PREDICTIVE ANALYTICS FOR SALES & LOYALTY OPTIMISATION AT *TURTLE GAMES*

Tilani Wijamunige - Data Analyst

30/09/2024

BUSINESS OVERVIEW

Turtle Games is a global game manufacturer and retailer, offering a product range that includes books, board games, video games, and toys. They manufacture and sell their own products, alongside sourcing and selling products from other companies.

Business Objective

To improve overall sales performance by leveraging data to optimize customer insights, enhance the loyalty program, and refine marketing strategies.

Analytical Objectives

This analysis addresses 2 key areas:

1. **Customer Behaviour and Loyalty Program Analysis:**
   - What relationships exist between customer demographics and spending behaviour?
   - How can customer segmentation using remuneration and spending scores inform marketing decisions?
   - What do customer reviews reveal about their experience with Turtle Games?
2. **Predictive Modelling and Program Effectiveness:**
   - Which model (Multiple Linear Regression or Decision Tree) better predicts loyalty points accumulation based on customer features?
   - How can Turtle Games improve its loyalty program and data collection practices?

ANALYTICAL APPROACH

Data Overview

Customer demographic and sales data from 2,000 customers were used for this analysis. A provided metadata file helped understand the dataset and rename some variables for clarity. During the data cleaning phase, duplicates and missing values were addressed using Python and R. Uniform variables, such as language (English) and platform (Web), were excluded as they provided no analytical value. A key limitation of this analysis is the lack of descriptive information about the products, as only product codes were provided.

1. Exploratory and Regression Analysis in Python and R

Python

The dataset was imported into Python, and libraries such as *NumPy*, *pandas*, *sklearn*, and *statsmodels* were used for descriptive and exploratory analysis. *Matplotlib* and *seaborn* were employed for visualization.

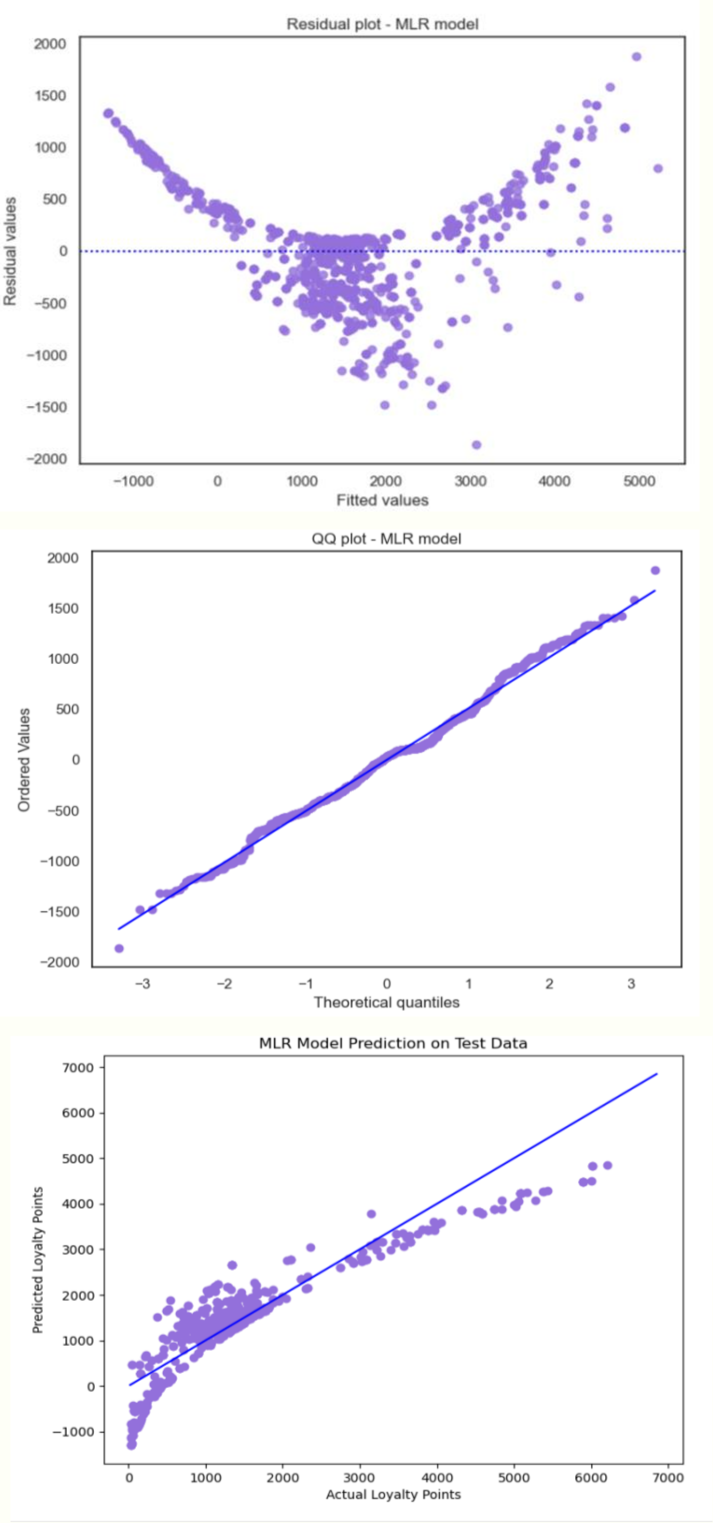
After data wrangling, simple linear regression models were developed to explore relationships between the quantitative variables: loyalty points (dependent variable), spending score, remuneration (income in £1,000), and age (independent variables). Multiple Linear Regression (MLR) was performed using the Ordinary Least Squares (OLS) method.

