Turtle Games Data Assignment 3: Predicting future outcomes

Report Prepared for Turtle Games Sept 2022

SEPTEMBER 12 **2022**

Turtle Games / LSE Career Accelerator
Authored by: Zikomo Smith

Executive Summary

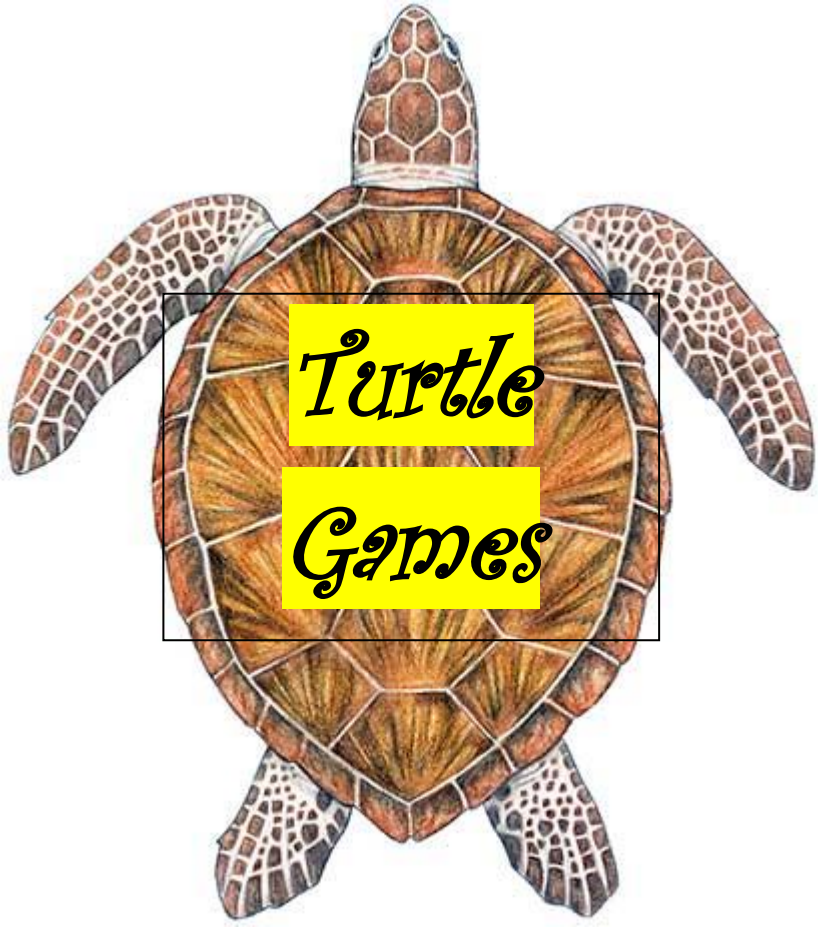
Background/context of the brief

Turtle Games wants to improve its overall sales performance by assessing customer trends to build a business strategy. Turtle Games must understand the insights that current customer and sales data provide, whether the datasets are reliable, and how to proceed with the data at their disposal.

