Turtle Games
Data analysis to help improve sales

Purpose of this report

This report is designed to allow Turtle Games to reperform our analysis either to validate the results or to extend it to additional data. It is therefore focused primarily on the analytical approach, assumptions, and questions.

Insights, patterns, and predictions are also covered, but for more detail on those please refer to our presentation of key insights.

Table of Contents

Context and purpose

The Turtle Games business challenge and goals

The overarching purpose for our data analysis is to identify and communicate insights that may help Turtle Games improve their overall sales performance.

To achieve our overarching purpose Turtle Games identified six related questions[1]:

1. How do customers accumulate loyalty points? More specifically, does the number of loyalty points accumulated by customers vary by customer age, remuneration, or spending score?
2. Which groups could customers be classified into based on their remuneration and spending score to enable more effectively targeted marketing campaigns?
3. How do customers tend to think and feel about video games supplied by Turtle Games? More specifically, what are the most common words used in customer reviews? What is the overall sentiment of customer product reviews? What are the 15 most common words in customer product reviews? What are the 20 most positive customer product reviews and what are the 20 most negative?
4. What is the contribution of individual video games to the sales in each of the customer geographic locations?
5. What is the statistical distribution of the sales data at the product-platform level, and at the aggregate product regardless of platform level? Specifically, how symmetric is the sales data around the mean (skewness)? How close to a normal distribution is the spread of sales values (kurtosis)?
6. How well can sales in one customer location predict sales in another?

[1] reformulated where appropriate

The data

We analysed customer reviews and sales data for 175 video games. Some of the 175 video games are available in multiple gaming platforms resulting in a total number of unique platform-game combinations of 352.

Sales segments based on the location of customers:

Customer location	Sales value - £'m	Sales value %
North America	886	47
European Union	579	31
Other countries	413	22
Total – global	1,878	100

As explained in more detail in the appendix, to maximise the probability of reaching valid insights and therefore the chances of enabling Turtle Games to make valid judgements and decisions, we filtered out 250 customer product review records relating to a total of 25 products which were not covered in the sales data.

We checked for obvious errors and omissions in the data in 'turtle_reviews.csv' and 'turtle_sales.csv',

by for example looking for missing values with .isnull()in Python and with sum(is.na() in R.

We removed redundant variables from the data as follows:

- From 'turtle_reviews.csv' we removed the review 'language' and 'platform' given that all reviews in the data were in English and online.
- From 'turtle_sales.csv' we removed the review 'Ranking', 'Year', 'Genre' and 'Publisher'. This was mainly to focus our analysis at this is initial stage. Further analysis could include these and other variables.

Other than the inconsistency between the two data sets mentioned above, we did not identify other obvious errors or omissions.

A list of more specific assumptions and questions is included in the appendix.

Libraries and functions

In both our Jupyter notebook for Python and our R script the libraries we installed are at the top, and we used comments throughout to explain the functions we used.

Loyalty points (week 1 – Python)

Analytical approach

To help Turtle Games answer the question

> 'How do customers accumulate loyalty points?'

we investigated the relationship between *loyalty points* as the outcome variable and respectively *age*, *remuneration*, and *spending score* as the predictor variables.

We started with looking at the correlation between each pair of variables.

Then we performed three simple linear regressions to understand how *loyalty points* may vary based on the variability of each of *age*, *remuneration*, and *spending score*.

A multiple linear regression would show how much of the variability in *loyalty points* remains unexplained after considering *age*, *remuneration*, and *spending score*. This analysis would be useful for exploring which variables together (*age*, *remuneration*, *spending score* and *others*) account for the variability of *loyalty points*. We did not consider it a priority to perform this analysis at this stage.

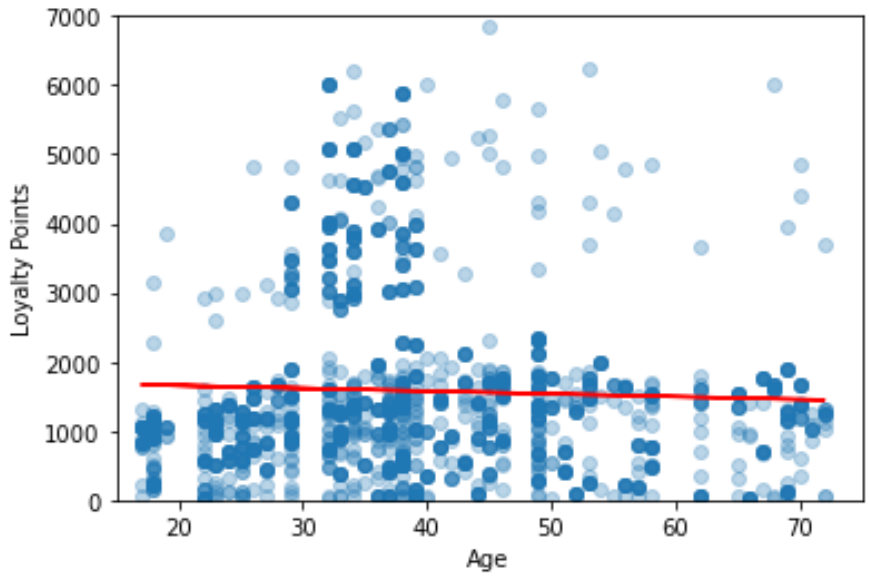
Visualisation and insights

The table below is a correlation matrix showing the correlation between each pair of variables.

	age	remuneration	spending_score	loyalty_points
age	1.000000	0.012766	-0.241476	-0.038950
remuneration	0.012766	1.000000	0.008998	0.628517
spending_score	-0.241476	0.008998	1.000000	0.658235
loyalty_points	-0.038950	0.628517	0.658235	1.000000

