LSE_DA301 Advanced Analytics for Organisational Impact: Final Assignment

Turtle Games:
Predicting Future Outcomes

Ryan Lin

Word Count: 1264

Introduction & Background

This report and the resulting findings are derived from two data sets provided by Turtle Games, which includes information about their sales data and customer metrics.

Dataset	Dataset Contents
Reviews turtle_reviews.csv	**Details of customer demographics and product reviews**: • Gender, age, remuneration, spending score, loyalty points, education, language, platform, review and summary across products.
Sales turtle_sales.csv	**Details of video games sold globally:** • Rank, product, platform, genre, publisher, and their sales across North America, Europe, and worldwide

Note: See meta data file titled metadata_turtle_games.txt for more information

Turtle Games are keen to develop a data-driven approach to determine how to improve overall sales performance by utilising customer trends (business problem) by exploring:

1. How customers accumulate loyalty points
2. How groups within the customer base can be used to target specific market segments
3. How customer reviews can be used to inform marketing campaigns
4. The impact that each product has on sales
5. How reliable the data is (normal distribution, skewness, or kurtosis)
6. The relationship between North American, European, and global sales

A variety of questions needs to be asked to better understand the project and the data provided. Additional areas to explore outside the scope of this project are also listed (next page).

Questions Related to This Project

- Who are the key decision makers and main internal & external stakeholders involved? Is it the CEO, Sales and Marketing Team?
- Who will the finding and insights be presented and shared with?
- What does success look like from what we can determine regarding this analysis?
- Who will be implementing the recommendations based of the resulting finding?
- What is the reason for trying to establish a relationship between NA, EU and Global Sales?

Questions Related to The Data

- Why isn't the individual product items included in the data set and only the product ID?
- Is there any relevance in exploring and analysing old games or games sold on old, obsolete devices?
- Global Sales ≠ NA + EU Sales. Are the remaining sales Rest-of-the-world or another market? Why isn't this data included?

Additional Areas to Explore

- Exploring the relationship between marketing spend and sales.
- Exploring the sentiment of reviews specific to individual products

Analytical Approach

Setting up GitHub Repository

A GitHub Repository was set up to host the Jupyter Notebook and R files as well as related documents used to complete the analysis.

The first half of the analysis on the turtle_reviews.csv data set was conducted in Python using Jupyter Notebook, while the second half was done on the turtle_sales.csv conducted in R and R Studio.

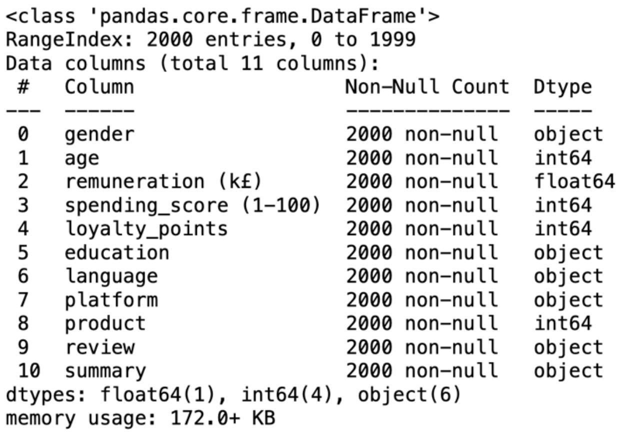
It is assumed the datasets imported in this analysis were all cleaned, sense checked for data types, missing values and duplicates. The metadata descriptive statistics were also explored, see the Jupyter notebook and R files for more details. The various libraries used are seen at the beginning of each document such as in Python numpy, pandas, matplotlib, seaborn, scipy, statsmodels, and scikit-learn, and in R packages, such as tidyverse, dplyr, ggplot2, which are essential for data analysis, visualisation, and modeling.

Customer Data (Python)

