TURTLE GAMES: CUSTOMER INSIGHTS AND LOYALTY ANALYSIS

Data-Driven Strategies for Sales Growth and Customer Engagement

Business context

Turtle Games is a company that both manufactures and sells its own product, while also sourcing and retailing products made by other companies. Its products include books, board games, video games, and toys.

The objective of the data analyst team is to improve overall sales performance through data analysis and customer reviews collected by Turtle Games. The analysis provided attempts to answer some questions posed by Turtle Games:

1. What influences 'loyalty points?
2. How can customers be clustered for marketing purposes?
3. How can social data be leveraged once customer reviews are collected?
4. Can prescriptive statistics demonstrate the suitability of the 'loyalty points' data to create predictive models?

Analytical Approach

The dataset consisted of **2,000 observations** and **11 variables**. The analysis was conducted using **Python** and **RStudio** to ensure robust statistical evaluation and cross-validation of findings.

Python Analysis

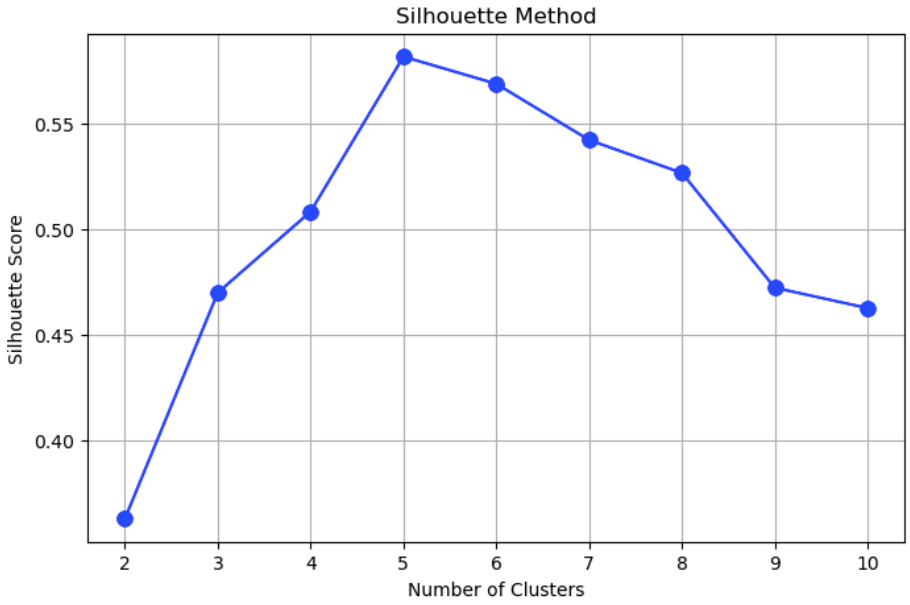
The analysis began using Python by importing the data frame provided by Turtle Games and focused my attention on **loyalty points, age, remuneration, and spending score** features. The data was clean; no null values were found.
- **NumPy** and **Pandas** were employed to work with arrays
- statistical models were used to obtain estimates and run performance tests

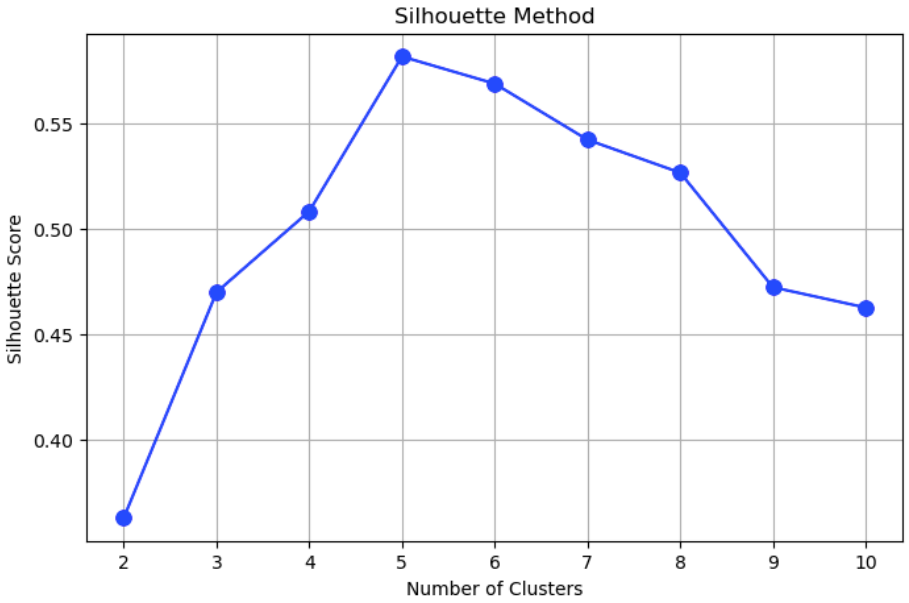
- **Matplotlib** was used for visualization
- **Scikit-learn** was used for statistical modelling and machine learning

Three regression models were developed using sklearn.model_selection to assess how *loyalty points* relate to the other variables.

