ANALYSIS REPORT

Turtle Games

Erin John

REPORT

Background and Context.

Turtle Games, a gaming corporation, is eager to decode customer behaviour for product optimization. Among the provided variables, age, remuneration, spending points, and loyalty points are considered paramount. The problem statement that guides the data analysis is as follows "To determine the overall sales performance of Turtle games by understanding their existing sales data and customer trends". This data analysis unfolds in a way to look at the customer trends and the sales data through Python analysis and R script respectively. Thus, what has occurred can be categorized into two goal points of the analysis.

- Understanding customer trends through their engagement with the business. Variables used here are 'income', 'spending nature', 'age of customer' and 'customer reviews.
- Understanding the potential patterns in their sales performance across locations. Variables used are 'sales of various locations', 'genre of games' etc.

The next sections of the report will cover the process of data analysis in detail.

Analytical Approach
Cleaning and Preparing the data for exploratory analysis and the steps of analysis.

- Understanding the metadata, or the data dictionary, is a crucial first step in any analysis process. This information provides insight into the structure, meaning, and relationships of the data, helping to guide subsequent steps.
- Upon importing each dataset into R and Python, cleaning was performed using various functions such as "Drop", "Rename", and "Converting". These operations aimed to remove redundant or irrelevant information that could potentially hinder the data analysis process. For instance, columns such as "education" and "year" were deemed irrelevant to the company's interest in understanding customer behaviour and were thus removed.
- The "Group By" and "Aggregate" functions in R were used to tidy up the dataset and streamline the analysis process. This involved grouping data based on specific criteria and summarizing key metrics, enabling easier interpretation and comparison of results.
- In Python, the next step for analysis which was done is to create regression models between the income, age and spending nature of the customer alongside the loyalty points (a proxy for the purchases made by the consumer).This was done to bring out possible correlations between these variables. To further understand whether any of these or all of these affect the purchases made by the consumer.
- To further explore the underlying patterns within the data, the K-means method of clustering was employed. This technique groups similar observations together based on their features, helping to identify common trends or clusters within the dataset. By clustering customers based on their 'remuneration' and 'spending score' was done specifically in this analysis.

The rationale behind this was the company could potentially group a certain section of customers and focus marketing efforts on them specifically.

- Natural Language Processing of the reviews were done next. Understanding their sentiment can further improve customer retention for the company. When conducting NLP analysis libraries imported include "Wordcloud"."Textblob" and so on.
- In R-script, the analysis specially focused on the visualisation of the sales data. The rationale behind the R-script was to bring to the stakeholders an easier and digestible understanding of their current sales performance.

The next section covers a descriptive idea about the visualisations that were used for the analysis.

Visualisation of data.