Assumptions of OLS - linearity, independence, homoscedasticity, and normality of residuals- were tested. The model's reliability was validated through a train-test split (70% training, 30% testing). The MLR in Python identified spending score and remuneration as the most significant predictors of loyalty points, while age was not statistically significant.

R

A similar approach was applied in R, using packages like *readr*, *dplyr, ggplot2*, *car*, and *lmtest*. Boxplots, scatter plots, and histograms were used to visualize the data and explore relationships, particularly focusing on gender, education, age, and product codes (n=200).

Based on insights from the Python MLR model, an additional MLR model in R was developed, excluding the age variable. Both models (Python and R), including age, showed a slightly better fit (higher R-squared and lower error metrics) than the R model excluding age.



Model Diagnostic Plots - MLR Model

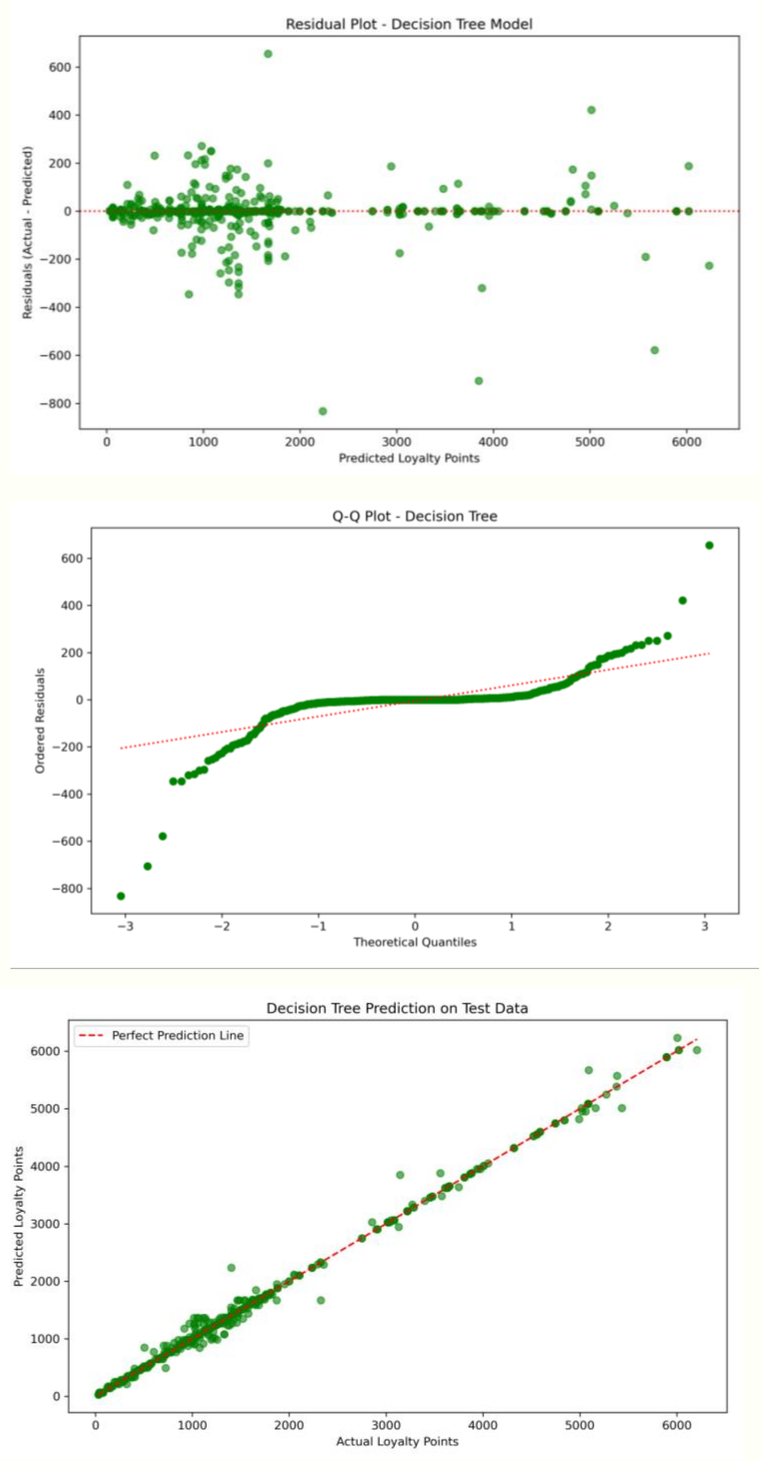
However, both models exhibited heteroscedasticity and non-normal residuals, indicating structural issues despite age's slight contribution as a predictor.