Kick Off Call Questions & Answers – Insights and Patterns

1. How do customers accumulate loyalty points? (Visualizations on pg.8)
    - Higher spending customers and richer customers tend to accumulate more loyalty points
    - Loyalty is less correlated with customers with a spending score over 60 and a remuneration of over £60K
2. How to group customers into specific market segments? (Visualizations on pg.9)
    - Aim to get customers with high remuneration to spend more money
    - Cluster analysis based on remuneration and spending suggests 5 distinct groups to target
3. how can customer review inform marketing campaigns? (Visualizations on pg.10)
    - Current review sentiment is positive – aim to understand the words customers use to describe Turtle Game's products and incorporate them into marketing campaigns.
    - Include customer sentiment as a marketing KPI to over time to assess whether marketing improves the perception of Turtle Games.
    - Reassess the engagement strategy of those with fewer loyalty point (0-2000) who tend to share the most comments
4. What impact does each product have on sales? (Visualizations on pg.11)
    - The long tail of product IDS account drive the most sales
    - Three of the top-selling products, IDs = 123, 254, & 948, sell almost double the amount in North America compared to Europe - increasing their sales in Europe to drive incremental revenue
5. Is there a relationship between North American, European, and global sales. (Analysis on pg.12)
    - Both regions have a significant impact on Global sales - North America is still impactful market
6. how reliable is the data? (Analysis on pg.13)
    - The sales data is not normally distributed and skewed towards lower selling products
    - Turtle games might need to look at market signals to assess sales strategy over current data – it might be skewed due to a current, incorrect sales strategy.
    - We need to refine our predictive models further to account for the long tail of products that are more difficult to assess

Table of Contents



1.Analytical approach

Python Analytical Approach (Pros/ Cons):

1. *Assess missing values*

```
In [145]: # Any missing values?
          #Determine the number of rows that contain missing values.
          #Create a dataframe with the rows that have missing values
          #Assess the shape of the dataframe consisting of rows that have missing values
          #We can see the reviews data has no rows with missing values
          reviews_na = reviews[reviews.isna().any(axis=1)]

          reviews_na.shape

Out[145]: (0, 11)
```

- Created dataframe that identified no missing values in the data
- Ensured the completeness of the dataset, which gives us confidence that regular data collection is taking place.
- Does not tell us whether data has been QA'd.

2. *Create linear regression models across multiple views*

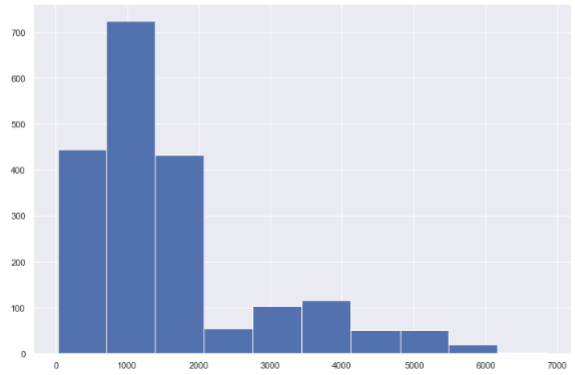
```
Predicted values:  [1659.01044907 1648.66145499 1621.06413743 ... 1579.6681611  1607.26547866
 1586.56749049]
Out[178]: [<matplotlib.lines.Line2D at 0x221bdce6190>]
```

```
In [179]: # Solution - age vs loyalty
```

- Allows us to assess correlation of a variety of factors
- Able to assess whether there was any difference between audiences (no significant difference found).
- Quite cumbersome and time consuming with current code.

3. *Use histograms to group data*

```
In [185]: #create histogram to understand the type of people sharing comments
          plt.hist(reviews['loyalty_points'])
Out[185]: (array([444., 724., 432.,  55., 103., 116.,  51.,  52.,  20.,   3.]),
           array([  25. ,  707.2, 1389.4, 2071.6, 2753.8, 3436. , 4118.2, 4800.4,
                  5482.6, 6164.8, 6847. ]),
           <BarContainer object of 10 artists>)
```

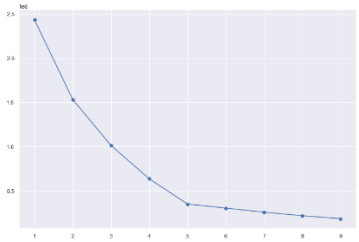


- Allows us to transform numeric loyalty data into a categorical feature of our customers.

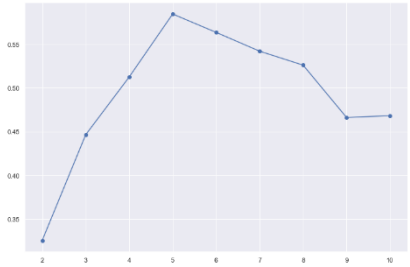
- Gives us insight that can inform a loyalty marketing strategy.
- Limited in expressiveness.

