Course 3 Assignment – Predicting Future Outcomes
Case Study: Turtle Games

Project Overview

In this assignment, we tackle Turtle Games' sales challenges using data analysis. Our goal: improve sales performance by understanding customer trends and optimising marketing. We explore various activities to answer critical questions, including data exploration, visualisation and statistical modelling. Our insights will empower Turtle Games in the competitive gaming industry.

Section 1: Making Predictions with Regression

In our analysis, we employed linear regression to investigate the relationships between loyalty points and key demographic and behavioural factors, including age, remuneration (in thousands of pounds), and spending score (on a scale of 1 to 100). The results of the regression analysis are summarised below (*Figure 1)*:

R-squared: 0.840

- The R-squared value indicates that approximately 84% of the variability in loyalty points can be explained by the variables included in our model.

Coefficients:

- Age: 11.0607 (p-value < 0.001) - A positive coefficient suggests that, on average, for each year increase in age, loyalty points increase by approximately 11.06 points.
- Remuneration (k£): 34.0084 (p-value < 0.001) - The positive coefficient indicates that higher remuneration is associated with higher loyalty points.
- Spending Score (1-100): 34.1832 (p-value < 0.001) - The positive coefficient suggests that individuals with higher spending scores tend to accumulate more loyalty points.

Model Significance:

- The overall model is statistically significant (F-statistic: 3491, p-value < 0.001), indicating that at least one of the independent variables is a significant predictor of loyalty points.

Regression Plot:

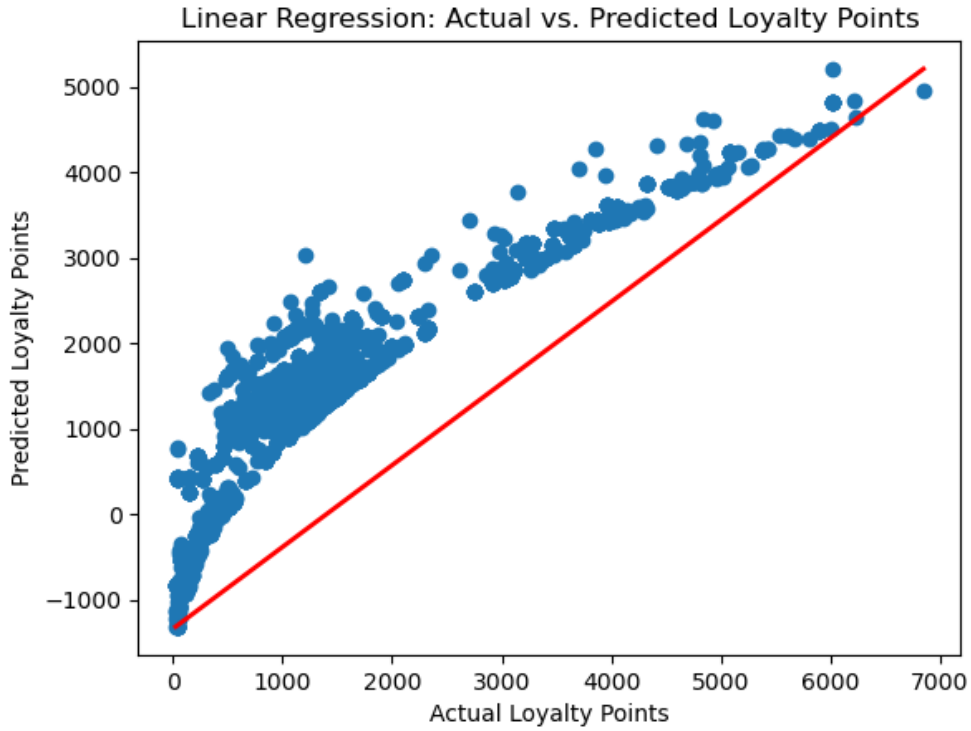

Figure 1: Actual vs Predicted Loyalty Points Linear Regression Graph

The regression plot displayed above illustrates the relationship between actual loyalty points and the predicted values based on the regression model. It visually demonstrates the positive linear association between the independent variables (age, remuneration, spending score) and loyalty points.

These findings suggest that age, remuneration, and spending score are valuable predictors of loyalty points, with higher values in these variables corresponding to higher loyalty points. These insights can inform our marketing strategies to target individuals with higher remuneration and spending scores, ultimately enhancing our loyalty program's effectiveness, and boosting overall sales performance.

Section 2: Making Predictions with Clustering

In this phase of our analysis, we explore clustering to reveal unique customer segments, allowing us to optimise tailored marketing strategies based on salary and spending score.

Elbow Method:

Using the Elbow method, we sought the ideal cluster count (k). By plotting inertia against k values (ranging from 1 to 10), we determined the "elbow point" where inertia reduction slowed. We found that k=3 marked this significant change, suggesting it as the optimal choice for k-means clustering. Beyond k=3, inertia reduction was marginal.

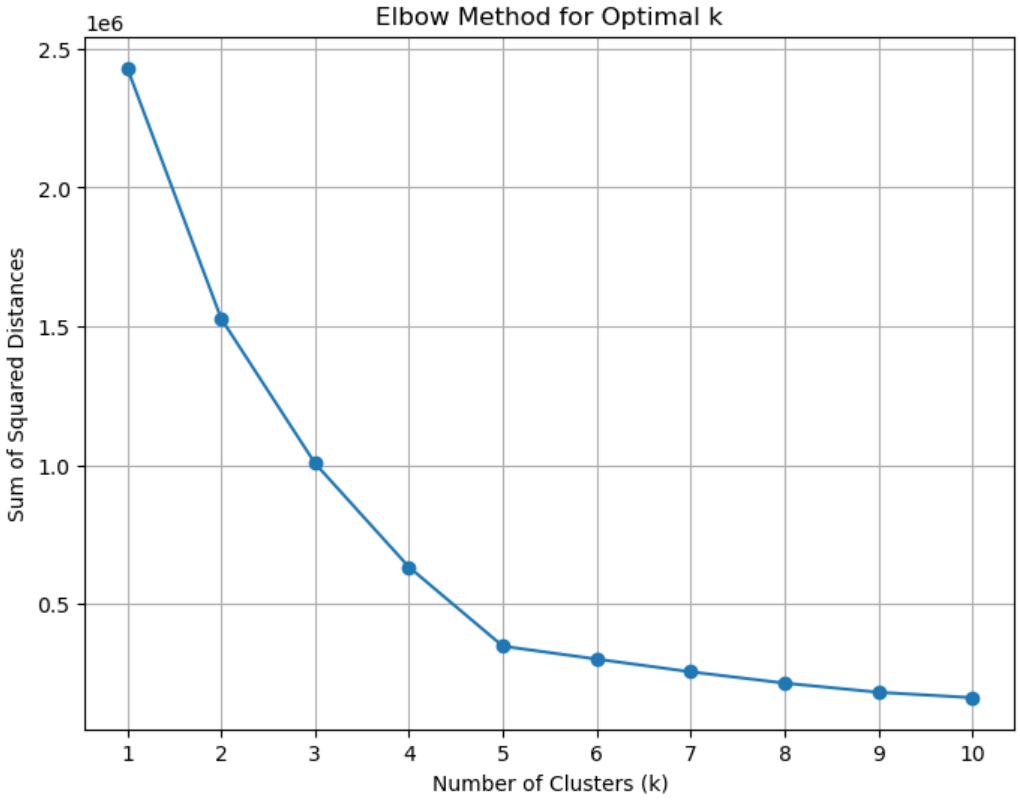

Figure 2: Elbow Method for Optimal k-value

Silhouette Method:

The Silhouette method assessed clustering quality with silhouette scores for k values (2 to 10). A higher score indicates well-defined clusters. We observed that k=5 yielded the highest silhouette score (0.583), indicating clearer separation among data points. Beyond k=5, clustering quality diminished.

