Nanwani_Nanik_DA301_Assignment_Report.pdf.

Predictive Analytics for Sales & Loyalty Optimisation at Turtle Games

BUSINESS OVERVIEW

Turtle Games is a global manufacturer and retailer specialising in books, board games, video games, and toys. The company produces and sells its own branded products while also sourcing items from external suppliers.

Business Objective:
To improve overall sales performance by leveraging data to enhance customer insights, optimise the loyalty programme, and refine marketing strategies.

Analytical Objectives:
This analysis focuses on two core areas:

1. **Customer Behaviour and Loyalty Programme Analysis**
   - What is the relationship between customer demographics and spending behaviour?
   - How can segmentation using remuneration and spending scores inform marketing decisions ?
   - What insights can be drawn from customer reviews about their experiences with Turtle Games?
2. **Predictive Modelling and Programme Effectiveness**
   - Which model—Multiple Linear Regression or Decision Tree—more accurately predicts loyalty points accumulation based on customer features?
   - How can Turtle Games improve its loyalty programme and enhance data collection practices?

ANALYTICAL APPROACH

Data Overview:
The analysis is based on demographic and sales data from 2,000 customers. A provided metadata file was used to better understand the dataset and rename certain variables for clarity. During data cleaning (conducted in Python and R), duplicates and missing values were addressed. Uniform variables—such as language (English) and platform (Web)—were excluded due to limited analytical value.

Limitations:
A key limitation is the lack of product-level detail; only product codes were provided, offering minimal descriptive information.

1. Exploratory and Regression Analysis in Python and R

Python

The dataset was imported into Python, using libraries such as NumPy, pandas, scikit-learn, and statsmodels for descriptive and exploratory analysis. Data visualisation was carried out with matplotlib and seaborn.

After data wrangling, simple linear regression models were developed to explore relationships between quantitative variables. The dependent variable was *loyalty points*, with *spending score*, *remuneration* (income in thousands), and *age* as independent variables.

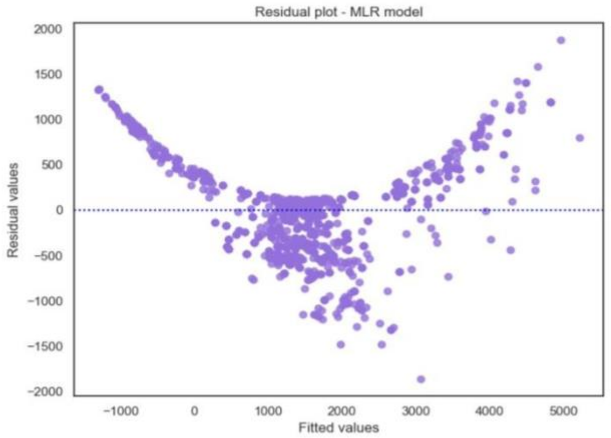
A Multiple Linear Regression (MLR) model was then built using the Ordinary Least Squares (OLS) method. Key OLS assumptions—linearity, independence, homoscedasticity, and normality of residuals—were tested to ensure model validity. Model reliability was evaluated using a 70/30 train-test split.

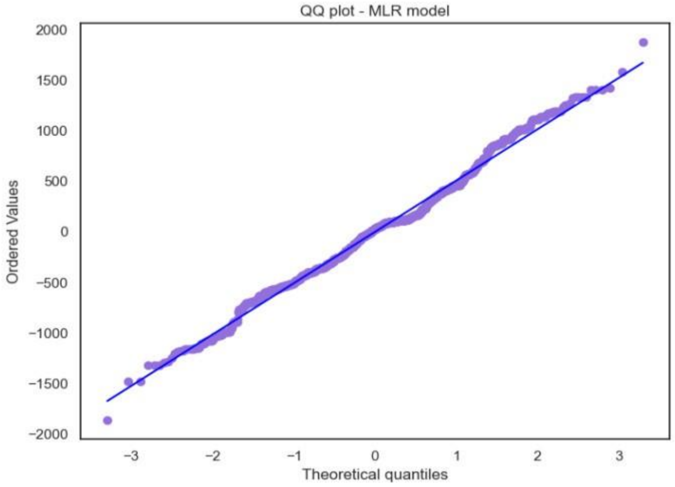
Results showed that *spending score* and *remuneration* were the strongest predictors of loyalty points. In contrast, *age* was not statistically significant.