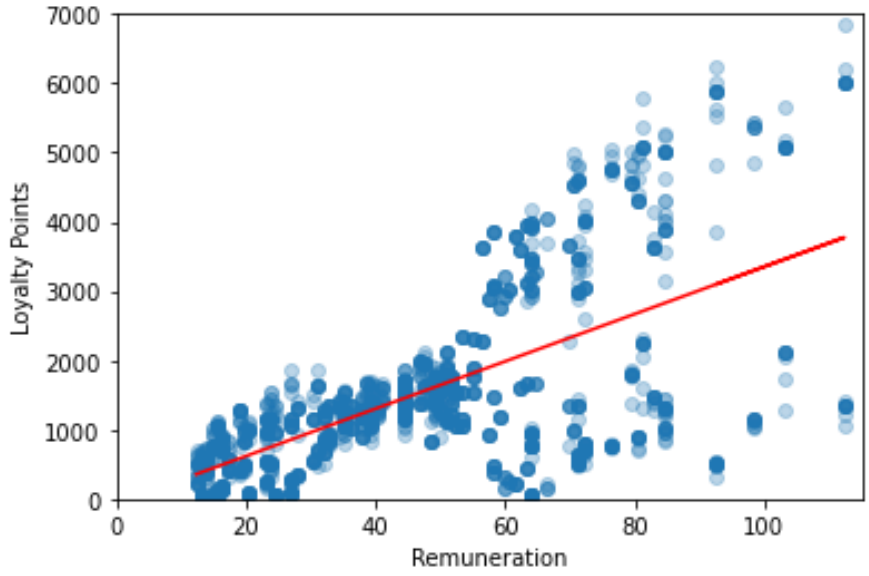
Age has a negligible negative correlation (-0.04) with *loyalty points* whereas *remuneration* and *spending score* have a relatively strong positive correlation with *loyalty points* of 0.63 and 0.66 respectively.

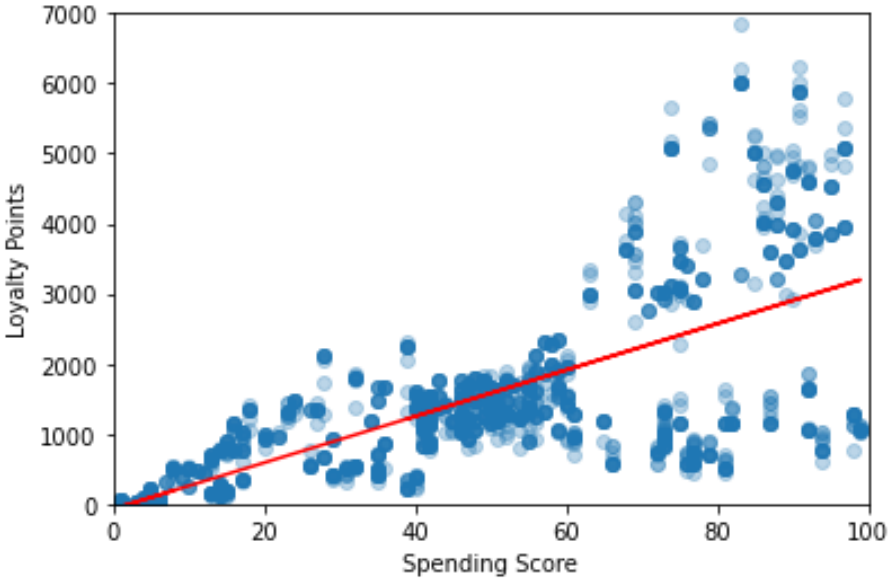
However, looking at the scatterplot for *age* and *loyalty points* below, we see a high concentration of customers in the age range of approximately 30 to 40 with high accumulations of loyalty points, suggesting that a non-linear model may be more effective in explaining the relationship between the two variables.



Our homoskedasticity tests for all three pairs of variables showed that the homoskedasticity assumption was not met, meaning that the variance of the loyalty points is not stable at all levels of each of the three predictor variables. A data transformation could address this issue. However, we did not consider it necessary to perform such a transformation at this stage of our analysis.

The two plots below show the relationship of loyalty points with remuneration and spending score respectively.




7000
6000
5000
4000
3000
2000
1000
0
Loyalty Points
0
20
40
60
80
100
Spending Score

Customer clusters (week 2 – Python)

