PREDICTING FUTURE OUTCOMES

An analysis conducted for
Turtle Games

Analyst
Michelle Halpin
9 September 2022

LS3 Course 3
Advance Analytics for Organisational Impact

Report Contents

Predicting Future Outcomes: Turtle Games Analysis

1. Analysis Background

1.1 Context

Turtle Games is a manufacturer and retailer with a global customer base. They manufacture and sell their own products as well as sourcing and selling products manufactured by other companies. Their product range includes books, board games, video games and toys.

1.2 Analysis Brief

The Marketing and Sales departments wish to improve overall sales performance by better understanding customer trends. Customer sales and review data has been provided to inform this analysis which explore 5 set business questions:

1. how customers accumulate loyalty points
2. how customer groups can be used to target specific market segments
3. how customer review social data can inform marketing campaigns
4. what is the impact each product has on sales
5. how reliable is data in terms of distribution, skewness and kurtosis
6. what relationships exists between different sales regions

2. Analysis Approach

2.1 Approach

As directed by Turtle Games, this analysis uses Python and R.

- questions 1 to 3 were analysed using Python
- questions 4 to 6 were analysed using R

At each stage of analysis and regardless of tool used, data was imported from source csv files, cleaned and shaped into appropriate formats for analysis and visualisation. These actions included checking for missing data, assessing data types, identification of possible outlier data points and subsetting data to relevant variable selection.

2.2. Python and R library and package selection

Questions 1 to 3 Python libraries selected and installed:

- **numpy and pandas** were used to facilitate data wrangling activity
- **statsmodel** enabled linear regression
- **matplotlib and seaborn** enabled plot and chart creation
- **sklearn and scipy** enabled k-means clustering analysis
- **nlk and os** enabled sentiment analysis

Questions 4 to 6 R packages selected and installed:

- tidyverse and reshape2 to clean and wrangle data
- ggplot2 to create plots and charts
- skimr and DataExplorer to create statistical overview reports
- moments and BSDA to assess data distribution normality
- forecast to perform regression and make predictions

Observations and findings were documented at relevant stages throughout analysis and can be accessed within the Jupyter Notebook and R Script.

3. Observations: Factors affecting customer loyalty point accumulation

Methodology: linear regression within Python

3.1 Exploratory analysis observations

Initial analysis and visualisation was conducted to gain an understanding of the customer demographic. This identified the following age, gender and salary demographics:

Age: Average: 40 / minimum: 17 / maximum: 72
Most customers are aged between 29 and 40

Gender: Similar number of female (56% / 1120) and male (44% / 880)
Similar salary and age ranges for female and male
Similar loyalty point accumulation patterns for female and male

Salary: Average: £48K / minimum: £12.3K / maximum: £112.3K
Most salaries are between £30.3K and £64K

Education: There are 5 types of education attainment level:
PhD, Postgraudate, graduate, Diploma, Basic

All education types apart from 'basic' have similar loyalty point distribution patterns which range from 500 to 2000.

'Basic' differs from other 4 types as it has a much broader and higher majority distribution (1500 to 4000).

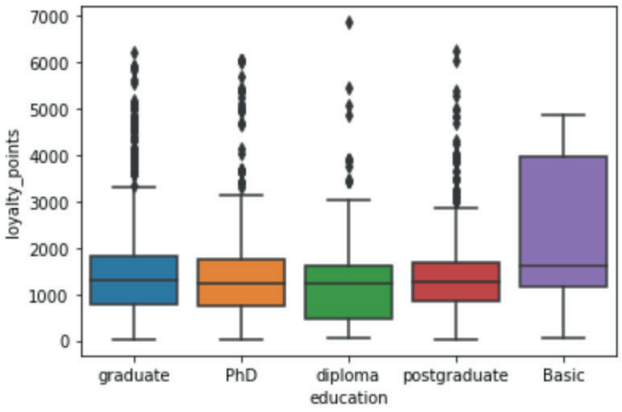

Fig 1.
Education attainment and loyalty point accumulation

Spend Score: Average: 50 / minimum: 1 / maximum: 99
Most scores are between 32 and 73

Loyalty point distribution

The histogram visualisation shows that lower point accumulation is more common that higher point accumulation; most counts of point can be seen to be between 0 and 2000.

