Turtle Games Analysis Report

1 Introduction

Turtle Games is a game manufacturer and retailer with a global customer base. The company manufactures and sells its own products, along with sourcing and selling products manufactured by other companies. Its product range includes books, boardgames, video games, and toys.

Turtle Games has a business objective of improving overall sales performance by utilising customer trends. This report aims to address the questions posed by Turtle Games to help them understand their customer base and the trends and reliability of their sales and reviews data.

2 Analytical Approach

The overall approach taken for the analysis includes a combination of descriptive and predictive statistics methods applied to the data provided by Turtle Games to understand the reliability of the data and to uncover the data trends.

Both Python and R was used for data analysis and visualizations. The Python code was written inside of a Jupyter Notebook for ease of use. A combination of statistical methods were utilized in the analysis including linear regression, k-means clustering, NLP (natural language processing) and EDA (exploratory data analysis).

2.1 Data Collection and Preparation

The data files were imported into Python using the pandas library and into R using base R functions to read the CSV files into R data frames. The data sets were checked for missing values and columns names were changed for consistency and ease of use. The customer reviews data was cleaned up for NLP by changing all letters to lower case and punctuation and duplicate entries were removed for accuracy.

2.2 Data Exploration

When exploring the data, key statistics such as unique data counts for categorical values and percentiles for numerical values were shown. The frequency of categorical values were inspected to understand the data accuracy and range. Combinations of related data points were visualized using scatter plots and pair plots to show distribution of variables. Clusters in similar data points were identified and tagged using k-means clustering.

To check anomalies in customer reviews data the counts of duplicate reviews were analyzed and most frequent duplicate reviews were investigated further. The cleaned data was tokenized using NLP and frequencies of the words were visualized using word clouds.

2.3 Predictive Modelling

Linear regression was used to understand how customers accumulate loyalty points. This was done using statsmodels library's ols method in Python.

Groups within the customer base were identified using k-means clustering via KMeans method of the sklearn library in Python.

Relations between sales data in individual regions were visualized with linear regression fit lines in scatterplots using ggplot2 in R via the use of 'lm' method to draw trendlines. Additionally, regression models of the same were created using the 'lm' method inside the stats package in R and predictions were forecasted using the created models.

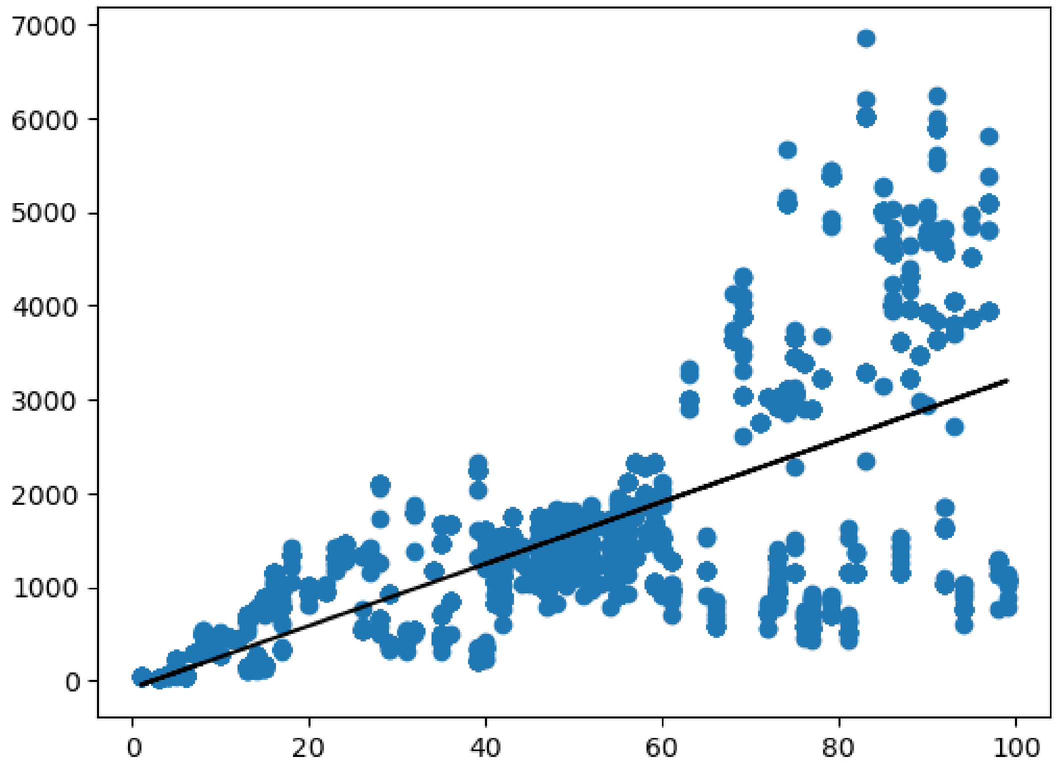
2.4 Reproducibility and Documentation

Markdown cells were utilized to explain each block of code in the Jupyter Notebook. The R code was explained using comment blocks for each important section of the analysis. To improve reusability variable names were selected carefully to allow reruns of data cleanup code without impacting other sections of the code.

3 Visualisations and Insights

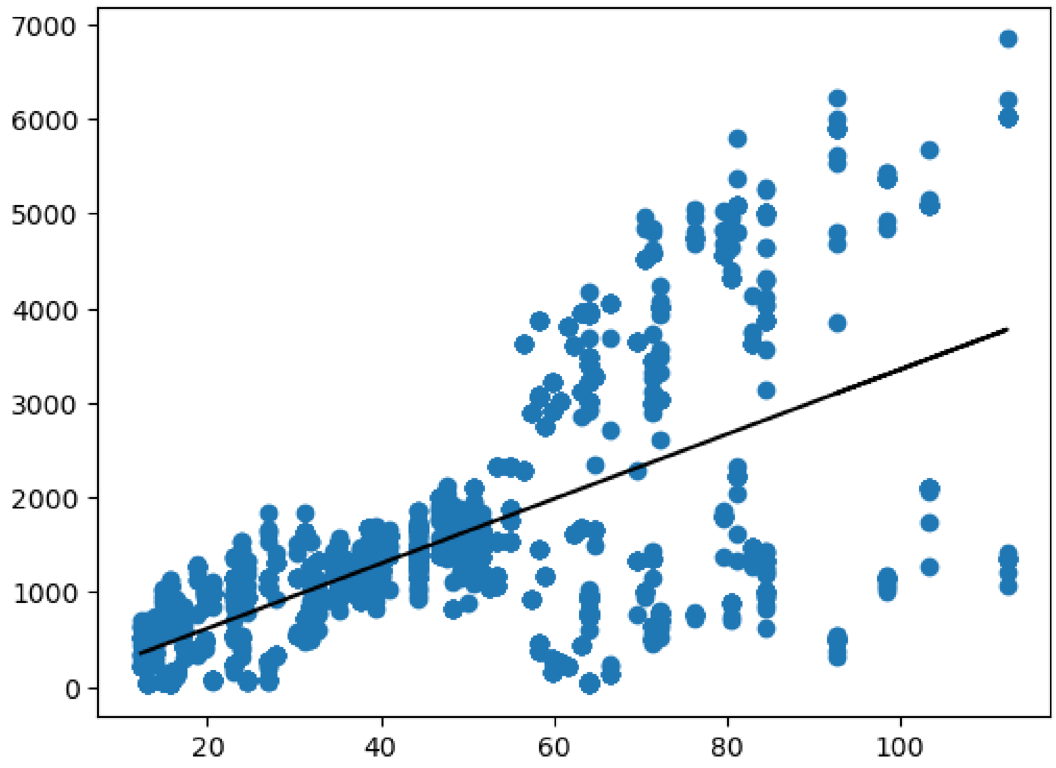
3.1 Customers loyalty points accumulation

