Analysis on customer behaviour and trends to accelerate Turtle Games sales

Business Context

Preface: We're part of a data analysis team working with Turtle Games, a global game manufacturer and retailer. Our mission is to enhance sales by understanding customer behaviour and trends across various product categories. Turtle Games collects sales and customer review data, and our analysis covers customer loyalty, market segmentation, social data's marketing impact, product sales, data reliability, and regional sales relationships.

Business Case:

Turtle Games seeks deeper insights into loyalty program performance and demographic-specific segments for targeted marketing. This involves analysing past campaigns and utilising social data for future insights. To drive overall sales growth, understanding sales dynamics across regions is vital for informed investment decisions.

Five questions have been identified to guide the structure of the comprehensive data analysis.

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

Analytical approach

Python: To analyse the relationship between loyalty points and customer behaviour and demographics, we imported the dataset "turtle_reviews.csv" containing customer data using Python, leveraging libraries like numpy, pandas, matplotlib, seaborn, sklearn, and statsmodels. Our initial data preparation involved cleaning by removing null values, verifying data types, handling outliers in descriptive statistics, and eliminating unnecessary columns such as language and platform.

To analyse the relationship between loyalty points and independent variables (income, spending, age), we employed linear and multilinear regression models to identify the most influential variables. For multilinear regression, we partitioned the dataset into 80%

training and 20% test data. We also conducted a Variance Inflation Factor analysis to detect multicollinearity, finding none due to low correlation among predictor variables. With an impressive adjusted R-squared value of 83.5%, we confidently concluded that the multilinear regression model effectively captured the underlying relationships.

As we delve deeper into customer segments, we refined our dataset by retaining only the "income" and "spend" columns while performing necessary data cleaning. To gain deeper insights, we utilised visualisations, creating histograms and scatter plots. For our k-means clustering analysis, an unsupervised machine learning technique, we employed the Elbow Method to pinpoint the optimal number of clusters with the smallest sum of squares between data points. Additionally, we harnessed the Silhouette Method to identify the number of clusters that exhibit the greatest separation from other clusters.

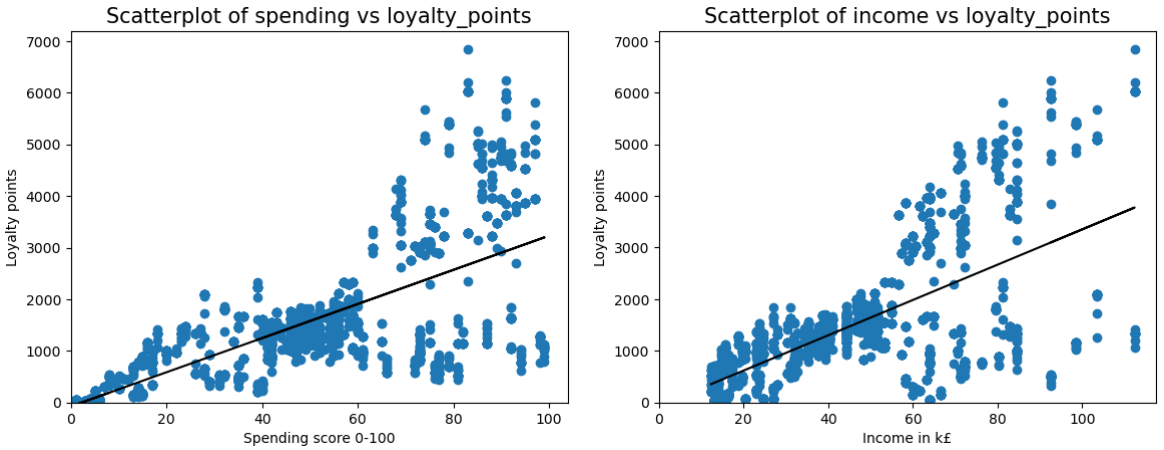
For processing customer reviews, we need libraries like NLTK, OS, WordCloud, Collections, TextBlob, and SciPy. The process involves data cleaning, including converting text to lowercase, removing punctuation, and eliminating duplicates. To extract meaningful insights from the text, we tokenize sentences and words, eliminate stopwords, and further stem and lemmatize words while assigning sentiment scores. The final output can be shown in a word cloud, plots and analysed with SentimentIntensityAnalyzer to derive their polarity and sentiment score.

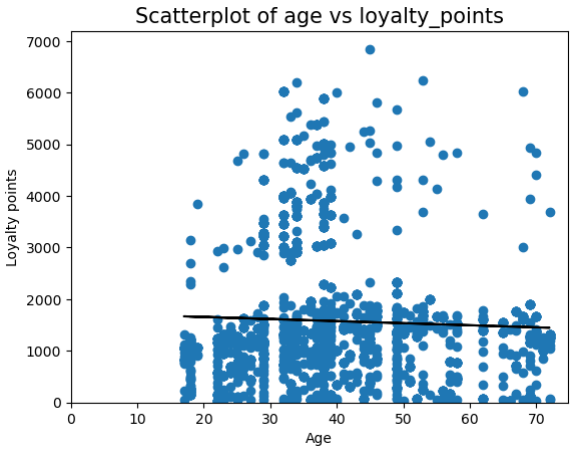
R and RStudio: To explore sales data relationships across regions, we used R and R Studio with libraries like tidyverse, dplyr, readxl, ggplot2, stats, moments, and BSDA. Our process involved data cleaning, aggregation of regional sales figures, and visual analysis with ggplot2. Regression models and sensitivity analysis helped identify correlations between EU and NA sales and their impact on global sales.

