Turtle Games

KW Data Analytics

24 April 2023

Contents

Table of Figures

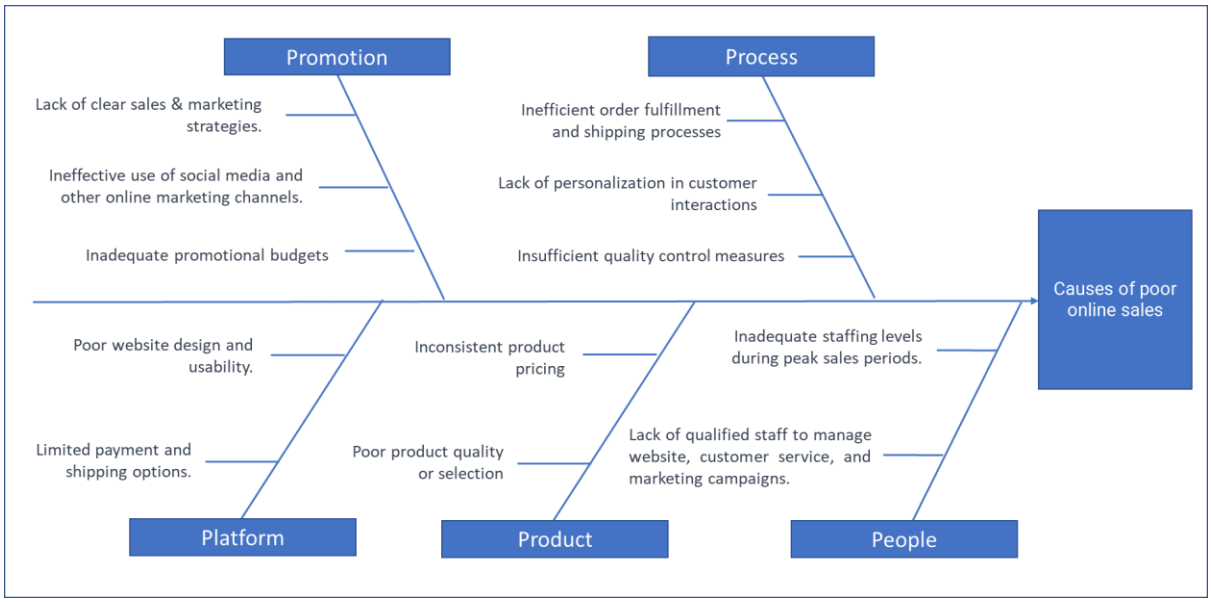
1. Background

Turtle Games has commissioned KW Data Analytics to conduct an analysis of customer trends in order to improve overall sales performance.

The marketing department wants to understand how to assess customers by loyalty, how to group customers for targeted campaigns, and how to best use online reviews for marketing and customer engagement.

The sales department wants to understand the impact of products on sales and how regional sales can predict Global sales. This will help optimize resource allocation and improve the effectiveness of sales tactics.

Figure 1. Fishbone diagram showing causes of poor online sales.



2. Analytical Approach

2.1 Data ingestion and wrangling

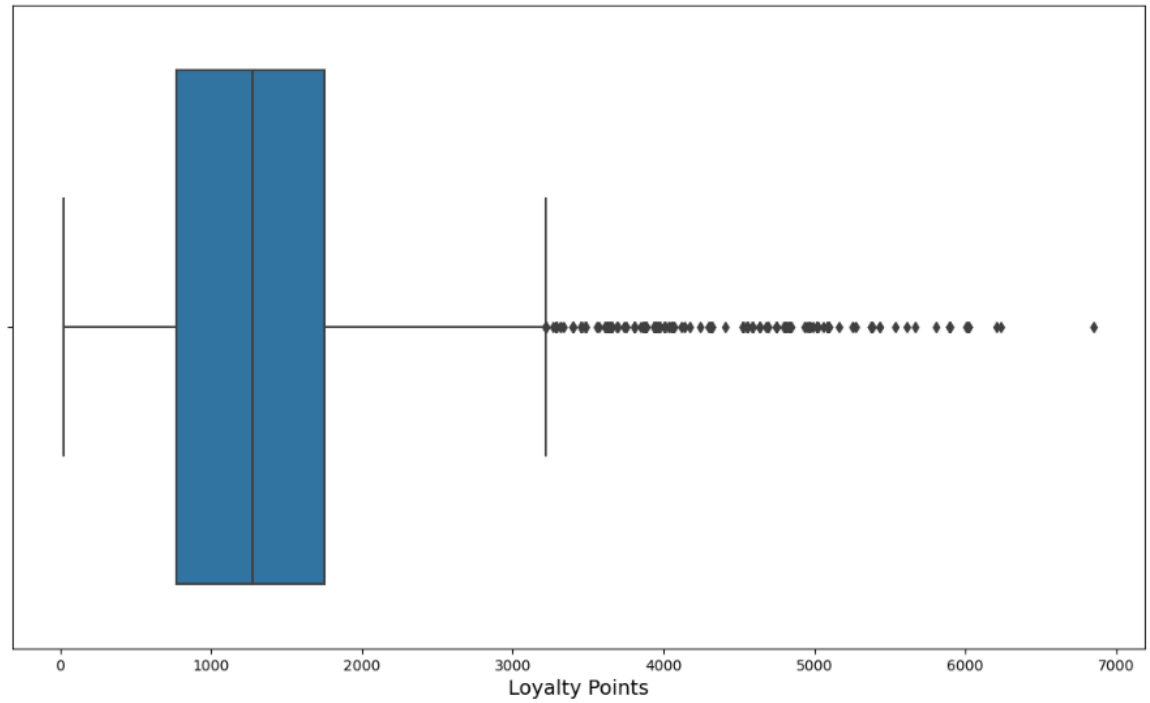
- Short code, descriptive function/variable names, concise syntax, and insightful commentary.
- Data files added to directories and libraries/packages imported.
- No unusual data, spelling errors, or duplicates found. Replaced two N/A turtle_sales Year values.
- For turtle_sales, Year and Product changed to categorial. New column, Other_sales, calculated.

2.2 Turtle_reviews data (analysed with Python)

2.2.1 Data exploration and outliers

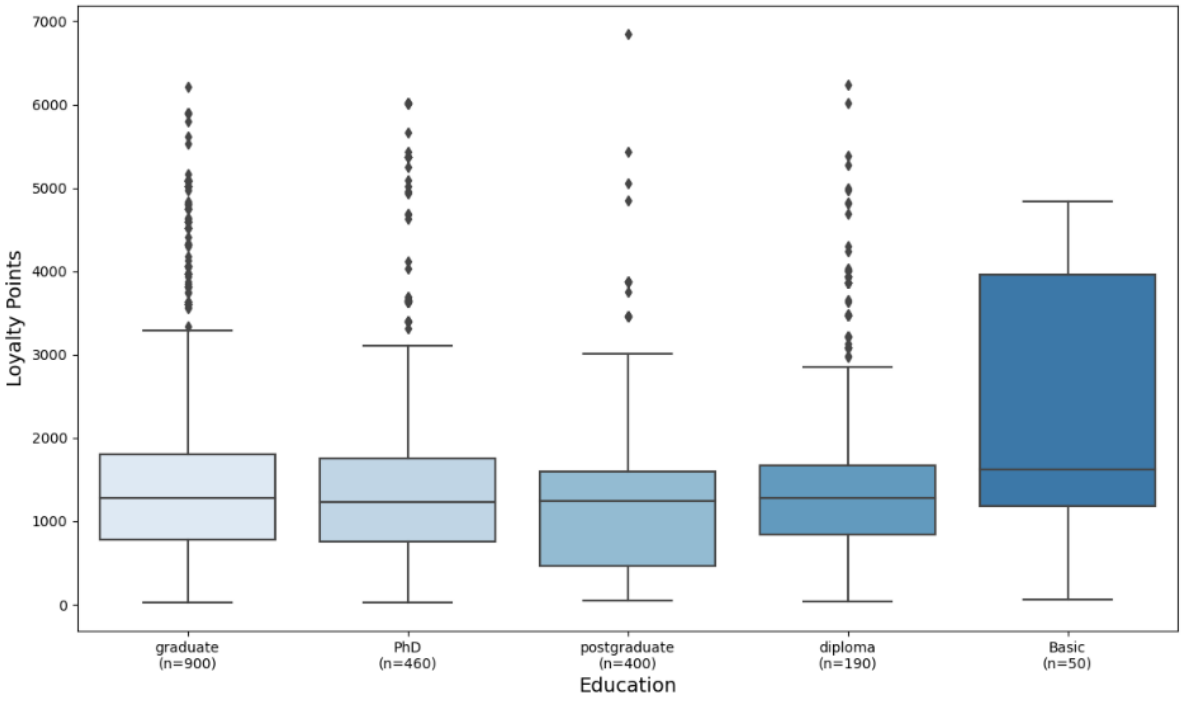
- No outliers detected for income, spending score, or age. Outliers for loyalty points not removed.

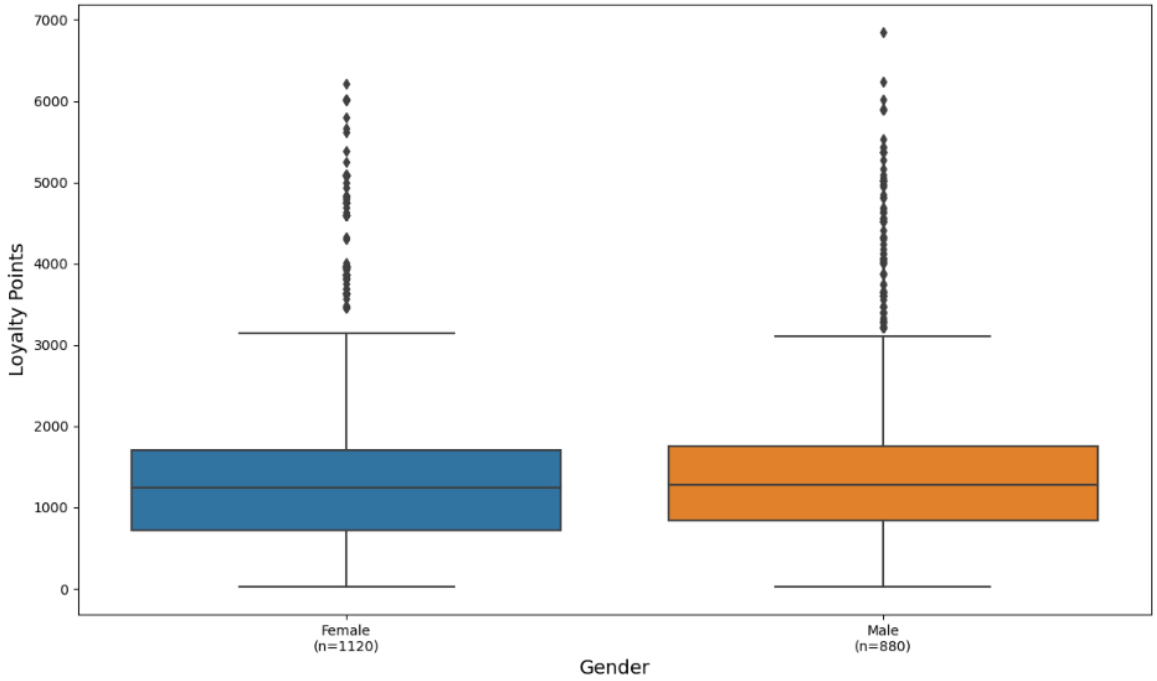
Figure 2: Boxplot showing outliers for Loyalty points.



- Gender showed no influence on loyalty points. Basic education had some impact, but observations too low.

Figure 3: Boxplot showing impact of gender (top) and education(bottom) on Loyalty points.





- Correlation between loyalty points and income and spending score, but not age.

Figure 4: Correlation matrix showing relationships between numerical variables.

