LSE DATA ANALYTICS CAREER ACCELERATOR COURSE 3 REPORT

BY: KEVIN WIJAYA OEY

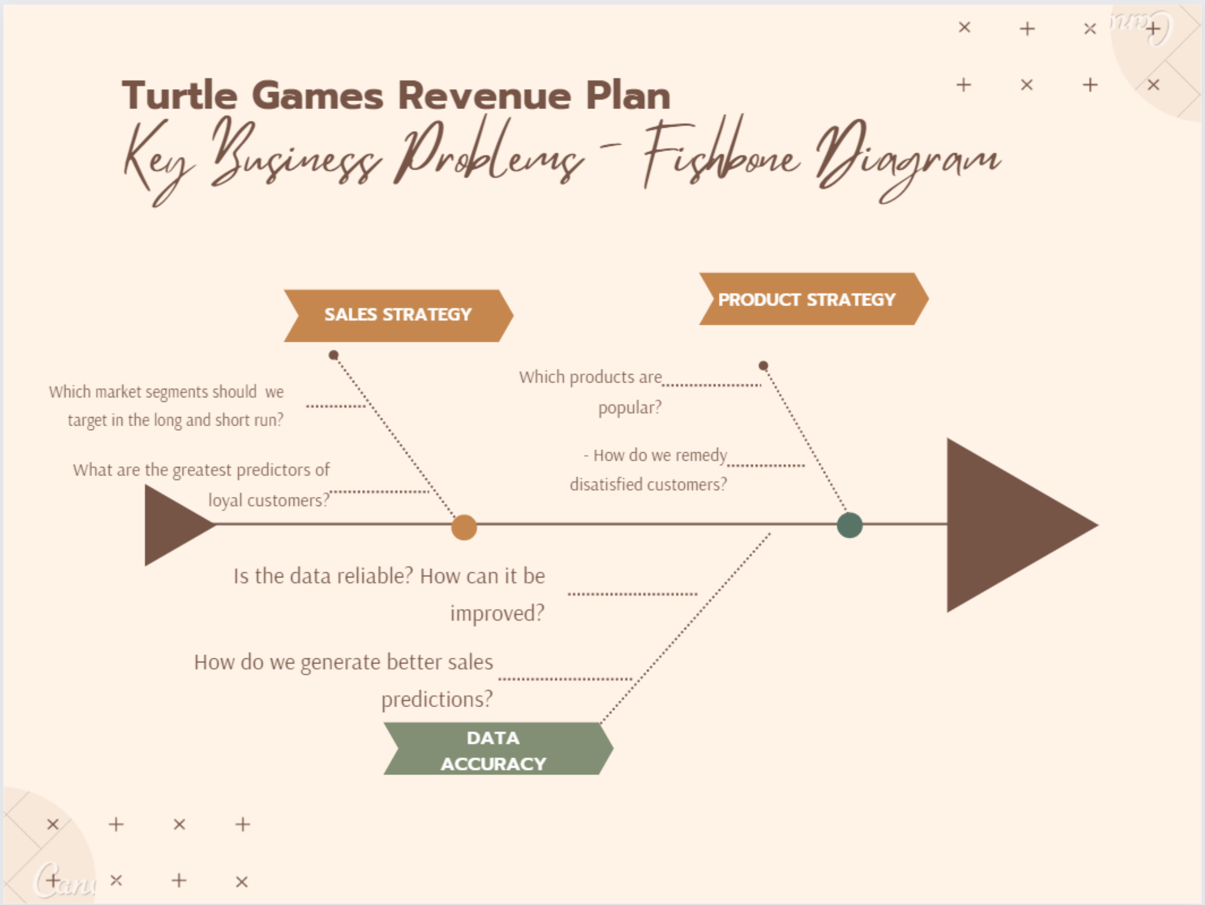
NUMBER OF WORDS: 1200 WORDS

(excluding citations)

BACKGROUND

Turtle Games sells in-house and externally sourced games, toys and books globally. They intend to maximise sales revenue by analysing social data and understanding customers to optimise strategies for:

- **Product Categorisation**, to find popular products and remedy customer complaints from bad products.
- **Optimal sales and incentivisation strategies**, by:
    - identifying strong predictors of loyalty points
    - Targeting lucrative market segments in the long and short run.
- **Evaluating data reliability and correlation** to gather more accurate data and create better predictions.



Turtle Games Revenue Plan

Key Business Problems - Fishbone Diagram

SALES STRATEGY

Which market segments should we target in the long and short run?

What are the greatest predictors of loyal customers?

PRODUCT STRATEGY

Which products are popular?

- How do we remedy disatisfied customers?

Is the data reliable? How can it be improved?

How do we generate better sales predictions?