A **Decision Tree Regressor** was initially fitted using mean_squared_error for evaluation. The first model overfitted the training data, suggesting limited generalizability. After applying **model pruning**, overfitting was mitigated while maintaining strong predictive performance.

For **customer clustering**, both the **elbow method** and **silhouette method** were applied, helping identify the optimal number of clusters.
For **sentiment analysis**, text cleaning and tokenization were performed using **Re** and **WordCloud** libraries. Duplicate entries were removed, and the most frequently used words were visualized. Polarity scores were then computed to analyze sentiment across both *reviews* and *summaries*, followed by an in-depth examination of the top 20 positive and top 20 negative examples.





RStudio Analysis

In RStudio, libraries such as **Skimr** and **ggplot2** were used for data summarization and visualization. After reviewing the dataset, the key features selected for modelling were *age*, *remuneration*, *spending score*, *gender*, *education*, and *loyalty points*.

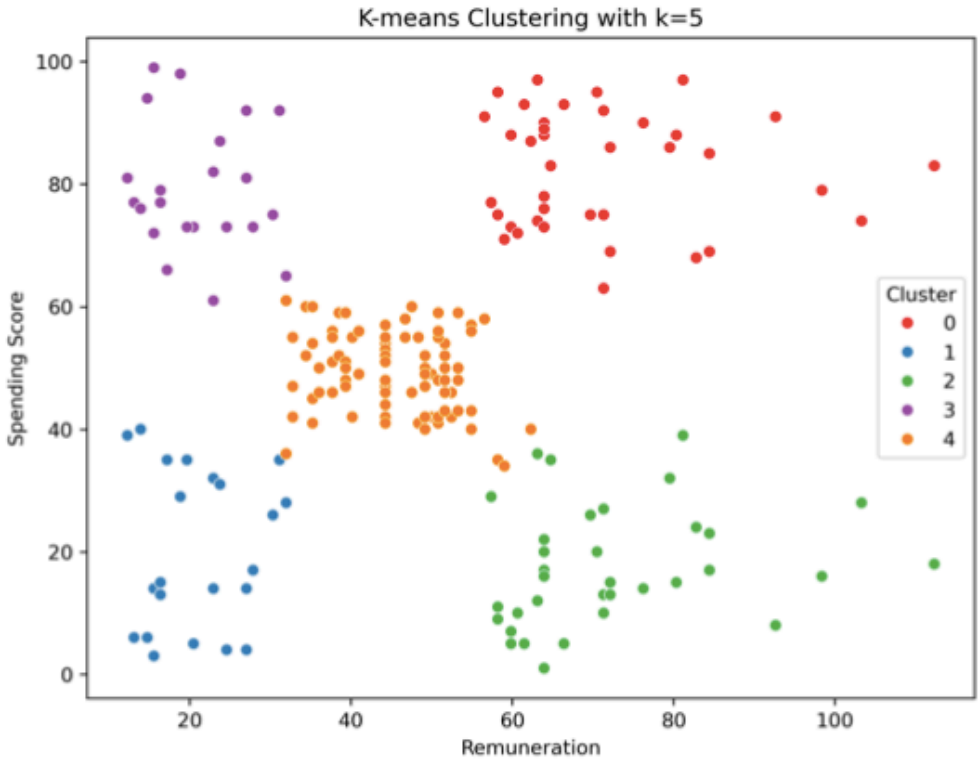
Descriptive statistics and visualizations (histograms, boxplots, scatterplots) were employed to understand variable distributions.
A **multiple linear regression model** was initially built using *spending score* and *remuneration*, as single-variable regressions suggested that *age* had limited explanatory power. However, when *age* was integrated into the multiple regression, a slight performance improvement was observed.

Residual plots were examined to verify the normality assumption of the model, confirming that no log-linear transformation was necessary.

