Turtle Games Report

LSE Data Analytics Career Accelerator

Maduraa Kannan
25th September 2022

Maduraa Kannan Turtle Games Report

Turtle Games, a games manufacturer and retailer, comprises of a global customer base. The company manufactures both its own products, and products manufactured by other companies. The products they sell include:

- Books
- Board games
- Video games
- Toys

The main objective of Turtle Games is for it to improve its overall sales performances.

I wanted to first find how customers accumulated loyalty points. I decided to use the data provided in turtle_reviews.csv to explore the linear relationship between the dependent variable, loyalty_points and the independent variables, age, remuneration, and spending_score in Python.

Firstly, I checked for missing data and duplicates within the dataset and then explored the dataset to get an understanding of Turtle Games and their customers. There were no missing or duplicated datasets. The initial data also provided useful information on the spread of age, spending score, remuneration and loyalty points of customers. Results from this analysis are recorded in the .pynb file.

OLS Regression Results

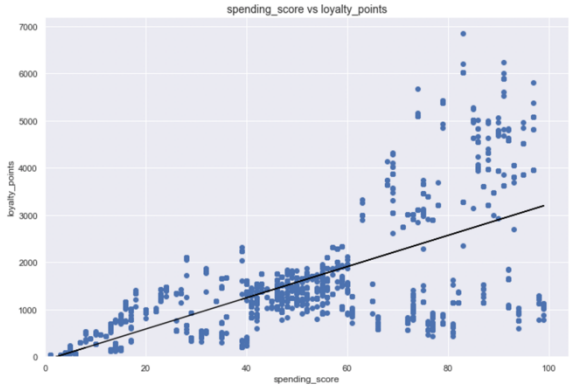
Dep. Variable:	y	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.606
Date:	Wed, 21 Sep 2022	**Prob (F-statistic):**	0.0577
Time:	23:26:55	**Log-Likelihood:**	-17150.
No. Observations:	2000	**AIC:**	3.430e+04
Df Residuals:	1998	**BIC:**	3.431e+04
Df Model:	1		
Covariance Type:	nonrobust		

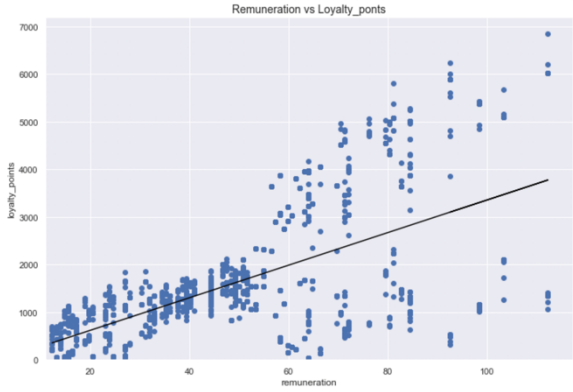
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1736.5177	88.249	19.678	0.000	1563.449	1909.587
x	-4.0128	2.113	-1.899	0.058	-8.157	0.131

Omnibus:	481.477	**Durbin-Watson:**	2.277
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	937.734
Skew:	1.449	**Prob(JB):**	2.36e-204

Figure 1: Age vs Loyalty_Points

I then evaluated the linear relationships. The correlation for age vs loyalty_score was negative however, the R^2 value (=0.002) for age vs loyalty_score indicated that the goodness of fit of the linear model was too poor to be used. Additionally, as the P-value > 5% (relationship between variables not statistically significant), and the regression line is underfitted (which leads to high biases and errors), the results from the regression analysis for age and loyalty_points cannot be used to predict loyalty_points.