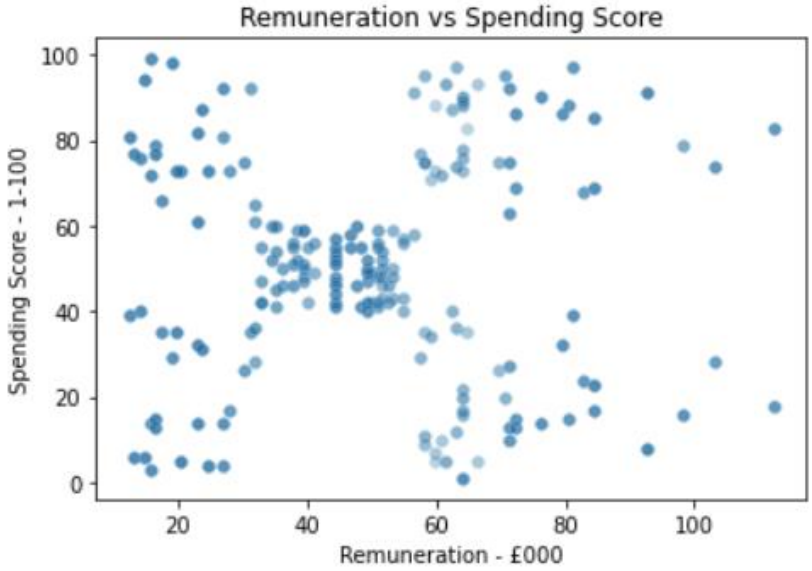
Analytical approach

To help Turtle Games answer the question

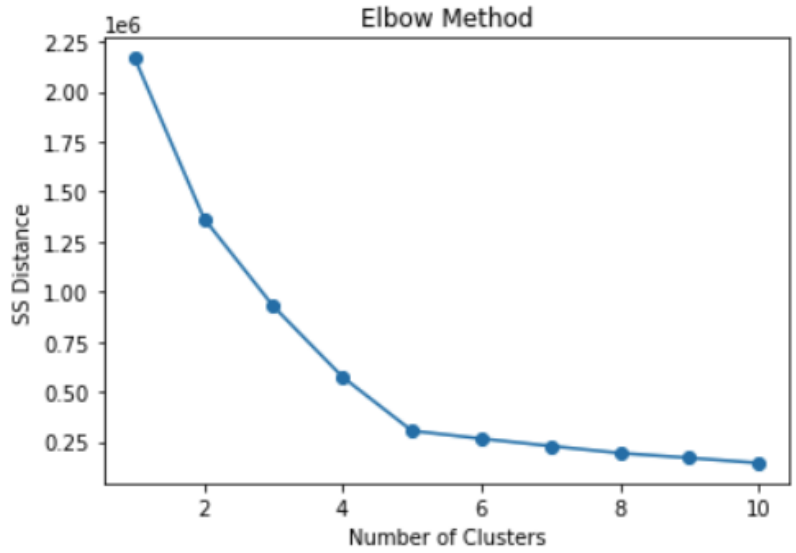
‘How can customers be allocated into categories for marketing purposes based on their *spending score* and *remuneration*?’

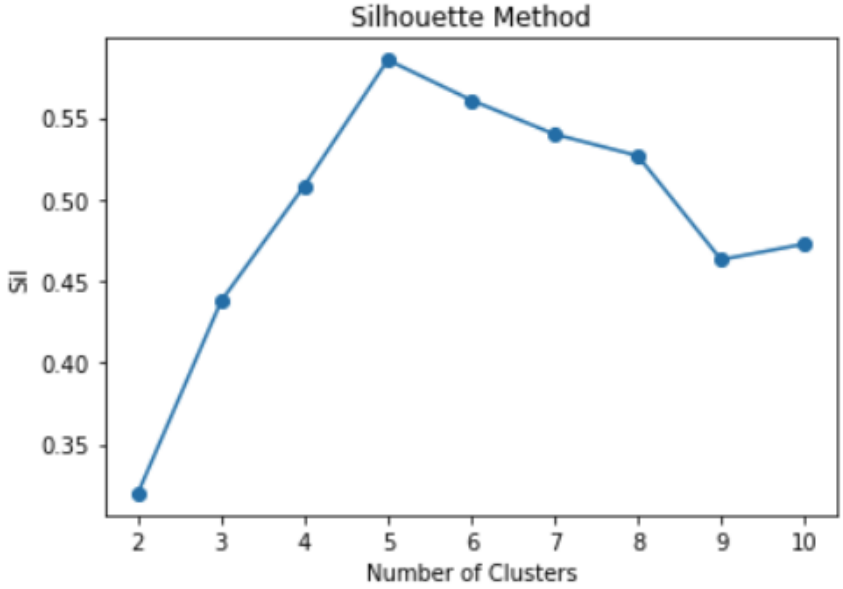
we used k-means clustering.

We started with the scatterplot shown below which gave us the impression of five or possibly seven clusters.



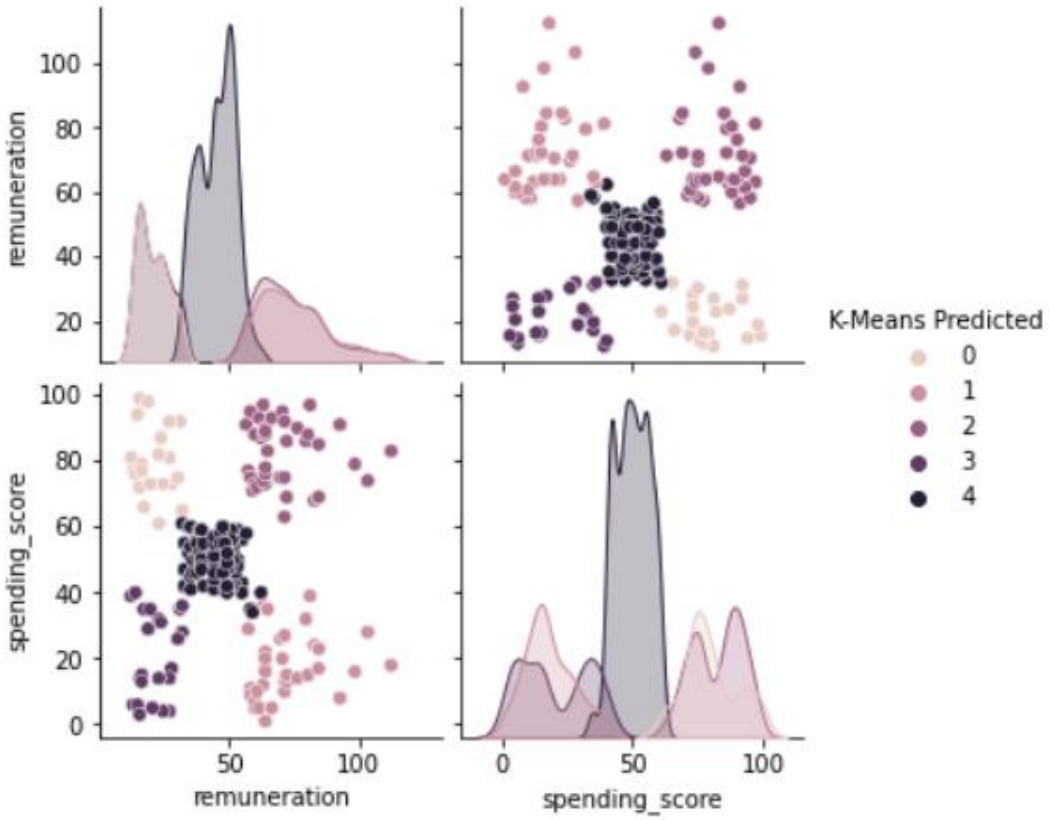
Then we used the Elbow and Silhouette methods the respective outputs of which are shown below.