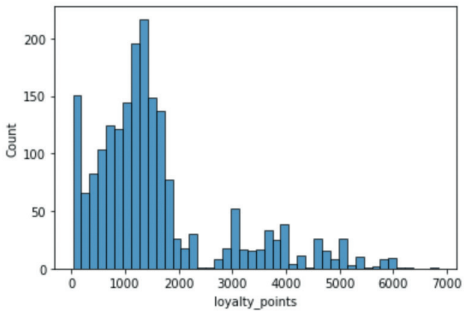

Fig 2.
Loyalty point distribution

3.2 Linear regression observations

- Spending score shows significant correlation to loyalty points
- Salary also shows correlation
- Age however does not appear to correlate with loyalty points

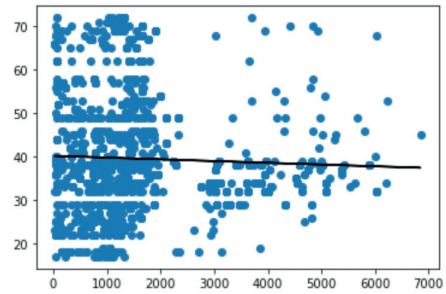

Fig 3.
Loyalty points and age regression

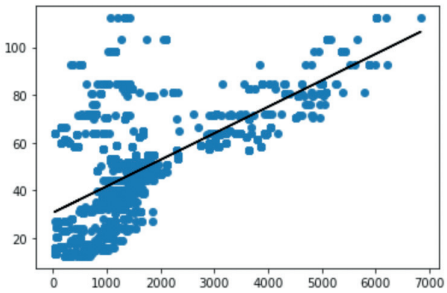

Fig 4.
Loyalty points and salary regression

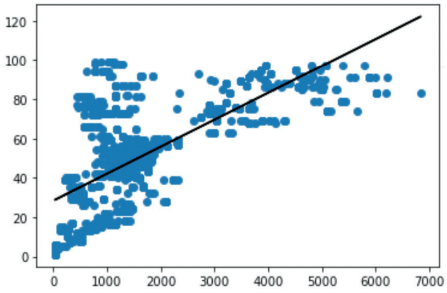

Fig 5.
Loyalty points and spend score regression

Analysis shows that higher salaries and higher spend score relate to increasing loyalty point accumulations. Detailed regression score outcomes justifying observations can be viewed in appendix 1.

4. Observations:
Customer clusters for market segmentation

Methodology: k-means clustering within Python

4.1 Exploratory analysis observations

Analysis indicates salary and spend score are significant factors affecting loyalty point accumulation. To better understand these factors k-means clustering was used to group customers using these variables.

As part of this approach exploratory scatter and pair plots highlighted the following:

- similar distributions of male and female salaries and spend score
- differing distributions of education attainment levels for spend score and salary; graduates enjoy higher salaries and exhibit higher spend score. Salary and spend score levels then decrease as educational levels decrease.

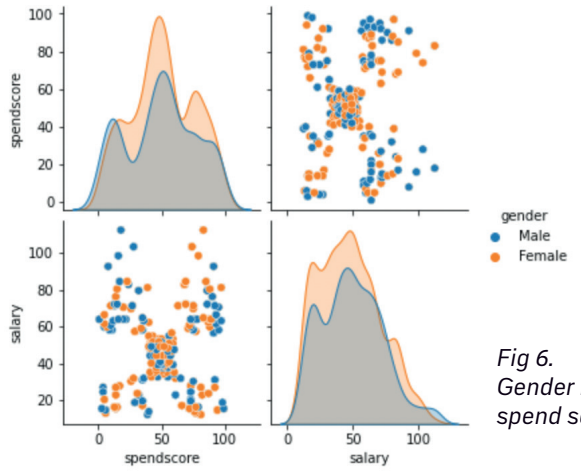

Fig 6.
Gender in relation to spend score and salary

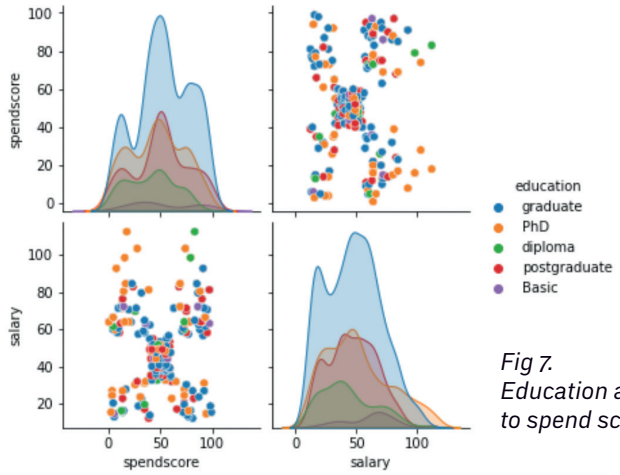

Fig 7.
Education attainment in relation to spend score and salary

4.2 K-means clustering observations

Model testing using Elbow and Silhouette methods. This indicated that 5 is the optimum number of customer clusters.