DATA ACCURACY

ANALYTICAL APPROACH

DATA WRANGLING

- Imported necessary libraries for calculation (pandas, matplotlib, math, tidyverse, moments), visualisation (seaborn), NLP (nltk, corpus)
- Removed Null values and duplicates in review data
- Subsetted review and sales data to retain relevant columns.
- Renamed ambiguous columns.
- Renamed cleaned datasets to distinguish them from prior versions.
- For Natural Language Processing, the review and sales text were:
    - Subsetted
    - changed to lowercase
    - Stripped of punctuation, duplicates and stopwords
    - Applied with reset.index( ), otherwise tokenized words cannot be appended into lists.
- Numeric text (eg Product IDs) were converted to factors or characters as they aren't continuous by nature.
- Grouped Platform by Console Family (eg PS1,PS2 are all Playstation consoles) and Console Type (Handheld, Home, or dedicated) to simplify segments in visualisations
- Grouped total sales by product and converted the dataframe to "long form" and used left join to add relevant categories (eg Console_Family). The subsequent data could be used to create facetplots.

DATA ANALYSIS

- Fitted various sales data variables (age,expenditure_rank,salary) against loyalty points as linear regression models to determine which variable can predict accumulation of loyalty points.
- Utilised k-means clustering justified by the elbow method, silhouette method and yellowbrick to determine critical customer segments pivotal to maximising revenue.
- Tokenised cleaned reviews into lists and strings, generated wordclouds to determine (in broad terms) popular words. Polarity and subjectivity scores were generated and visualised to generate frequency distributions to indicate the distribution of customer sentiment.
- Created charts faceted by Console_family and segmented by sales region (North America, Europe, Global) to determine how products influence sales across console families.
- Determined data reliability by testing for normality via kurtosis,skewness, shapiro-wilk tests and qqplots. Multiple tests were used to override conflicting results and justify normality.
- Tested a multiple regression model to predict Global Sales based on North American and European sales.

VISUALISATIONS AND INSIGHTS

RATIONALE

- Linear regression utilised scatterplots as:
    - variables used were continuous
    - It shows a broad overview, identifies outliers, clusters
    - A line-of-best-fit can be added to show overall data trajectory.

- Multivariate faceted scatterplots provide a broad overview of multivariate relationships across different categories, expediting the identification of identify similar / dissimilar trends and outliers across console families before drilling down.

- Wordclouds and Histograms were used for textual analysis as they emphasise the most prominent words or sentiments easily, while understanding other distributions.

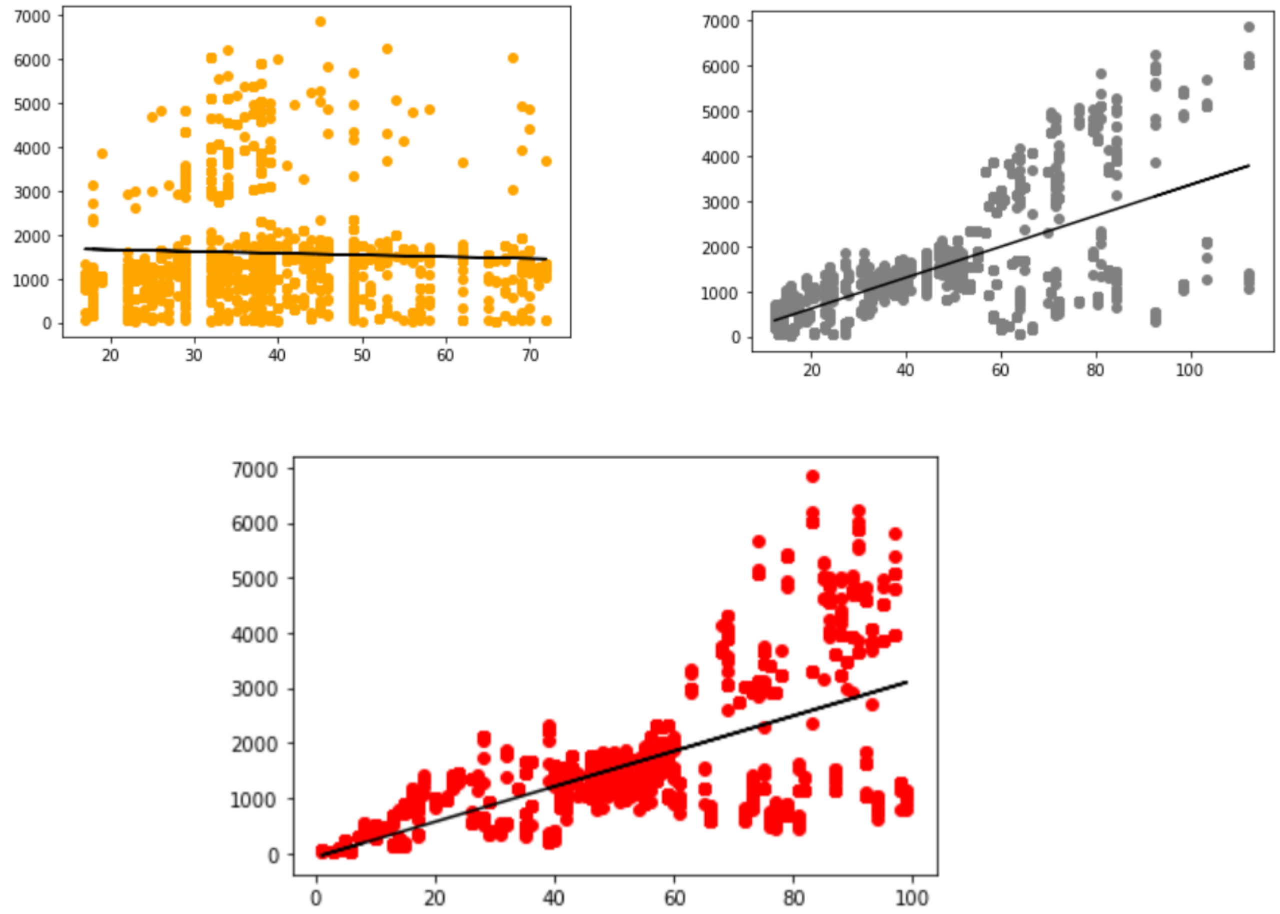
- Pairplots were used in k-means clustering to show bivariate relationships between salary and expenditure, and to visualise ideal k-means predicted clusters.