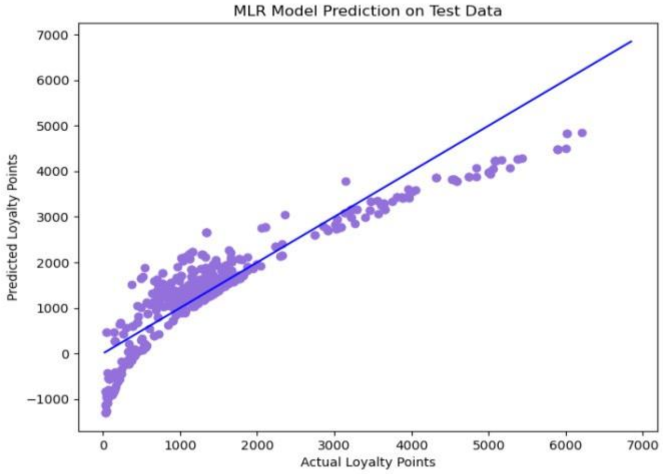
R

A similar process was conducted in R, using packages including readr, dplyr, ggplot2, car, and lmtest. Boxplots, scatter plots, and histograms were used to visualise relationships—particularly among gender, education, age, and product codes (n = 200).

Building on insights from the Python analysis, an additional MLR model was created in R excluding *age* as a predictor. Both Python and R models that included *age* showed marginally better fit, indicated by slightly higher R-squared values and lower error metrics, than the R model excluding *age*.







However, both models exhibited signs of heteroscedasticity and non-normal residuals, indicating potential structural issues within the data—despite age offering a minor contribution as a predictor.

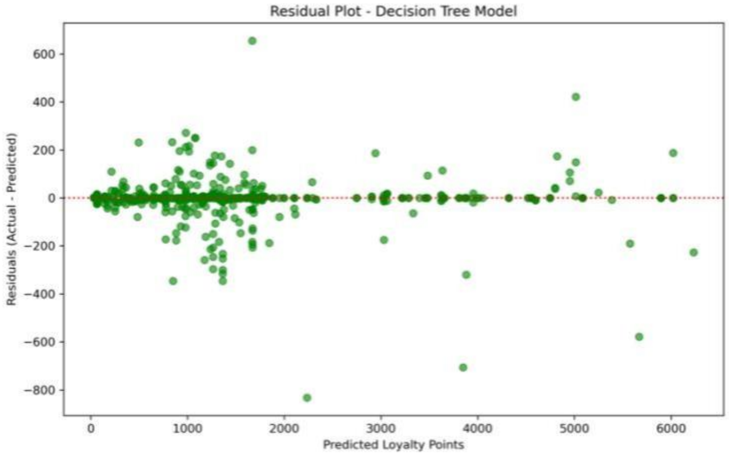
2. Decision Tree Model in Python

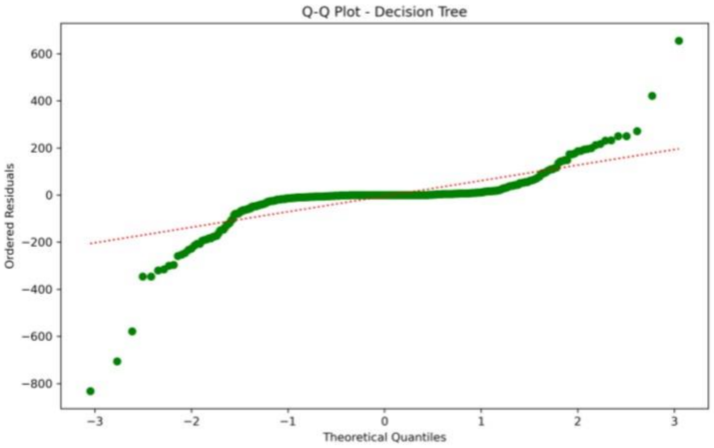
A `DecisionTreeRegressor` was used to explore the impact of both quantitative and categorical variables on loyalty points. The original five levels of the *education* variable were consolidated into three broader categories for simplicity, and dummy encoding was applied to both *education* and *age*. The dataset was split 70/30 for training and testing.