The clustering model also showed that customers earning between 35K and 55K are the largest cluster of customers showing most similar attributes. These customers have a spendscore of between 40 and 60.

Two other customer clusters both earn 60K to 120K; one cluster (blue colour) has a lower spend score of 20 to 40 when compared to the other cluster (black colour) which has a higher spend score of 60 to 100. Both clusters appear similar in size and appear to be the second largest.

The remaining two customer clusters appear similar in size but smaller than the other three. Both earn lower salaries of up to 40K; one cluster (green colour) has a lower spend score of 20 to 40 when compared to the other cluster (orange colour) which has a higher spend score of 60 to 100.

It can also be observed that consistently close alignment with cluster centroid can only be observed within the red cluster i.e. those earning between 35K and 55K with a spend score of between 40 and 60. All other clusters exhibit a broader, less consistently aligned spread from their centroids.

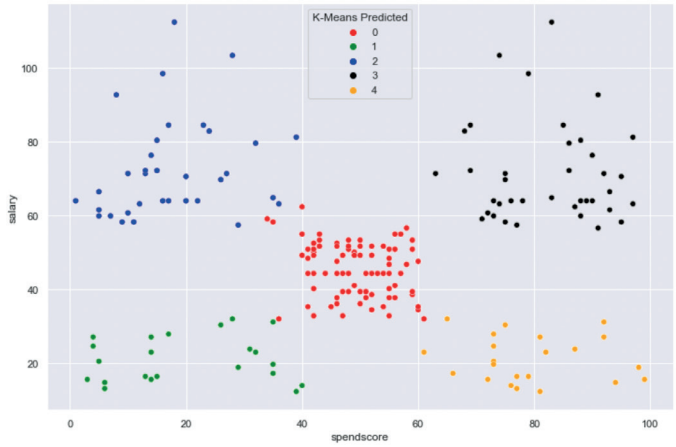

Fig 8.
k-means clustering outcomes for spend score and salary

Full rationale for this choice can be found within appendix 2.

5. Observations: Social review data to inform marketing campaigns

Methodolgy: Natural Language Processing (NLP) in Python

Appendix 5 details NLP analysis steps

5.1 NLP Analysis Approach

In order to better understand how online customer review and summary data can shape marketing campaigns, analysis used NLP to identify the:

- 15 most common words used in customer reviews
- 20 most positive customer reviews
- 20 most negative customer reviews

5.2 Most common words

With 1572 mentions, analysis shows that 'Game' is the most popular word customer use in product reviews and summaries. Other most popular words include: fun, one, great, play, like. The top 15 most common words are:

Word	Frequency
game	1572
fun	504
one	482
great	467
play	453
like	420
really	301
get	300
tiles	296
book	291
cards	283
love	280
would	265
good	262
new	253

Table 1.
15 most popular words used in customer reviews and summaries

Wordclouds for customer reviews and summaries were also created to gain further insight into how customers think about Turtle Game products.

Fig 9.
Combined customer review and summary wordcloud

Fig 10.
Customer review wordcloud

Fig 11.
Customer summary wordcloud

5.3 Most positive reviews

Most positive review make reference to toys for children as well as creative / therapeutic products for adults. All top 20 comments look to have been correctly categorised as positive.

Review	Polarity	Subjectivity
perfect	1.00	1.00
my daughter loves her stickers awesome seller thank you	1.00	1.00
perfect for tutoring my grandson in spelling	1.00	1.00
the best part i see is the box what a wonderfully diverse and rounded set for the cost i am so happy and as the dm you know that if i am happy my players are happy	0.88	0.86
great quality very cute and perfect for my toddler	0.82	0.92
the pictures are great ive done one and gave it to a friend of mine who likes dragons	0.80	0.75
great seller happy with my purchase 5 starrrr	0.80	0.86
great easter gift for kids	0.80	0.75
these are great	0.80	0.75
bought this because i wanted it all these dd games are great	0.80	0.75
husband seems happy with it	0.80	1.00
great accessory to use with the playing mat	0.80	0.75
great price arrived on time with no damage will be a great addition to my collection	0.80	0.75
this is a great accessory to the starter set i would recommend this to anyone who owns the starter set	0.80	0.75
my granddaughter loves these so happy to find peppa pig items for her	0.80	1.00
great doll to go with the book animals cant wait to read book with the doll to the grandkids	0.80	0.75
a great creation tool it helps me concentrate	0.80	0.75
prompt service and a great product	0.80	0.75
this is a great tool to have at hand when playing quiddler	0.80	0.75
this is a great product i use it as a therapeutic tool and it has been very effective	0.79	0.86