- Boxplots were used to compare medians across different console families and pinpoint outliers.

- Where possible visualisations have been annotated with titles, subtitles, legends, axes and colourblind friendly palettes to improve accessibility.

RESULTS

LINEAR REGRESSION (loyalty points vs age, salary and expenditure - clockwise)



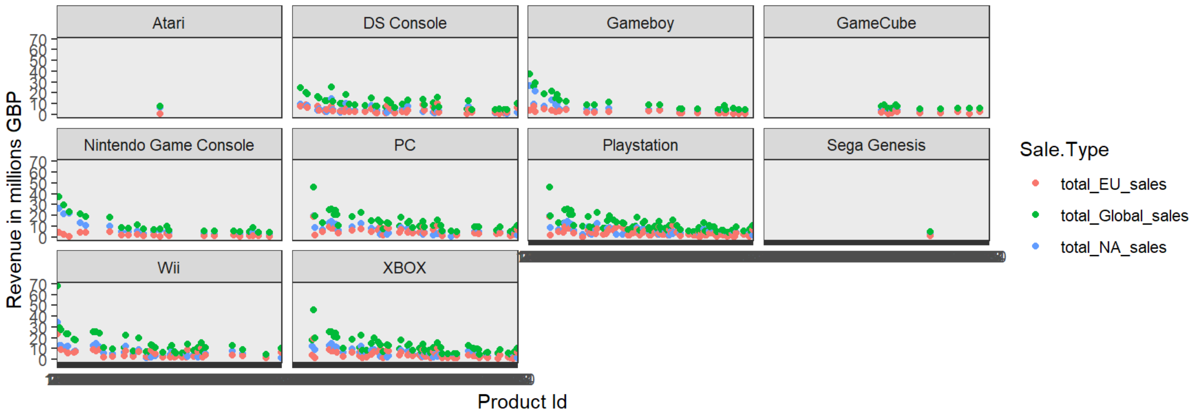
Loyalty points correlate:

- Negatively with age
- Positively with salary and expenditure rank

FACETPLOTS

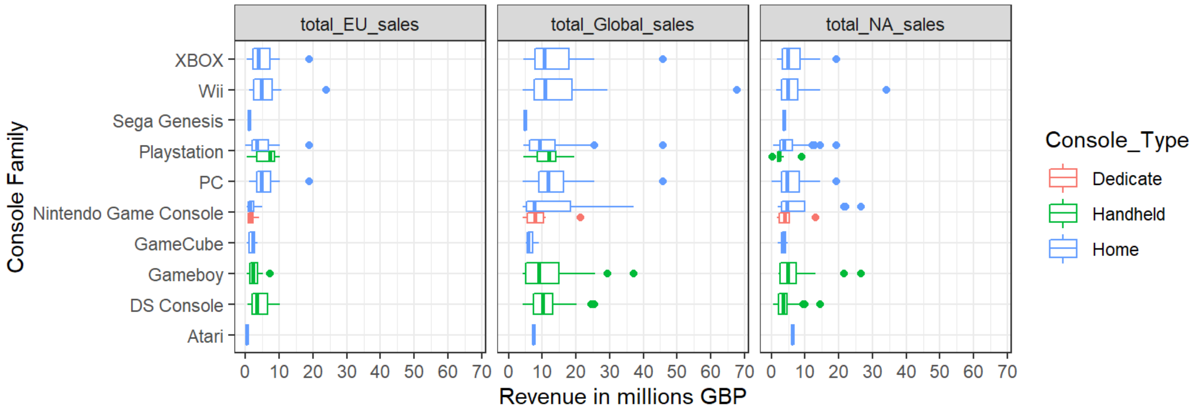
Breakdown of Regional Sales by Console Family