4. *Employ elbow and silhouette methods for clustering analysis*

- Elbow

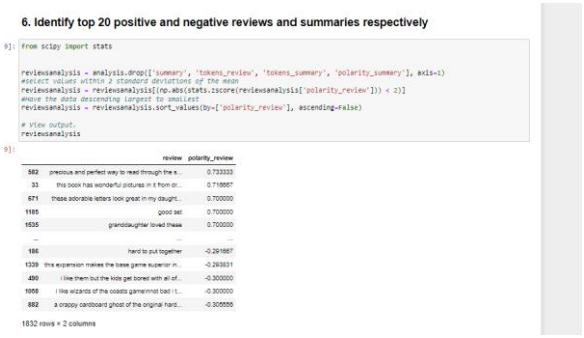


- Silhouette



- Use of multiple clustering methods producing similar results supports choice of 5 clusters.
- Requires more analysis to see whether we can effectively target clusters in reality.

5. *Remove outliers for assessing top tweets*



6. Identify top 20 positive and negative reviews and summaries respectively

```
from scipy import stats

reviewsanalysis = analysis.drop(['summary', 'tokens_review', 'tokens_summary', 'polarity_summary'], axis=1)
#select values within 2 standard deviations of the mean
reviewsanalysis = reviewsanalysis[(np.abs(stats.zscore(reviewsanalysis['polarity_review'])) < 2)]
#move the data descending largest to smallest
reviewsanalysis = reviewsanalysis.sort_values(by=['polarity_review'], ascending=False)

# View output.
reviewsanalysis
```

	review	polarity_review
562	precious and perfect way to read through the a...	0.733333
33	this book has wonderful pictures in it from dr...	0.716667
671	these adorable letters look great in my daught...	0.700000
1185	good set	0.700000
1535	granddaughter loved these	0.700000
...	...	...
186	hard to put together	-0.291667
1339	this expansion makes the base game superior in...	-0.293831
490	i like them but the kids get bored with all of...	-0.300000
1058	i like wizards of the coasts gamemnot bad i t...	-0.300000
882	a crappy cardboard ghost of the original hard...	-0.305556

1832 rows × 2 columns

- Used stats module to remove tweets with a sentiment 2 standard deviations outside the mean
- Allows us to focus on non-outlier tweets
- Turtle Games may miss out on negative sentiment it needs to monitor.

R Analytical Approach

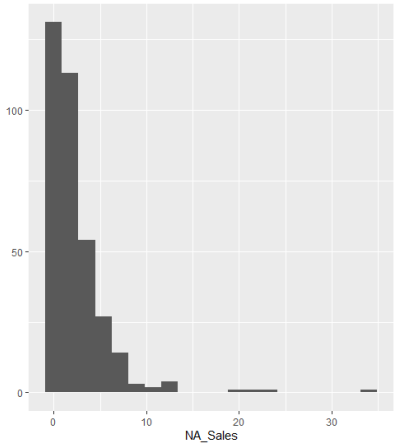
6. *Drop unnecessary columns*

```
# Create a new data frame from a subset of the sales data frame.
# Remove unnecessary columns.
salesfinal <- subset (turtle_sales, select = -c(Ranking, Year, Genre, Publisher))
```

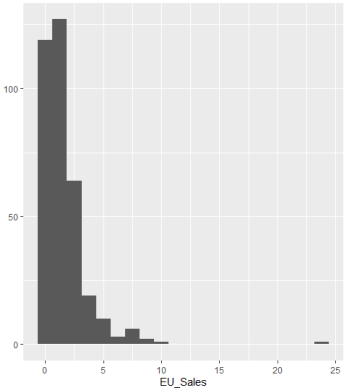
- Used subset function to create a new dataframe and drop unnecessary columns
- Allows us to focus on most pertinent business metrics.

