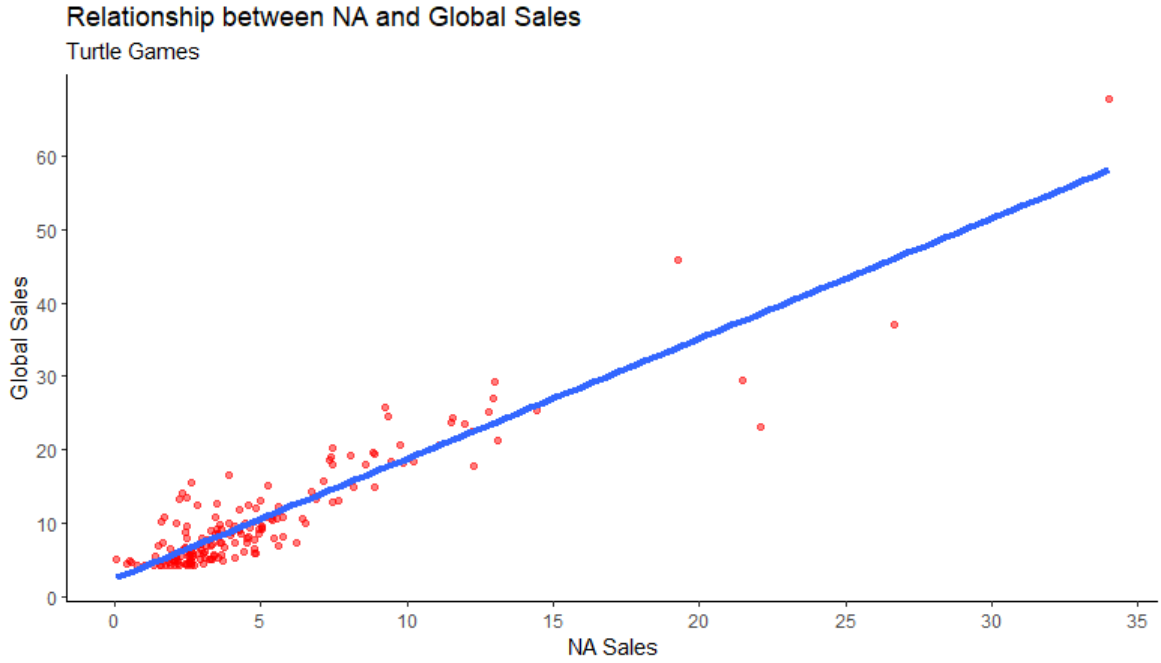
The P value is p < 2.2e-16 for the Shapiro Wilk test thereby indicating that the data is statistically significant and thus we may reject the null hypothesis if there is any.

The skewness and kurtosis of the sales data indicates that the data is highly positively skewed and a kurtosis value of around 17 indicates that the data may have a very high peak and many outliers.

The correlation coefficient among all three sales data is positive indicating that if sales value increase for NA and EU, then Global Sales will also increase. The correlation is the strongest for NA and Global sales hence indicating a strong positive relation.

Scatterplots were used to identify the relation between the sales as they are continuous variables. A positive relation was observed between NA and Global Sales where most of the values were concentrated on one end of the tail between 5M to 10M. A positive relation was also displayed for EU and NA sales indicating a linear relationship between the two.

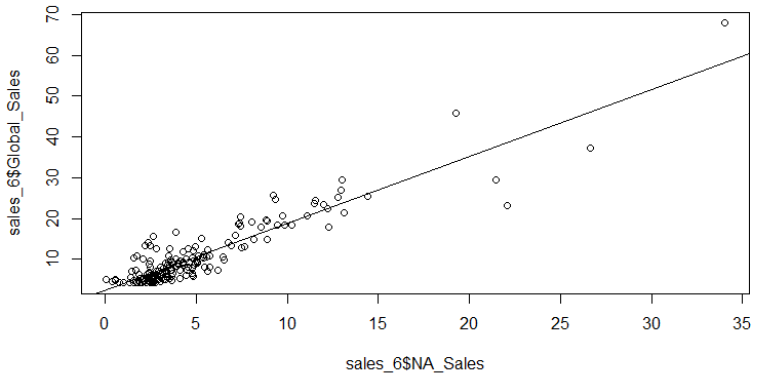




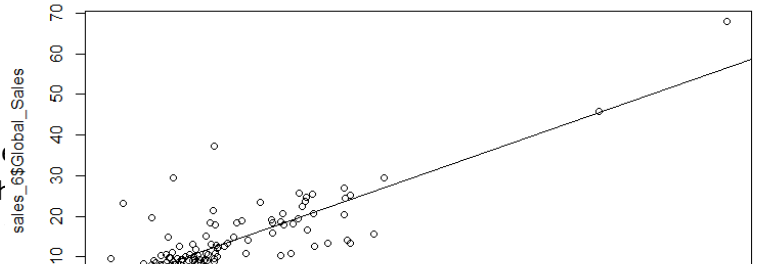
Linear Regression using R Studio

There is a strong positive correlation between NA and Global sales (0.916), followed by EU and Global Sales(0.848) and NA and EU Sales (0.6) which is slightly moderate.