Spending vs loyalty points



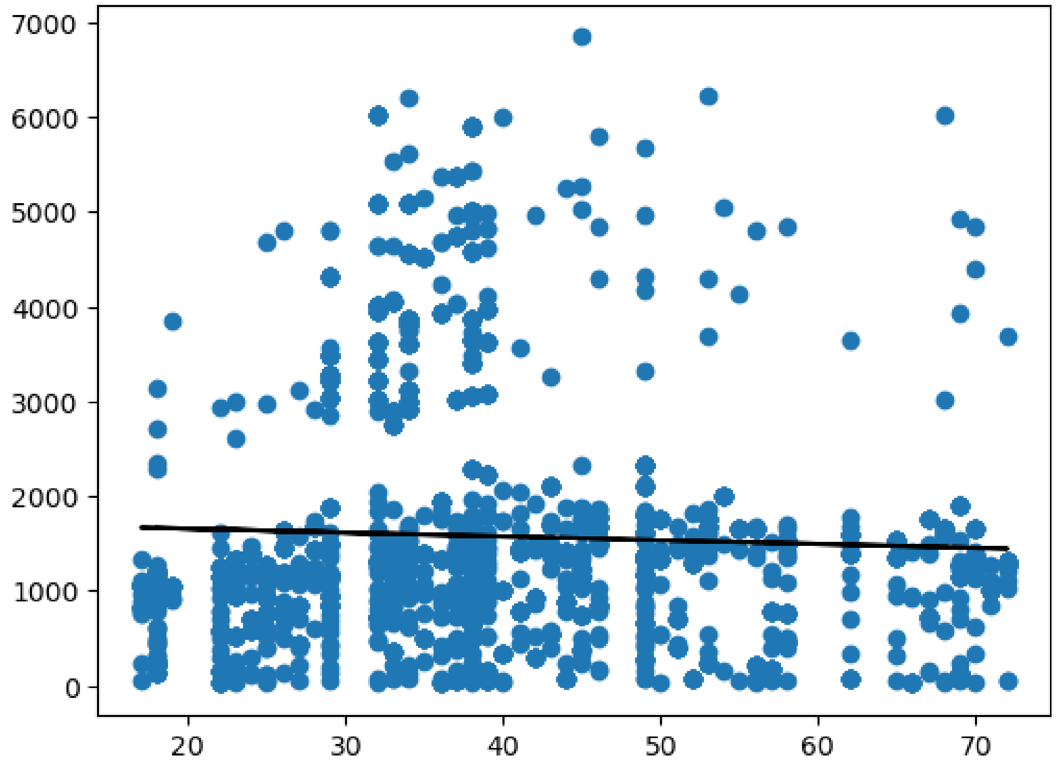
Customers accumulate loyalty points at a rate of 1 point per 33 USD spent.

Remuneration vs loyalty points



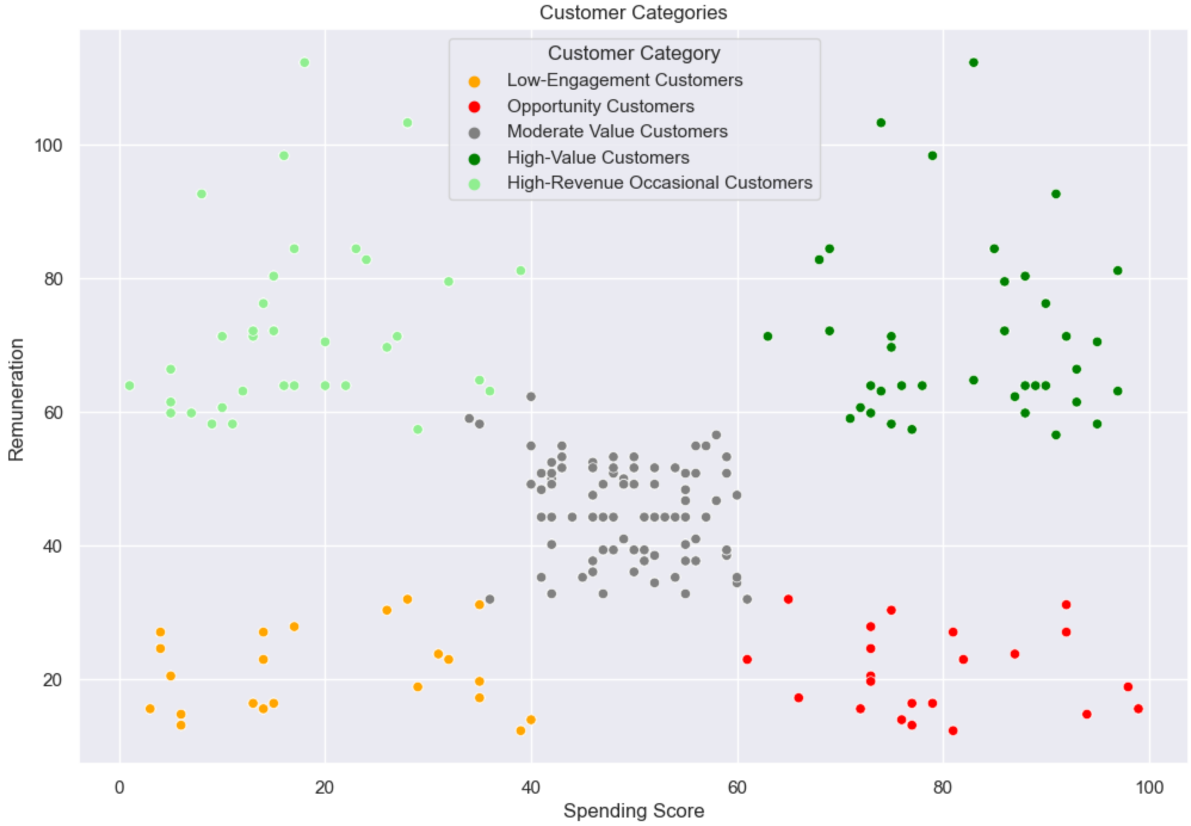
Turtle Games remunerates at a rate of 34.1 USD per customer loyalty points.

Age vs loyalty points



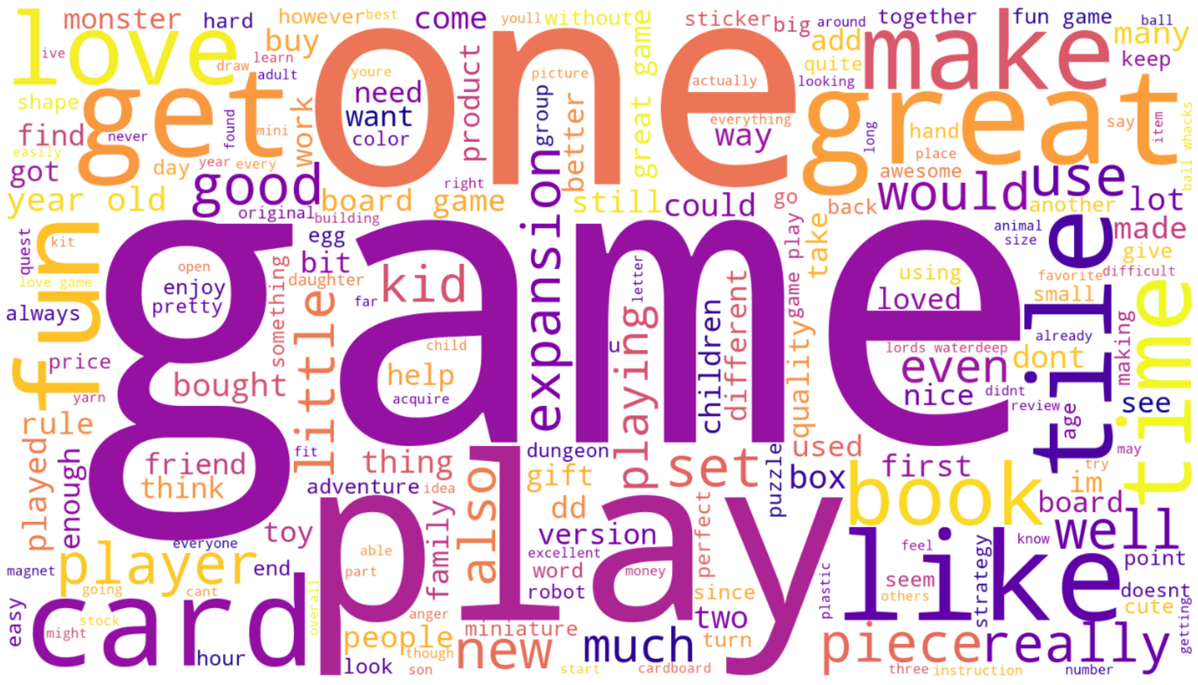
No significant correlation between age and loyalty points.

