COURSE 3
ADVANCED ANALYTICS FOR ORGANISATIONAL IMPACT

Predicting Future Outcomes

ANAND JOSHI

24/07/2023
Word count (excluding “Contents”, “References”, “Appendix”)**: 1137**

Contents

Context

Business Scenario

Turtle Games are a game manufacturer and retailer with a global customer base. They aim to improve their overall sales performance by leveraging customer trends, which includes analysing data from sales and customer reviews of their own products as well as those sourced from other companies.

Business Problem

Turtle Games needs to optimise their sales performance by leveraging customer loyalty points, identifying target market segments, utilising social data for marketing campaigns, understanding the impact of products on sales, assessing data reliability, and analysing the relationship between North American, European, and global sales. Additional questions that will be considered in the analysis include:

- What are the key factors influencing customer loyalty and how can Turtle Games enhance their loyalty program to drive repeat purchases and increase customer retention?
- How can Turtle Games effectively leverage customer reviews and social data to identify emerging trends, preferences, and sentiments to inform product development and marketing strategies?

Analytical approach

Python in Jupyter notebook

1. Data Import and cleaning: Libraries used: pandas for data manipulation, and numpy for mathematical operations. Functions used: 'pd.read' and to import data as pandas data frames. Employed 'info()', 'shape()', 'dtypes()', describe(), and isnull().sum() functions to ensure data integrity.
2. Customer Segmentation: Scikit-learn's K-Means: Performed K-means clustering to segment customers into distinct market segments.
3. Social Data Analysis for Marketing: NLTK (Natural Language Toolkit): Utilised for tokenisation, stop words removal, and sentiment analysis. Created a word cloud to visualise frequent words in customer reviews.
4. Visualisation: Employing data visualisation libraries like Matplotlib to create meaningful scatterplots, boxplots, and histograms to visualise trends within the customer base.

Extensive libraries and functions were used to conduct the analysis for this project. (See Jupyter notebook provided for the full in-depth pythonic code analysis of the above points).

R script in RStudio

1. Impact of Each Product on Sales: To determine the impact of each product on sales, the **dplyr** library was used. Grouping the sales data by product and using the **sum()** function to calculate the total sales. The **ggplot2** library was utilised to visualise the sales distribution using histograms, box plots, and scatter plots, gaining insights into the sales performance of each product. Additionally, a summary was created, enabling a clear understanding of the sales impact of individual products.
2. Data Reliability and Distribution: To assess the reliability of the data, we performed data cleaning and descriptive statistics. Missing values and inconsistencies were handled using functions like **na.omit()** and **unique()**. Using the **skewness()** and **kurtosis()** functions to assess skewness and kurtosis, which indicated that the sales data was not normally distributed but exhibited leptokurtic distributions with heavy tails.
3. Relationship Between North American, European, and Global Sales: To investigate this relationship, correlation analysis was used. The **cor()** function calculated the correlation matrix, and **ggplot2** produced scatter plots with trend lines to visualise the correlations between sales columns. The correlation values indicated positive correlations between the three variables. Scatter plots visually represented the relationships, showing how sales in different regions tend to move together concerning total global sales.

Many other libraries and packages were also used to conduct the analysis for this project. (See R script provided for the full in-depth R code analysis of the above points).

Visualisation and insights

Visualisations

Rational:

- Scatterplot – Visualise data that shows the relationship between two quantitative variables measured for the same individuals, i.e., EU sales vs NA sales.
- Histogram – Represents grouped data. Frequency density is used to plot a histogram, i.e., polarity and sentiment scores.
- Boxplot – Displays important statistical measures such as the median, quartiles, and potential outliers. Boxplots are particularly useful for detecting skewness, variability, and the presence of outliers in the data, i.e., global sales by product.

Insights

Scatterplots:

- Scatterplots visually display the relationship, correlations or patterns between loyalty points and age, income, and spending score. These graphical representations allow for a quick understanding of how these factors interact and impact customer behaviour.
- The scatterplots comparing Europe sales, North America sales, and global sales against each other help in understanding the sales distribution across different regions. It provides insights into market trends, potential growth opportunities, and how each region contributes to overall global sales. (See Fig. 1, 2 , 3 in Appendix A).

