Report – Turtle games analysis to improve overall sales performance

Background/context of the business

Turtle Games is a game manufacturer and retailer with a global customer base. The company manufactures and sells its own products, along with products manufactured by other companies. Its product range includes books, board and video games and toys. Turtle Games has a business objective of improving overall sales performance by utilising customer trends.

Turtle Games wants to understand:

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

Analytical approach

Turtle games have provided two data sets to support the analysis; 'turtle_reviews', which contains customers reviews in English from their website and turtle_sales which contains sales figures for their video game products.

Turtle games requested that the analysis to answer questions 1-3 was undertaken in python. The following sense check and data preparation was undertaken:

Action	Output	Conclusion / Notes
Import csv file 'turtle_reviews'	Data imported as data frame	Data successfully imported
Review metadata	N/A	There is no indication as to how the 'summary' is obtained compared to the 'review'. Based on the content it appears to be separate and in addition to the review, rather than a portion of the review.
Check record number	2000 records with 11 variables	The key columns necessary to explore the business question were present Language is always English, and platform is not required.
Check data types (consistency)	Most auto assigned data types appear to match data.	Product ID number has been auto assigned the integer type, but is really a categorical variable. No need to changes data types initially.
Check for missing data (completeness)	None	There were no missing values in the data frame.
Check on key descriptive statistics (accuracy)	(see Jupyter notebook)	Key stats show that data set has a wide range for age, income, spending score and loyalty point score. This suggests that the data provides a wide sample of customers. No obvious anomalies.

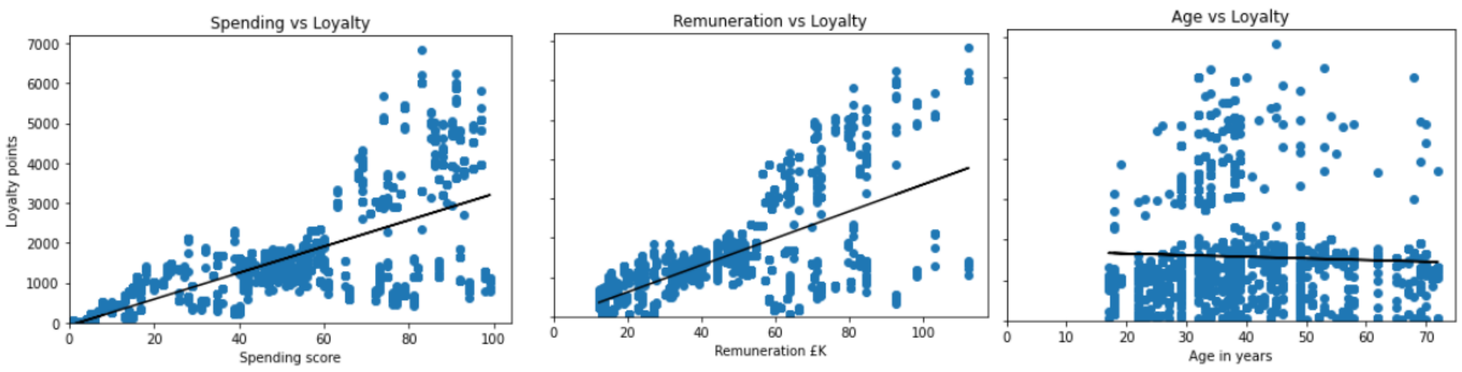
Turtle games requested that the analysis to answer questions 4-6 was undertaken in R. The following sense check and data preparation was undertaken:

Action	Output	Conclusion / Notes
Import turtle_sales csv	Data imported as a data frame	Data successfully imported
Review metadata	NA	Sales data appears to be for video games only, rather than all games. Sales figures are in GBP millions. Global sales is the sum of a number of regions, but only the break down for the North American and European regions are provided.
Check record number	352 observations, 9 variables	Not all the variables are required. New data frame created without Ranking, Year, Genre or Publisher.
Check data types (consistency)	Most auto assigned data types appear to match data.	Product ID number has been auto assigned the integer type, but is really a categorical variable. No need to changes data types initially.
Check for missing data & locate (completeness)	2 missing values	Although there are 2 rows with missing values which will be left in, as they are in the year column which will be removed, as it is not required.
Check for duplicates	None	No duplicate records present in data set.
Check on key descriptive statistics (accuracy)	(See R script)	North American sales appear to be higher on average than European sales.

Full details of the specific analysis required to respond to each of the questions is detailed in sections 1-6 below.

1. How do customers accumulate loyalty points?

A regression analysis using an Ordinary Least Squares (OLS) test was performed on the data to see if there was a relationship between customer loyalty points, their spending score, income or age and to what extent these variables explain loyalty point acquisition.



Spending vs Loyalty

It was expected to find a correlation between spending points and loyalty points because, as per the metadata, the loyalty points are awarded 'based on the point value of the purchase, converting the monetary value to point purchases'. While the spending score is 'based on the customer’s spending nature and behaviour.'

The graph shows a weak positive correlation between the spend score and loyalty, but is not a great predictor, especially as after 60 spending points there is no correlation. Therefore do not extrapolate beyond 60. The p-value is highly significant so we can reject the null hypothesis and conclude there is a statistically significant relationship between the two scores. The R-squared value indicates that 45% of the loyalty points score is explained by the spending behaviour (spending score) of the customers. The higher the spending score the more loyalty points up to a spending score of 60.

Remuneration vs Loyalty

There is a strong positive correlation between income (remuneration) and loyalty points. The p-value is significant, showing there is a statistically significant relationship between the two. The R-squared value indicates that 38% of the loyalty points score is explained by the income (remuneration) of the customers. The greater the income of the customer the more loyalty points they have up to an income of about £55,000 after which income doesn't predict loyalty point score.

Age vs Loyalty

There is no relationship between age and loyalty points. This is supported by the graph which shows an almost horizontal line. Additionally, there is a p-value of 3.61 indicating a lack of statistical significance and an R-squared score of just 2%.

