Data Analytics Report for Turtle Games London School of Economics in partnership with FourthRev

By Ted Malumbe
8th July 2022

Table of Contents

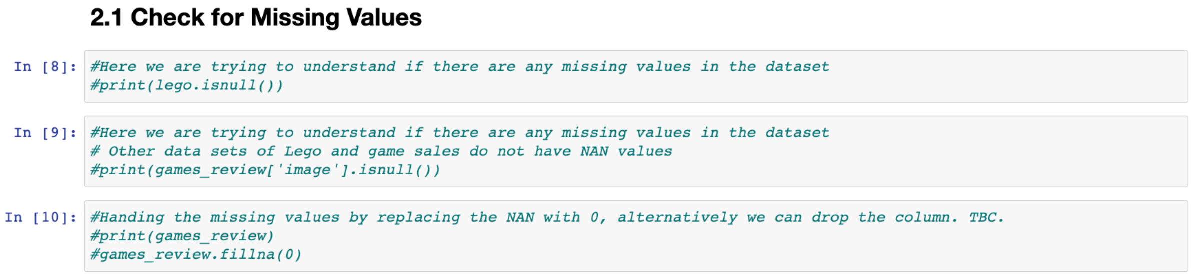
Background and Business Problem:

The purpose of this report is to assist Turtle games in addressing some of its key business objectives. These objectives can be summarized as the points listed below.

- Determine the optimal price for specific demographic to sell their products
- Understand the sentiments of customers across various products made
- Identify those customers that will most likely leave a review on the products
- Identify the most expensive product purchased by a particular group of customers
- Predict global sales using North America and European sales for the next year

Data wrangling and analysis:

Before any analysis was taken, I carried out the essential step of data wrangling that involved identifying duplicate values, null values, and missing values. With this information, I cleaned the data. For example, I changed missing values from NAN to 0 and deleted duplicate values from the dataset. The image below is an extract of one of those steps carried out in Python. Within the R environment, I leveraged on the use of the tidyverse and dplyr for my data manipulation following further instructions by (George, 2021).



2.1 Check for Missing Values

```
In [8]:  #Here we are trying to understand if there are any missing values in the dataset
         #print(lego.isnull())

In [9]:  #Here we are trying to understand if there are any missing values in the dataset
         # Other data sets of Lego and game sales do not have NAN values
         #print(games_review['image'].isnull())

In [10]: #Handing the missing values by replacing the NAN with 0, alternatively we can drop the column. TBC.
         #print(games_review)
         #games_review.fillna(0)
```

Image 1: Jupyter Notebook Extract Python

Dealing with these elements were important so as not to distort the dataset and provide incorrect recommendation or conclusions. Although more data wrangling was carried out, using location specific columns such as “review text” in game reviews. I could be confident incorrect data from other columns would not impact the column I was focusing on.

To be able to fully leverage the use of Linear Regression and Multiple Linear Regression, I applied the libraries such as statsmodels and sklearn. Using these libraries helped answer the business objectives of setting the optimal price. Whilst libraries such as NLTK and word cloud aided in analyzing the qualitative text and illustrated the sentiments of the customers graphically. This was valuable information for the business to understand what the customer thinks of their product offering. I used the reference from (Geek for Geek, 2022) in applying the regression models.

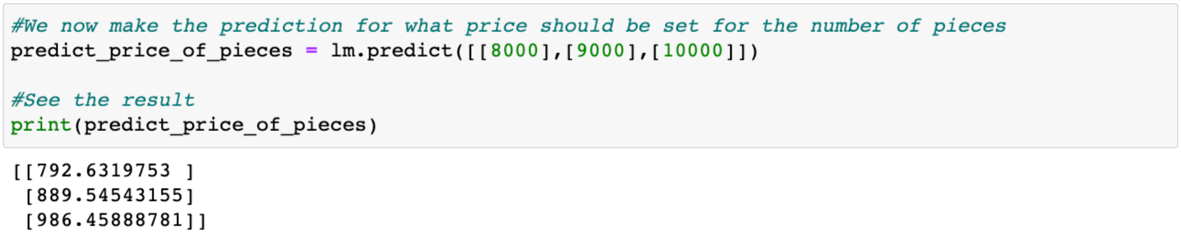
Visualization and insights:

As the analysis began with linear regression and multiple linear regression, the logical plot was a scatter plot to understand the relationship between the x and y variables as well as a linear regression plot to determine the strength of the relationship between the two variables. This approach gave a clear indication of for example, how an increase in Lego pieces developed an increase in price.

