**Turtle Games Analysis**

**Author:** Matt Rose

**Date:** Thursday 8th September 2022

**Version**: 1.0

## Version

|  |  |  |  |
| --- | --- | --- | --- |
| **Author** | **Version** | **Date** | **Comments** |
| Matt Rose | 0.1 | 8th September 2022 | Initial Draft |
| Matt Rose | 1.0 | 8th September 2022 | Final Version |

Table of Contents

[Version 2](#_Toc113539751)

[Scenario 3](#_Toc113539752)

[Approach & Insights 3](#_Toc113539753)

[Accrual of Loyalty points 3](#_Toc113539754)

[Customer Clustering 4](#_Toc113539755)

[Social Sentiments 6](#_Toc113539756)

[Product Effect on Sales 8](#_Toc113539757)

[Data Reliability 10](#_Toc113539758)

[Relationship between Sales and Location 12](#_Toc113539759)

[Additional Exploration 12](#_Toc113539760)

**Scenario**

Turtle Games is a games manufacturer and retailer and supply customers on a global scale. They manufacturer and sell their own label games whilst also acting as an intermediary for other games manufacturers, selling products which they purchase from competitors.

Through the analysis of available data, Turtle Games stakeholders want to understand the following:

* 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
* Impact on sales by product
* How reliable the data is
* What the relationship is (if any) between geographical locations and global sales.

**Approach & Insights**

### Accrual of Loyalty points

Initial Analysis using linear regression techniques against the Turtle Games data shows that there is clear linear relationships between loyalty points, spending score and salary, however there is probably deeper analysis needed by implementing multiple linear regression to show even further trend analysis between the columns.

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated | Chart, scatter chart  Description automatically generated |
| Chart, scatter chart  Description automatically generated | |

Figure - Graphs to show the relationships between Loyalty Points, age, spending score and remuneration

There was no linear relationship between loyalty points and age. This would be useful to use as an independent variable in a multiple linear regression to see if it has any influence on the model when coupled with other independent variables.

When plotting the datapoints and the regression line on each model, it is interesting how the higher that the data moved in the independent variable, the more erratic the data becomes with values showing both below and above the trend line. Ideally a further data cleanse would be performed to remove any outliers outside of the IQR which may drive a more accurate model and trend line.

### Customer Clustering

To establish groups within the customer base which can be targeted for specific marketing campaigns, we first use k-means clustering. This allows the ability to determine the optimal number of customer clusters.

Prior to running the predictive models, the plot of the score and renumeration data led to a scatterplot which initially implies a 5-cluster model.

Chart, scatter chart

Description automatically generated

Figure - Scatterplot of remuneration and score

Continuing the analysis to determine usefulness, the use of the Elbow and Silhouette methods shows that we should consider 4, 5 and 6 clusters as this is where the line curves before trailing.

|  |  |
| --- | --- |
| Chart, line chart  Description automatically generated  Figure - Elbow Method | Chart, line chart  Description automatically generated  Figure - Silhouette Method |

Following the analysis of each of the different numbers of clusters outlined from the above figures, the resulting optimal cluster is deemed as 5 and is shown below. This clearly indicates the various clusters of customer salary vs score meaning that Turtle Games could specifically market certain products to specific demographics of customer.

Chart, scatter chart

Description automatically generated

Figure - scatter plot of 5 cluster model

### Social Sentiments

Consumer data and opinion has never been more powerful than it is today. Using Python, we can get to the heart of customer sentiments and provide the ability to analyse how Turtle Games is seen as well as how the products are reviewed. This is important data to allow Turtle Games to react, pre-empt and identify improvement processes in their business to deliver as high a customer satisfaction as possible – thus leading to repeat business.

The analysis has been performed over a series of review and summary data, stripping out common ‘non-useful’ words to deliver a word cloud of the most frequently used words that Turtle Games customers are using. As is shown below, most of the words used are seemingly positive, with words like ‘game’, ‘great’ and ‘fun’ being used in a lot of the reviews. Whilst this is a good start initially as it identifies the most frequent words, further analysis on the more negative words would be very helpful in determining what the biggest challenges that customers face.