Limitations: While multicollinearity was not identified in this dataset, it's important to note that the presence of multicollinearity can depend on the dataset in use. The conclusions drawn from this analysis are contingent on the specific dataset and parameters applied, which may not generalise to all scenarios.

Visualisation and insights:

How do customers accumulate loyalty points?





Based on scatter plots, we observe a positive correlation between income and spend with loyalty points, as data points move away from the best-fit line as income and spend increase. For age, there is no clear relationship, although the best-fit line indicates a slight negative correlation. This is reflected in the R-squared values: spend explains 45.2% of variance, income 38%, and age only 0.2% for loyalty points.

The table below displays our multilinear regression results, explaining 83.3% of loyalty point variance, surpassing individual variables. Using this model, a one-point change in spend, income, or age can drive 34, 33, and 11 loyalty point changes. For instance, a user with a spending score of 60, income of 50, and age of 20 is likely to generate 1754 loyalty points. Turtle Games can focus more on spend and income demographics, given their strong correlation with loyalty points, and model their impact on future loyalty points.demographics given their strong correlation with loyalty points and model the impact of these variables on future loyalty points.

OLS Regression Results

Dep. Variable:	loyalty_points	R-squared:	0.835
Model:	OLS	Adj. R-squared:	0.835
Method:	Least Squares	F-statistic:	2692.
Date:	Sun, 08 Oct 2023	Prob (F-statistic):	0.00
Time:	19:48:33	Log-Likelihood:	-12227.
No. Observations:	1600	AIC:	2.446e+04
Df Residuals:	1596	BIC:	2.448e+04
Df Model:	3		
Covariance Type:	nonrobust		

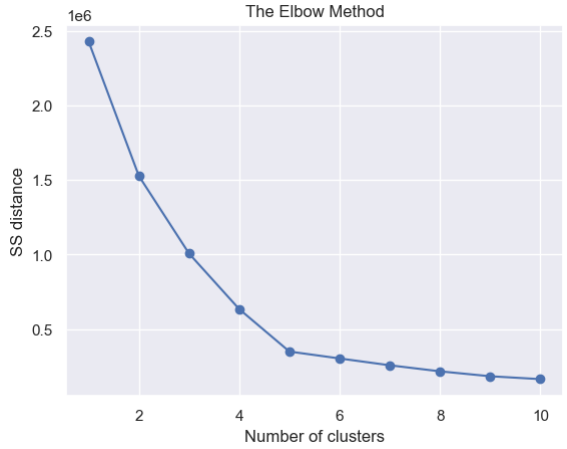
	coef	std err	t	P>|t|	[0.025	0.975]
const	-2203.3411	58.253	-37.824	0.000	-2317.601	-2089.081
spending_score	34.2661	0.503	68.101	0.000	33.279	35.253
renumeration	33.6612	0.554	60.806	0.000	32.575	34.747
age	10.9350	0.944	11.579	0.000	9.083	12.787

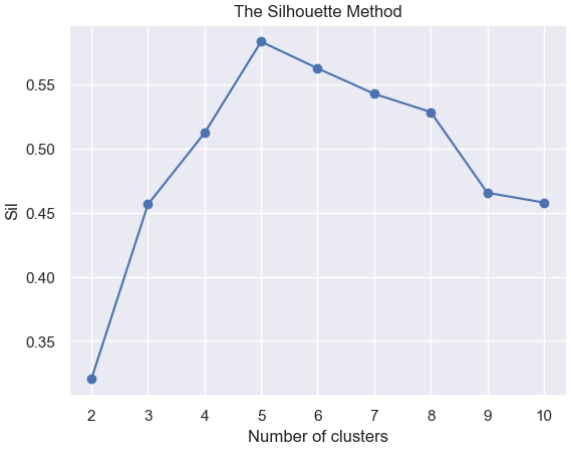
Omnibus:	21.363	Durbin-Watson:	1.984
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23.304
Skew:	0.238	Prob(JB):	8.70e-06
Kurtosis:	3.352	Cond. No.	381.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

How groups within the customer base can be used to target specific market segments?

With the help of the Elbow Method and Silhouette Method explaining the number of clusters required for the smallest distance among data points within clusters and the largest distance between clusters, we can see that 5 clusters are ideal for our k-means cluster analysis.