How Users Accumulate Loyalty Points

Importing and Exploring Data

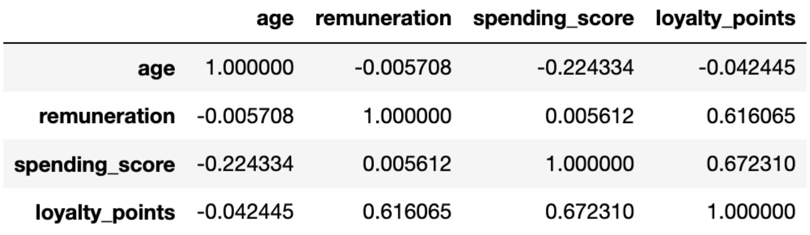
The provided customer demographic and review data (turtle_reviews.csv) was sense-checked and metadata descriptive statistics was explored, the column names are also shown below:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 11 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   gender                  2000 non-null   object 
 1   age                     2000 non-null   int64  
 2   remuneration (k£)       2000 non-null   float64
 3   spending_score (1-100)  2000 non-null   int64  
 4   loyalty_points          2000 non-null   int64  
 5   education               2000 non-null   object 
 6   language                2000 non-null   object 
 7   platform                2000 non-null   object 
 8   product                 2000 non-null   int64  
 9   review                  2000 non-null   object 
 10  summary                 2000 non-null   object 
dtypes: float64(1), int64(4), object(6)
memory usage: 172.0+ KB
```

No missing values or duplicates were found. Redundant columns were dropped, and the column headings were modified for ease of reference. A copy of the cleaned DataFrame was saved to csv, imported and sense-checked and used for the rest of the analysis.

A multiple linear regression model using statsmodels functions was created to evaluate the possible linear relationships between loyalty points and age/remuneration/spending scores to determine whether these can be used to predict the loyalty points. The correlation coefficients of the variables in question are seen below:



	age	remuneration	spending_score	loyalty_points
age	1.000000	-0.005708	-0.224334	-0.042445
remuneration	-0.005708	1.000000	0.005612	0.616065
spending_score	-0.224334	0.005612	1.000000	0.672310
loyalty_points	-0.042445	0.616065	0.672310	1.000000

Any relevant or significant findings are discussed:

Age and Remuneration (-0.0057.): Very weak, almost negligible negative correlation. This might imply that age has little influence on how much a person is paid in this context.

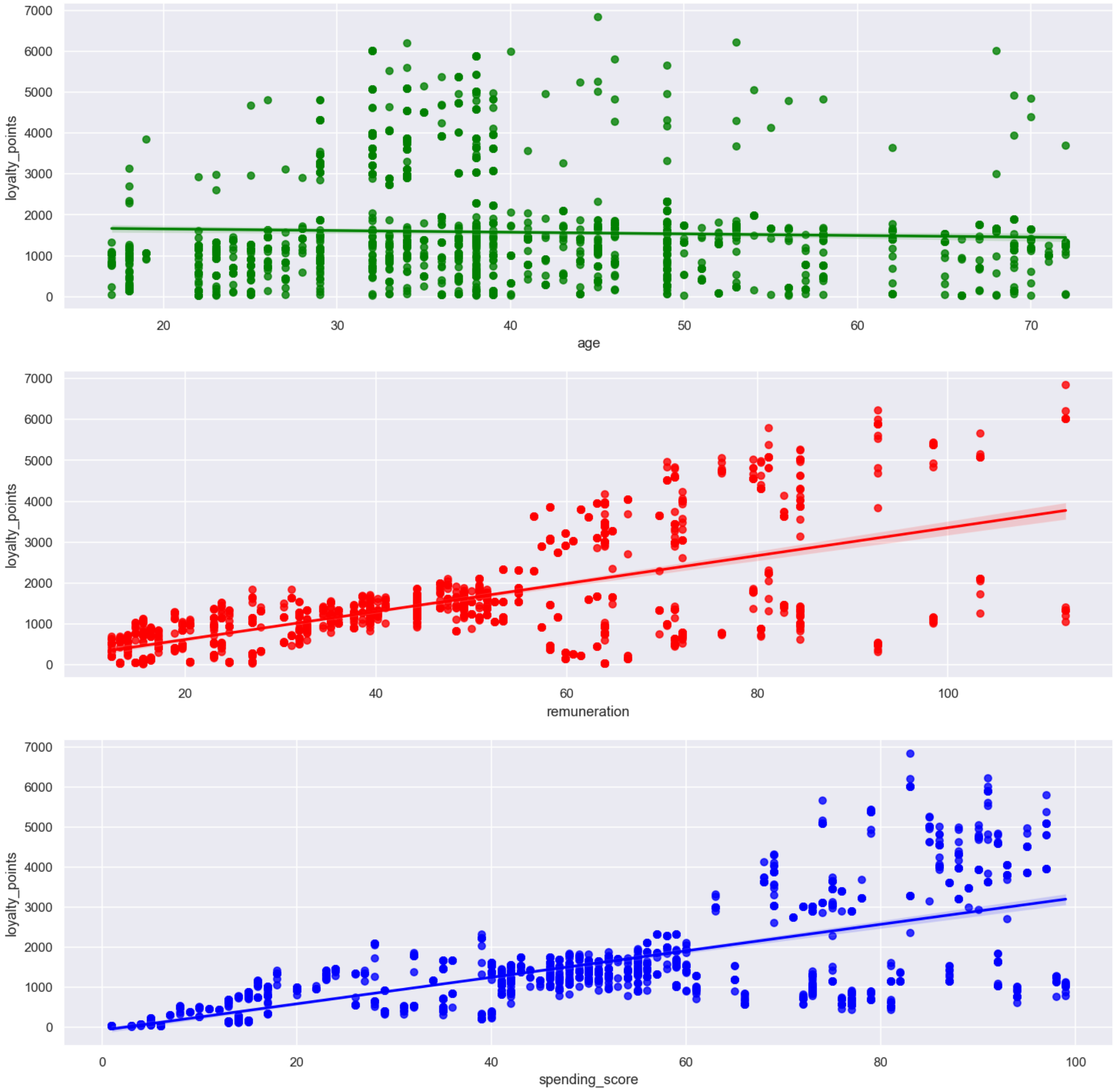
Remuneration and Spending Score (0.0056): Very weak positive correlation. Earning more doesn't necessarily mean a higher spending score, or vice versa.

Spending Score and Loyalty Points (-0.042445): Very weak, almost negligible negative correlation. Implies that age has little influence on loyalty points.

Spending Score and Loyalty Points (0.6723): Moderately strong positive correlation. This could mean that customers who have higher spending scores tend to have more loyalty points, which could indicate that they are more loyal or frequent customers.

Remuneration and Loyalty Points (0.6161): Moderately strong positive correlation. This suggests that individuals with higher remuneration (salary) tend to have more loyalty points, implying that higher-paid individuals may be more loyal customers.

To relationships are visualised below:



In our analysis, we observed that the correlation between age and loyalty points is weak, while remuneration and spending score exhibit a more significant relationship.

Before conducting an OLS regression, we examined the distribution of the dependent variable, "loyalty points," to understand its characteristics, as linear regression assumes normally distributed residuals, primarily for the dependent variable.

