Analysis Report for Turtle Games

Er (Sophia) Xu

A. Background

Turtle Games is a game manufacturer and retailer with presence globally. The company's products include books, board games, video games and toys. The objective of the analysis is to improve overall sales performance by utilizing customer trends.

We aim to answer the following questions of Turtle Games

- How customers accumulate loyalty points
- How groups within the customer base can be used to target specific market segments
- How social data can be used to inform marketing campaigns
- The impact that each product has on sales
- The reliability of the data
- The relationships between NA, EU and Global Sales

B. Analytical Approach

1) Python

a. Importing libraries and conduct data cleaning

The following libraries were imported to conduct the analysis: *numpy, pandas, statsmodels.api, statsmodels.stats.api, sklearn, matplotlib, seaborn*

After importing and exploring the csv file, I checked for missing values and dropped unnecessary columns.

b. Linear regression

I created three linear regression models to evaluate the possible linear relationships between loyalty points and age/remuneration/spending scores.

c. Clustering

I used the Silhouette and Elbow methods to determine the optimal number of clusters for k-means clustering. Then, I evaluated the usefulness of three values for *k* (4, 5 and 6). The selection of k=5 can be justified by the evaluation process.

d. Using NLP to analyze customer sentiments

I started the analysis by keeping the relevant columns ("review" and "summary"). Then I prepared the data for NLP, including changing the data to lower case, joining the elements, replacing punctuations and dropping duplicates in respective columns.

To create word clouds for both columns, I first installed WordCloud. Then tokenization was applied on relevant columns. I plotted the WordCloud image after adjusting the size, background and font size of the image. I removed alphanumeric characters and stop words. Finally, I reviewed the sentiment scores for the respective columns and also plotted histograms of polarity. The top 20 positive reviews and the top 20 negative reviews were identified from this analysis.

2) **R**

I started with installing library tidyverse in R studio and importing the csv file “turtle sales”. After reviewing and exploring the data, I removed unnecessary columns. Then I used qplot to create scatterplots, histograms and boxplots to get some initial insights into the sales data.

Then I performed Shapiro-Wilk test on all of the sales data, determined the skewness and kurtosis of the sales data. I checked the correlation between the sales data columns.

Finally, I used ggplot2 and generated the types of plots that best suit the sales data and the insights to address Turtle Games’ questions.

C. Visualization and Insights

1) The linear regression models

a. Age and loyalty points

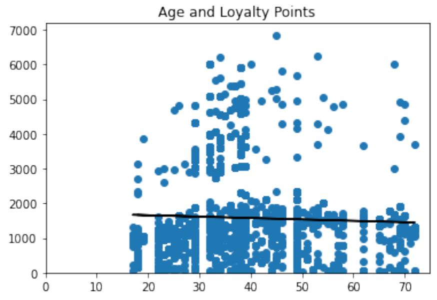
The R-square value of 0.002 indicates the proportion of the total variation in loyalty points that is explained by the total variation of age is limited.

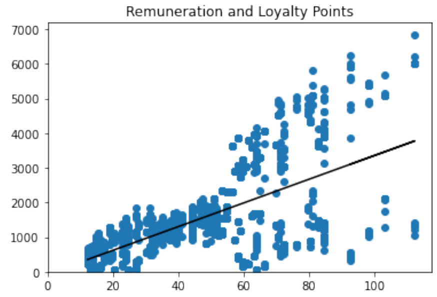
b. Remuneration and loyalty points

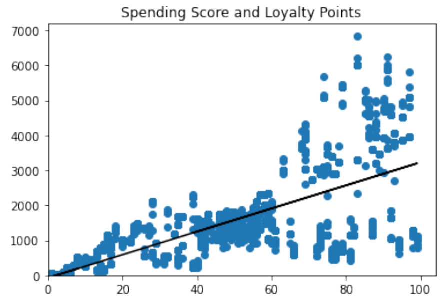
The correlation co-efficient 34.19 indicates a strong relationship between the two variables.

c. Spending score and loyalty points

The standard error value of 0.814 indicates that the data points are relatively close to the regression line and the regression equation can be used to predict with some confidence.