Source: turtle_sales.csv



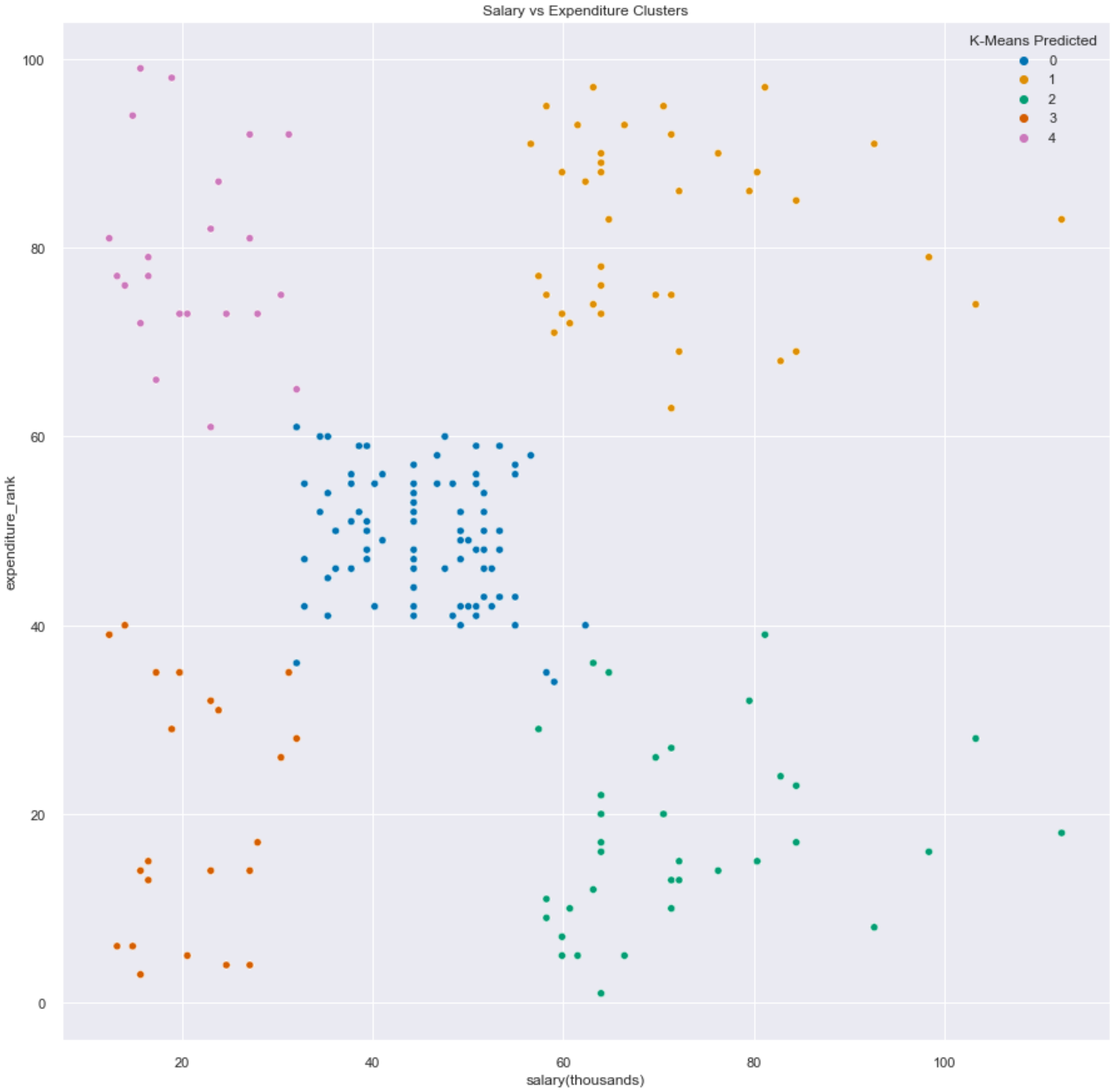
Boxplot of Sales by Region segmented Console Family & Console Type

Source: turtle_sales.csv



- Low Product IDs from North American, European and Global sales generate very high sales (and vice versa) across different console_families.
- DS Consoles, Gameboys, Playstation, PC Games, Nintendo Game Consoles, Wii and XBOX games generate the highest revenue. Outliers are also consistent across regions in these three consoles.
- Turtle Games should stop selling Gamecube, Atari and Sega Genesis consoles as they are not popular.

K-MEANS CLUSTERING



Assumptions:

- highest revenue comes from the densest clusters
- Expenditure is relative (not 1:1) to salary

