Turtle Games Technical Report

Background

Turtle Games is a game manufacturer and retailer with a global customer base. The company manufactures and sells its own products, along with sourcing and selling products manufactured by other companies. Its product range includes books, board games, video games, and toys. The company collects data from sales and customer reviews. Turtle Games has a business objective of improving overall sales performance by analysing and considering customer trends. To improve overall sales performance, Turtle Games has developed an initial set of questions, which are:

- How do customers engage with and accumulate loyalty points?
- How can customers be segmented into groups and which groups can be targeted by the marketing department?
- How can text data (e.g. social data such as customer reviews) be used to inform marketing campaigns and make improvements to the business?
- Can descriptive statistics provide insights into the suitability of the loyalty points data to create predictive models (e.g. normal distribution, skewness, or kurtosis) to justify the answer.

Analysis + Visualisations

To support TG, both Python and R were utilised to gain deeper insights into customer behaviour. They were used to reveal patterns in customer activity. The visualizations generated were aimed to provide clear insights to help TG better understand their customers and make informed business decisions. These visualisations not only highlighted customer preferences but also identified areas for improvement in their marketing and product offerings.

Python

First, Python was used to analyse turtle_reviews.csv, which contained data from 2000 customers. The dataset was imported into a Jupyter workbook, alongside the libraries, which included:

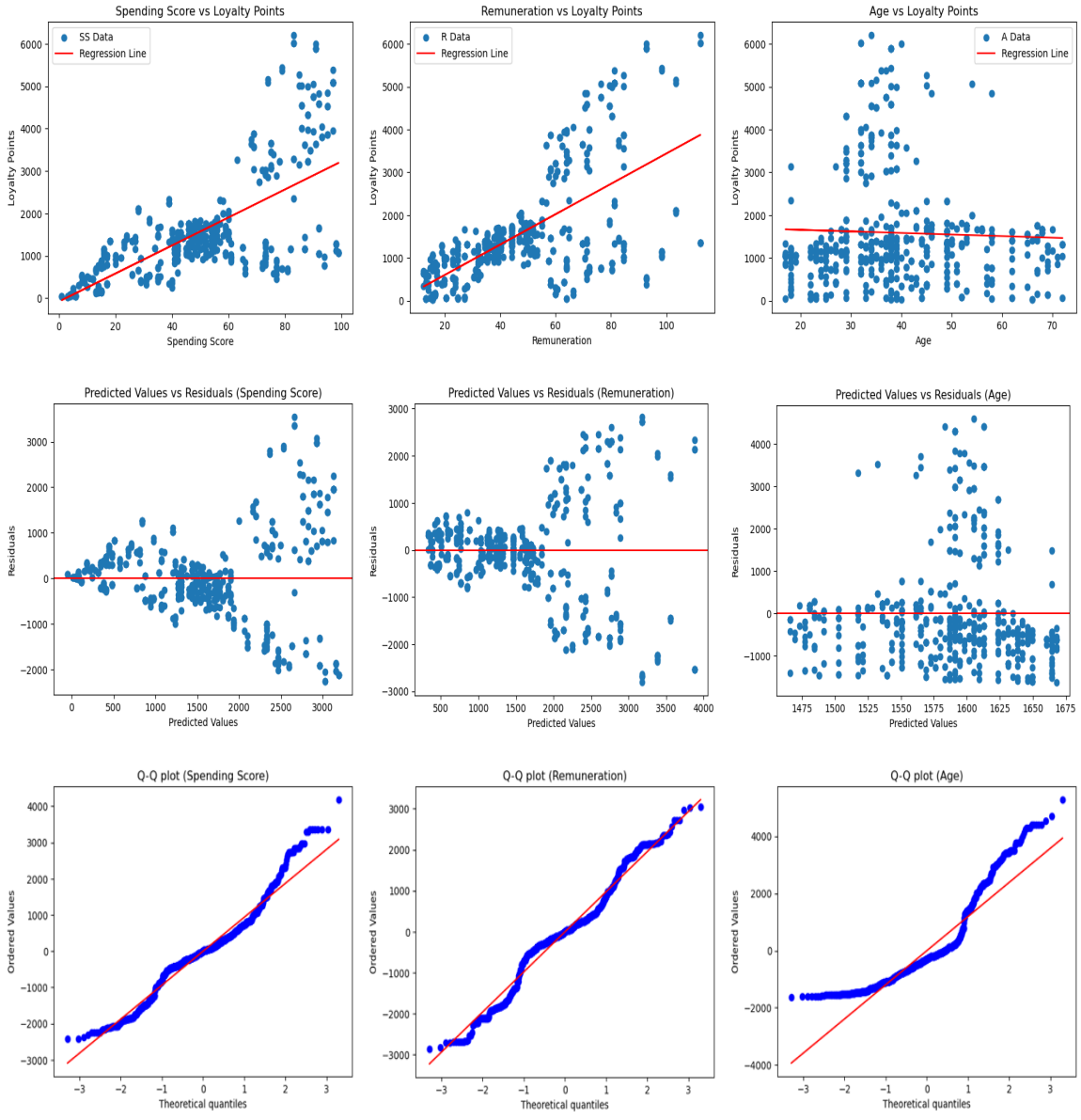
- NumPy
- Pandas
- Sklearn
- Statsmodels
- Matplotlib
- Seaborn