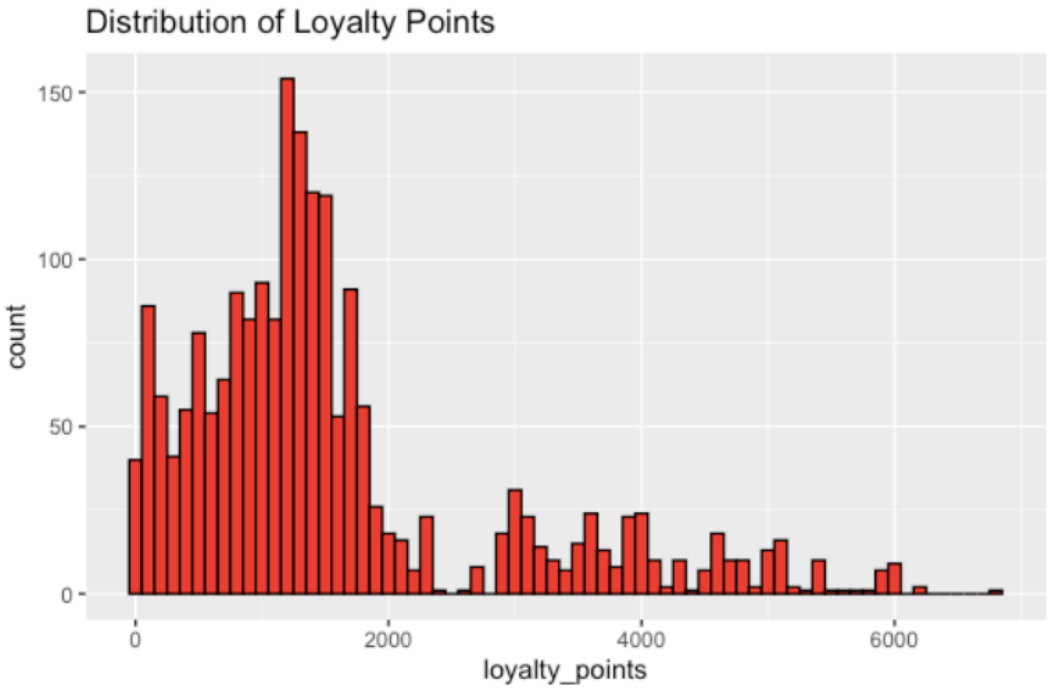
Visualization and Insights

Visualizations were designed to emphasize data distribution patterns and support marketing decision-making.

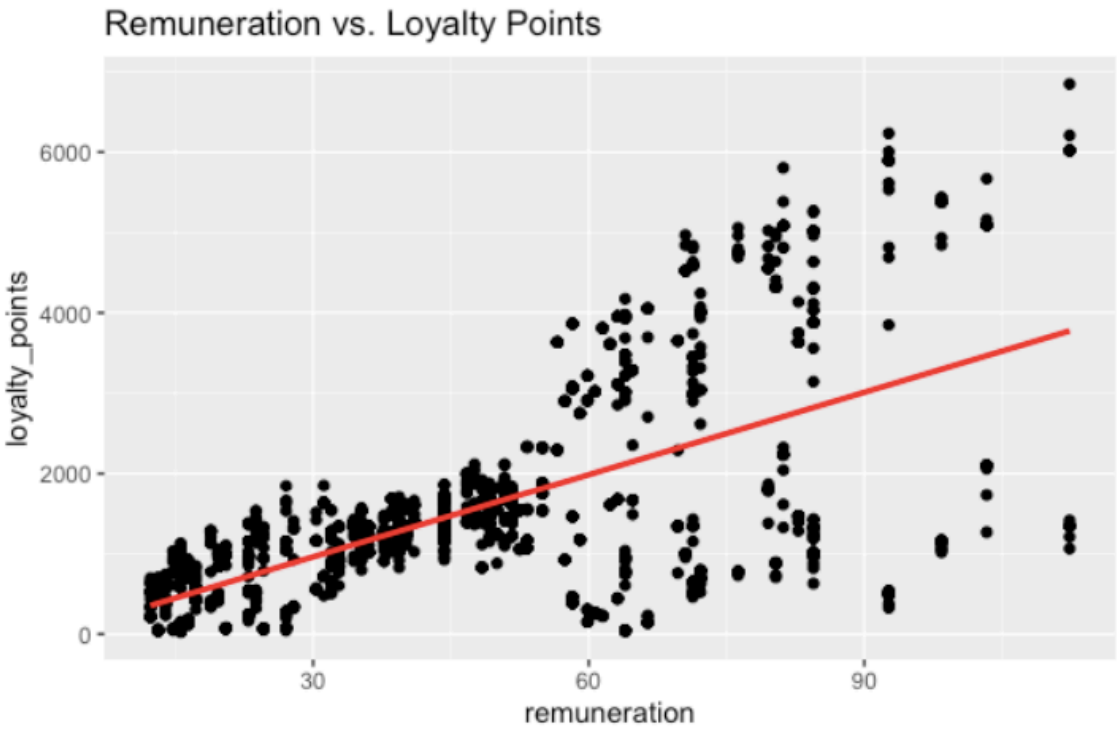
- **Clustering Analysis:**
  The elbow and silhouette methods indicated that **five clusters** provided the clearest segmentation. These clusters represented distinct customer profiles — from *high-income, high-spending* to *low-income, low-spending* — offering valuable input for personalized marketing strategies.