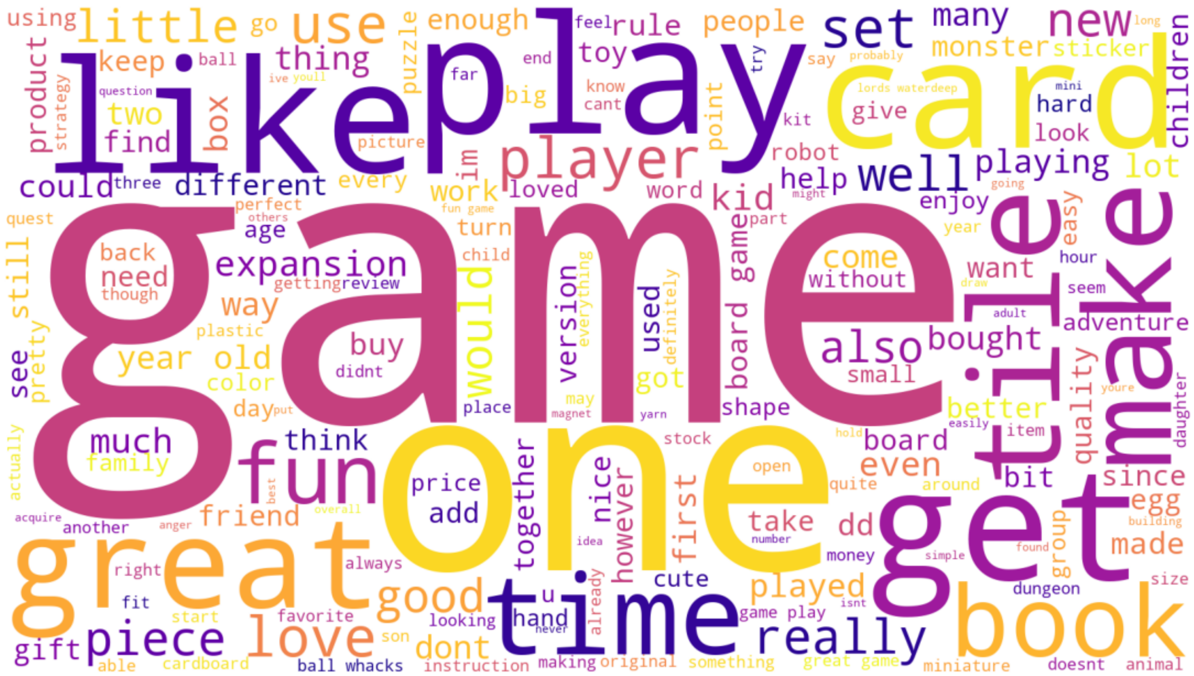
The k-means clustering clearly shows 5 customer segments:

- Cluster 0, the highest density cluster with moderate salary and expenditure rank.
- Cluster 1, high salary and high expenditure.
- Cluster 2, high salary and low expenditure.
- Cluster 3, low salary and low expenditure.
- Cluster 4, low salary and high expenditure.

In the next 1-6 months, Turtle Games should allocate their marketing budget to all clusters. In the long run (over 6 months), they should allocate the most marketing budget targeting Clusters 0 and 1, as:

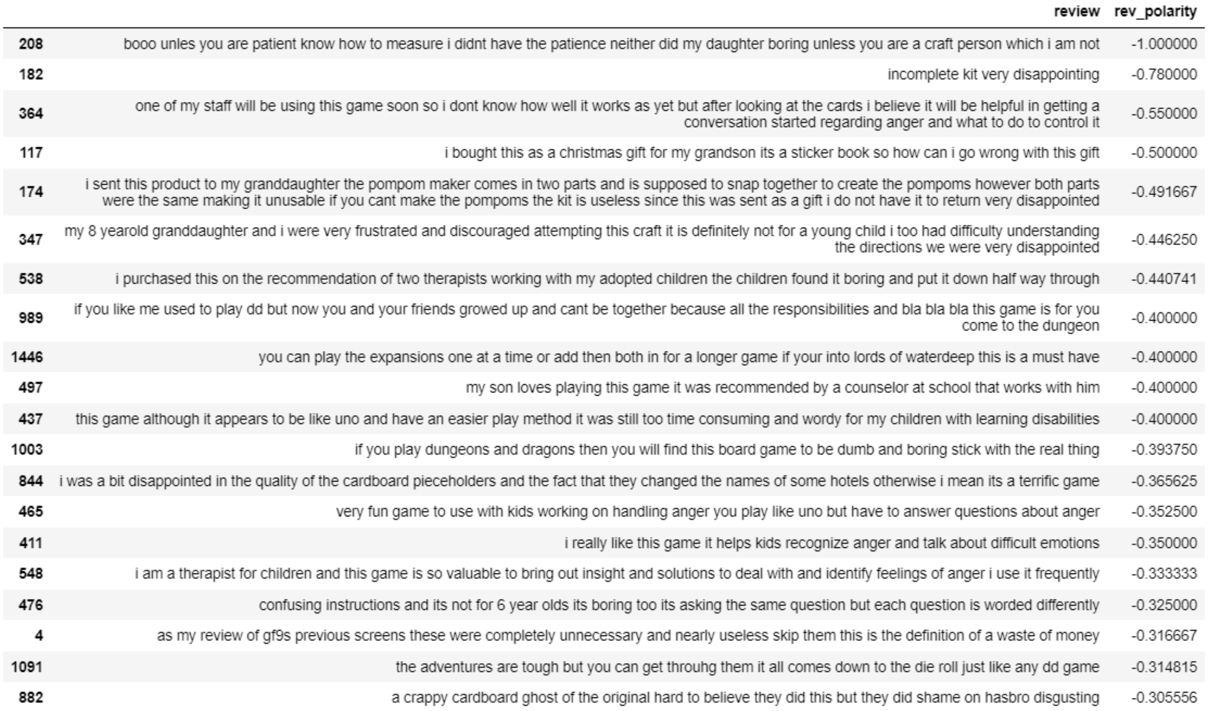
- Cluster 2 isn't cost effective.
- Clusters 3 and 4 won't supply business revenue the long run.
- It is cheaper to retain existing customers than find new ones.