7. *Use histograms to compare game sales*

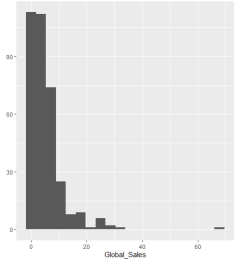
NA Sales



EU Sales



Global Sales



- Allows us to see skew of data and where majority of products fall in terms of sales
- Allows us to identify outlier sales values and long tail
- Does not have predictive power and does not account for market trends

8. Use multiple linear regression to make predictions

```
#############################################################################

# 3. Create a multiple linear regression model

# Multiple linear regression model.
multi.fit = lm(Global_Sales_sum~EU_Sales_sum+NA_Sales_sum+Product, data=productsales)
summary(multi.fit)

#############################################################################

# 4. Predictions based on given values
# Compare with observed values for a number of records.

# Compare with observed values for a number of records.
# Predict based on NA_Sales_sum of 34.02 and EU_Sales_sum of 23.80.

Question1 <- data.frame(EU_Sales_sum=c(23.80), NA_Sales_sum=c(34.02), Product=c(107))

predict(multi.fit, newdata=Question1)
```

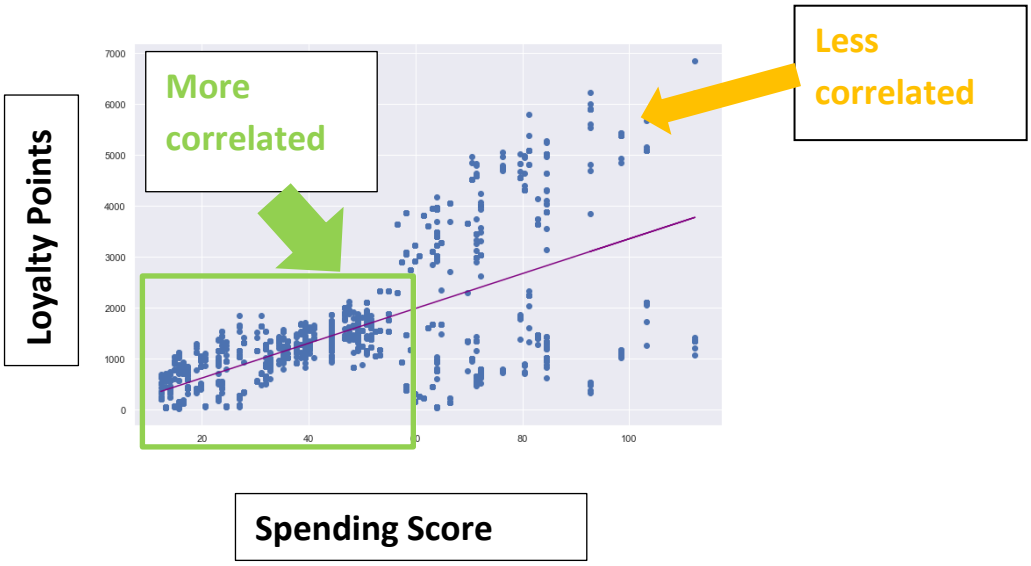
- Allows us to predict sales based on region
- Allows for topline insight into relative importance/ potential sales targets per region
- Does not allow for more granular analysis into factors driving sales

2.Visualizations & Insights

How do customers accumulate loyalty points?