Table 2.
20 most positive customer reviews

5.4 Most negative reviews

Analysis of the most negative reviews shows that customers are concerned with the quality of product construction and promoted product age levels. However, analysis also shows limitations of the NLP model in terms of accurate polarity categorisation. Almost half of the most negative reviews do not in fact appear to be negative. A reason for this may be that some products are intended to support therapeutic activities. Customers include the word ‘anger’ within reviews when describing these activities in a positive way.

The 20 most positive and negative customer summaries can viewed in appendices 3 and 4.

Review	Polarity	Subjectivity
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	-1.0	1.00
incomplete kit very disappointing	-0.78	0.91
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	-0.55	0.30
i bought this as a christmas gift for my grandson its a sticker book so how can i go wrong with this gift	-0.50	0.90
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	-0.49	0.43
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	-0.44	0.53
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	-0.44	0.48
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	-0.40	0.40
my son loves playing this game it was recommended by a counselor at school that works with him	-0.40	0.40
if you like me used to play dd but now you and your friends growed up and cant be together because all the responsibilities and bla bla bla this game is for you come to the dungeon	-0.40	0.40
you can play the expansions one at a time or add then both in for a longer game if your into lords of waterdeep this is a must have	-0.40	0.40
if you play dungeons and dragons then you will find this board game to be dumb and boring stick with the real thing	-0.39	0.55
i was a bit disappointed in the quality of the cardboard pieceholders and the fact that they changed the names of some hotels otherwise i mean its a terrific game	-0.36	0.71
very fun game to use with kids working on handling anger you play like uno but have to answer questions about anger	-0.35	0.26
i really like this game it helps kids recognize anger and talk about difficult emotions	-0.35	0.45
i am a therapist for children and this game is so valuable to bring out insight and solutions to deal with and identify feelings of anger i use it frequently	-0.33	0.30
confusing instructions and its not for 6 year olds its boring too its asking the same question but each question is worded differently	-0.32	0.53
as my review of gf9s previous screens these were completely unnecessary and nearly useless skip them this is the definition of a waste of money	-0.32	0.32
the adventures are tough but you can get throuhg them it all comes down to the die roll just like any dd game	-0.31	0.51
a crappy cardboard ghost of the original hard to believe they did this but they did shame on hasbro disgusting	-0.30	0.76

Table 3.
20 most negative customer reviews

5.5 Sentiment averages and distribution

Review Comments:

- Average polarity score: 0.18
- Average subjectivity score 0.51

Most review comments are moderately positive and subjective.

Summary Comments:

Average polarity score: 0.27
Average subjectivity score 0.48

Most summary comments are moderately positive and subjective.

When viewing distributions however differences are observed between review and summary comments; review comments appear more normally distributed that summary comments. Summary sentiment is mostly neutral to positive but summary subjectivity does not appear to form a understandable trend or pattern.