3.2 GROUPS WITHIN THE CUSTOMER BASE

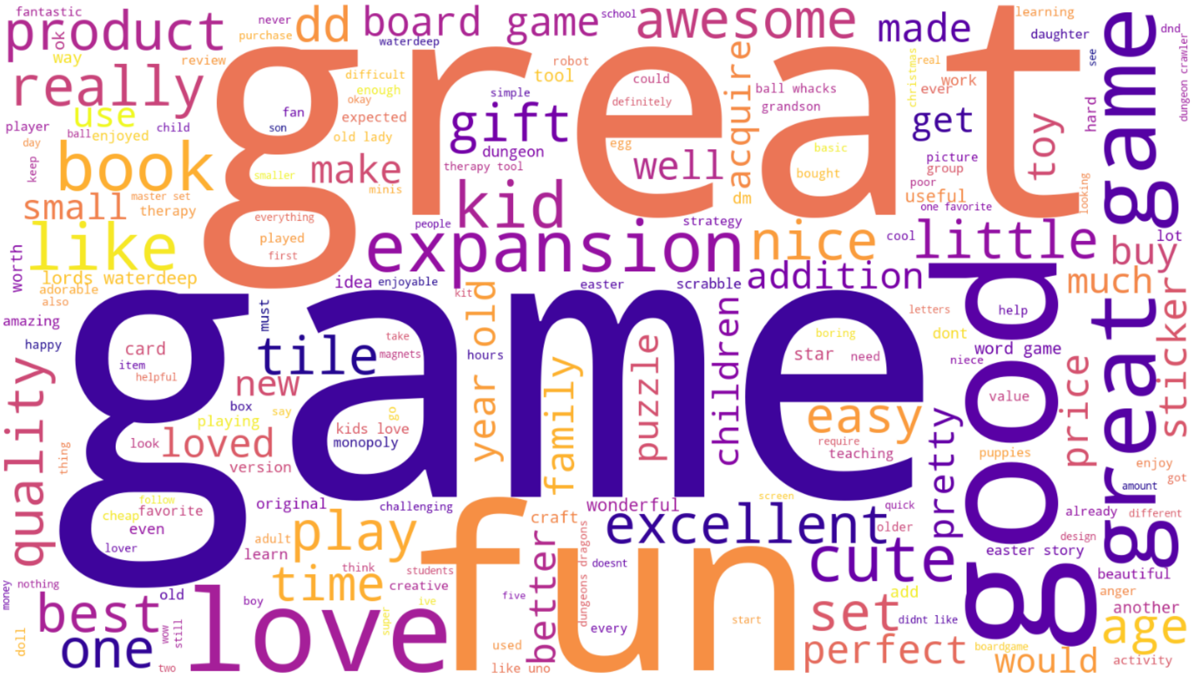


3.3 SOCIAL DATA (CUSTOMER REVIEWS)

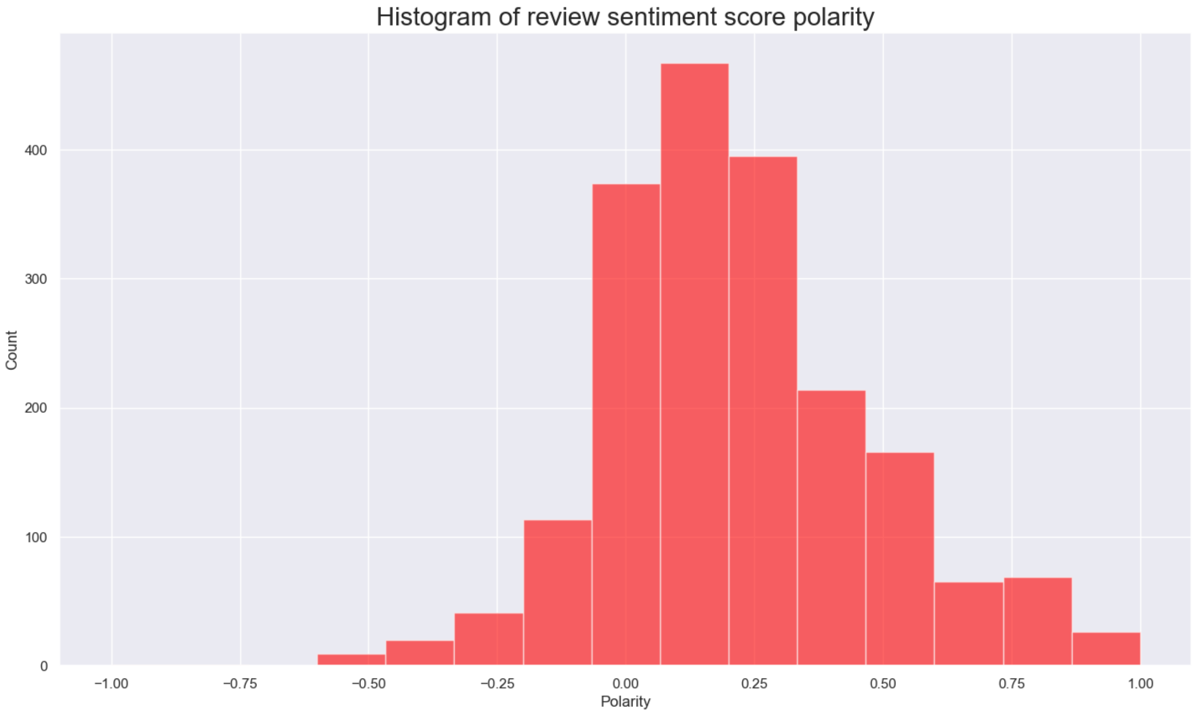
Reviews word cloud