Optimal Number of Clusters:

Our analysis using both the Elbow and Silhouette methods converged on the same result: k=5 appears to be the optimal number of clusters. This choice strikes a balance between meaningful cluster separation and complexity.

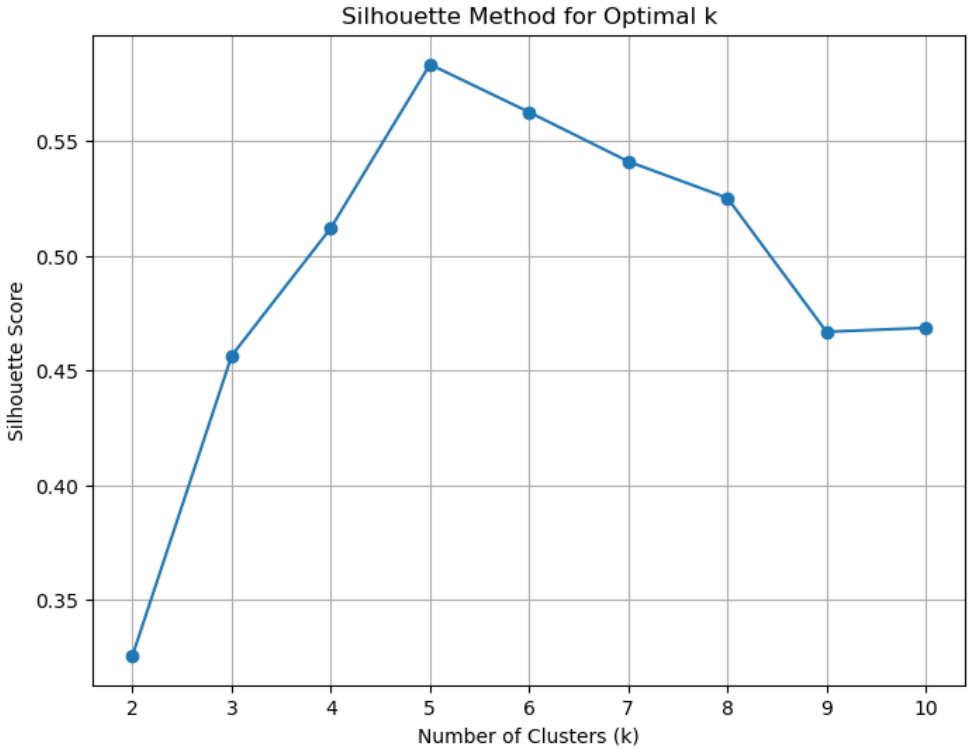

Figure 3: Silhouette Method for Optimal k-value

Cluster Insights:

Our analysis unveiled five customer segments based on salary and spending score:

- Cluster 0 (Moderate Income, Reasonable Spending): 774 customers. Tailored marketing and loyalty programs can enhance loyalty.
- Cluster 1 (Low Income, Minimal Spending): 271 customers. Targeted initiatives are needed to boost spending and engagement.
- Cluster 2 (High Income, Low Spending): 330 customers. Exclusive products or incentives can unlock their potential.
- Cluster 3 (High Income, High Spending): 356 customers. Premium offerings and personalised experiences are key.
- Cluster 4 (Low Income, High Spending): 269 customers. Understanding preferences can enhance engagement.

These insights inform marketing strategies, customer retention, and sales performance improvement.

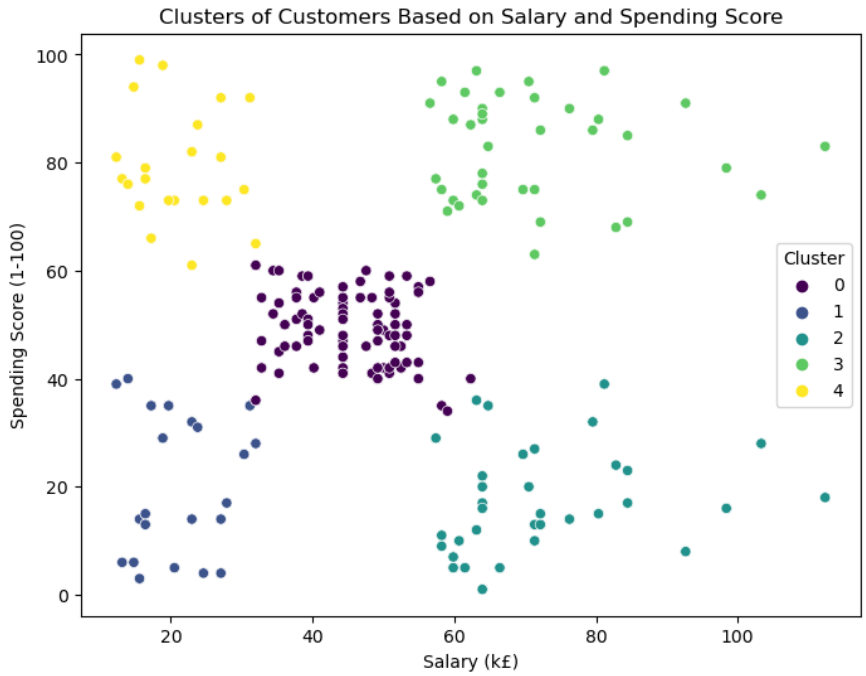

Figure 4: Clusters of Customers Based on Salary and Spending Score

Section 3: Analysing Customer Sentiments with Reviews

In our customer sentiment analysis from online product reviews, we discovered insights that can profoundly shape Turtle Games' marketing strategies.

Commonly Used Words:

To derive these insights, we harnessed the power of customer reviews. Through captivating word clouds (Figures 5 and 6), we visually highlighted frequently occurring themes. We also condensed customer sentiment into a table (Table 1) featuring the top 15 frequently used words. This table illuminated critical themes and topics that deeply resonate with customers.

Figure 5: Word cloud from 'review' column.

Figure 6: Word cloud from 'summary' column.

	Word	Frequency
0	game	2308
1	great	1169
2	fun	987
3	stars	891
4	five	704
5	one	607
6	play	554
7	like	522
8	love	509
9	good	473
10	really	379
11	book	359
12	get	348
13	tiles	337
14	kids	334

Table 1: Most frequently used words in 'review' and 'summary' columns combined.

Sentiment Polarity Distribution:

Examining the sentiment polarity distribution of the reviews using a histogram, revealed an intriguing pattern (*Figure 7)*. The majority of reviews tended to exhibit positive sentiment, with a peak in the polarity score around 0.875 to 1.0. This distribution underscores the importance of catering to a diverse range of customer opinions and experiences.