After analysing the distribution of the original loyalty points using a histogram, Q-Q plot, and the Shapiro test, we found a significant departure from normality. While normality of residuals is an assumption, it's less critical than other assumptions (outlined below), and deviations from normality can be acceptable with a reasonably large sample size.

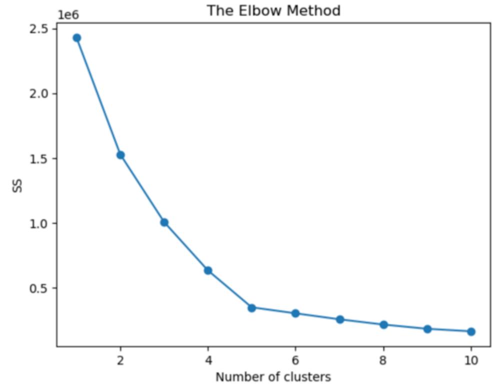
For the OLS regression, it's crucial to assess assumptions of linearity, homoscedasticity, and normality of residuals, which was done by examining residuals. The original loyalty points and log-transformed loyalty points models exhibited violations of linearity and homoscedasticity, but after applying a Box-Cox transformation, these issues improved.

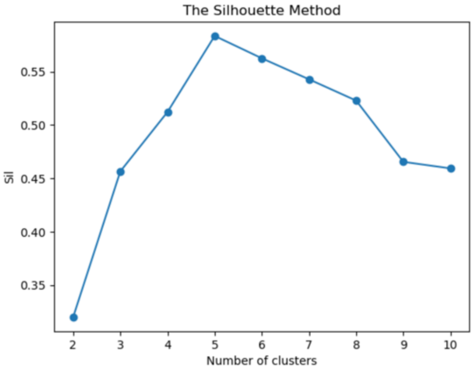
However, the presence of heteroscedasticity in the residuals, as indicated by the Breusch-Pagan test, suggests variability in residuals across different independent variable values. This can impact the reliability and interpretability of regression results, but addressing this issue goes beyond the project's scope.

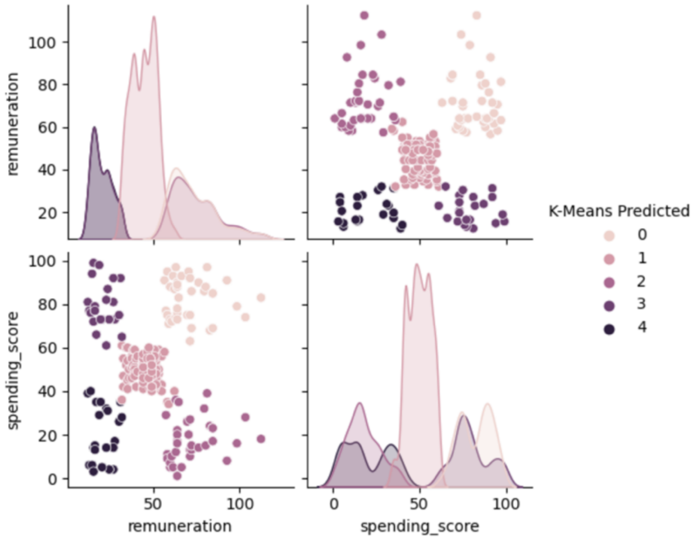
The OLS regression of the Box-Cox transformed model showed promising results with higher R-squared values and improved performance, although heteroscedasticity remains a concern. Considering this, there is concern with the reliability of using this model to make prediction related to turtle games objectives. Another limitation is the omission of outlier analysis due to stakeholder uncertainty in defining what is considered an outlier.

Clustering with k-means

A structured process for k means clustering was followed, starting with data exploration and visualisation and using both the Elbow and Silhouette methods to determine the optimal number of clusters.





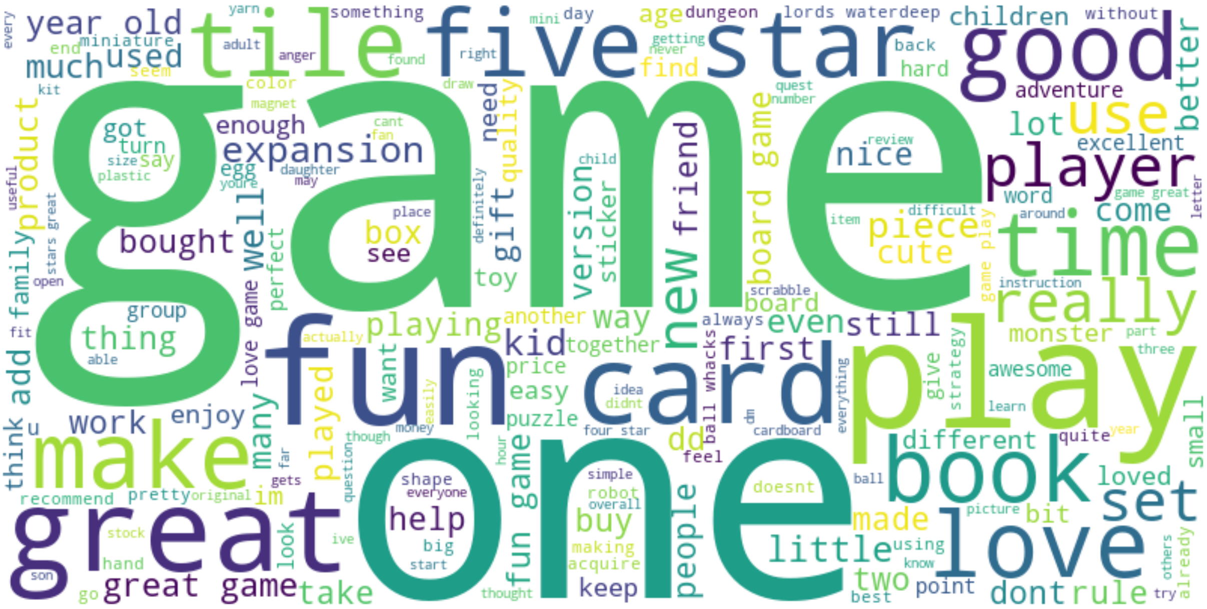


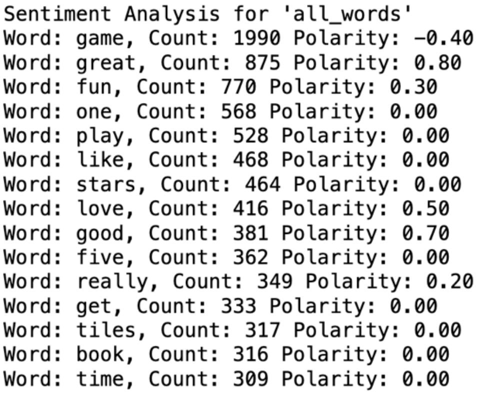
The analysis indicates that clustering the data into five distinct segments provides a balanced and interpretable solution for the data set. This can inform marketing and business strategies tailored to different customer segments such as discounts and vouchers tailored to low earner-high spender and more personalised marketing to high earner-low spender segments.

Sentiment Analysis of Customer Reviews

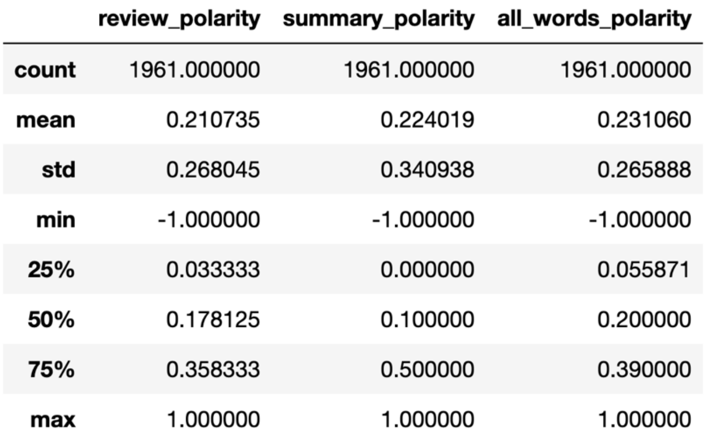
The customer reviews were loaded and explored, necessary packages were imported and the required NLTK resources were downloaded. The data was prepared for NLP by converting text to lowercase, removing punctuation, and eliminating duplicates. Tokenisation and word cloud generation are performed, followed by frequency distribution analysis and sentiment polarity calculation for key words. The sentiment subjectivity was also assessed and histograms created for review, summary, and combined text data. Finally, we identified and displayed the top 20 positive and negative reviews and summaries based on sentiment polarity. By investigating the top positive and negative reviews, this process provides valuable insights for understanding customer sentiment towards different products and platforms.

The wordclouds for the review and summary token column without punctuation and stopwords is generally positive with words such as great, fun, excellent five stars:



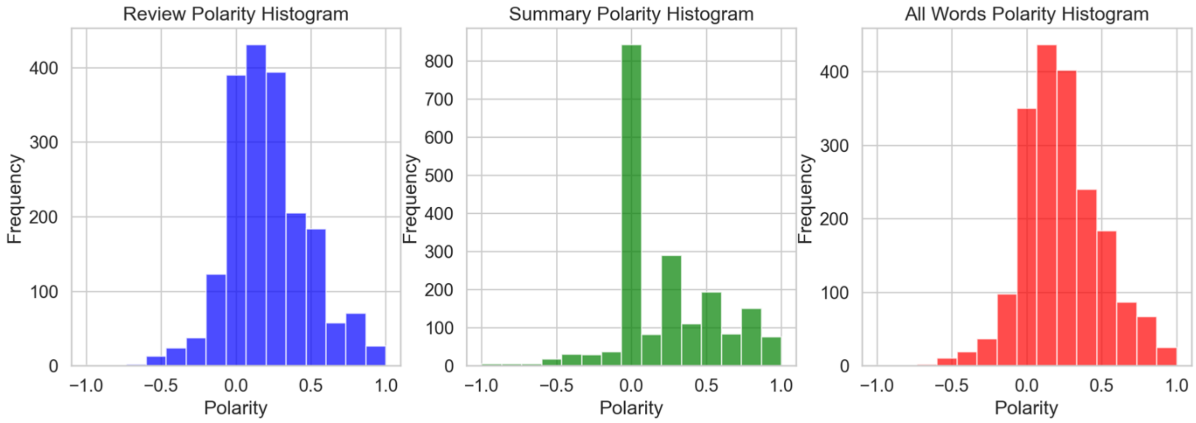