2. Decision Tree Model in Python

A DecisionTreeRegressor was used to explore both quantitative and categorical variables' impact on loyalty points. The five levels of the 'education' variable were aggregated into three categories for simplicity, and dummy encoding was applied to both education and age. The data was split into a 70/30 ratio for training and testing.

The initial tree model fit the training data perfectly, indicating overfitting, which was evident in the test set's error metrics. Fivefold cross-validation was performed, confirming consistent performance across training subsets. Pre-pruning techniques (max_depth=5, min_samples_leaf=10) reduced complexity but led to some underfitting, as evidenced by higher error rates due to the pruning constraints.

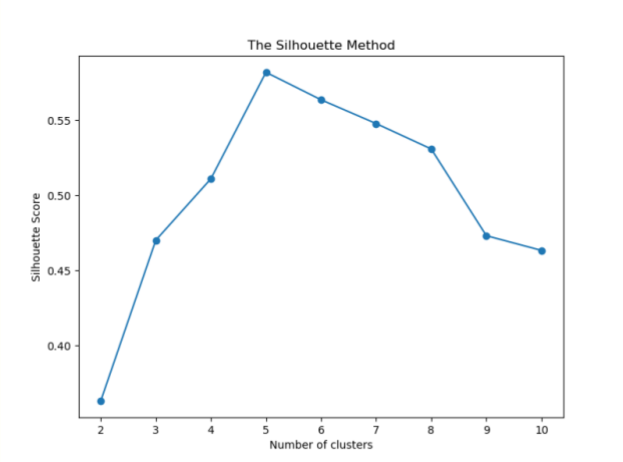
Finally, the tree model was optimized using Cost Complexity Pruning (ccp_alpha). The post-pruned model achieved a strong balance between accuracy and complexity, with significant predictive power (99.5%).

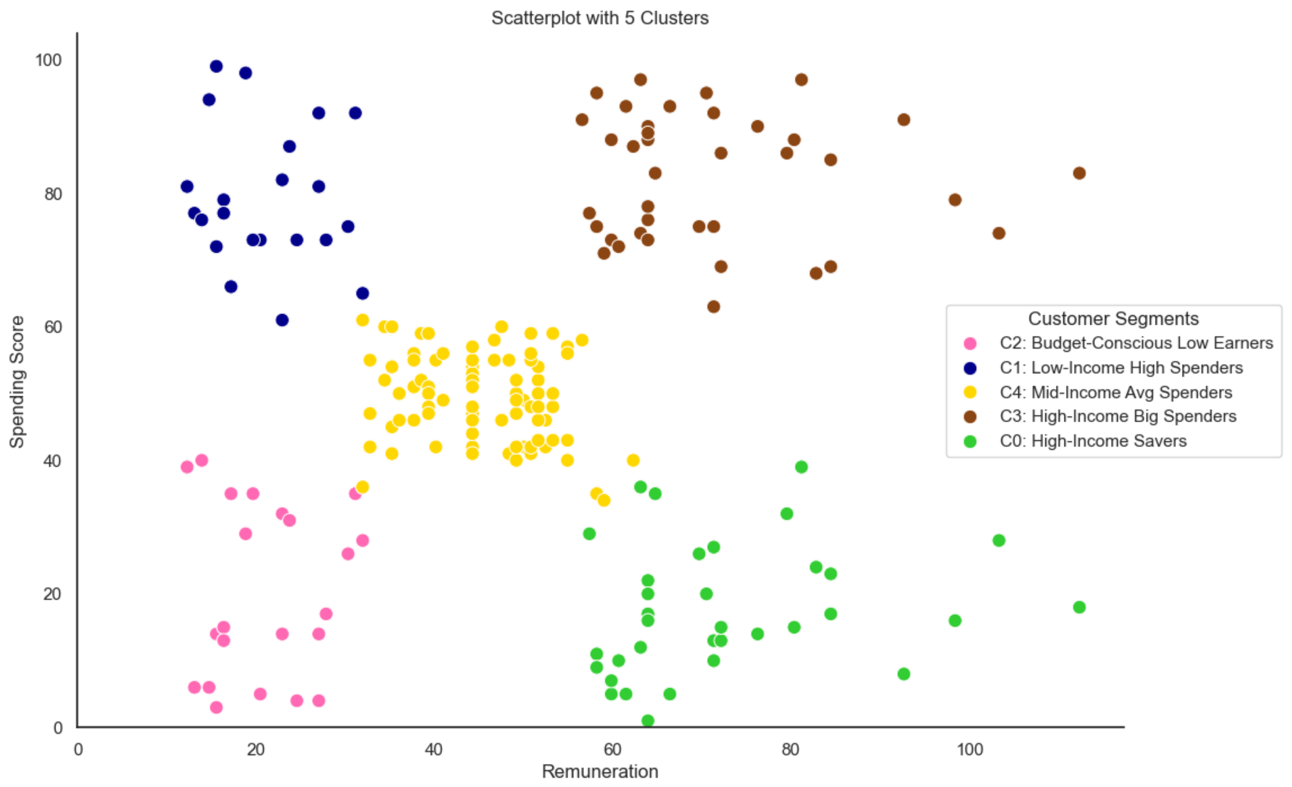


Model Diagnostic Plots - Decision Tree Model

3. Customer Segmentation Using K-Means Clustering in Python

K-Means clustering was applied to segment customers, based on remuneration and spending scores. Since these variables had different ranges (standard deviations: remuneration £23,120K; spending score 26.09), they were normalized using StandardScaler. The silhouette and elbow methods indicated that five clusters (k = 5) provided the most distinct and well-defined customer groups, particularly among higher-income customers.