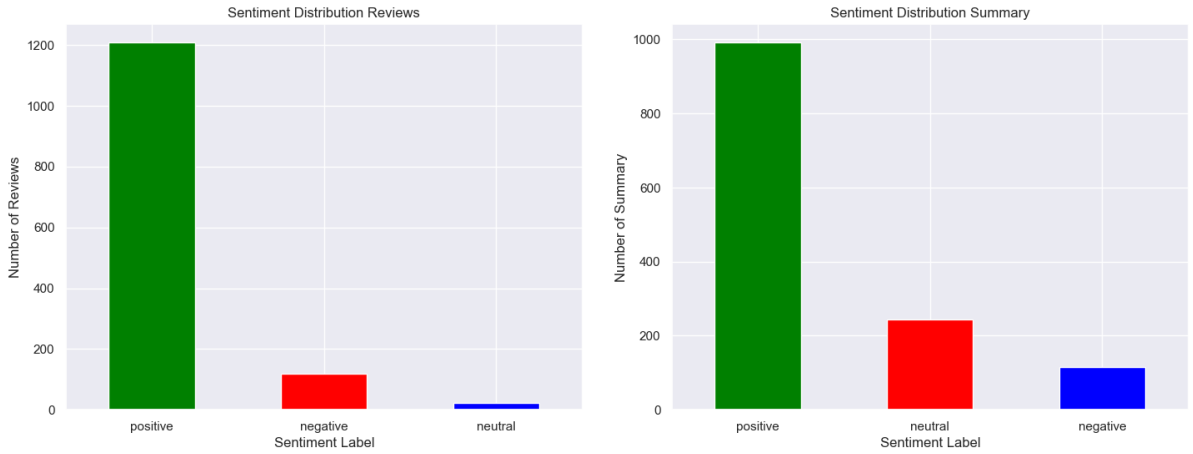


The below graph showing the 5 clusters based on spend and income helps us to identify 5 customer segments namely:

- **Segment 0: "Luxury Enthusiasts"**
  - Annual Income: 65k to 112k, Spending Score: 60-100
  - High-income customers with strong spending habits, seeking luxury, personalization, and premium products. Target with exclusive, high-quality offerings and tailored loyalty programs.
- **Segment 1: "Balanced Shoppers"**
  - Annual Income: 35k to 65k, Spending Score: 40-60
  - Moderate-income customers with practical spending habits. Appeal with affordability, quality, and convenience. Offer mid-range products with versatile features.
- **Segment 2: "Savvy Savers"**
  - Annual Income: 65k to 112k, Spending Score: 60-100
  - Higher-income customers prioritizing savings over luxury. Emphasize value, quality, and long-term benefits. Consider loyalty programs or savings discounts.
- **Segment 3: "Spending Optimists"**
  - Annual Income: 10k to 38k, Spending Score: 60-100
  - Lower-income customers with high spending scores. Highlight affordability, discounts, and budget-friendly options. Offer flexibility, sales, and value bundles.
- **Segment 4: "Budget Essentials"**
  - Annual Income: 10k to 38k, Spending Score: 0-40
  - Customers with lower incomes and cautious spending. Focus on cost-saving options, essentials, and budget-friendly items. Improve the in-store experience for increased spending.

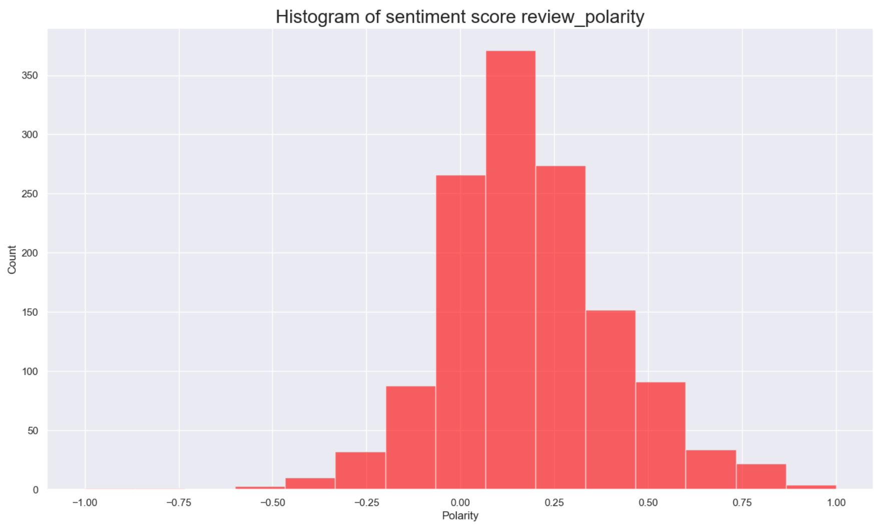
How social data (e.g. customer reviews) can be used to inform marketing campaigns?

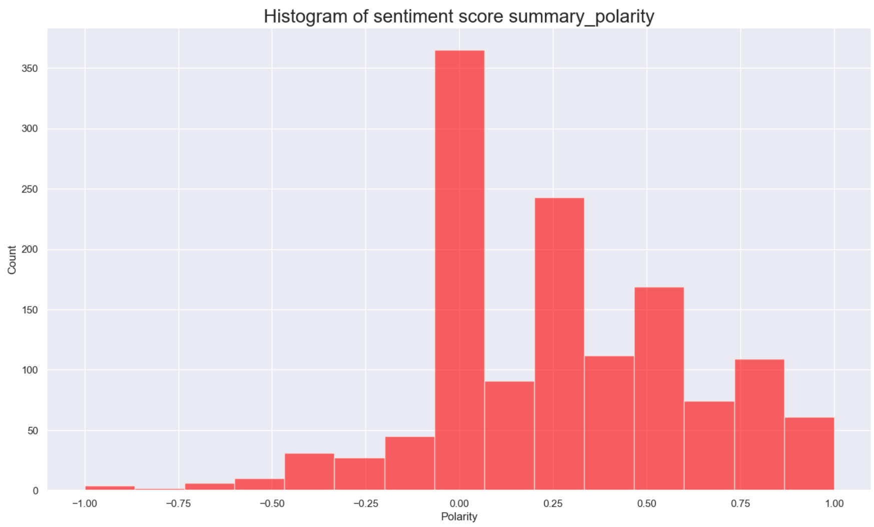
Customer reviews and summaries reveal a predominantly positive sentiment, offering valuable insights for Turtle Games' future planning. By linking these insights to specific products and customer profiles, the company can improve targeting and messaging strategies. Frequent words like "game," "great," and "fun" underscore this positive sentiment, reflecting enjoyable customer experiences. Accurate sentiment categorization in the top 20 reviews is an impressive achievement, highlighting the effectiveness of sentiment analysis and the importance of customer feedback. In conclusion, Turtle Games can utilise this positive sentiment to refine targeting, enhancing customer satisfaction and sales.