```
Sentiment Analysis for 'all_words'
Word: game, Count: 1990 Polarity: -0.40
Word: great, Count: 875 Polarity: 0.80
Word: fun, Count: 770 Polarity: 0.30
Word: one, Count: 568 Polarity: 0.00
Word: play, Count: 528 Polarity: 0.00
Word: like, Count: 468 Polarity: 0.00
Word: stars, Count: 464 Polarity: 0.00
Word: love, Count: 416 Polarity: 0.50
Word: good, Count: 381 Polarity: 0.70
Word: five, Count: 362 Polarity: 0.00
Word: really, Count: 349 Polarity: 0.20
Word: get, Count: 333 Polarity: 0.00
Word: tiles, Count: 317 Polarity: 0.00
Word: book, Count: 316 Polarity: 0.00
Word: time, Count: 309 Polarity: 0.00
```

The polarity for sentiment analysis indicates the distribution of sentiment scores within a dataset.

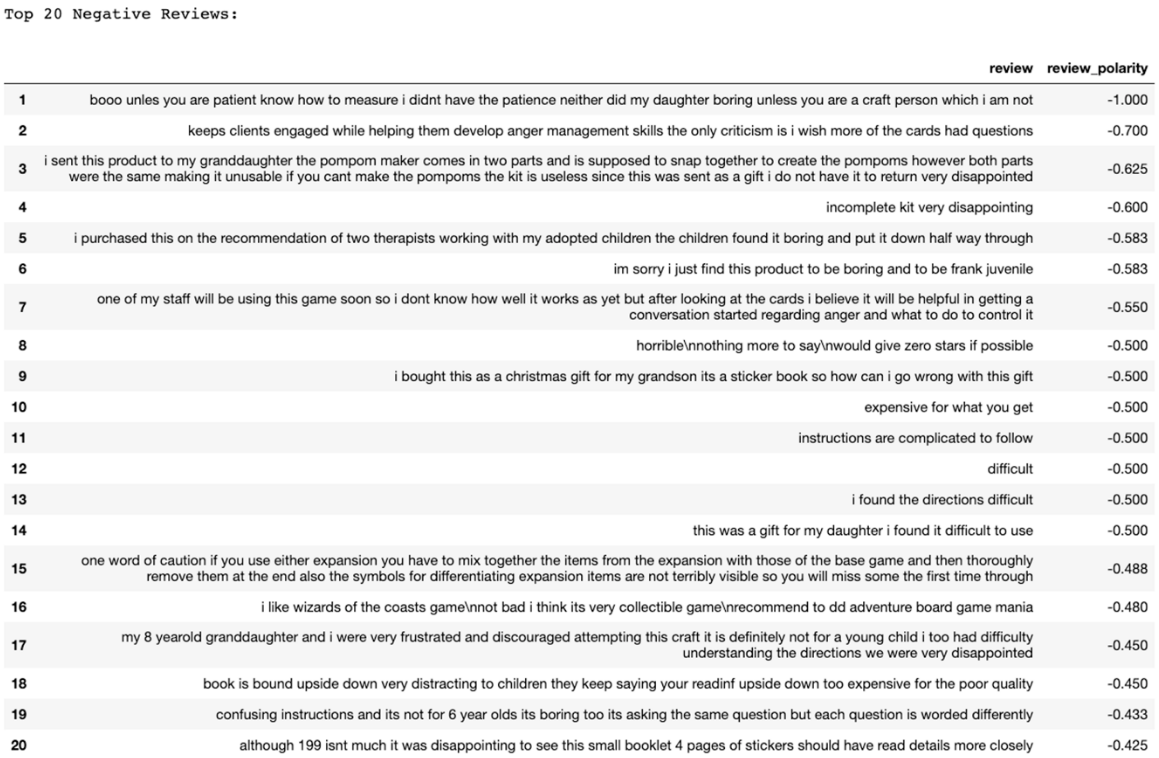


	review_polarity	summary_polarity	all_words_polarity
count	1961.000000	1961.000000	1961.000000
mean	0.210735	0.224019	0.231060
std	0.268045	0.340938	0.265888
min	-1.000000	-1.000000	-1.000000
25%	0.033333	0.000000	0.055871
50%	0.178125	0.100000	0.200000
75%	0.358333	0.500000	0.390000
max	1.000000	1.000000	1.000000

The mean of 0.21 for review polarity and 0.22 for summary polarity indicates that on average the sentiment score lean towards the positive side. The SD of 0.27 and 0.34 respectively, shows the degree of variability in the sentiments scores, suggesting only slight variations in the dataset, greater in summary column. Regarding the quartiles, 25th percentile has values close to 0, indicating neutral sentiment in the dataset. While the 50th (0.18 & 0.1), and 75th percentile (0.358 and 0.5) imply there are more positive sentiments scores than negative, predominantly positive sentiment.



Analysing the top positive and negative reviews highlighted recurring customer concerns, such as difficulties understanding the games and missing or faulty components. Turtle Games can use this information to improve product quality and enhance customer satisfaction. It is recommend to gather customer data to personalise support for negative reviewers.