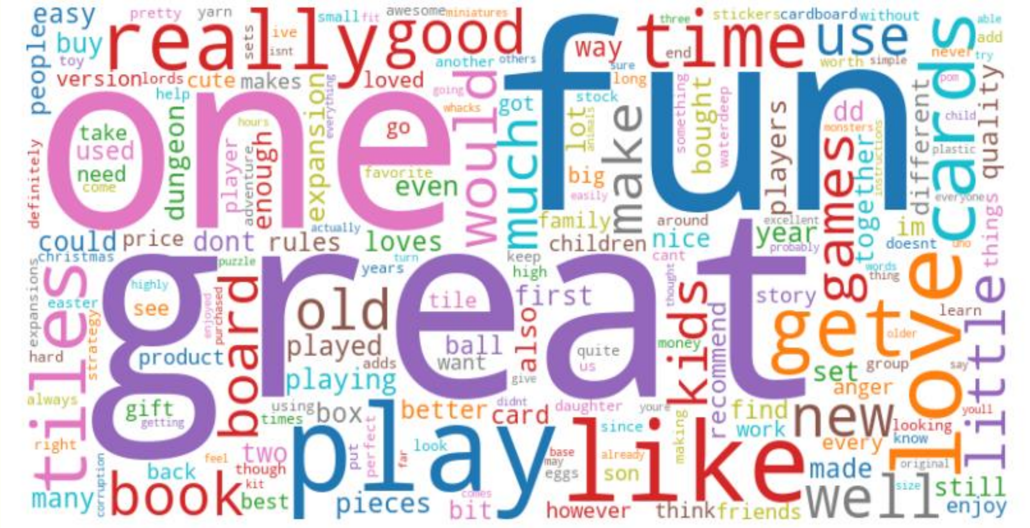




4. Natural Language Processing (NLP) and Sentiment Analysis

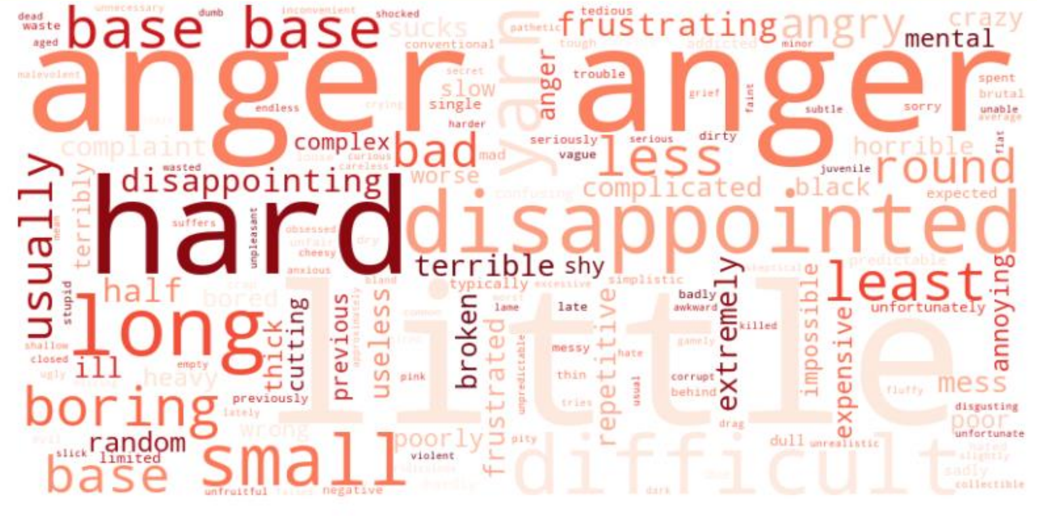
Customer reviews and review summaries were analysed for sentiment using Python’s *TextBlob* library. Duplicates were removed to ensure data quality, and tokenization was performed to break the text into individual words for further analysis. The WordCloud library visualized frequently occurring words, offering a high-level view of customer sentiment.

To improve accuracy, common ‘stopwords’ and domain-specific terms like "game" were removed. Polarity (ranging from -1 for negative to 1 for positive sentiment) and subjectivity (ranging from 0 for objective to 1 for subjective) were extracted to gain deeper insight into customer opinions.



WordCloud – Overall Customer Reviews

While both review columns were initially analysed, the full customer reviews provided richer insights and were thus prioritized for polarity analysis. TextBlob's ability to handle longer reviews performed better than VADER for this dataset. Positive reviews typically featured terms like "fun" and "great," while negative reviews highlighted issues such as "complexity" and "difficulty with instructions."



