**Executive Summary**

To improve overall sales performance, Turtle Games has come up with an initial set of questions. You’ll explore these questions in greater depth through the weekly assignment activities. Turtle Games wants to understand:

Approaches to support the answer analyzing data given from different aspects:

|  |  |
| --- | --- |
| 1. Data Source Exploration | 1. how customers accumulate loyalty points  * read the meta data for definition of scores * *analyse the descriptive statistics from the data frame*  1. *investigate trend and patterns against different numerical and categorical variables* |
| 1. Relational analysis between individual definitions of data | 1. how groups within the customer base can be used to target specific market segments:  * *Regression analysis* * *KMeans clustering for unlabeled classifier* * *Natural Language Process (NLP) to transform textual data into sentimental definition for quantitative analysis*  1. *Multi-method approaches:*  * *OLS, Random Forest Regressor* |
| 1. Trends and Patterns | The impact that each product has on sales   * Data type and natures: normality, skewness, Kurtosis * Model by Ordinary Least Square method * Reliable of data  1. Multiple Regression 2. Breusch Pagen Test 3. VIF comparison 4. Hypothesis Testing 5. Goodness of Fit test |
| 1. Insights and Explanations | 1. Relationship(s) is/are (if any) between North American, European, and global sales 2. Utilise social data (e.g. customer reviews) for marketing campaigns |

1. **Data Source Exploration**

Two data analytic platform would be used for the whole analysis:

* Python
* R

Data frame Summary:

1. 2 data frames shared common key - Product IDs

2. No missing data in the files.

3. Number of observations:

turtle\_reviews: 2000

turtle\_sales: 352

Data Type Check:

Product column contains the unique product codes such that there is no ordinal meaning relating to the product descriptions but saved as numerical type in the turtle\_reviews.csv file.

Same product ID would have several Ranking in turtle\_sales.csv file.

Data Exploration

1. Categorical Data

A graph of different colored columns

Description automatically generated

Observations:

* The plot indicated an interesting pattern that the samples whose education background of 'Basic' within age 30-35 would have highest loyalty points.
* While there were a number of samples with educational background of graduate level or above falls into 15-20 age group. Generally speaking, the age at which most people receive their PhD varies from person to person, but typically falls between the mid-twenties and early thirties.

1. Numerical Data

A graph with a line

Description automatically generated

A diagram of a graph

Description automatically generated

* The range between min and max of the loyalty points was quite large, suggested checked if transformation of the variable could give better result of estimation.
* The ranges for age, remuneration and spending score were quite close.

A graph with blue dots

Description automatically generated

* There was some pattern identified between Annual Income and spending scores.

1. **Trends and Patterns**

Meta Data Interpretation

* According to the meta data, loyalty points are the score based on the point value of the purchase, converting the monetary value to point value, and the point value of an action (purchase).
* Since the purchase of any goods would be influenced by the spending behavior as well as the income level of any individual, which were both available in our given data set.

1. *Patterns Identified with Loyalty Points:*

* Correlation Matrix

A red and white squares with numbers

Description automatically generated

* *The loyalty points were highly correlating to spending scores and remuneration; while with age did not because its correlation coefficient was near to zero.*
* *OLS Regression Analysis*

|  |  |
| --- | --- |
| *A graph with blue dots  Description automatically generated* | *A graph with blue dots  Description automatically generated* |
| Estimated coef:  (const): -75.052663 (sp\_score) 33.061693  standard errors:  (const): 45.930554 (sp\_score) 0.814419  *R-Squared:* 0.452 | Estimated coef:  (const): -65.686513, (annual\_inc(k£)): 34.187825  standard errors:  ( const): 52.170717 (annual\_inc(k£)): 0.977925  *R-Squared:* 0.380 |
| A graph with a red line  Description automatically generated |  |
| Estimated coef:  (const): 1736.517739 (age) -4.012805  standard errors:  (const): 88.248731 (age) 2.113177  *R-Squared:* 0.002 |  |