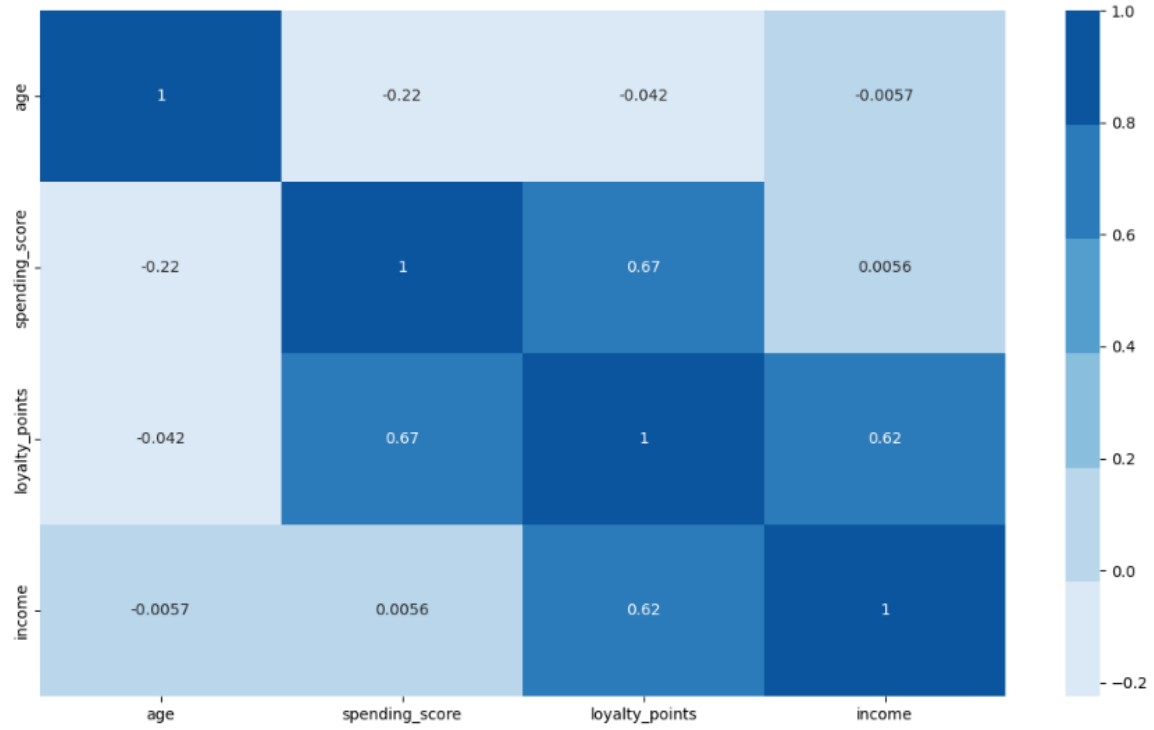
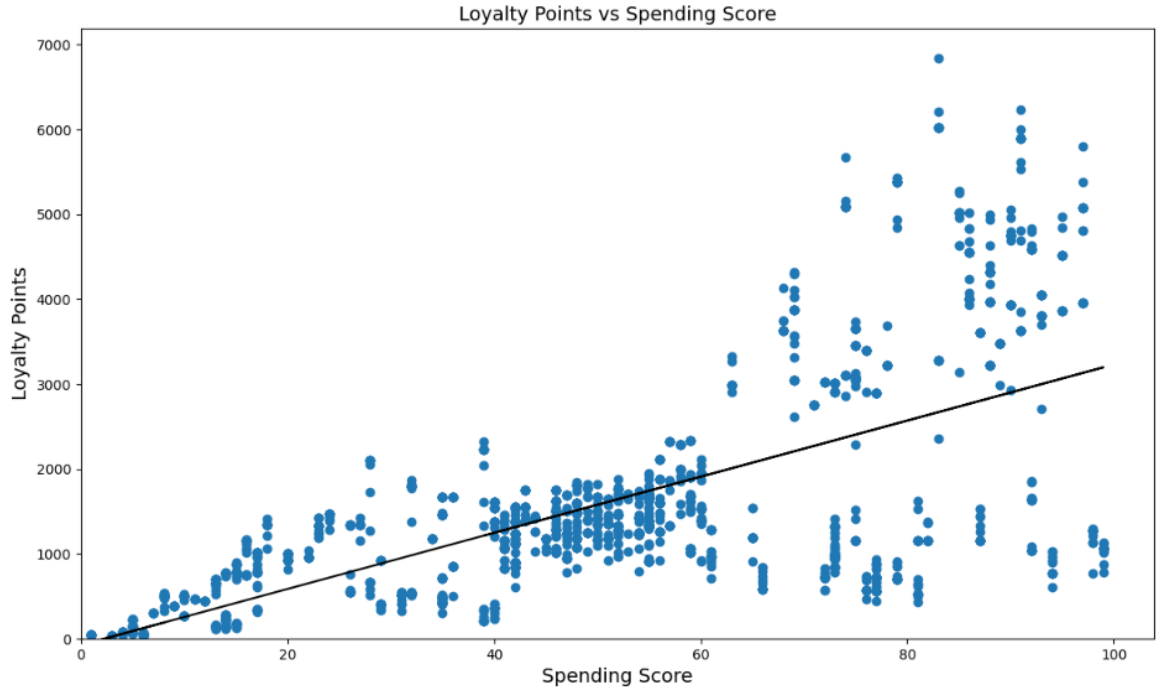


Figure 5: Scatterplot showing relationship between age and loyalty points.

2.2.2 Simple Linear regression with OLS method

- Models created and regression lines fitted for loyalty points versus spending score, income, and age.
- Accuracy assessed using R squared and data checked for heteroscedasticity.
- Spending score and income showed heteroscedasticity. Columns log transformed and new models created.
- Similar models created with machine learning and LinearRegression() function (requested by company).

Figure 6: Cone shaped scatterplot (heteroscedasticity) for loyalty points vs spending score.



2.2.3 Multiple Linear regression

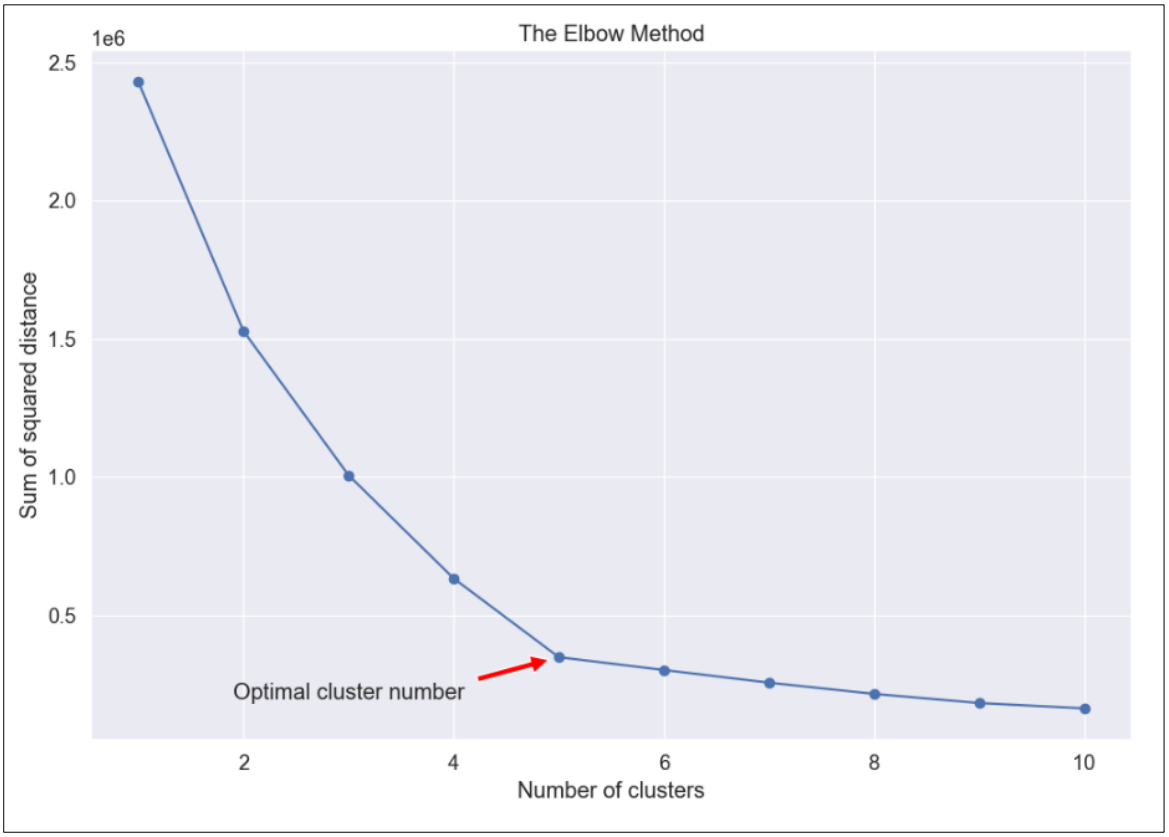
- Model fitted and loyalty points predicted using spending score and income.
- Training and test data created (70:30), model fitted using training data, and loyalty points predicted for test data. Accuracy assessed using R squared and other parameters.
- Data checked for heteroscedasticity and multicollinearity.

2.2.4 Clustering of customer groups by income and spending score

- Income and spending score clusters assessed using pairplots.
- Elbow and silhouette methods used to determine optimal cluster number (k = 5)
- K-mean clustering performed using different k values (4-6).

Figure 7: Pairplots showing distribution of income vs spending score.

Figure 8: Elbow method showing optimal cluster number as five.



2.2.5 Sentiment analysis of customer reviews using natural language processing

- Text pre-processed and tokenised. Data concatenated from each row, word frequency determined, and wordclouds generated.

Figure 9: Wordcloud for Review column before stopwords removed.

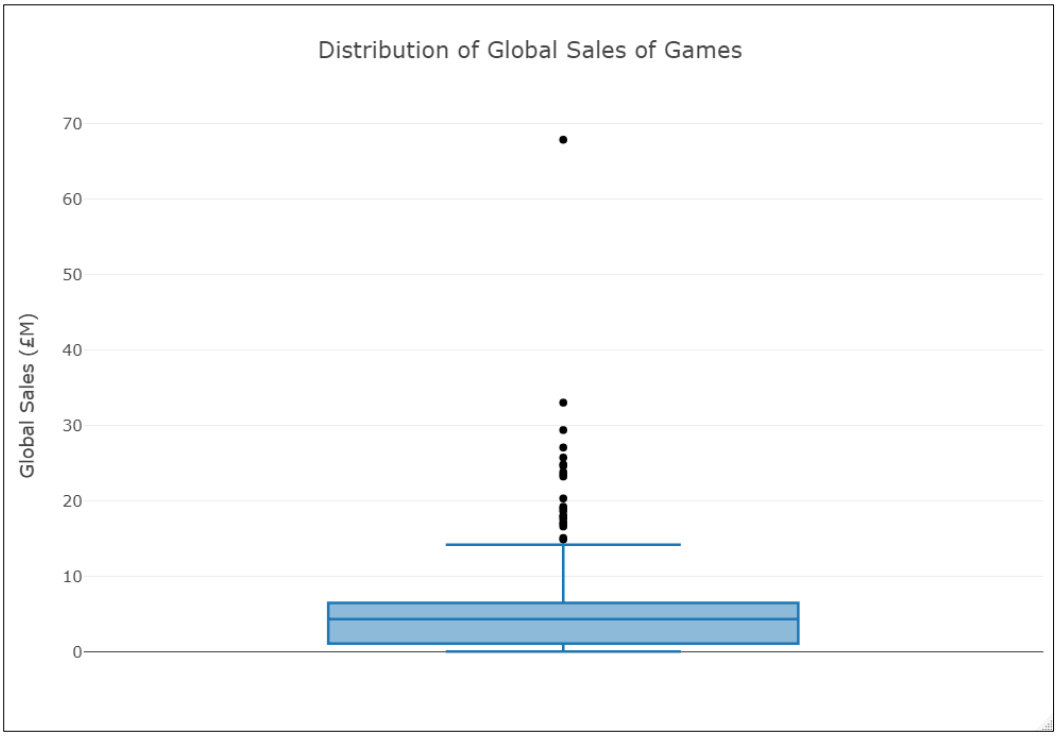
- Stopwords removed and new wordclouds generated. Most common words extracted, and polarity determined.
- Polarity and subjectivity calculated for all reviews and summaries and histograms used to visualise distribution of overall sentiment.
- Top 250 positive and negative reviews and summaries by polarity identified, with wordclouds and frequency distribution for top terms generated. Certain negative words investigated further.
- Top 20 products by negative polarity identified. Certain products investigated further.