```
Top 20 Negative Reviews:
```

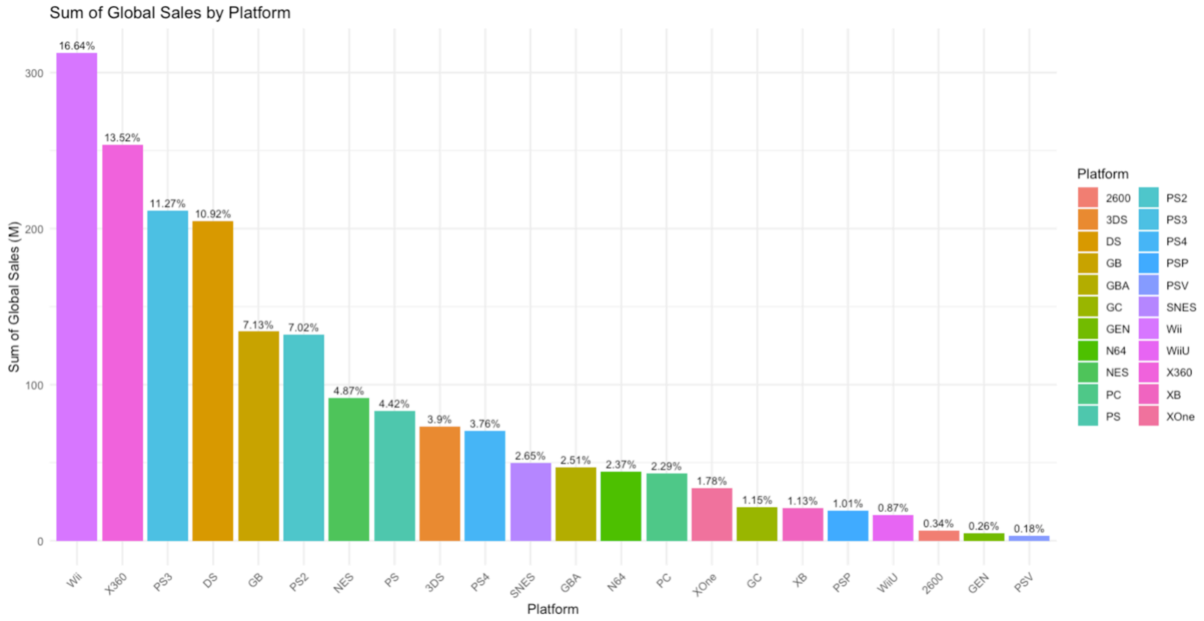
	review	review_polarity
1	booo unles you are patient know how to measure i didnt have the patience neither did my daughter boring unless you are a craft person which i am not	-1.000
2	keeps clients engaged while helping them develop anger management skills the only criticism is i wish more of the cards had questions	-0.700
3	i sent this product to my granddaughter the pompom maker comes in two parts and is supposed to snap together to create the pompoms however both parts were the same making it unusable if you cant make the pompoms the kit is useless since this was sent as a gift i do not have it to return very disappointed	-0.625
4	incomplete kit very disappointing	-0.600
5	i purchased this on the recommendation of two therapists working with my adopted children the children found it boring and put it down half way through	-0.583
6	im sorry i just find this product to be boring and to be frank juvenile	-0.583
7	one of my staff will be using this game soon so i dont know how well it works as yet but after looking at the cards i believe it will be helpful in getting a conversation started regarding anger and what to do to control it	-0.550
8	horrible\nnothing more to say\nwould give zero stars if possible	-0.500
9	i bought this as a christmas gift for my grandson its a sticker book so how can i go wrong with this gift	-0.500
10	expensive for what you get	-0.500
11	instructions are complicated to follow	-0.500
12	difficult	-0.500
13	i found the directions difficult	-0.500
14	this was a gift for my daughter i found it difficult to use	-0.500
15	one word of caution if you use either expansion you have to mix together the items from the expansion with those of the base game and then thoroughly remove them at the end also the symbols for differentiating expansion items are not terribly visible so you will miss some the first time through	-0.488
16	i like wizards of the coasts game\nnot bad i think its very collectible game\nrecommend to dd adventure board game mania	-0.480
17	my 8 yearold granddaughter and i were very frustrated and discouraged attempting this craft it is definitely not for a young child i too had difficulty understanding the directions we were very disappointed	-0.450
18	book is bound upside down very distracting to children they keep saying your readinf upside down too expensive for the poor quality	-0.450
19	confusing instructions and its not for 6 year olds its boring too its asking the same question but each question is worded differently	-0.433
20	although 199 isnt much it was disappointing to see this small booklet 4 pages of stickers should have read details more closely	-0.425

Furthermore, random samples confirmed the alignment of polarity and subjectivity with individual reviews, with many comments being neutral. This suggests the need to examine neutral reviews in more detail to identify potential areas for improvement.

Analysis of Sales Data (R)

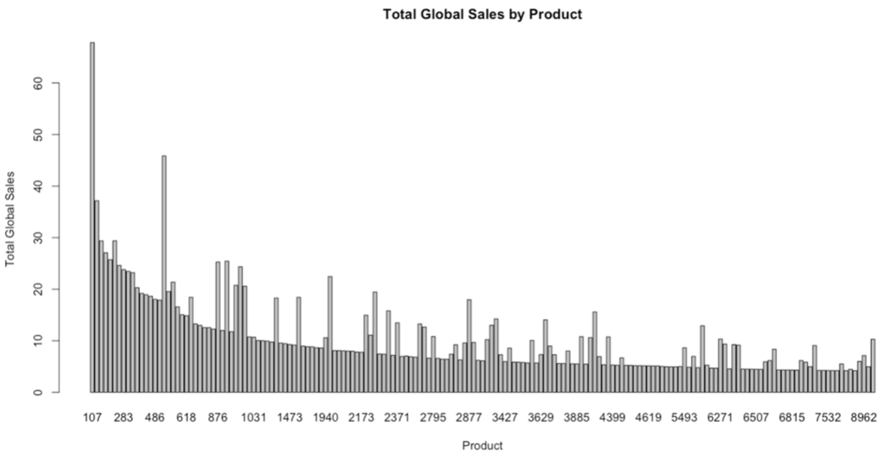
Exploratory Analysis of Sales Data – Impact of Product ID and Platform

When analysing the sales data, initial exploratory analysis was conducted and visualisations created to understand the relationship between sale and product ID and Platform using group_by, as well as the different sales columns of NA_sales, EU_sales and Global_sales.

