[*Venson Chou*](https://unsplash.com/@venson_chou) on[Unsplash](https://unsplash.com/?utm_source=bulb&utm_medium=referral)

**Advanced Analytics - 'Turtle Games' Case Study**

[**HMHaroon Miah**](https://www.bulbapp.com/haroonmiah/)

In this case study, I was provided with **data**from a manufacturer and retailer called **Turtle Games**. They have a **global customer base**, selling various products including **books, toys, board games** and **video games**. Turtle Games collect their own **sales data** as well as **customer review** data and they have tasked us to leverage this data to provide **insights**into their overall **sale performance** and how this can be improved.

For this assignment, I had to use **R/RStudio** and **Python**to conduct the data analysis and demonstrate my findings.

Other key areas to analyse:

* how customers accumulate loyalty points
* how groups within the customer base can be used to target specific market segments
* how social data (e.g. customer reviews) can be used to inform marketing campaigns
* the impact that each product has on sales
* how reliable the data is (e.g. normal distribution, skewness, or kurtosis)
* what the relationship(s) is/are (if any) between North American, European, and global sales.

Data files provided:

1. **turtle\_reviews.csv** includes details on customer gender, age, remuneration, spending\_score, loyalty\_points, education, language, platform, review and summary across products.
2. **turtle\_sales.csv** includes details of video games sold globally, such as the rank, product, platform, genre, publisher, and their sales across North America, Europe, and worldwide.

Firstly, to understand how customers accumulate **loyalty points**, we will investigate **relationships**with **spending scores, renumeration** and **age**. **Linear regression** is a type of predictive analytics tool that can help demonstrate the relationship between a dependent variable (y or loyalty points) and an independent variable (x or spending scores/renumeration/age).

Figure 1: OLS Regression results and Scatter plot to show Spending Score vs Loyalty points with Regression line.

Figure 2: OLS Regression results and Scatter plot to show Renumeration vs Loyalty points with Regression line.

Figure 3: OLS Regression results and Scatter plot to Age vs Loyalty points with Regression line.

As we can see from figures 1-3, the r-squared values are very low (the closer to 1, the better) for all three models; 0.45, 0.38 and 0.002, repectively. This tells us that only 45%, (for spending score vs loyalty points, in figure 1), of the total variation in the data is explained by the regression model.

In the scatter plots, there is a slight positive relationship or linearity between spending score vs loyalty points and renumeration vs loyalty point; the regression lines help portray this also. However, there is no relationship whatsoever for age vs loyalty points and this shouldn’t be pursued further.

As many variables are involved, it may be worth considering a multiple linear regression model instead.

Building on from the work we have done with the linear regression models, let’s focus on the usefulness of renumeration and spending scores to identify groups within the customer base that can be used to target specific market segments. We can use k-means clustering for this.

To determine the optimal number of clusters to use from our dataset, we can use the Elbow or Silhouette methods, which suggested between 4-6 clusters to test out.

Figure 4: The Elbow method to determine optimal number of clusters.

Figure 5: The Silhouette method to determine the optimal number of clusters.

My interpretation of all the above methods/models suggest that 5 clusters provide the most distinct separation and grouping. This is one of the drawbacks to unsupervised models such k-means clustering; results can be subjective and if you are not experts in the field, you do not know for sure what the labels should be.

Figure 6: k-means model using 5 clusters.

You can see from the k-means model in Figure 6 that the clusters/groups are well defined and there isn’t a huge overlap with the data points from the different clusters.

* Cluster 1 (red) -> higher income group but spending more
* Cluster 2 (green) -> average income group and average spending group
* Cluster 3 (blue) -> higher income group but spending less
* Cluster 4 (black) -> lower income group but spending more
* Cluster 5 (orange) -> lower income group but spending less

To better understand the customer reviews, we can use Natural Language Processing to conduct sentiment analysis to identify the 15 most common words used and also the top 20 positive and negative reviews received.

Once the data was imported and cleaned up, we were able to tokenise the words and create a word cloud.

Figure 7: Word cloud of the online customer reviews.

We can see from the word cloud, the words that stand out are predominately positive. To confirm this, we were able to plot the top 15 most commonly used words (Figure 8) and then create a histogram to show the sentiment score polarity (Figure 9).

Figure 8: Count plot for the top 15 words used in customer product reviews.

Figure 9: Histogram of sentiment score polarity of reviews.

We can see that these do indeed confirm a positive sentiment towards the products and the polarity is beyond the neutral and positive ends of the scale.