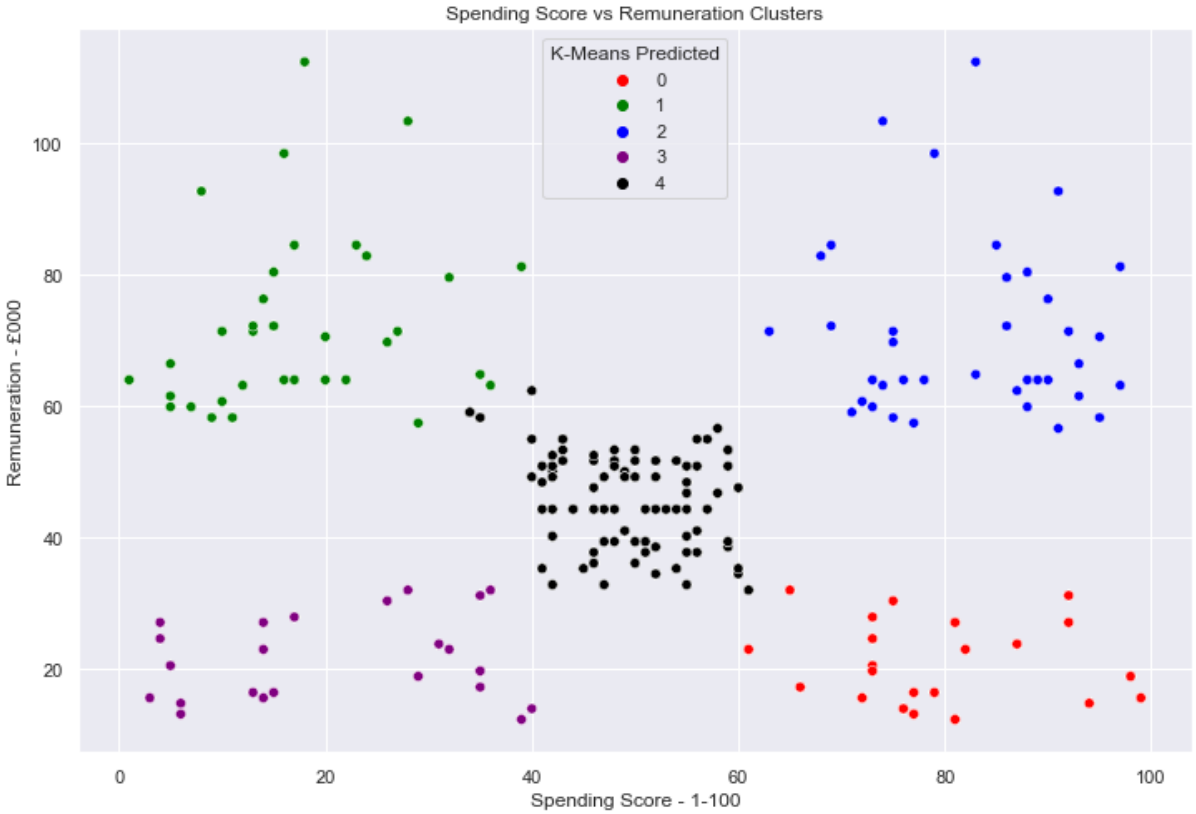
Both methods seemed to strongly suggest five as the optimal number of clusters.

We examined data visualisations and the number of members per cluster based on three, five and seven clusters and concluded that, subject to discussion with the marketing team at Turtle Games, five seems the most appropriate.



Visualisation and insights

We prepared the graph below for the purposes of our discussion with the Turtle Games marketing team.



We look forward to hearing the reaction from the marketing experts.

Our naive interpretation from a marketing strategy perspective is that a classification along the lines of the plot above could support a marketing strategy whereby customers are classified into classes for marketing campaign purposes based on a high-medium-low scale for remuneration and spending score.

Customer thoughts and feelings (week 3 – Python)

Analytical approach

To help Turtle Games answer the questions

> ‘What words do customers tend to use when talking about Turtle Games video games?’
>
> ‘Overall, how positive or negative are customer product reviews?’
>
> ‘What are the top 20 positive and top 20 negative customer reviews?

we used natural language processing.

Visualisation and insights

The fifteen most common words used by customers in their product reviews based on the *full* review comments are:

word (frequency)		
game (1192)	one (416)	play (392)
fun (376)	great (343)	like (336)
get (266)	really (254)	cards (239)
book (230)	tiles (227)	time (224)
would (223)	new (216)	well (214)

The fifteen most common words used by customers in their product reviews based on the *summary* review comments are:

word (frequency)		
game (248)	great (210)	fun (166)
good (73)	love (58)	like (48)
kids (43)	cute (39)	expansion (39)
book (35)	old (31)	really (26)
best (25	one (25)	excellent (25)

We also prepared the word clouds below which may be a an additional and more impactful way to present the most common words results to the Turtle Games marketing team.

Word cloud based on *full* product review comments

Word cloud based on summary product review comments

From both data sets (full and summary reviews) the most common words seem to uniformly include only positive and neutral words.

In the interests of keeping this report relatively short we did not reproduce the top 20 most positive and top 20 most negative full and summary reviews. These are available in the Jupyter Notebook.

In summary, based on the *full* product review comments the median compound score is positive .84 with a first quartile also positive .6. Based on the *summary* product review results the median compound score is positive .51 with a first quartile neutral 0.

Contribution of individual video games to the sales in each of the customer geographic locations (week 4 – R)

Analytical approach

To help Turtle Games answer the question

'How do individual video games contribute to the sales in each of the customer geographic location categories used for marketing purposes?'