Using the Word cloud and counter, we could quickly and visually identify the words most frequently used by customers and get a broad overview of customer feedback. The histogram used also provided a quantifiable component to the customers responses. Which matched the counter and word cloud information previously retrieved.

Pattern and Predictions:

Some of the observations seen was that older people will be prepared to pay higher price for Lego set with more pieces. Since a linear relationship could be explained between the number of Lego pieces and the price. This provides the opportunity for the organization to increase the price. The below extract gives an example of prices that can be set for different number of pieces of 8000,9000 and 10000 pieces of Lego.



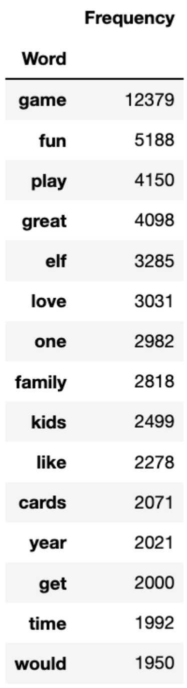
```
#We now make the prediction for what price should be set for the number of pieces
predict_price_of_pieces = lm.predict([[8000],[9000],[10000]])

#See the result
print(predict_price_of_pieces)
```

```
[[792.6319753 ]
 [889.54543155]
 [986.45888781]]
```

Image 2: Jupyter Notebook Extract Prediction

Based on the sentiment analysis using the word cloud we can determine out of the top 15 statements. Positive comments were largely represented as shown in the list below. That expresses words such as “Fun” “Great” and “Love” as the most common words. This provides positive feedback towards the general sentiments of the consumer to the products.



	Frequency
Word	
game	12379
fun	5188
play	4150
great	4098
elf	3285
love	3031
one	2982
family	2818
kids	2499
like	2278
cards	2071
year	2021
get	2000
time	1992
would	1950

Image 3: Jupyter Notebook Extract Sentiment Analysis

The histogram below goes on further to show the polarity in responses by the customers, overall showing a greater distribution towards positive sentiments.