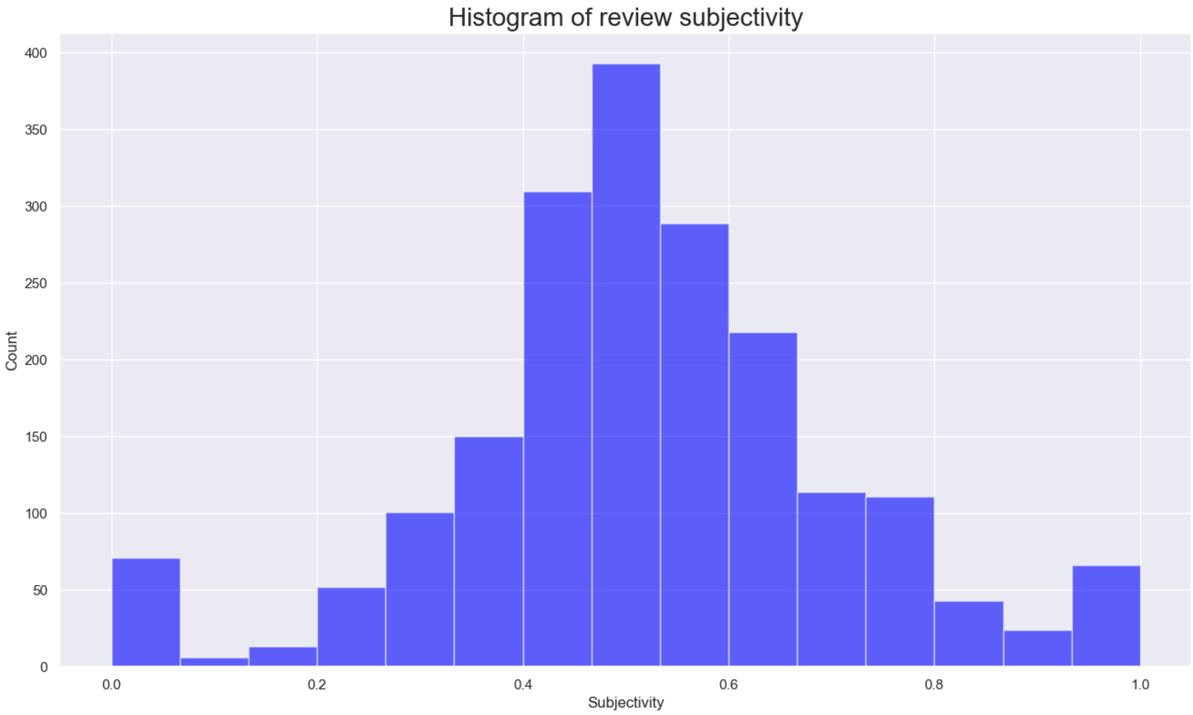


Summaries word cloud

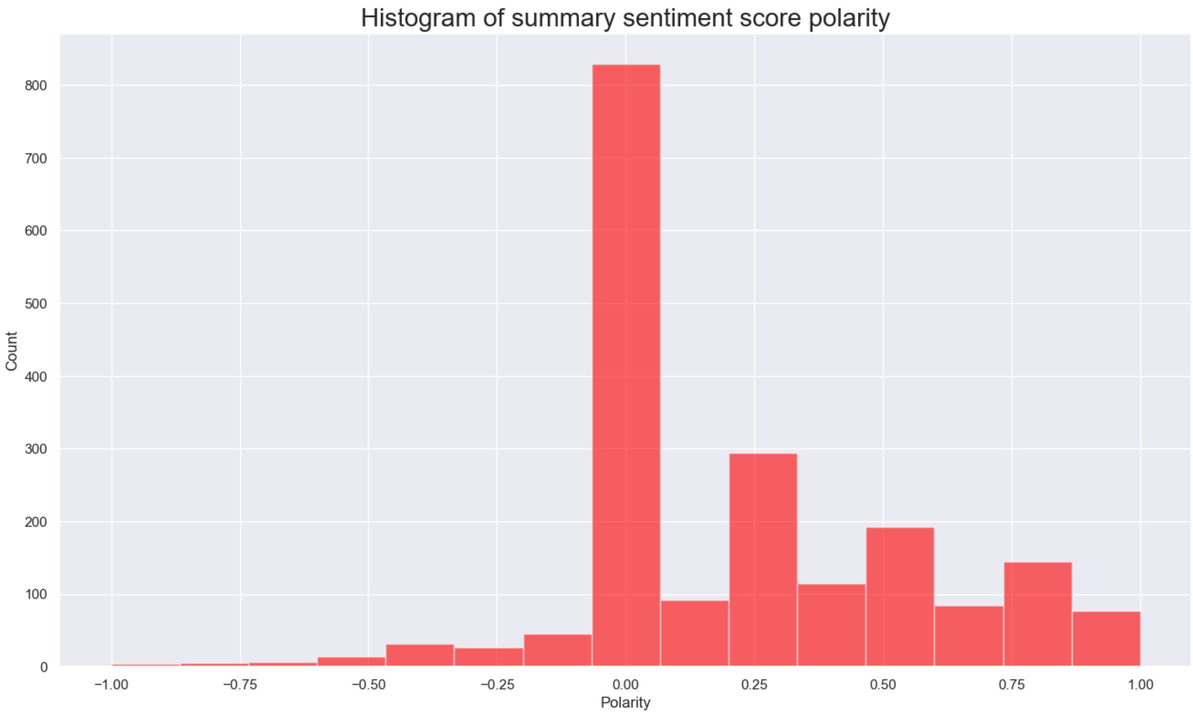


Polarity and subjectivity

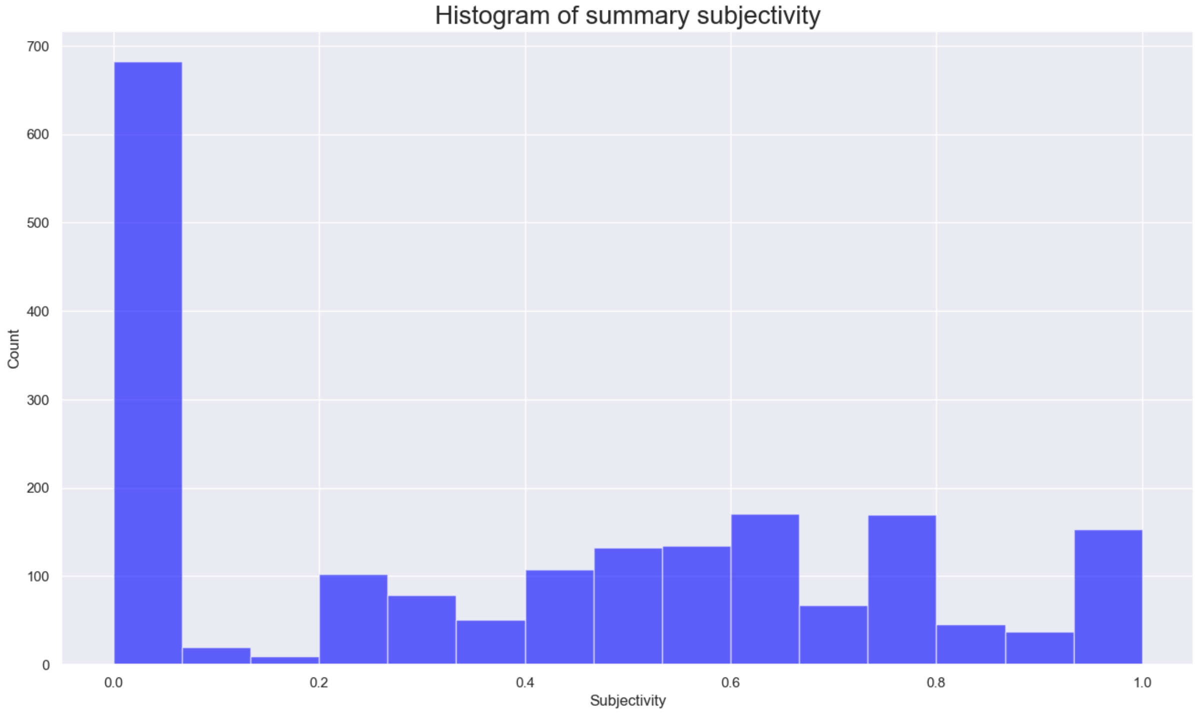