2.3 Turtle_sales data (analysed with R)

2.3.1 Data exploration and outliers

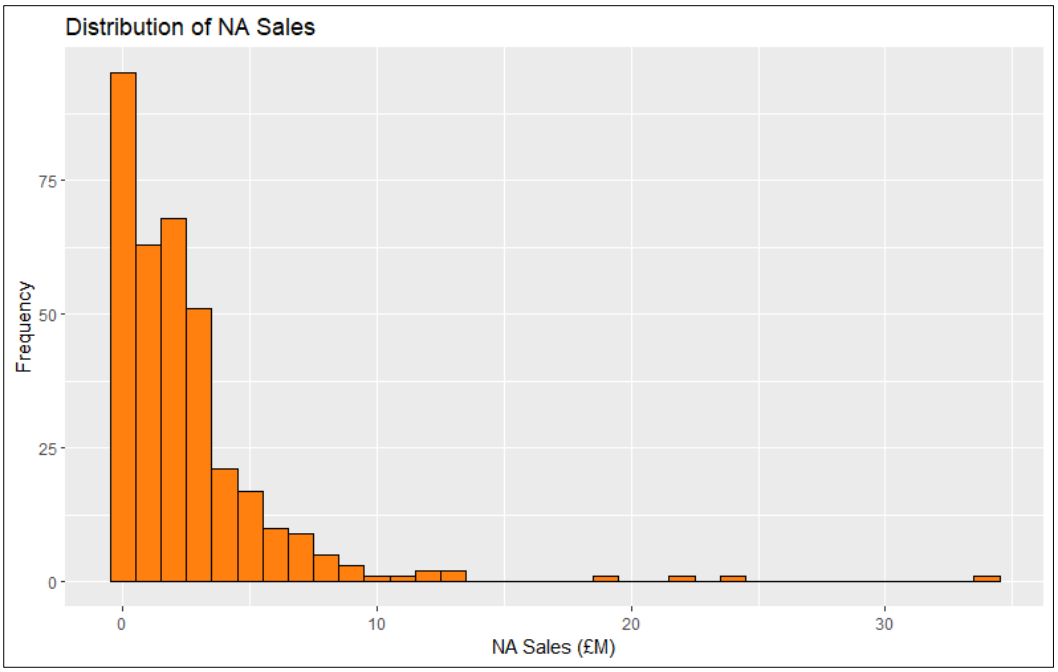
- Outliers seen in all sales columns but not removed.

Figure 10: Boxplot showing outliers for Global sales.



- Histograms used to see distribution of sales columns, and all skewed heavily to the right.

Figure 11: Histogram showing distribution of NA Sales.



- Boxplots assessed distribution and impact of platform, genre, publisher, and year on Global sales.
- Barcharts created to see game count by platform, genre, publisher, and year. Faceted barcharts compared global and regional grouped sales by platform, genre, publisher, and year.

2.3.2 Impact of product on sales

- Stacked barchart created for top 50 products globally, split by region.
- Barcharts created for top 25 products by region, as well as products with ≤0.2% sales.

2.3.3 Determine the reliability of the data sets

- Sales columns assessed for normality using Q-Q Plots with references lines for normal distribution.
- Data assessed using Shapiro-Wilk tests and checked for skewness and kurtosis.
- Correlation between Global sales and NA and EU sales determined.

2.3.4 Simple Linear regression

- Models of Global sales against EU sales and NA sales created and regression lines fitted.
- Accuracy determined using R squared, MSE, RMSE, and MAE.
- Checked for heteroscedasticity.

2.3.5 Multiple Linear regression

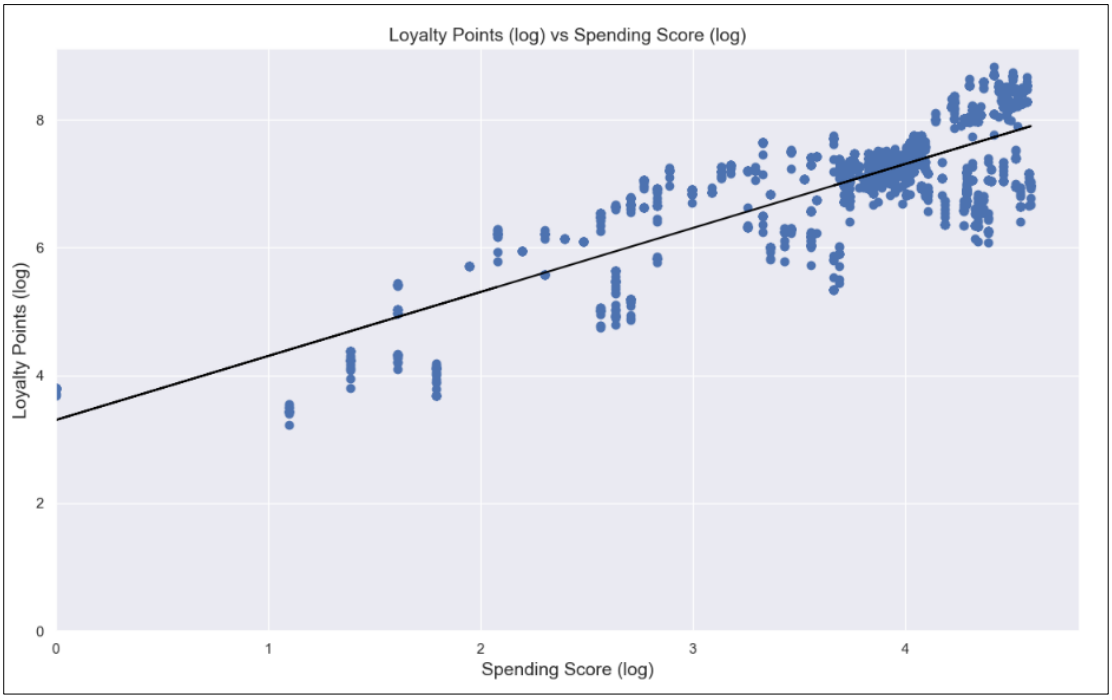
- Four models generated of Global sales versus EU sales and NA sales, and scatterplots showed predicted versus observed values.
- Accuracy determined using R squared, adjusted R squared, MSE, RMSE, and MAE.
- Checked for heteroscedasticity and multicollinearity.
- Accuracy of predicted values against observed assessed for the best model.

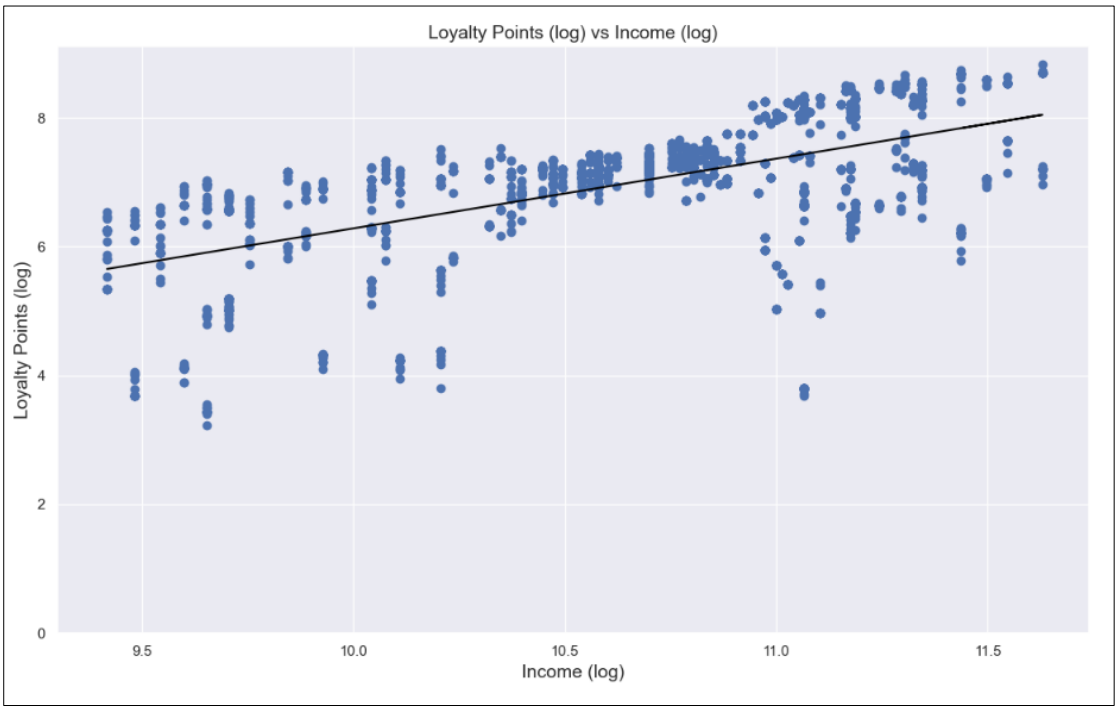
3. Visualisations and Insights

3.1 Factors that influence accumulation of loyalty points by customers

- Simple linear regression showed positive relationships between loyalty points and income and spending score. R squared indicated they were not strong predictors of loyalty points.

Figure 12: Simple linear regression for loyalty points versus spending score (log transformed).





- Multiple linear regression with spending score and income had an R squared of 98% and was therefore stronger for predicting loyalty points. Error metrics were lower, with no multicollinearity or heteroscedasticity.

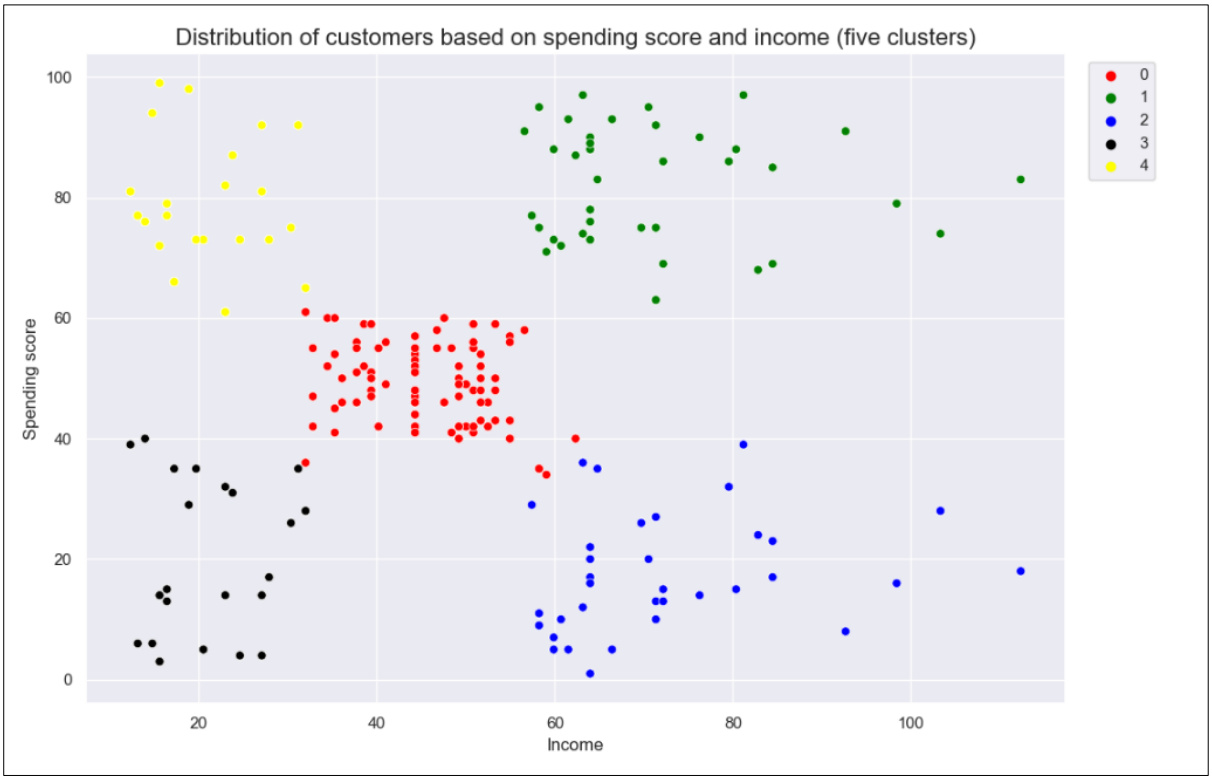
Table 1: Accuracy parameters of simple versus multiple linear regression models for predicting loyalty points.