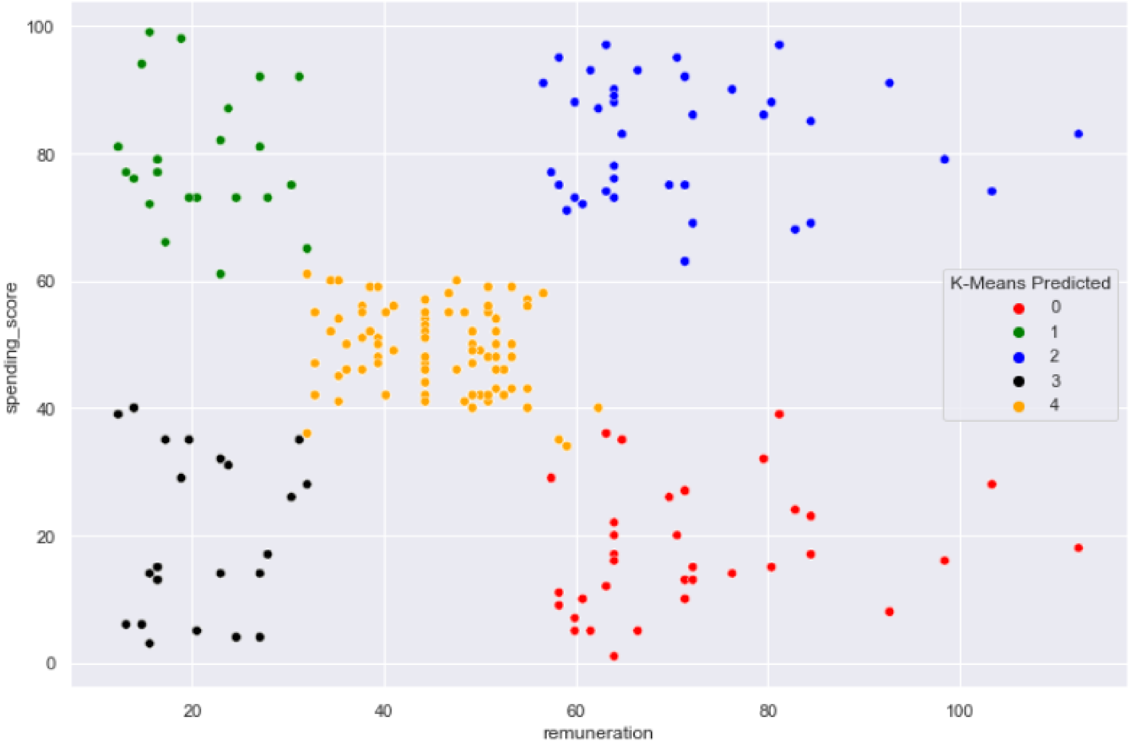
Conclusion

Combined, spending and income (remuneration) do a good job of explaining loyalty point acquisition. Further analysis based on these variables can be taken forward to answer the other business questions.

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

An initial visualisation, based on remuneration and spending scores, suggests an obvious grouping that would support a cluster analysis. This was verified using k-means clustering and the Elbow and Silhouette methods to establish the optimum number of 5 clusters. Additionally, k-means was tried using 4 and 6 clusters to verify that 5 provided the best groupings for further analysis.

Conclusion



There are 5 distinct clusters of customers based on remunerations and spending score. Cluster 0 shows a high annual income but low spending, it would be worth exploring this cluster more to understand if there are any other variables that define this cluster and what might move these customers toward greater spending. We could also compare them against cluster 2 customers who earn similar incomes to cluster 0 but spend significantly more.

3. How can customer reviews be used to inform marketing campaigns?

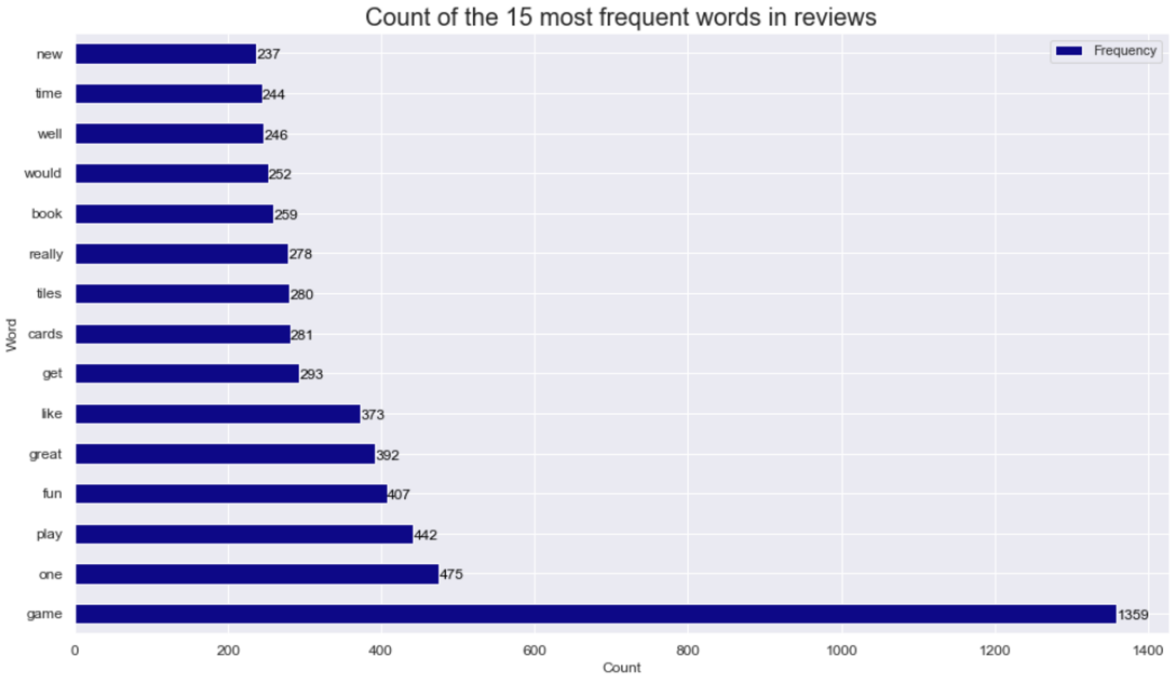
Natural Language Processing (NLP) techniques were used to analyse the review data and provide the information requested by the marketing department. As they appear to be distinct, both the review and summary columns were prepared and analysed.