Word Cloud of Most Common Lemmatized Words Reviews
game
play
one
card
great
tile
make
time
fun
player
book
use
set
really
love
piece
little
good
new
well

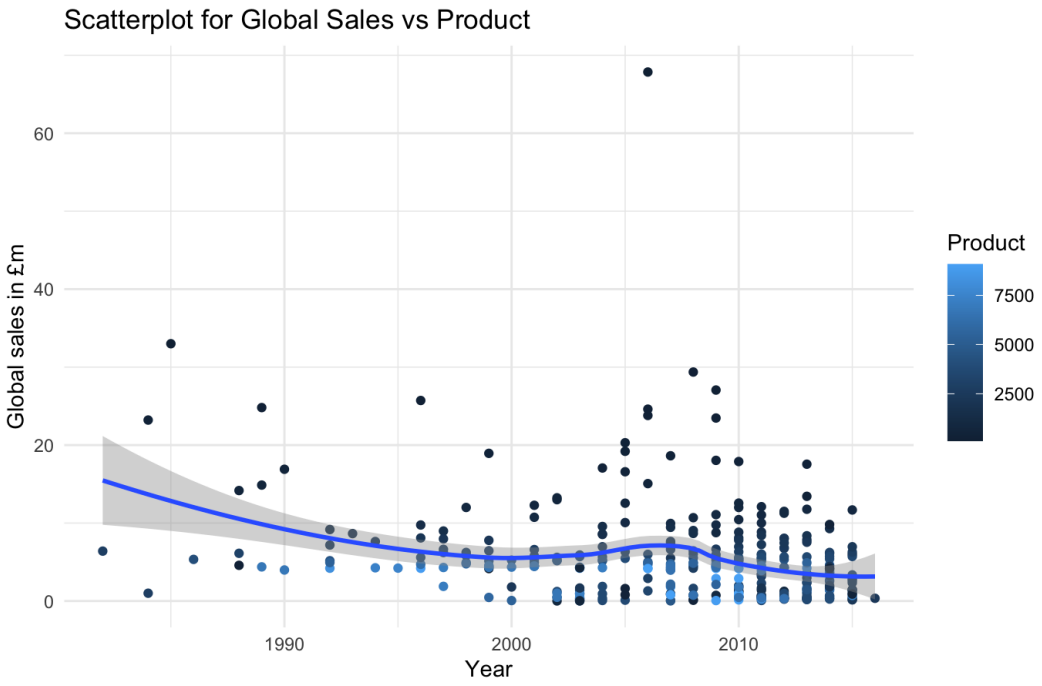

Word Cloud of Most Common Lemmatized Words Summary
game
great
fun
love
good
expansion
kid
cute
book
product
excellent
quality
gift
nice
awesome
tile
little
perfect
easy
child