WORDCLOUDS



The cleaned Wordclouds reveal that the words "game", "like", "book", "tile", "play", "make", "great" and "time" are the largest, suggesting that games and books are popular among customers.

TOP 20 REVIEWS



	review	rev_polarity
208	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
182	incomplete kit very disappointing	-0.780000
364	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
117	i bought this as a christmas gift for my grandson its a sticker book so how can i go wrong with this gift	-0.500000
174	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
347	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
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	-0.440741
989	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	-0.400000
1446	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	-0.400000
497	my son loves playing this game it was recommended by a counselor at school that works with him	-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	-0.400000
1003	if you play dungeons and dragons then you will find this board game to be dumb and boring stick with the real thing	-0.393750
844	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	-0.365625
465	very fun game to use with kids working on handling anger you play like uno but have to answer questions about anger	-0.352500
411	i really like this game it helps kids recognize anger and talk about difficult emotions	-0.350000
548	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	-0.333333
476	confusing instructions and its not for 6 year olds its boring too its asking the same question but each question is worded differently	-0.325000
4	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	-0.316667
1091	the adventures are tough but you can get throuhg them it all comes down to the die roll just like any dd game	-0.314815
882	a crappy cardboard ghost of the original hard to believe they did this but they did shame on hasbro disgusting	-0.305556

The top 20 positive reviews and summaries show that customers are satisfied with their purchases. The negative reviews show that customers are mostly dissatisfied by:

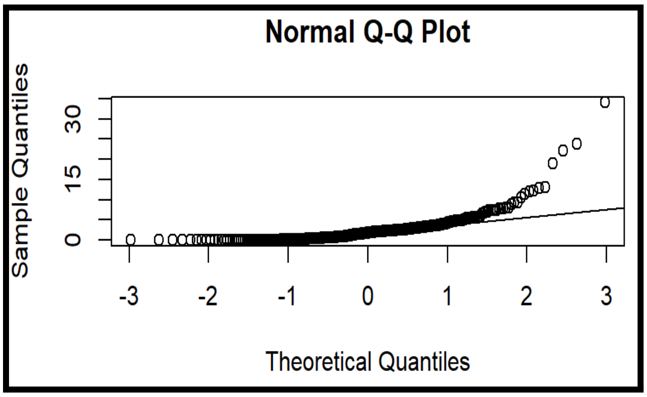
- incomplete product packaging
- unclear instructions
- products not working properly
- incomplete kits
- unclear messaging (eg games need patience)
- copied games or quality issues

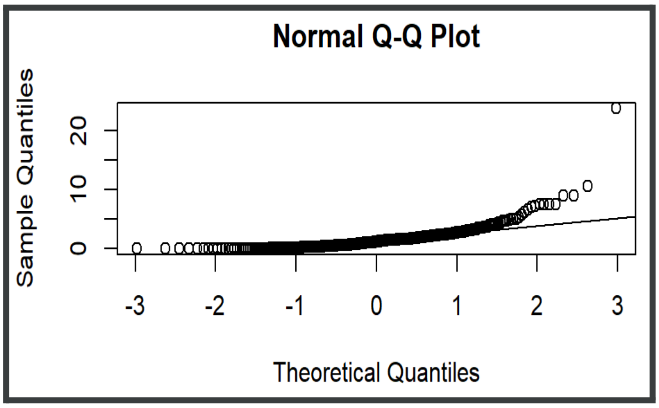
LIMITATIONS

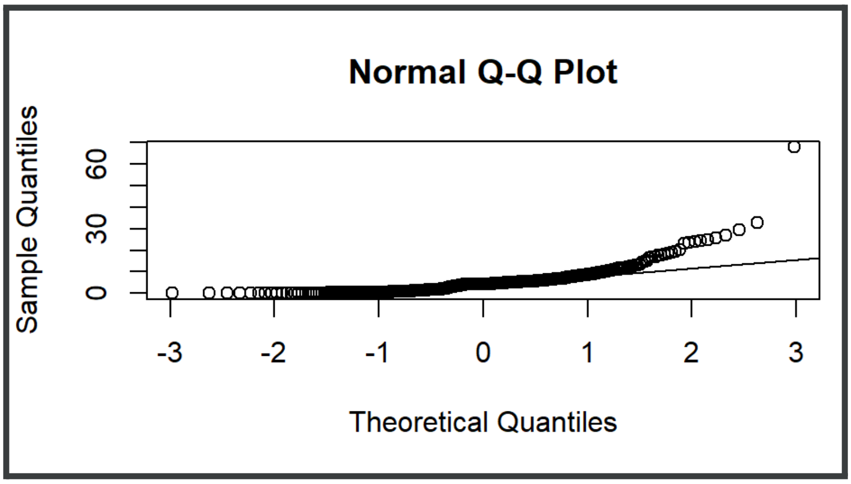
- Sentiment Analysis via wordclouds, review and summary data lacked contextual analysis. Conflicting statements lead inaccurate polarity scores, as evidenced by:
  - half the negative reviews actually being positive, although all negative summaries were actually negative. This difference suggests that customers favour summaries for negative comments over reviews.
- All variables in the linear regression models have low R-squared values (<0.70) indicating that they are a poor fitly fitted.