*Further Suggestion:*

*The R-square for all of above regression models were a bit low, such that it might be not accurate enough for predictive analysis.*

*In the following section, we would adopt a machine learning model – an ensembled method by Random Forest Regressor.*

1. Unlabeled Classifier

K-Means Clustering Method- *Determine possible number of clusters:*

|  |  |
| --- | --- |
| A graph with blue dots  Description automatically generated | A graph with blue lines and numbers  Description automatically generated |
| Both methods suggested the possible number of clusters should be 5 and the following viz was quite clear for to justify this decision. | |

A graph of different colored dots

Description automatically generated

No. Obs

0 774

3 356

2 330

1 271

4 269

**Clusters Effect on variables**

**A screenshot of a graph

Description automatically generated**

* Further Insight with K-Means Cluster

With the K-mean identified clusters did clearly identify pattern between loyalty points and other variables concerned which help for better prediction.

1. Textual Data Analysis

The approach to analyze textual data would be through Word Frequency and Sentiment Analysis to evaluate any insight from customers’ reviews.

* Word Frequency

The most common words were ‘game’, attributed to the fact that the reviews and summaries were concerned with the game they had bought.

Other words like ‘great’, ‘fun’, ‘play’, ‘love’, ‘good’, ‘like’, ‘nice’, were relating to customers’ general comments about the game purchased.

Some more specific words like ‘cards’, ‘tiles’,’book’, ‘kids’, would reveal some features about the game.

On the whole, the word frequency check did show a positive sentiments from the customers and it is valuable to convert it into a word cloud for the marketing concept about the products.

A bar graph with different colored bars

Description automatically generatedA bar graph with different colored bars

Description automatically generated

* Word Cloud for a general concept for marketing campaign

A close up of words

Description automatically generated

* Sentiment Analysis

In this section, two lexicon based processes: TextBlob Sentiment and Vader Sentiment, was used to analyse the textual data with scores unveiling the sentiment of the sentence:

Both processes produce an index ranging from -1 to 1 , 1 being positive and -1 being negative. We compare the two outcome with a histogram as below:

|  |  |
| --- | --- |
| A graph of a bar graph  Description automatically generated | A graph with blue and white bars  Description automatically generated |
| TextBlob Sentiment Score Distribution (Reviews) | Vader Sentiment Score Distribution (Reviews) |

The last comparison of the top 20 positive and negative sentiment scores with the Review contents, it seems sometimes Textblob was making mistake when referring to the actual texts but Vader was more accurate to differentiate between positive and negative meanings.

* Finally we joined the turtle\_sales.csv files by the product ID and analyze the average sentiment scores against each product’s publisher which pointed out that most customer comments were relating to Nintendo’s products and most of them enjoyed the purchase.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | vader\_sentiment | vader\_sentiment | tblob(review) | tblob(review) |
|  | mean | count | mean | count |
| Publisher |  |  |  |  |
| Nintendo | 0.64026918 | 756 | 0.219660222 | 756 |
| Sony Computer Entertainment | 0.668585475 | 179 | 0.216268436 | 179 |
| Electronic Arts | 0.687925333 | 150 | 0.268538894 | 150 |
| Microsoft Game Studios | 0.610619333 | 150 | 0.205040172 | 150 |
| Activision | 0.645052273 | 132 | 0.212859455 | 132 |
| Take-Two Interactive | 0.627749398 | 83 | 0.249989098 | 83 |
| Ubisoft | 0.653512857 | 70 | 0.223374745 | 70 |
| Capcom | 0.6683125 | 40 | 0.278740186 | 40 |
| Square Enix | 0.5631925 | 40 | 0.114739997 | 40 |
| Konami Digital Entertainment | 0.639156667 | 30 | 0.248419012 | 30 |

1. Random Forest Regressor

* An importance feature indication contributing to the loyalty points found that the spending score from the customer was the most important one, and then following the cluster group number 3 from our KMean’s prediction.
* Mean Absolute Percentage Error: 0.35%
* RMSE - Mean Absolute Error= 0.053769986078316956 - 0.022741451674507064 = 0.031
* [RMSE - Mean Absolute Error] close to 0 indicate that there are no large errors in the forecast