Spending Score vs Loyalty Points

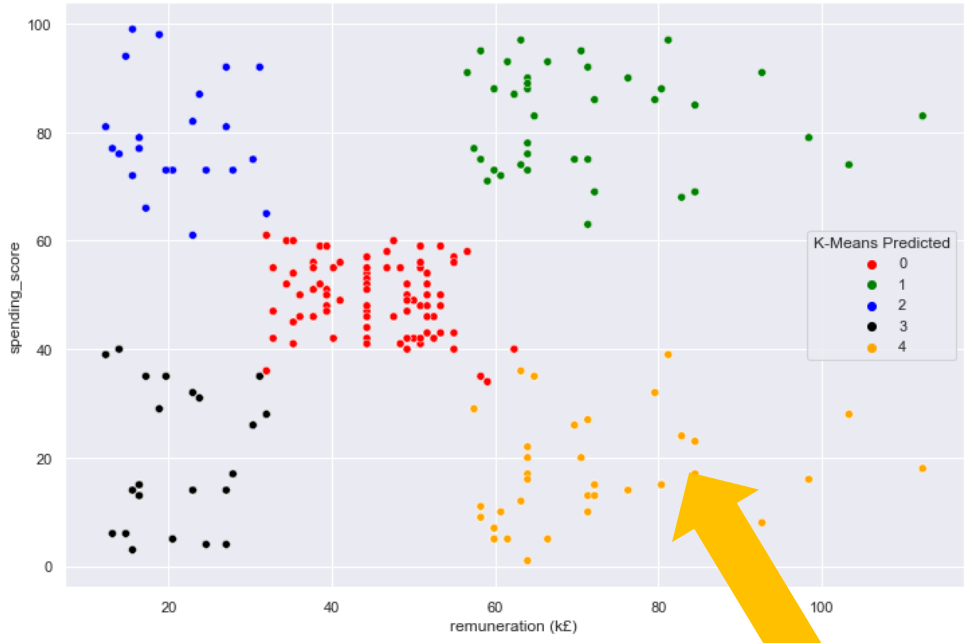
Remuneration vs Loyalty Points



[Return to exec summary](#)

How to group customers into specific market segments?

- Cluster analysis grouped by customer spending score relative to customer remuneration.

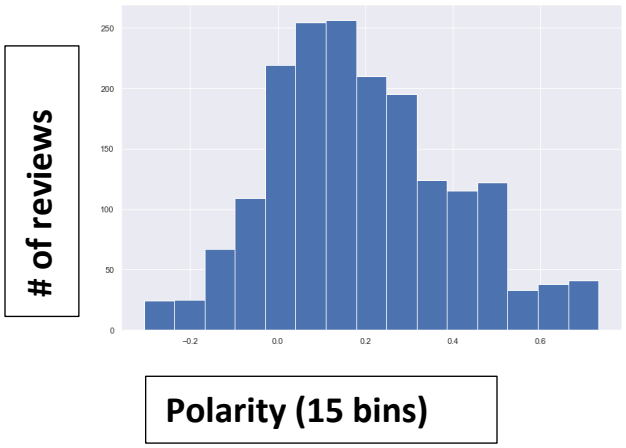


Target this group of wealthier customers who spend less

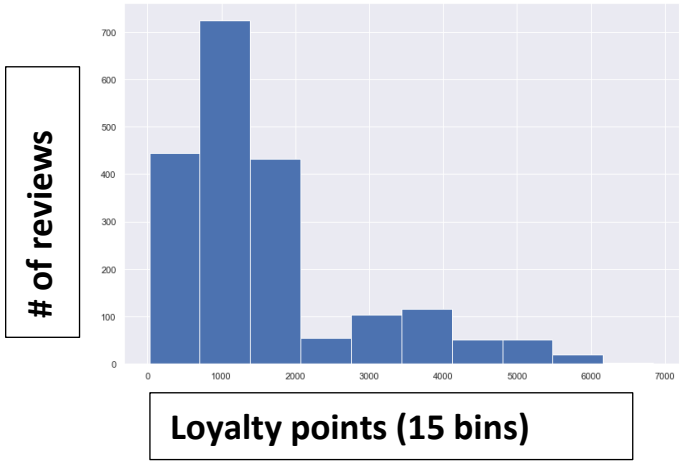
[Return to exec summary](#)

how can customer review inform marketing campaigns?