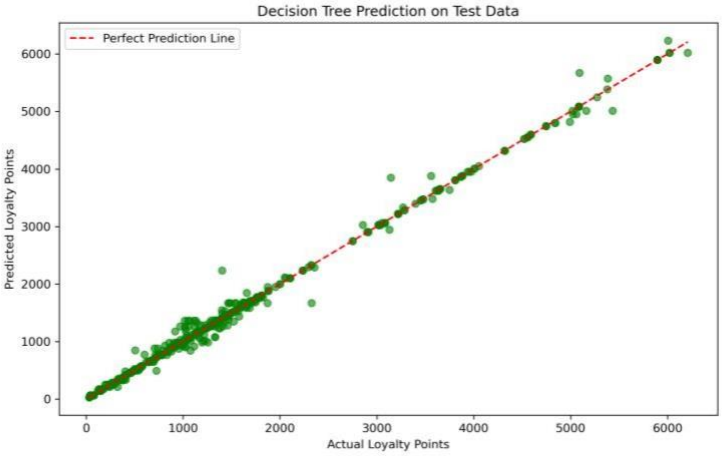
The initial model fit the training data perfectly, which indicated overfitting—confirmed by higher error metrics on the test set. To address this, fivefold cross-validation was performed, which confirmed consistent model performance across training subsets.

Pre-pruning techniques (`max_depth=5`, `min_samples_leaf=10`) were applied to reduce complexity. However, this introduced some underfitting, as reflected in increased test error due to the pruning constraints.

To further refine the model, **Cost Complexity Pruning** (`ccp_alpha`) was used. The resulting post-pruned model achieved an optimal trade-off between complexity and accuracy, demonstrating strong predictive power with an accuracy of 99.5%.