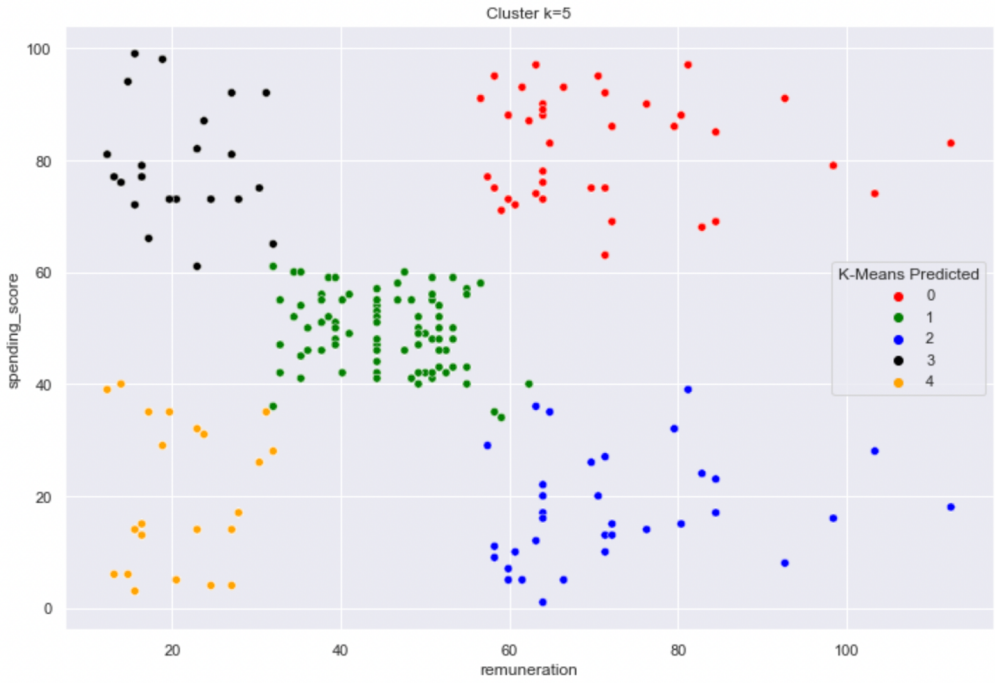
The linear relationships between remuneration and loyalty_points and spending_score and loyalty_points, however, shows a positive correlation with a good R^2 value (0.380 and 0.452 respectively) which indicate a good fit, and p-values indicating the relationship between the variables are statistically significant.

Next, I wanted to explore how graphs within the customer base can be used to target specific market segments. For this, I used k-means clustering to identify the optimal number of clusters and then applied and plotted the data using the created segments.

The optimal number of clusters was identified using the elbow method, followed by the silhouette method. I explored three values for *k*:

- K=4 – too few a cluster causing two clusters to become one clusters.
- K=5 – perfect fit
- K=6 – too many clusters which may result in an inaccurate analysis.

Having chosen k=5, I then plotted this data into a scatterplot using the seaborn library.



The segmentation is as follows (highest to lowest observations):

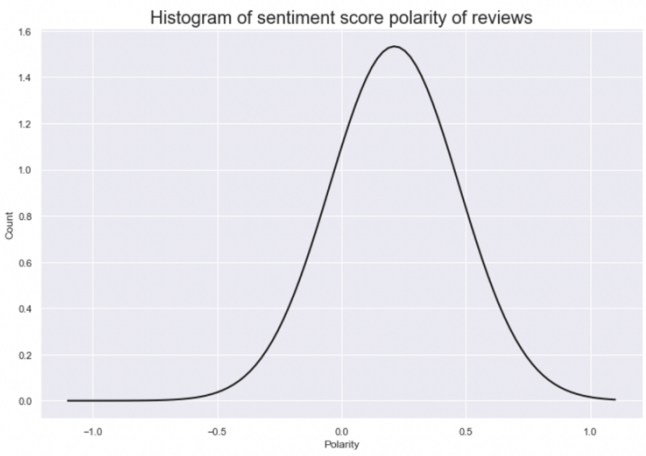
- Green (774)
- Red (356)
- Blue (330)
- Yellow (271)
- Black (269)

To improve overall sales, the marketing campaigns may want to focus more on the yellow and blue clusters (total 601 observation) as they have a low spending_score despite the remunerations being like the black and red clusters (total 625 observations) who have a higher spending_score. Questions for the marketing team may be: why do this cluster not

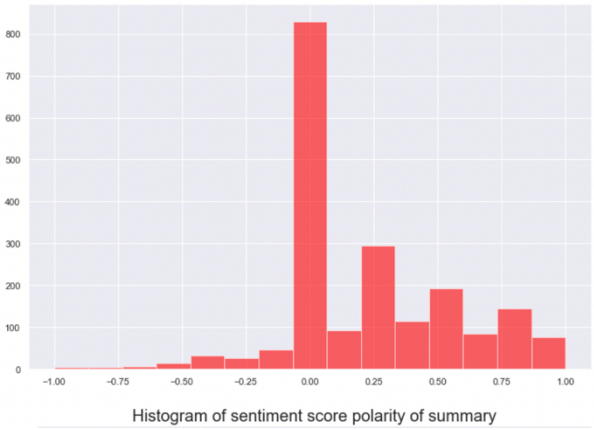
have as high a spending_score? Is there a way for the marketing to target these clusters and increase their spending_score?