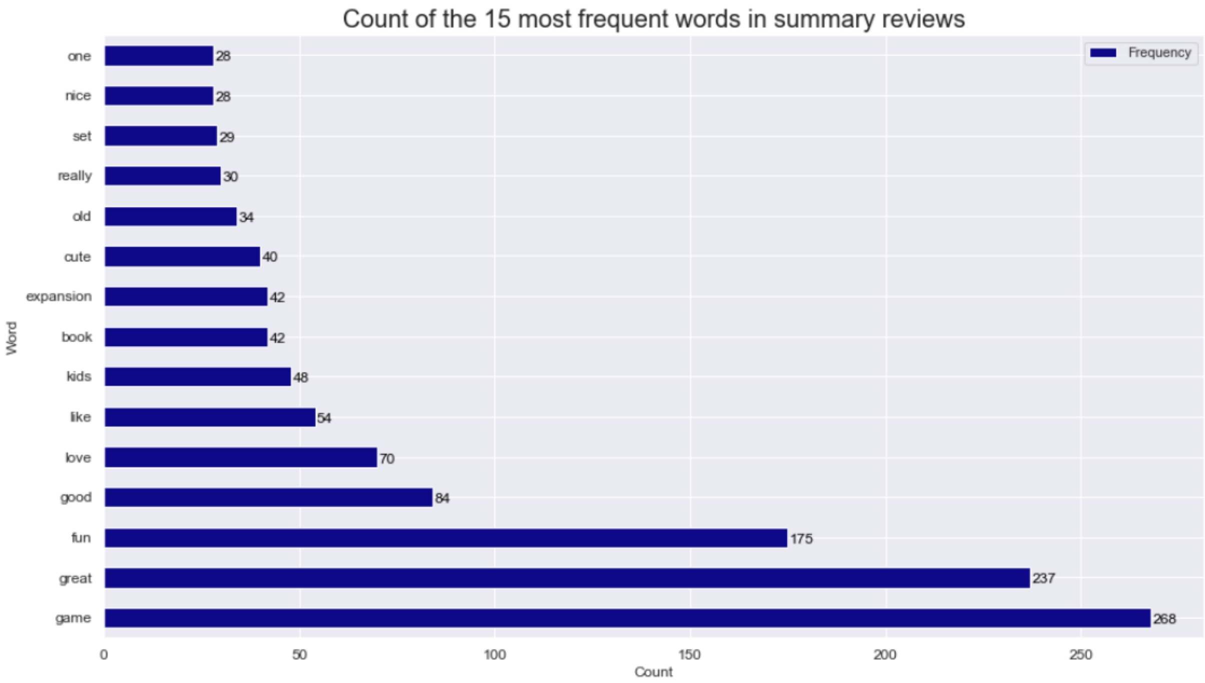
To convert the sentences into useable data, they were transformed into chunks of information that Python functions can correctly identify.

This was achieved by:

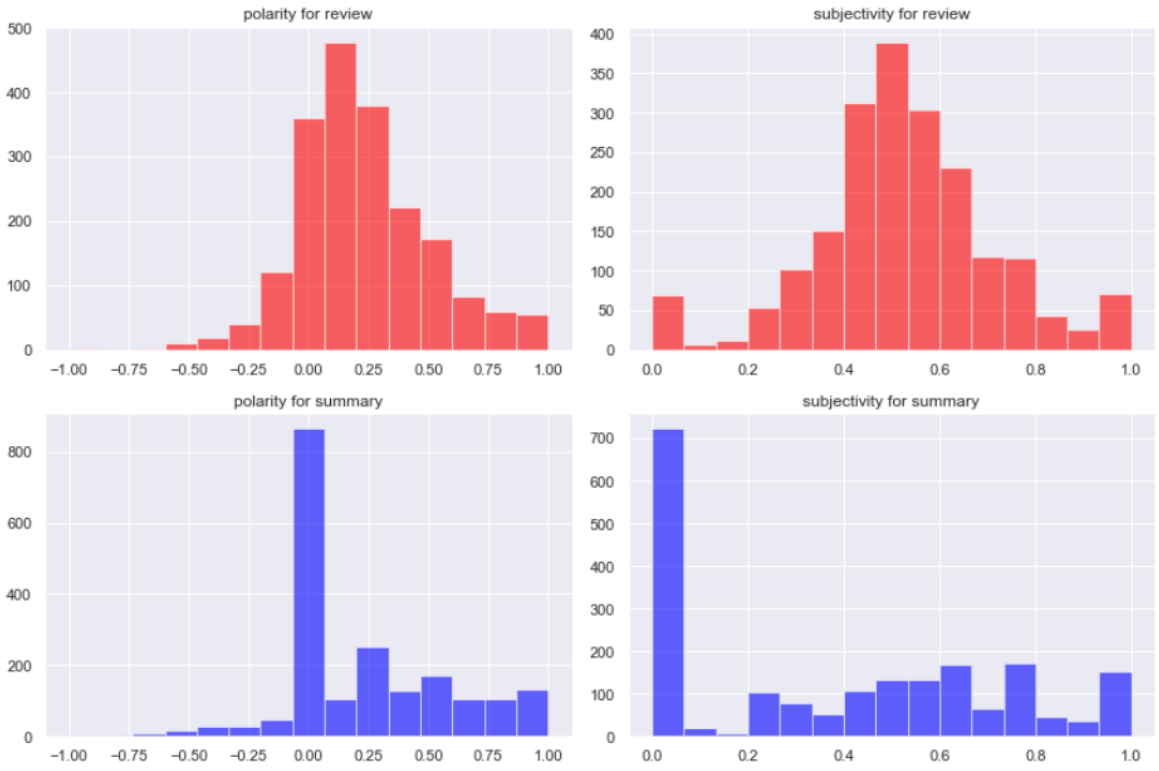
1. Changing all words to lower case.
2. Replacing punctation with blank spaces.
3. Checking for and removing duplicates.
4. Transforming sentences into individual words using tokenisation
5. Removing stop-words like ‘and’ or ‘the’, which appear frequently in sentences but provide no value in sentiment analysis.

The 15 most common words used in online product reviews and summary reviews is shown below. We can see that if we exclude the word game (which has no value), the top 5 words from each column are similar.





Addtitionally, sentiment analysis was performed using ‘Textblob’ to score the comments for polarity (how positive, neutral or negative they are) and subjectivity (to quantify the amount of personal opinion and factual information contained in the text). The results were plotted in histograms for comparison:



Polarity score = scale of 0 (negative) to 1 (positive) **Subjectivity score** = scale of 0 (objective) to 1 (opinion)

(Note that the Y axis on each chart is slightly different but the shape and X axis are the important features)

The histograms show that both the review and summary reviews are skewed positively, indicating more positive than negative reviews. While the subjectivity scores mainly zero for summary, indicating more factual information. Summaries show a more dispersed response ranging from fact to public opinion. Subjectivity scores can be challenging to interpret, but could provide scope for further analysis.

The polarity score was used to filter the comments to get the 20 most positive and negative reviews, the full lists are included in annex 1. Textblob has achieved mixed results in correctly identifying positive and negative comments, for example :

"My son loves playing this game. It was recommended by a counselor at school that works with him."