Finally, we were able to show the top 20 positive and negative reviews. One of the issues with using NLP is that the interpretations of the words/reviews used can be incorrect. For instance, if we look at the list of the top 20 negative reviews (Figure 10), we can see that some of the reviews are actually positive but the model has incorrectly grouped them in the negative group.

Figure 10: List of top 20 negative sentiment reviews with positive reviews mixed.

In relation to the sales data, it’s difficult to determine the impact a product has on the sales based on the data that we have available. We only have the data for the year the product was released and not trend data of sales over time. However, grouping the data by “Product”, we are able to identify the highest and lowest selling products in the 2 regions. This will help the marketing team to tailor their marketing efforts on the right products to improve sales.

Figure 11: Bar chart to show the highest selling products in £millions for North America

Figure 12: Bar chart to show the highest selling products in £millions for Europe

Figure 13: Boxplot to show the number of games sold (£millions) for each region.

We can see from Figure 13 that North America has a higher median number of sales compared to Europe and all regions have many outliers. The Histogram helped to visualise this also.

Figure 14: Histogram to show the distribution for Global sales.

Figure 14 shows that the data is skewed to the right and so the few larger values raise the mean higher than the median. This is not a normal distribution. We can confirm this by running the Shapiro-Wilk test; skewness and the kurtosis functions.

Figure 15: Shapiro Wilk test results for all regions

Figure 15 shows that the p-value is less than 0.05 for all three regions, meaning that the null hypothesis should be rejected and therefore the sales data is not normally distributed.

Figure 16: Skewness and Kurtosis scores for all regions

Figure 16 shows us that the skewness values are positive indicating that the tail is on the right side of distribution and the Kurtosis is leptokurtic (heavy-tailed), meaning that data produces more outliers than a normal distribution.

If data is not normally distributed it could impact any machine learning models that we use for predictive analytics. We would need to remove any outliers, but this could reduce the data sample and so it would be best to get a larger data set so that we can get a more distribution.

To confirm if there are any relationships between North American, European, and Global sales, we are able to create a linear regression model.

Figure 17: Using cor() function to create correlations between the variable, North American, European, and global sales

The cor() values show us that there is a positive relationship between all 3 regions but the strongest correlation (91.6%) is between North American Sales and Global Sales. By plotting the relationships, we are able to visualise this relationship (Figure 18).

Figure 18: Scatter plot for the correlation between North American Sales and Global sales (£millions)

As there are multiple variables, a multiple linear regression model is more appropriate to assess the effect of the variables on the dependent variable.

Figure 19: Using cor() function to show the correlations between the North American, European, and global sales in a multiple linear regression model.

Figure 20: Summary of the multiple linear regression model.

The summary statistics for the multiple linear regression show us that we can be very confident with our model; the p-values are very low (<0.05), and the adjusted R-squared value is very high (96.6%). This tells us that EU\_Sales and NA\_Sales are highly significant variables in relationship to Global sales.

We can use this model to test its ability to predict the global sales and compare with what was actually observed.

Figure 21: Using our multiple linear regression to compare predicted global sales values to what was actually observed.

From our analysis, we have been able to identify the **key groups of customers** shopping at Turtles Games, based on **renumeration**and **spending scores** and the **best groups** to **target**to boost sales (Groups 1 and 0 from Figure 6).

Our **sentiment analysis** demonstrated that customers had predominantly **positive**sentiments towards the company and the sentiment polarity distribution was more neutral to positive. We were able to identify the highest and lowest selling products for all regions and so Turtles Games will be better informed when it comes to marketing so that they promote the most popular products.

In terms of relationships relating to sales data, our analysis showed that there was a positive relationship with North American, Europeans sales and Global Sales; although, **North American Sales**had the strongest relationship with Global sales and so this would be a key market to focus on also. We were able to predict the Global sales with very strong accuracy using a multiple linear regression model. Questions were raised regarding the distribution of the data and so to further our analysis, we would suggest getting data from a bigger sample that is more representative of the normal distribution or data for a longer time frame.

Shared February 21, 2023

0 Comments

0Likes

**[PREV](https://www.bulbapp.com/u/data-analytics-on-covid-19-vaccination-data-using-python)**

**[Data Analytics on COVID-19 Vaccination data using Python](https://www.bulbapp.com/u/data-analytics-on-covid-19-vaccination-data-using-python)**

**[VIEW COLLECTION](https://www.bulbapp.com/u/lse-data-analytics-career-accelerator)**