Histograms:

- Histograms effectively showcase the distribution of polarity sentiment scores across customer reviews. By analysing the shape and spread of the histogram, the business can

gauge the overall customer satisfaction level and identify potential areas for improvement in products or services.

- Furthermore, it can aid in understanding the frequency and intensity of positive, negative, or neutral sentiments expressed by customers, providing a comprehensive view of customer feedback. (See Fig. 4 in Appendix B).

Boxplots:

- Boxplots offer a clear summary of the number of sales for each product, providing a visual representation of the median, quartiles, and outliers. This allows for easy comparison of sales performance across different products.
- By analysing boxplots, businesses can identify their best-selling products and pinpoint underperforming ones, enabling targeted strategies to boost overall sales. (See Fig. 8 in Appendix D).

These graphical representations offer valuable insights into customer behaviour, sentiment analysis, and product sales, facilitating the identification of areas for improvement and optimisation to enhance sales performance effectively.

Patterns and predictions

The results from the linear regression model indicate that 'age' has limited predictive power for 'loyalty points,' as the coefficient is negative and not statistically significant. However, 'Income' and 'spending score' are significant predictors, showing a positive relationship with 'loyalty points.' Customers with higher incomes and spending scores tend to have more loyalty points. The customer base can be divided into five distinct clusters based on income levels and spending behaviour, providing opportunities to target specific market segments effectively. (See Fig. 1, 2, 3 in Appendix A).

Social data analysis conducted reveals a generally positive sentiment, but also a diversity of opinions among customers, ranging from extreme negative to extreme positive. Most reviews fall within a moderately positive to neutral sentiment range.

Scatterplots of "EU Sales" against "NA Sales" shows a weak positive correlation, indicating that some products may perform better in one region compared to the other. The histogram of "Global Sales" exhibits positive skewness, indicating that there are more products with lower global sales than those with higher global sales.

In addition to this, the box plot of "Global Sales" by "Platform" identifies the gaming consoles "Wii," "NES," and "GB" as the most popular products in terms of global sales, with a few exceptional products as outliers. (See Fig. 8 in Appendix D).

Furthermore, the correlation matrix demonstrates positive correlations between North American and European sales with total global sales, implying that sales in these regions tend

to move together concerning overall global sales. The strongest correlation is observed between ‘Total_NA_Sales’ and ‘Total_Global_Sales.’ (See Fig. 5, 6, 7 in Appendix C).

The predictions from the model suggest that higher ‘NA_Sales_sum’ and ‘EU_Sales_sum’ led to higher predicted global sales, while lower values result in lower global sales predictions. Overall, the data and model support a positive relationship between North American, European, and global sales, where sales in one region are positively associated with sales in other regions. (See Fig. 5, 6, 7 in Appendix C).

References

- FourthRev, LSE_DA301_Advanced Analytics for Organisational Impact_C1_2023, Available from: LSE_DA301_Advanced Analytics for Organisational Impact_C1_2023 (fourthrev.com)

- Tutorial republic, Available from: Python Tutorials – Real Python

- Python Programming, Available from: Python Programming Tutorials

- R Tutorial, Available from: R Tutorial (w3schools.com)

- Introduction to R programming, Available from: R Programming Tutorial for Beginners - [# Updated 2023] (intellipaat.com)