	Simple Linear Regression		Multiple Linear Regression
	Loyalty points vs spending score (log)	**Loyalty points vs income (log)**	**Loyalty points vs spending score and income (log) (test data)**
R squared	0.67	0.34	0.98
Adjusted R squared	0.67	0.34	0.98
RMSE	0.82	0.83	0.14
MSE	0.34	0.69	0.02
MAE	0.48	0.61	0.11
Heteroscedasticity	Yes (reduced)	No	No

3.2 Grouping customers to target specific market segments

- K-means clustering created five customer segments.
  - **Low Budget**: low income / low spending score
  - **Spenders**: low income / high spending score
  - **Average**: average income / average spending score
  - **Savers**: high income / low spending score
  - **Ideal**: high income / high spending score

Figure 13: Scatterplot showing the five customer clusters based on spending score and income.



3.3 Using online customer reviews for marketing campaigns

- Wordcloud and frequency distribution showed most common words in reviews (e.g., 'great,' 'fun') and summaries (e.g., 'stars,' 'five,' 'great').

Figure 14: Wordcloud (no stopwords) (top) and frequency distribution (bottom) for reviews.

Figure 15: Wordcloud (no stopwords) (top) and frequency distribution (bottom) for summaries.

- The 250 most positive reviews/summaries identified additional words (e.g., 'perfect,' 'excellent,' 'best,' 'wonderful'). The 20 most positive reviews/summaries printed to identify opportunities to thank customers online.

Figure 16: Wordcloud for top 250 positive reviews by polarity.

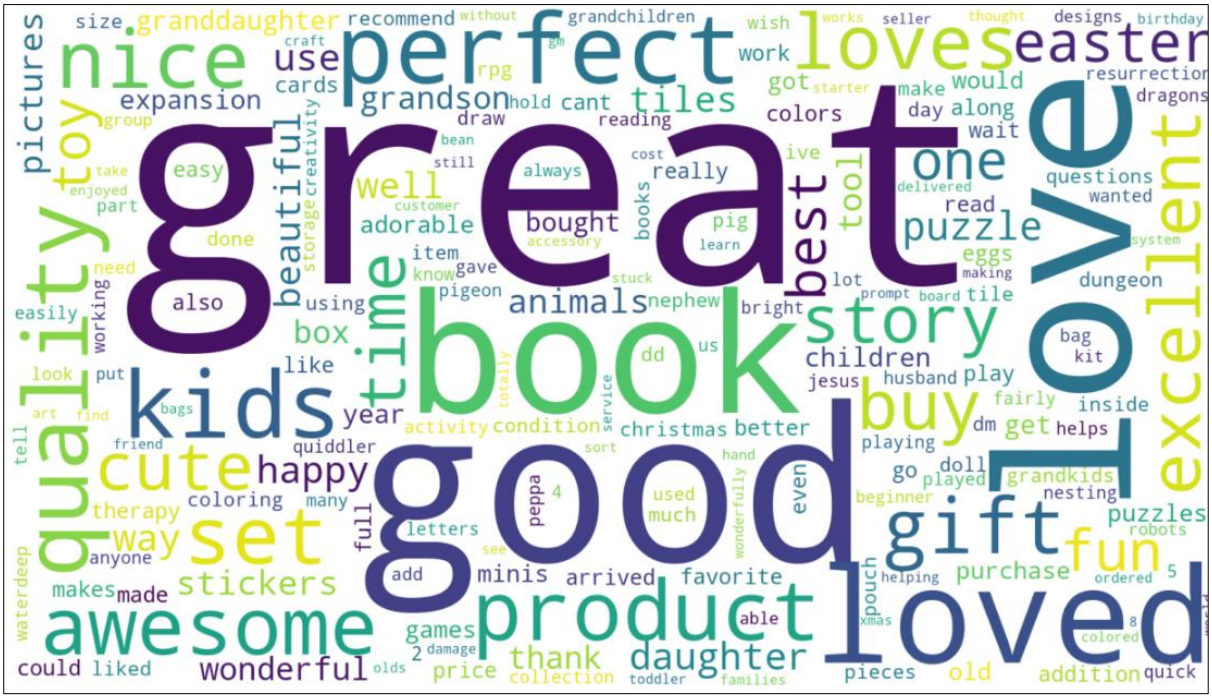
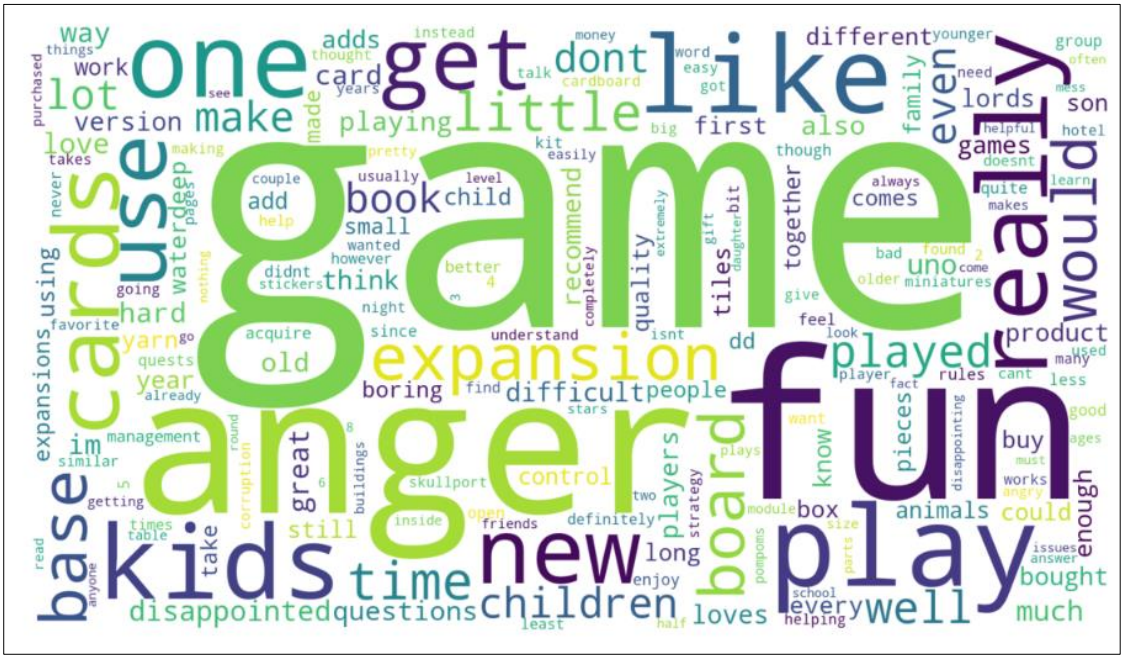


Figure 17: Wordcloud for top 250 positive summaries by polarity.

- The 250 most negative reviews/summaries identified terms for investigation. ‘Anger' incorrectly classified as negative, whereas summaries containing 'disappointed' or ‘disappointing' provide opportunities for customer engagement. Top negative reviews/summaries printed to identify opportunities to address product or customer service issues.

Figure 18: Wordcloud for top 250 negative summaries by polarity.



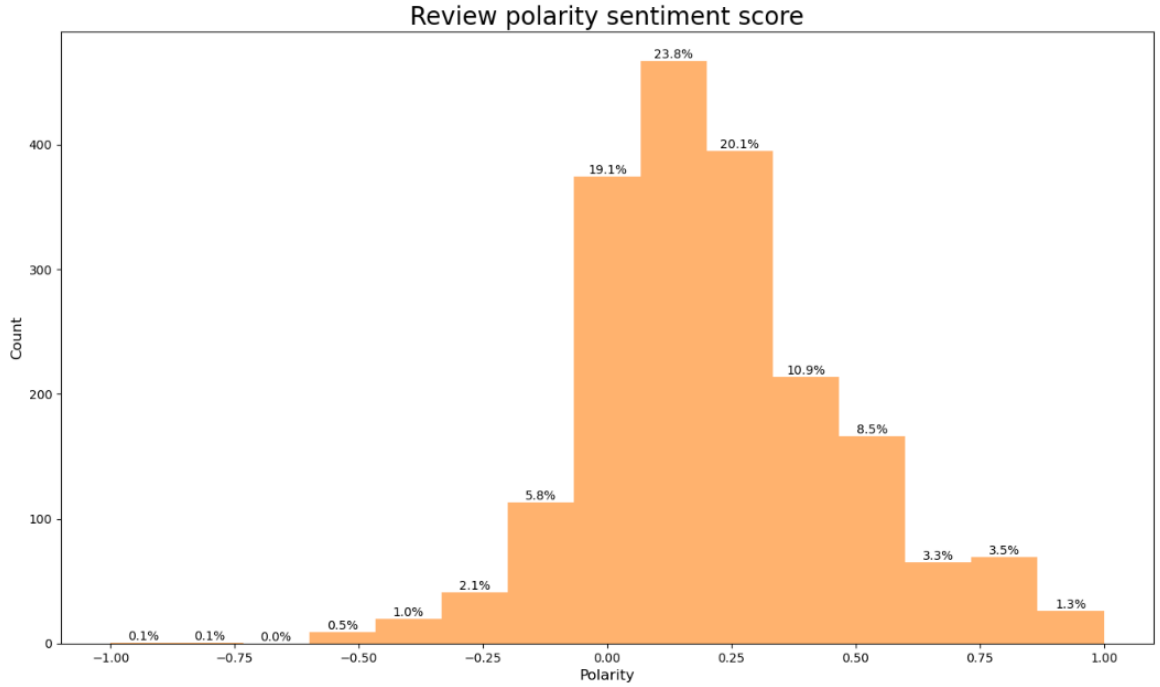
- Products with most negative reviews/summaries identified. Not negative in some cases and others revealed opportunities to improve products, e.g., better quality materials.