This sounds quite positive, but was given -0.4 rating indicating negative sentiment. Further analysis by individuals who know the products would be beneficial to extract the value from these comments.

Most of the negative comments seem to be about poor instructions or missing parts. It would be good to look at this further and see if it was linked to specific types of products and take action to resolve the issues or perhaps stop selling this product. A number of customers refer to the fact that a product was recommended by a therapist and it could be useful to explore what product/s these are and whether there is a market that could be successfully targeted in this area. Additionally, aspect based sentiment analysis might create a clearer view by classifying the aspect / element of the game that a customer liked or disliked.

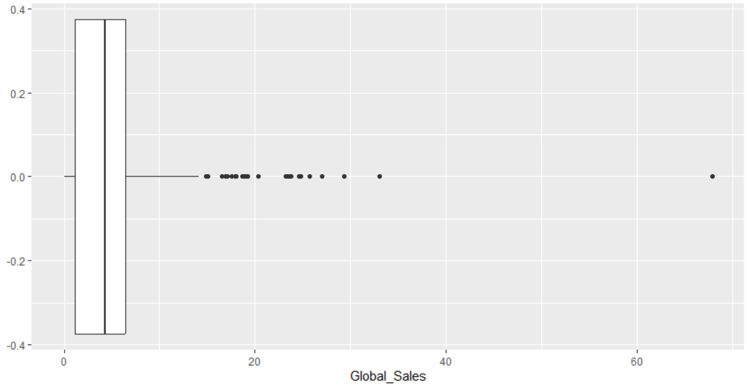
Conclusion

NLP is a growing field and needs to be managed with care when drawing conclusions from the analysis. It can provide a way to quickly gather a high-level overview of customer sentiment towards a product or company. However human language is extremely complex and requires context to understand, as such the various sentiment analysis options do not always categorize sentiment correctly. However, unlike humans they are consistent, which means they will always use the same methodology (as opposed to 2 humans that might categorize the same sentence differently), which can be extremely valuable for repeated testing and assessment on the success or failure of interventions to boost sales.

4. What impact has each product had on sales?

Simple plots were created in R script file to get a sense of the data. Some products were sold for several years, so the data has been aggregated by product.

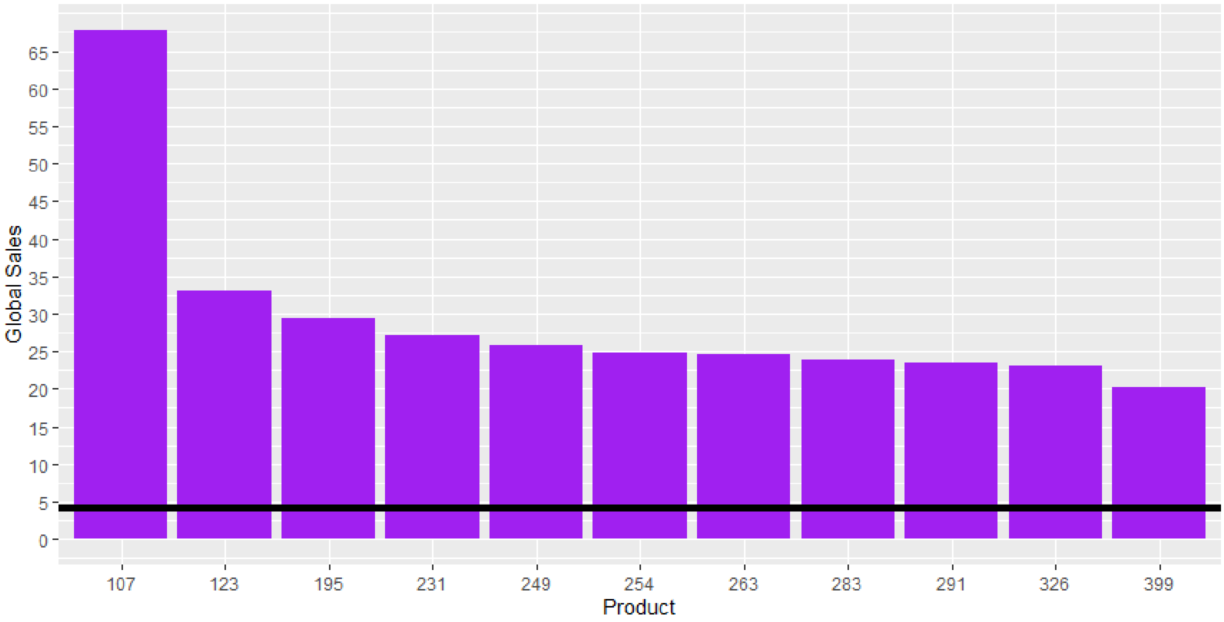
Global Sales



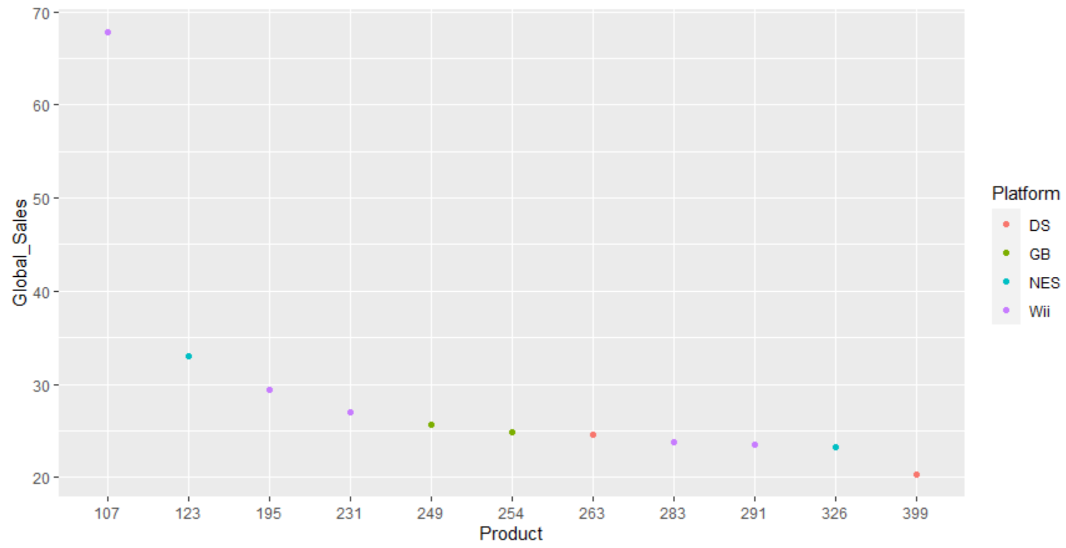
The above chart shows that most video games sell less than the median value of £4.32m. A small selection of games sell between £10m and £20m. The data contains an extreme outlier that sold over £60m.