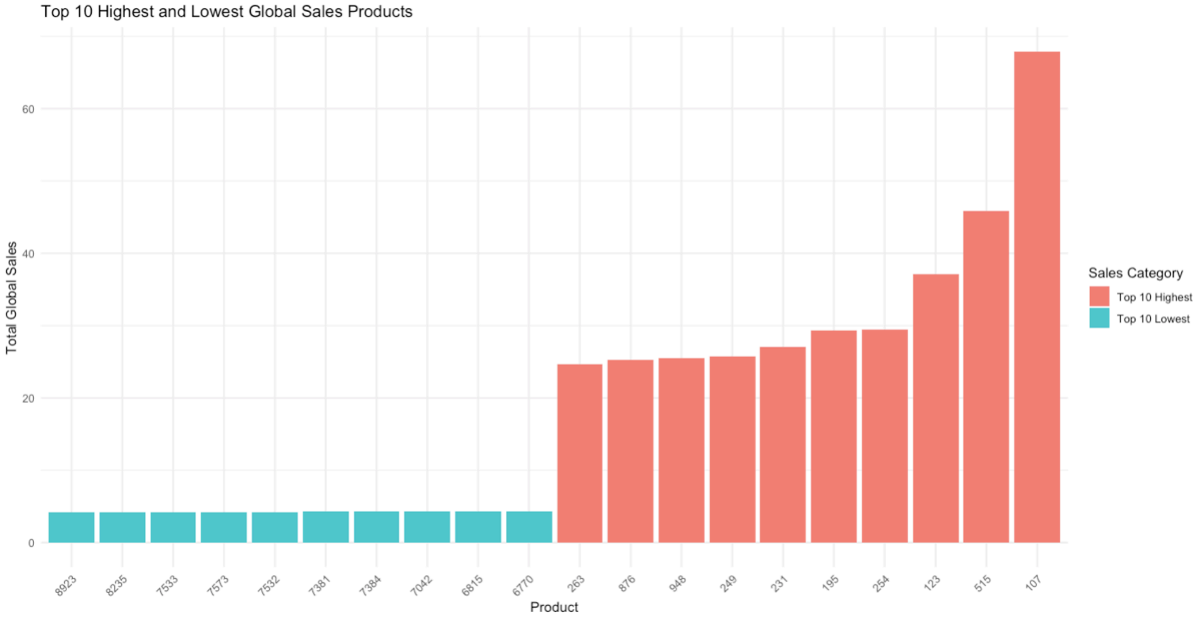


Bar charts and boxplots were created to delve into sales by platform, identifying both the top and bottom-selling products. The analysis revealed insights into sales per product and platform. The mean global sales amounted to 5.335 million, with NA and EU sales averaging 2.510 million and 1.644 million, respectively.

Notably, the highest-selling platforms for global sales included Wii, Xbox360, and PS3, while in NA, X360, Wii, and PS3 were the top performers, and in EU, Wii, PS3, and X360 stood out.

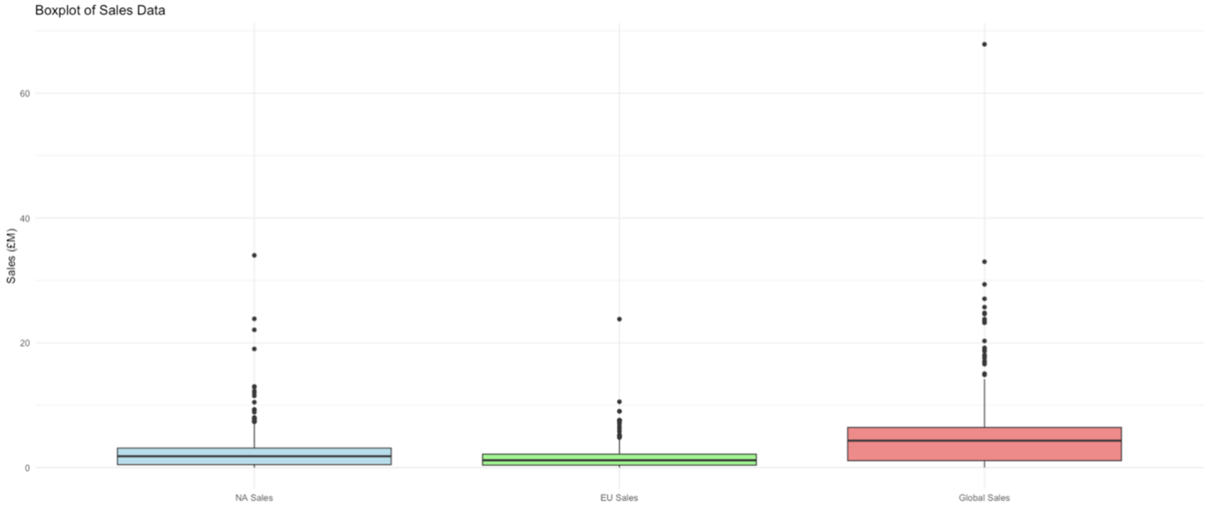


Concerning product ID, our analysis found that lower product ID numbers were associated with higher total global sales. However, since we lacked information about the products corresponding to these product IDs, our analysis was limited to displaying the top and lowest global sales by product IDs. This analysis indicated that the lowest-selling products consistently hovered around 4.2 - 4.3 million in sales. It is advised that Turtle games provide the corresponding product names so that a more in-depth analysis can be done on the highest and lowest selling products.


Top 10 Highest and Lowest Global Sales Products

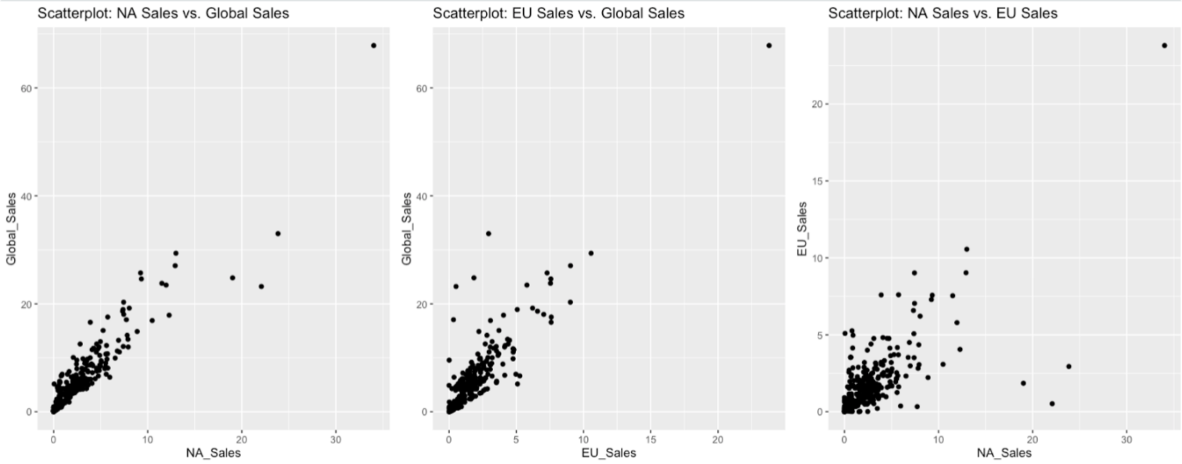
Exploratory Analysis of Sales Data – Relationship of Sales Data

Investigating the relationships between the different sales columns, the focus was on determining the normality of the data, starting with Q-Q plots to visualize the distribution and Shapiro-Wilk tests to assess normality. Skewness and kurtosis are calculated to understand the data's shape and heaviness of tails. Lastly, correlations between different sales columns are explored, to provide insights into relationships between sales regions using scatterplots with trend lines to visualize correlations between sales columns.

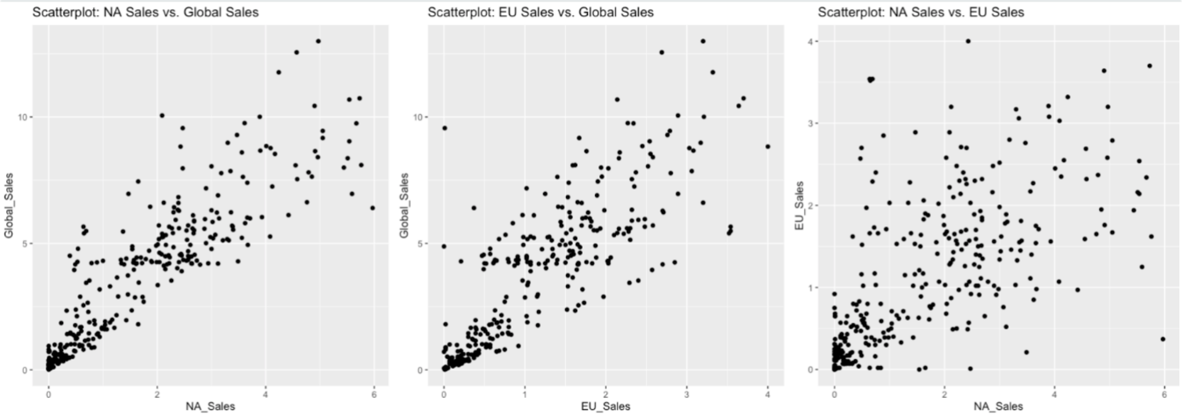

Boxplot of Sales Data

The data was heavily skewed to lower sales amount in the 0 – 10 million range with outliers present, outliers were identified and removed using the IQR method, followed by the creation of scatterplots again to visualise relationships without outliers.

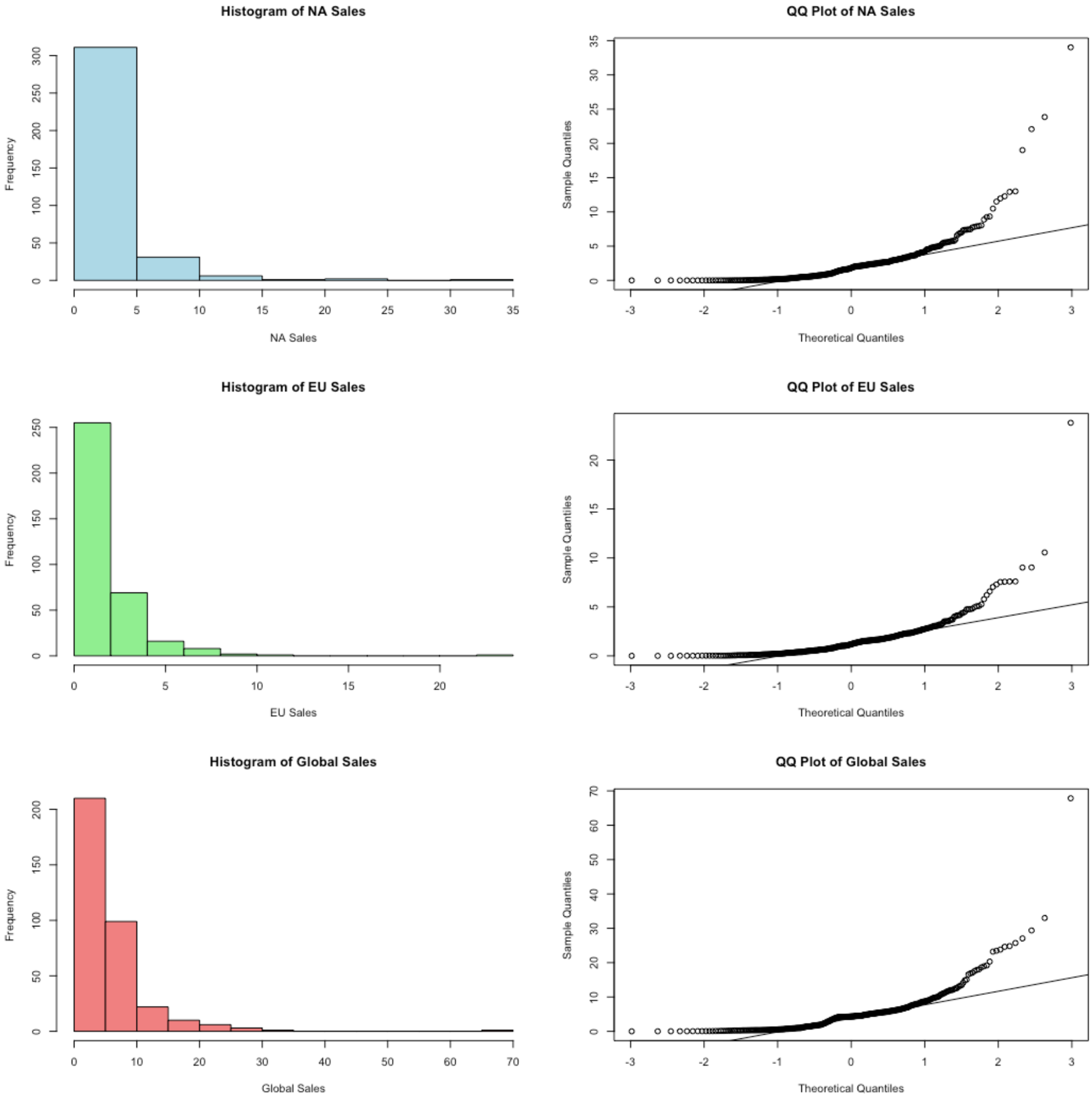
Without Outliers Removed



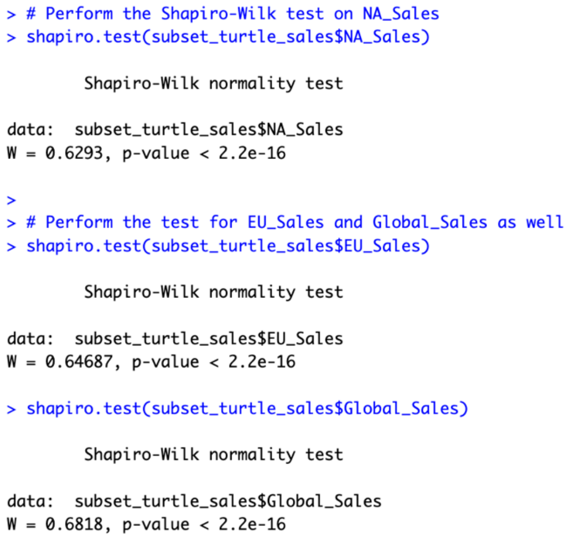
With Outliers Removed



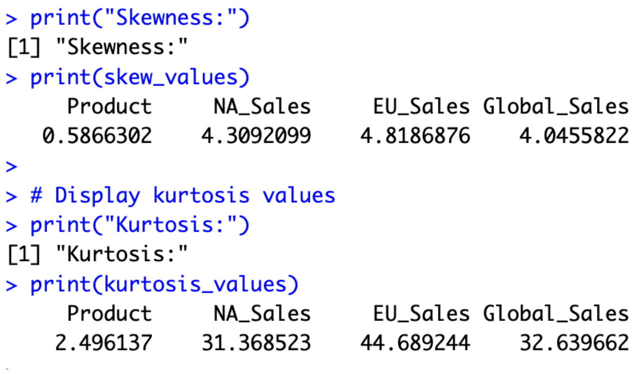
Subsequent steps involved the creation of histograms and boxplots to gain insights into the distribution and spread of sales data. The QQ plots indicated positive skewness with pronounced tails, supported by significantly high kurtosis values. The Shapiro-Wilk tests indicated non-normal data distribution, typically indicated by a low p-value (usually less than 0.05).



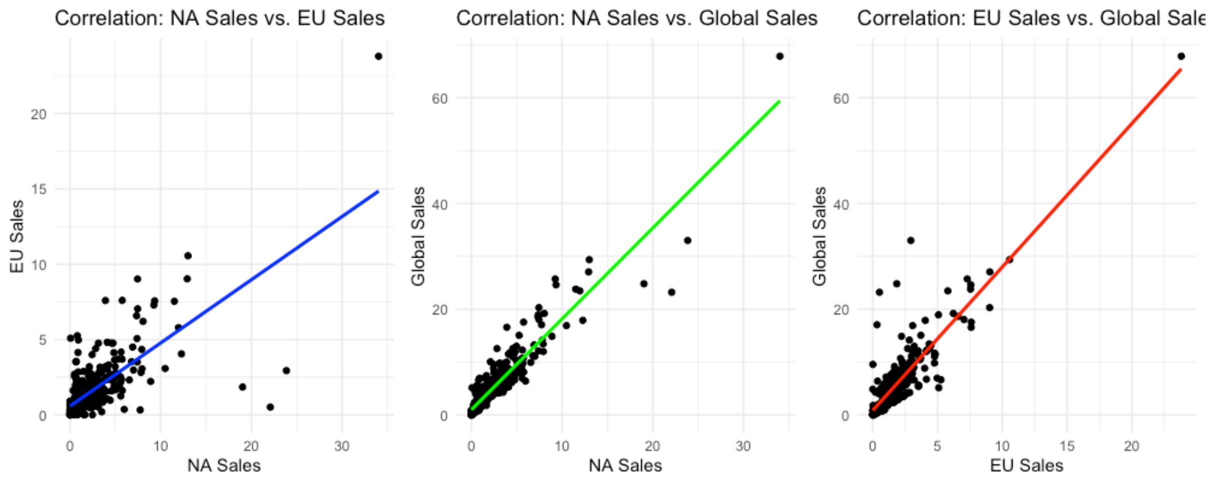
The original sales data can be seen above, it suggests that the data is heavily skewed to the right where the majority of results are lower amounts of global sales. The QQ plots indicate the data is positively skewed with heavy tails indicated by the significantly high kurtosis values (below). This may have implication towards the performance and reliability of the regression model used explained below.

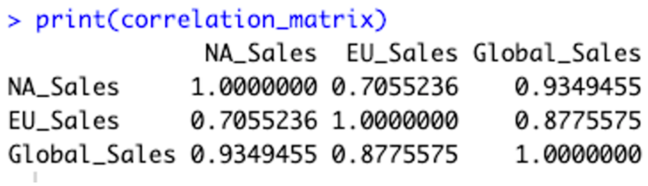