Figure 19: Reviews for product 3436 with words related to product quality highlighted.

review
0
Smaller than we thought, kind of disappointed in it.
1 As an adult crafter who likes simple projects to relieve stress, I enjoyed this product very much, even though there are a f
ew minor flaws. I'll list some pros, cons, and suggestions below.\n\nPros:\nThe book provides detailed directions for seven dif
ferent dog breeds in addition to abbreviated suggestions for how to create an unlimited number of breeds on your own.\nBook dir
ections are clear and detailed with large, accurate pictures.\nThe pompom maker is sturdy and easy to use. I had no problem cre
ating pompoms easily my first try.\nThe kit comes with enough basic supplies to make four adorable puppies, provided you take y
arn measurements into account\nI was very satisfied with how cute my puppies turned out!\n\nCons:\nThe basic school glue provid
ed was not strong enough to hold the heads to the bodies or the felt ears to the heads.\nThe punch-out dog carriers were too sm
all for all of my pups.\nThe kit is probably a bit too challenging for most eight-year-olds. I'd recommend it for ages ten and
up.\nWhile there is enough yarn provided to make four dogs, there is not enough to make any four you want. Some dogs require to
o much yarn if you want to create four dogs total.\nThe measuring tape on the back of the book measures in inches and centimete
rs, while the book directions are in yards.\n\nSuggestions:\nUse a hot glue gun to hold the dog together. I used superglue, but
my puppies are a little fragile.\nTake yarn measurements into account when choosing pups or just buy more yarn at a craft stor
e.\nIf you wrap the yarn around the height of the book, you'll have half a yard of yarn. Use this cheat mentioned in the book i
nstead of the confusing ruler.\nWhen the book says to cut the yarn short, be sure to cut it quite short or your pup might look
a little messy and shaggy.\n\nOverall I really liked this kit and would definitely recommend it.
2
My son loves playing this game. It was recommended by a counselor at school that works with him.
3
great for my 5 yr old
4
The game tiles, board, and tile stands are all made of paper. After using few times it will not sustain. Paper board tiles will
move on the board making the game messy and inconvenient to manage. Its a shame what they have done to such a brilliant game.
5
I was attracted to this game and its counterparts (Castle Ravenloft and Legend of Drizzt) mainly because there is a one-player
option. There are plenty of times where I'd like to play but no one is available to play. Single or multi player, it seems to b
e a solid game.\n\nMy only complaints are that the punch out tiles seem to be a BIT less of quality when compared to other big
box games like Descent Second Edition. I noticed when I was punching out the tiles, there was a lot of debris on the table. May
be a poor cut in the sheet, but annoying in any event. They also are prone to tearing if you aren't careful when taking them ou
t. But once they are punched out, they seem to be pretty solid. The playing cards are also a bit thinner than other games. Not
too big of a deal since most players buy sleeves for all their cards, so that I can live with. Lastly, the adventure book and r
ule book are pretty flimsy.\n\nThe minis are of great quality and there are a lot of them. Wizards of the Coast never disappoin
ts in this department IMHO.\n\nAll in all, if the punch outs are carefully taken out and the cards and books are stored properl
y, there should be no issue.\n\nFor a sale price of $38, I couldn't pass this up, but I would be VERY hesitant to pay full blow
n retail price.\n\nDefinitely worth picking up for solo adventures or a good solid game night.

- Overall positive sentiment distribution for reviews/summaries. Most reviews had positive polarity, while most summaries were neutral or positive. Average polarity scores were similar (0.21 for reviews, 0.22 for summaries). For subjectivity, review distribution was spread evenly, whereas summaries were more fact based.

Figure 20: Histograms of distribution of polarity for reviews (top) and summaries (bottom).



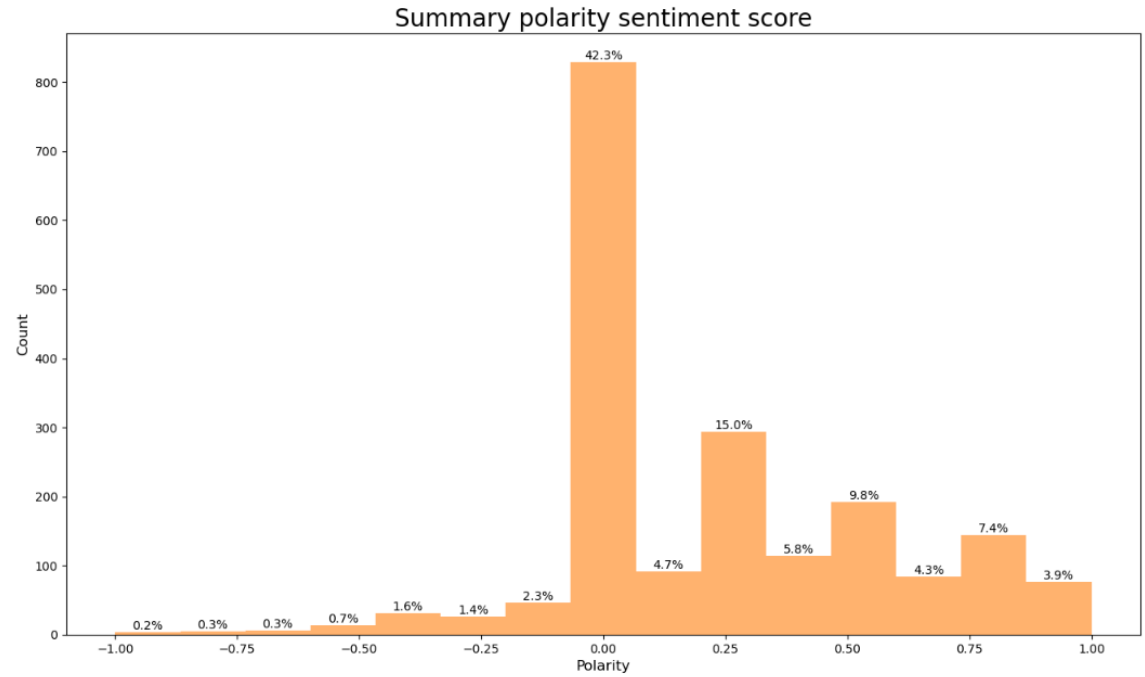
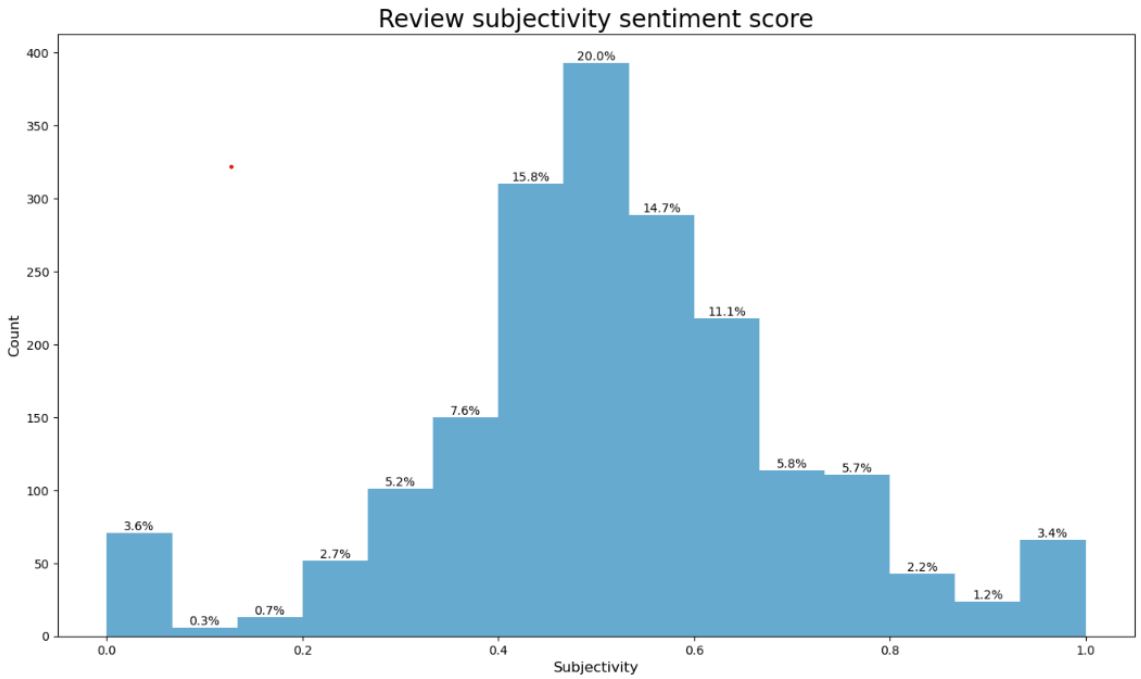
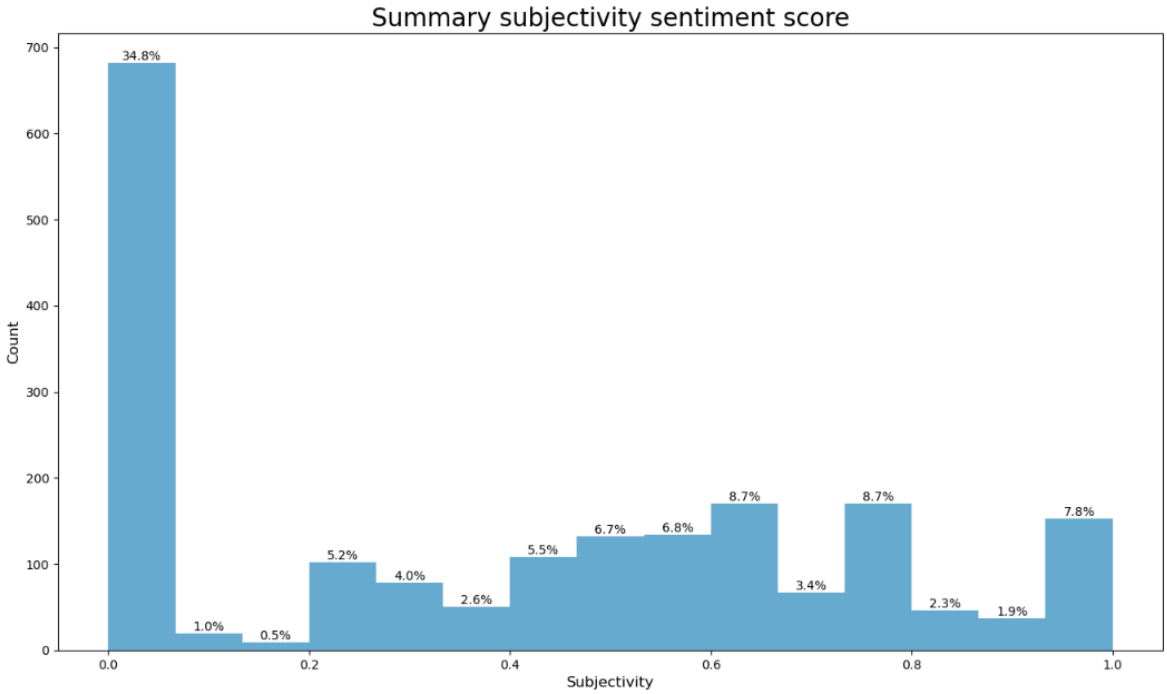


Figure 21: Histograms of distribution of subjectivity for reviews (top) and summaries (bottom).

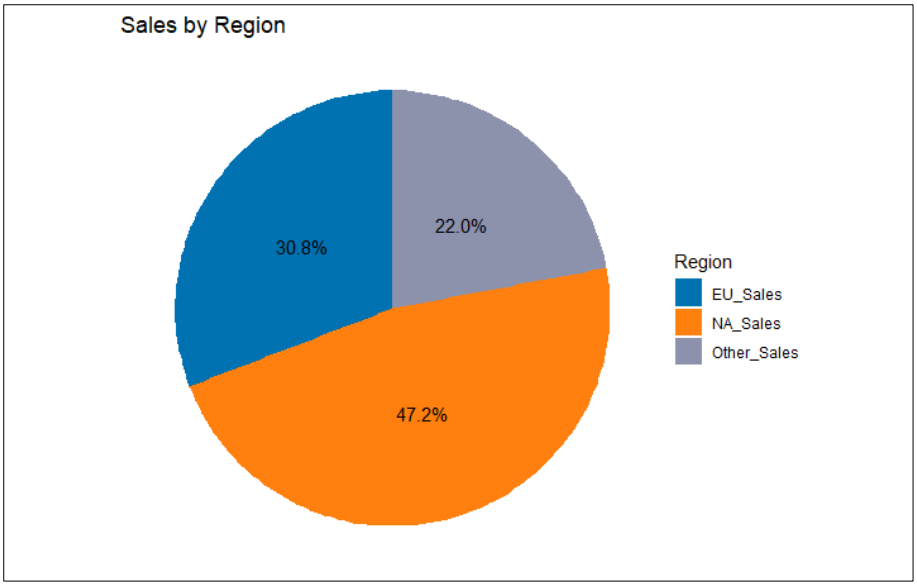




3.4 Analysis of product sales trends

- North America had the highest proportion of sales (47%).

Figure 22: Total sales by region.



- Boxplots showing impact of variables on Global sales highlighted different genre had less effect compared to platform, publisher, and year.

Figure 23: Boxplot showing impact of platform on Global sales.