The dataset was sense-checked for missing/duplicate values. The redundant columns were removed. Python used the processes below to perform the analysis and produce the visualisations.

Linear Regression

One thing TG was interested in learning was how the users accumulated loyalty points, which lead to the investigation of whether there were any existing relationships between the loyalty points with the age, remuneration and spending scores. That is where regression came in - to explore the potential relationships between the dependent variable (loyalty points) and the independent variables (age, remuneration, spending scores). This included splitting the dataset into train and test sets. Multiple linear regression was also performed. 3 graphs were designed for each independent variable:

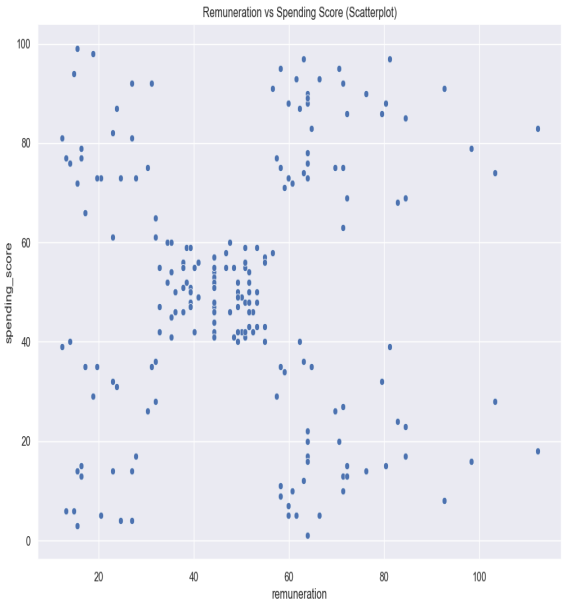
- 1st row: Relationships between the loyalty points and the independent variables.
- 2nd row: Relationships between the residuals with the predicted values.
- 3rd row: Q-Q plots to check if the residuals follow the normal distribution.

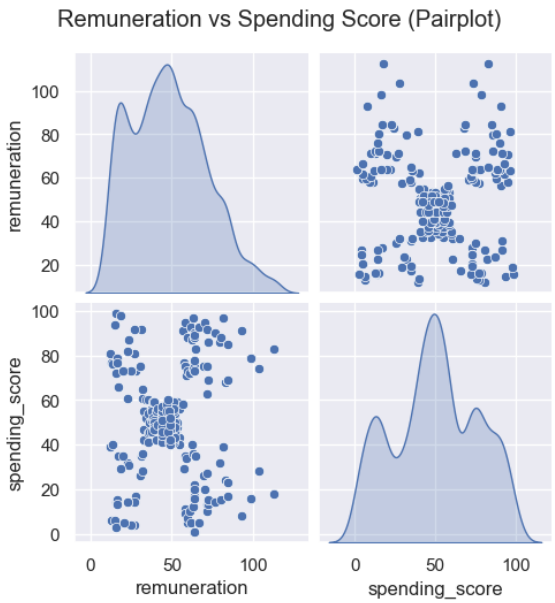


K-Means Clustering

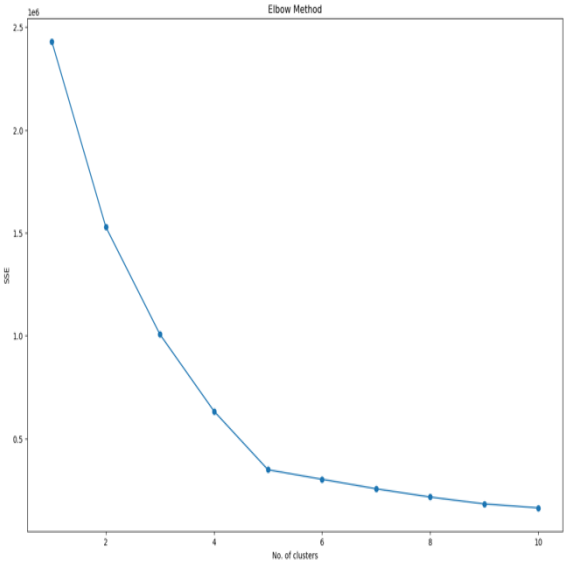
Clustering was utilised to segment the customers – TG wanted to learn if remuneration and spending score can be used to understand their customers' purchasing behaviour.