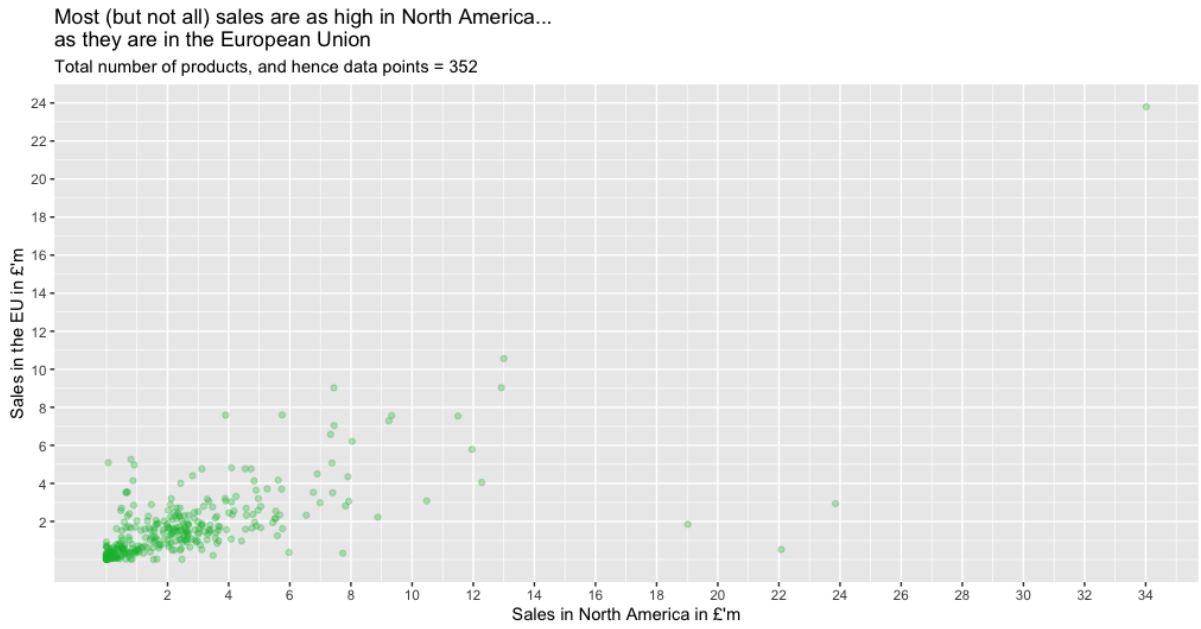
we produced a set of scatterplots, histograms, and boxplots.

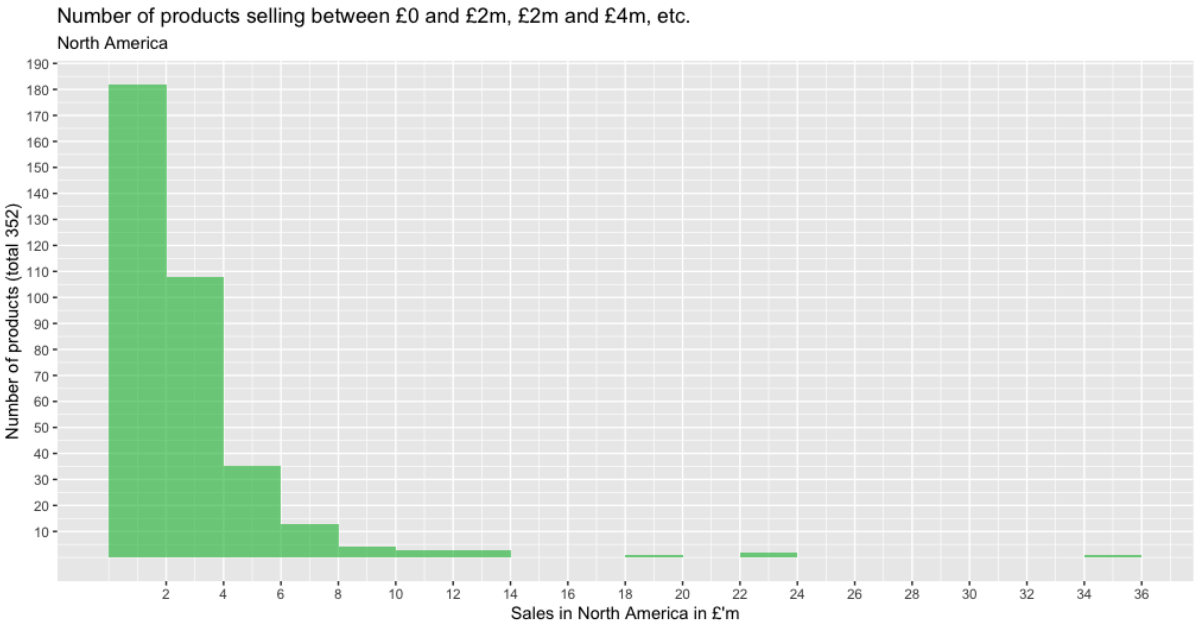
In addition to the request to visualise the data for North America, the European Union, and globally, we also prepared visualisations for sales in other countries. Other countries at around 22% of global sales could play a significant part in the strategy for improving sales.

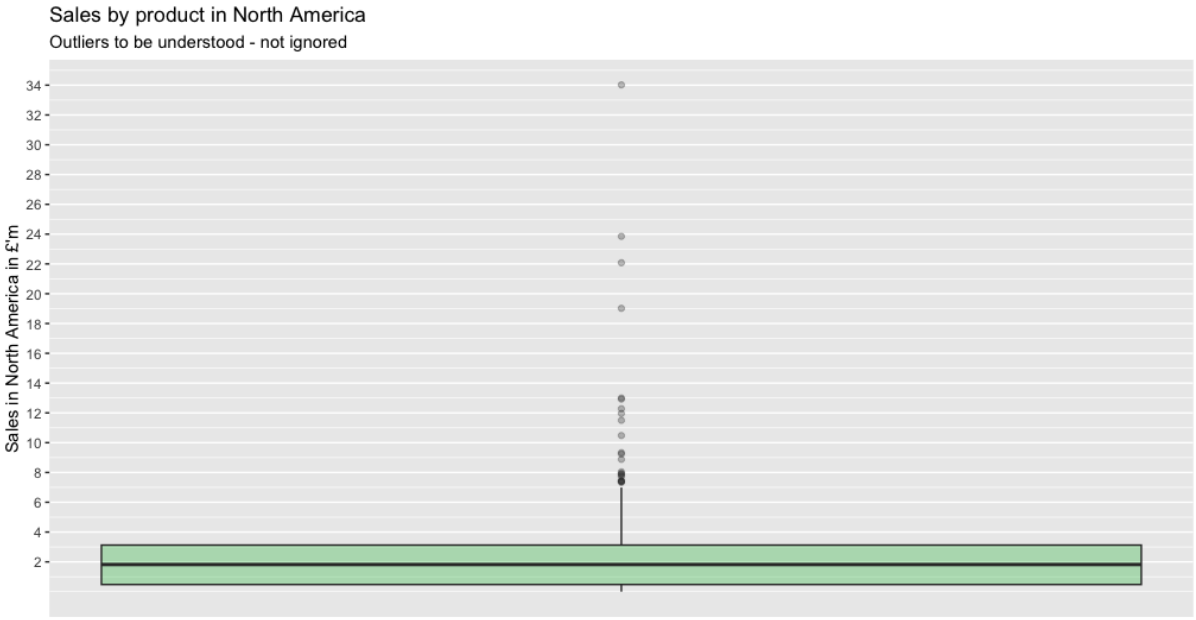
In addition, by calculating the sales for other countries (global sales minus sales in NA and the EU) we were able to gain additional evidence that there are no obvious data issues given none of the products shows a negative value for sales in other countries.

