Introduction

Turtle Games wishes to improve sales performance. To achieve this, I'll analyse their datasets to recommend how to utilise customer trends to increase sales.

After importing the necessary libraries, I imported 'turtle_reviews.csv' into a DataFrame 'reviews'. I confirmed no missing values were present. I dropped and renamed several columns to improve clarity.

How do customers accumulate loyalty points?

I defined the dependent variable y as loyalty_points, and the independent variables x1-x3 as spending_score, remuneration and age. I created 3 OLS models of y against each x, and printed each model's summary to obtain OLS regression results. Using each model summary's coefficients, I created 3 linear regression models of predicted loyalty_points for each x. I created 3 scatterplots to visualise loyalty_points against each x, alongside their respective regression lines (see Fig. 1-3 below). The coefficient on age had a p-value above 0.05, meaning the null hypothesis that age's coefficient equals 0 is not rejected at 5% significance level. Fig. 3 confirms no relationship between age and loyalty_points. The p-values for spending_score and remuneration coefficients were statistically significant: ceteris paribus, a 1-unit increase in spending_score and £1k increase in remuneration leads to a 33.0617 and 34.1878 increase in loyalty_points respectively. Fig. 1-2 confirm initially there's a linear relationship between loyalty_points and both spending_score and remuneration. However, for spending_scores above 60 and remuneration above £50k, there doesn't appear to be any relationship with loyalty_points.

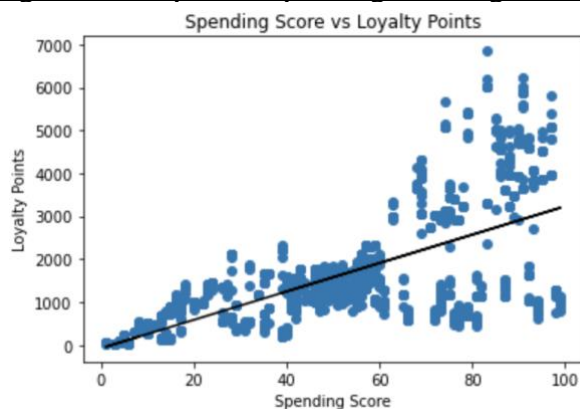Fig. 1: Scatterplot of Spending Score against Loyalty Points

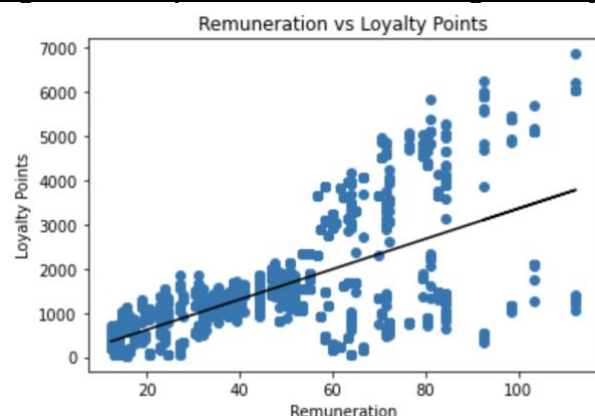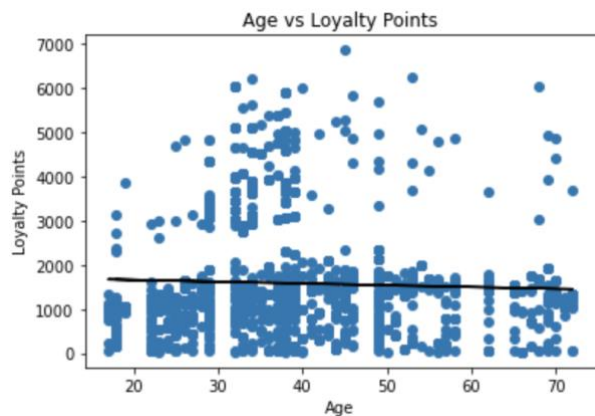
Fig. 2: Scatterplot of Remuneration against Loyalty Points

Fig. 3: Scatterplot of Age against Loyalty Points

How to target market segments?
I created plots of remuneration against spending_score (Appendix Fig. 4-5), suggesting 5 clusters within the dataset. To identify the optimal number of clusters to utilise, I employed Elbow and Silhouette methods. The Elbow chart (Appendix Fig. 6) appeared to become linear at 5 clusters, and the highest value in the Silhouette chart (Appendix Fig. 7) occurred at 5 clusters, suggesting 5 was the optimal number of clusters. To confirm this, I evaluated usefulness of using 5-7 clusters. Of the 3 pairplots created (Appendix Fig. 8-10), the k-means model with 5 clusters featured closely grouped clusters clearly separated from each other, unlike the other models. I utilised 5 clusters in the final model, and visualised them in the below scatterplot.

Fig. 11: Market Segments
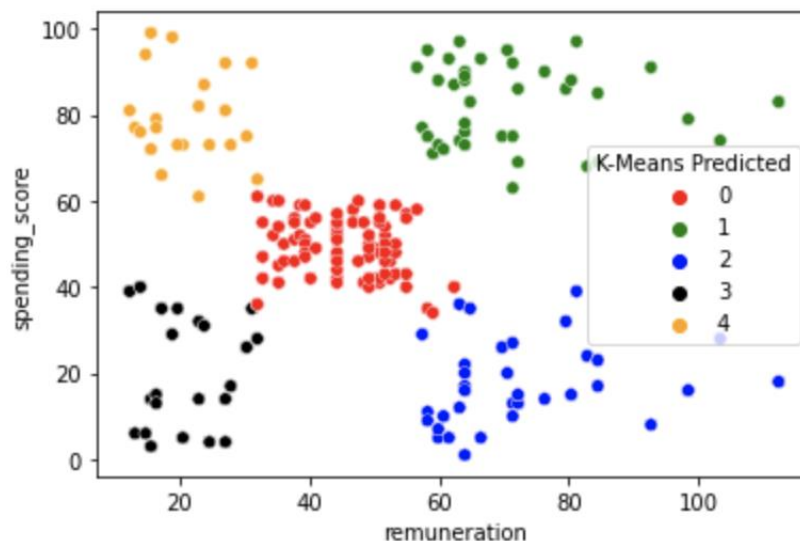
These 5 clusters can be interpreted as follows:

Market Segment	Strategies to improve sales
1. Low remuneration and spending scores	Target customers with discount offers to improve sales

2. Low remuneration but high spending scores	Target customers with discount offers to provide financial support with aim of retaining customers
3. Medium levels of remuneration and spending scores	Use combination of discounts and frequent purchase rewards
4. High remuneration but low spending scores	Target customers with rewards from repeat purchases to improve sales
5. High remuneration and spending scores	Offer customers high-end exclusive products to further incentivize sales

Utilising reviews to inform marketing

I prepared review and summary columns for NLP, applied tokenisation, created one list of review and summary tokens, removed alphanumeric characters and stopwords, and created the below wordcloud for this list to visualise most utilised words across both columns. I identified the 15 most common words across both columns and their polarity scores in the below table.

Fig. 12: Review and Summary wordcloud

15 most common words across review and summary columns

Rank	Word	Frequency	Polarity score
1	game	1990	-0.4
2	great	875	0.8
3	fun	770	0.3
4	one	568	0.0
5	play	528	0.0
6	like	468	0.0
7	stars	464	0.0
8	love	416	0.5
9	good	381	0.7
10	five	362	0.0
11	really	349	0.2
12	get	333	0.0

13	tiles	317	0.0
14	book	316	0.0
15	time	309	0.0

Words like 'great' have positive polarity scores, and should be utilised in marketing to evidence positive customer experiences. I identified the top 20 positive reviews and summaries (Appendix Fig. 13-14): some should be highlighted in marketing to confirm high-quality experiences to customers. I identified the top 20 negative reviews and summaries (Appendix Fig. 15-16), highlighting difficult instructions and boring products. This feedback should be employed to improve products and sales.

How do products impact sales?

In RStudio, I imported the tidyverse library and the turtle_sales dataset, and then removed several unnecessary columns: Ranking, Year, Genre and Publisher. I utilised scatterplots to determine relationships between the numerical variables: Product Code and each Sales variable. The smoothing curve for the below scatterplots resembles negative exponential curves, which means the highest Sales belonged to Products with low Product Codes. Turtle Games should thus prioritise Products with low Product Codes to improve Sales.