Appendix

Appendix A

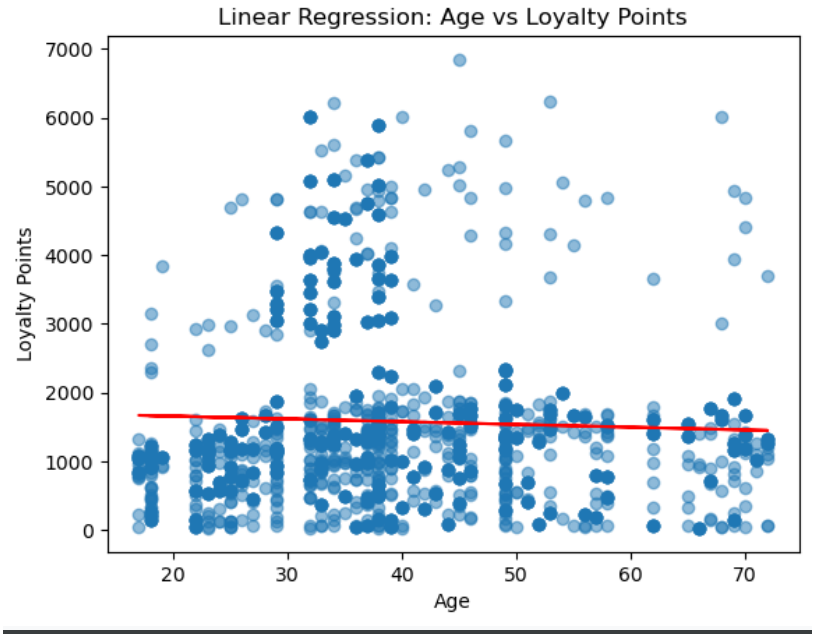

Fig.1: Linear Regression: Age vs Loyalty points.