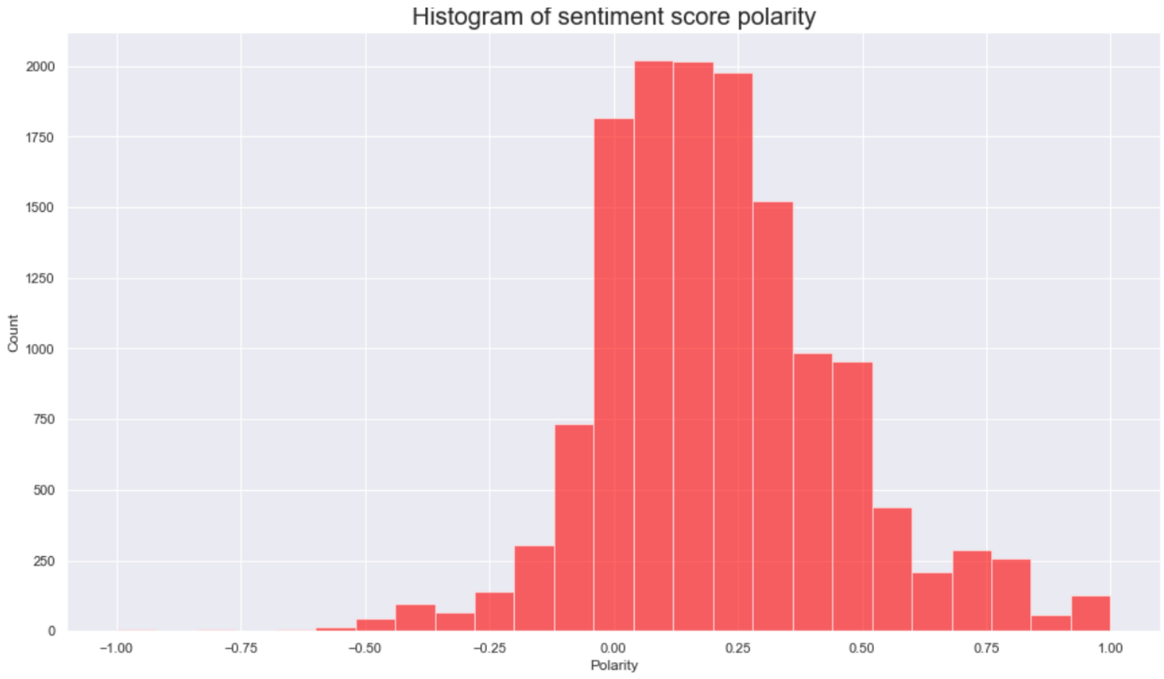
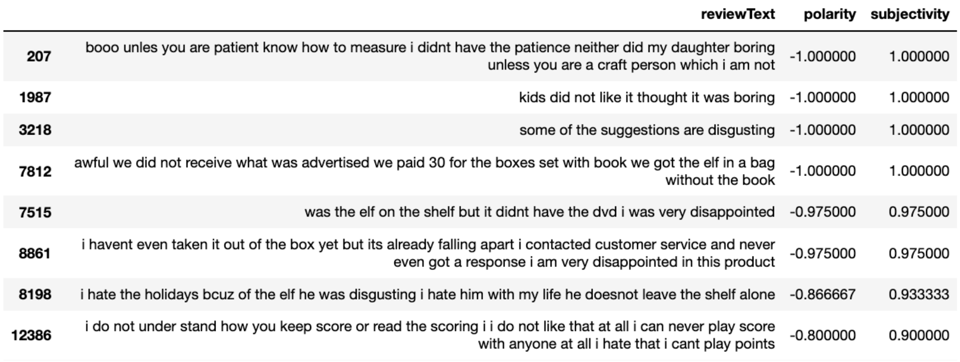


Image 4: Jupyter Notebook Extract Polarity Score

However, to improve the product the business must consider the negative reviews, below show a summarized list of the negative feedback with more summarized in the Jupyter Notebook. Some of the comments referred to the game as boring and product packaging. These comments can be grouped and then be sent to the respective business department to address further how they can improve the product.



	reviewText	polarity	subjectivity
207	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	1.000000
1987	kids did not like it thought it was boring	-1.000000	1.000000
3218	some of the suggestions are disgusting	-1.000000	1.000000
7812	awful we did not receive what was advertised we paid 30 for the boxes set with book we got the elf in a bag without the book	-1.000000	1.000000
7515	was the elf on the shelf but it didnt have the dvd i was very disappointed	-0.975000	0.975000
8861	i havent even taken it out of the box yet but its already falling apart i contacted customer service and never even got a response i am very disappointed in this product	-0.975000	0.975000
8198	i hate the holidays bcuz of the elf he was disgusting i hate him with my life he doesnot leave the shelf alone	-0.866667	0.933333
12386	i do not under stand how you keep score or read the scoring i i do not like that at all i can never play score with anyone at all i hate that i cant play points	-0.800000	0.900000

Image 5: Jupyter Notebook Extract Negative Reviews

With further study through R we can see (image 6) that the age bracket between 5 and 10 are the most likely to leave a review as they represent the highest number of reviews. This provides an opportunity for the business to focus on this demographic as they are the ones interacting with the product the most.