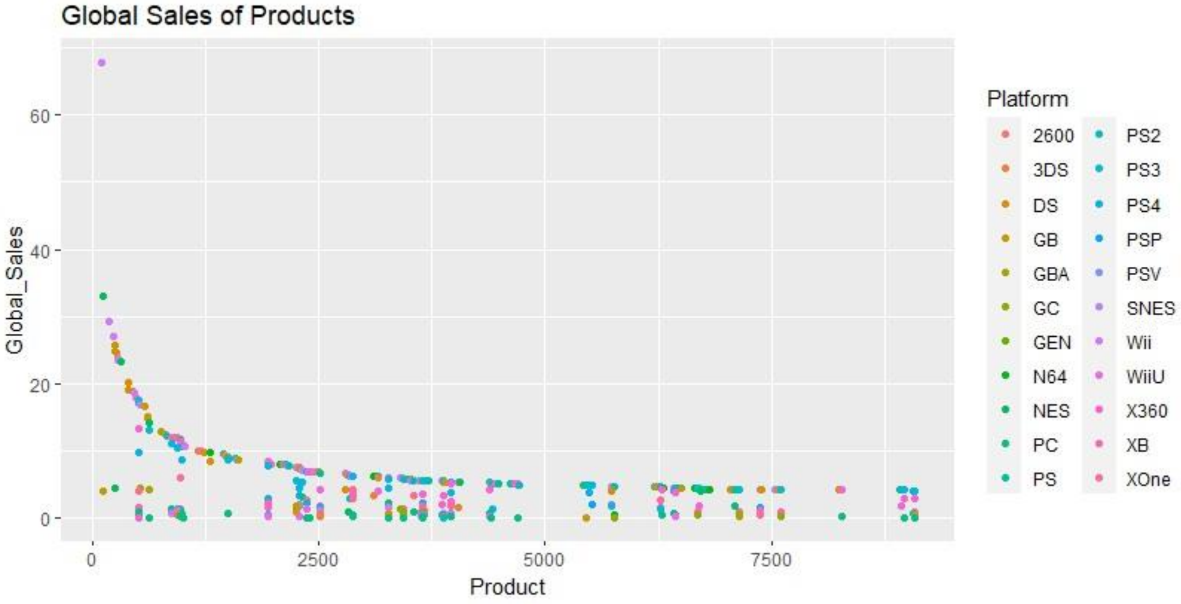
Histogram of review subjectivity

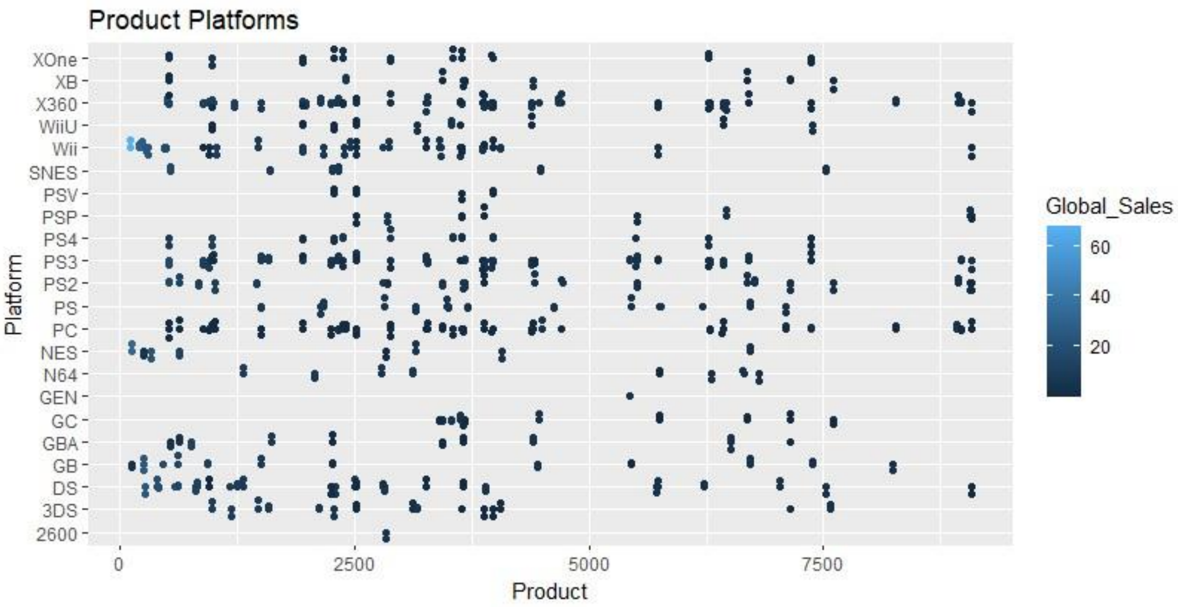


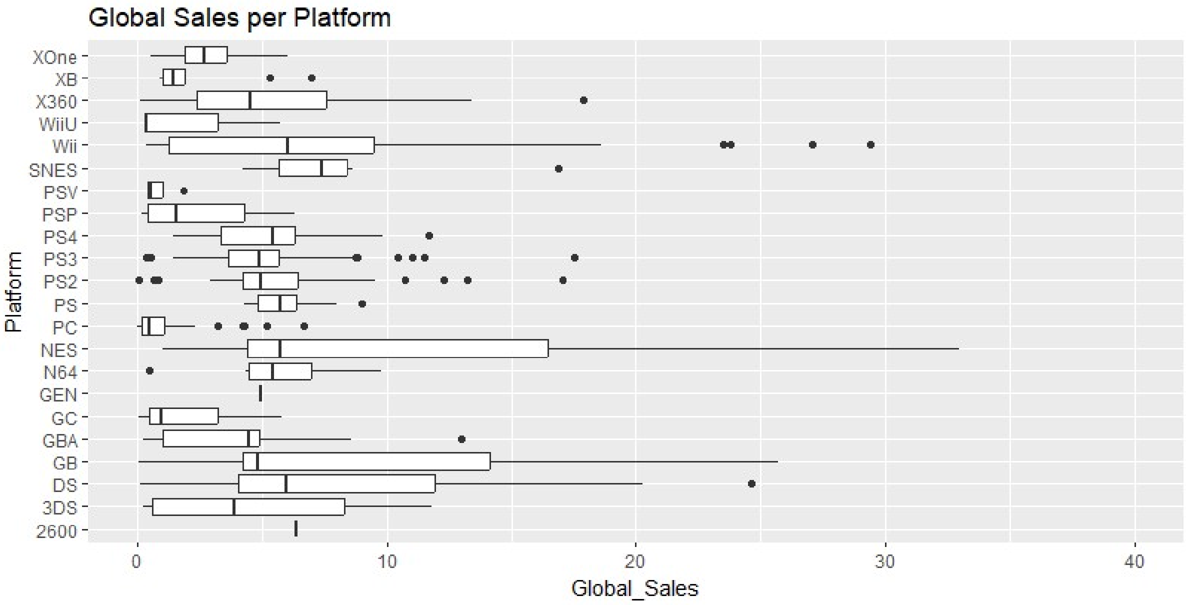
Histogram of summary sentiment score polarity