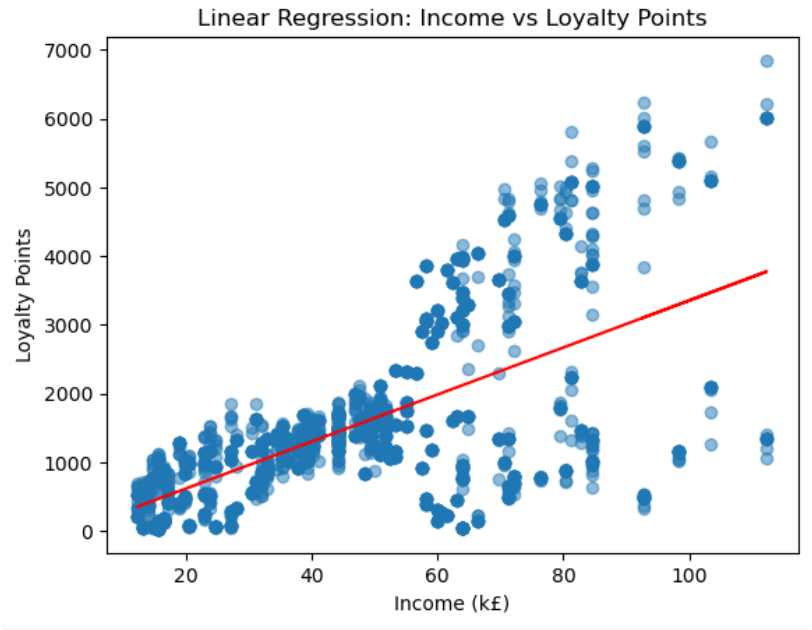

Fig.2: Linear Regression: Income vs Loyalty points

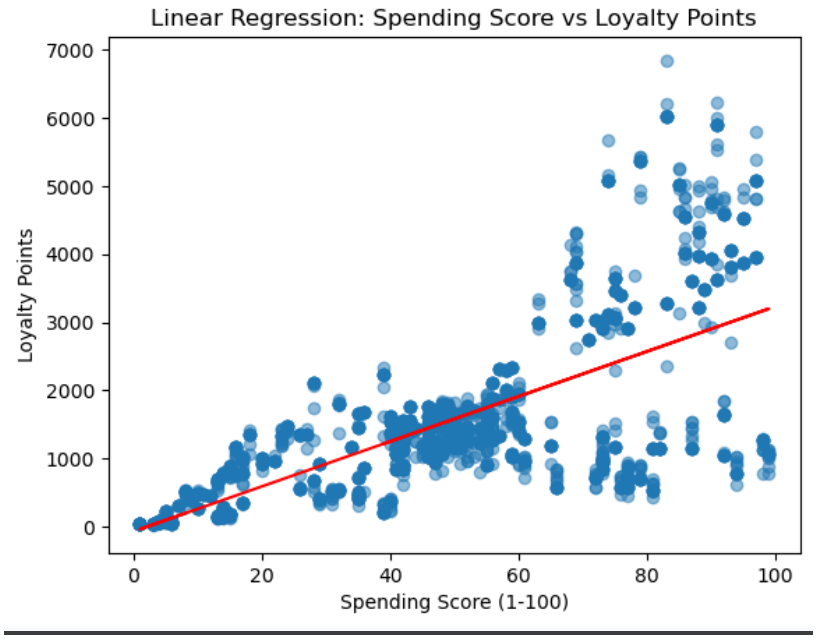

Fig.3: Linear Regression: Spending score vs Loyalty points

Appendix B

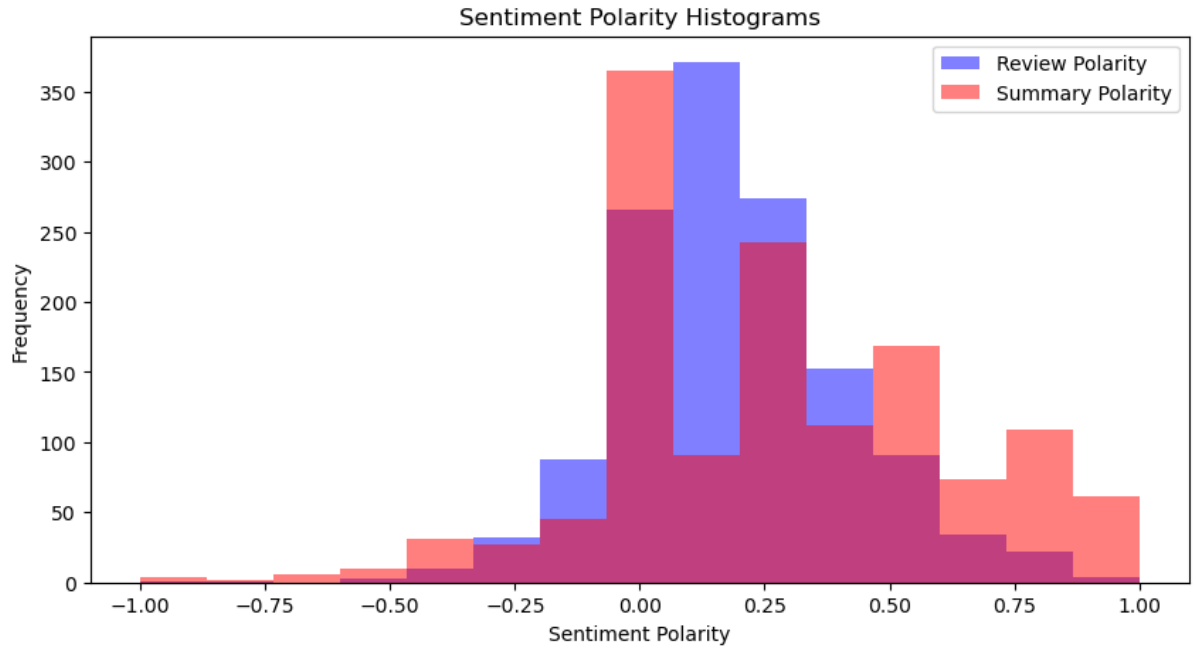

Fig 4. Sentiment Polarity for review and summary columns.