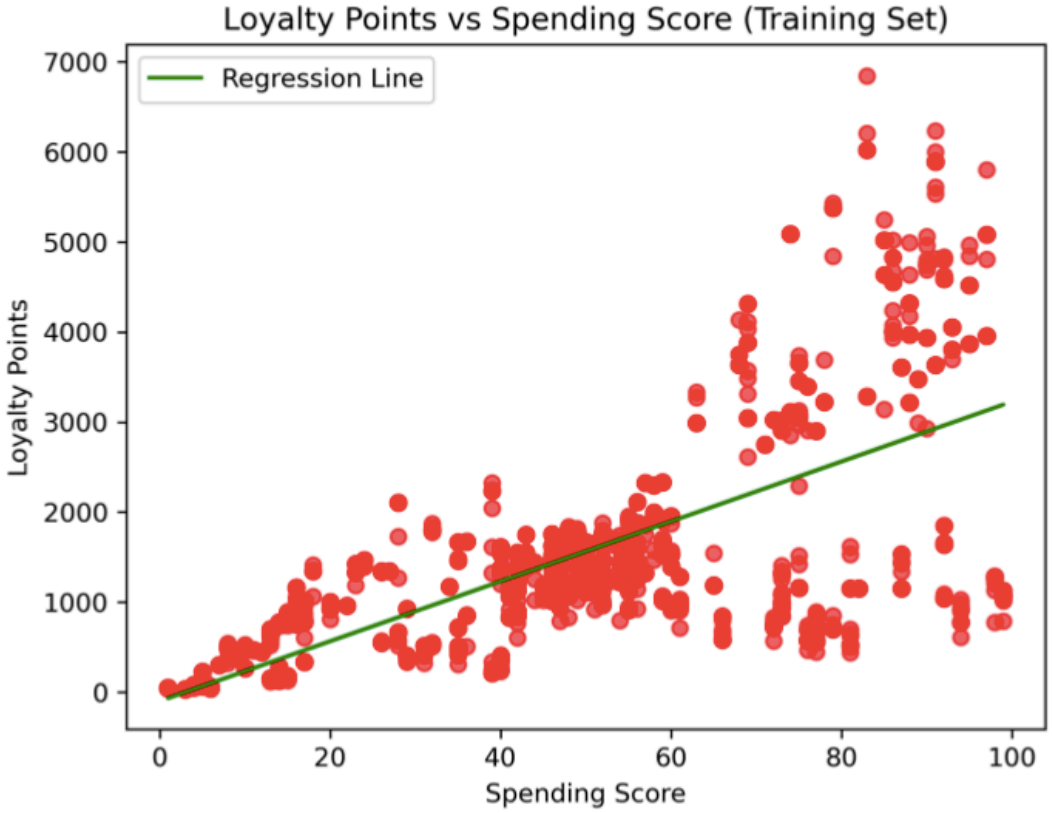


- **Distribution Analysis:**
  Histograms and boxplots revealed that *loyalty points* follow a **right-skewed distribution**, where most customers accumulate relatively low points while a small number earn disproportionately high points. Outliers were more frequent among well-educated customers.