SALES DATA RELIABILITY AND NORMALITY

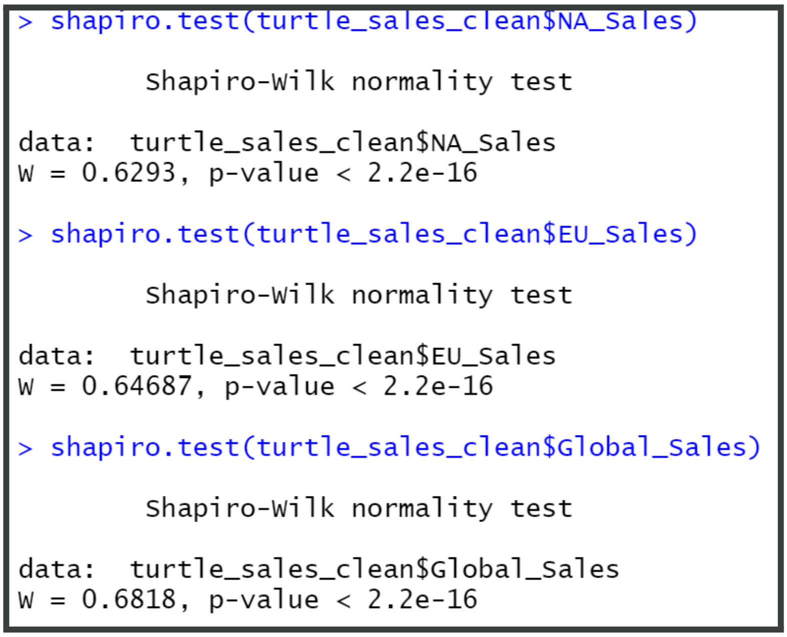
Analysis shows the data is **unreliable** and **not normal**.





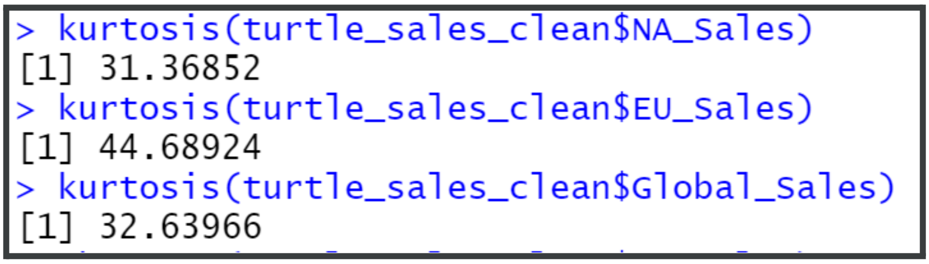


- qqplots (North America - top left; Europe - top right, Global - bottom) show that some data points lie in a straight line and circle around the line of best fit. Normality is ambiguous and needs to be checked via other tests.



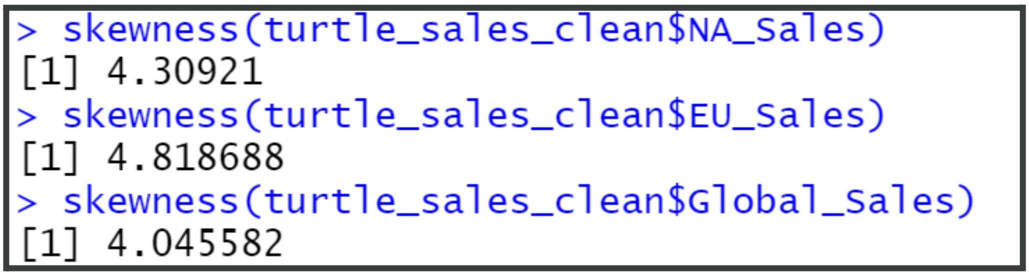
```
> shapiro.test(turtle_sales_clean$NA_Sales)

        Shapiro-Wilk normality test

data:  turtle_sales_clean$NA_Sales
W = 0.6293, p-value < 2.2e-16

> shapiro.test(turtle_sales_clean$EU_Sales)

        Shapiro-Wilk normality test

data:  turtle_sales_clean$EU_Sales
W = 0.64687, p-value < 2.2e-16

> shapiro.test(turtle_sales_clean$Global_Sales)

        Shapiro-Wilk normality test

data:  turtle_sales_clean$Global_Sales
W = 0.6818, p-value < 2.2e-16
```