WordCloud – Negative Customer Reviews based on Polarity

KEY INSIGHTS

- **Decision tree outperforms MLR with higher accuracy and lower MSE**

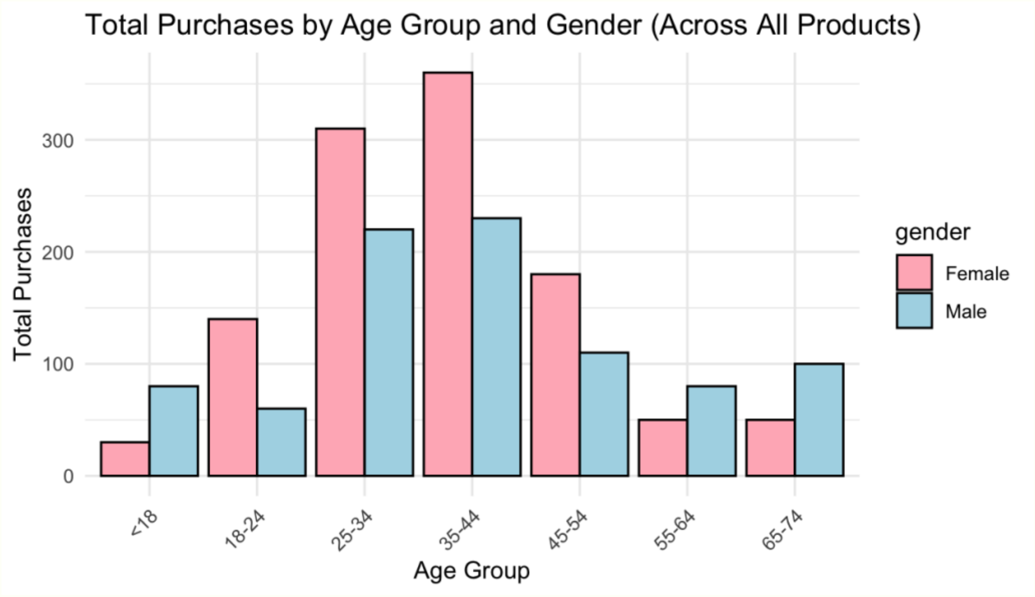
  The MLR model explains 83% of loyalty point variance but struggles with non-linearity, resulting in a high MSE of 275,278, limiting predictive accuracy. In contrast, the pruned Decision Tree Regressor captures 99.5% of variance with a much lower MSE of 7,878.91, improving accuracy and handling non-linear relationships.

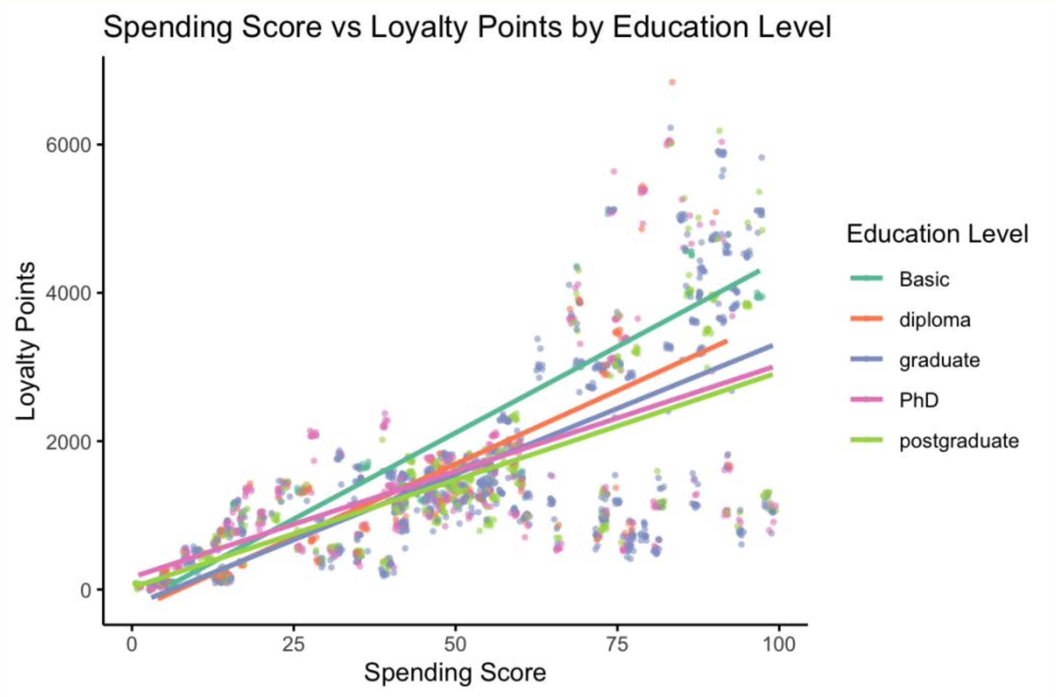
- **Positive sentiment and Five customer groups**

  K-means clustering identified 5 distinct customer groups with varied spending behaviours. Sentiment analysis revealed over 80% positive reviews, with negative feedback frequently related to product complexity.

- **Highest spenders aged 25- 44, education impacts behaviour**

  The largest customer group was aged 25- 44, with spending declining after age 50. Customers with basic education had the highest spending scores, while those with graduate and PhD degrees showed lower, but more consistent, spending behaviour.