I then explored how social data can be used to inform marketing campaigns. For this, I analysed customer sentiments with reviews. I used NLP to identify the 15 most common words used in online product reviews, and the top 20 positive and negative reviews received from the website.

I used a word cloud to create a visual representation of the most common words. It showcases a lot of positive words such as “five stars”, “fun”, “love” and “good”, and gives an indication of the kinds of products that customers are into, such as “game”, “book” and “toy”. The word cloud also has kid – could this be a target audience that the marketing team of turtle games should focus on?

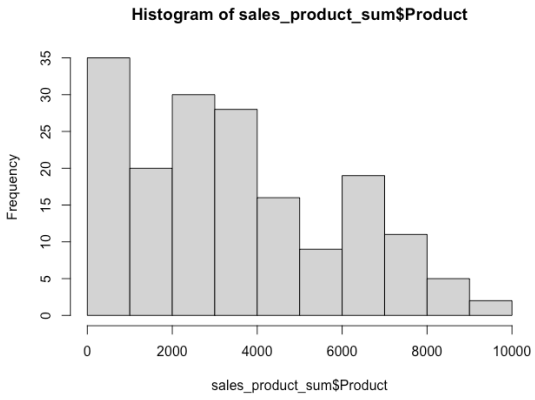


Maduraa Kannan Turtle Games Report



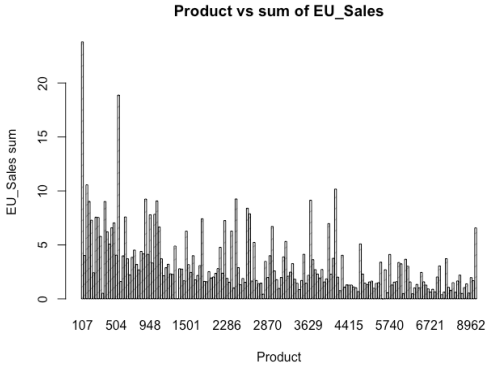
The results from the histogram of sentiment score polarity of reviews showed that the reviews are negatively skewed. There are more positive reviews than negative reviews. The results from summary showed the scores to be more neutral in comparison. This could be due to the lack of sarcasm captured in the summary. However, the summary doesn't capture the entirety of the review and so both analyses must be completed for accuracy. The reviews can provide insight to the game manufacturer and retailer about what they're doing right versus what they need to improve on and provide an overall insight into how the company is doing.

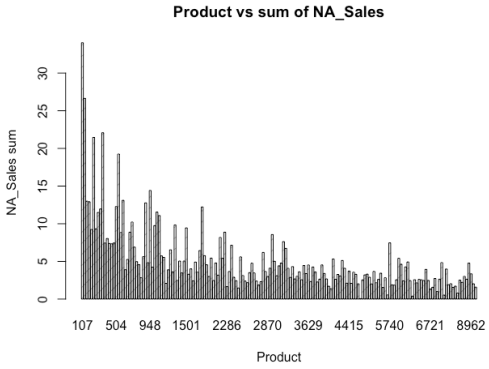
For the next stage of analysis, I used turtle_sales.csv, and performed the steps on R. I wanted to get an understanding of the impact each product has on sales, so I first created a histogram to showcase how many of each product was sold.