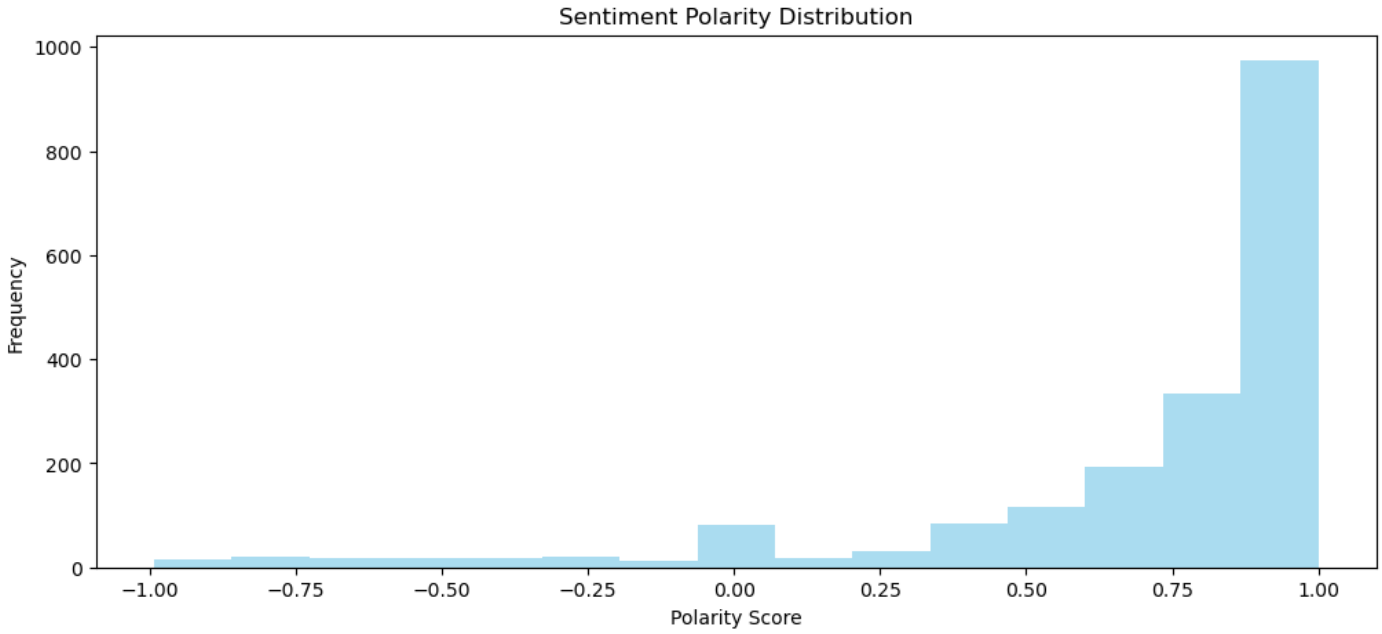

Figure 7: Histogram showing Sentiment Polarity Distribution

Top Positive and Negative Reviews:

Our analysis extracted top 20 positive and negative reviews, covering a range of customer experiences. Positive reviews highlighted product quality and customer service, while negative ones mentioned product bugs and slow support, showing room for improvement. In conclusion, our sentiment analysis guides Turtle Games' marketing efforts with insights from commonly used words and sentiment polarity. Top reviews offer actionable guidance for satisfaction enhancement and product refinement, empowering data-driven decisions and better marketing campaigns.

Section 4: Visualising data to gather insights

In our data analysis using R, we streamlined the dataset by removing unnecessary columns (Ranking, Year, Genre, Publisher) and employed essential libraries (ggplot2, dplyr, tidyr) for data visualisation.

We created scatterplots, histograms, and boxplots, providing visual insights into sales data. These visuals help identify regional trends, platform preferences, and outliers, enabling informed decisions for sales optimisation.

These plots are invaluable for Turtle Games, offering clear insights into sales performance. In our presentation, we'll explain how they reveal regional platform preferences, outlier detection, and distribution characteristics. These insights drive strategies which can enhance marketing, distribution, and growth opportunities.