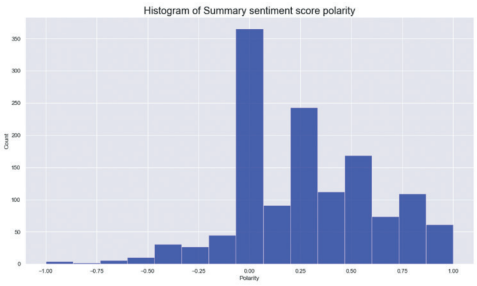

Fig 12. Sentiment Polarity: Summary

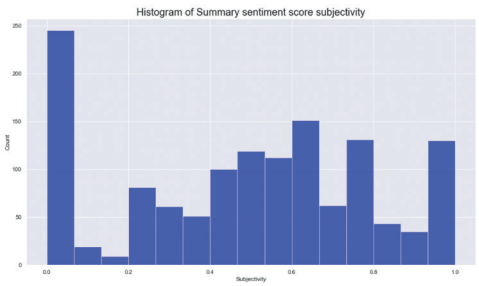

Fig 13. Sentiment Subjectivity: Summary

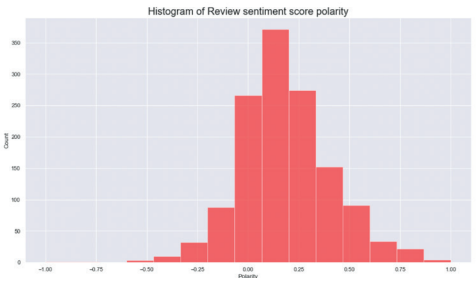

Fig 14. Sentiment Polarity: Review

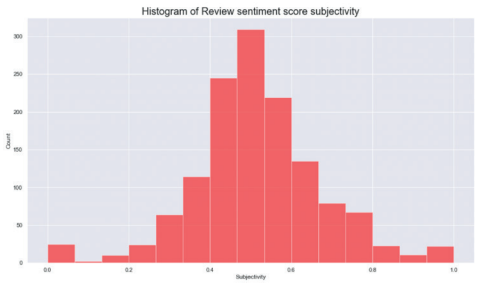

Fig 15. Sentiment Subjectivity: Review

6. Observations: Impact of product on sales

Methodolgy: Visualisation and analysis within R

6.1 Exploratory visualisation observations

In order to better understand the impact of product ID on sales, exploration analysis and visualisation was conducted with R.

Analysis shows that Turtle Games sales occur in three regions; Europe, North America and Global. Initial analysis via a range of plot approaches (scatter, histogram and box) indicate that sales decrease when plotted against increasing (0 to 9000) product ID.

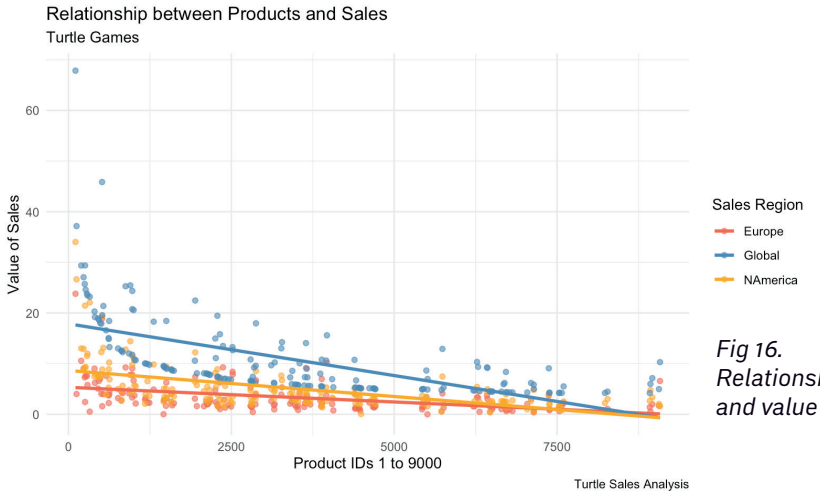

Fig 16.
Relationship of Product ID and value of sales by region

Plot visualisations also reveal the presence of potential data outliers within sales regions:

- Global sales: outlier identified at 60+ sales
- EU Sales: outlier identified at 20+ sales
- NA Sales: outlier identified at 30+ sales

Appendix 6 shows exploratory scatter plots for all sales regions.

Other visualisations indicate:

- The top 5 selling plaftorms are: X360, PSC, PC, Wii, DS
- Non-normal sales distribution observed in all regions
- Sales distribution is leptokutic (heavy tailed)
- Most sale values are less than £5

6.2 Product popularity observations

Further analysis identified the top 10 selling 'popular' products by region:

Products popular in all regions:
107, 515, 195, 876, 231 and 515

Products popular in only one region:

- Europe: 3967, 2371, 3645, 979, 399
- North America: 326, 535
- Global area: 249, 263

Products not popular in Europe:
123, 254, 948 are not popular in Europe.