A graph with blue squares

Description automatically generated

1. The impact that each product has on sales

* Data type and natures

A graph with red and blue dots

Description automatically generated

* The sales of games for both EU & NA region mostly occurred with product sales below 3 millions pounds with similar pattern.
* EU sold more than NA on average

A graph with red lines

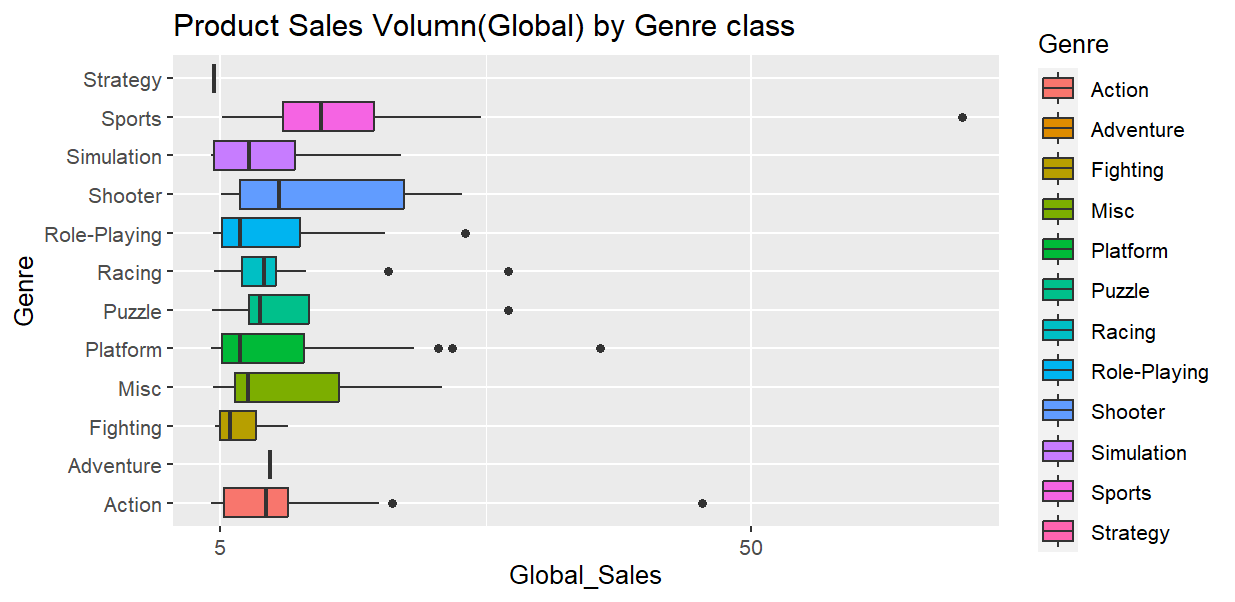
Description automatically generated

* Global sales distributed evenly and regional data showed skewness towards smaller per product sales amount.

A graph with a line and a red line

Description automatically generated

* Products launched with multiple platforms showed more global sales on average compared to the single platform.



The average Global Sales for the top 3 genre games were : Sports, Shooter and Puzzle.

* Analytical model: Ordinary Least Square

How reliable the data is (e.g. normal distribution, skewness, or kurtosis):

* Statistical test:
* Multiple Regression
* Breusch Pagen Test

The p-value is 0.1718 which means failed to reject homeskedasticity

* VIF comparison

EU\_Sales: 1.627488

NA\_Sales: 1.627488

Both VIF indicated there were not multicollinearity in the model

* Hypothesis Testing

Null Hypothesis: EU\_Sales = 0 & NA\_Sales = 0

The F-statistics with P-value <2.2e-16, so it rejected the null and the model should be used.

**Observations and Insight:**

The model showed statistically significant reliability but the product launched with platforms would also increase Global Sales and thus would be preferred over the model just based on sales data from Europe and North America region.