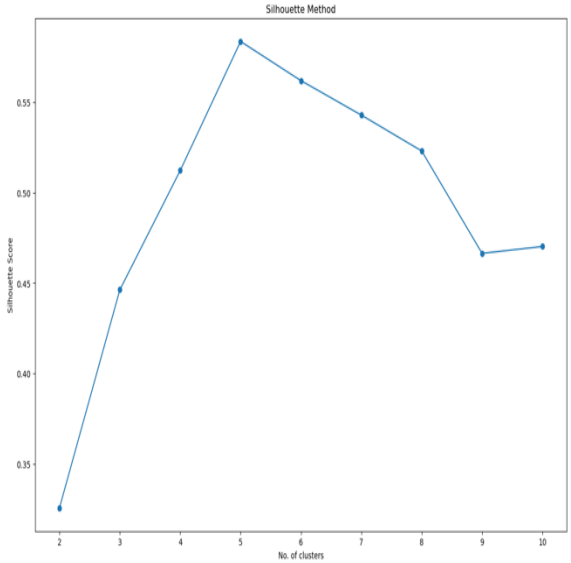
Started out by exploring if there was a relationship between the spending score and remuneration as shown below.



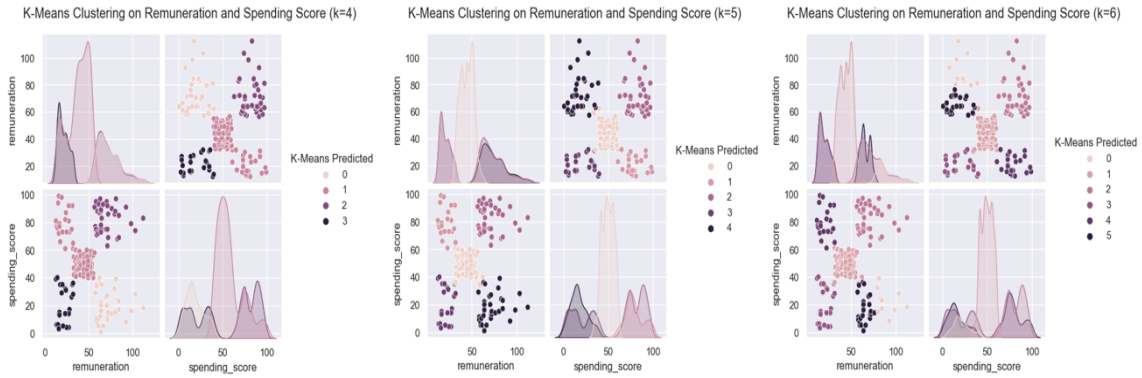


To determine the right number of groups to segment the customers, both the elbow and silhouette methods were applied.





Evaluated the model by experimenting with different values for the cluster number.