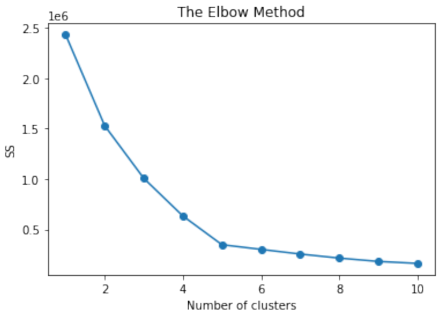


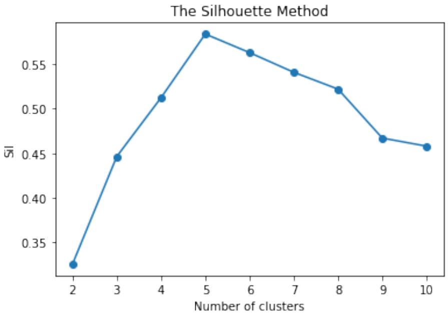




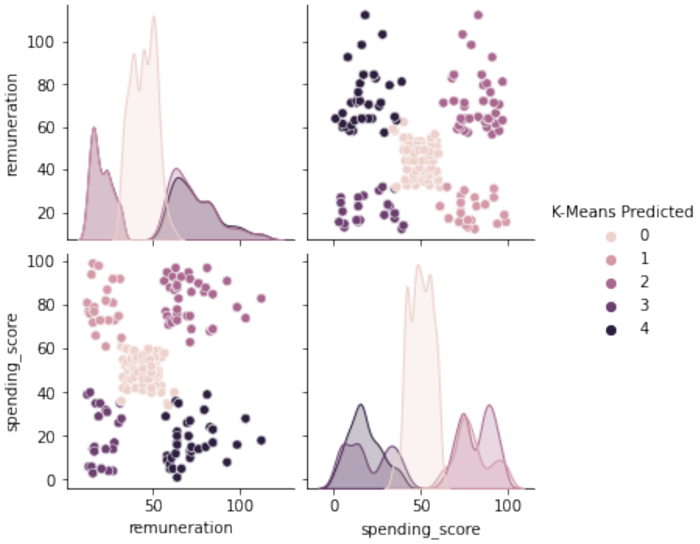
2) Clustering

Both Silhouette and Elbow methods indicate the optimal number of clusters for k-means clustering is 5.





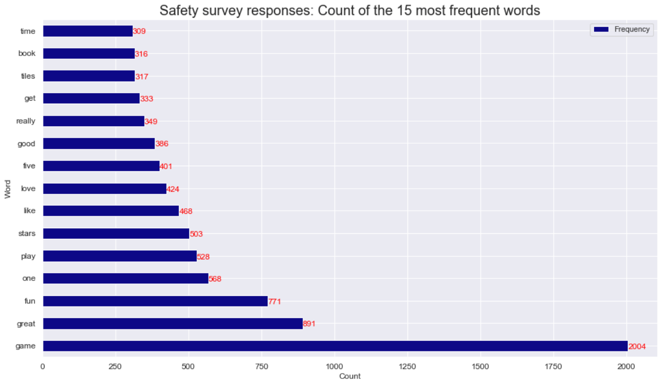
After evaluating three values of k (4, 5 and 6), I found k = 5 (five clusters) gives the best result, as shown below.



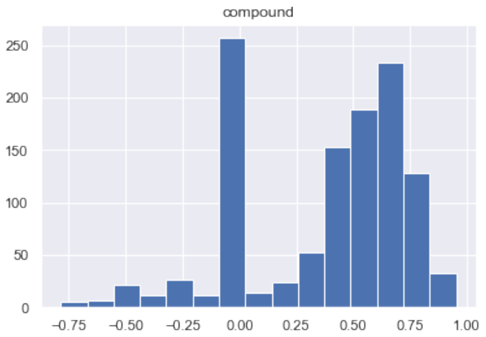
3) **Sentiment Analysis**

The 15 most frequently used words in the customer review are the following: *game, great, fun, one, play, stars, like, love, five, good, really, get, tiles, book, time*, as shown in the word cloud image.

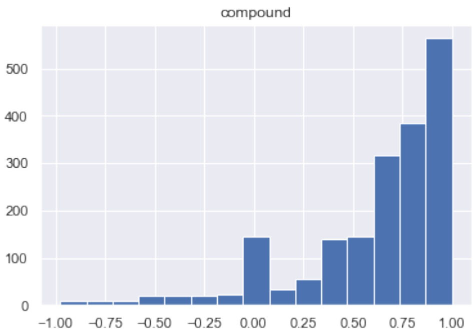
game
fun
one
play
great
five
star
love
tile
good
card
make
time
player
book
kid
set
piece
little
expansion
children
help
different
will
well
played



Sentiment analysis for the summary column: 1165 instances were analyzed. Overall, there are more positive sentiments than negative sentiments (mean = 0.38, 50th percentile = 0.50), but there is also a significant portion of neutral sentiments (25th percentile = 0).

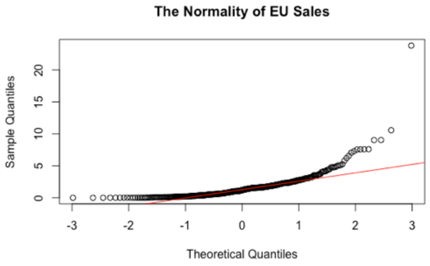


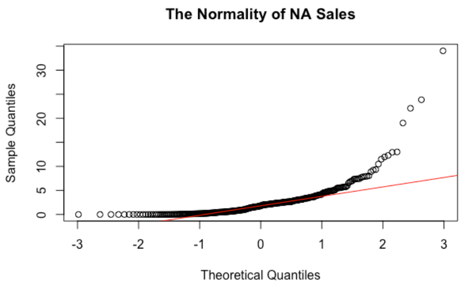
Sentiment analysis for the review column: 1893 instances were analyzed. The mean is a positive value (0.61). The 25th percentile, 50th percentile and 75th percentile are all positive values, indicating there are more positive sentiments than negative sentiments.