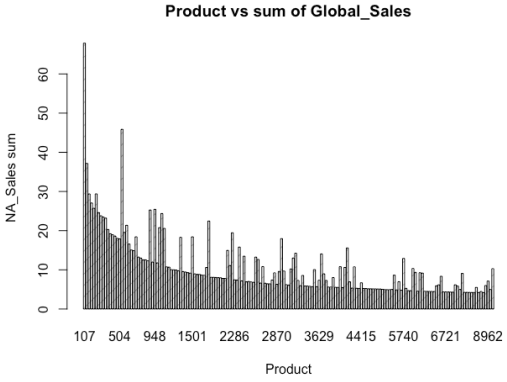


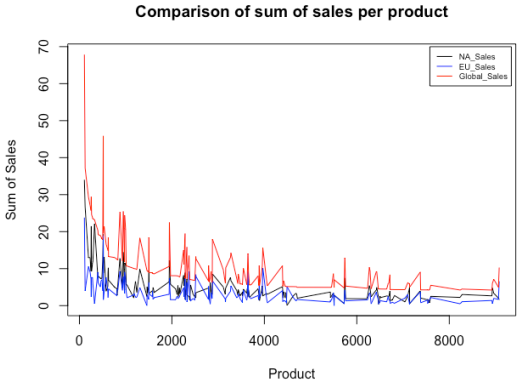
The histogram shows products are skewed to the right. Meaning the products with the lower ID number are sold the most.

The bar plots (Product vs sales) show that NA_Sales had a higher number of sales than EU_Sales. Is this a result of more people being a fan of these types of products in North America compared to the EU? Or is there a difference between the marketing that has led to this result?



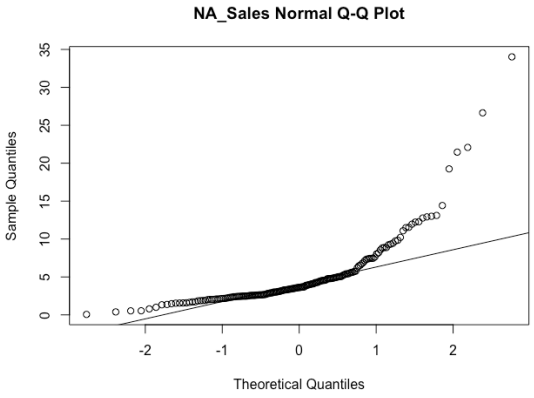


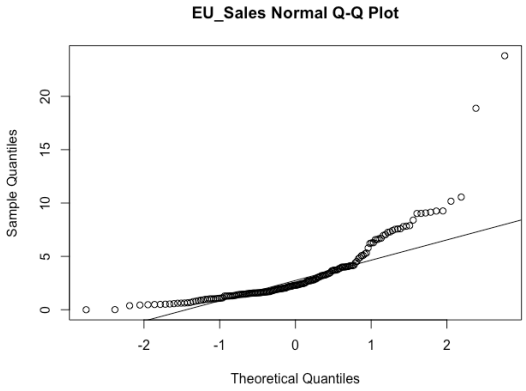




Graphs show similar trends amongst all sales, and follow the trend discovered in histogram of products.

Additionally, QQ plots, Shapiro-Wilk, and the skewness and Kurtosis of sales data showed that the data was not normally distributed and was heavily skewed. The diagram suggests the skewness was to the right.