Platform vs. NA Sales scatterplot

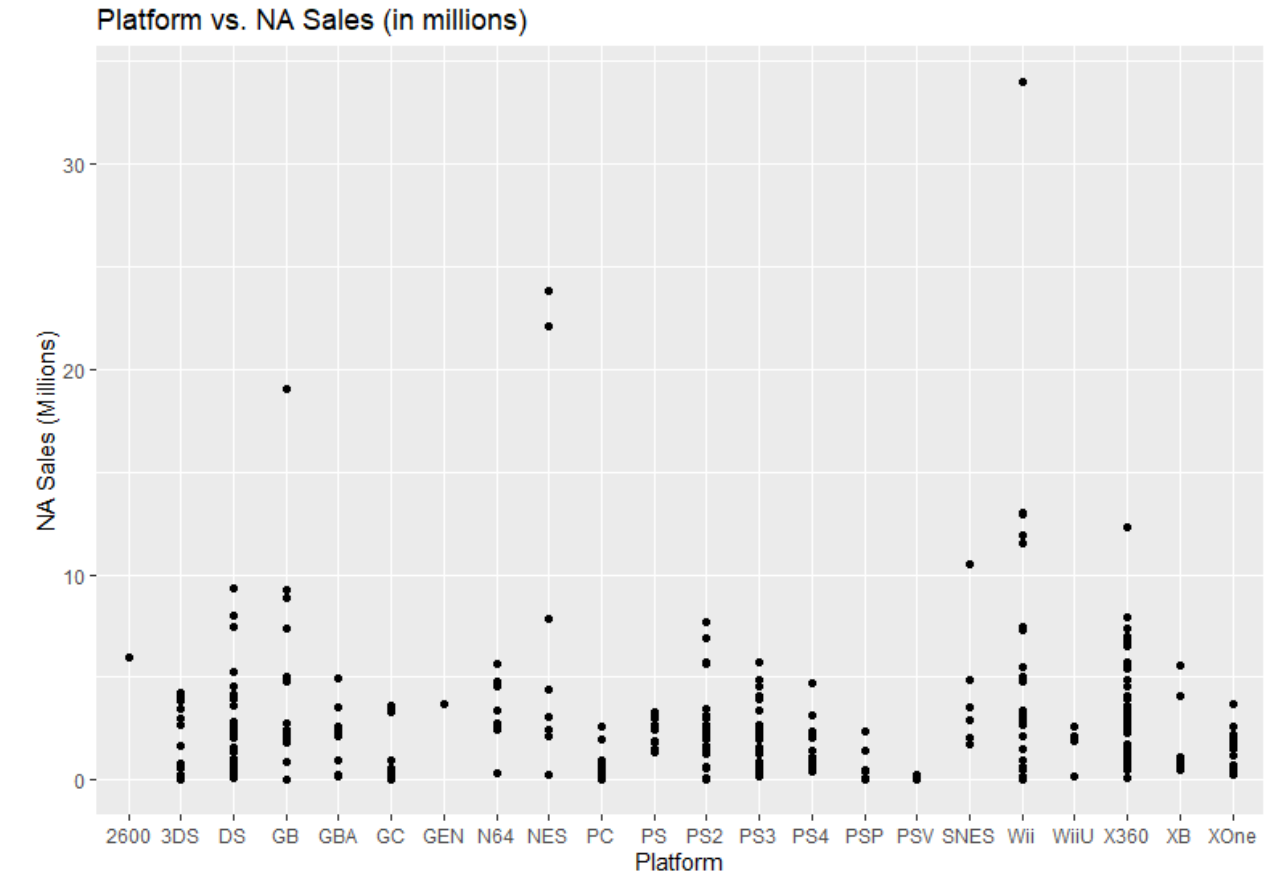

Figure 8: Scatterplot of Platform vs NA Sales

Platform vs. EU Sales scatterplot

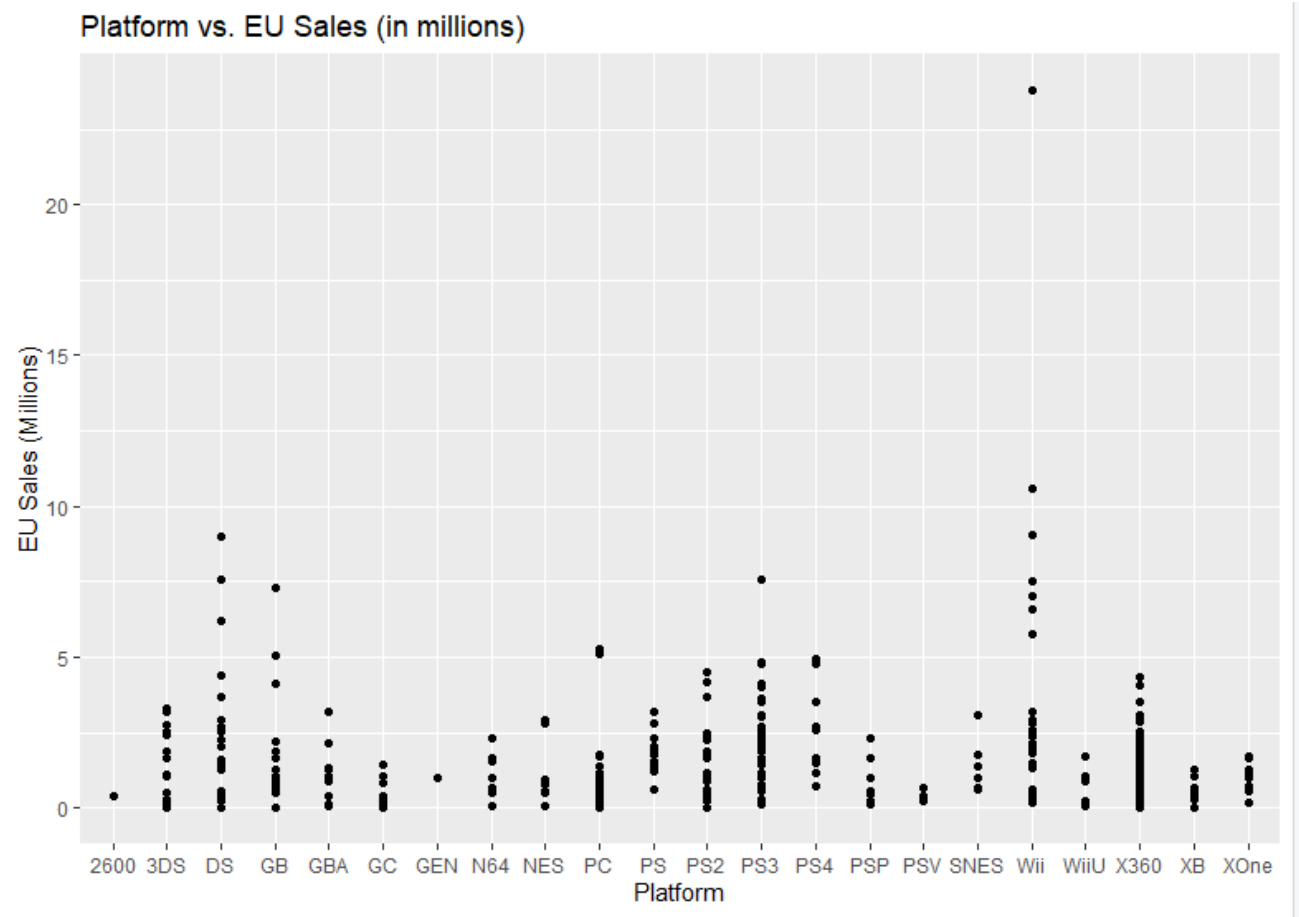

Figure 9: Scatterplot of Platform vs EU Sales

Platform vs Global Sales scatterplot

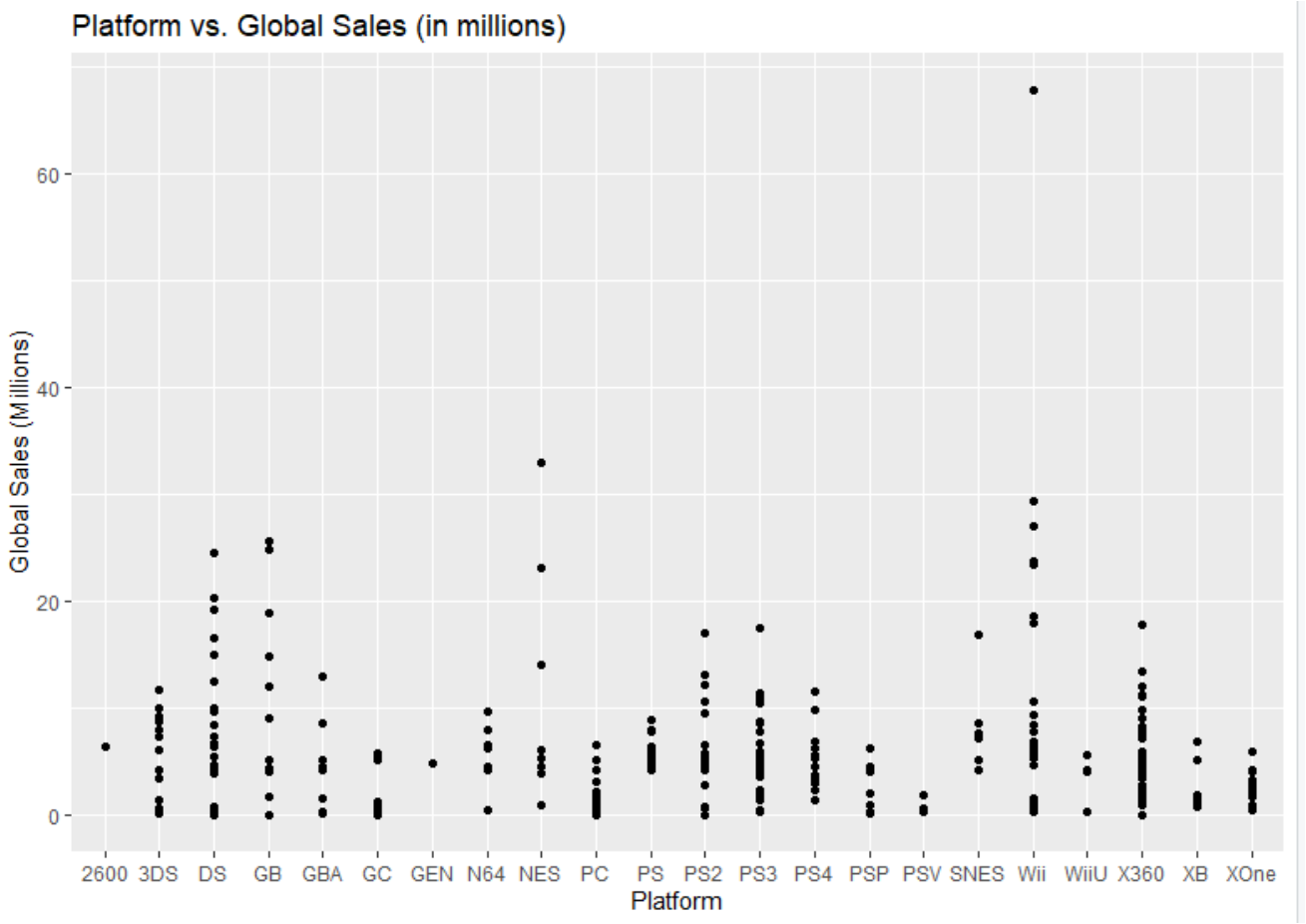

Figure 10: Scatterplot of Platform vs Global Sales

NA Sales Histogram

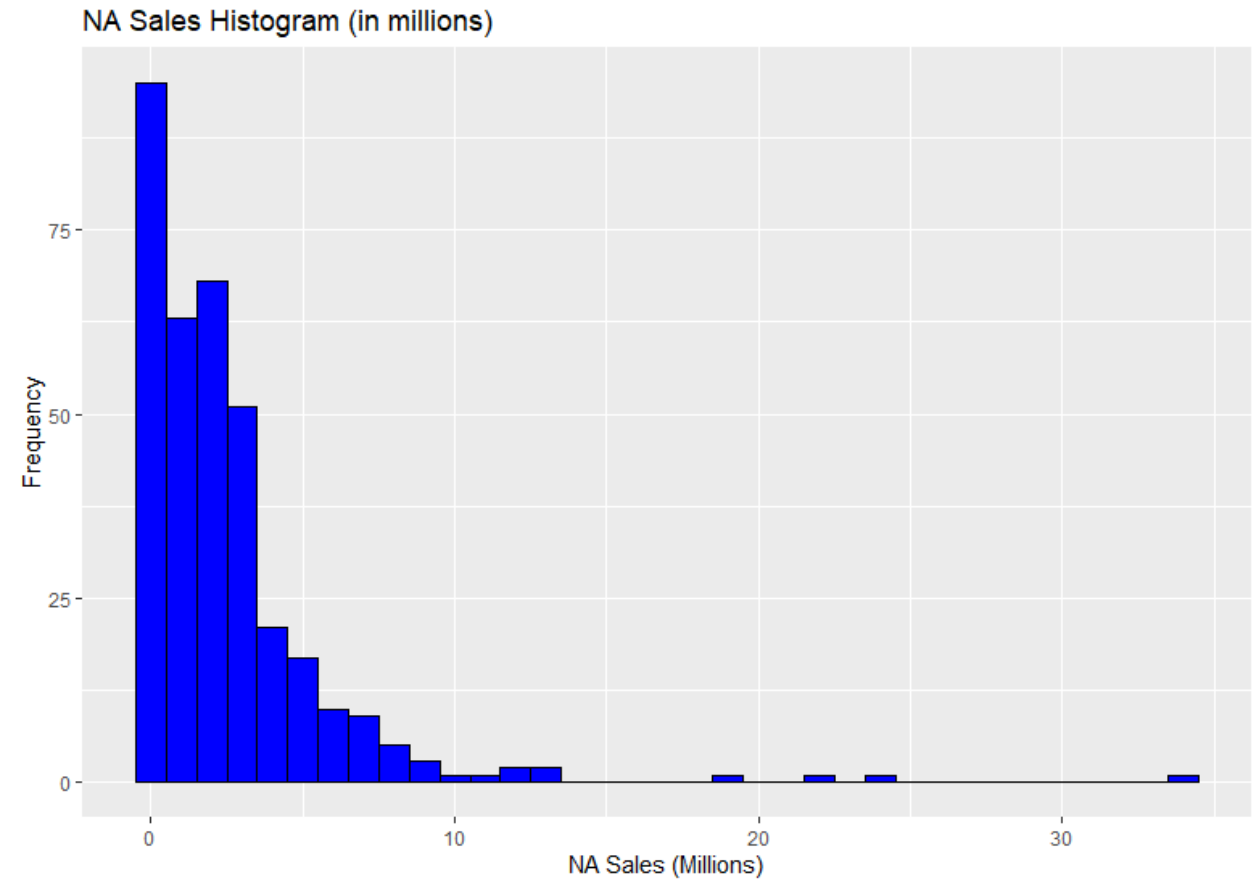

Figure 11: Histogram of NA Sales

EU Sales Histogram

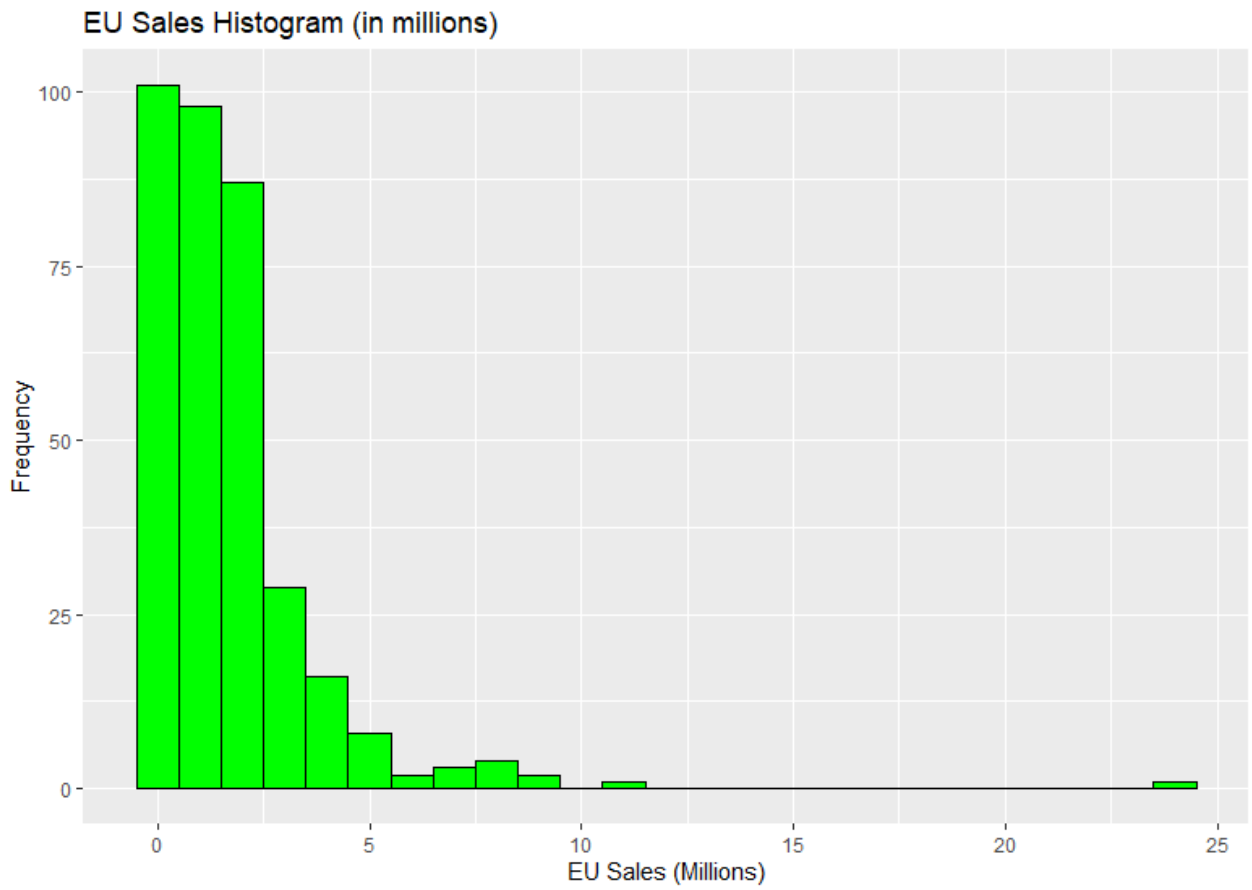

Figure 12: Histogram of EU Sales

Global Sales Histogram

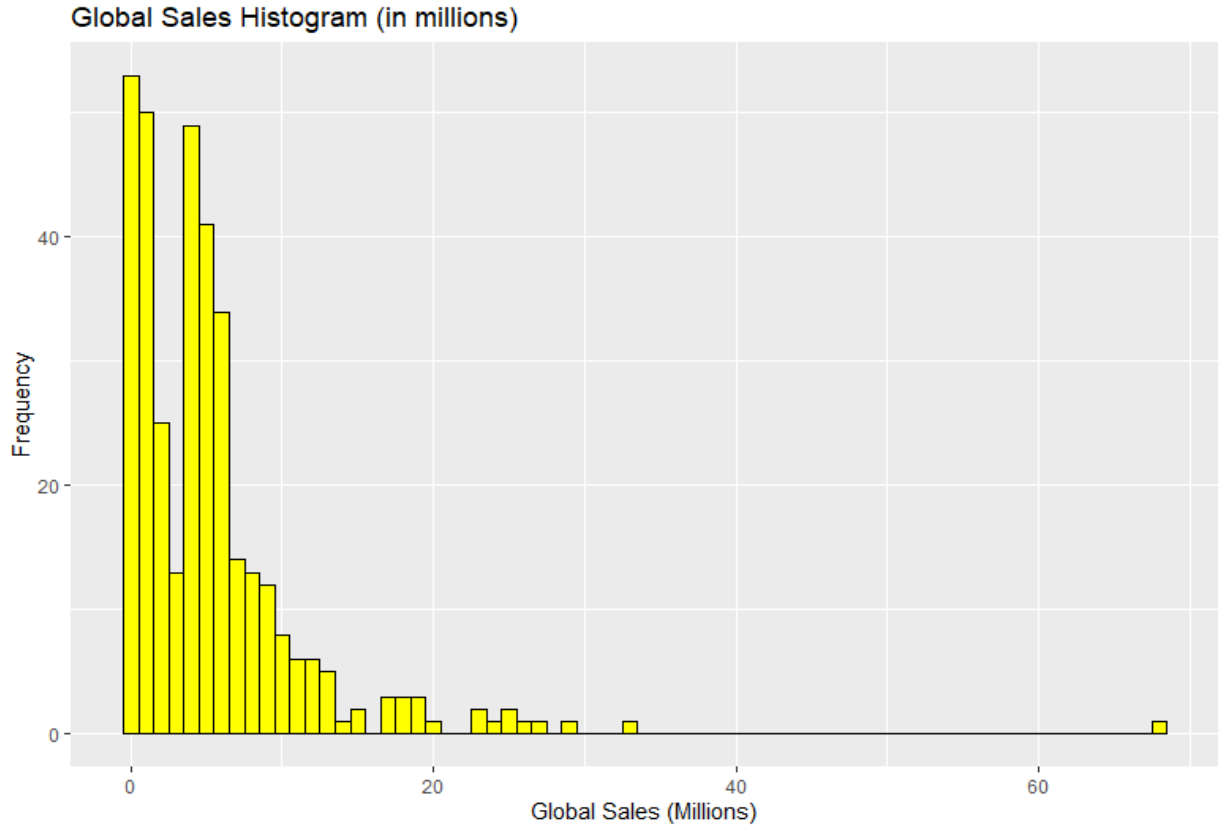

Figure 13: Histogram of Global Sales