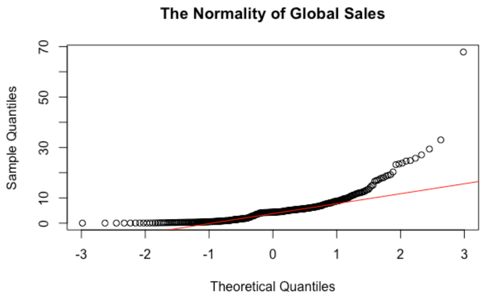


4) The reliability of the data

From the Shapiro Wilk tests, we found consistent patterns across all three columns of sales data: 1) p value less than 0.05 2) skewness is not close to 0 and 3) kurtosis is over 3, indicating that the sales data are not normally distributed and are fat-tailed.

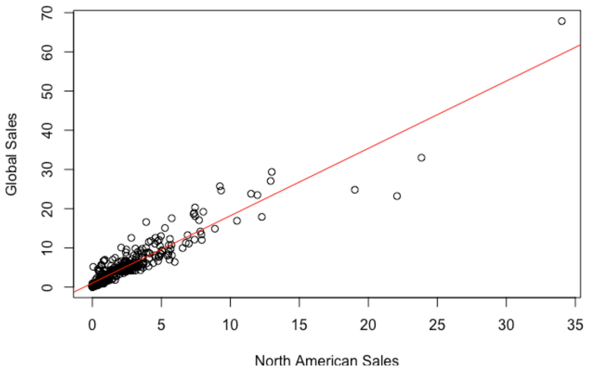






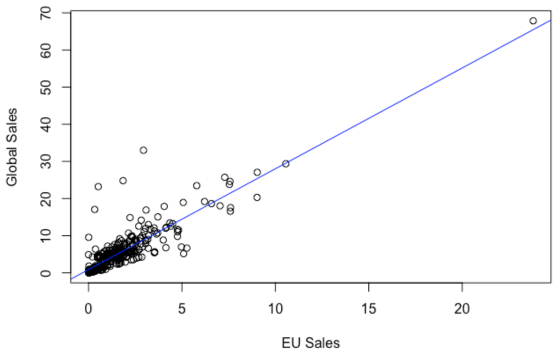
D. Patterns and predictions

1) Simple Linear Regression Model: NA and Global Sales
   - Multiple R-squared: 0.8741
   - Adjusted R-squared: 0.8738
   - P value <0.05

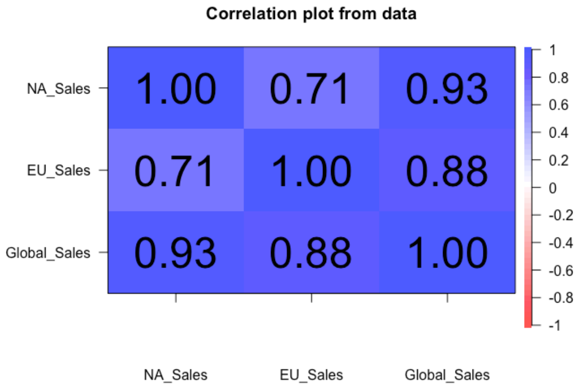


2) Simple Linear Regression Model: EU and Global Sales
   - Multiple R-squared: 0.7701
   - Adjusted R-squared: 0.7695
   - P value <0.05

3) Multiple Linear Regression Model: NA, EU and Global Sales
   - Multiple R-squared: 0.9687
   - Adjusted R-squared: 0.9685
   - P value < 0.05



The multiple linear regression model is a strong model, with the highest adjusted R-squared value, indicating North American sales and EU sales together are good explanatory variables of Global Sales. However, the two independent variables NA sales and EU sales are correlated (correlation = 0.71), implicating the issue of multicollinearity.



To test the accuracy of the multiple linear regression model, I compared the predicted value of global sales and the actual observed value.

Number	**NA Sales**	**EU Sales**	**Global Sales (Predicted)**	**Global Sales (Observed)**
1	34.02	23.80	71.47	67.85
2	3.93	1.56	6.86	6.04
3	2.73	0.65	4.25	4.32
4	2.26	0.97	4.13	3.53
5	22.08	0.52	26.43	23.21

Final recommendations to Turtle Games:

1. Remuneration and spending scores are explanatory variables of loyalty points
2. Turtle Games can segment customers into 5 clusters based on remuneration and spending scores. They can increase sales by concentrating their marketing dollars on customers with high remuneration and low spending scores.
3. Overall, the social data shows positive sentiments towards Turtle Games' products. The management team can benefit from reviewing more data points, given the large proportion of neutral sentiments. Also, they can dig deeper into potential operational issues revealed in the negative comments.
4. The sales in North America and Europe both have significant impact on global sales. Turtle Games need to pay attention to sales in both regions, with North America at a higher priority.