Appendix C

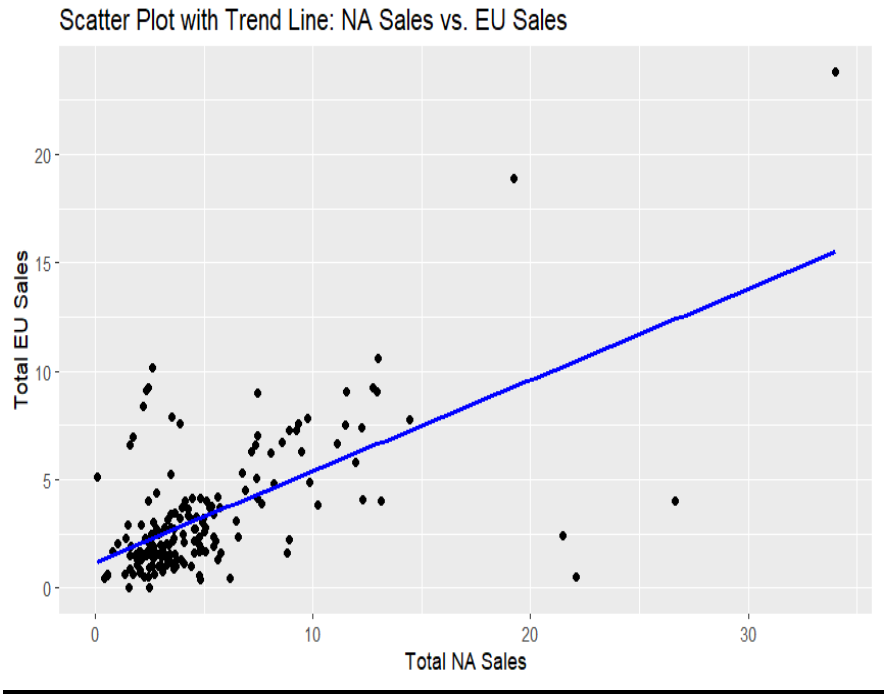

Fig. 5: Scatterplot with trend line: NA sales vs. EU sales

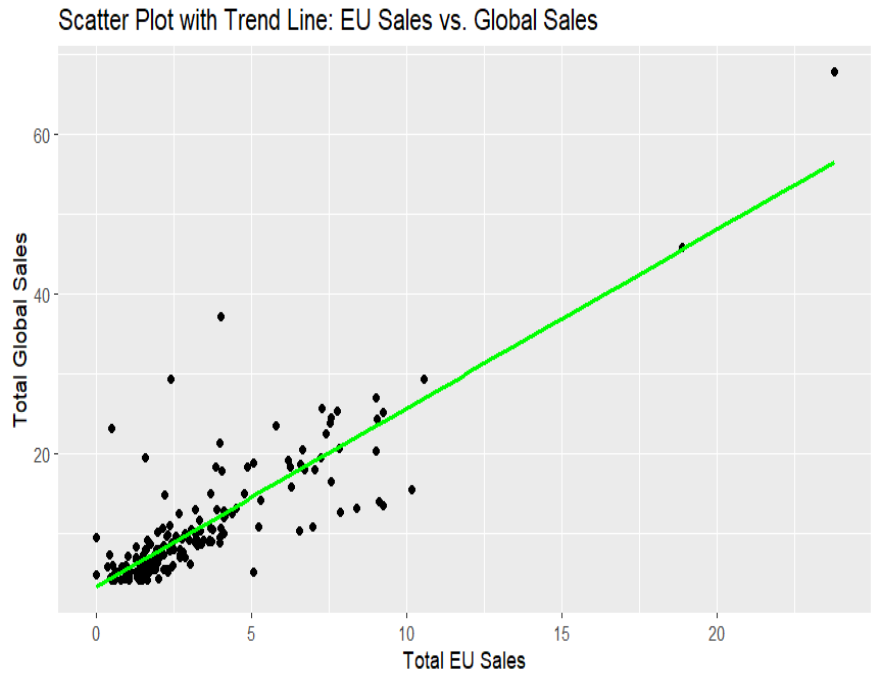

Fig. 6: Scatterplot with trend line: EU sales vs. Global sales.