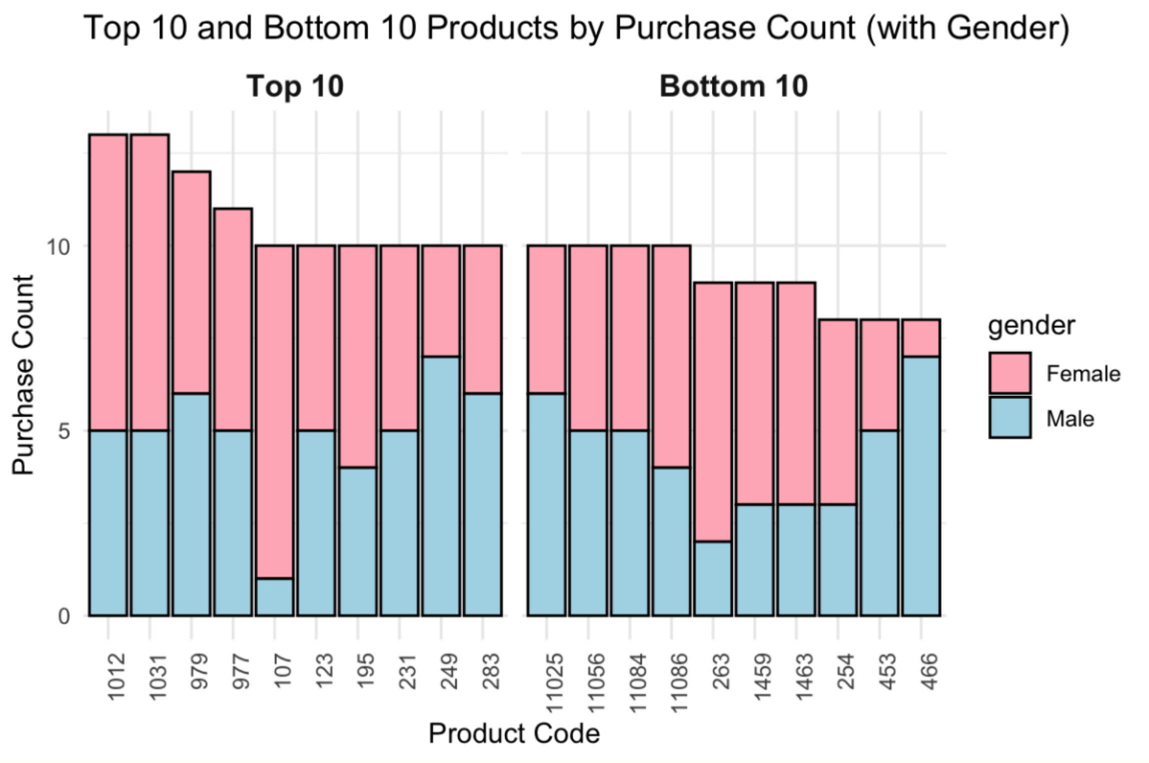


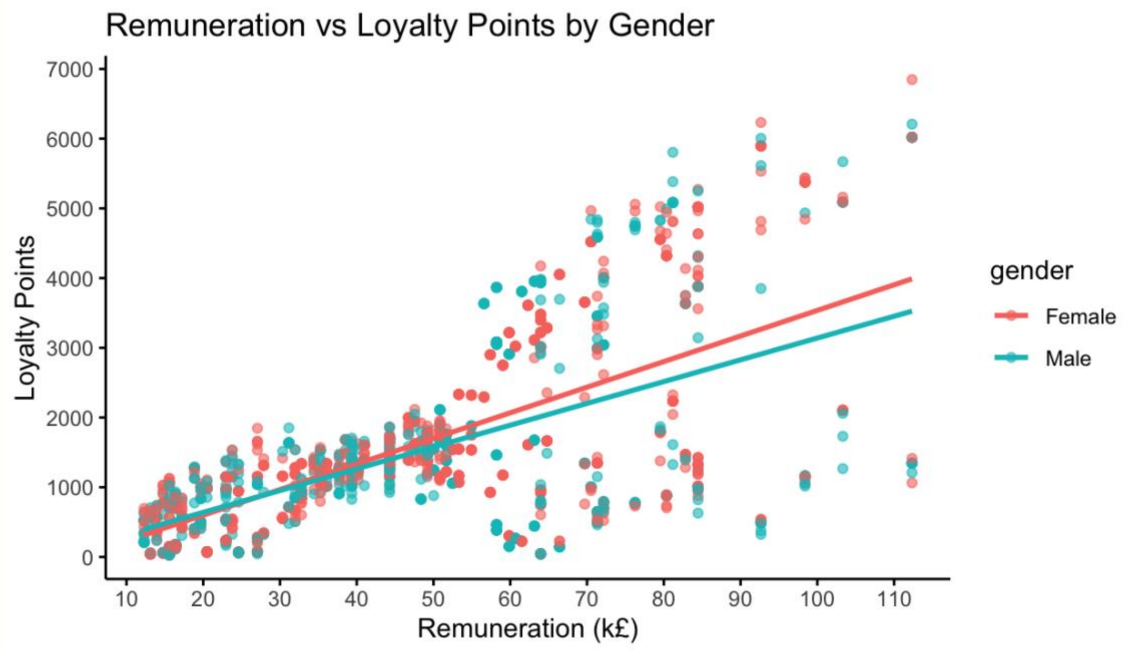
- **Higher Loyalty Points for Basic Education Level**

  Customers with a basic education accumulate significantly more loyalty points compared to those with higher education, indicating potential differences in engagement with loyalty programs.