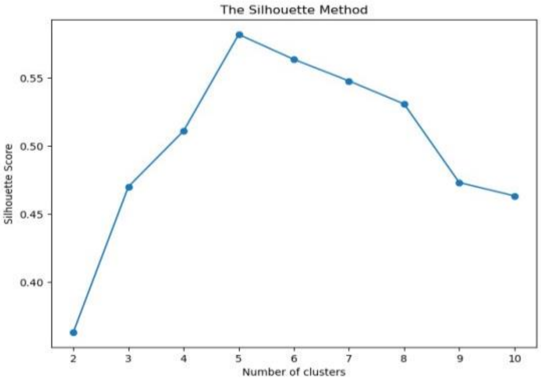


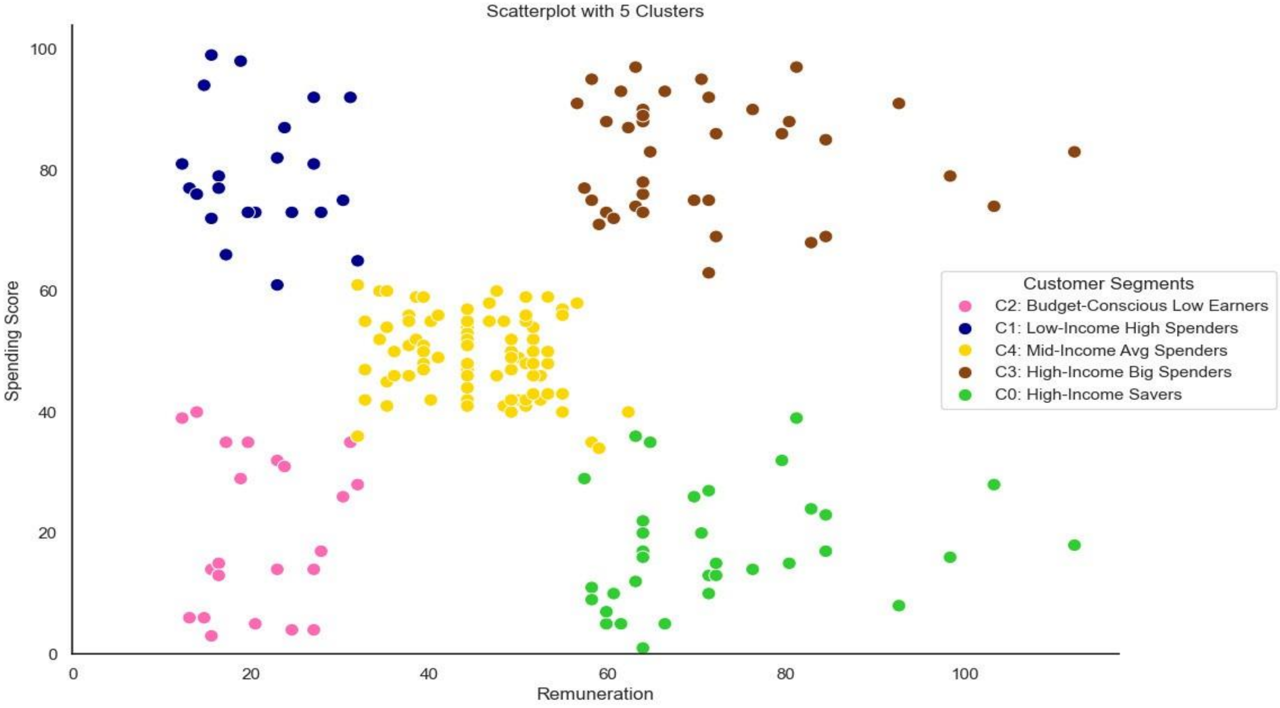


3. Customer Segmentation using K-Means Clustering in Python

K-Means clustering was used to segment customers based on their *remuneration* and *spending scores*. Given the difference in scale between the two variables (standard deviations: remuneration = 23.12K; spending score = 26.09), the data was normalised using StandardScaler to ensure balanced weighting.

Both the **silhouette** and **elbow methods** confirmed that **five clusters (k = 5)** offered the most distinct and meaningful segmentation—particularly among higher-income customer groups—enabling clearer identification of behavioural patterns.





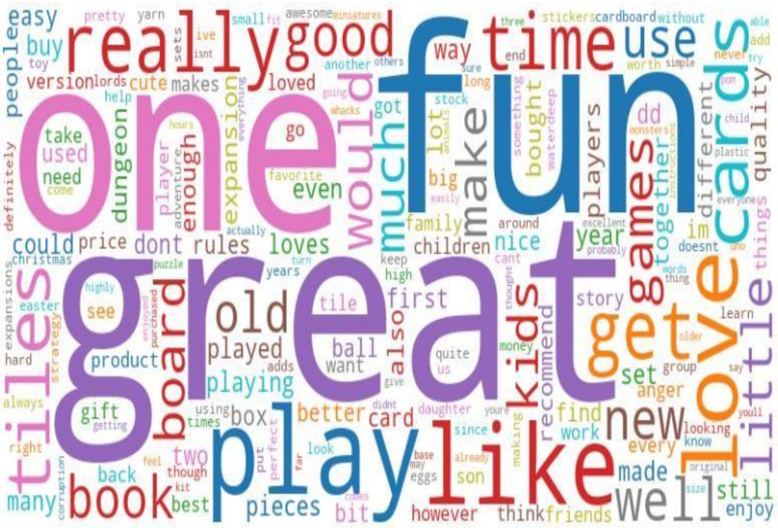
4. Natural Language Processing (NLP) and Sentiment Analysis

Customer reviews and review summaries were analysed using Python's **TextBlob** library to extract sentiment insights. To ensure data quality, duplicate entries were removed, and tokenisation was applied to break the text into individual words for further analysis. The **WordCloud** library was used to visualise the most frequently occurring terms, providing a high-level overview of customer sentiment.

To enhance the clarity of insights, common stopwords and domain-specific terms such as *"game"* were removed. Two sentiment metrics were extracted:

- **Polarity**: Ranges from -1 (completely negative) to 1 (completely positive)
- **Subjectivity**: Ranges from 0 (objective) to 1 (subjective)

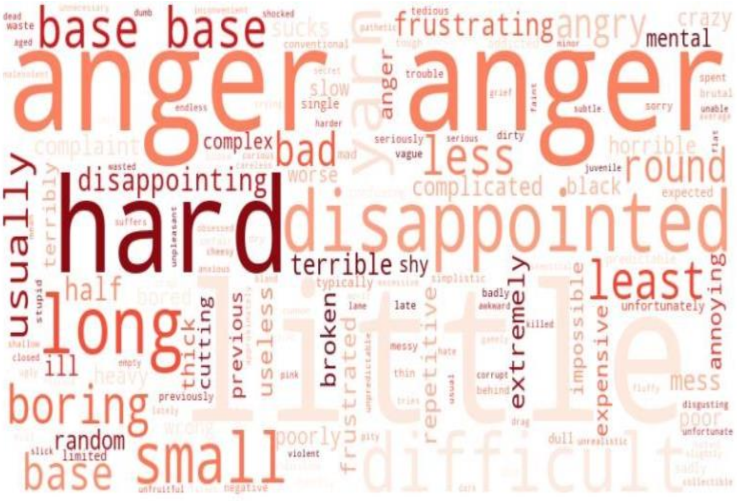
WordCloud – Overall Customer Reviews



WordCloud – Negative Customer Reviews (Filtered by Polarity)