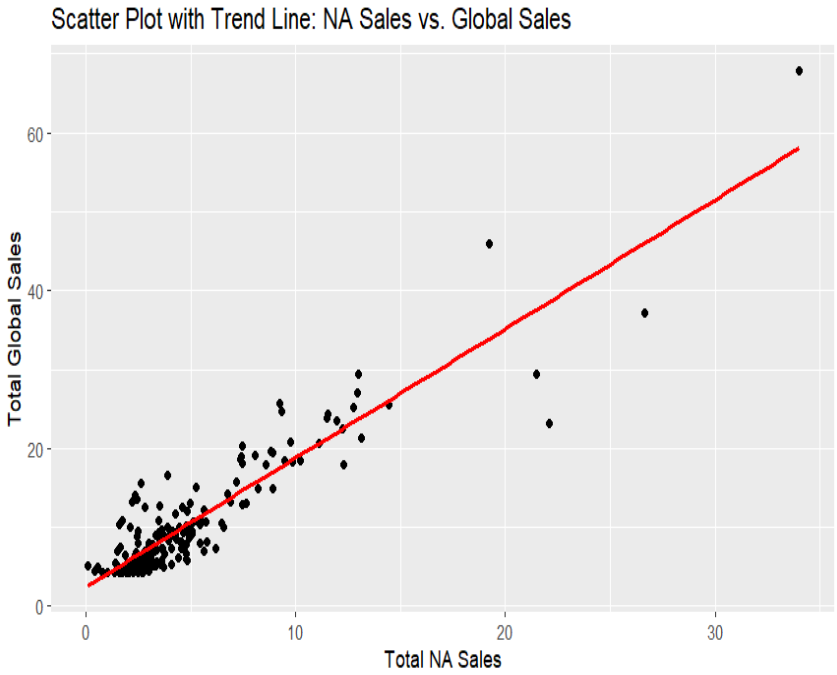

Fig 7: Scatterplot with trend line: NA sales vs. Global sales.

Appendix D

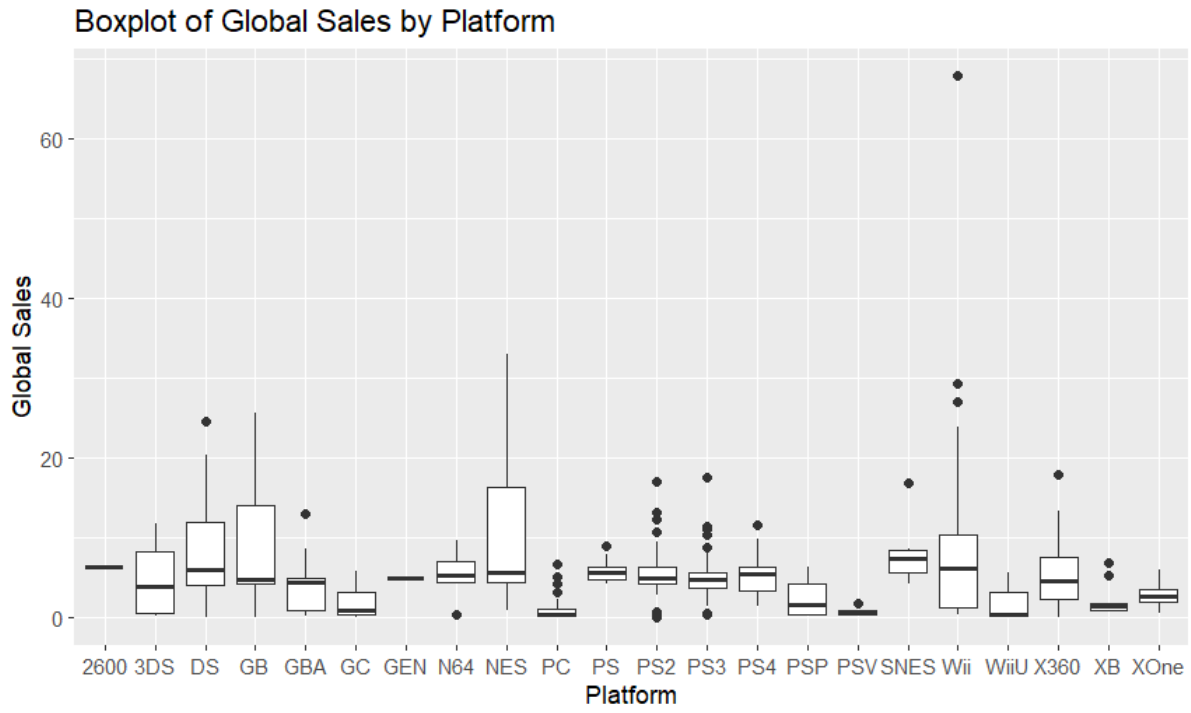

Fig. 8: Boxplot of Global sales by Platform.