Fig. 17: North American Sales per Product Code
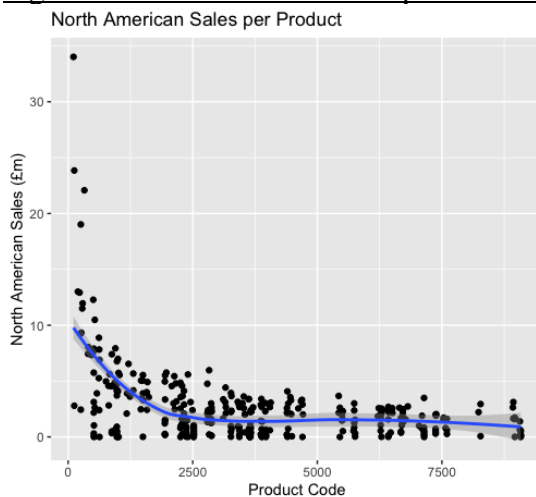
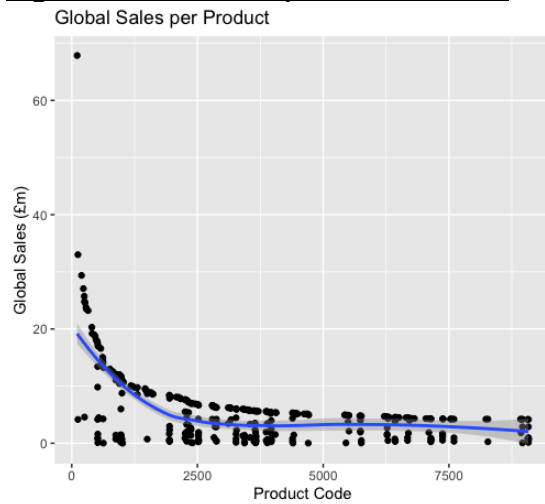Fig. 18: European Sales per Product Code

Fig. 19: Global Sales per Product Code
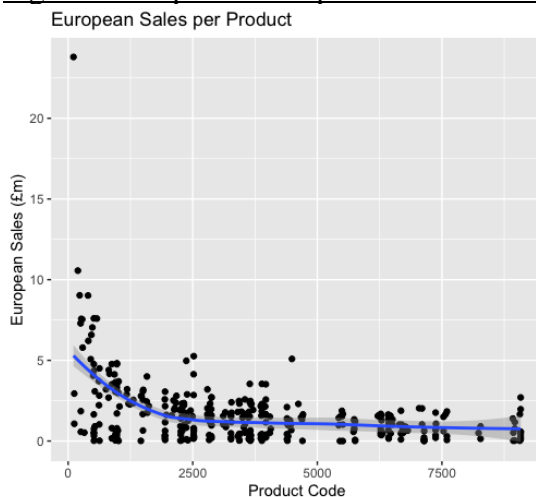
Global Sales per Product

I utilised a histogram to determine the Platforms with the most Products: the below histogram indicates these Platforms to be the Xbox 360, PS3 and PC. Turtle Games should prioritise Products on these Platforms to improve sales, as they are the most popular Platforms. I also generated histograms of each Sales variable to determine their distribution: the below Sales histograms indicate that each Sales variable is positively skewed.

Fig. 20: Most popular Platforms

Number of Products per Platform

Fig. 21: Distribution of North American Sales
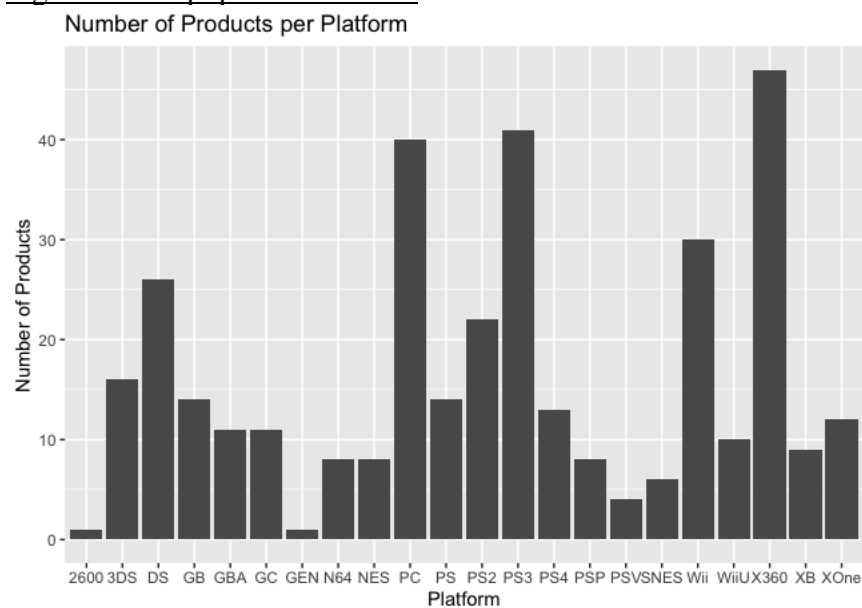
Distribution of North American Sales

Fig. 22: Distribution of European Sales
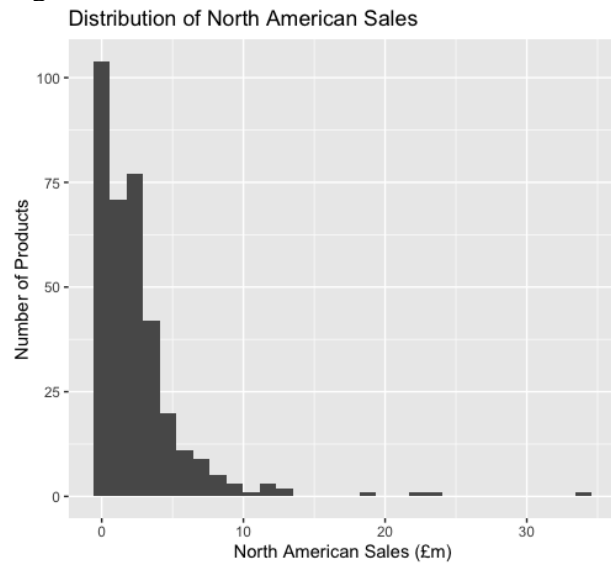
Distribution of European Sales

Fig. 23: Distribution of Global Sales
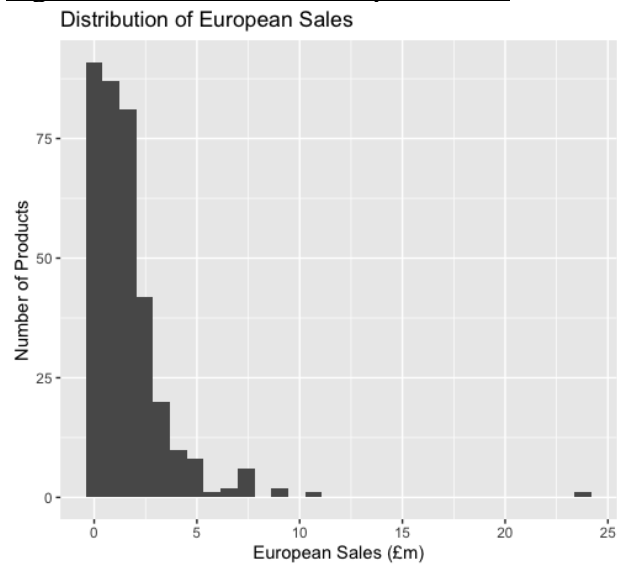
Distribution of Global Sales
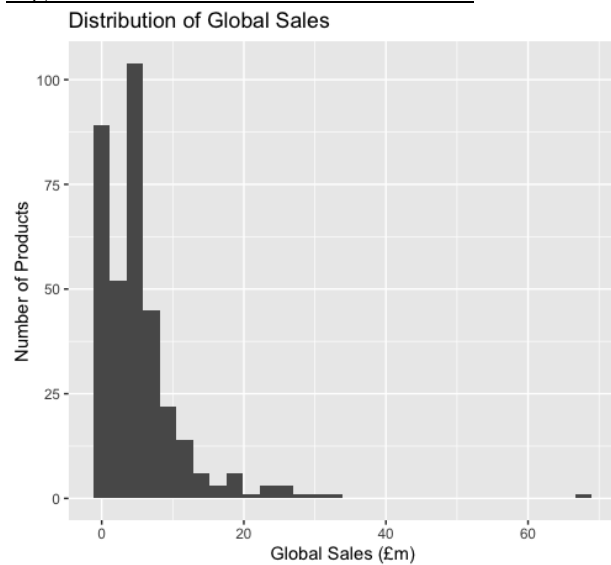
I utilised boxplots to determine the distribution of each Sales variable by Platform. In each of the below Sales boxplots, there was an extreme positive outlier for the Wii Platform, so I would like to further investigate Sales amongst Products on the Wii Platform.