```
> # Perform the Shapiro-Wilk test on NA_Sales
> shapiro.test(subset_turtle_sales$NA_Sales)

        Shapiro-Wilk normality test

data:  subset_turtle_sales$NA_Sales
W = 0.6293, p-value < 2.2e-16

>
> # Perform the test for EU_Sales and Global_Sales as well
> shapiro.test(subset_turtle_sales$EU_Sales)

        Shapiro-Wilk normality test

data:  subset_turtle_sales$EU_Sales
W = 0.64687, p-value < 2.2e-16

> shapiro.test(subset_turtle_sales$Global_Sales)

        Shapiro-Wilk normality test

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



```
> print("Skewness:")
[1] "Skewness:"
> print(skew_values)
   Product    NA_Sales    EU_Sales Global_Sales
 0.5866302   4.3092099   4.8186876    4.0455822
>
> # Display kurtosis values
> print("Kurtosis:")
[1] "Kurtosis:"
> print(kurtosis_values)
   Product    NA_Sales    EU_Sales Global_Sales
  2.496137   31.368523   44.689244    32.639662
```





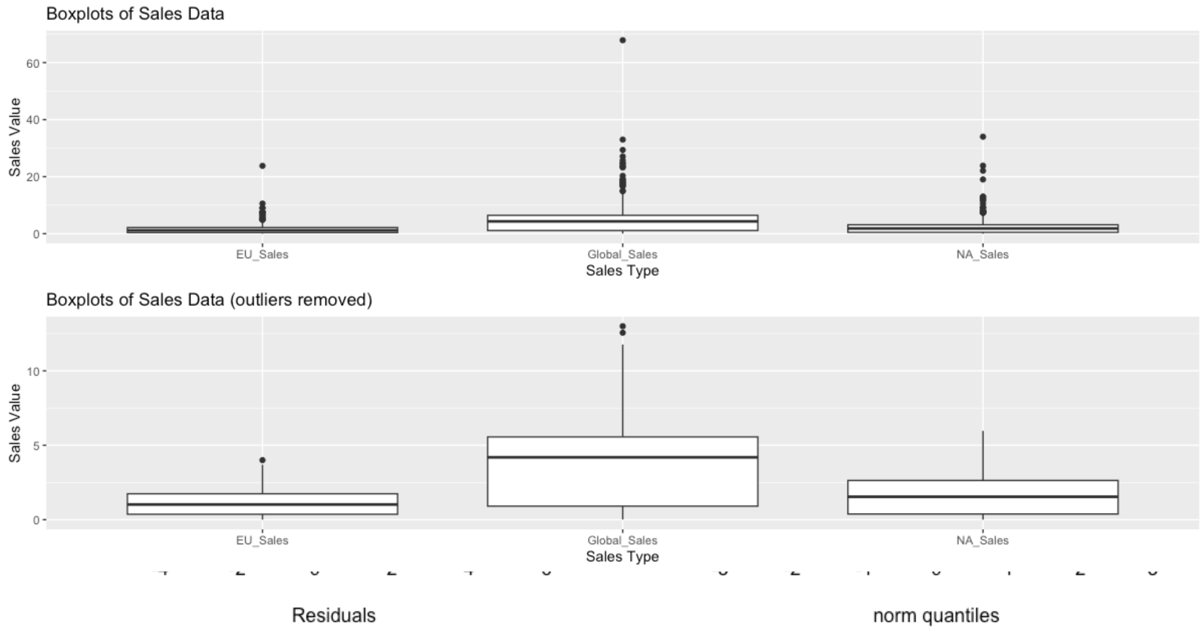
```
> print(correlation_matrix)
              NA_Sales  EU_Sales Global_Sales
NA_Sales     1.0000000 0.7055236    0.9349455
EU_Sales     0.7055236 1.0000000    0.8775575
Global_Sales 0.9349455 0.8775575    1.0000000
```

The analysis also revealed strong positive correlations between various sales regions. NA and Global Sales exhibited a very strong positive correlation (0.93), with NA driving the majority of sales. A moderately strong positive correlation (0.88) existed between EU and Global Sales, as well as between NA and EU Sales (0.71). These strong positive correlations implied that sales in these regions were positively related, where an increase in one type of sales corresponded with increases in others, aligning with expectations.