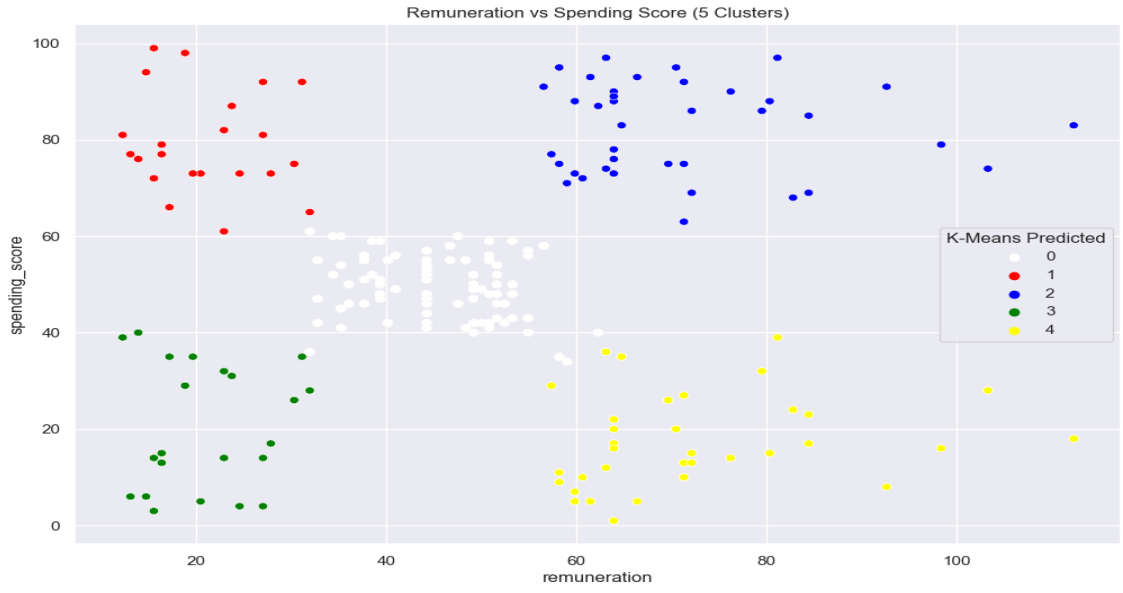


Chose to keep the number of clusters at 5 for the following reasons:

- k = 4: Risk of merging the distinct spending behaviours.
- k = 6: Introductions of unnecessary fragmentations without any useful insights.
- k = 5: Visualisations produced the most interpretable and distinct customer segments.

This created 5 separate groups within the customers regarding their remuneration and spending score:

- Mid-earning mid spenders
- Low-earning big spenders
- High-earning big spenders
- Low-earning small spenders
- High-earning small spenders



Natural Language Processing

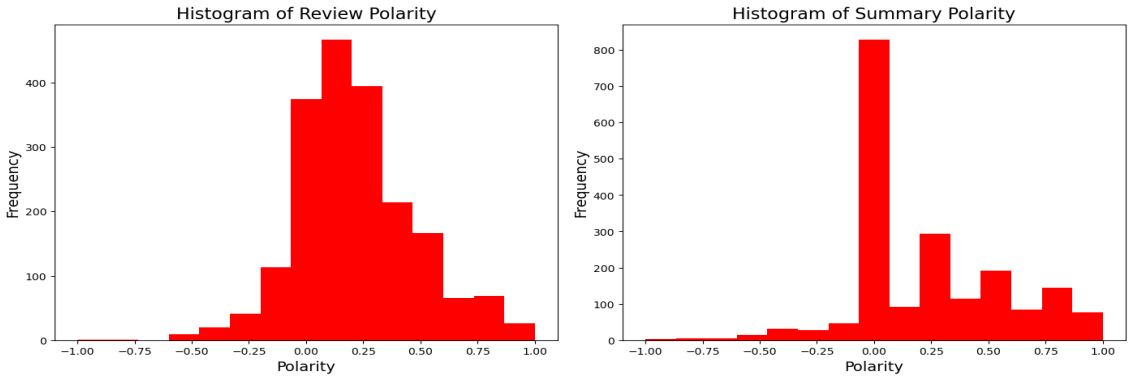
The NLP techniques were used to perform sentiment analysis on the reviews and summaries. To ensure maximum accuracy, the datasets were:

- Checked for missing and duplicate values.
- All the letters were changed to lowercase to reduce variability.
- Punctuation removed for simplification.

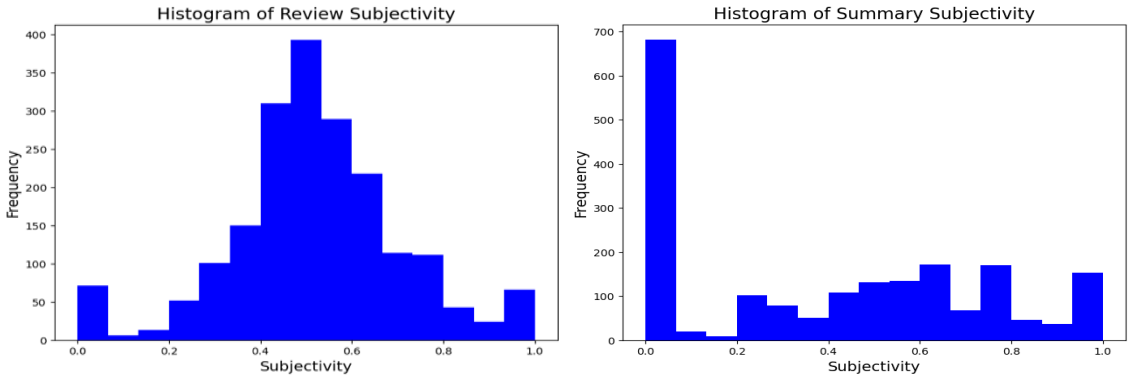
Next step was tokenisation – breaking the dataset down to words. This led to the formation of these word clouds to gain a deeper insight of how the customers truly felt about the company.