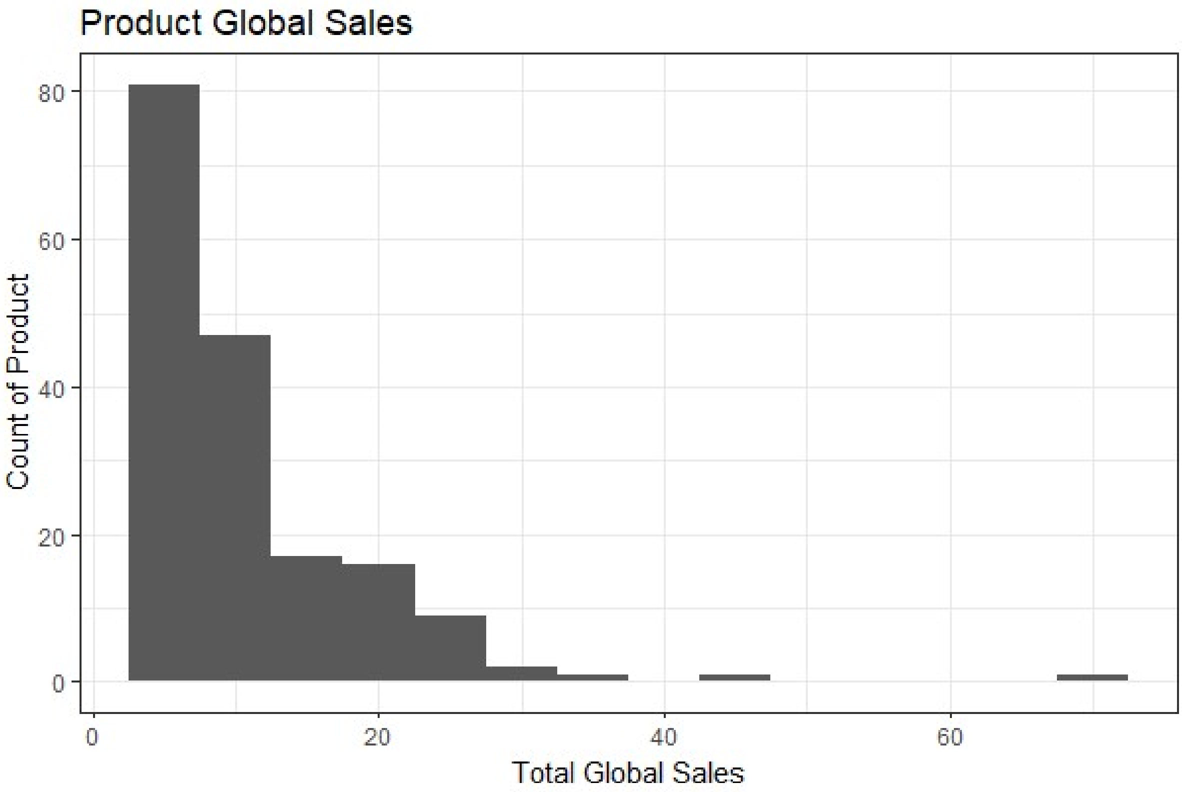


3.4 The impact of each product on sales

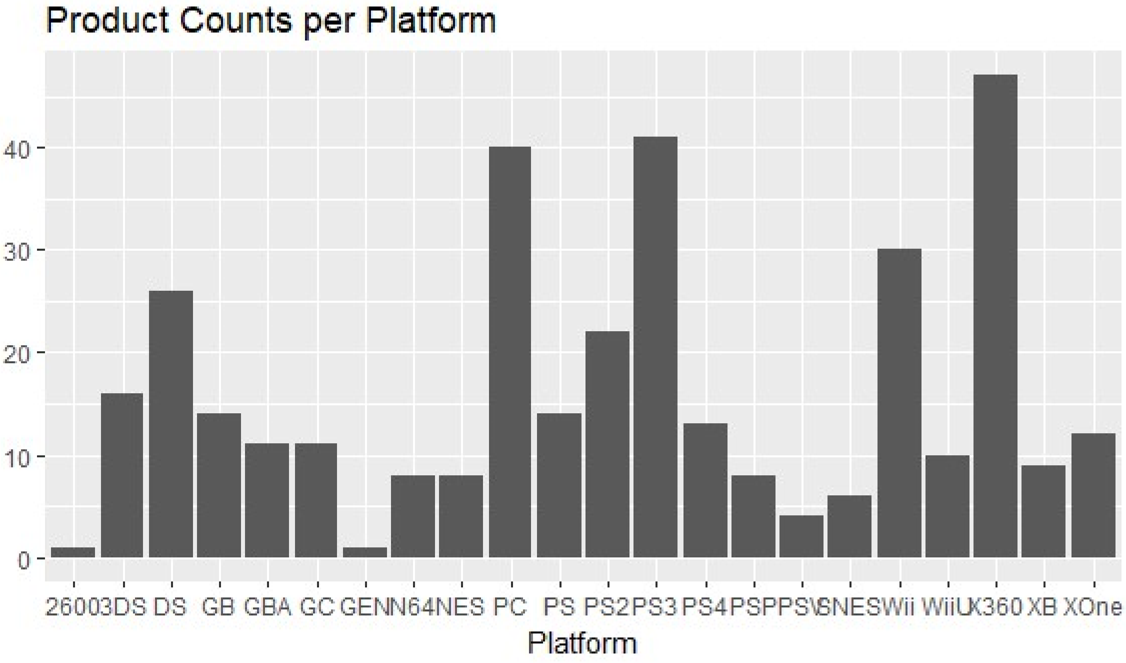



Product Platforms
Platform
XOne
XB
X360
WiiU
Wii
SNES
PSV
PSP
PS4
PS3
PS2
PS
PC
NES
N64
GEN
GC
GBA
GB
DS
3DS
2600
0
2500
5000
7500
Product
Global_Sales
60
40
20

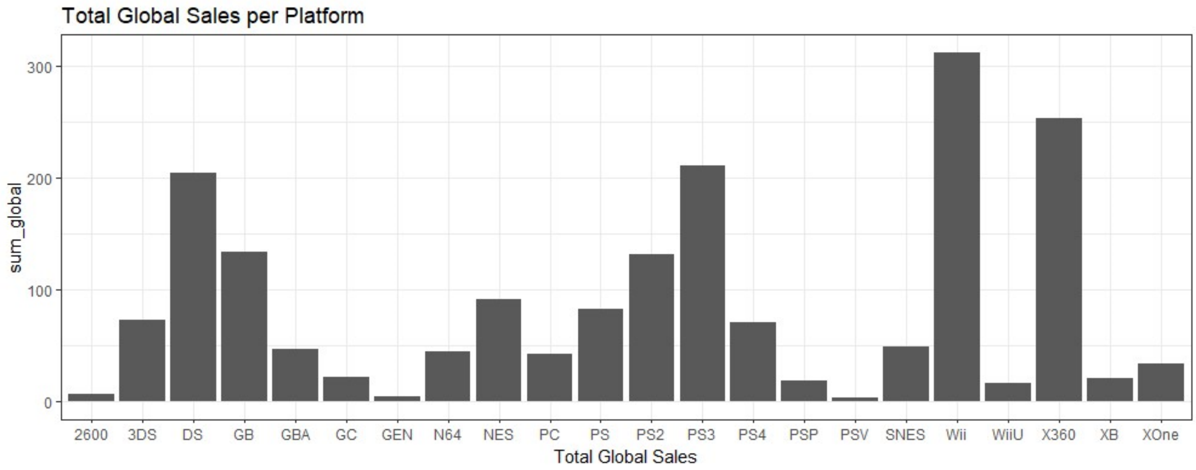

Global Sales per Platform
Platform
XOne
XB
X360
WiiU
Wii
SNES
PSV
PSP
PS4
PS3
PS2
PS
PC
NES
N64
GEN
GC
GBA
GB
DS
3DS
2600
0
10
20
30
40
Global_Sales