An additional plot was created to explore the high selling products. The following products sold over £20m, make up 17% of global sales and 5 of the 11 products analyzed are on the Wii platform.

Highest Selling Products



Highest Selling Products with platform

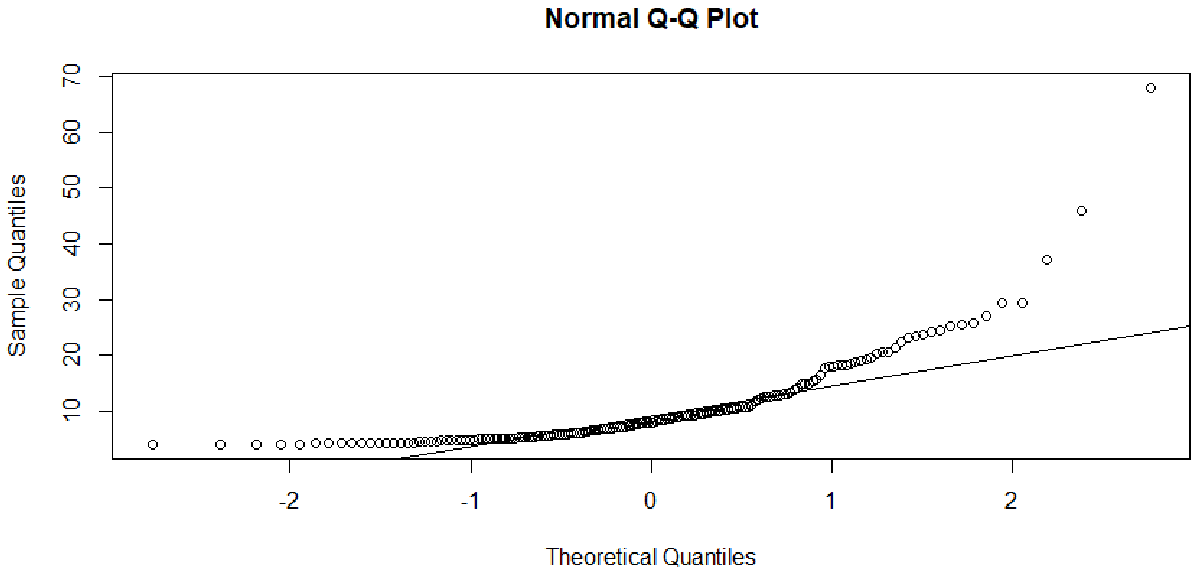


<u>Conclusion</u>

There is a lot of variation between products and further analysis is required to understand which products contribute most to sales, for example are there particular genres or platforms that perform well. In addition, it would be helpful to compare this to company profits and understand the relationship between the two, in order for the marketing department to target their efforts effectively.

5. How reliable is the data ?

EDA indicated that data is not normally distributed and this can be seen in the qunatile quantile plot below using global sales figures. This shows a poor correspondence between quantiles, as the data does not follow the line and large number of points are above and below the line:



To test the goodness of fit a Sharpiro-Wilks test (from the R library 'Moments') was performed on each of the sales figures to verify the shape of the data. Highly significant p values of much < 0.05 (for all 3 sales groups) confirm that data shows significant deviation from a normal distribution.

To assess the deviation and shape of the data, the skew and kurtosis was assessed using the standard R functions. All 3 sales columns show high positive skew and Leptokurtic or heavy tailed distribution. This indicates that that the variance is caused by a few extreme values that are more extreme than expected, which can be clearly seen on the chart with a number of points over the line above theoretical quantile 1. As the data shows such a strong skew and many data points outside of the line, the decision was taken not to try to normalise the data with a log approach. This would likely yield a result that still did not conform to normal distribution, instead, further testing and analysis will take the data shape into account.

Conclusion

The statistical tests support the chart evidence and show that the data is reliable but skewed by the presence of a small number of products with sales that are far above the average.

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

As the data is not normally distributed the Spearman's correlation test was selected to test for correlation between the sales. Spearman's correlation is suitable for monotonic data and is robust to outliers, which are present in this data set.

European sales, Global sales and North America sales showed highly significant and very strong positive correlation of 0.82 and 0.79 respectively (note 1 is equivalent to perfect correlation). This indicates that Global Sales will increase along with sales in North America and Europe.

North America sales and European sales shows highly significant positive correlation (0.53) and indicates sales in these regions will increase together. Note that the correlation is less strong and does reflect the value of the increase in either region.

In addition, linear and multi linear regression models were built to see if sales values in 1 or 2 regions could reliably predict sales in other regions given the strong correlations.

Linear models for Global Sales using either North American or European sales as the predictor showed limited success. Both indicated the presence of heteroscedasticity (most likely impure heteroscedasticity) meaning that the model is incorrectly specified, due to the absence of an important variable.

To solve the short comings of the linear models, a multi linear model was built. A quick correlation check with all available numerical variables indicated that only North America and European sales figures were worth including. Although product numbers appeared to show some negative correlation it is not a true numerical value and so cannot be used in the model.

The star ratings and p values for North American and European sales indicates that these are very significant variables that explain global sales. The R squared value indicates that the 2 variables explain 97% of the variation in Global Sales price. While the adjusted r-squared value also shows 97% indicating that the r-squared value has not been overly inflated by the addition of the second variable and still exceeds the value of either linear model. Model 4 is based on multiple linear regression and is the best fit to predict Global Sales compared to models 1 and 2.