- The first set of visualisations took place in the python notebook. The need for this arose initially in the regression analysis. While the OLS tables are quite descriptive, scatter plots will give a robust overview of the relationship between the variables. Throughout the regression analysis 'loyalty points' were in the y-axis as it is the dependent variable which will be insightful in understanding what affects the customer purchases. For the scatter plots the colour of the data points and the regression line were kept uniform to be able to maintain consistency in one section of the analysis.
- Another important visualisation is that of the cluster groups, which is colour coded and labelled to better understand the density of the customer groups within each zone.
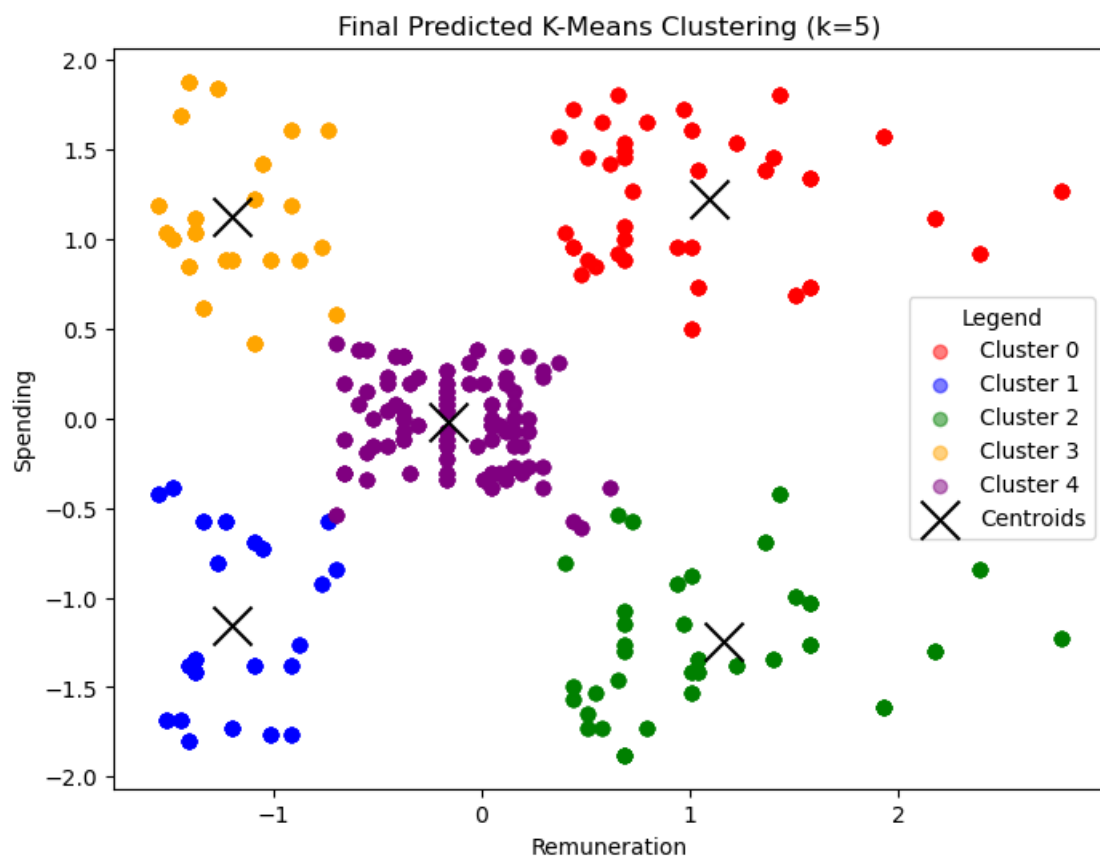
- In RStudio, I utilized data visualization tools like histograms, box plots, and scatter plots to glean insights into our sales data. Histograms allowed me to visualize the distribution of European sales, North American sales revealing trends such as skewness and central tendency. Box plots provided a deeper understanding of the sales distribution, highlighting measures like median, quartiles, and potential outliers.
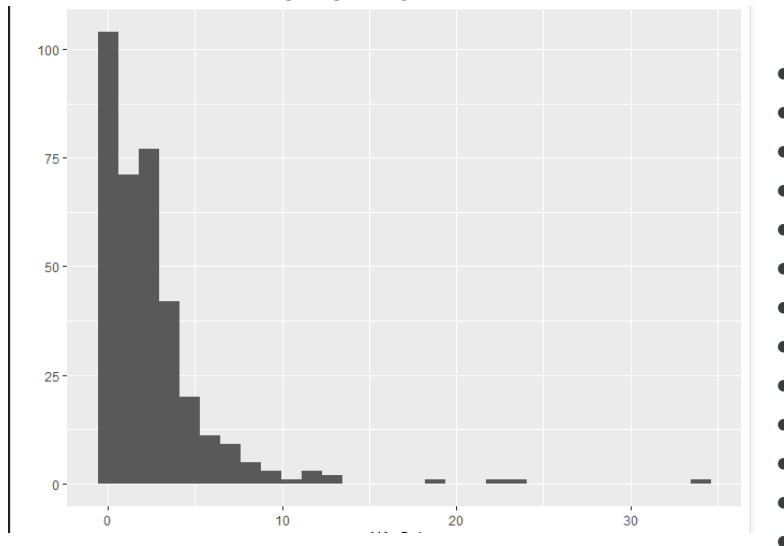
- For the NLP analysis there were two sets of visualised data that was produced. These were the word clouds to give a wholistic view about the overall sentiment that the customers possess towards the products. The analysis that can be understood from the word clouds were :
  The word cloud that was generated for both the summaries and reviews have a positive shade to it. So it is safe to say that the products in general are quite well received.

Patterns and Predictions

The analysis started out with the objective of guiding Turtle game manufacturer and retailer better understand their current environment so as to improve their performance in the future. The analysis was split into two avenues, firstly the customer trends were analysed and secondly the sales data was analysed. The insights from both these analysis is as follows.
- There is a significant positive relationship between the income group and the spending nature of a customer and the total amount of purchases they make.
- Through further clustering of the spending nature and the income group of the customer one can say that as most of the customers that the company has come from middle to above income group and they engage in spending more as well.
- From the NLP analysis it can be derived that the most recurring negative sentiment is that the game is quite boring and requires a lot of patience. The positive statements are quite general, but what can be inferred from it is that the

games are suitable for younger customers and the sizing options are convenient as well.
- From the visualisations in the R-studio what can be very well interpreted is that fewer games that achieve very high sales figures. There is a possibility of saturation in the gaming market.