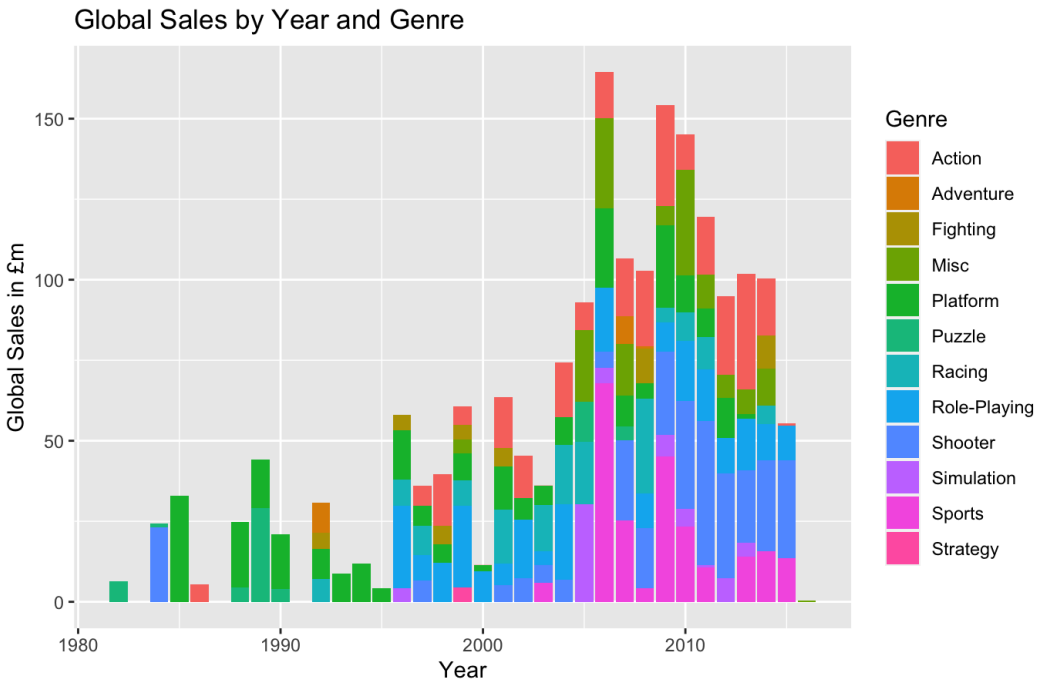

Histogram of sentiment score review_polarity
Count
Polarity
0
50
100
150
200
250
300
350
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

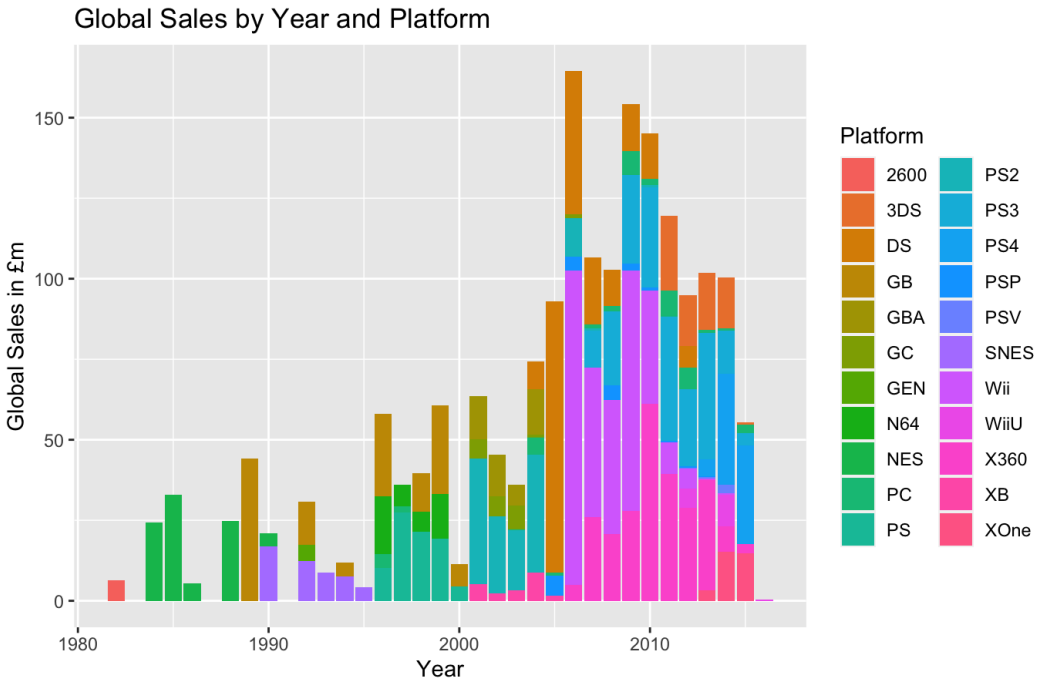

Histogram of sentiment score summary_polarity
Count
Polarity
0
50
100
150
200
250
300
350
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

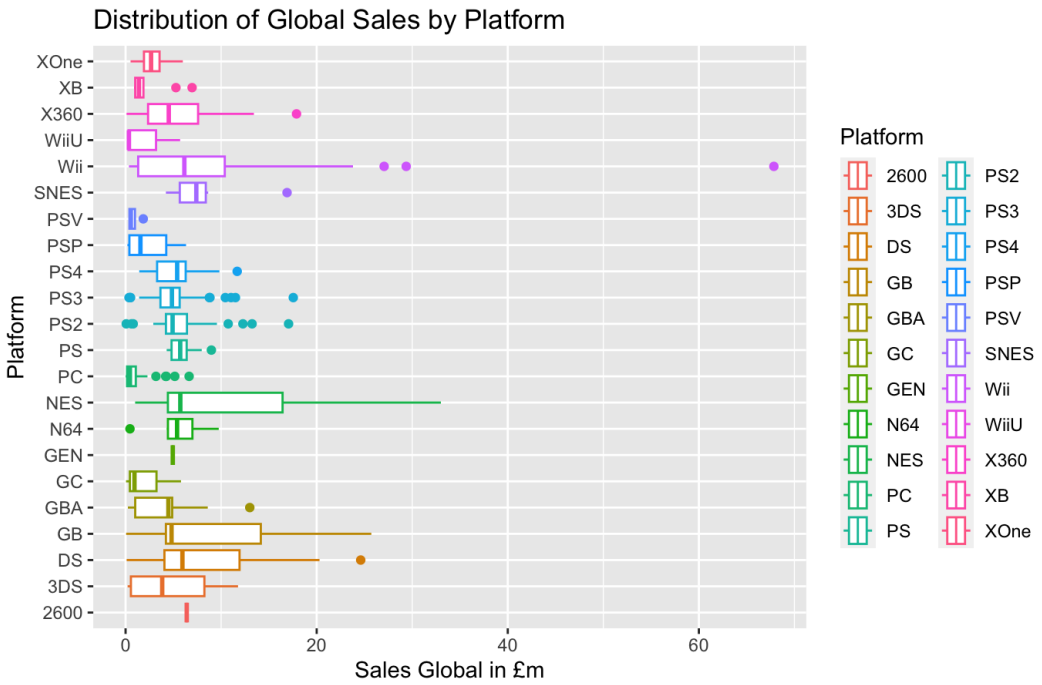
The impact that each product has on sales?