The word clouds were generated again but without the alphanumeric characters and stop-words so the analysis can focus on the more important words as shown below.

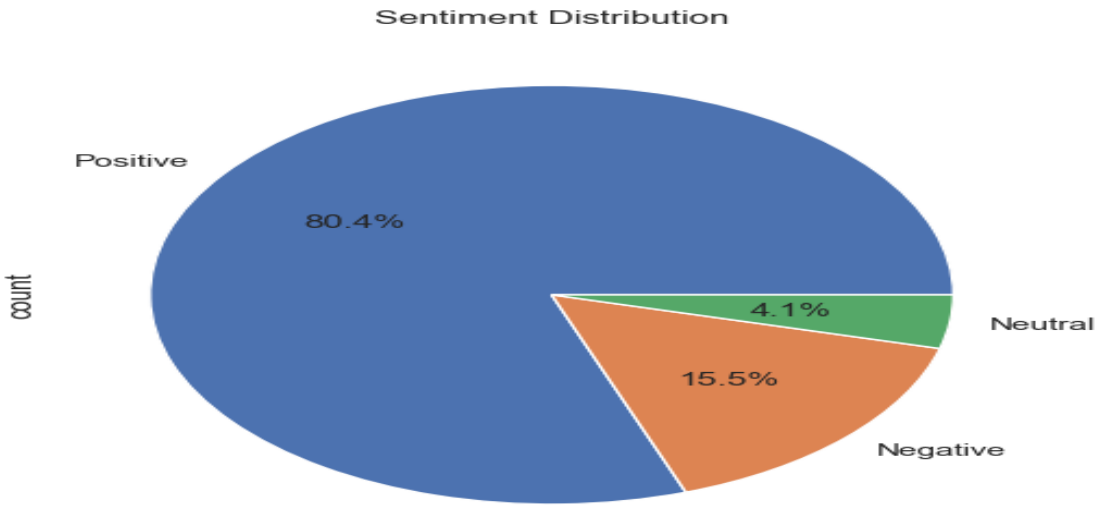
The sentiment analysis results were plotted on histograms. The polarity shown in the histograms below range from -1 (negative sentiments) to 1 (positive sentiments).



Two more histograms were plotted but based on subjectivity – 0 (objectivity) to 1 (subjectivity).



Applied a pie chart to work out the overall sentiment of the customers' feelings towards Turtle Games.



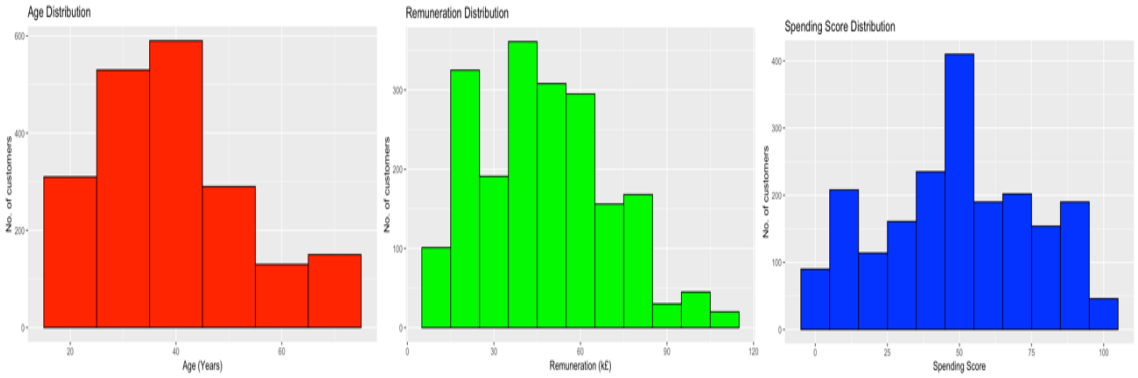
R

Turtle Games preferred R for the analysis due to its workflow systems. The clean version of turtles_reviews.csv, which was created during the Python analysis, was imported into RStudio. The following libraries were loaded:

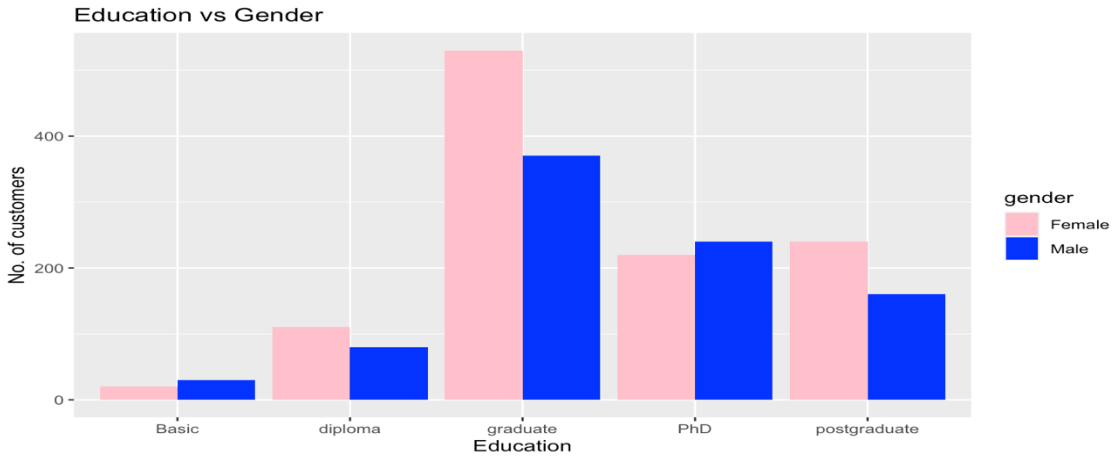
- tidyverse
- skimr
- DataExplorer
- moments

Just like with Python, the dataset was sense-checked, which included searching for missing values. R used the below visualisations to provide useful insights for TG.

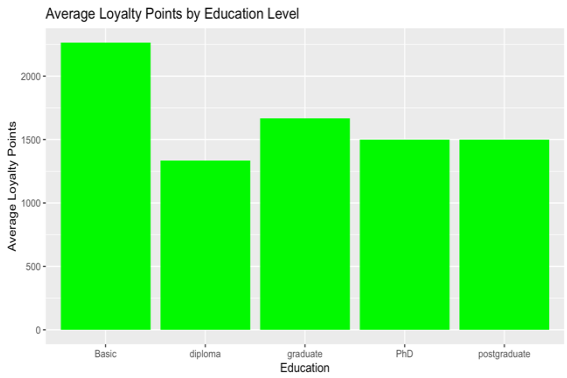
The histograms were used to visualise the distribution of each of the independent variables to gain more information about the customers.



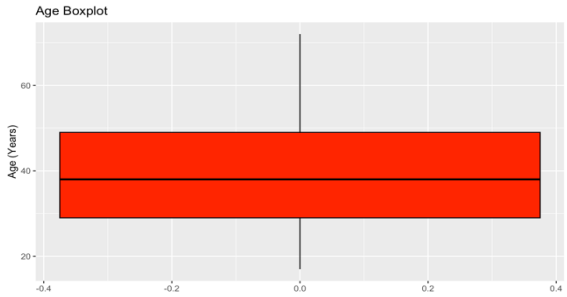
The bar chart below provided an overview of customer distribution of 'education' and 'gender' together to further explore which demographics TG appease to the most.

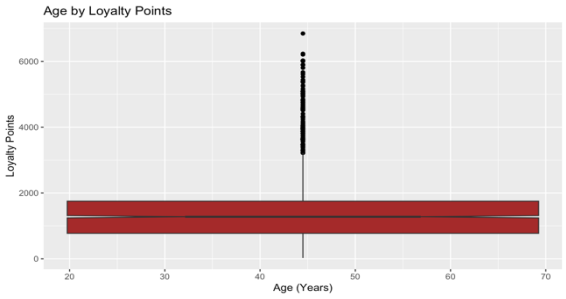


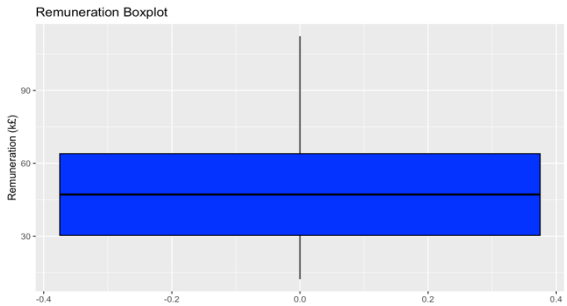
Used bar charts again to see which demographics accumulated the best loyalty points averages.