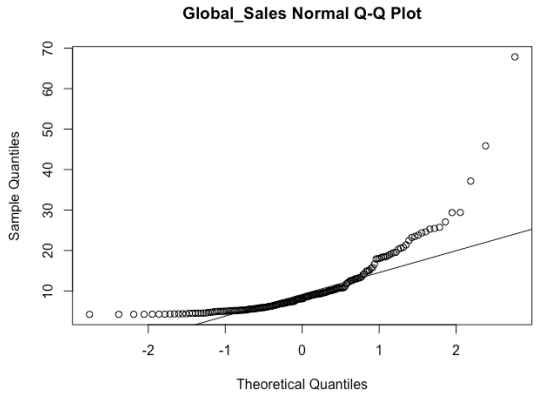


Figure 2 Q-Q plots all show the right skew

```
data:  sales_product_sum$NA_Sales
W = 0.69813, p-value < 2.2e-16

>
> # Shapiro-Wilk test for EU_Sales.
> shapiro.test(sales_product_sum$EU_Sales)

        Shapiro-Wilk normality test

data:  sales_product_sum$EU_Sales
W = 0.74058, p-value = 2.987e-16

>
> # Shapiro-Wilk test for Global_Sales.
> shapiro.test(sales_product_sum$Global_Sales)

        Shapiro-Wilk normality test

data:  sales_product_sum$Global_Sales
W = 0.70955, p-value < 2.2e-16
```

```
##########################################################################
# Performing a Shapiro-Wilk test on all the sales data.
##########################################################################

# Shapiro-Wilk test for NA_Sales.
shapiro.test(sales_product_sum$NA_Sales)

# Shapiro-Wilk test for EU_Sales.
shapiro.test(sales_product_sum$EU_Sales)

# Shapiro-Wilk test for Global_Sales.
shapiro.test(sales_product_sum$Global_Sales)

# A p-value <= 0.05 indicates that the test rejects the hypothesis of normality.
# In the case of these sales data, all the p-values are less than 0.05, and thus
# show that the datasets aren't normally distributed.
```

```
# Skewness and Kurtosis for NA_Sales.
skewness(sales_product_sum$NA_Sales) # = 3.048198>1, so highly skewed data.
kurtosis(sales_product_sum$NA_Sales) # = 15.6026>2, so not normally distributed.

# Skewness and Kurtosis for EU_Sales.
skewness(sales_product_sum$EU_Sales) # =  2.886029>1, highly skewed data.
kurtosis(sales_product_sum$EU_Sales) # = 16.22554>2, not normally distributed.

# Skewness and Kurtosis for Global_Sales.
skewness(sales_product_sum$Global_Sales) # = 3.066769>1, highly skewed data.
kurtosis(sales_product_sum$Global_Sales) # = 17.79072>2, not normally distributed.

# Kurtosis value shows the curves to be too peaked.
```

Non-normality does not indicate unreliableness of the data.

```
Call:
lm(formula = Global_Sales ~ EU_Sales + NA_Sales, data = sales_product_sum)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4156 -1.0112 -0.3344  0.6516  6.6163

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.04242    0.17736   5.877 2.11e-08 ***
EU_Sales     1.19992    0.04672  25.682  < 2e-16 ***
NA_Sales     1.13040    0.03162  35.745  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.49 on 172 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.9664
F-statistic:  2504 on 2 and 172 DF,  p-value: < 2.2e-16
```

Figure 3 Multiple regression for Global_Sales based on NA_Sales and EU_Sales

Relationships between NA_Sales, EU_Sales, and Global_Sales were explored through the multiple regression analysis above. There was a positive correlation between Global_Sales and EU_Sales, and Global_Sales and NA_Sales.

```
# Function to calculate predicted Global_Sales_Sum.
Global_predicted <- function(NA_Sales_sum, EU_Sales_sum) {
  Global_predicted_sum <- coef(GL_EU_NA)[3]*NA_Sales_sum +
    coef(GL_EU_NA)[2]*EU_Sales_sum +
    coef(GL_EU_NA)[1]
  return(Global_predicted_sum)
}

# a) NA_Sales_sum = 34.02 and EU_Sales_sum = 23.80
Global_predicted(34.02, 23.80) # Observed value =67.85, predicted value = 68.06

# b) NA_Sales_sum = 3.93 and EU_Sales_sum = 1.56
Global_predicted(3.93, 1.56) # observed = 6.04, predicted = 7.36

# c) NA_Sales_sum of 2.73 and EU_Sales_sum of 0.65
Global_predicted(2.73, 0.65) # Observed = 4.32, predicted = 4.908

# d) NA_Sales_sum of 2.26 and EU_Sales_sum of 0.97
Global_predicted(2.26, 0.97) # Observed = 3.53, predicted = 4.76

# e) NA_Sales_sum of 22.08 and EU_Sales_sum of 0.52
Global_predicted(22.08, 0.52) # Observed = 23.21, predicted = 26.626
```

The screenshot above shows that the predicted value from the multiple regression analysis isn't too far off the actual observed value. This shows that the model is a good predictor.

To conclude, there is evidence to suggest loyalty_points of customers is dependent on remuneration and spending_score. The higher the spending_score, the higher the loyalty_points, and the larger the remuneration, the higher the loyalty_points. I could have done a multiple linear regression analysis to give further insight.

The segmentation of customer base can be used by marketing teams to group similar customers, and intentionally target groups to maximise profit.

Social data can be used by marketing campaigns to get an indication of the products they can use to attract customers, such as games which is a popular word used in reviews. They can also use this to get an overall idea of how well the company is doing and improve on areas by utilising negative reviews as constructive feedback.

Products with the smaller number code sell more everywhere.

The data set are non-normal data points; however, this does not make it unreliable. It can be transformed to a normalised dataset by using logarithmic transformation.

Finally, the sales data have a positive correlation between one another. Global_Sales is positively affected by NA_Sales and EU_Sales. So, the higher the NA_Sales, and EU_Sales,

the higher the Global_Sales. To increase Global_Sales further, NA_Sales and EU_Sales would need to be increased.