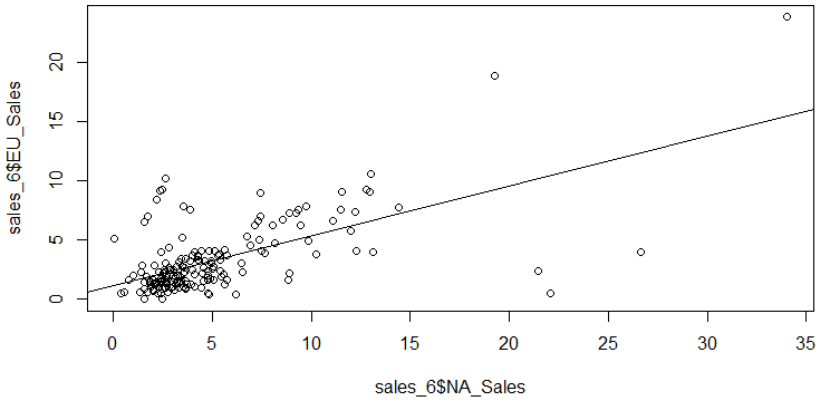
The coefficient for the relation between NA and Global Sales is 1.63 indicating that Global sales will increase by 1.63 with every increase in NA Sales. The R square value of 83.5% indicates a good fit while the p value below 0.05 indicates NA sales being highly significant.



The coefficient for the relation between EU and Global Sales is 2.23 indicating that Global sales will increase by 2.23 with every increase in EU Sales. The R square value of 72% indicates a good fit while the p value below 0.05 indicates EU sales being highly significant.

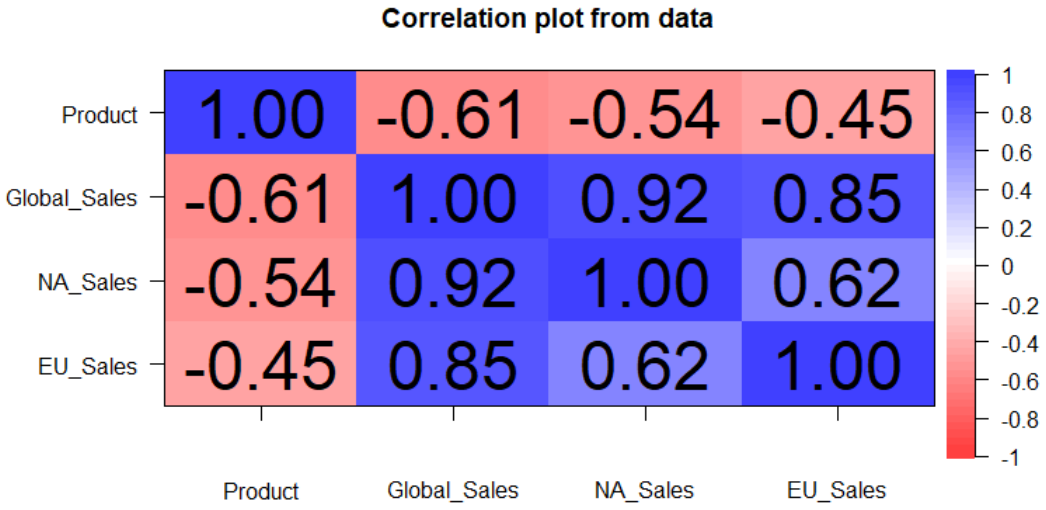


The coefficie t EU sales will increase by 0.42 with ; not indicate a good fit while the p



A multiple linear regression line was created in order to determine the relationship between Global Sales and NA and EU Sales. A correlation plot indicated a positive strong correlation between Global

and NA as well as EU Sales where the relation between Global and NA sales is stronger. However the model would have been stronger if the data was available for different regions and may have been more granular by being spread out monthly rather than annually.



Further Observations and Insights

I believe the data could have been more comprehensive in terms of Sales. There could have been more information regarding Sales from regions within NA and the EU. It is useful for Turtle Games to know which region within NA or EU is making the least and most amount of sales and hence target customers accordingly through their marketing campaigns as the data extracted is quite broad.

The extreme tail end where most of the values are concentrated between 0.01M and 1M should be focused on. It can be observed that the products published by Nintendo make the most sales especially in NA. The best selling products belong to the platform of Wii.

There are some very prominent outliers below the linear regression line in the plot comparing the relation between NA and Global Sales and above the line in the plot comparing relation between EU and Global Sales.

The Scatterplot displaying the relationship between NA_Sales and EU_Sales is heavily dense at the tail indicating that the data is positively skewed. There is an extreme outlier which can be observed in all the plots. The relation between Global and NA Sales however is more linear and follows a positive direction.