6.3 Summary values of product sales

Average sales by region:
Europe: £3.36, NAmerica: £5.06, Global: £10.73

Minimum sales by region:
Europe: £0, NAmerica: £0.06, Global: £4.20

Maximum sales by region:
Europe: £23.80, NAmerica: £34.02, Global: £67.85

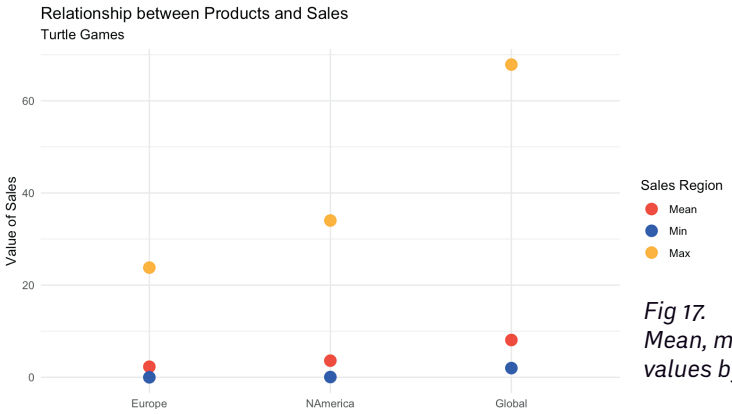

Fig 17.
Mean, max and min sales values by region

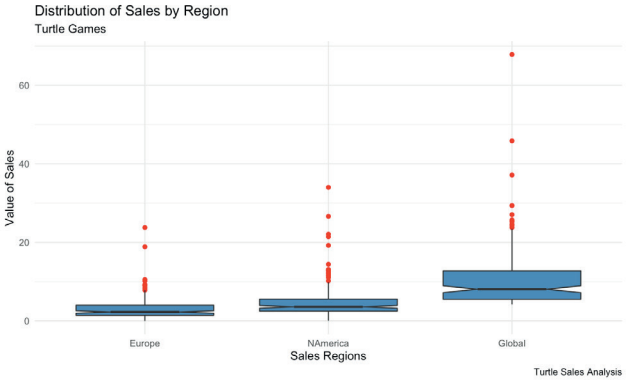

Fig 18.
Sales distributions by region

Visualisations suggest highest sales in the Global region with sales higher in North America than in Europe.

7. Observations: Data reliability in terms of distribution, skewness and kurtosis

Analysis Methodolgy: DataExplorer Report, QQ-plots, Shapiro-Wilk tests in R

7.1 QQ Plots

Quantile quantile (QQ) plots and Shapiro-Wilk tests were conducted to verify data reliability in terms of distribution, skew and kurtosis. These tests were necessary steps to ensure reliability of sales price predictive analysis.

The Data Profiling Report generated within R (Appendix 7) suggests issues exist with normality of data distribution.

The creation of QQ plots reveal 'S' shaped curved data points, right skewed 'fat' tailed data points which indicate non-normal distribution.

7.2 Shapiro-Wilk tests

The P-value for all three regions was less that 0.05 indicated non-normal data distribution. In addition, all regional data was highly skewed to the right with very high leptokurtic kurtosis also observed. Appendix 8 details the full Shapiro-Wilk test results and interpretations.

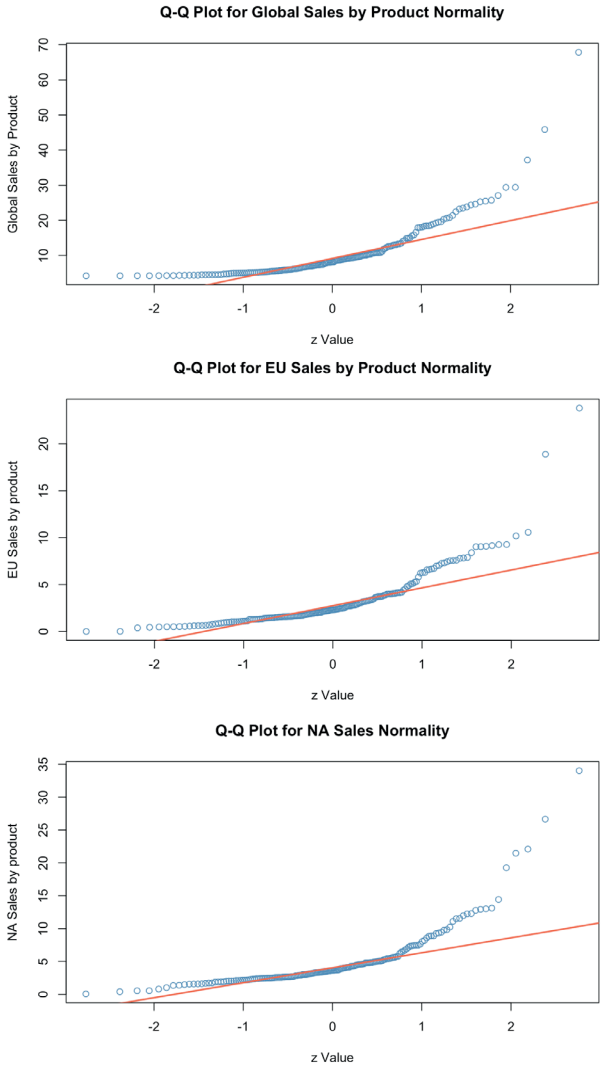

Fig 19.
QQ-plots of regional sale normality

8 Observations: Regional sales predictions

Analysis Methodolgy: linear and multiple regression in R

8.1 Data preparation

To enable prediction of product sale values within differing regions, data was tested and prepared using:

- correlation tests that determine R coefficient
- linear regression between product and sales region
- log transformations to strengthen predication model building

Detailed score results can be viewed in Appendix 9. Summary observations are as follows:

Positive, strong correlations can be observed between all regions with strength of correlation ranked as follows:

1. North America and Global - strongest correlation (R = 0.92)
2. Europe and Global - second strongest correlation (R = 0.86)
3. Europe and North America - third, strongest correlation (R = 0.85)

8.2 Linear regression: regional sales relationship with product

In all regions the P-value is less than 0.05 indicating regional sales are a significant factor. The probability of 't' is also less than 0.05 indicating that the regression model is a good fit. Appendix 10 shows full linear regression results before and after log transformations.

R-squared outcomes before and after log transformation:

R-square Europe

Before log transformation: explains 20.47% of variability
After log transformation: explains 23.81% of variability

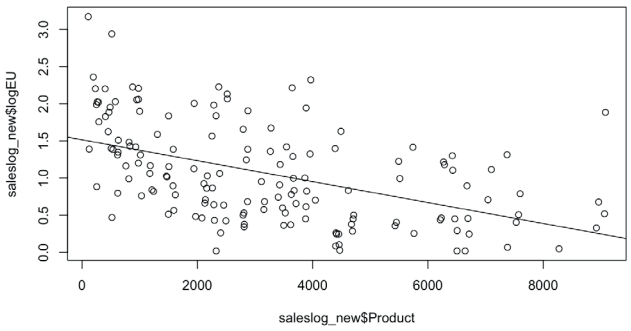

Fig 20.
Europe / Product sales regression

R-square North America

Before log transformation: explains 29.54% of variability
After log transformation: explains 39.1% of variability

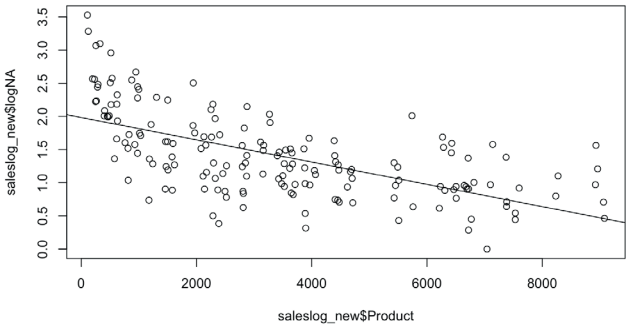

Fig 21.
North America / Product sales regression

R-square Global

Before log transformation: explains 36.74% of variability
After log transformation: explains 54.85% of variability

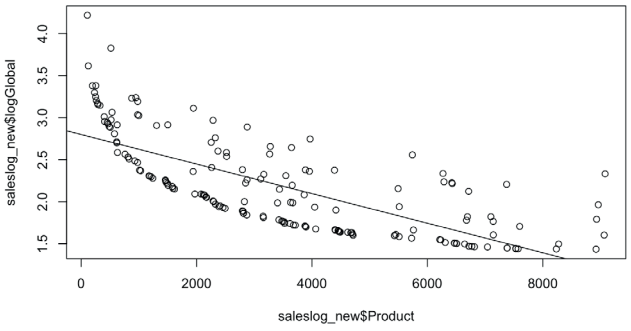

Fig 22.
Global region / product sales regression

At 54.85%, the global sales region appears most significant in terms of accounting for product sale outcomes.

Charts also show the presence of outliers and that significant 'noise' / residual dispersal exists within all three sales regions. The global sales region indicates least dispersal and strongest trend.

8.3 Multiple linear regression

To further explore and consolidate findings, multiple linear regression was applied to the sales and product data.

8.3.1 Product and sales region correlation

Testing between sales region and product ID showed stronger correlations exist between sales regions than exist between product and region:

Strongest correlations are between regions:

- Correlations in order from high to low (near to + 1 or -1):
- Global / North America: 0.92
- Global / Europe: 0.85
- North America / Europe: 0.62

Weaker correlations are between product and regions:

- Correlations in order from high to low (near to + 1 or -1):
- Product / Global: -0.61
- Product / North America: -0.54
- Product / Europe: -0.45

8.3.2 Model creation and testing

Three regression models (A, B and C) were created and tested on product and sales region. Model A proved to be the strongest:

Model A Product and Global: 97.09%
Model B Product and Europe: 77.97%
Model C Product and North America: 85.44%

The Model A adjusted R-squared is closest to 1 (100%). P-values and probability of 't' also indicate all models are a good fit. Full regression scores can be viewed in Appendix 11.

