Assignment: Predicting Future Outcomes

Turtle Games: Data Analysis Case Study

Background:

Turtle Games is a game manufacturer and retailer with sales across the globe and wide product range. The purpose of this data analysis project is to provide a better understanding of the Sales and Customers of Turtle games with the goal of increasing overall sales performance.

The datasets used for data analysis are turtle_reviews and turtle_sales. The dataset: turtle_reviews provides us with the reviews that customers have provided surrounding the games sold along with certain details of customers such age, gender, remuneration and spending score. This dataset allows us to get a deeper understanding as to the demographics of users and how to cater to customers better.
The dataset: turtle_reviews provides us with the sales data of games and details of the games such as ranking, platform and region-wise sales. This allows us to analyse which regions provide better sales and understand how we can improve offerings accordingly.

Analytical Approach

Pandas was used to analyze the turtle_reviews data. The redundant columns were dropped during data analysis and outliers were removed to prevent the arising of inaccurate data. There were no null values within the turtle_reviews data which was checked using the isna() function.

Using the matplotlib, I was able to create the regression plots comparing loyalty points with age, remuneration and spending score. From the sklearn library, I imported k-means which enabled the use of elbow and silhouette method to accurately identify the best number of clusters to categorize the remuneration against spending score data. On achieving the ideal number for clustering, pairplot from the seaborn library was used to create a pairplot to visualise the clustered distribution of data according to the k-means clustering.

The dataset was further filtered for the reviews and summary columns. The data was also filtered by lowercase conversion , punctuation & duplicates removal. A word cloud was created which included stopwords and alphanumeric words by tokenizing the data. The stopwords and alphanumeric words were further filtered out and word clouds were recreated for these columns. The most common words were identified and depicted using a horizontal bar plot. The vaderSentiment tool was installed to enable analysis of reviews and summaries to identify as to whether they were negative or positive. The average sentiment polarity was calculated for the reviews and summary columns. The top 20 most positive & negative reviews and summaries were identified and presented. Finally, a histogram was created to display the Sentiment Polarity across reviews and summary columns.

Rstudio was used to analyse the turtle_sales data. The packages used within the data were tidyverse and moments. The is.na() function was used to identify any missing values within the dataset and the data was sense-checked using the dim(), str() and summary() functions. The qplot function was used to plot various types of graphs to visualise sales globally, within NA and within EU. The maximum and minimum sales through sales columns was identified using the min and max functions and mean for the global sales was calculated. This data was filtered for outliers and grouped by its Product IDs with the sales being summed according to their Product ID. The normality of the dataset was tested by q-q plot, Shapiro-Wilk Test, Skewness and Kurtosis (using the moments package). The correlation between the sales columns are identified. Further, simple & multiple linear regression was performed on the sales columns to predict global sales.

Visualisation and Insights

The visualisations within Pandas notebook include:

Linear Regression of Reviews Data:
The visualisations surrounding linear regression between loyalty points against age, remuneration and spending score allows us to understand which demographic of customers are the most beneficial for Turtle Games.

Eblow and Silhouette Method Visualisations:
These visualisations enable us to visually identifty the ideal number of 5 clusters for the dataset.

Pairplot for Kmeans Clustering:
The pairplot visualise shows us how the 5 clusters are distributed in scatterplot and linegraph visualisations of remuneration against spending score and vice versa.

Word Clouds:
The world clouds enable us to visually identify the most common words with them being the largest in size within the word cloud to least common occupying the least amount of space within the clouds.

Histogram of Sentiment Polarity of Reviews:
The histogram of sentiment polarity displays the distrubution customer sentiments in their reviews. The histogram displays that reviews are highly postive.

Histogram of Sentiment Polarity of Summary:
The histogram of sentiment polarity displays the distrubution customer sentiments in their reivew summaries. The histogram displays that review summaries are mostly postive with many of instances neutral sentiments.

The visualisations within the R Script include:

Scatterplots and Bar plots of Global Sales, NA Sales and EU Sales
The scatterplot and bar plots of sales helps understand the number of instances of sales according to sale size globally as well as within the NA and EU Regions. These plots showed a positive or right skew with low instances of high sales and high instances of small sales.

Box plot of Sales:
The box plot on Global Sales and NA sales showed a pattern of 1st quartile, median and 3rd quartile falling within the 5 sales and low instances of sales greater than 10. It also displayed a number of outliers within the dataset.

Visualisations on dataset grouped by Product ID:
Scatterplot: The scatterplot continued to show a right skewed line with many instances of higher values across the scatterplot.
Bar plot: The bar plot showed a less intense right skew with signs of symmetry across the distribution.
Box Plot: The box plot showed a slightly higher 1st quartile, median and 3rd quartile falling within 5 to 10 sales. It also displays low instances of outliers.

Normal q-q plot for Global, NA Sales and EU Sales:
The q-q plot indicated a slight curve and the quantiles fell close to the line. This indicates a right skew in the distribution.

Regression models for Sales Columns
The regression model plots enable us to visually understand how a change in the sales of one region affect the sales of other region. There was a strong positive correlation identified between the sales columns.

Patterns and Predictions

The patterns identified are:

Linear regression between loyalty points against age, remuneration and spending score identified that:

- There is a negatively linear relationship between loyalty points and age.
- There is a positive linear relationship between loyalty points and remuneration and between loyalty points and spending score.

This means Turtle games must focus on younger customers with higher remuneration and spending score.

Histogram of sentiment polarity among reviews express very strong positive sentiments.
Histogram of sentiment polarity among review summaries express strong positive sentiments along with a large number of neutral sentiments.
Turtle Games can focus on those products with the most positive reviews and discontinue products with the most negative reviews to enhance sales performance.

Linear regression models of Global sales, NA Sales and EU sales have positive linear relationship have a positively linear relationship with each other. Hence, Turtle Games can focus on NA and EU regions as they play a large role in the Global sales.