Although both the summary and full review columns were initially considered, the full customer reviews yielded richer sentiment insights and were prioritised for analysis. **TextBlob** was chosen over VADER due to its better handling of longer and more nuanced text within the dataset.

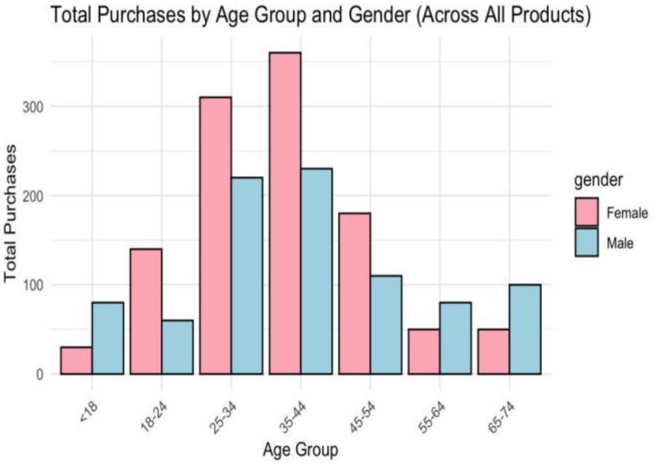
Positive reviews commonly included words such as *"fun"* and *"great,"* indicating strong customer satisfaction. In contrast, negative reviews often referenced issues like *"complexity"* and *"difficulty with instructions."*



This visual isolates terms from reviews with negative polarity, helping Turtle Games pinpoint recurring customer pain points and areas for improvement.

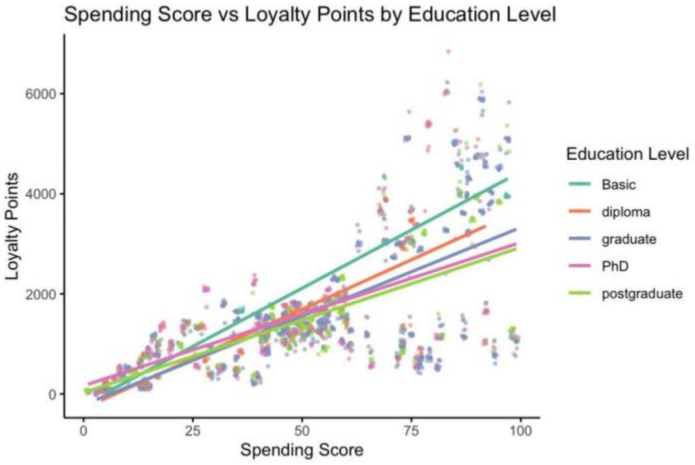
KEY INSIGHTS

- **Decision Tree Outperforms MLR in Accuracy and Predictive Power**
  The Multiple Linear Regression (MLR) model explains 83% of the variance in loyalty points but struggles with non-linearity, resulting in a high Mean Squared Error (MSE) of 275,278. In comparison, the pruned Decision Tree Regressor captures 99.5% of the variance with a significantly lower MSE of 7,878.91—demonstrating improved accuracy and better handling of complex, non-linear relationships.

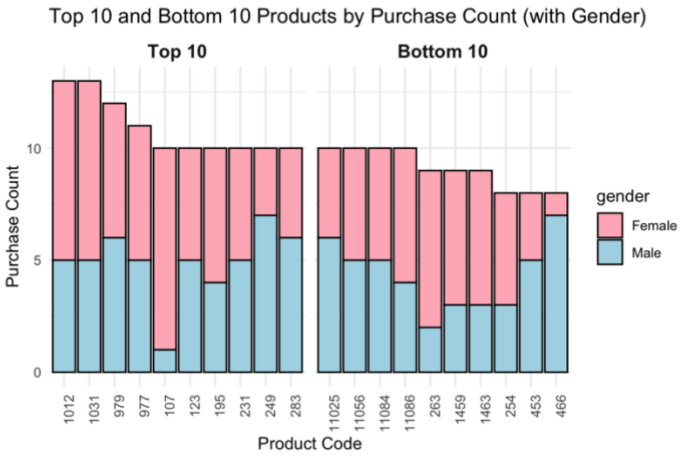


- **Strong Positive Sentiment and Five Distinct Customer Segments**
  K-Means clustering revealed five unique customer segments with diverse spending behaviours. Sentiment analysis showed that over 80% of customer reviews were positive, while most negative feedback was related to product complexity.