- **Females dominate loyalty points, prefer products 1012/1031**

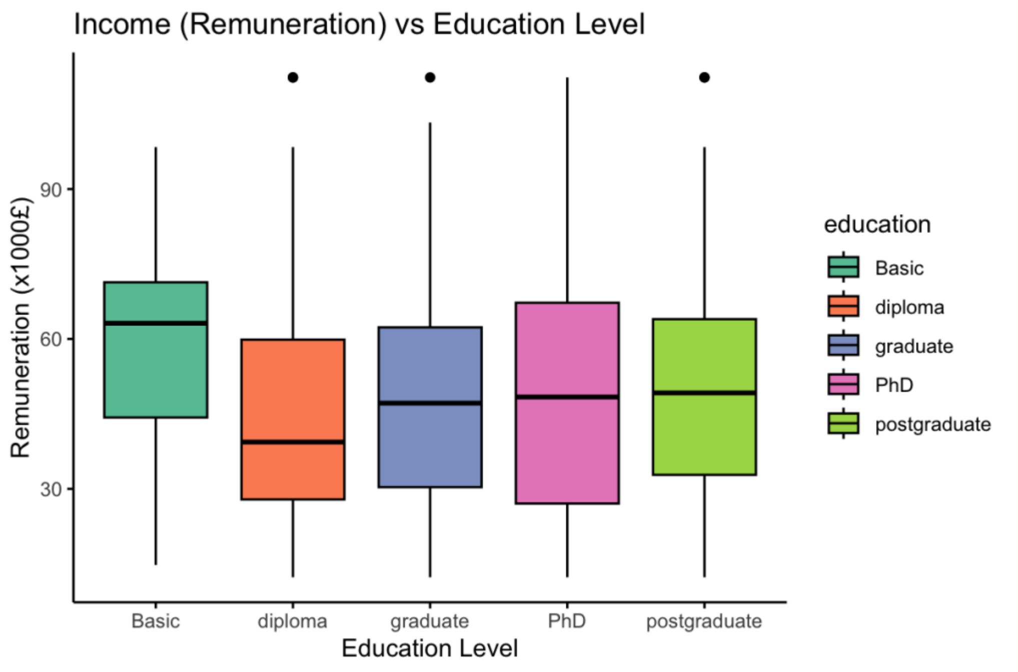
  Females, representing 56% of the customer base, accumulate more loyalty points than males. Products 1012 and 1031 were particularly popular with female customers. Low-performing products like 11056 and 11084 had balanced gender engagement but low overall interest.