Boxplot of NA Sales by Platform

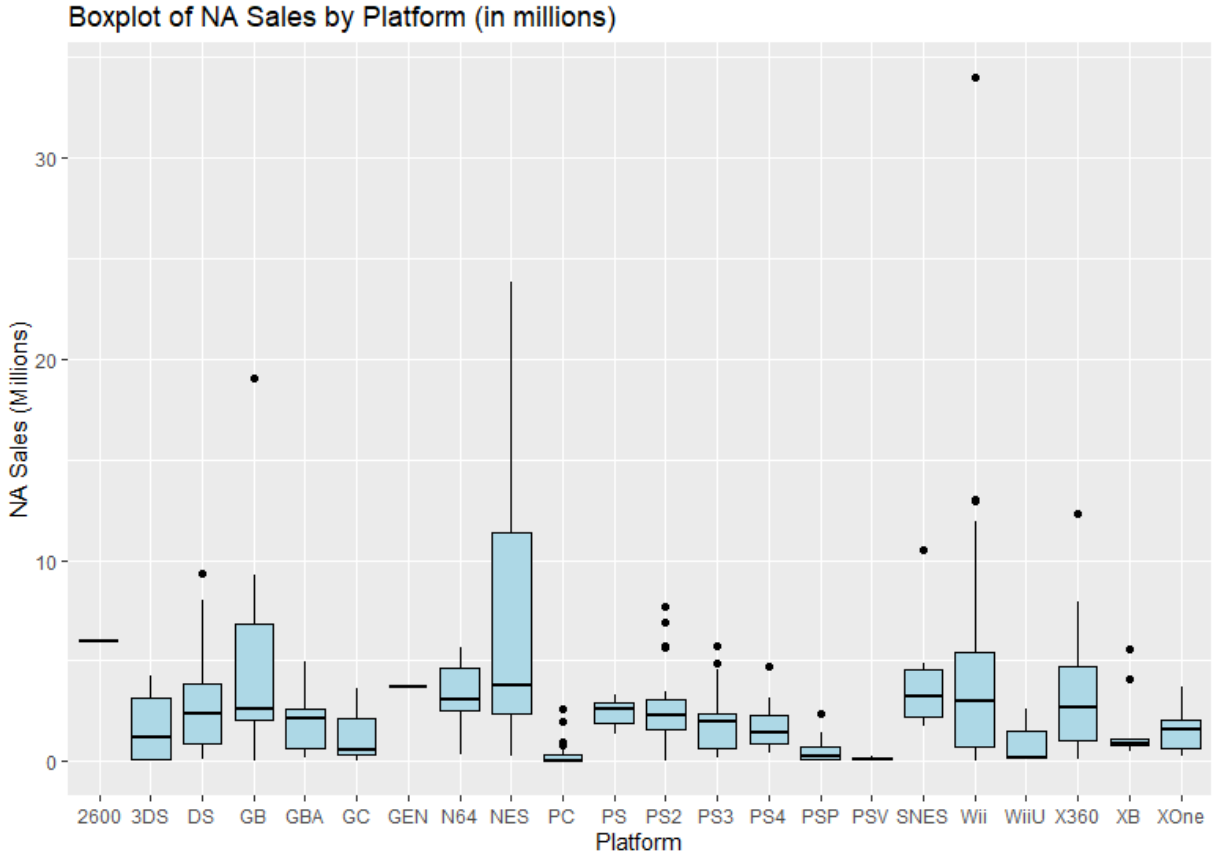

Figure 14: Boxplot of NA Sales by Platform

Boxplot of EU Sales by Platform

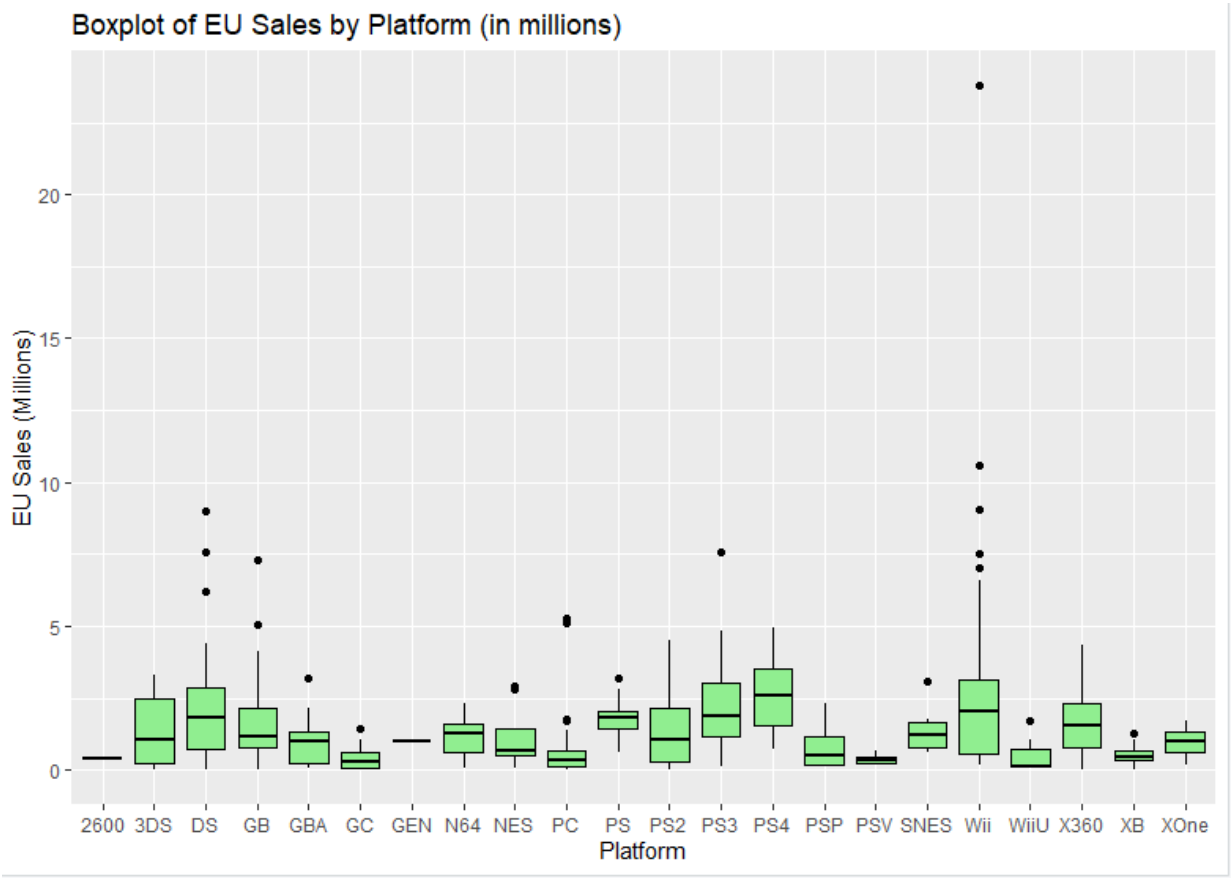

Figure 15: Boxplot of EU Sales by Platform

Boxplot of Global Sales by Platform

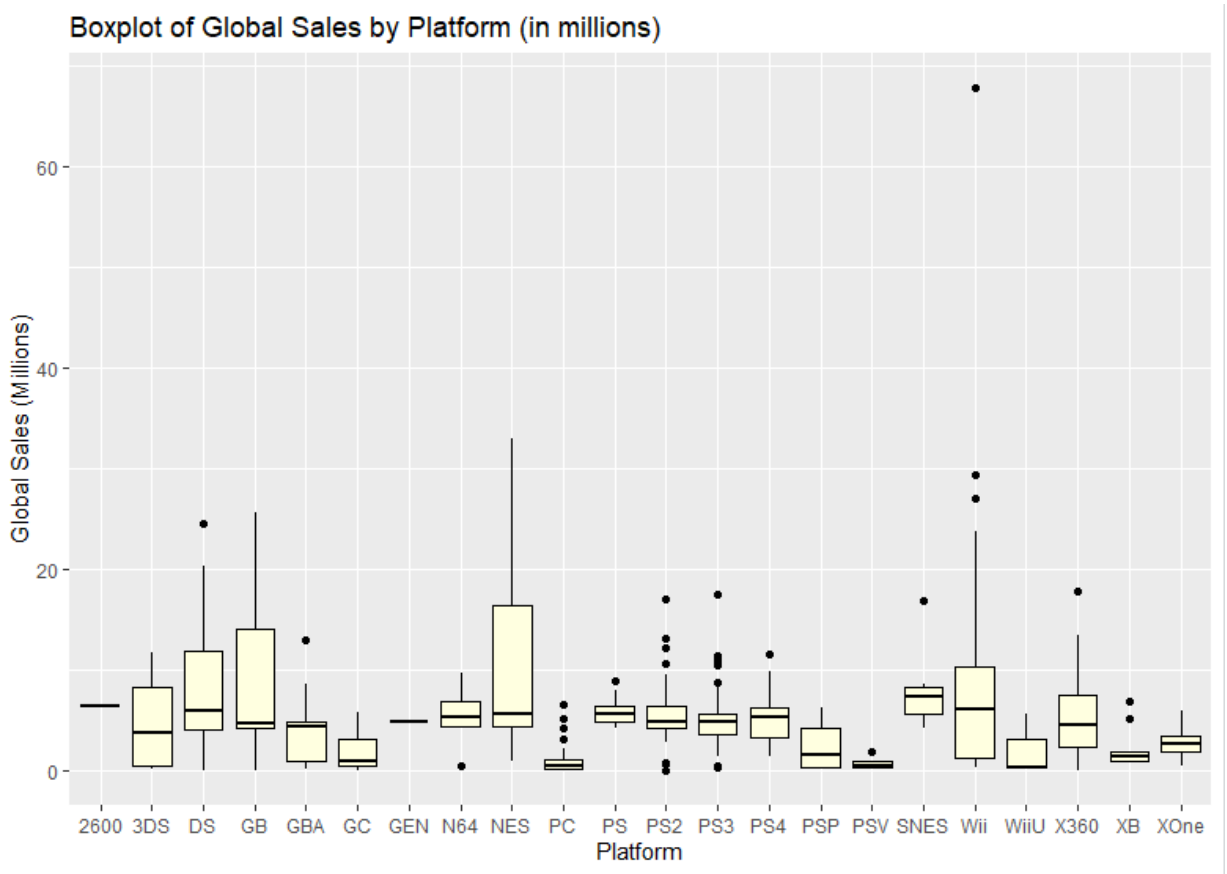

Figure 16: Boxplot of Global Sales by Platform

Section 5: Cleaning, Manipulating and Visualising the Data

In this section, we delve into the Turtle Games sales dataset to assess its reliability and gain crucial insights. We examine data normality, impact per product, skewness, kurtosis, and correlations to inform data-driven decisions. Understanding the dataset's characteristics is essential for effective business strategies.

Impact on Sales per Product

We calculated the total sales per product, revealing that product sales vary significantly. The minimum total sales for a product are 4.2 million units, while the maximum reaches 67.85 million units. This demonstrates the diverse performance of Turtle Games' products in the market.

Q-Q Plot Analysis

We used Q-Q plots to visually assess the normality of the sales data for North America (NA Sales), Europe (EU Sales), and global (Global Sales) regions. The Q-Q plots provide strong visual evidence that the sales data for Turtle Games does not follow a normal distribution. The pronounced curvature in these plots suggests that the data contains outliers with exceptionally high sales values. Understanding these deviations from normality is crucial for selecting appropriate statistical methods and addressing potential outliers in future analyses and decision-making processes

Normality Assessment

The Shapiro-Wilk normality tests for all three sales columns (NA Sales, EU Sales, Global Sales) showed extremely low p-values (< 2.2e-16), indicating strong evidence against normality. This means that the sales data does not follow a normal distribution. The Q-Q plots further visually confirmed the deviation from normality.