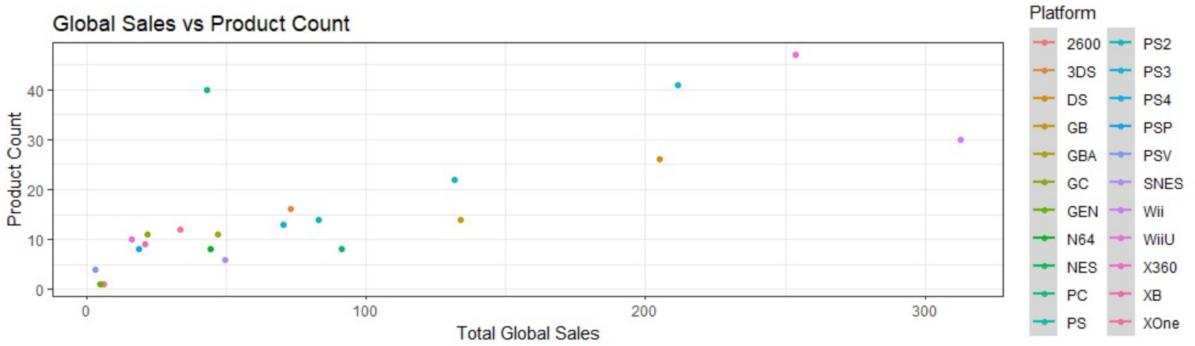


Product Global Sales



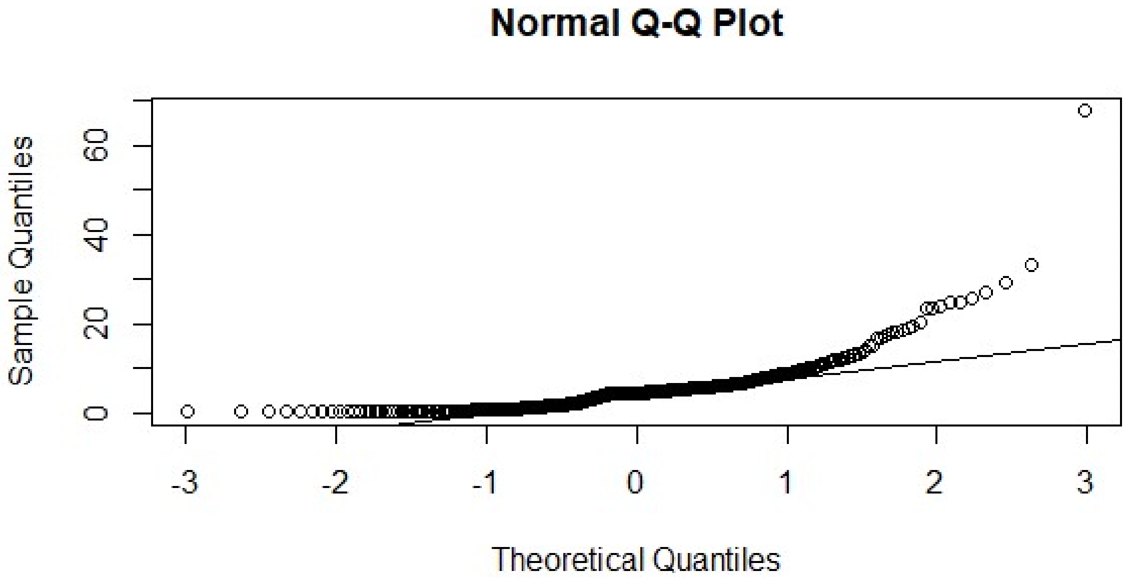
Product Counts per Platform





3.5 Data reliability

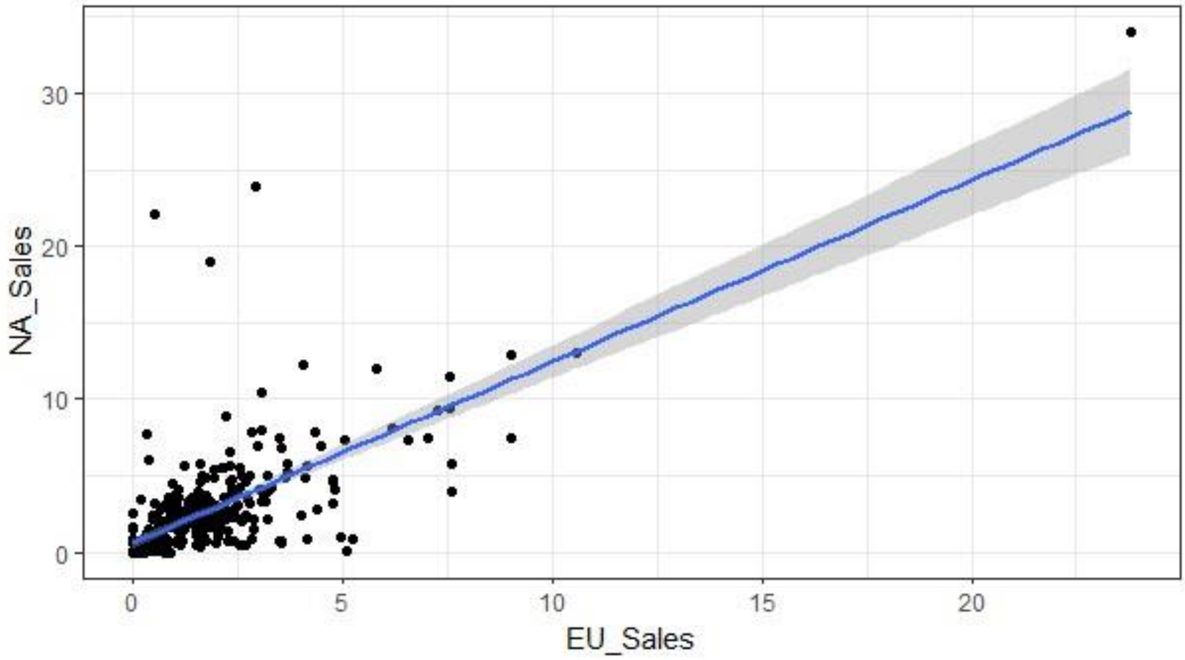
Global sales data normality

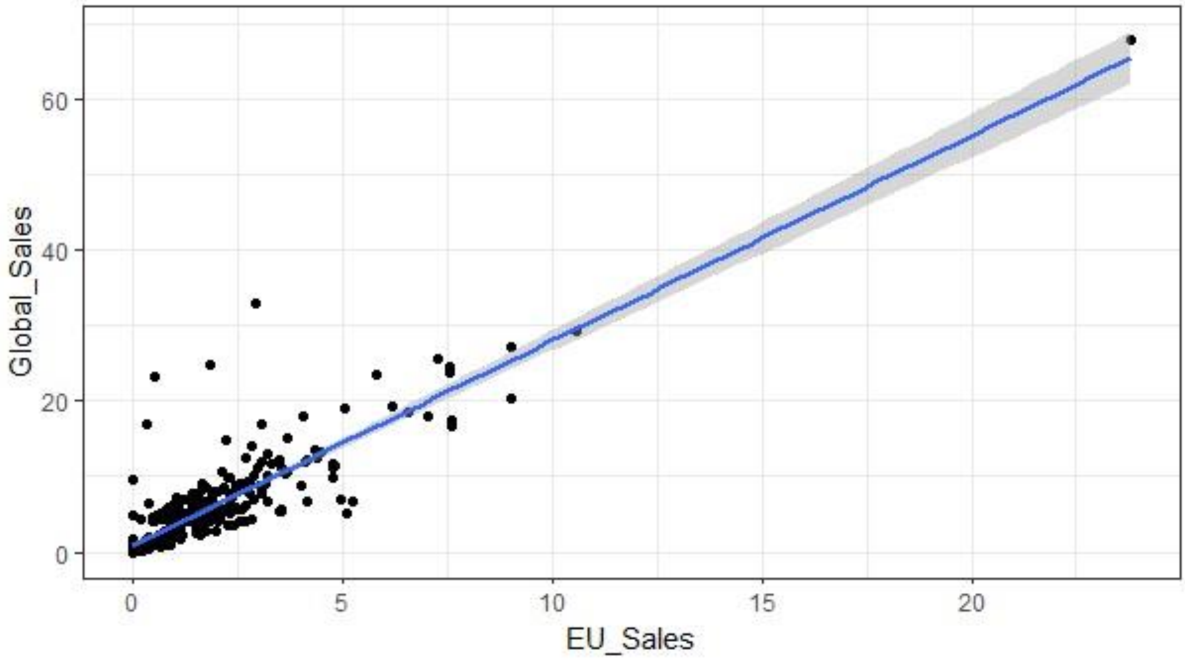


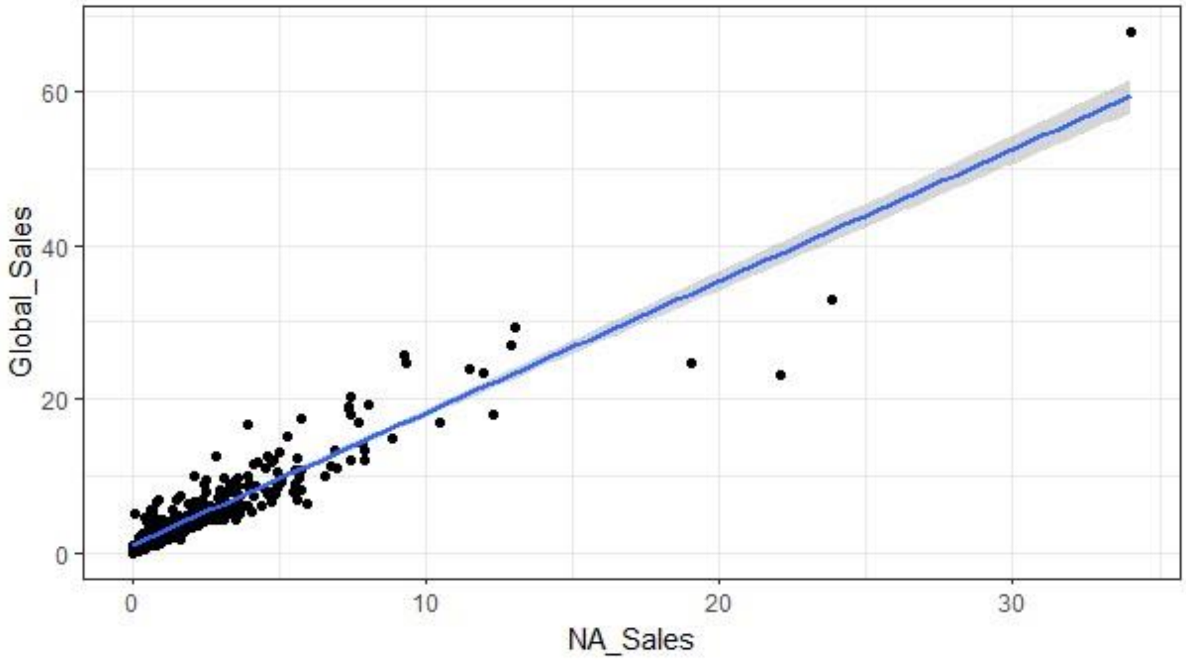
```
	Shapiro-Wilk normality test

data:  ts2$Global_Sales
W = 0.6818, p-value < 2.2e-16
```