Visualisation and insights

Below we reproduce a representative sample of the three different types of plots we prepared for this part of our analysis.







Although not reproduced here, the respective scatterplots, histograms, and boxplots reflecting the data for the EU, other countries and global are similar to those illustrated above. The sales performance of individual video games seems relatively consistent across locations.

Statistical distribution of the sales data at the product-platform level, and at the aggregate product level (week 5 – R)

Analytical approach

To help Turtle Games answer the question

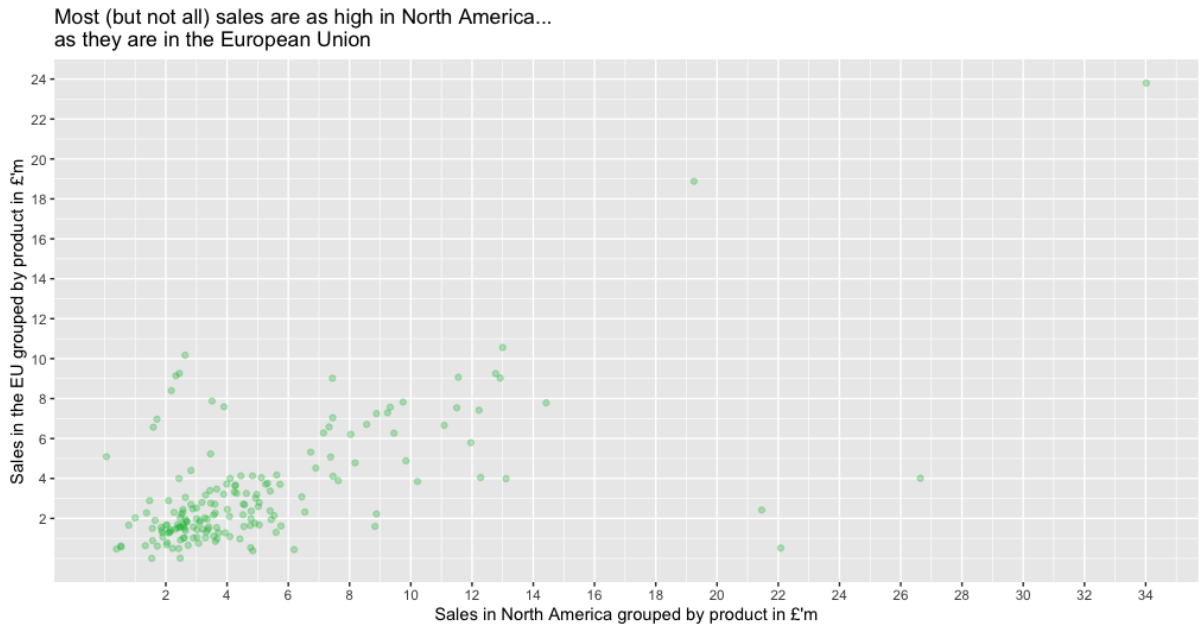
'What is the statistical distribution of the sales data at the product-platform level, and at the aggregate product regardless of platform level?'

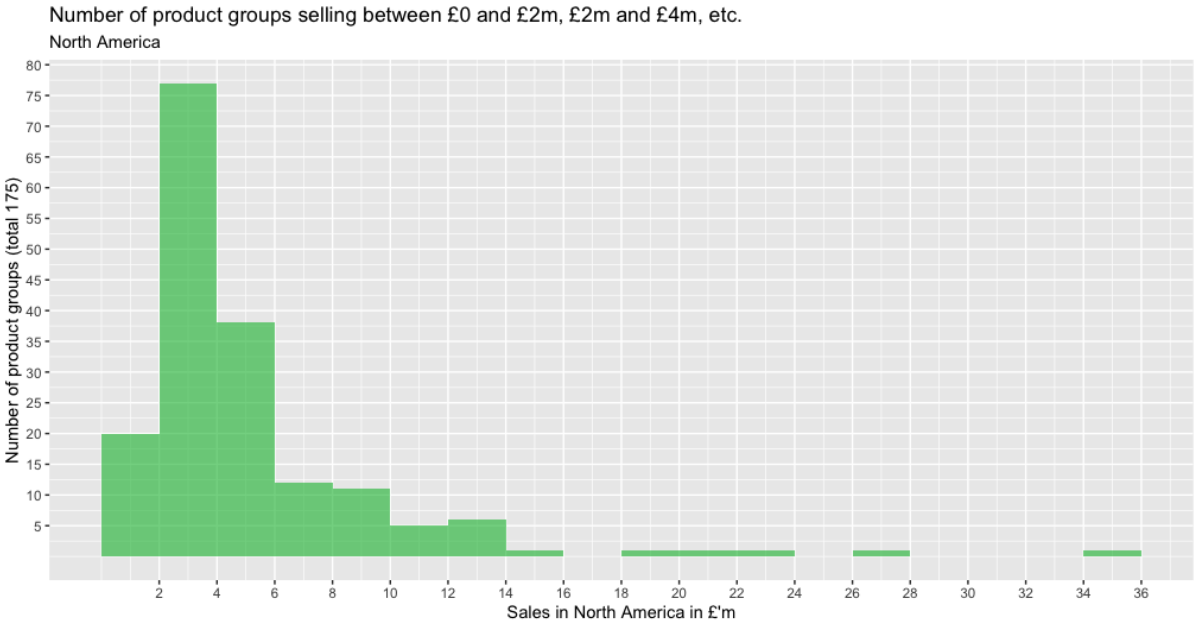
we followed the steps listed below:

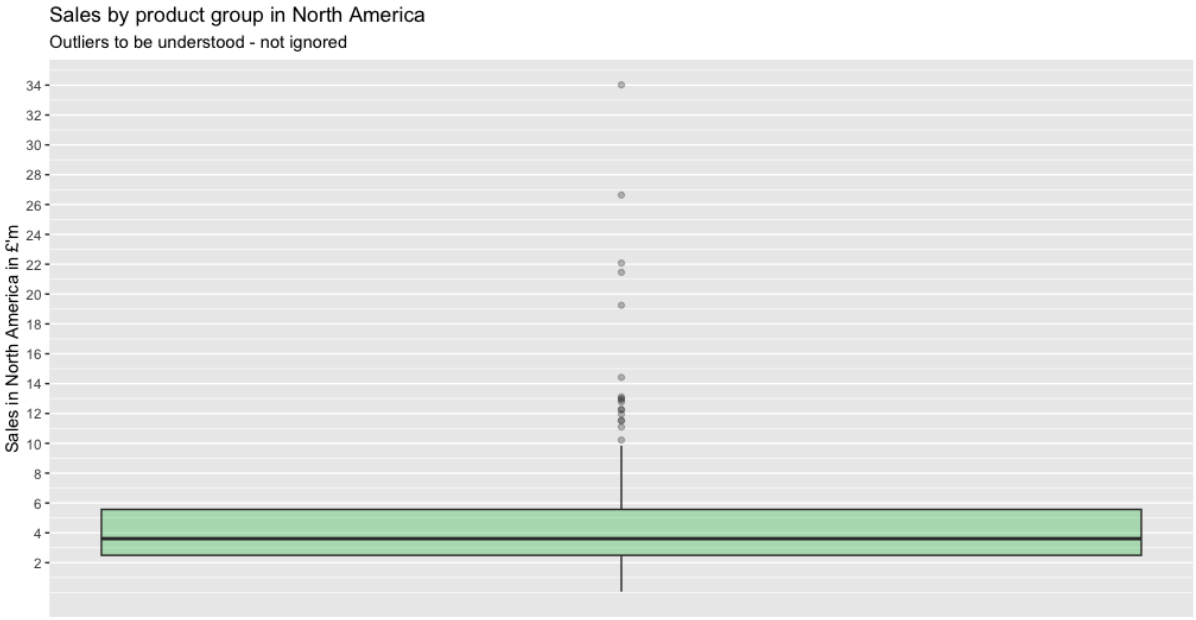
1. Used the group_by() function to create product group sales regardless of gaming platform
2. Created scatterplots, histograms, and boxplots as we did for the data at the individual video game level
3. Created Q-Q plots, used the Shapiro-Wilk test, and calculated the Skewness and Kurtosis for each data set at both the unique video game / gaming platform combination, and at the video game group level regardless of platform.
4. Calculated the correlation between sales at different locations.

Visualisation and insights

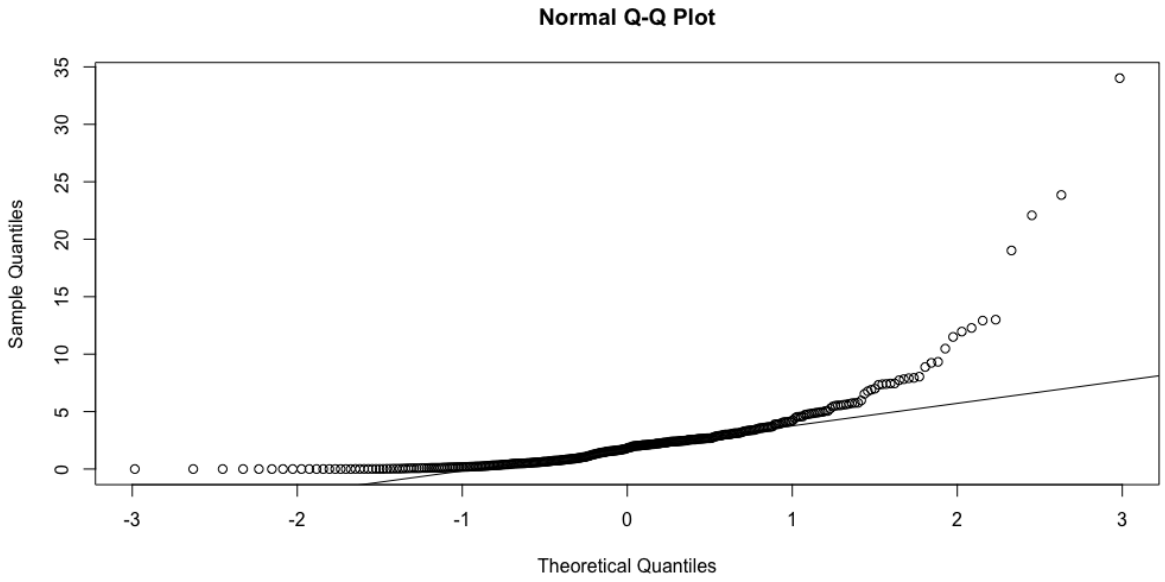
Below we reproduce a representative sample of the results of our analysis.




Number of product groups selling between £0 and £2m, £2m and £4m, etc.
North America
Number of product groups (total 175)
Sales in North America in £'m


Sales by product group in North America
Outliers to be understood - not ignored
Sales in North America in £'m

Q-Q plot for sales of video games in North America (without grouping by product)



The Q-Q plot above is similar for those for the other data sets and it indicates a heavy tail on the high end of sales values (some video games sell at many standard deviations above the mean) and a light tail at the low end of sales values.

The Shapiro-Wilk test was statistically significant for all data sets indicating that normal distribution cannot be assumed.

The results of our calculations of Skewness were consistent across all data sets with values above 4 indicating positively skewed data (0 would indicate normal distribution).