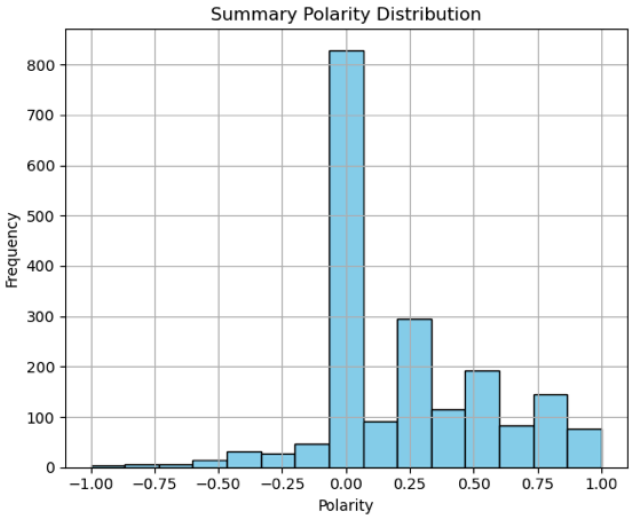
- **Correlation and Regression Insights:**
  Scatterplots with regression lines illustrated:
  - A **strong positive correlation** between *loyalty points* and *income* (r = 0.62).
  - A **strong positive correlation** between *loyalty points* and *spending score* (r = 0.67).
  - A **weak negative correlation** between *age* and *loyalty points* (r = –0.04).
  - A **slightly negative relationship** between *age* and *spending score*.

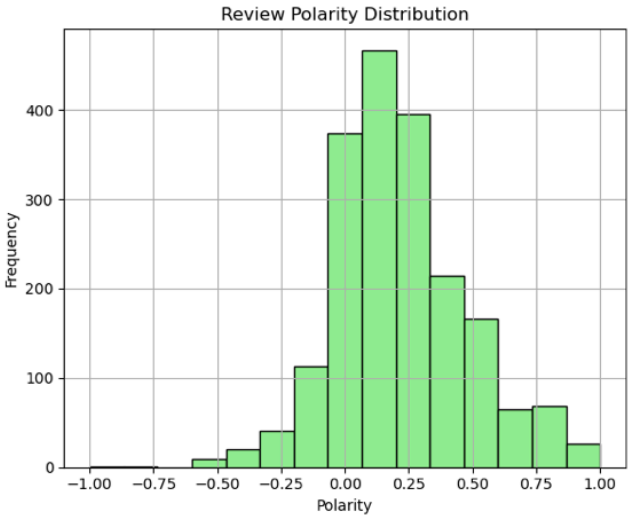

Remuneration vs. Loyalty Points
loyalty_points
6000
4000
2000
0
30
60
90
remuneration


Loyalty Points vs Spending Score (Training Set)
Regression Line
Loyalty Points
7000
6000
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
Spending Score

- The multiple linear regression model explained **83.99% of the variance**, validating its suitability for predictive applications. The decision tree model achieved an approximate **91% accuracy** when classifying new data.

- **Sentiment Analysis:**
  Word clouds and polarity histograms provided qualitative insight into customer feedback. Common positive terms such as *"stars"*, *"five"*, and *"great"* indicated strong satisfaction and brand loyalty.
  Polarity distributions were **left-skewed**, confirming a predominance of positive sentiment. However, negative terms like *"boring"* and *"disappointing"* revealed isolated areas for product improvement.





Patterns and Predictions

The findings highlight several key insights:

- **Determinants of Loyalty:**
  Loyalty points are strongly influenced by both **spending score** and **remuneration**. Wealthier customers tend to spend more and consequently earn higher loyalty points.

- **Customer Segmentation:**
  Five distinct clusters emerged from the analysis:

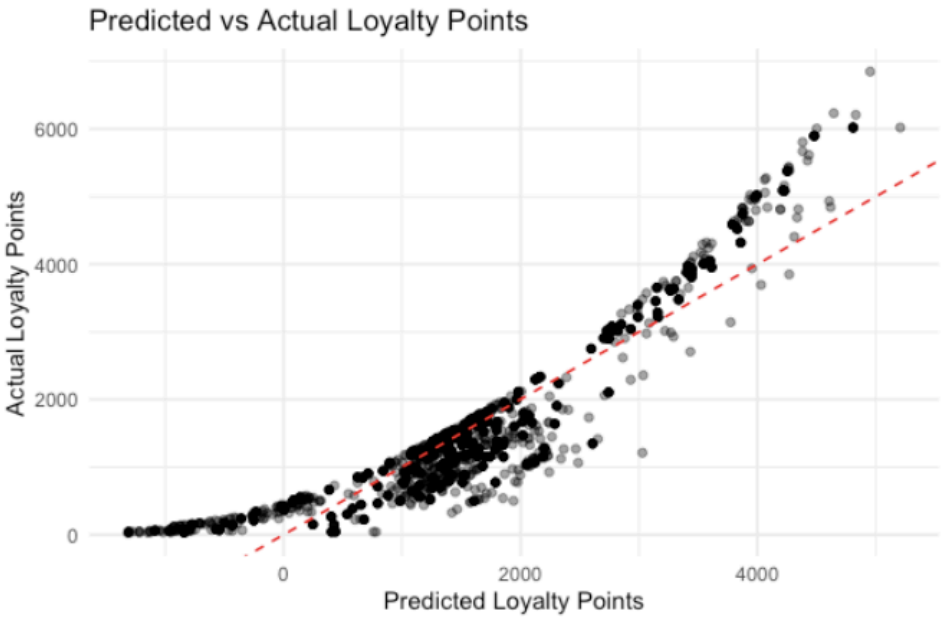
  1. High remuneration, high spending
  2. Low remuneration, low spending
  3. High remuneration, low spending
  4. Low remuneration, high spending
  5. Medium remuneration, medium spending