These graphs portray popular platforms such as X360, Wii, NES, PS4, DS, and 3DS, revealing considerable sales variability, evident in the distribution graph. Notably, products with lower IDs stand out as outliers, signifying robust sales performance. The smoothed fit line underscores a pre-2000 era of low global sales and high error margins, followed by a post-2000 surge in sales and product volume. The bar charts illustrate a significant growth in volume, genre diversity, and platform variety around 2010, driven by technological advancements in gaming and global accessibility.





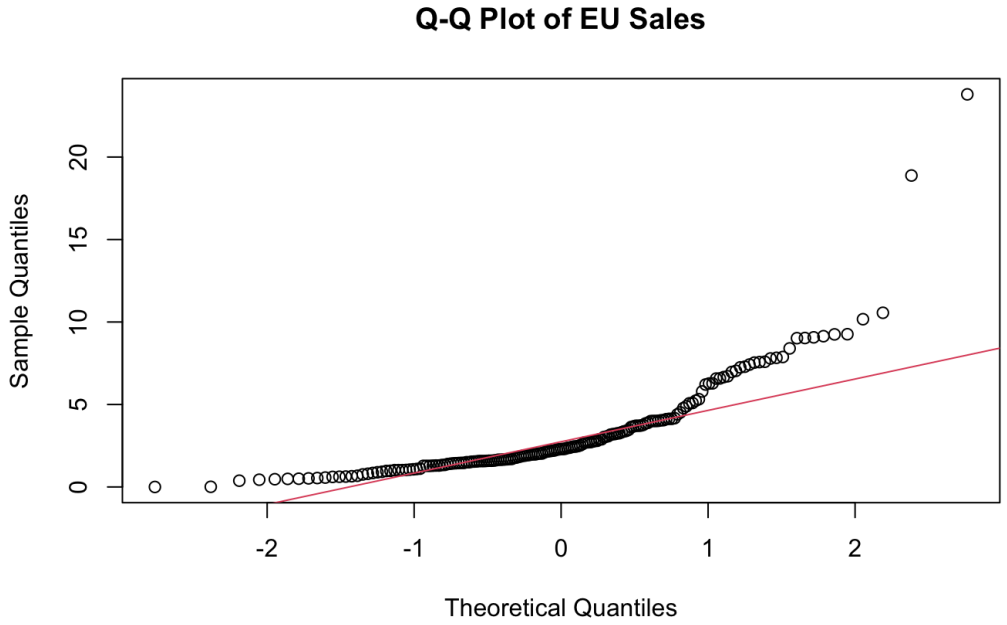


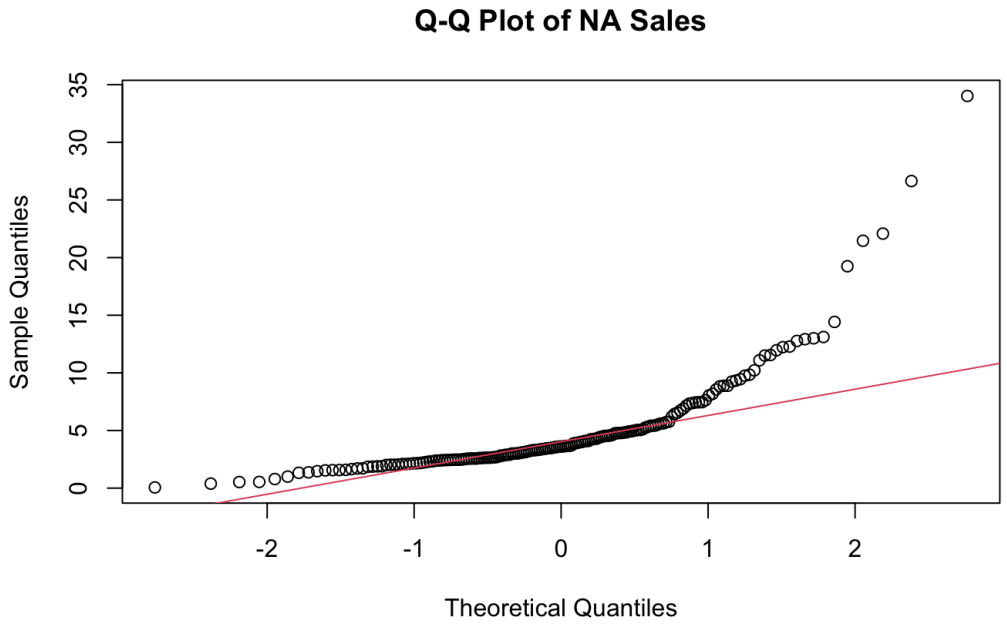


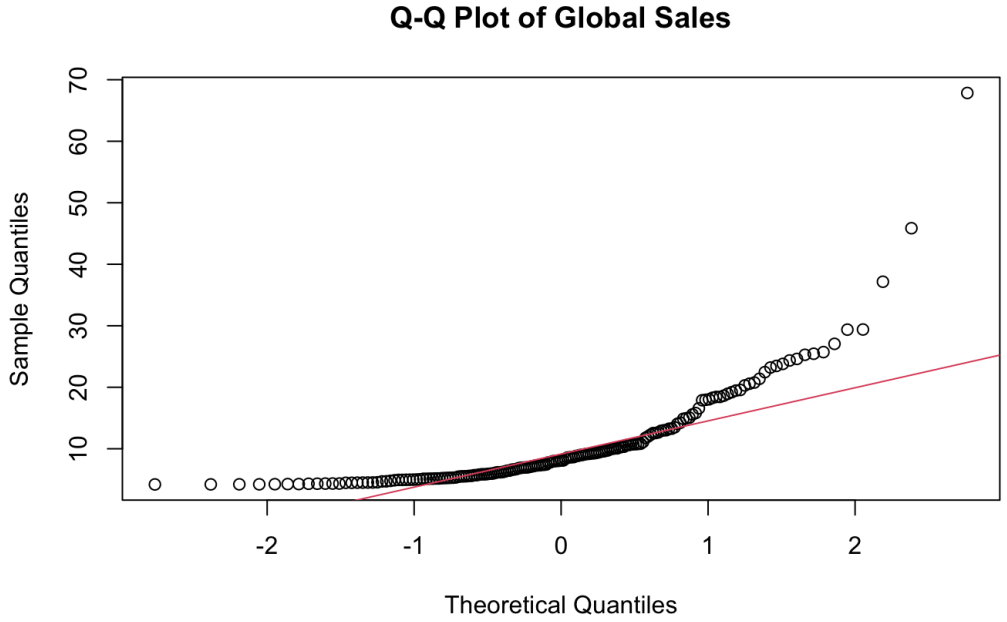
How reliable the data is (e.g. normal distribution, skewness, or kurtosis)?

- EU sales has a skewness of 2.89, Kurtosis of 16.23 and Shapiro p-value lower than 5%.
- NA sales has skewness of 3.05, Kurtosis of 15.60 and Shapiro p-value lower than 5%.
- Global sales has a skewness of 3.07, Kurtosis of 17.79 and Shapiro p-value lower than 5%.
- Based on these results Q-Q Plots we can derive that sales do not follow a normal distribution given their Shapiro-Wilk test with significant results and p-value lower than 0.05.

- The Kurtosis for all sales numbers of more than > 15 indicate that the distribution has heavy tails and that there are more extreme values in the tails compared to a normal distribution. This suggests that the distribution has a greater number of outliers or extreme values.
- A positive skewness value above 3 indicates that the distribution is skewed to the right. In a positively skewed distribution, the tail on the right side of the distribution is longer or fatter than the left side. This suggests that there may be a concentration of data points on the left side of the distribution, with some extreme values stretching the distribution to the right.