Text

Description automatically generated

Figure - Word cloud from customer reviews

In addition to the word cloud, we have been able to determine the sentiment scores as well as identify the top and bottom 20 reviews for Turtle Games. Plotting the sentiment scores on a histogram shows that we have majority of positive or neutral sentiment, with only a few showing as negative sentiment. This shows that Turtle games is, for the most part, doing a good job with its customers, but further analysis can be performed on the negative sentiments to identify improvement areas.

Chart, histogram

Description automatically generated

Figure - Review sentiment histogram

Graphical user interface, text, application

Description automatically generated

Figure - Top 20 reviews based on sentiment

Table

Description automatically generated with medium confidence

Figure - Bottom 20 reviews based on sentiment

### Product Effect on Sales

To undertake the analysis of the effects of Products on Sales, there was a switch from Python to R to make the most of the statistical libraries that R can access. The initial plot of Products against Global Sales show that the lower, more original products with lower product IDs contribute far more to Global Sales than newer products with higher Product ID’s. **This is predicated on the assumption that lower product IDs are the original product ranges of Turtle Games.**

Chart

Description automatically generated

Figure - Product vs Global Sales

In addition, we can see the relationship where the products have the biggest impact in EU and NA Sales, which follow the same trajectory as the Global Sales. This indicates that both EU and NA Sales follow the global pattern and that there are consistent products that Turtle Games sells which are globally adopted.

Chart, scatter chart

Description automatically generated

Figure - Products sold vs EU Sales

Chart, scatter chart

Description automatically generated

Figure - Products sold vs NA Sales

When continuing with the analysis we can also see which platforms also contribute the most to the global sales with also displaying outliers – in this instance NES, GameBoy and Wii have been the most impactful on Global Sales.

Chart

Description automatically generated

### Data Reliability

Fundamentally, the key question that needs to be answered is “how reliable is the data?”

In order to provide the answer to this question, we turn again to R and identify data normality and skewness using a variety of statistical methods. We focused primarily on QQ (quartile-quartile), Shaprio-Wilk, skewness and kurtosis tests to identify how normalised the data is and whether it can be trusted for the purpose of the analysis we are performing.

The QQ plot shows that when we look at the EU, NA and Global Sales they all show that the data aligns to some sort of normality due to the fact that there isn’t a large deviation from the trendline. That said, as the sales grow, the deviation becomes larger suggestion a move away from normalised data.

Chart, line chart

Description automatically generated

Figure - Global Sales QQ Plot for normalisation

The Shapiro-Wilk test however, gives us an interesting outcome. The p value is showing as significantly less than 0.05 which means that the data is not normally distributed for Global Sales.

Text

Description automatically generated

Skewness and Kurtosis values are shown as follows for NA, EU and Global Sales, this indicates a data set which is highly peaked, which moves away from normalised data and shows there is a right-skewed output of this data. Further work on the data set should be undertaken to bring normalisation and skewedness back towards the 0-centre line.

Text

Description automatically generated with medium confidence

### Relationship between Sales and Location

To determine the relationship between NA and EU sales to Global Sales, we ran several linear and multiple regressions to show that there is a high correlation between the NA and EU sales and Global Sales.

Having run the linear models for EU and NA Sales against Global Sales we see that there is an R-squared value of 77.01% and 87.4%. When running a multiple linear regression with a combination of both NA and EU sales, this R-squared value jumps to 96.8%. This means that a significant number of NA and EU sales can be used to explain the Global Sales, so the combination of both geographical regions make up most of the Global Sales.

It would be interesting to see how the EU and NA sales break down further into countries/states.

**Additional Exploration**

* Multiple linear regression to be performed against the loyalty points to see if there is a combination which more accurately predicts the combination that will enhance loyalty point acquisition.
* Breakdown of EU and NA states/countries to have a more granular view of where the sales make up is and where to target campaigns
* Cleanse the data further to get a better, more normalised data set.
* Run a multiple linear regression of loyalty points, using score, age and remuneration to see if the age has any additional influence over loyalty points.