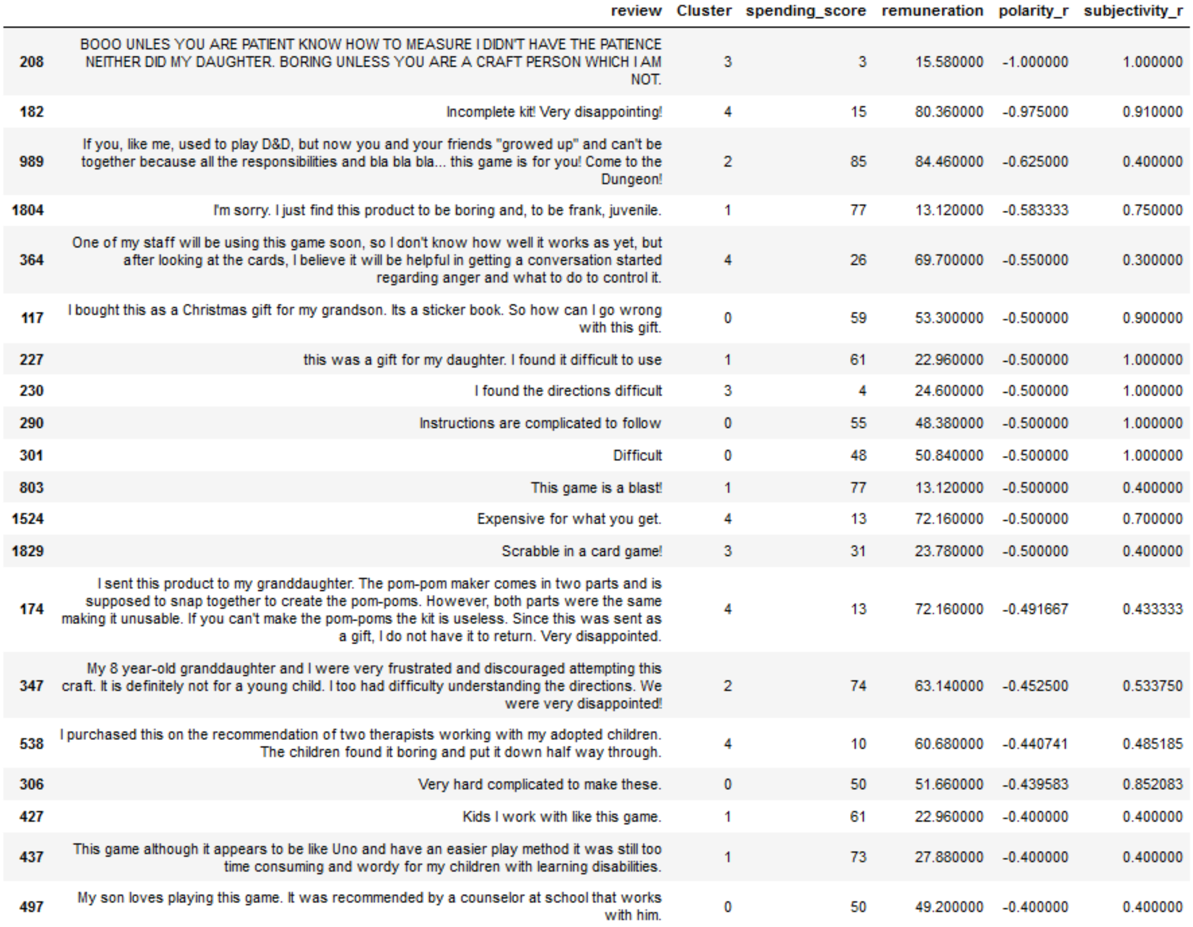
Conclusion

Based on the analysis undertaken so far it is possible to observe the following patterns:

- Customer income and spending score are good predictors of loyalty point acquisition.
- Customers group into 5 clear market segments by income and spending score and these groups can be further explored to provide additional information for marketing.
- Customer reviews show a positive sentiment overall and can be used in 2 valuable ways
    - To perform high level stock take if sentiment towards the company or particular products
    - To provide context to sentiment and identify areas for improvement or exploration.
- Individual video game products show a huge range of sales with some selling less than £1m and others selling more than £20m. It would be valuable to explore this further by genre, platform and country.
- The data is reliable, but heavily skewed by the presence of a small number of high selling games.
- North American and European sales show a positive correlation, but sales are smaller in Europe.
- Global sales can largely be predicted on the basis of sales in North American and Europe.

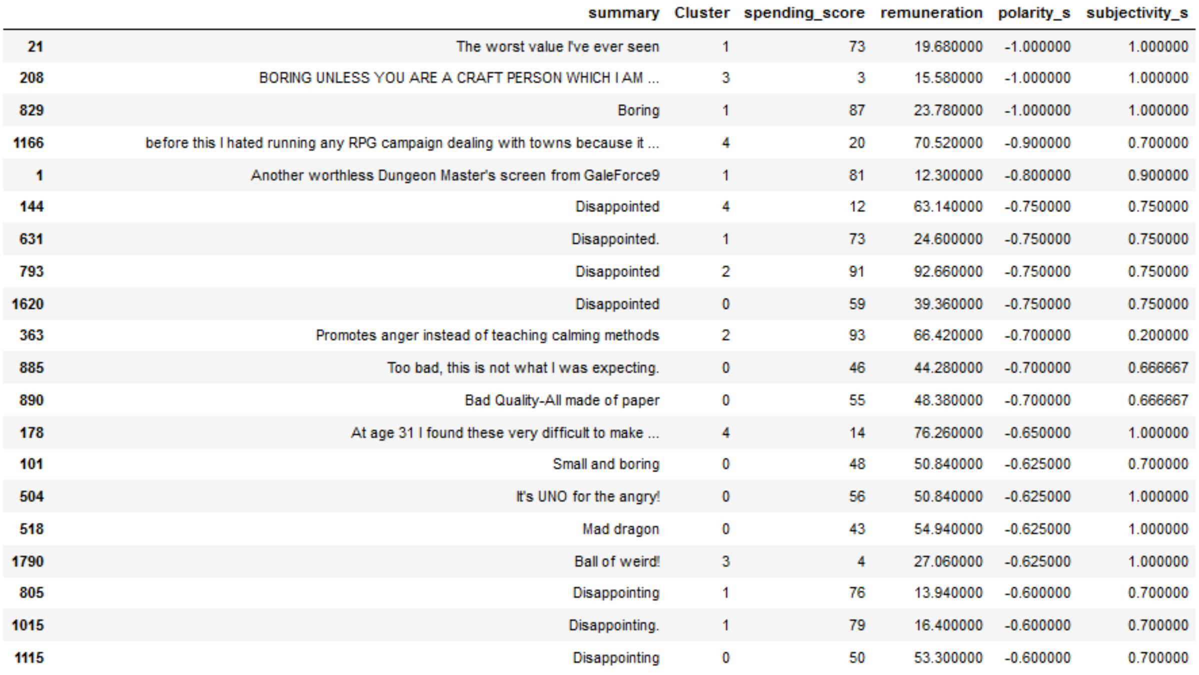
ANNEX 1 – Positive and negative review.