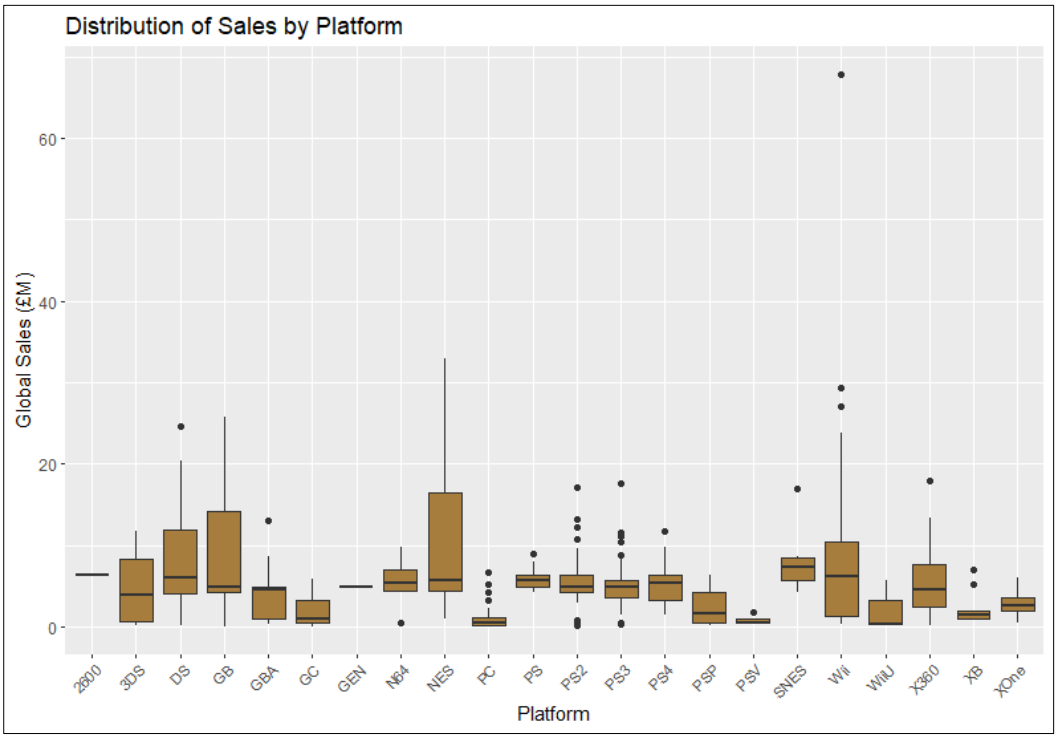
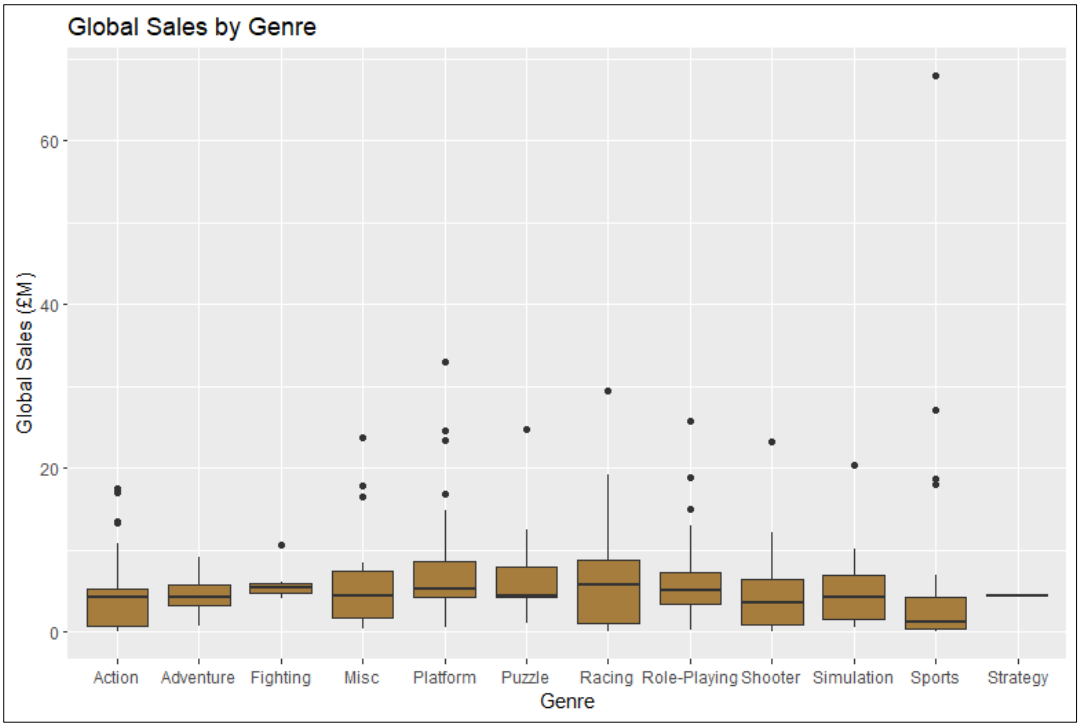
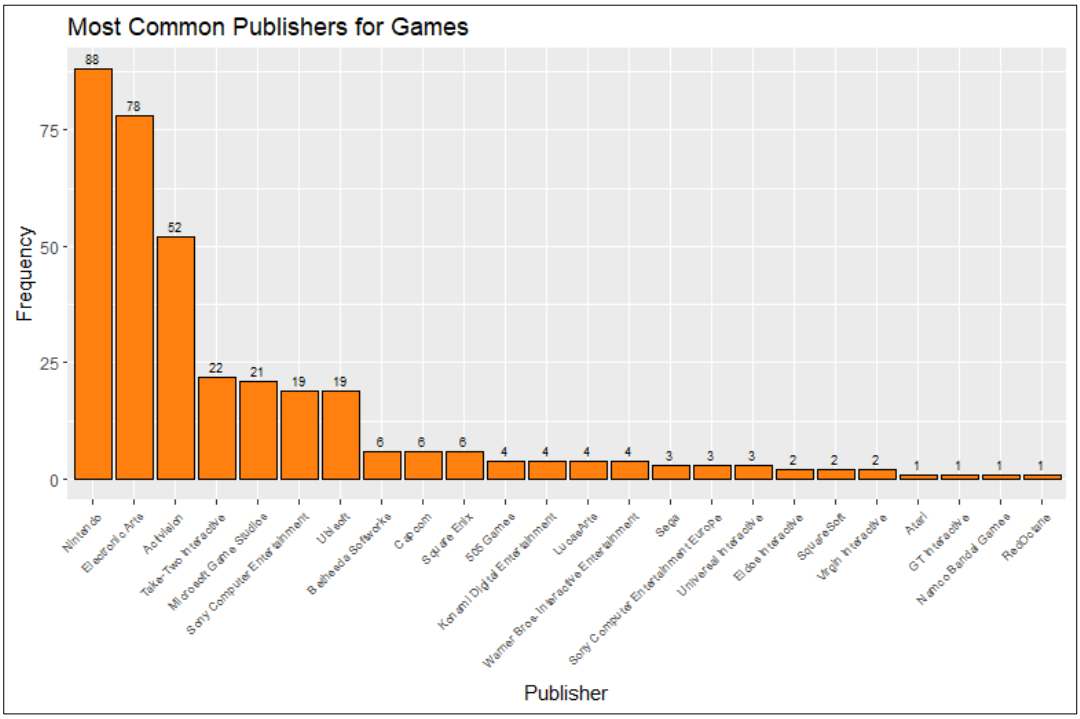


Figure 24: Boxplot showing impact of genre on Global sales.



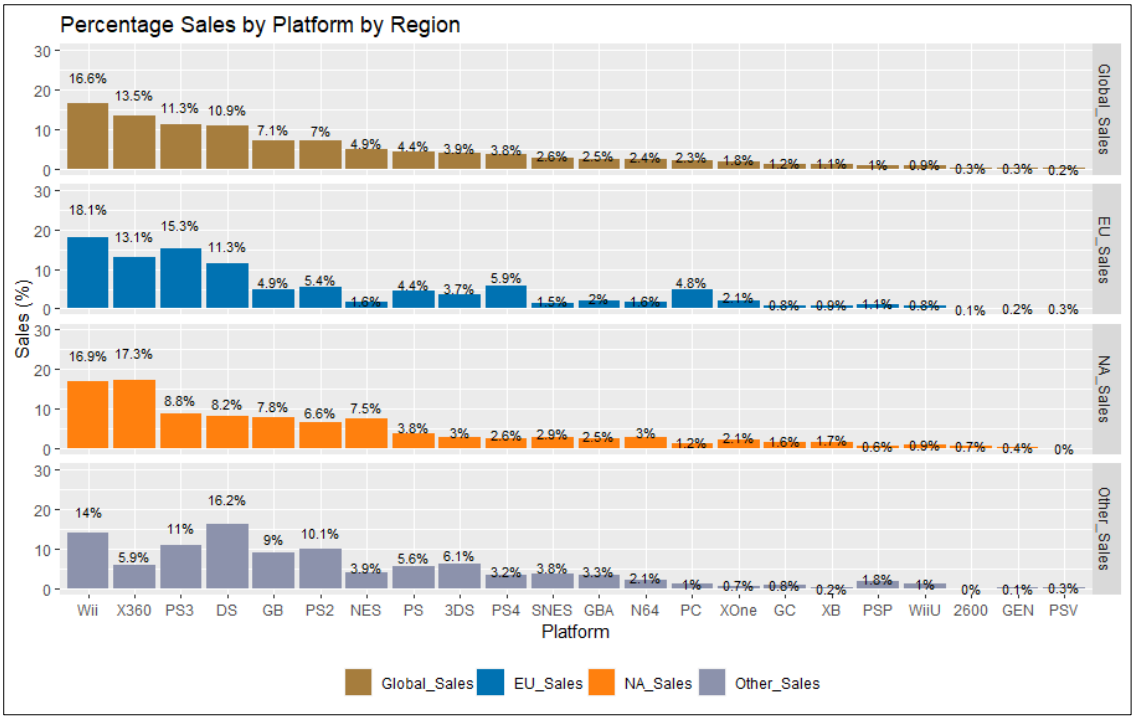
- Barcharts show the count of games of platform, publisher, genre, and year.

Figure 25: Barchart showing count of games by publisher.



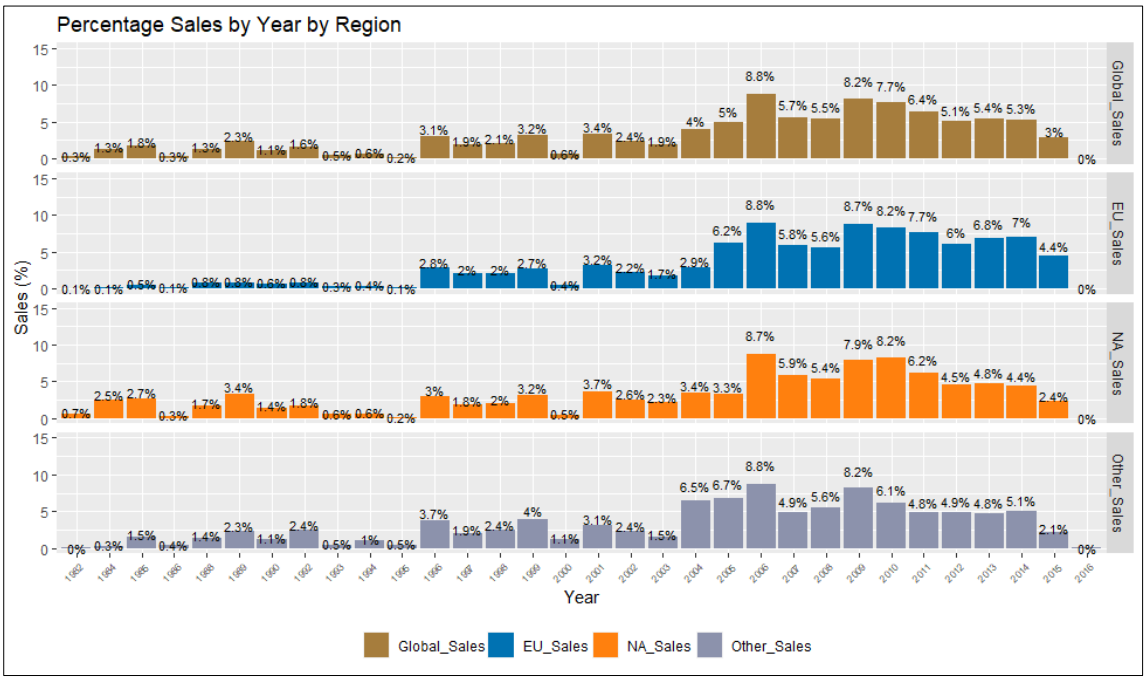
- Global and regional sales compared by platform, genre, publisher, and year. For platform, Wii accounted for 17% of Global sales, followed by Xbox 360 (14%). Similar trends for EU and NA sales, but DS was top for Other Sales.