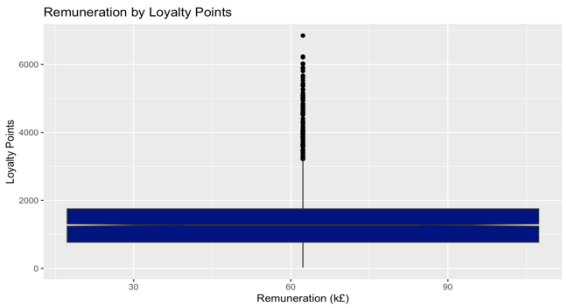


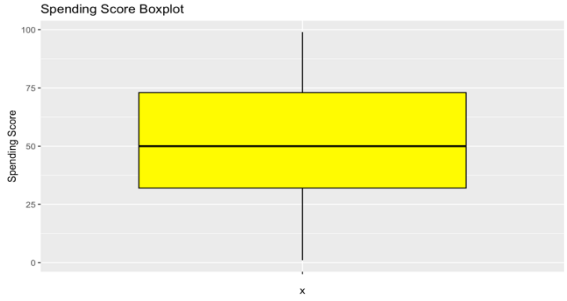
These boxplots were produced to visualise the distribution of the independent variables separately and loyalty points across each of them. The purpose was to detect the presence of any outliers. Although the majority followed the distribution, there were some outliers detected, particularly with the spending score and remuneration. These outliers were retained, as they were true reflections of the customers' behaviour and would even further support the analysis.