Correlation of numeric variables in sales data set

```
             Product NA_Sales EU_Sales Global_Sales
Product         1.00    -0.40    -0.39        -0.44
NA_Sales       -0.40     1.00     0.71         0.93
EU_Sales       -0.39     0.71     1.00         0.88
Global_Sales   -0.44     0.93     0.88         1.00
```

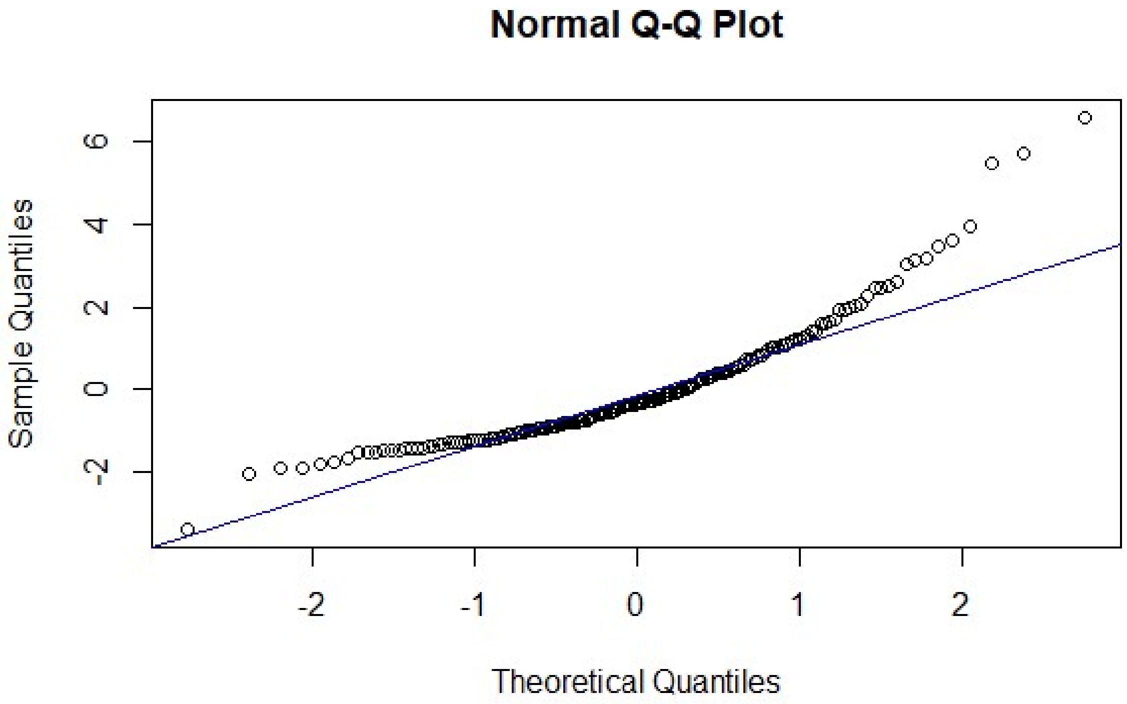




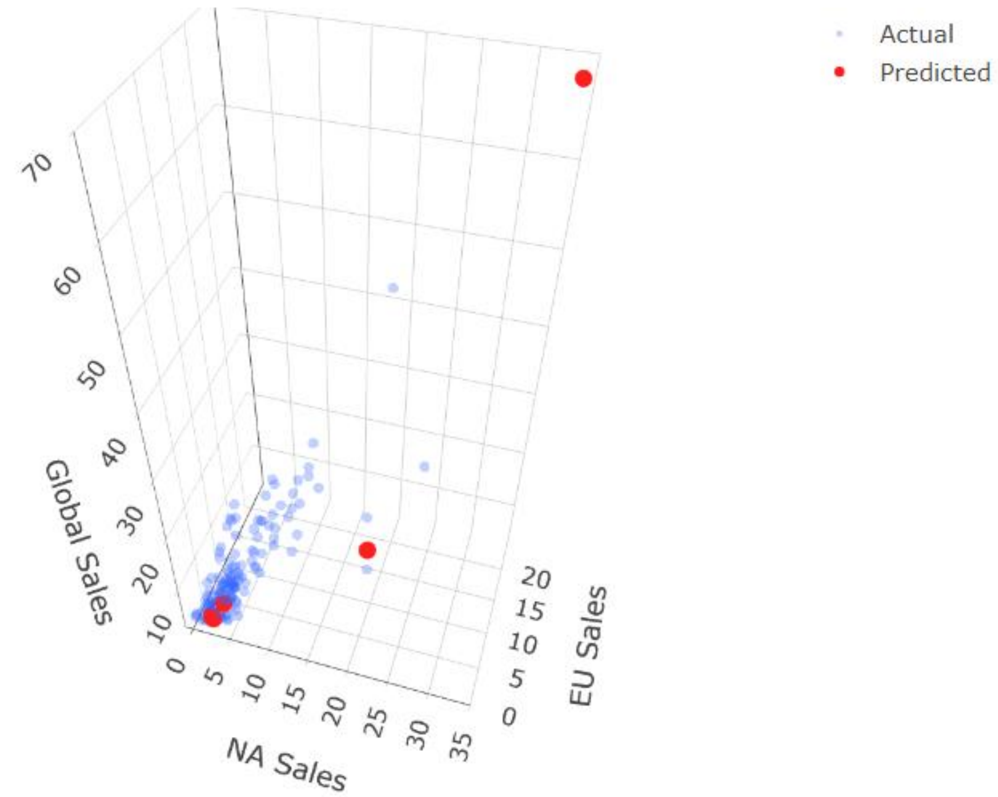


3.6 Relationship(s) between North American, European, and global sales

Q-Q plot for multiple linear regression model with global sales as the dependent variable and North American and European sales as independent variables



Forecasting using the above model



4 PATTERNS AND PREDICTIONS

1. The data reveals strong correlations between spending score, remuneration, and loyalty points, while the age-loyalty points correlation is insignificant. The linear regression models for spending score and remuneration explain 45.2% and 38% of the data's variation, respectively, with statistically significant slopes. Customers accumulate loyalty points at a rate of 1 point per 33 USD spent while remuneration is at a rate of 34.1 USD per customer loyalty points.
2. The clustering analysis reveals five customer groups: High-Value Customers, Opportunity Customers, Low-Engagement Customers, and Moderate Value Customers.
3. Customer reviews are positive and subjective, while review summaries are neutral and objective. Word clouds show that the products are popular with kids and adults alike with most reviews using positive words to describe products such as “great” and “fun”.
4. Wii has an outlier product with very high sales. PC, GC, and XOne have many products with low sales. NA sales make up most of the global sales number, and SNES platform products have the highest mean sales globally.
5. Sales data is not normally distributed and the data has a heavy tail. EU and NA sales are heavily correlated with global sales, as expected. The “product” column is unexpectedly correlated with the sales columns, which suggests that the products might have been partially numbered based on their sales performance. EU and NA sales have a significant correlation between them, which suggests that sales performance of products across these continents are similar.

6. A multiple linear regression model created using the global sales as the dependent variable and North American and European sales as independent variables can explain 96.64% of the variance in the dependent variable. This is a very good fit and suggests that global sales can be predicted reasonably well using only NA and EU sales data.

5 RECOMMENDATIONS

1. Every loyalty point that customers accumulate translates to a 1.1 USD of profit. Benefits awarded by loyalty points should be adjusted to cost the company less than 1.1 USD to be profitable.
2. This identified customer groups can be used to target specific market segments with tailored marketing strategies. For example, Opportunity Customers could be targeted with campaigns to increase their engagement and remuneration, while High-Revenue Occasional Customers could be targeted with retention campaigns.
3. Customer reviews can be used to inform marketing campaigns by highlighting positive aspects, identifying areas for improvement, targeting specific market segments, and building trust and credibility.
4. PC platform has many products but poor global sales performance. Turtle Games can consider reducing the number of PC products to focus on the most popular and profitable ones to simplify inventory management and make it easier for customers to find the products they are looking for.
5. The strong correlation between EU and NA sales data suggests that sales performance for products is similar across these two regions. However, it is still important to analyse sales data by smaller regions to identify any trends or differences that may exist. This information can be used to develop more targeted marketing and sales strategies.
6. Turtle Games can use the multiple linear regression model to predict global sales performance using sales data from North America and Europe.