Background and business context

The aim of this analysis was to help Turtle Games to improve its sales performance by understanding customers' opinions and purchasing behaviours. Turtle games would like to understand its customers better by gaining an insight into the makeup of its customer base, how customers collect loyalty points and customer opinions. This information will help Turtle Games to tailor its services to meet customer preferences which will sustain and grow the number of customers. Furthermore, Turtle Games would like to understand the structure of its sales data. This will help Turtle Games to use predictive data analysis to create a sales strategy that captures market trends.

Analytical approach

Single and Multiple Linear Regression

I used linear regression to explain to what extent the variance of a certain dependent variable within the turtle data set could be explained by independent variables. This helped me to identify what factors could explain the variation in loyalty points. I found that both remuneration and spending score had P values that were less than 0 demonstrating their significance in affecting the variance of loyalty points. I tested and found no multicollinearity between spending score and remuneration, this strengthened the accuracy of my results because multicollinearity can cause R-value to overestimate the proportion of variance caused by independent variables.

I also used single linear regression in R to determine whether sales data from North America and the EU was responsible for the variance of turtle games' global sales.

K-means clustering

I used K-means clustering to identify different customer groups in the turtle reviews data set. K-means clustering groups data points around a certain number of centroids. Each data point will belong to the cluster with the nearest mean. Methods such as the Elbow method help to identify the optimum number of clusters by computing the sum of square distances that each data point is from the centroid for a range of different possible clusters. At the elbow point, there is an optimisation point where the returns of having more clusters is no longer significant.

NLP

NPL allowed me to identify customers' sentiments towards turtle games.

I cleaned data in preparation for NPL by:

- Using a list comprehension within a lambda function to apply lowercase to each word in the review and summary column.
- I replaced punctuation with blank spaces using the str.replace() method.
- I dropped the duplicates in the columns using the drop_duplicates() function
- I then created a separate data frame for each column and used a for loop to string the comments within the review and summary column into a string variable.
- I then tokenized the words and eliminated stop words by downloading and applying the corpus module in combination with a list comprehension and for loop to iterate through data to remove stop words.

Firstly I used the TextBlob sentiment analysis to assess the polarity of customer reviews and to find the most negative and positive reviews. However, after finding inaccuracies in this method I applied the Vader model as this model has been trained to understand sentiment in social media. This means it can identify nuances in language such as colloquialisms and repetitive words.

Exploratory Data Analysis (EDA)

I used EDA in R to determine the reliability of the turtle sales data.

- I created and plotted a qqnorm function to determine if the data is normally distributed. The straighter the plot on the qqnorm plot the more normally distributed the data will be.
- I used the Shapiro-Wilk test, if the test produces a p-value less than 0.05 we reject the null hypothesis that the data is normally distributed.
- The Skewness and the Kurtosis tests determine whether the data has longer/fatter or heavier/lighter scales retrospectively.

Visualisation and Insights

Single and Multiple Linear Regression

I started by using a heat map to determine the level of correlation between variables in the review and sales data set. This helps to identify which variables to use in the regression tests. Then scatter plots and best-fit lines visualise the predicted values resulting from a linear combination of the predictors.

K-means clustering and visualisations

I started by using simple 2D clustering to visualise possible clusters in the turtle reviews data set. However, I found that 3D clustering increased the amount of

information that could be gained from the analysis as clusters could include three different variables.

Natural Language Processing

Cloud maps were used to visualise the most common words in the data set. Words that were more common appear in larger letters.

I then used matplotlib to create clear bar charts with the most used words.

I then used an orange histogram to demonstrate the distribution of polarity scores produced by the TextBlob model and a black histogram to compare the distribution of compound Vader scores produced by the Vader model.

Finally, I used a red histogram to demonstrate the distribution of comments Vader scores that mentioned words such as challenging and a green histogram to show the distribution of comments Vader scores that mentioned words such as simple.

Visualisations in R

I used the group_by(), arrange() and head() to wrangle data to create plots that could reveal certain insights.

I created simple bar charts and scatter plots to analyse the difference between EU, Global and North America (NA) sales. I chose blue to represent the EU, red to represent North America and green to represent Global sales as I believe that to some extent these colours represent these regions.

I created a visualisation using a time series in R which helped me to identify customers' seasonal purchasing behaviours.

Patterns and predictions that address Turtle Games Business Objectives:

- In multiple linear regressions where loyalty points is the dependent variable, the R^2 value is 82% which shows how change in loyalty points can be mostly explained by remuneration and spending score.

- 3D k-means clustering revealed that Turtle Games should create a marketing campaign to target younger customers with more income who are willing to spend more. A cluster group revealed how some younger customers have less income but they’re willing to spend the same amount as high income

groups, bundles and discounts could help to retain this group with less purchasing power.

- Using the Vader model to perform sentiment analysis I found that Turtle Games should create a marketing campaign around the sale of gifts. Furthermore, customers enjoy both challenging and easy games. Turtle Games should use different marketing strategies to capture these different customer preferences.

- I found that Turtle Games should shape its sales strategies around platforms that are trending. The sales of different products go up or down depending on when they've been released. The PS4 is the most recent platform to have a spike in sales.

- Plotting the data on a Q-Q plot, carrying out the Shapiro-Wilk, skewness and kurtosis tests revealed that the data was not normally distributed. You can take steps such as sqrt(x) for positively skewed data to normalise the data to create more consistency in the data set.

- Finally, I found that EU sales and NA sales account for 97% of the variation in global sales using multilinear regression in R.