Skewness and Kurtosis

Skewness measures indicated right-skewed distributions for all three sales columns, with values around 4. This skewness suggests that the data is concentrated towards lower sales values but has a long tail of higher sales. High kurtosis values (around 30 to 41) implied heavy tails in the distributions, indicating the presence of outliers and the potential for extreme sales values.

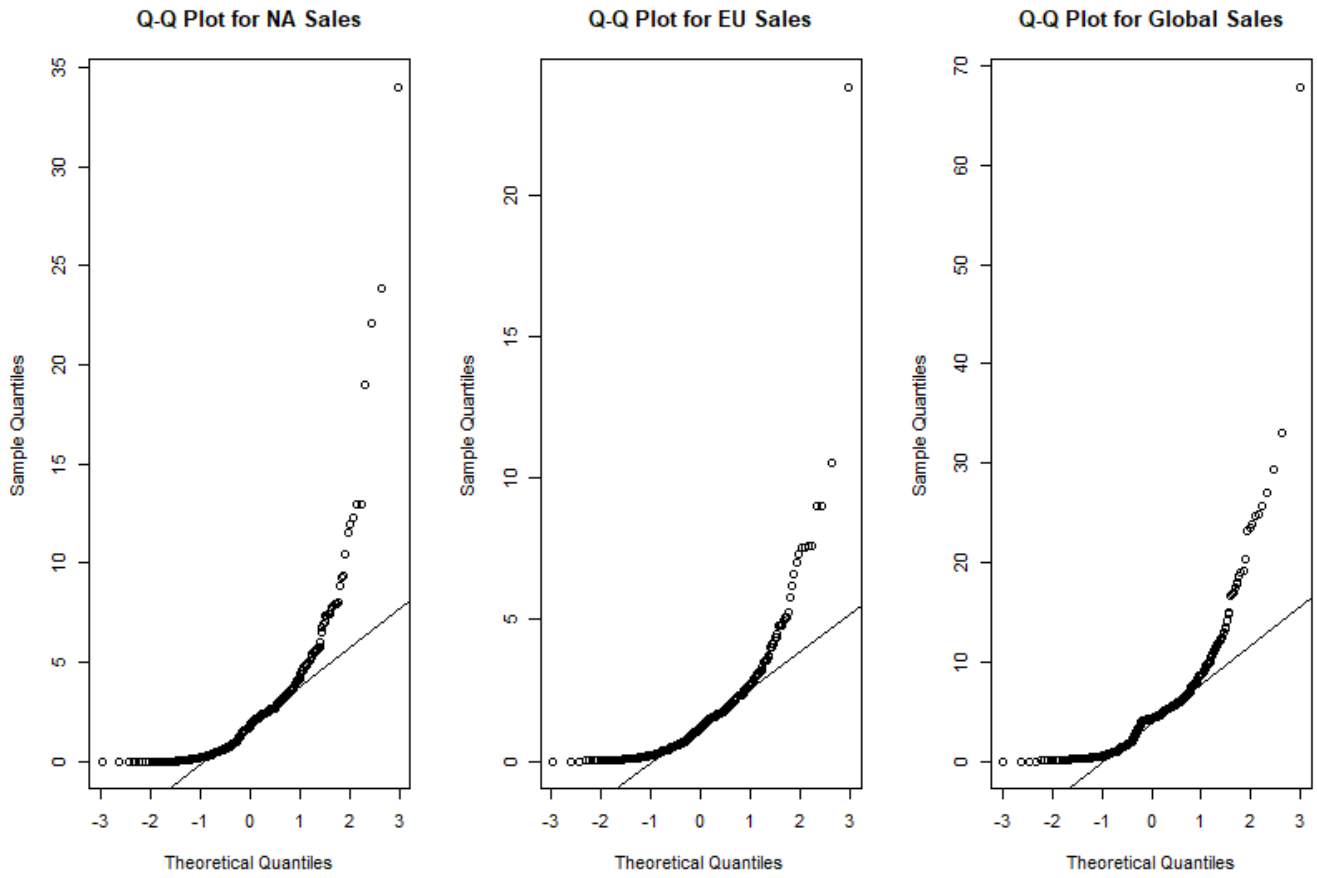

Figure 17: QQ Plot Analysis

Correlation Analysis

Correlation analysis revealed strong positive associations among sales columns. NA Sales correlated strongly with EU Sales (0.706) and Global Sales (0.935), while EU Sales and Global Sales had a robust positive correlation of 0.878, suggesting sales in one region are tied to others. This insight informs strategic decisions for Turtle Games.

Section 6: Making Recommendations to the business

In this section, an analysis has been conducted on Turtle Games' sales data using linear regression models and correlation analysis to yield insightful findings that are invaluable from a business perspective.

Simple Linear Regression Model

We initially built a simple linear regression model to analyse NA Sales' relationship with Global Sales. Key insights from the model's summary include significant coefficient values (p-value < 0.001) for the intercept and NA Sales. The R-squared value of 0.8741 indicates that 87.41% of Global Sales variability can be attributed to NA Sales, making it a valuable predictor for sales planning and expansion strategies.

Correlation Analysis

The correlation matrix demonstrates the relationships between sales columns. Notably, NA Sales has a strong positive correlation with both EU Sales (0.7055) and Global Sales (0.9349). Similarly, EU Sales exhibit a positive correlation with Global Sales (0.8776). These correlations imply that success in one region tends to translate into success in others, offering opportunities for coordinated marketing efforts and cross-regional strategies.

Multiple Linear Regression Model

Our multiple linear regression model highlights significant coefficients for NA Sales and EU Sales (p-value < 0.001), affirming their substantial impact on Global Sales. With an R-squared of 0.9687, this model explains about 96.87% of Global Sales variability, offering accurate predictions based on both regions.

Predictions for Business Planning

Finally, we utilised the multiple regression model to predict Global Sales based on provided NA and EU Sales values. These predictions offer immediate value to the business.

- Scenario Planning: Turtle Games can use these predictions to anticipate the impact of specific sales scenarios, allowing for more informed decision-making.
- Resource Allocation: By knowing the expected Global Sales outcomes for different NA and EU Sales scenarios, the business can allocate resources and marketing efforts more effectively.
- Risk Mitigation: Understanding potential sales outcomes in advance enables proactive risk management and the identification of high-potential growth scenarios.

Business Recommendations

- Cross-Region Strategy: Leverage interregional correlations for coordinated marketing and sales efforts to maximise overall sales.
- Outlier Management: Analyse and manage outliers, particularly high-value ones, to ensure accurate sales forecasts.
- Model Improvement: Continuously refine regression models, considering data characteristics like non-normal distribution, for enhanced accuracy.
- Market Expansion: Explore potential markets like North America for expansion through targeted advertising or localised game offerings.

Further Exploration

Further investigation into the impact of specific game genres, platforms, or marketing campaigns on sales can provide deeper insights. Additionally, the temporal aspect, such as trends over time, can be explored to adapt strategies accordingly.

In conclusion, the findings from this analysis serve as a foundation for evidence-based decision-making at Turtle Games. By leveraging the relationships between regions, addressing outliers, and refining predictive models, the business can optimise sales strategies and seize growth opportunities in the dynamic gaming industry.