Fig. 24: Distribution of North American Sales per Platform
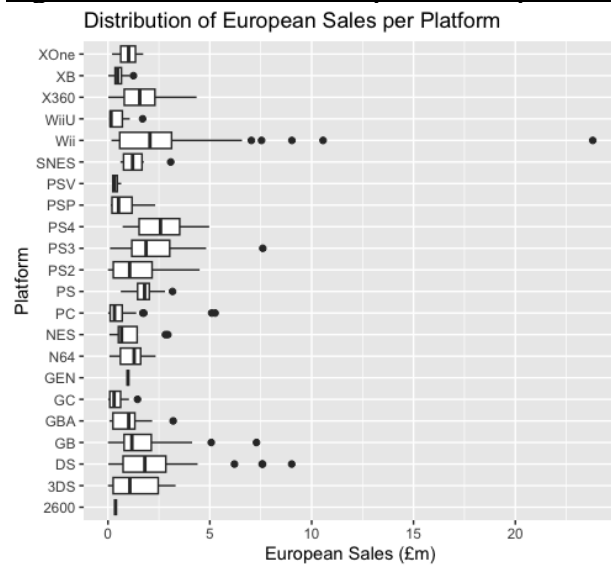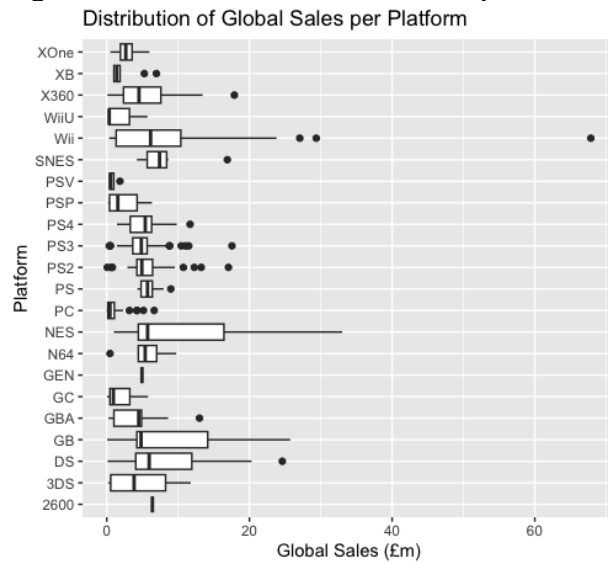

Distribution of North American Sales per Platform

Fig. 25: Distribution of European Sales per Platform
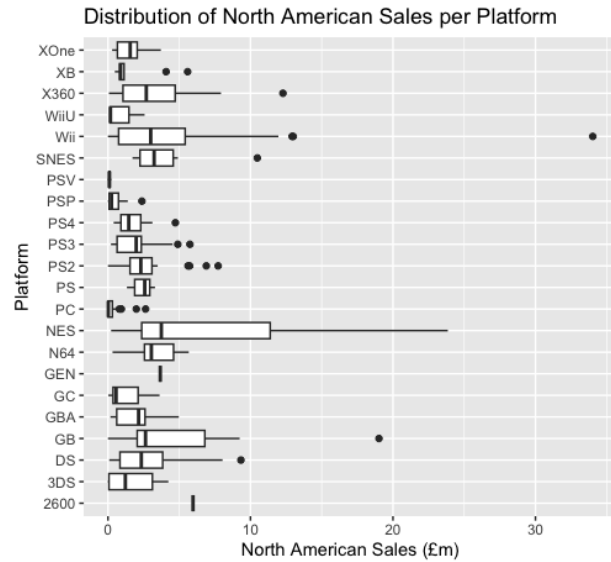
Distribution of European Sales per Platform

Fig. 26: Distribution of Global Sales per Platform

Distribution of Global Sales per Platform

Data reliability

I created a new DataFrame (sales2) that grouped turtle_sales data based on Product and summated each Sales variable per Product. For each Sales variable in sales2, I performed Shapiro-Wilk tests and rejected the null hypotheses that each Sales variable was normally distributed due to p-values below 0.05, meaning each Sales variable is non-normal. This was confirmed by the below Q-Q plots of these Sales variables, as all the points don't lie on a straight line. Consequently, correlation can't be determined between the different Sales variables. I used skewness() and kurtosis() to determine each Sales variable is extremely positively skewed and heavy-tailed. This is confirmed by the below KDP of Global Sales from sales2. Utilising ggplot() on the turtle_sales and sales2 data, I generated improved versions of the visualisations presented in the previous section (Appendix Fig. 33-42), confirming my previous insights. I generated a below multivariate scatterplot to inform Turtle Games of which platforms the highest selling products globally belong to. I generated the below violin plot of Global_Sales per the 5 most popular Platforms to inform Turtle Games of the summary statistics and distribution of Global_Sales on the platforms they should target.

Fig. 27: Q-Q plot of North American Sales per Product
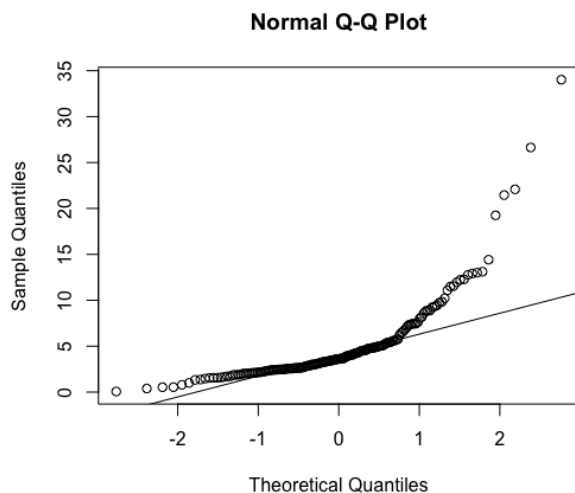

Normal Q-Q Plot

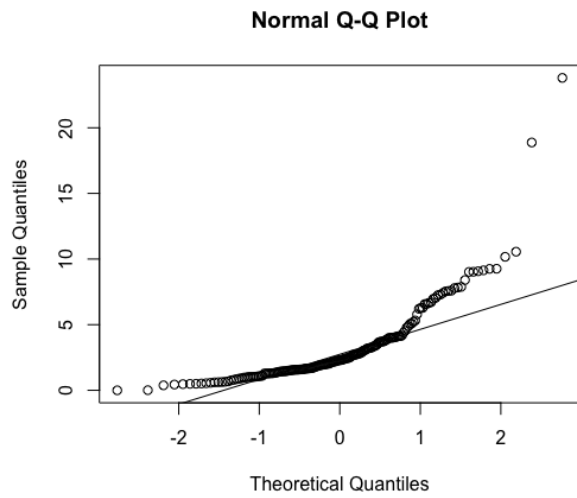Fig. 28: Q-Q plot of European Sales per Product
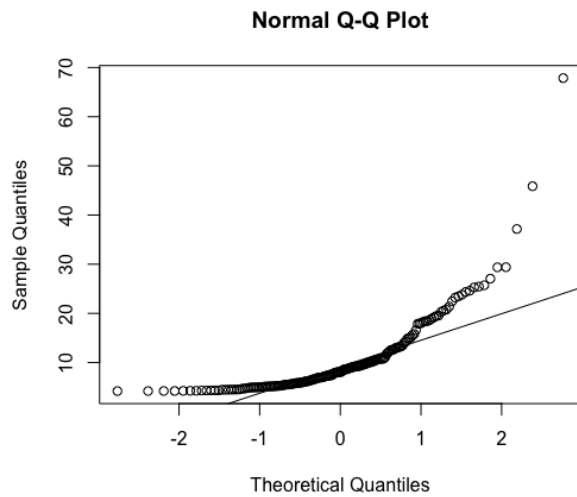