Figure 26: Global sales versus regional sales by platform.



- For year, trends were similar but a slightly higher proportion of sales of older games in the US highlights a trend for retro gaming.

Figure 27: Global sales versus regional sales by year.



- Impact of product on sales investigated. Products 107 and 515 were top two globally and appeared in top five in all regions, but other top products differed. Regional differences also for products with ≤ 0.2% sales.

Figure 28: Top 50 products by Global sales split by region.

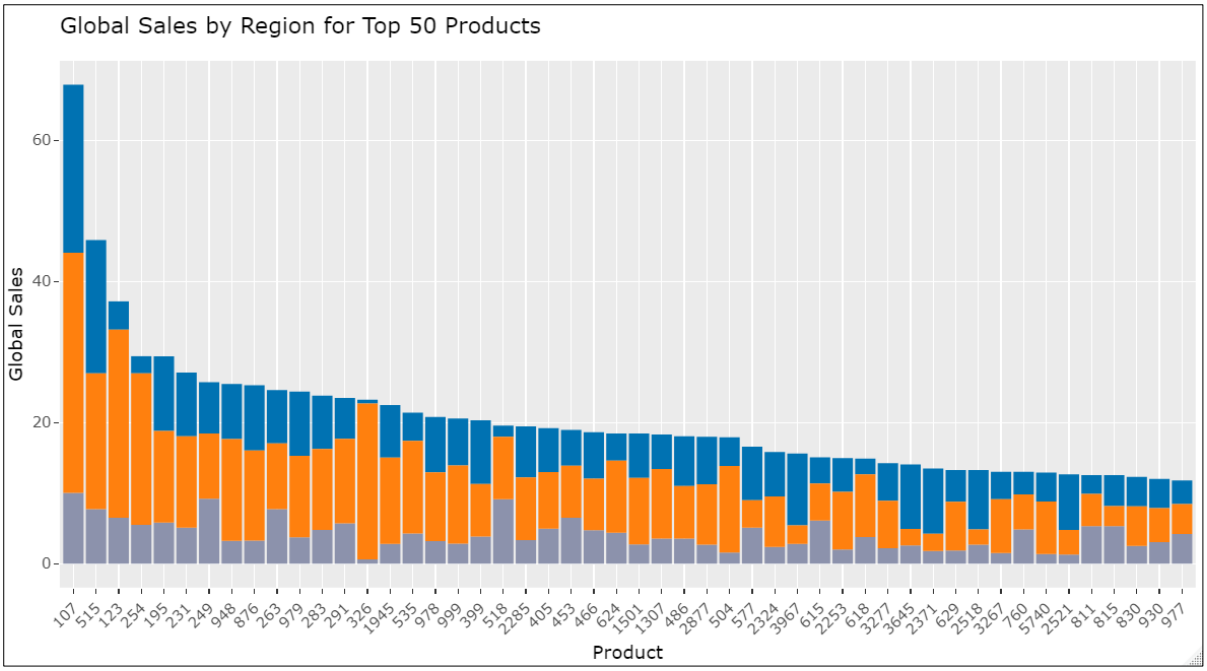


Figure 29: Top 25 products by NA sales.

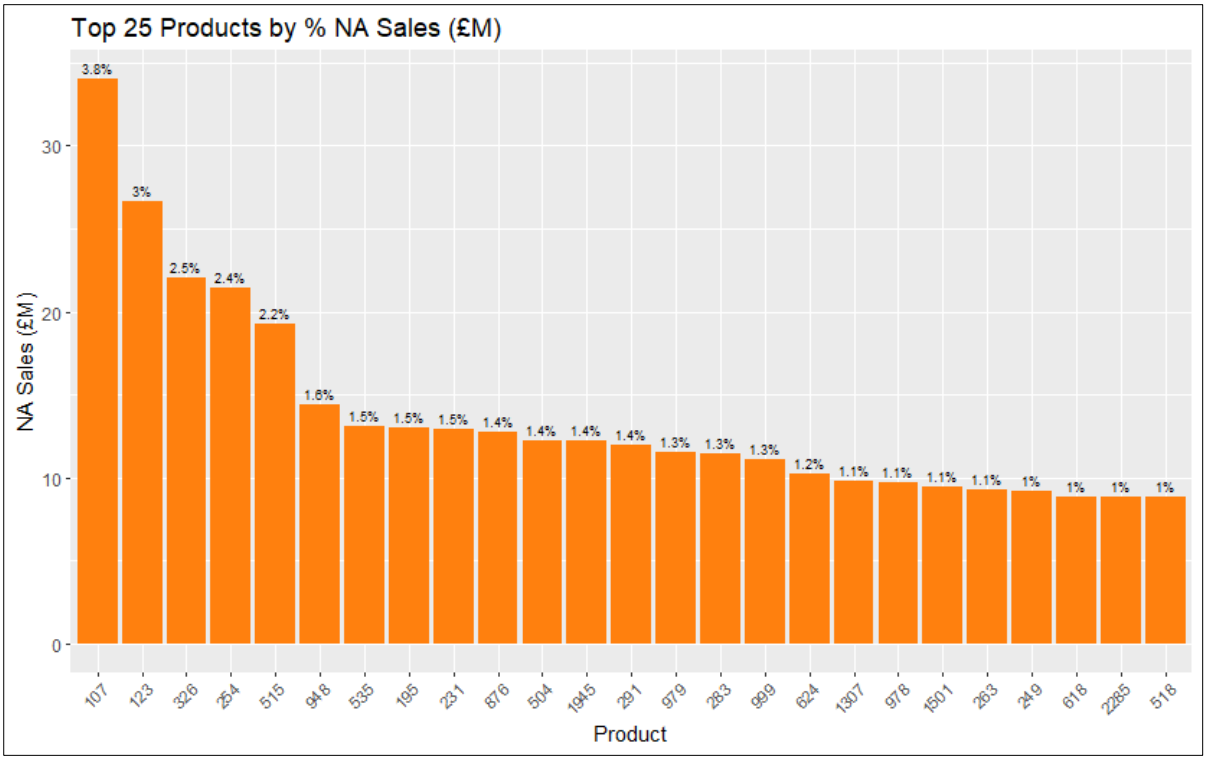
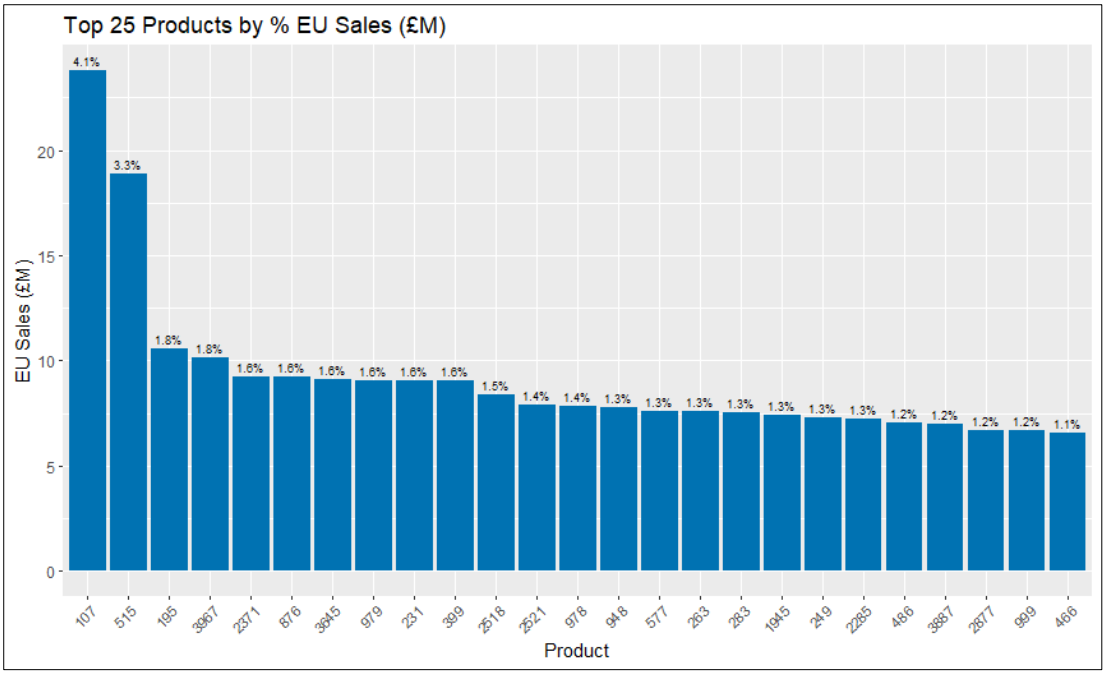


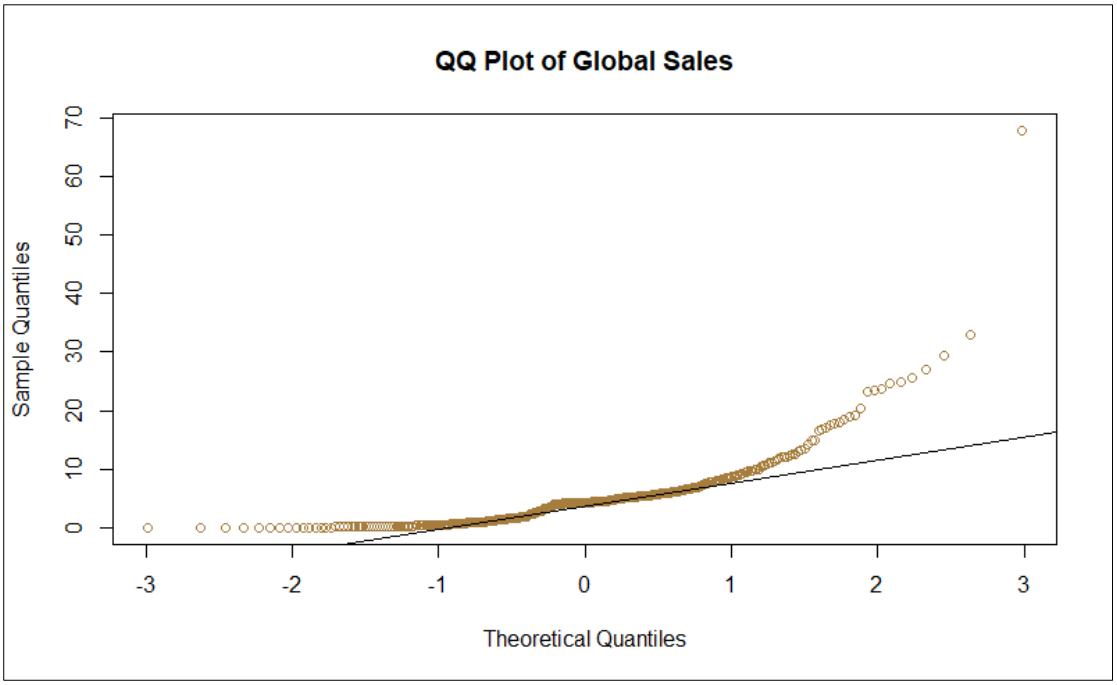
Figure 30: Top 25 products by EU sales.