Review Polarity



of reviews by loyalty points



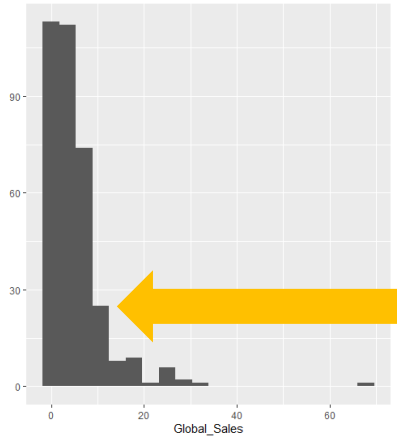
Top 20 tweets

	summary	polarity_summary
1507	sturdy bright colors awesome	0.850000
155	beautiful	0.850000
1764	perfect item i loved it	0.850000
642	beautifully made	0.850000
35	beautiful coloring book	0.850000
40	so beautiful	0.850000
309	she made four beautiful puppies from the kit a...	0.850000
703	great quality very cute and perfect for my tod...	0.816667
1190	great start for any wargamer looking for orcs ...	0.800000
508	great	0.800000
1262	great for the price	0.800000
199	great product darling puppies	0.800000
1783	great puzzle toy	0.800000
516	great therapist tool	0.800000
1327	great expansion	0.800000
1472	great expansion	0.800000
1805	great for your coffee table	0.800000
1280	a great tile set for any fantasy gaming group	0.800000
1175	another great dungeon command set	0.800000
1473	great expansion set	0.800000

[Return to exec summary](#)

What impact does each product have on sales?

Long tail of products



Long tail of products

Top Selling Products, NA & EU Comparison

Product	NA_Sales_sum	EU_Sales_sum	Global_Sales_sum
<int>	<dbl>	<dbl>	<dbl>
107	34.0	23.8	67.8
515	19.2	18.9	45.9
123	26.6	4.01	37.2
254	21.5	2.42	24.4
195	13	10.6	
231	12.9	9.03	24.1
249	9.24	7.29	25.7
948	14.4	7.79	25.4
876	12.8	9.25	25.3
263	9.33	7.57	24.6

Large delta between NA & EU

[Return to exec summary](#)

Is there a relationship between North American, European, and global sales.

> cor(productsales$EU_Sales_sum, productsales$Global_Sales_sum)

EU correlation: 0.8486148

> cor(productsales$NA_Sales_sum, productsales$Global_Sales_sum)

NA correlation: 0.9162292

(1 is the highest correlation coefficient a market can achieve)

Return to exec summary

how reliable is the data?

Shapiro-Wilk normality test (NA)

W = 0.69813, p-value < 2.2e-16

Shapiro-Wilk normality test (EU)

W = 0.74058, p-value = 2.987e-16

Shapiro-Wilk normality test (Global

W = 0.70955, p-value < 2.2e-16

Skewness NA
3.048198
Kurtosis NA
15.6026
Skewness EU
2.886029
Kurtosis EU
16.22554
Skewness Global
3.066769
Kurtosis Global
17.79072

P-Values suggest the data is not normally distributed and heavily skewed left (towards a long tail of lower selling products.)

Return to exec summary

3.Patterns & Predictions

Based on the 5 scenarios Turtle Games provided, below are the global sales predictions by product:

Scenario	EU Sales (M)	NA Sales (M)	Predicted Global Sales (M)	Actual Global Sales (M)
1(Product ID107)	23.80	34.02	66.3587	67.8
2(Product ID99)	3.93	1.56	8.645582	6.04
3(Product ID176)	2.73	0.65	6.282032	4.32
4(Product ID258)	2.26	0.97	6.078803	5.6
5(Product ID326)	22.08	0.52	28.58208	23.21

- These predictions are based on the existing sales strategy which leverages a long tail of smaller selling products.
- These predictions do not account for seasonality or market shifts.
- Predictions overshoot sales, particularly for products with fewer sales.
- We recommend reassessing the model, given that so many products sell less.

Return to exec summary

Thank You