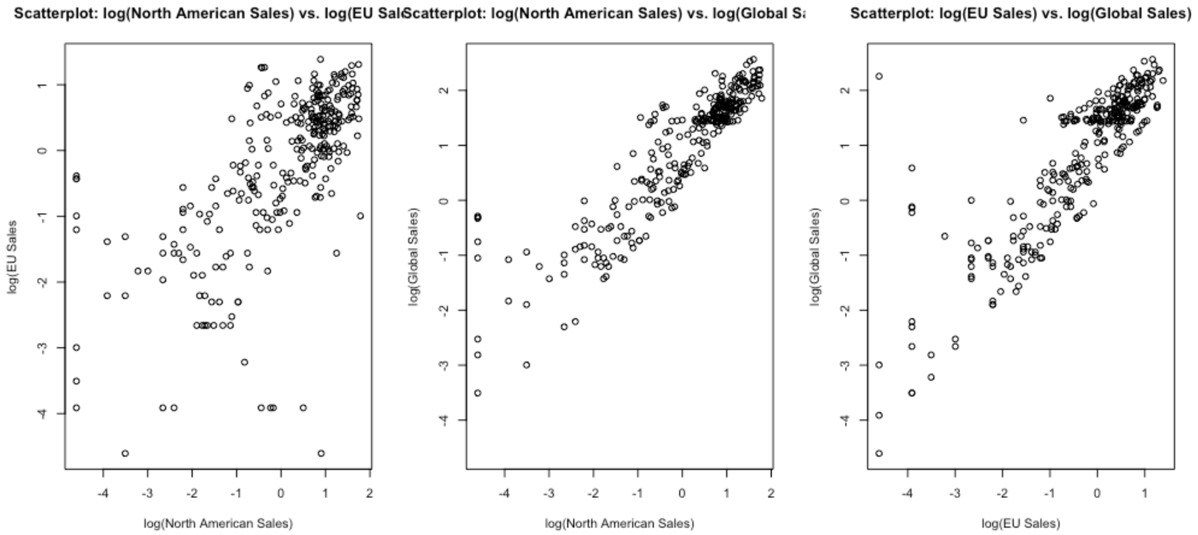
The positive skewness of the data holds significance in how turtle games can identify pricing and marketing decisions, such as bundling lower selling with higher selling products.

Linear Regression Analysis of the Relationship Between Sales Data

The initial analysis of linear regression on sales data revealed several important considerations. The non-normality of data distribution, positive skewness, and kurtosis indicated the presence of extreme values or outliers, potentially affecting the regression model. Strong correlations between sales columns suggested potential multicollinearity, making it challenging to disentangle the individual effects of predictors in linear regression modelling.

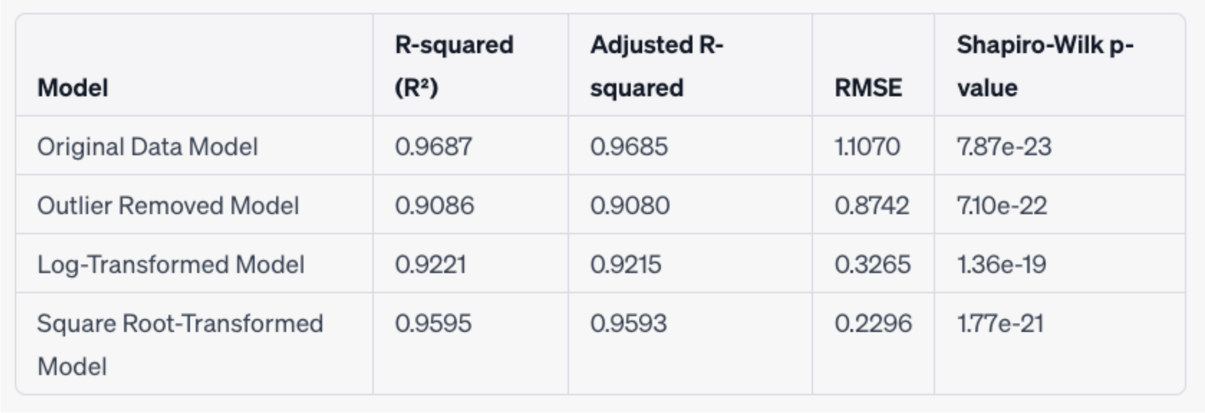


Due to the interdependence of predictors in the sales data, a decision was made to employ multiple linear regression to explore how the three sales variables interact and influence the outcome. To improve model fit, outliers were removed, resulting in a better distribution in the histogram and improved residuals in the QQ plot. However, the presence of a small p-value suggested that residuals did not follow a normal distribution.



Log transformations were attempted but did not resolve the non-normal distribution issue. The positively skewed data also raised concerns about non-constant variance of residuals (heteroscedasticity), violating the assumption of normally distributed residuals with constant variance. Further data transformations, such as square root transformations are considered to address these issues.

Having created and tested the performance of 4 separate models the findings are shown below:



Model	R-squared (R²)	Adjusted R-squared	RMSE	Shapiro-Wilk p-value
Original Data Model	0.9687	0.9685	1.1070	7.87e-23
Outlier Removed Model	0.9086	0.9080	0.8742	7.10e-22
Log-Transformed Model	0.9221	0.9215	0.3265	1.36e-19
Square Root-Transformed Model	0.9595	0.9593	0.2296	1.77e-21

In the analysis, the original data model has a high R-squared value but non-normally distributed residuals and a relatively high RMSE. The outlier-removed model has a lower R-squared value but lower RMSE. The log-transformed model shows a good R-squared value, normally distributed residuals, and a lower RMSE compared to the original data model. The square root-transformed model has the highest R-squared value, normally distributed residuals, and the lowest RMSE among all models. Depending on the priority, the square root-transformed model excels in relationship strength, while the log-transformed and square root-transformed models both exhibit normally distributed residuals. If a balance between relationship strength and RMSE is desired, the square root-transformed model is the top choice.

The predicted values obtained represent estimated global sales based on 'NA_Sales_sum' and 'EU_Sales_sum' inputs. Higher 'NA_Sales' and 'EU_Sales' values correspond to higher predicted global sales, as seen with sets of data inputs. The square root-transformed model's coefficients for these variables are statistically significant, with an R-squared value of approximately 0.9595, indicating a strong relationship and substantial variance explanation. Residual analysis shows that residuals are approximately normally distributed, meeting a key linear regression assumption. The adjusted R-squared value is around 0.9593, providing a more conservative estimate of the model's goodness of fit. The square root-transformed model demonstrates a strong fit to the data, supported by statistical significance and diagnostic tests. However, issues with heteroskedasticity should be considered with performance testing on new data being advisable.

Limitations

Finally, limitations of the existing data as well as recommendations for further areas to explore are listed below.

Limitations of Data and Analysis

- As both the sales and customer data is not continuous time series data, it impossible to determine trends over time or make predictions. For example, NES has a relatively hig total sales, but is a very old console and was discontinued in August, 1995, making it irrelevant to current or future marketing efforts.
- Additionally as we can't speak to the stakeholders to confirm business objectives, we are unsure whether to include or ignore games sold past a certain date or on obsolete consoles.
- Some of the top negative comments actually contained positive sentiment but it was still given a -ve polarity.
- Concerns persist about heteroskedasticity and non-normality of residuals, necessitating further exploration.

Recommended Future Actions and Areas to Explore

- Identify individual users from negative reviews to determine churn rate and offer tailored services or discounts to high-purchasing customers.
- Match product IDs with corresponding products to gain a comprehensive view of sales by product.
- Investigate discrepancies between Global sales and the sum of NA and EU sales to uncover unaccounted-for sales.
- Estimate predicted sales using categorical variables like Genres and platforms and assess their impact on revenue.
- Focus on products from the last 5-10 years due to the age of the data.
- Link products with customers for more targeted marketing strategies, particularly for high-spending genres.
- Merge data sets using product IDs to link negative reviews to individual products and across platforms.
- Develop an interactive dashboard for Turtle Games to enhance understanding of sales and customer data, identifying new opportunities and explore changes live. over time