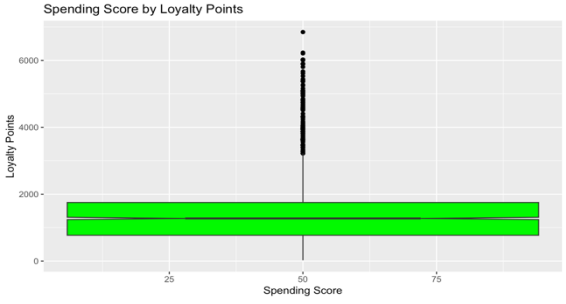




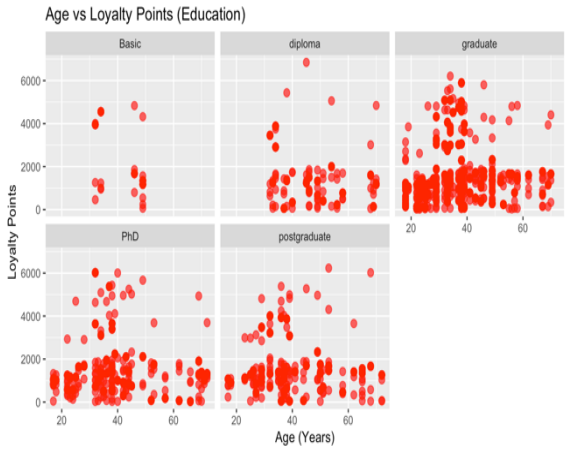


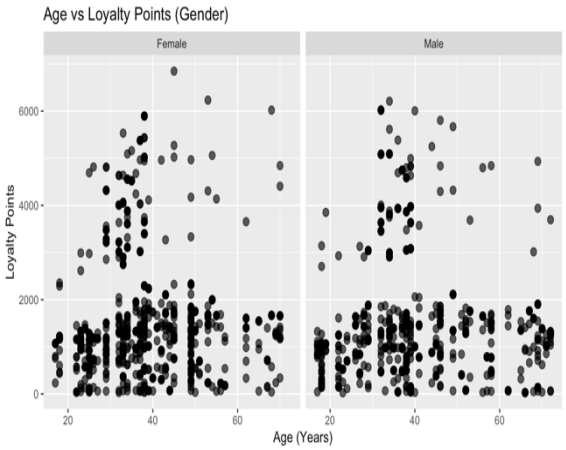


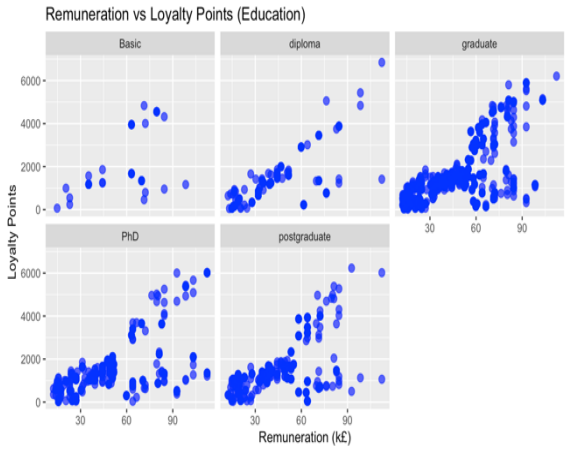


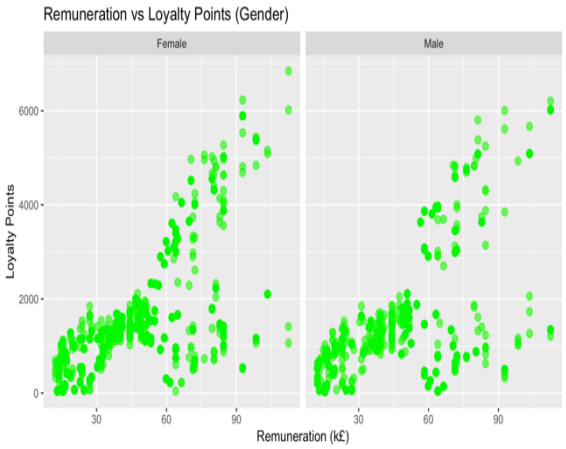


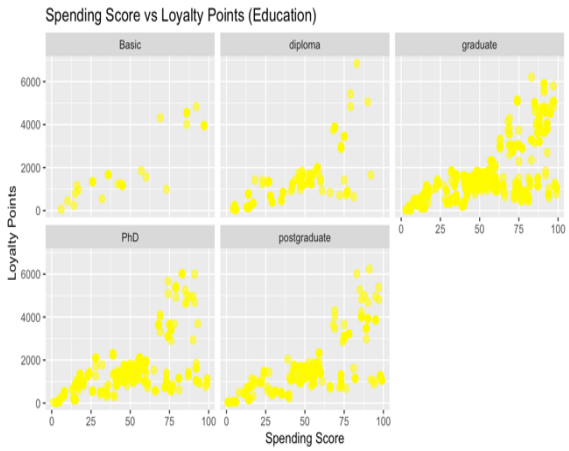
Scatter plots were used to explore any possible relationships between the loyalty points and the independent variables. The visualisations were based on the education and gender to explore any patterns within the customer segments.

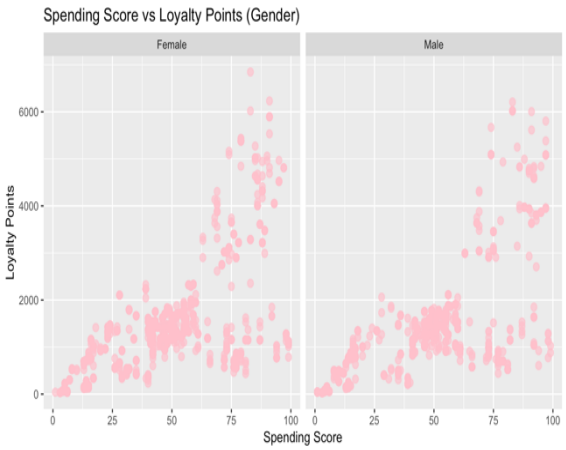




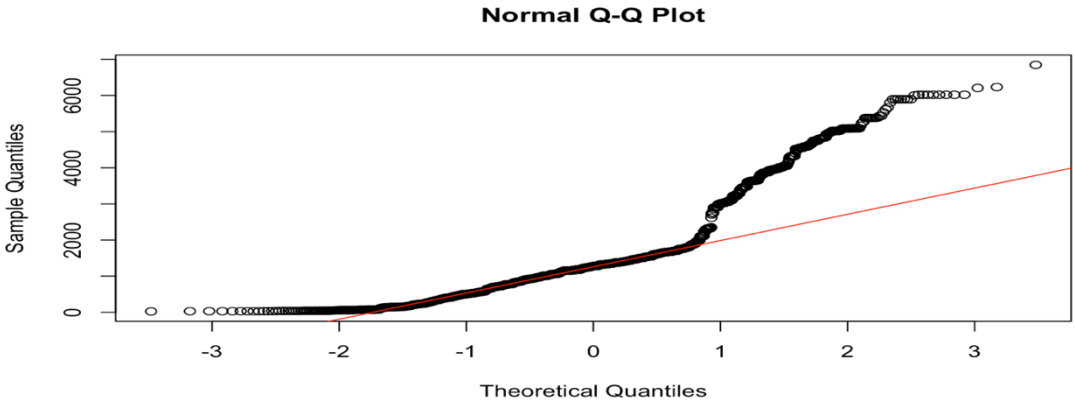








Next step with R was to perform statistical analysis on the loyalty points data. This included testing the normal distribution with the Shapiro-Wilk test. This led to the below q-q plot.



After the q-q plot, the data was also checked for:

- Skewness
- Kurtosis
- Correlations

A multiple linear regression model was also used to assess the relationship between the loyalty points and the predictors. Created many models with different combinations of the predictors to find the best fit and highlight the main contributors to the loyalty points accumulation.

Recommendations

After the analysis from both Python and R, it became quite clear the accumulation of the loyalty points was related more to the customers' behaviour rather than their demographics. Therefore, the recommendations for Turtle Games to improve their business are:

- Adopt behaviour-based targeting.
- Tailor loyalty programs by customer segment.
- Prioritise high-spending customers.
- Promote high-performing products categories.