Negative review



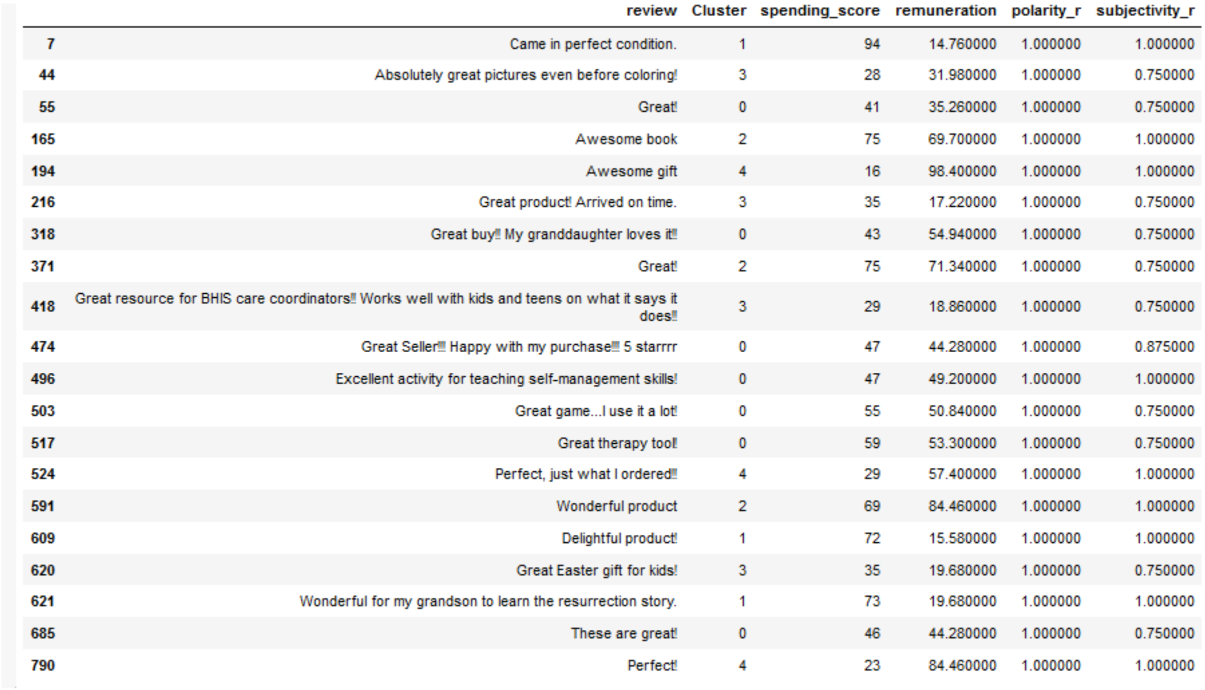
	review	Cluster	spending_score	remuneration	polarity_r	subjectivity_r
208	BOOO UNLES YOU ARE PATIENT KNOW HOW TO MEASURE I DIDN'T HAVE THE PATIENCE NEITHER DID MY DAUGHTER. BORING UNLESS YOU ARE A CRAFT PERSON WHICH I AM NOT.	3	3	15.580000	-1.000000	1.000000
182	Incomplete kit! Very disappointing!	4	15	80.360000	-0.975000	0.910000
989	If you, like me, used to play D&D, but now you and your friends "growed up" and can't be together because all the responsibilities and bla bla bla... this game is for you! Come to the Dungeon!	2	85	84.460000	-0.625000	0.400000
1804	I'm sorry. I just find this product to be boring and, to be frank, juvenile.	1	77	13.120000	-0.583333	0.750000
364	One of my staff will be using this game soon, so I don't 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.	4	26	69.700000	-0.550000	0.300000
117	I bought this as a Christmas gift for my grandson. Its a sticker book. So how can I go wrong with this gift.	0	59	53.300000	-0.500000	0.900000
227	this was a gift for my daughter. I found it difficult to use	1	61	22.960000	-0.500000	1.000000
230	I found the directions difficult	3	4	24.600000	-0.500000	1.000000
290	Instructions are complicated to follow	0	55	48.380000	-0.500000	1.000000
301	Difficult	0	48	50.840000	-0.500000	1.000000
803	This game is a blast!	1	77	13.120000	-0.500000	0.400000
1524	Expensive for what you get.	4	13	72.160000	-0.500000	0.700000
1829	Scrabble in a card game!	3	31	23.780000	-0.500000	0.400000
174	I sent this product to my granddaughter. The pom-pom maker comes in two parts and is supposed to snap together to create the pom-poms. However, both parts were the same making it unusable. If you can't make the pom-poms the kit is useless. Since this was sent as a gift, I do not have it to return. Very disappointed.	4	13	72.160000	-0.491667	0.433333
347	My 8 year-old 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!	2	74	63.140000	-0.452500	0.533750
538	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.	4	10	60.680000	-0.440741	0.485185
306	Very hard complicated to make these.	0	50	51.660000	-0.439583	0.852083
427	Kids I work with like this game.	1	61	22.960000	-0.400000	0.400000
437	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.	1	73	27.880000	-0.400000	0.400000
497	My son loves playing this game. It was recommended by a counselor at school that works with him.	0	50	49.200000	-0.400000	0.400000