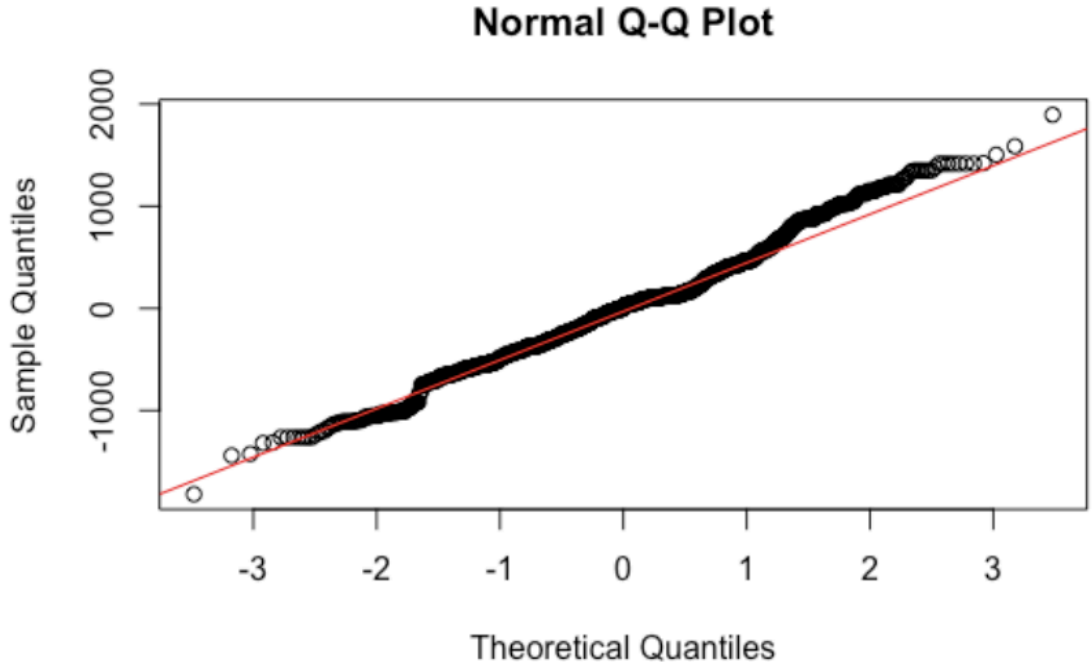
- These segments enable the marketing team to design **targeted incentives** and tailor promotional strategies for each group.

- **Model Reliability:**
  The multiple regression model explains nearly 84% of the variance, while the decision tree demonstrates strong predictive capability with a 91% accuracy rate.

- **Customer Sentiment:**
  Overall customer sentiment is **overwhelmingly positive**, indicating a solid brand reputation. However, the presence of occasional negative feedback provides an opportunity for continuous product and service improvement.

To validate the model's assumptions, the normality of the residuals was assessed. As shown in the graph below, the residuals are generally distributed around the red line, indicating that the normality assumption is reasonably satisfied. This suggests that a log-linear transformation may not achieve a better fit





Conclusion

The analysis confirms that *remuneration* and *spending score* are the strongest predictors of *loyalty points*, while *age* contributes marginally. The models developed

can effectively predict customer loyalty behaviour and inform marketing segmentation.

By integrating clustering and sentiment analysis, Turtle Games gains a holistic view of its customer base — enabling **data-driven marketing**, **targeted loyalty programs**, and **strategic product enhancements** that strengthen customer engagement and boost overall sales performance.