Fig. 29: Q-Q plot of Global Sales per Product

Fig. 30: KDP of Global Sales per Product

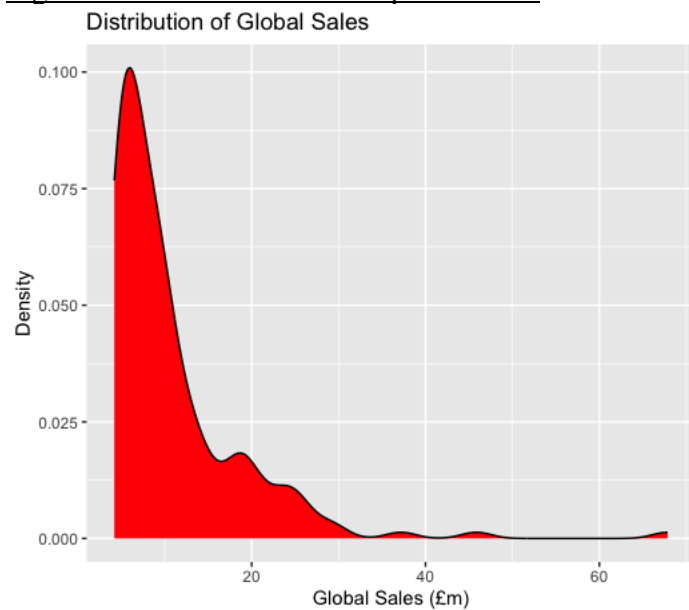Fig. 31: Global Sales per Product Code and Platform

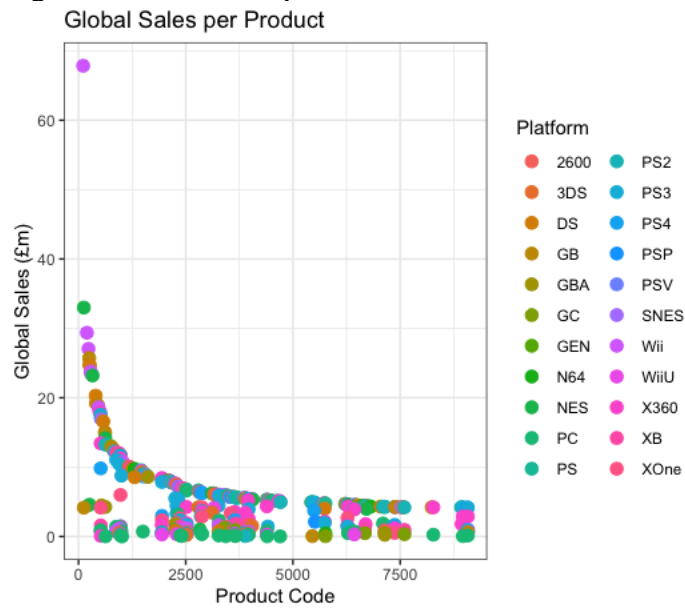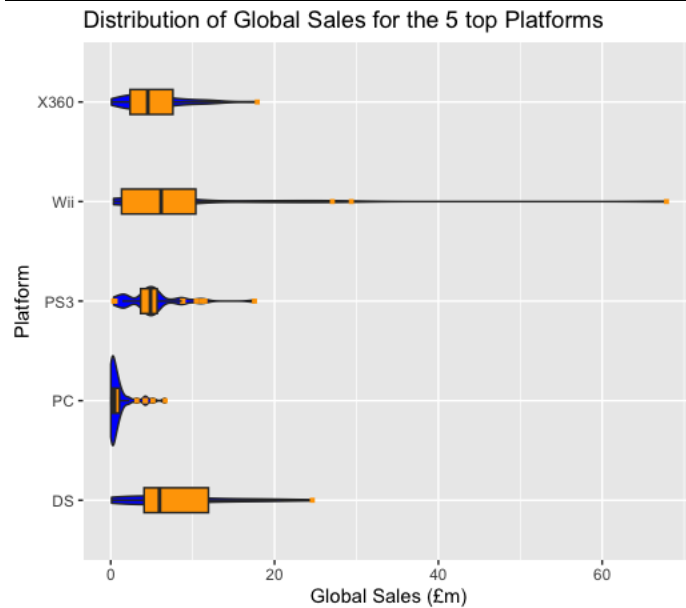Global Sales per Product

Fig. 32: Summary Statistics and Distributions of Global Sales per 5 most popular Platforms

Distribution of Global Sales for the 5 top Platforms

Relationships

To determine what relationships exist between the Sales per Product variables (sum_NA_Sales, sum_EU_Sales and sum_Global_Sales), I first created a correlation matrix that indicated that sum_Global_Sales was more strongly correlated with sum_NA_Sales versus sum_EU_Sales As there're 3 Sales variables, I utilised a multiple linear regression model to determine the effect of sum_NA_Sales & sum_EU_Sales on sum_Global_Sales. As expected, the coefficients on sum_NA_Sales & sum_EU_Sales were both extremely statistically significant and positive. Interestingly, sum_EU_Sales had a larger coefficient (1.20) than sum_NA_Sales (1.13), suggesting Turtle Games should slightly prioritise Europe to improve Sales. While prior analysis has indicated that the Sales variables are non-normal and very positively skewed, the t-tests utilised should be robust against this non-normality.

The R-squared and adjusted R-squared values were very strong at around 97%. I also predicted global sales based on the below values to test model accuracy.

Test Case	sum_NA_Sales value	sum_EU_Sales value	Predicted Global Sales	Confidence Interval Lower Bound	Confidence Interval Upper Bound
1	34.02	23.80	68.056548	66.429787	69.683310
2	3.93	1.56	07.356754	07.099418	07.614090
3	2.73	0.65	04.908353	04.614521	05.202185
4	2.26	0.97	4.761039	04.478855	05.043223
5	22.08	0.52	26.625558	25.367353	27.883763

The actual Global Sales for Test Case 1 (£67.85m) fell within the Confidence Interval, but actual Global Sales for Test Case 5 (£23.21m) didn't fall within the Confidence Interval despite being only £3m off the predicted value. Overall, this model is a decent predictor of Global Sales, but can be improved by adding in more numerical variables such as review scores.

Conclusion

I recommend that Turtle Games employs the following strategies to improve sales and overall financial performance.

Strategy	**Effect**
1. Tailor different strategies to each of the 5 Remuneration – Spending Score Market Segment	This will improve revenue generated from each market segment
2. Use some of the top 20 positive Reviews and Summaries in Marketing	This will evidence that Turtle Games' products are high-quality and attract new customers
3. Prioritise Products with low Product Codes	3. This will improve sales and revenue
4. Prioritise Products sold on the Xbox 360, PS3 and PC Platforms	4. As these are the most popular platforms, this will improve sales and revenue
5. Slightly prioritise European Sales over North American Sales	5. This will efficiently improve Global Sales

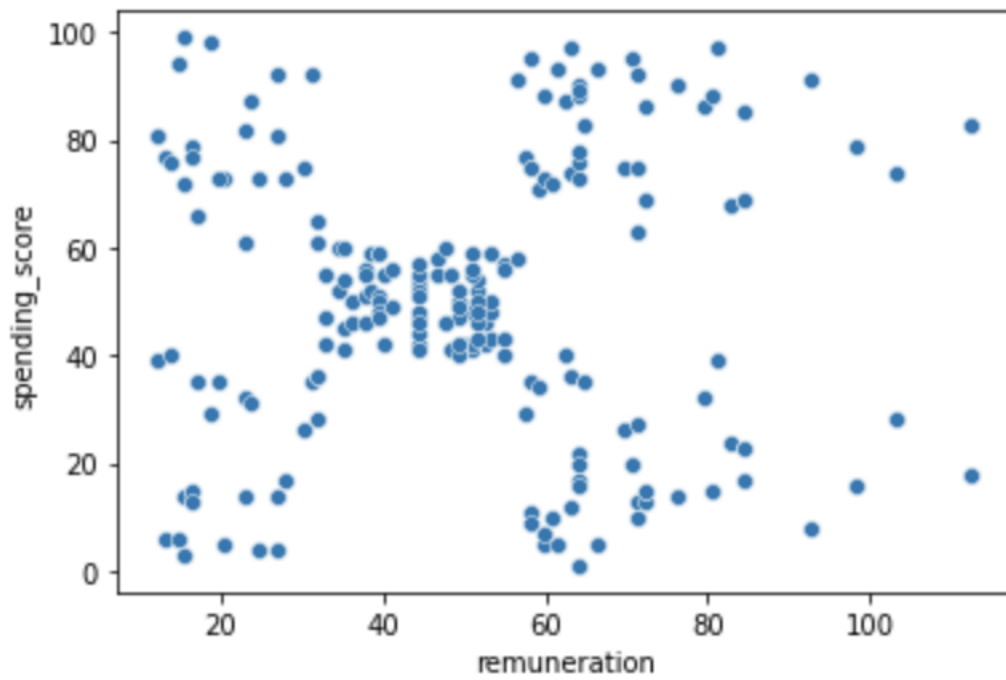Fig. 4: Scatterplot of Remuneration against Spending Score:

Fig. 5: Pairplot of Remuneration against Spending Score:

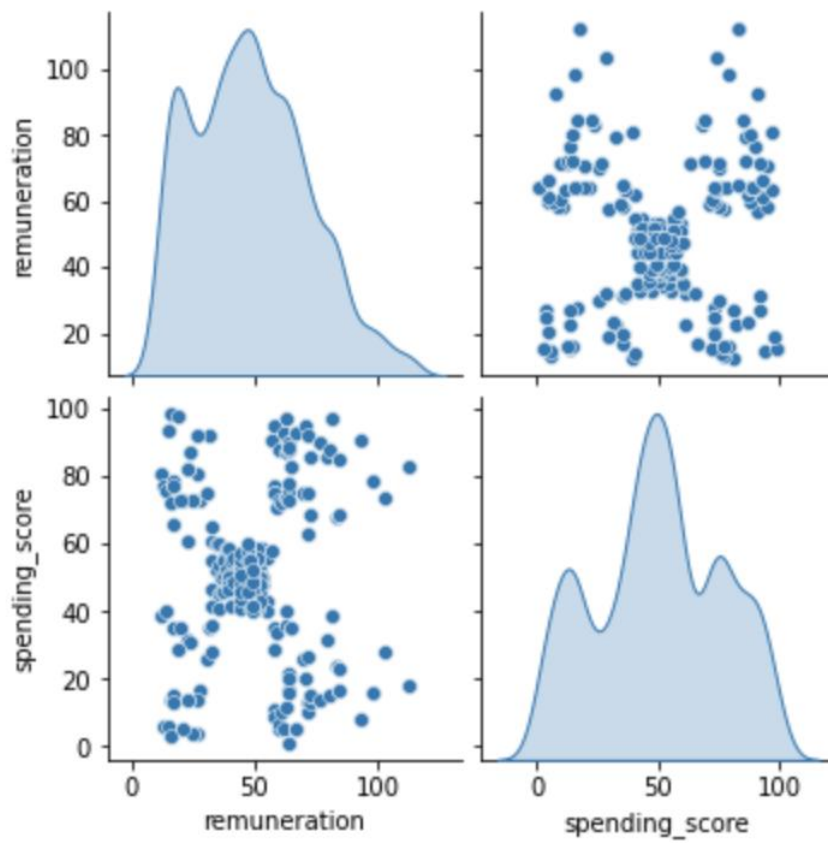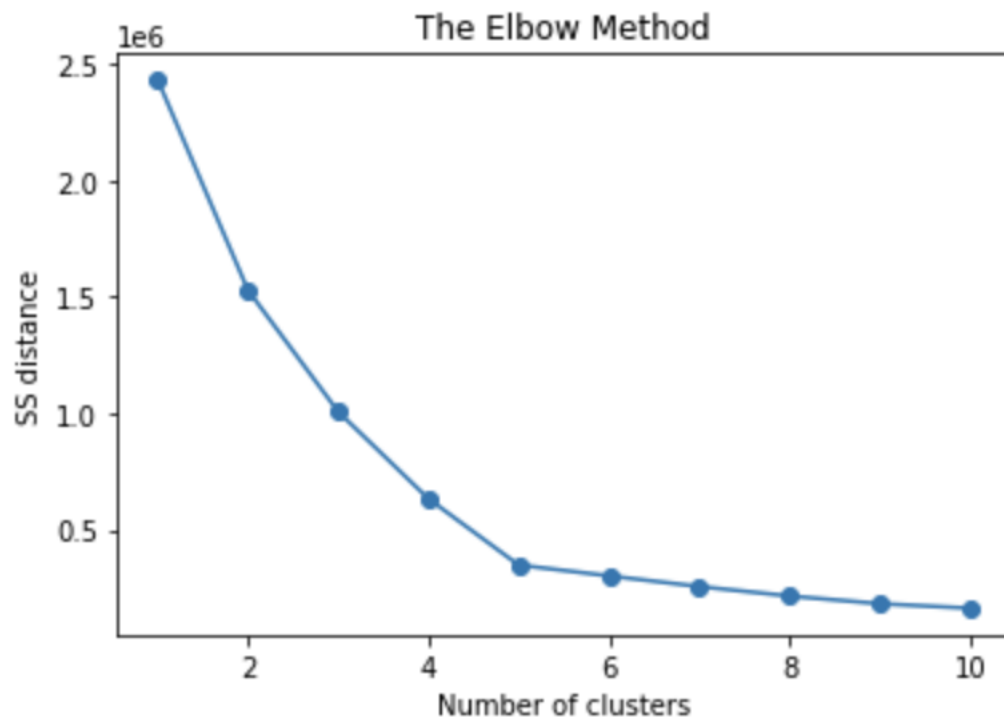Fig. 6: Elbow chart