Negative summary review



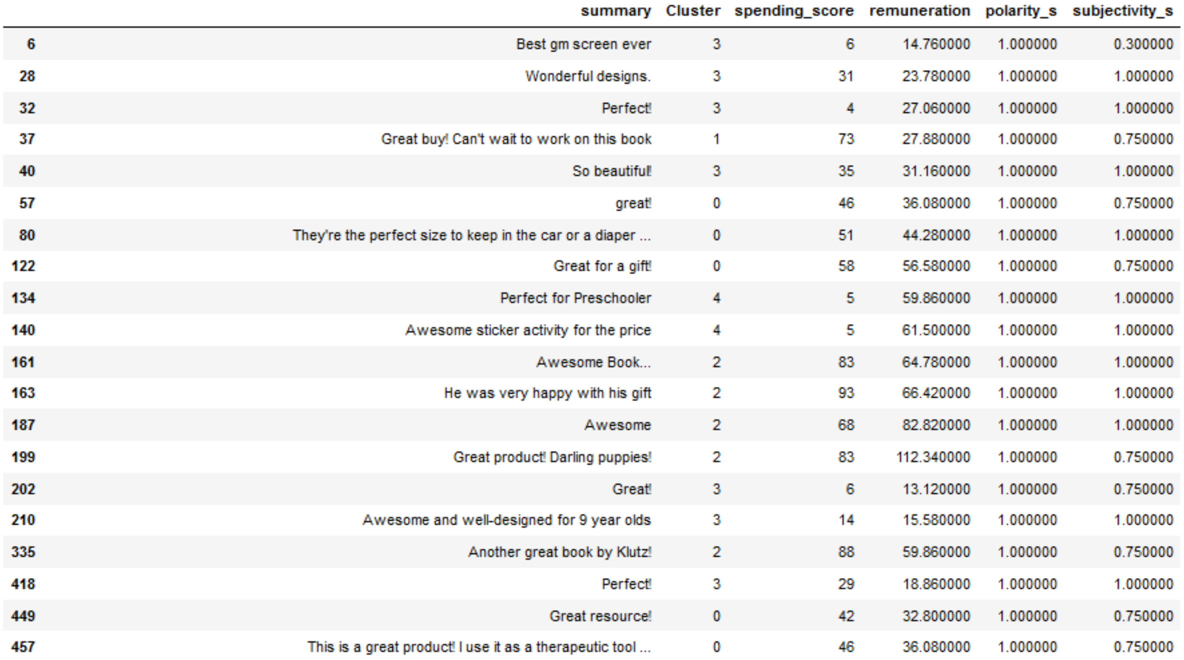
	summary	Cluster	spending_score	remuneration	polarity_s	subjectivity_s
21	The worst value I've ever seen	1	73	19.680000	-1.000000	1.000000
208	BORING UNLESS YOU ARE A CRAFT PERSON WHICH I AM ...	3	3	15.580000	-1.000000	1.000000
829	Boring	1	87	23.780000	-1.000000	1.000000
1166	before this I hated running any RPG campaign dealing with towns because it ...	4	20	70.520000	-0.900000	0.700000
1	Another worthless Dungeon Master's screen from GaleForce9	1	81	12.300000	-0.800000	0.900000
144	Disappointed	4	12	63.140000	-0.750000	0.750000
631	Disappointed.	1	73	24.600000	-0.750000	0.750000
793	Disappointed	2	91	92.660000	-0.750000	0.750000
1620	Disappointed	0	59	39.360000	-0.750000	0.750000
363	Promotes anger instead of teaching calming methods	2	93	66.420000	-0.700000	0.200000
885	Too bad, this is not what I was expecting.	0	46	44.280000	-0.700000	0.666667
890	Bad Quality-All made of paper	0	55	48.380000	-0.700000	0.666667
178	At age 31 I found these very difficult to make ...	4	14	76.260000	-0.650000	1.000000
101	Small and boring	0	48	50.840000	-0.625000	0.700000
504	It's UNO for the angry!	0	56	50.840000	-0.625000	1.000000
518	Mad dragon	0	43	54.940000	-0.625000	1.000000
1790	Ball of weird!	3	4	27.060000	-0.625000	1.000000
805	Disappointing	1	76	13.940000	-0.600000	0.700000
1015	Disappointing.	1	79	16.400000	-0.600000	0.700000
1115	Disappointing	0	50	53.300000	-0.600000	0.700000

Positive review



	review	Cluster	spending_score	remuneration	polarity_r	subjectivity_r
7	Came in perfect condition.	1	94	14.760000	1.000000	1.000000
44	Absolutely great pictures even before coloring!	3	28	31.980000	1.000000	0.750000
55	Great!	0	41	35.260000	1.000000	0.750000
165	Awesome book	2	75	69.700000	1.000000	1.000000
194	Awesome gift	4	16	98.400000	1.000000	1.000000
216	Great product! Arrived on time.	3	35	17.220000	1.000000	0.750000
318	Great buy!! My granddaughter loves it!!	0	43	54.940000	1.000000	0.750000
371	Great!	2	75	71.340000	1.000000	0.750000
418	Great resource for BHIS care coordinators!! Works well with kids and teens on what it says it does!!	3	29	18.860000	1.000000	0.750000
474	Great Seller!!! Happy with my purchase!!! 5 starrrr	0	47	44.280000	1.000000	0.875000
496	Excellent activity for teaching self-management skills!	0	47	49.200000	1.000000	1.000000
503	Great game...I use it a lot!	0	55	50.840000	1.000000	0.750000
517	Great therapy tool!	0	59	53.300000	1.000000	0.750000
524	Perfect, just what I ordered!!	4	29	57.400000	1.000000	1.000000
591	Wonderful product	2	69	84.460000	1.000000	1.000000
609	Delightful product!	1	72	15.580000	1.000000	1.000000
620	Great Easter gift for kids!	3	35	19.680000	1.000000	0.750000
621	Wonderful for my grandson to learn the resurrection story.	1	73	19.680000	1.000000	1.000000
685	These are great!	0	46	44.280000	1.000000	0.750000
790	Perfect!	4	23	84.460000	1.000000	1.000000

Positive summary review



	summary	Cluster	spending_score	remuneration	polarity_s	subjectivity_s
6	Best gm screen ever	3	6	14.760000	1.000000	0.300000
28	Wonderful designs.	3	31	23.780000	1.000000	1.000000
32	Perfect!	3	4	27.060000	1.000000	1.000000
37	Great buy! Can't wait to work on this book	1	73	27.880000	1.000000	0.750000
40	So beautiful!	3	35	31.160000	1.000000	1.000000
57	great!	0	46	36.080000	1.000000	0.750000
80	They're the perfect size to keep in the car or a diaper ...	0	51	44.280000	1.000000	1.000000
122	Great for a gift!	0	58	56.580000	1.000000	0.750000
134	Perfect for Preschooler	4	5	59.860000	1.000000	1.000000
140	Awesome sticker activity for the price	4	5	61.500000	1.000000	1.000000
161	Awesome Book...	2	83	64.780000	1.000000	1.000000
163	He was very happy with his gift	2	93	66.420000	1.000000	1.000000
187	Awesome	2	68	82.820000	1.000000	1.000000
199	Great product! Darling puppies!	2	83	112.340000	1.000000	0.750000
202	Great!	3	6	13.120000	1.000000	0.750000
210	Awesome and well-designed for 9 year olds	3	14	15.580000	1.000000	1.000000
335	Another great book by Klutz!	2	88	59.860000	1.000000	0.750000
418	Perfect!	3	29	18.860000	1.000000	1.000000
449	Great resource!	0	42	32.800000	1.000000	0.750000
457	This is a great product! I use it as a therapeutic tool ...	0	46	36.080000	1.000000	0.750000