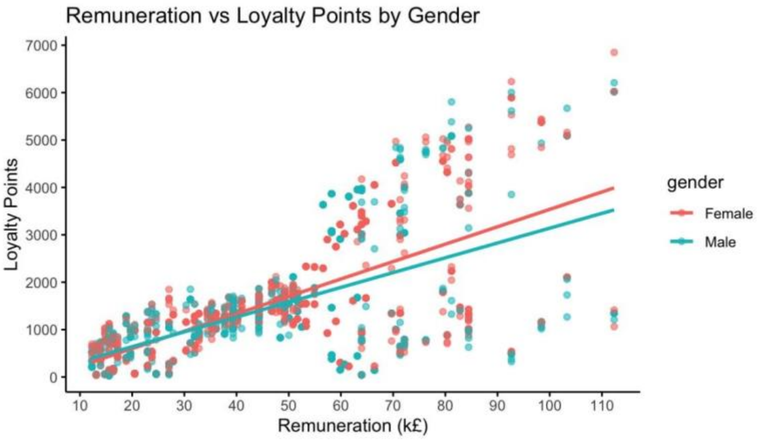
- **Top Spenders Aged 25–44; Education Level Influences Spending**
  The largest customer group fell within the 25–44 age range, with spending declining notably after age 50. Customers with only basic education recorded the highest spending scores, whereas those with graduate or PhD qualifications tended to spend less overall, but demonstrated more consistent purchasing behaviour.



- **Higher Loyalty Points Among Customers with Basic Education** Customers with a basic education level accumulate significantly more loyalty points than those with higher education qualifications. This suggests differing levels of engagement with the loyalty programme, potentially reflecting varying motivations or purchasing patterns.



- **Female Customers Lead in Loyalty Points and Product Preference** Women make up 56% of the customer base and consistently accumulate more loyalty points than men. Products **1012** and **1031** are particularly favoured by female customers. In contrast, low-performing products like **11056** and **11084** showed balanced gender interest but generated limited overall engagement.



- **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. **Leverage Customer Segmentation with Decision Tree Modelling**
   Segment customers based on behavioural patterns and apply Decision Tree models within each group to more accurately forecast loyalty point accumulation. These insights can inform personalised marketing strategies tailored to each segment's preferences.
2. **Prioritise Female Customers Aged 25–44**
   Concentrate marketing efforts on this high-value demographic, which demonstrates the strongest purchase behaviour. Link popular products such as 1012 and 1031 to loyalty incentives to drive retention and repeat purchases.
3. **Boost Engagement Among Highly Educated Customers**
   Although graduate and PhD holders exhibit consistent spending, they appear under-engaged with the loyalty programme. Introducing exclusive perks or premium-tier rewards may help capture their interest and increase participation.
4. **Promote High-Performing Products Strategically**
   Elevate visibility of top-performing items like 5510, 6466, and 9080—products that resonate strongly with high-income customers and significantly contribute to revenue.
5. **Re-evaluate Underperforming Products**
   Items such as 10270 and 2173 warrant further review. Consider whether targeted promotions could improve their performance or whether they should be retired from the catalogue.
6. **Customise Loyalty Incentives by Customer Segment**
   Design loyalty rewards that reflect customer segment profiles—offering premium rewards to high-income groups, and value-focused incentives to budget-conscious customers—to enhance overall engagement.
7. **Establish a Structured Customer Feedback System**
   Implement feedback mechanisms (e.g., product reviews, surveys, Net Promoter Score) focused on product experience and satisfaction. This will provide valuable data for future sentiment analysis using advanced NLP techniques.
8. **Enhance Product Categorisation for Deeper Insights**
   Improve the dataset by categorising products by type (e.g., toys, books, board games) alongside product codes. This will enable more detailed analysis of purchasing trends across demographics and seasons.