The Elbow Method

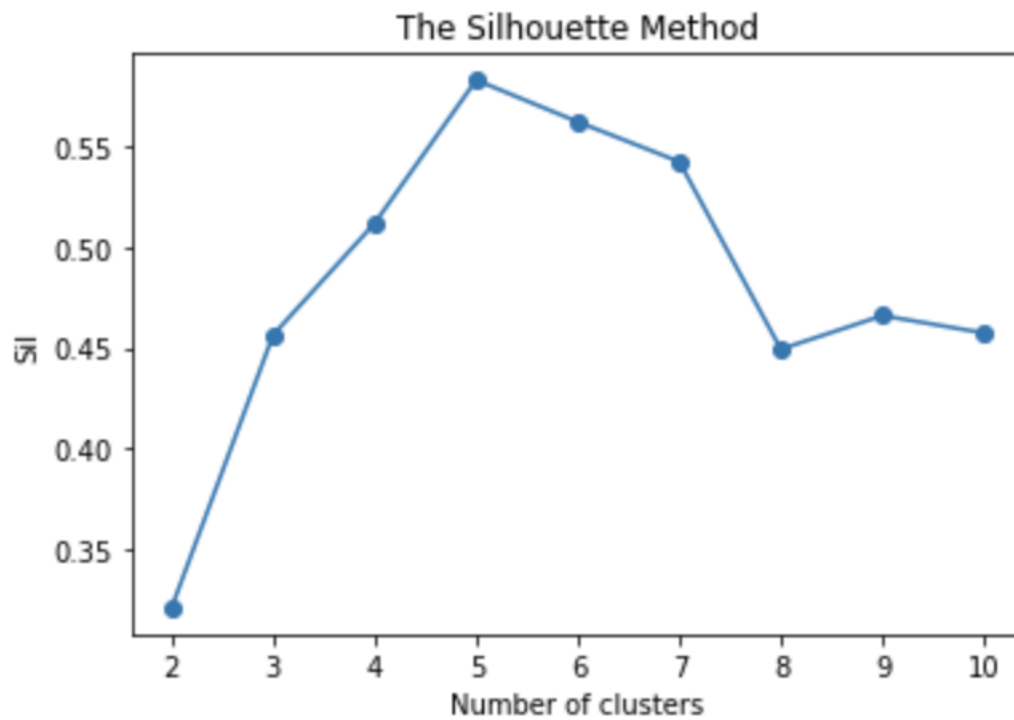Fig. 7: Silhouette chart

The Silhouette Method

Fig. 8: K-means model with 5 clusters

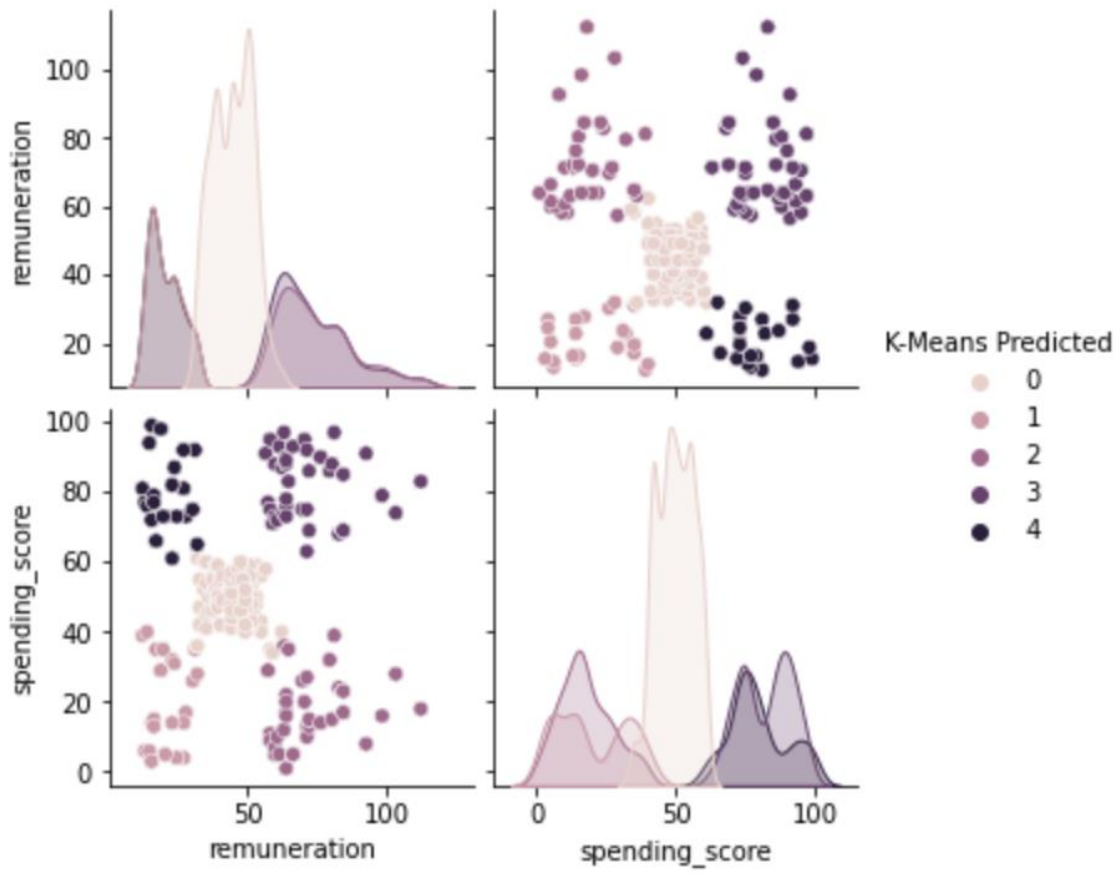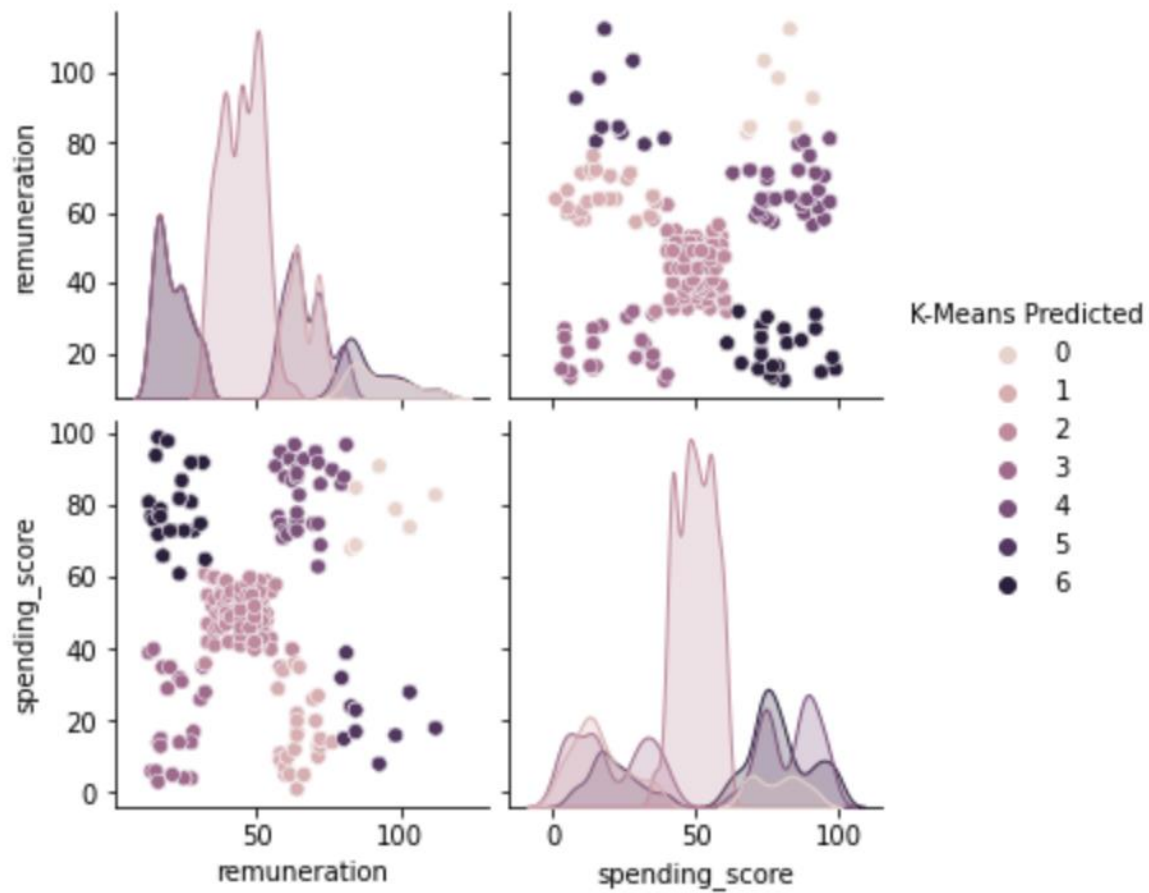Fig. 9: K-means model with 6 clusters
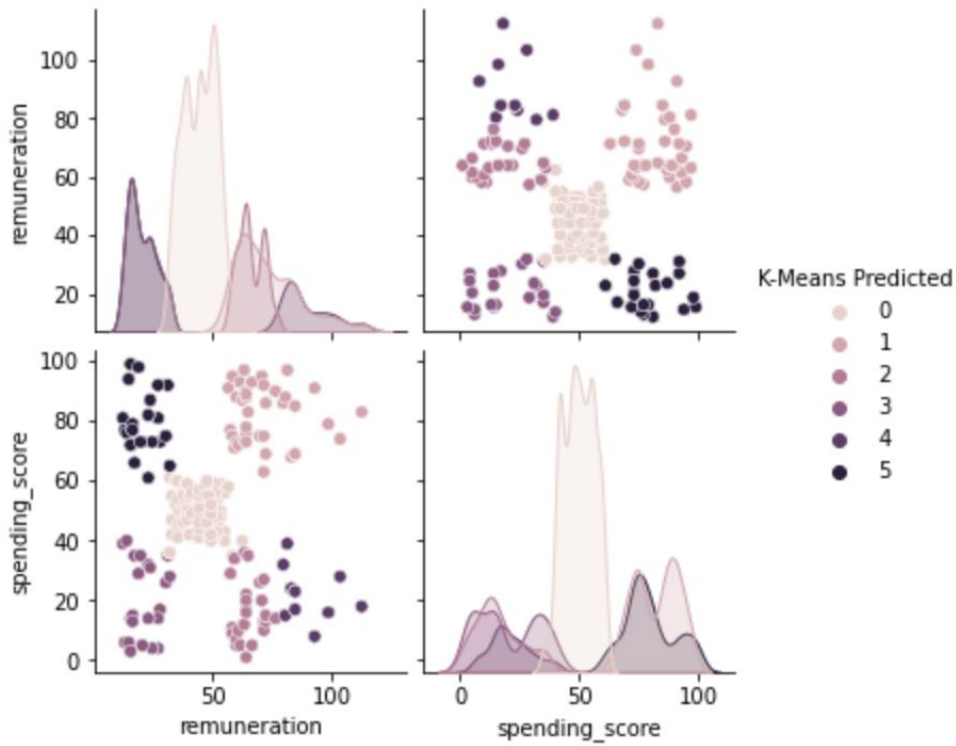
Fig. 10: K-means model with 7 clusters