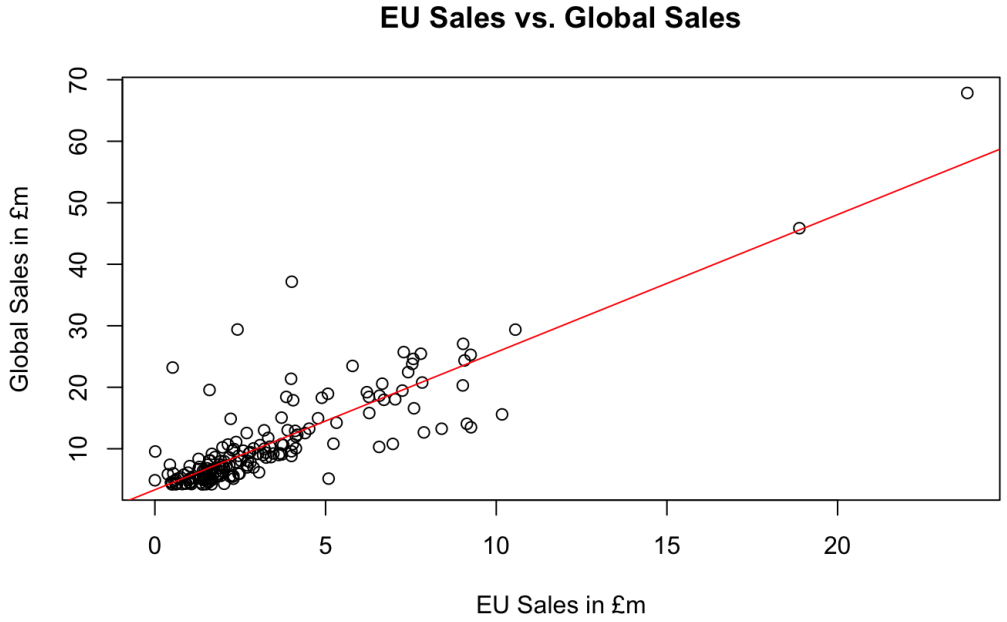


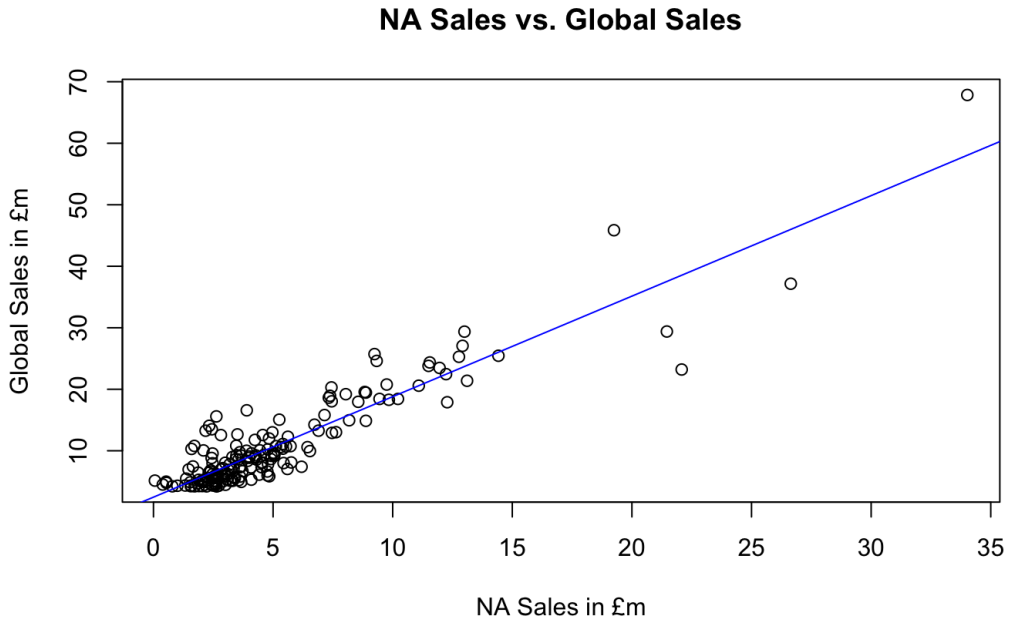




What the relationship(s) is/are (if any) between North American, European, and global sales?

- EU sales has a positive correlation with global sales of 0.84.
- North America sales has a positive correlation with global sales of 0.91.
- Both EU and North American sales have an impact on Global sales and are independent variables that are statistically significant given their p-value close to 0%. Both EU and NA variables have a high adjusted R-Squared of 96% based on multilinear regression and thus account for the majority of global sales variance. Both regions have a positive coefficient, meaning an increase in NA and EU sales will benefit global sales, which is a given, based on the fact that global sales is cumulative of all sales worldwide.
- When testing our multilinear regression model for accuracy with a testing sample of 5, we can see that our predictions are aligned with the test data in a majority of instances with little variance, i.e. EU sales of £23.8m and NA sales of £34.0m results in a predicted value of 68.1, very close to our actual data of £67.8m.
- Limitations: Multilinear regression models may not be appropriate for highly correlated independent variables.





Patterns and predictions

Throughout our analysis we were able to identify key patterns such as the following:

1. Income, spend and age of customers are highly correlated with their loyalty points explaining almost 83% of the variance in loyalty points.
2. Turtle games has 5 customer groups split into Luxury Enthusiasts, Balanced Shoppers, Savvy Savers, Spending Optimists, Budget Essentials with their own spending habits and respective income.
3. Customer reviews and summaries convey predominantly positive sentiment, providing valuable insights for Turtle Games to enhance targeting and messaging

strategies by linking these insights to specific products and customer profiles, with frequent words like "game," "great," and "fun" reflecting enjoyable experiences and accurate sentiment categorization in the top 20 reviews demonstrating the effectiveness of sentiment analysis, ultimately enabling Turtle Games to improve targeting, increase satisfaction, and boost sales.

4. Post-2000, video game sales surged, resulting in a broader range of genres and titles. Turtle Games can boost sales by targeting popular products across key platforms and genres.
5. EU sales and North America sales exhibit positive correlations with global sales (0.84 and 0.91, respectively), both serving as statistically significant independent variables with p-values near 0%. In a multilinear regression model, both regions have a high adjusted R-squared of 96%, explaining a substantial portion of global sales variance.

Recommendation:

In summary, Turtle Games can boost customer targeting by aligning offerings with income and spending patterns, leveraging correlations with loyalty points and utilising positive sentiment from reviews. Prioritising North America and the EU can drive long-term revenue growth. Key actions:

- **Enhance Loyalty Programs:** Customise and strengthen loyalty programs for each customer segment, fostering long-term loyalty and repeat sales.
- **Segmented Targeting:** Tailor marketing and products to customer segments based on income and spending for increased engagement and sales.
- **Leverage Positive Sentiment:** Utilise positive reviews in marketing to enhance customer satisfaction and sales.
- **Popular Genres and Platforms:** Invest in popular video game genres and platforms to expand the product range and boost sales.
- **Forecasting:** Utilise the multilinear regression model to predict global sales using EU and NA sales data, enhancing clarity and improving business planning.
- **Continuous Data Analysis:** Monitor customer purchase data for evolving trends and adapt marketing strategies in real-time to drive sales growth.