The results of our calculations of Kurtosis were consistent across all data sets with values above 30 (3 would indicate normal distribution).

Predicting global sales based on sales in NA and the EU (week 6 – R)

Analytical approach

To help Turtle Games answer the question

> ‘How well can sales in one customer location predict sales in another?’

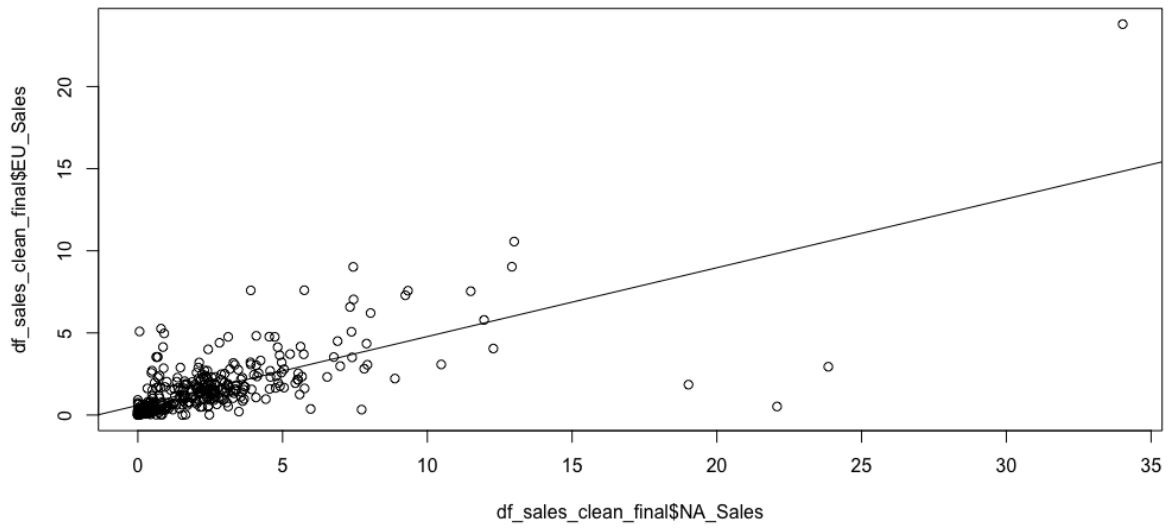
we created a simple linear model to represent the relationship between sales in North America and the EU, and a multiple linear model to represent the relationship between sales globally (outcome variable) and sales in North America and the EU (respectively predictor variables).

We also used the multiple regression model to predict the global in five instances of known NA and EU sales values.

Visualisation and insights

Below is a sample of our results.

Using sales in North America to predict sales in the EU.



Coefficients:

(Intercept)	NA_Sales
0.5891	0.4192

Using sales in NA and the EU to predict sales globally.

Coefficients:

(Intercept)	NA_Sales	EU_Sales
0.2217	1.1554	1.3420

Individual predictions

Product	Platform	NA_ Sales	EU_ Sales	Global_ Sales	Other	Predicted_ Global	Observed_v_ Predicted	Percent
107	**Wii**	34.0	23.8	67.9	10.0	71.5	-3.6	-5%
326	**NES**	22.1	0.5	23.2	0.6	26.4	-3.2	-12%
3267	**X360**	3.9	1.6	6.0	0.6	6.9	-0.8	-12%
6815	**N64**	2.7	0.7	4.3	0.9	4.2	0.1	2%
2877	**X360**	2.3	1.0	3.5	0.3	4.1	-0.6	-15%

Appendix

Understanding the data

Turtle Games is a game manufacturer and retailer with a global customer base. The company manufactures and sells its own products, along with sourcing and selling products manufactured by other companies.

Its product range includes books, board games, video games, and toys.

Turtle Games provided us with two data sets, and metadata for the two sets combined.

Copies of the original two data set files and the metadata file are at our GitHub repository. Their names are:

'turtle_reviews.csv'

'turtle_sales.csv'

‘metadata_turtle_games.txt’

Do both data sets relate to video games only?

The metadata makes it explicit that ‘turtle_sales.csv’ relates to video games only given one of the variables is ‘Platform’ described as ‘*the video game console on which the video game was launched*’.

However, the metadata is not clear about whether ‘turtle_reviews.csv’ relates to video games only.

Our investigation using Excel filtering of unique values and sorting to compare the list of products in ‘turtle_reviews.csv’ to the list of products in ‘turtle_sales.csv’ showed that all 175 products in the latter are included in the former. However, ‘turtle_reviews.csv’ includes an additional 25 product codes (range 9119 to 11086) not included in ‘turtle_sales.csv’.

In addition, from looking at some of the review comments we noted that at least some of the 25 product codes relate to products other than video games.

We decided to filter out the reviews data relating to these 25 products to avoid the risk of introducing noise into our analysis, especially in the context of the customer clustering and review comments sentiment analysis.

Assumptions and questions

1. What are the controls in place to ensure the accuracy, completeness, and validity of the data in ‘turtle_reviews.csv’ and ‘turtle_sales.csv’.

2. Which time period is covered by each of the two data sets?

3. What was the approach and rationale for selecting the sample of customer reviews in ‘turtle_reviews.csv’?

   Is the sample representative of the Turtle Games global customer base?
   Which countries are the customers represented in the sample of product reviews based in?
   Their remuneration may not be comparable if based in countries with varying cost of living.

4. The metadata describes 'spending_score' as *'a score (between 1 and 100) assigned to each customer based on the customer's spending nature and behaviour'*.

   How are *'spending nature'* and *'spending behaviour'* measured?
   How is *'spending_score'* mapped to *'spending nature'* and *‘spending behaviour'* measures?
   Are spending scores comparable across customers? They might not be if for example spending scores depend on the length of time since becoming a customer.

5. The metadata describes 'loyalty_points' as *'a score based on the point value of the purchase'*.

   How are 'loyalty_points' mapped to 'purchase value' (for example, is it one loyalty point per £1 of purchase)?

6. Why does the data in ‘turtle_sales.csv’ not include games manufactured by Turtle Games?