Fig. 13: Top 20 positive reviews

	review	review_polarity
7	came in perfect condition	1.000000
164	awesome book	1.000000
193	awesome gift	1.000000
489	excellent activity for teaching selfmanagement skills	1.000000
517	perfect just what i ordered	1.000000
583	wonderful product	1.000000
601	delightful product	1.000000
613	wonderful for my grandson to learn the resurrection story	1.000000
782	perfect	1.000000
923	awesome	1.000000
1119	awesome set	1.000000
1150	best set buy 2 if you have the means	1.000000
1159	awesome addition to my rpg gm system	1.000000
1282	its awesome	1.000000
1380	one of the best board games i played in along time	1.000000
1523	my daughter loves her stickers awesome seller thank you	1.000000
1580	this was perfect to go with the 7 bean bags i just wish they were not separate orders	1.000000
1684	awesome toy	1.000000
1689	it is the best thing to play with and also mind blowing in some ways	1.000000
1695	excellent toy to simulate thought	1.000000

Fig. 14: Top 20 positive summaries

	summary	summary_polarity
6	best gm screen ever	1.000000
28	wonderful designs	1.000000
32	perfect	1.000000
80	theyre the perfect size to keep in the car or a diaper	1.000000
133	perfect for preschooler	1.000000
139	awesome sticker activity for the price	1.000000
160	awesome book	1.000000
162	he was very happy with his gift	1.000000
186	awesome	1.000000
209	awesome and welldesigned for 9 year olds	1.000000
412	perfect	1.000000
468	excellent	1.000000
536	excellent	1.000000
541	excellent therapy tool	1.000000
572	the pigeon is the perfect addition to a school library	1.000000
591	best easter teaching tool	1.000000
639	wonderful	1.000000
643	all f the mudpuppy toys are wonderful	1.000000
649	awesome puzzle	1.000000
654	not the best quality	1.000000

Fig. 15: Top 20 negative reviews

	review	review_polarity
207	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	-1.000000
181	incomplete kit very disappointing	-0.780000
1773	im sorry i just find this product to be boring and to be frank juvenile	-0.583333
362	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	-0.550000
116	i bought this as a christmas gift for my grandson its a sticker book so how can i go wrong with this gift	-0.500000
226	this was a gift for my daughter i found it difficult to use	-0.500000
229	i found the directions difficult	-0.500000
289	instructions are complicated to follow	-0.500000
300	difficult	-0.500000
1501	expensive for what you get	-0.500000
173	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	-0.491667
345	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	-0.446250
531	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	-0.440741
305	very hard complicated to make these	-0.439583
421	kids i work with like this game	-0.400000
430	this game although it appears to be like uno and have an easier play method it was still too time consuming and wordy for my children with learning disabilities	-0.400000
490	my son loves playing this game it was recommended by a counselor at school that works with him	-0.400000
795	this game is a blast	-0.400000
798	i bought this for my son he loves this game	-0.400000
814	was a gift for my son he loves the game	-0.400000

Fig. 16: Top 20 negative summaries

	summary	summary_polarity
21	the worst value ive ever seen	-1.000000
207	boring unless you are a craft person which i am	-1.000000
819	boring	-1.000000
1148	before this i hated running any rpg campaign dealing with towns because it	-0.900000
1	another worthless dungeon masters screen from galeforce9	-0.800000
143	disappointed	-0.750000
623	disappointed	-0.750000
785	disappointed	-0.750000
1591	disappointed	-0.750000
361	promotes anger instead of teaching calming methods	-0.700000
875	too bad this is not what i was expecting	-0.700000
880	bad qualityall made of paper	-0.700000
177	at age 31 i found these very difficult to make	-0.650000
100	small and boring	-0.625000
511	mad dragon	-0.625000
797	disappointing	-0.600000
1001	disappointing	-0.600000
1099	disappointing	-0.600000
1773	disappointing	-0.600000
991	then you will find this board game to be dumb and boring	-0.591667