- All p-values from the Shapiro Wilk test are less than 0.05. Thus we reject the null hypotheses that the data for all 3 sales columns are normally distributed.



```
> kurtosis(turtle_sales_clean$NA_Sales)
[1] 31.36852
> kurtosis(turtle_sales_clean$EU_Sales)
[1] 44.68924
> kurtosis(turtle_sales_clean$Global_Sales)
[1] 32.63966
```

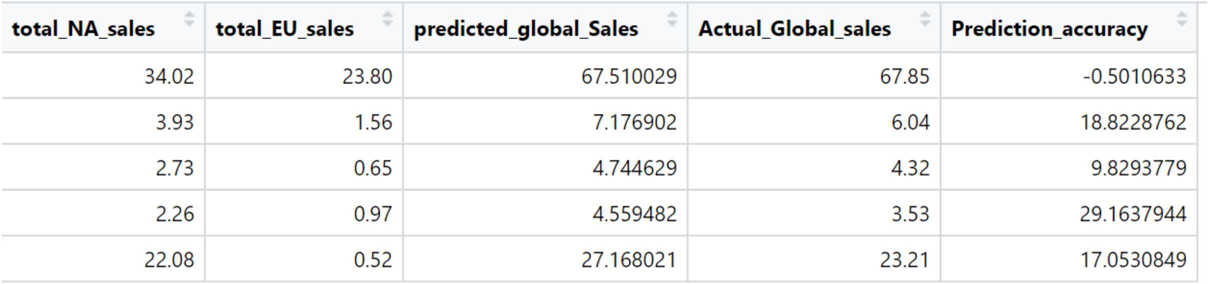
- The skewness levels for all sales data are larger than 1, indicating right positive skewness; data points are biased to higher values and not normal.



```
> skewness(turtle_sales_clean$NA_Sales)
[1] 4.30921
> skewness(turtle_sales_clean$EU_Sales)
[1] 4.818688
> skewness(turtle_sales_clean$Global_Sales)
[1] 4.045582
```

- The kurtosis levels for all sales regions are higher than 3, indicating a leptokurtic distribution that has abundant extreme outliers compared to a normal distribution.

MULTILINEAR REGRESSION MODEL



total_NA_sales	total_EU_sales	predicted_global_Sales	Actual_Global_sales	Prediction_accuracy
34.02	23.80	67.510029	67.85	-0.5010633
3.93	1.56	7.176902	6.04	18.8228762
2.73	0.65	4.744629	4.32	9.8293779
2.26	0.97	4.559482	3.53	29.1637944
22.08	0.52	27.168021	23.21	17.0530849

- Prediction_accuracry shows the **percentage difference between** predicted_global_sales and Actual_global_sales
- Further fine tuning is needed as the average prediction accuracy is 15.1% - ie predictions are overestimated.

RECOMMENDATIONS

PRODUCT & SALES STRATEGY

- Contact disappointed customers and remedy their concerns.
- Investigate the games that have product issues and triangulate pain points - is it a supplier issue, or are we mispackaging things?
- Review unclear or misleading instructions and packaging.
- Continue to market games and books, especially, Playstation, Wii, XBOX, PC and Gameboy consoles to customers with high salary and high expenditure and moderate salary and moderate expenditure. This is especially pertinent as customers who spend more have higher loyalty points.
- Stop selling Gamecube, Atari and Sega Genesis consoles.

DATA ACCURACY

- Check the feedback systems as there are duplicate reviews.
- Provide dates of customers' purchases so we can identify which customers return over the past 6 months.
- Add more fields, such as regions, store names and vendor suppliers for further granularity.

TO EXPLORE

- Test the linear regression model with other variables.
- Provide contextual analysis in addition to sentiment analysis for more accurate results.
- Apply sentiment analysis on data scraped from social media sites and competitors' sites.

REFERENCES

1. https://cmdlinetips.com/2018/03/how-to-filter-a-pandas-dataframe-based-on-null-values-of-a-column/
2. https://www.machinelearningplus.com/pandas/pandas-duplicated/#:~:text=duplicated()%20method%20is%20used,row%20is%20duplicate%20or%20unique.
3. https://www.statology.org/pandas-rename-columns/
4. https://dzone.com/articles/kmeans-silhouette-score-explained-with-python-exam#:~:text=Silhouette%20score%20is%20used%20to,each%20sample%20of%20different%20clusters
5. https://www.statology.org/seaborn-title/
6. https://www.analyticsvidhya.com/blog/2021/06/nlp-sentiment-analysis/
7. https://datatofish.com/lowercase-pandas-dataframe/
8. https://www.geeksforgeeks.org/python-program-to-convert-a-list-to-string/
9. https://www.geeksforgeeks.org/how-to-rename-columns-in-pandas-dataframe/
10. How to convert integer to string in R? (projectpro.io)
11. Case when in R using case_when() Dplyr - case_when in R - DataScience Made Simple
12. Acquiring New Customers Is Important, But Retaining Them Accelerates Profitable Growth (forbes.com)
13. List of best-selling game consoles - Wikipedia
14. How to Use pivot_longer() in R - Statology