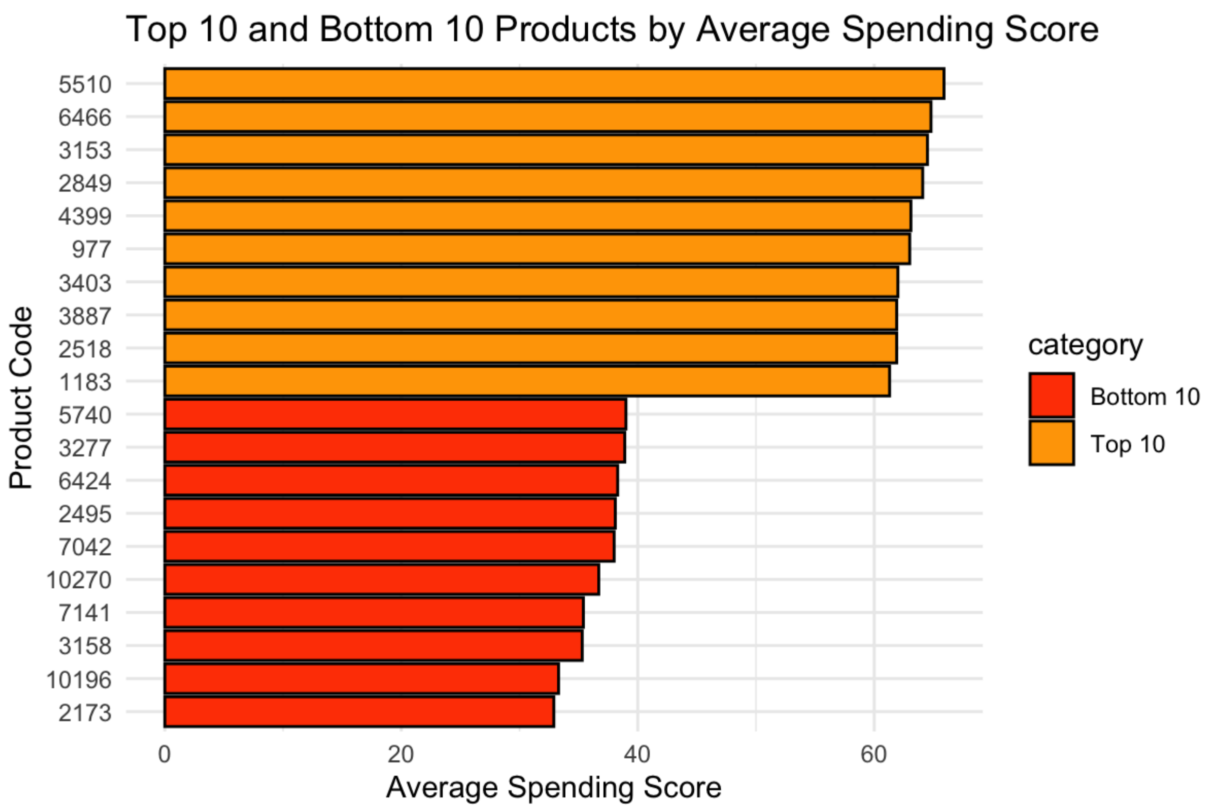

Top 10 and Bottom 10 Products by Purchase Count (with Gender)
Top 10
Bottom 10
Purchase Count
0
5
10
Product Code
1012
1031
979
977
107
123
195
231
249
283
11025
11056
11084
11086
263
1459
1463
254
453
466
gender
Female
Male


Remuneration vs Loyalty Points by Gender
Loyalty Points
0
1000
2000
3000
4000
5000
6000
7000
Remuneration (k£)
10
20
30
40
50
60
70
80
90
100
110
gender
Female
Male

- **High-income, educated customers prefer premium products, ideal for targeted loyalty rewards.**

  High-income customers (earning above £50,000), particularly those with higher education, favour "possible premium products" like 5510 and 6466, making them ideal candidates for targeted loyalty rewards.





RECOMMENDATIONS

1. **Use Customer Segmentation & Decision Tree model for Loyalty Predictions**

   Segment customers by behaviour, apply Decision Tree analysis to each group for accurate loyalty point predictions, and craft personalised marketing campaigns based on these insights.

2. **Target 25-44 female customers**

   Focus marketing efforts on this demographic, which shows the highest purchase activity. Products like 1012 and 1031 can be tied to loyalty rewards to increase repeat purchases.

3. **Engage Graduate and PhD Customers**

   Despite their consistent spending, graduate and PhD customers are under-engaged with loyalty programs. Offering exclusive benefits and premium rewards could increase their participation.

4. **Promote high-performing products**

   Prioritize marketing campaigns for premium products like 5510, 6466, and 9080, which appeal to high-income customers and serve as key drivers of sales.

5. **Reassess low-performing products**

   Products like 10270 and 2173 should either be promoted more aggressively or phased out. Further analysis is recommended to assess low engagement and whether targeted promotions could improve performance.

6. **Tailor rewards to boost customer engagement**

   Tailor loyalty programs to specific customer segments. Offer premium rewards to high-income customers and value-driven incentives to lower-income segments to boost overall engagement.

7. **Implement Feedback System for actionable insights**

   Introduce structured feedback system (online surveys/product ratings/ Net Promote surveys) targeting product quality, usability, and satisfaction. This will yield more actionable data for future sentiment analysis using advanced NLP models.

8. **Improve product categorisation**

   Record detailed product categories (e.g., toys, books, board games) alongside product codes to enable more granular analysis of purchasing trends by gender, age, and seasonality.

Reference

1. Sarthak, 2020, *Cost Complexity Pruning in Decision Trees Using Scikit-Learn*, Analytics Vidhya, viewed 20 September 2024, https://www.analyticsvidhya.com/blog/2020/10/cost-complexity-pruning-decision-trees/..

2. Dipanjan Sarkar, 2021, *Sentiment Analysis with VADER or TextBlob*, Analytics Vidhya, viewed 24 September 2024, https://www.analyticsvidhya.com/blog/2021/01/sentiment-analysis-vader-or-textblob/..

3. CRAN, n.d., *The Comprehensive R Archive Network*, viewed 28 September 2024, https://cran.mirror.ac.za/.

4. Matplotlib, n.d., *Visualization with Python*, viewed 30 September 2024, https://matplotlib.org/stable/index.html.

5. NICE Satmetrix, n.d., *Net Promoter Score*, viewed 25 September 2024, https://www.netpromoter.com/know/.

6. Pandas, 2024, *Pandas Documentation (Version 2.2.2)*, viewed 10 April 2024, https://pandas.pydata.org/docs/user_guide/index.html#user-guide.

7. Seabold, S. and Perktold, J., n.d., *Statsmodels: Statistical Models in Python*, viewed 30 September 2024, https://www.statsmodels.org/stable/index.html.

8. SurveyMonkey, n.d., *The Complete Guide to Customer Segmentation*, viewed 15 September 2024, https://uk.surveymonkey.com/market-research/resources/the-complete-guide-to-customer-segmentation/.

9. Tidyverse, n.d., *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, viewed 27 September 2024, https://ggplot2.tidyverse.org/.

10. Tidyverse, n.d., *Tidyverse*, viewed 30 September 2024, https://www.tidyverse.org/.