8.4 Sales predictions

Model A was used to make predictions on five sales within the global region. Results observed showed a mixed success rate; only three out of the 5 tests produced predicted sale values that fell within the lower and upper accuracy range. Therefore model prediction ability can be said to have a 60% success rate.

Full prediction results can be viewed in Appendix 12.

9 Analysis recommendations

9.1 Loyalty point accumulations

Customer gender and age do not appear to impact upon how customers collect loyalty points. However, analysis shows that higher salaries and higher spend score correlate with higher loyalty point levels. Therefore in order to achieve greater loyalty point accumulation, analysis recommends higher earning customers with higher spend scores are considered in marketing campaings. It should also be noted educational attainment levels may further inform customer identification. Customers with a basic level of educational attainment demonstrate broadest and highest levels of loyalty point accumulation.

9.2 Customer groupings for market segmentation

When considering which customer groupings to inform marketing segmentation, 3 groups can be recommended:

Group 1:

Customers earning £35K to £55K with 40 to 60 spend score
This is group represents the largest market segment. Although spend score and salary are not the highest, this largest volume of customer offers significant scale of sales opportunity.

Group 2:

Customers earning 60K to 120K with 60 to 100 spend score
Although not as large as group 1, this group should alse be considered for segmentation. It enjoys the highest salaries and has the highest spend scores.

Group 3:

Customers earning up to 40K with 60 to 100 spend score
This third group is the smallest of the 3 groups and has the lowest salaries. However, this analysis recommends that it be included in segmentation as it also has highest spend scores of 60 to 100.

In addition to these 3 groups, analysis advises that segmentation can be further informed by targetting customers based on:

- an equal focus on male and female customers
- those ages 29 to 49 (largest customer group)
- those earning the average £48K salary
- postgraduate educational attainment level

As it appears more postgraduates enjoy highest salary and spend scores, it is advised that this insight is also used to inform customer segmentation.

9.3 Social review data to inform marketing campaigns

Analysis of the most common words used within social reviews shows a customer inclination towards fun, happy, friendly and family orientated interactions with books, cards, games and crafting sets. When analysing the most positive reviews, two types of interaction can be observed; grandparents buying products for young grandchildren and adults

using therapy related game products. Therefore this analysis can advise that future marketing campaigns should target these two customer demographics. Also, a lighthearted, fun and family-led approach should inform the tone and brand of these campaigns.

When considering the negative reviews customers provided, valuable insights can be observed; customers are concerned with the quality of product construction and promoted product age levels. Analysis advised further investigation takes place to establish solutions to remedy game construction quality and provide age appropriate guidelines for products.

A final recommendation for use of social data to inform campaigns relates to NLP model limitations discovered during analysis. This recommendation is that the model be futher refined to exclude words such as 'anger' as these can be used by customers when providing postivie feedback that relates to therapy related games and products.

9.4 Impact of product on regional sales

Analysis shows highest sale values and volumes occur within the global sales region and lowest sale values and volumes occur within Europe. Turtle Games may wish to consider the sales opportunities offered within this largest market when planning stock distribution and marketing campaigns.

This analysis identifes which products are popular within which region. Although products 107, 515, 195, 876, 231 and 515 are popular in all regions, this analysis recommends additional product promotion focus on products 249 and 263 which are only popular with the most profitable global region.

9.5 Data reliability

Analysis demonstrated that the data supplied is not normally distributed which may impact upon predictive modelling outcomes. Analysis identified outlier data points within each of the 3 sales regions. Therefore a recommendation of this analysis is further investigation to discover the reason for these outliers with a view to excluding them from future analysis.

9.6 Predicting sales

Predictive modelling and testing showed strong correlations between sales regions. The strongest relationship was observed within the global region and the weakest within the Europe region. This finding further strengthens the recommendation for Turtle Games to prioritise sales within this region.

However, when modelling was used to predict sales values, only 3 out of the 5 predictions fell within an acceptable accuracy range. Therefore this analysis recommends caution when using this analysis model to predict future sales values within the global sales region.

This analysis also recommended that factors such as the non-normal data distribution and the existence of potential outliers be futher investigated before modelling is used to predict sales values.