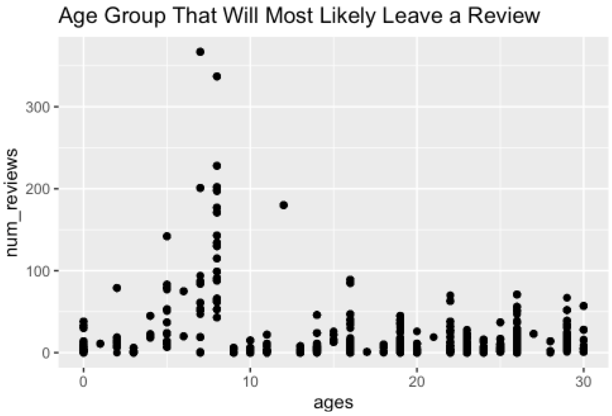


Image 6: R Script Extract Most Likely to leave a Review

And the graph below (image 7) shows in the demographic of 25-year-old the top right column is the most expensive product purchased.

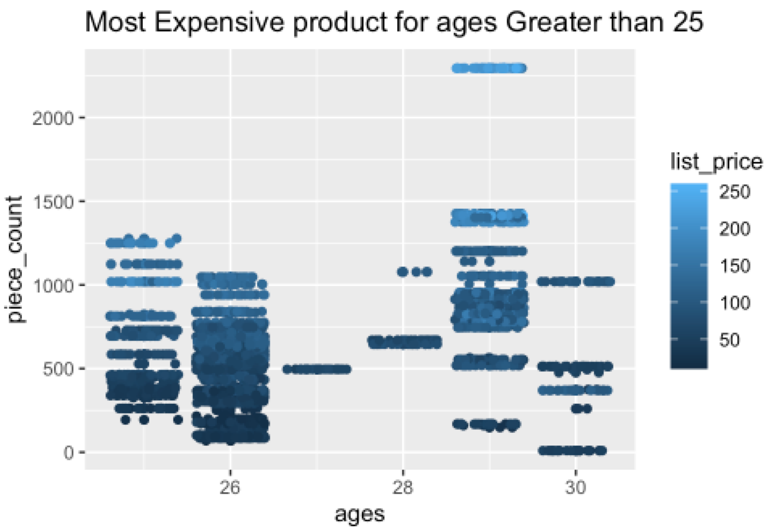


Image 7: R Script Extract Most Expensive Product Greater than 25

Whilst the graph below (image 8) shows the most popular product (2000 pieces) between the ages group of less than or equal to 25. Greater attention should be given to this LEGO set that is displaying an abnormal number of reviews.

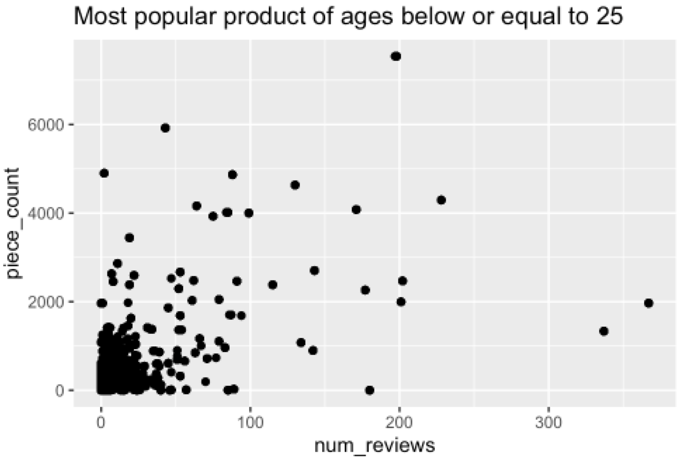


Image 8: R Script Extract Most Popular Product less than 25

Within the R scripts, I provide a collection of predicted sales for the next year considering the global, north America and European sales. This information can be used by the business to forecast the sales of next year. One approach used by the Institute of Business Forecasting is the Sales Ratio Model (IBF, 2022) that support this approach.

Bibliography:

George, M., 2021. *Advancing into Analytics: From Excel to Python and R.* 1st ed. Boston: Oreilly.
IBF, 2022. *Sales Ratio Model.* [Online]
Available at: https://ibf.org/knowledge/glossary/sales-ratio-model-242
[Accessed 6 July 2022].
Geek for Geek, 2022. *Linear Regression (Python Implementation).* [Online]
Available at: https://www.geeksforgeeks.org/linear-regression-python-implementation/
[Accessed 6 July 2022].