Fig. 33: North American Sales per Product

North American Sales per Product

17

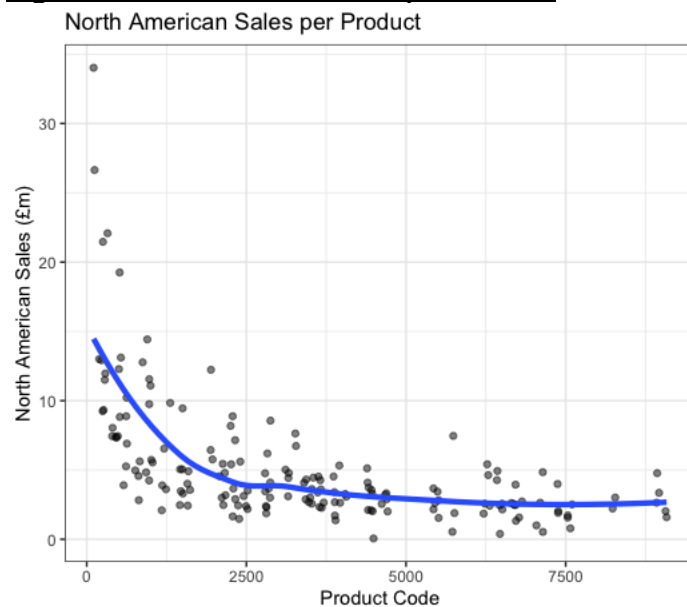Fig. 34: European Sales per Product

European Sales per Product

Fig. 35: Global Sales per Product
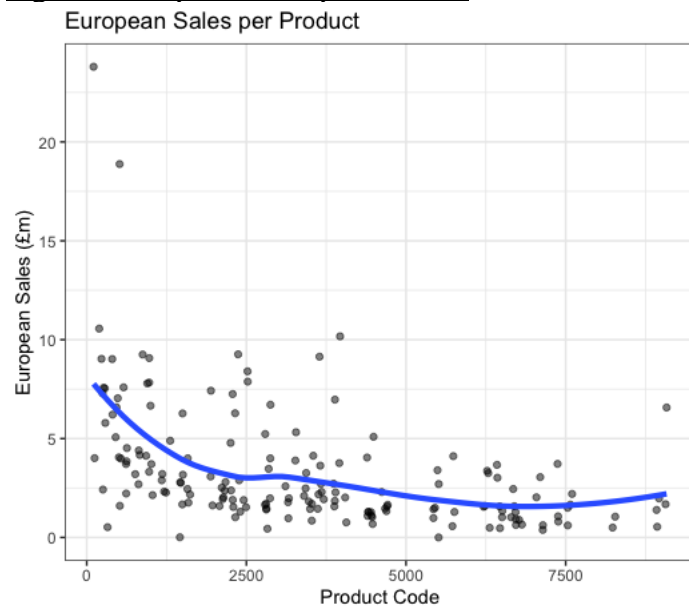

Global Sales per Product

Fig. 36: Distribution of North American Sales per Product

Distribution of North American Sales per Product
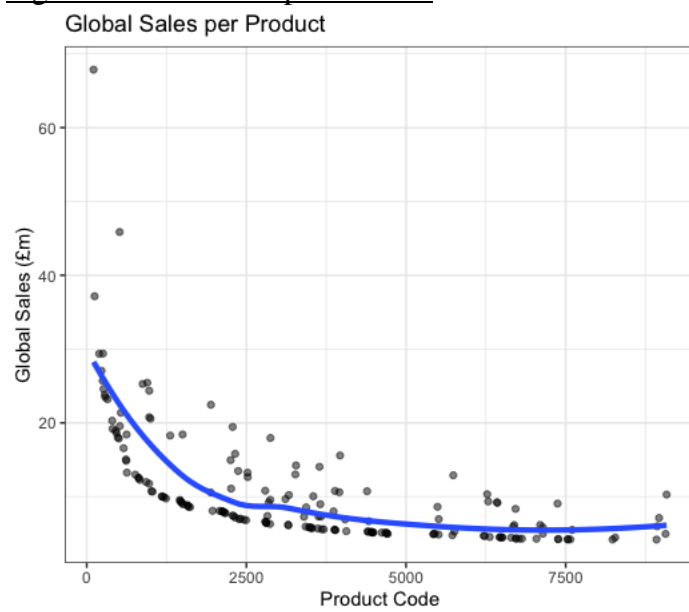
Fig. 37: Distribution of European Sales per Product
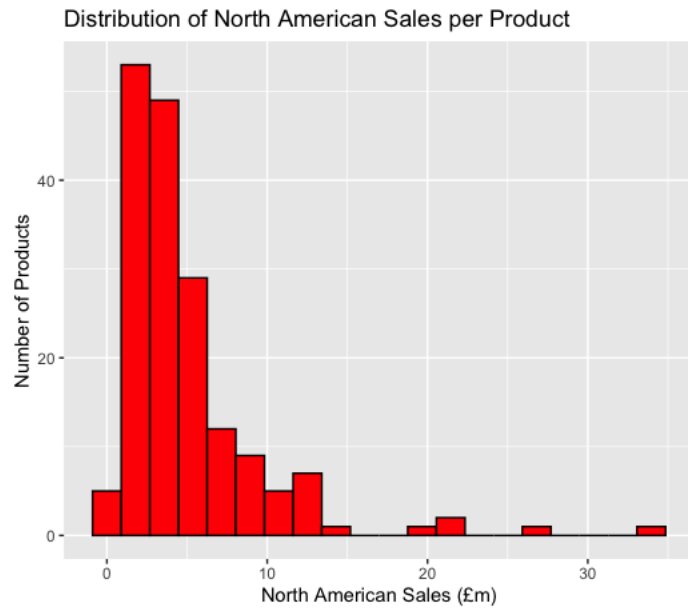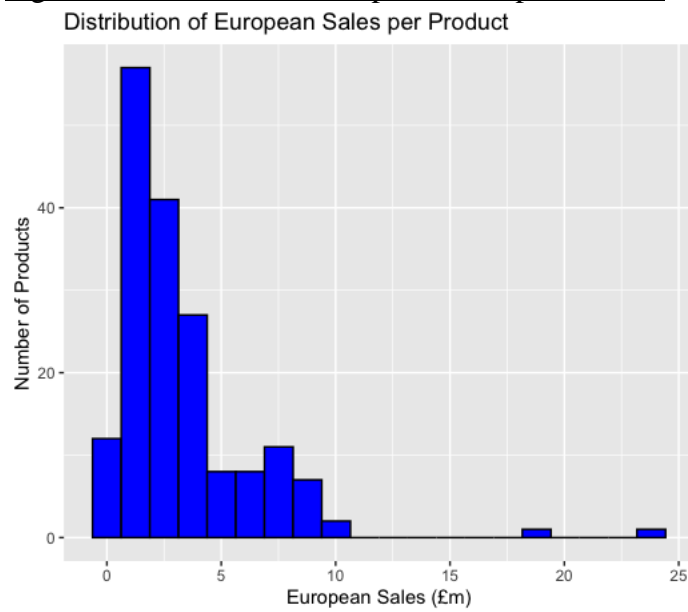

Distribution of European Sales per Product

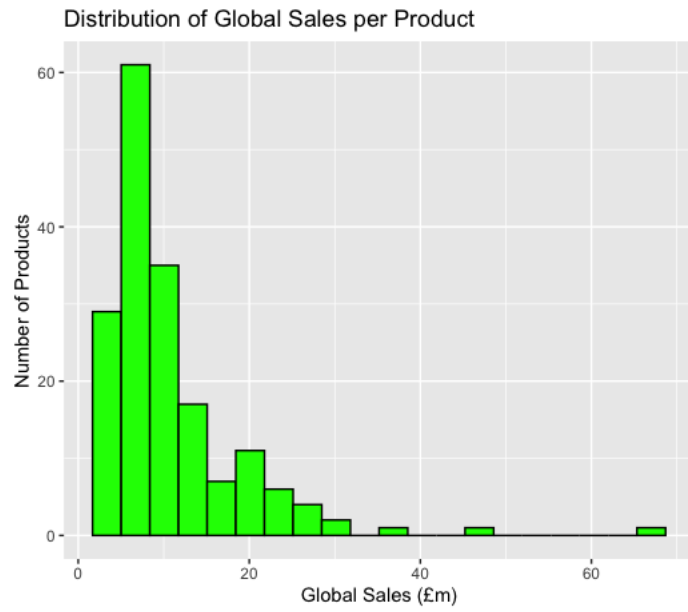Fig. 38: Distribution of Global Sales per Product

Distribution of Global Sales per Product

Fig. 39: Products by Platform
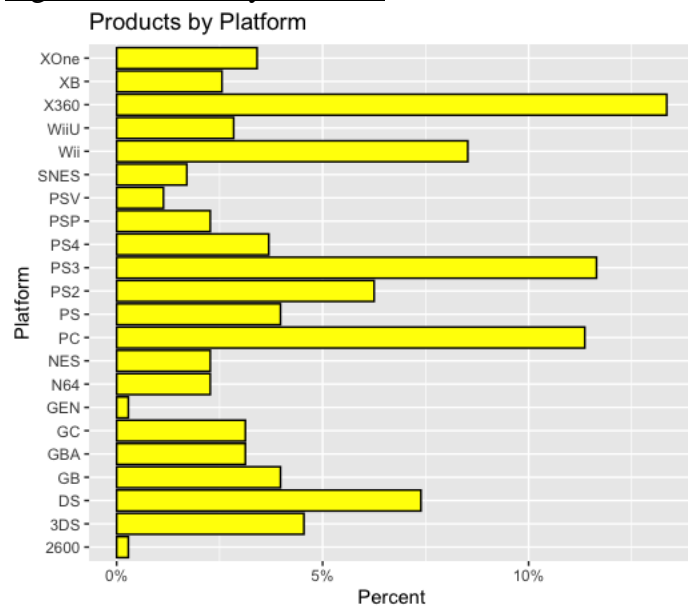

Products by Platform

Fig. 40: North American Sales by Platform

North American Sales by Platform

Fig. 41: European Sales by Platform
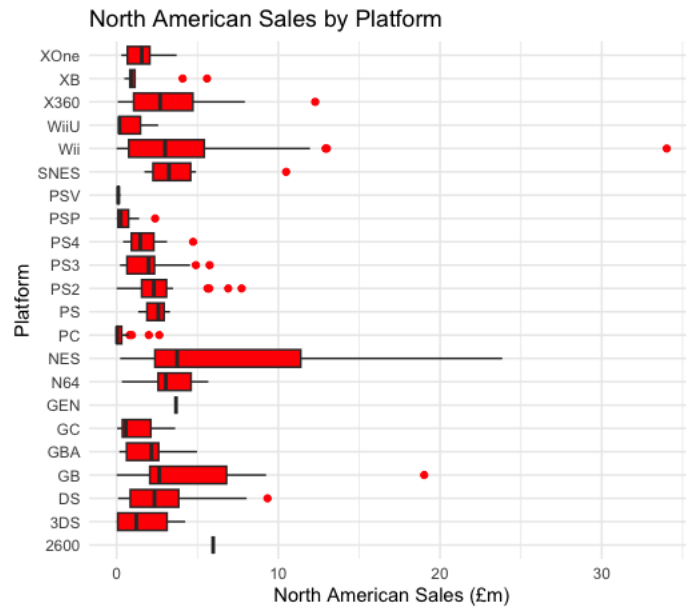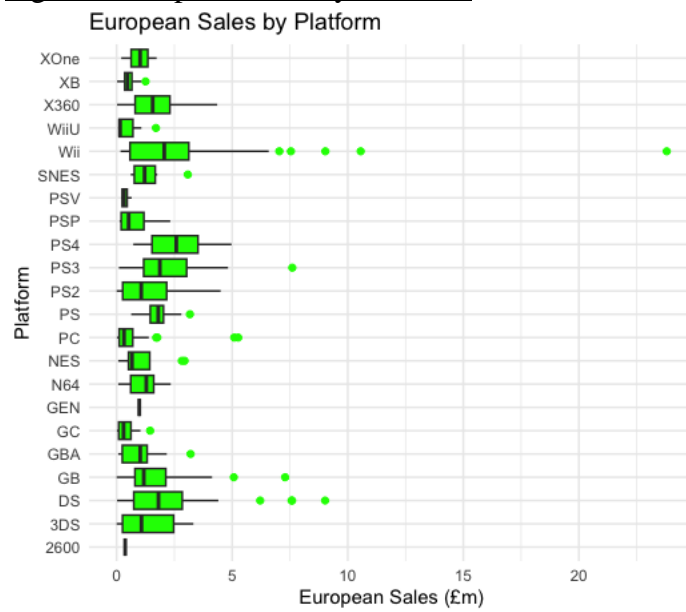

European Sales by Platform

Fig. 42: Global Sales by Platform

Global Sales by Platform