3.5 Reliability of the datasets

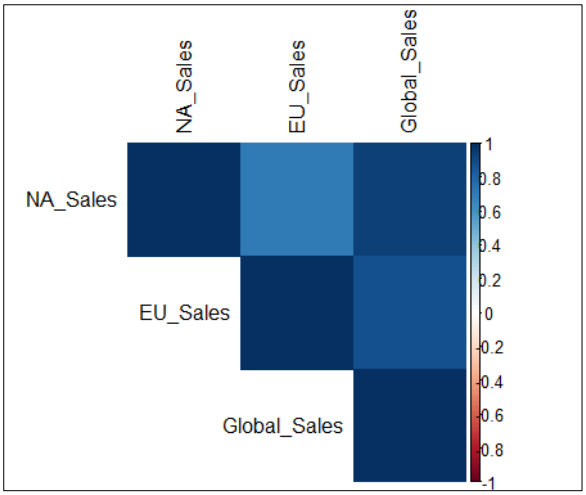
- Results from Shapiro-Wilk tests indicate that sales columns were significantly non-normal. Distributions were heavily skewed and highly leptokurtic.

Figure 31: Q-Q plot for Global sales showing significant right skew.



- Usually normalise before regression but sales columns have similar scales, and dependent and independent variables are strongly correlated. Therefore, normalisation not performed.

Figure 32: Correlation matrix for sales columns showing strong positive relationships.



3.6 Predicting Global sales with regional sales

- Simple linear regression models were a good fit. R-squared values acceptable but multiple linear regression performed to generate a better model for predicting Global sales.

Figure 33: Simple linear regression for Global sales versus EU sales.

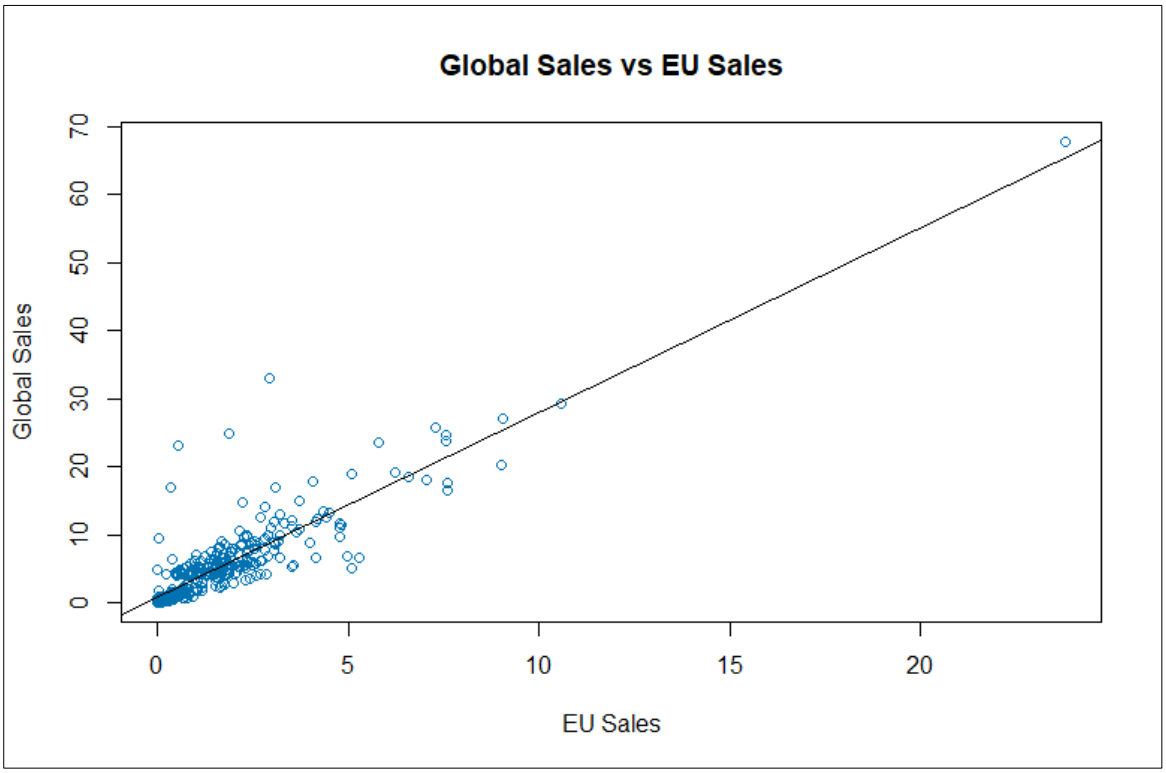
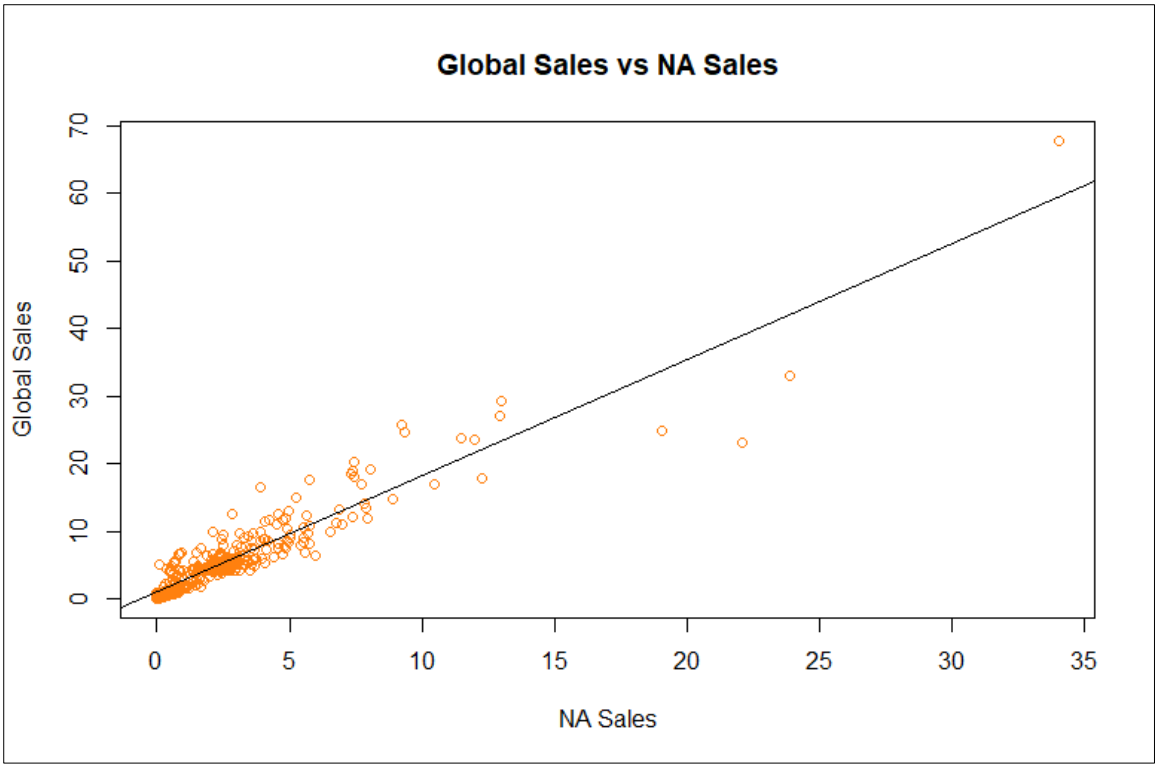


Figure 34: Simple linear regression for Global sales versus NA sales.

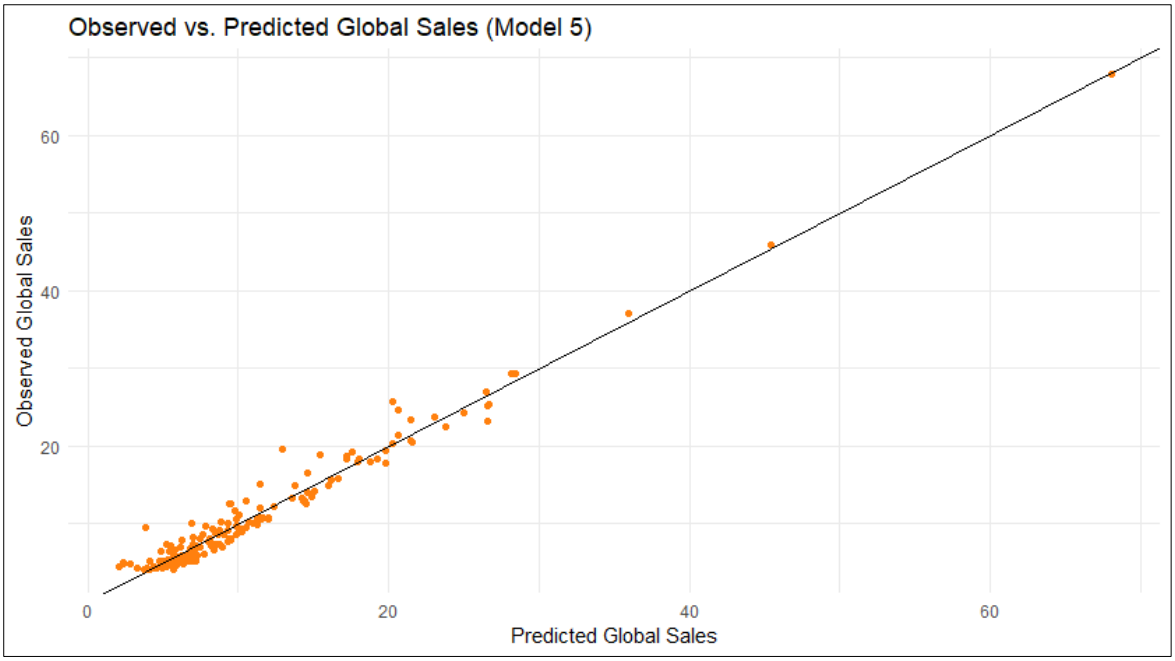


- Four multiple linear regression models created and compared. Based on accuracy parameters, Model 5 (grouped data by product) was most accurate for predicting Global sales based on EU and NA sales.

Table 2: Accuracy parameters of simple versus multiple linear regression models for predicting Global sales.

	Simple Linear Regression		Multiple Linear Regression			
	Model 1 EU Sales	**Model 2 NA Sales**	**Model 3 (Original data)**	**Model 4 (No outliers)**	**Model 5 (Grouped)**	**Model 6 (Grouped and no outliers**
R squared	0.77	0.87	0.97	0.92	0.97	0.91
Adjusted R squared	0.77	0.87	0.97	0.92	0.97	0.91
RMSE	2.99	2.23	1.11	0.90	1.48	1.40
MSE	8.99	4.93	1.23	0.82	2.18	1.97
MAE	1.73	1.45	0.69	0.58	1.10	1.04
Heteroscedasticity	No	Yes	Yes	No	No	No

Figure 35: Scatterplot showing observed versus predicted Global sales values using Model 5.



4. Patterns and Predictions

- **Multiple linear regression model** can predict how spending score and income impact **loyalty points** accumulation. The marketing department can use this to segment customers into "loyalty" groups for targeted campaigns. "Low loyalty" customers (lower income and spending score) require acquisition tactics e.g., increase value of loyalty points towards next purchase. Reward “high loyalty" customers (high income and spending score) e.g., early access to new products or exclusive sales.
- Marketing teams can use **customer clusters** for targeted campaigns. Spenders and Ideal are most "loyal" and generate most revenue. Reward them with enhanced loyalty programs, exclusive discounts, and referral schemes. Those with low spending score require acquisition strategies, such as free postage or discounts on next purchases, tailored to different income levels.
- From **sentiment analysis**, common words from positive reviews can be integrated into marketing campaigns and search engine optimisation. Turtle Games can thank customers online for positive reviews, particularly loyal customers. Negative reviews can identify pain points, create opportunities to resolve issues, and highlight why products have negative reviews.
- **Analysis of sales by product** will help Turtle Games understand which factors have the most impact on game sales, as well the highest and lowest selling products by region to support sales strategies and streamline local product offerings.
- **Multiple linear regression** for predicting **Global sales** with regional sales will allow sales teams to make better decisions, optimize resource allocation, and improve effectiveness of sales tactics.
- Areas of **further exploration and limitations**:
    - Time series data for more accurate sales forecasts.
    - Relationship between product review sentiment and sales.
    - Addition of sales data for other products (currently only video games).
    - Data quality issue - reviews of same ID number involve different products.