A. Background

Turtle Games, a global gaming leader, aimed to boost sales by analysing customer engagement, loyalty data, and segmentation insights. I led a data-driven analysis using Python and R to deliver actionable insights that supported their marketing and retention strategies.

B. Business question

- How are loyalty points engaged with and accumulated?
- How can customers be segmented into groups, and which groups can be targeted by the marketing department?
- How can text data be utilized to inform marketing campaigns and support business improvements?
- Can descriptive statistics be used to provide insights into the suitability of the loyalty points data for creating predictive models?

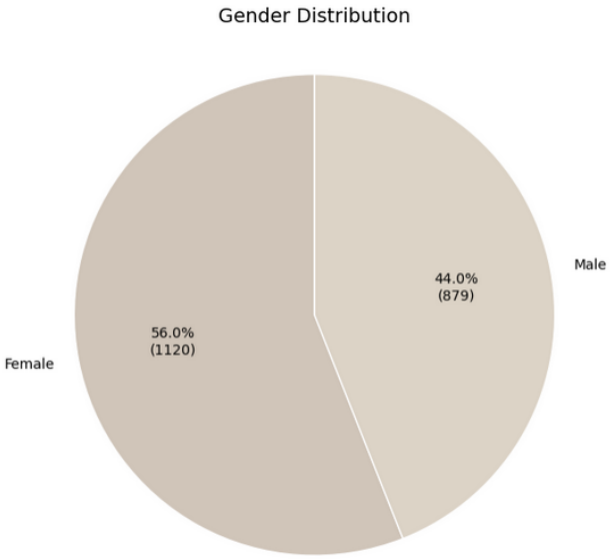
C. Analytic Approach

Part 1: Data Cleaning and Wrangling

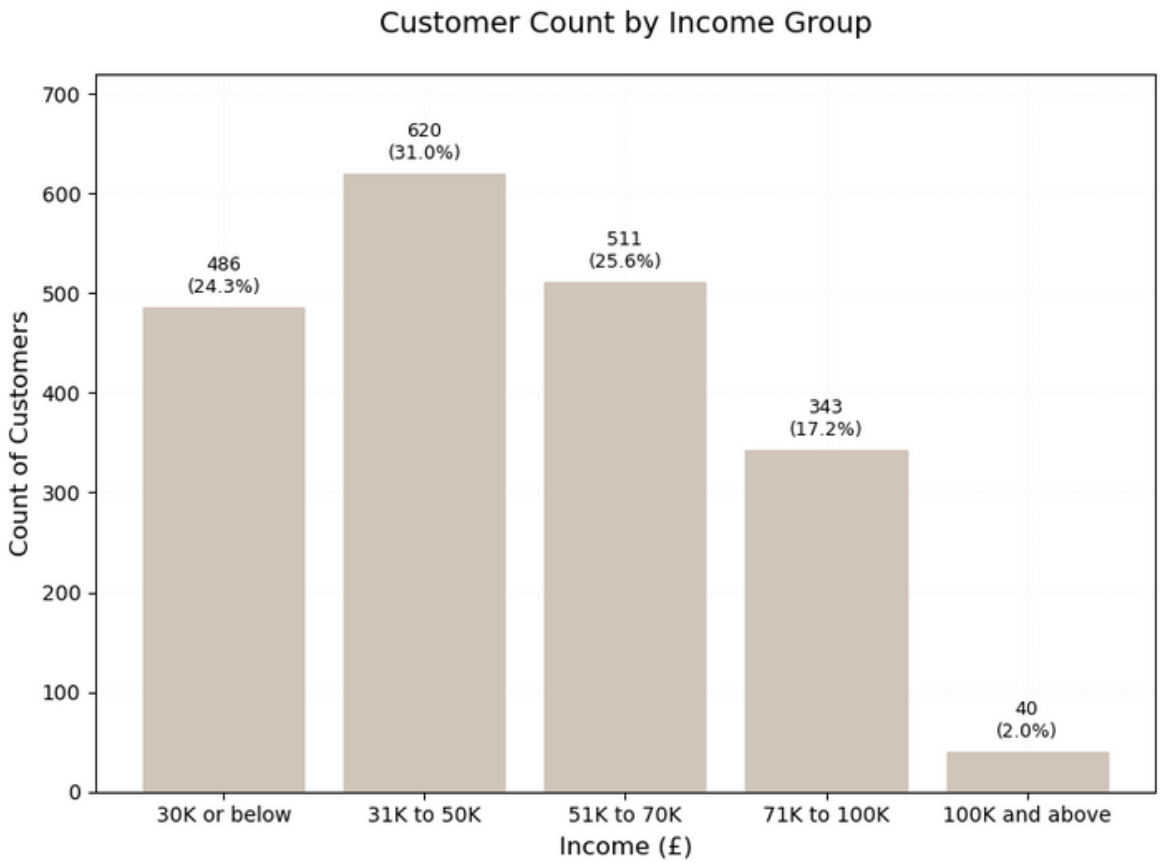
Ensured data completeness by checking for missing values using Python libraries. The dataset was generally clean, with no significant null entries or inconsistencies. Additionally, redundant or irrelevant columns were identified and removed or renamed to streamline the analysis and improve data efficiency for subsequent processing and modelling.

Part 2: Customer Demographics:

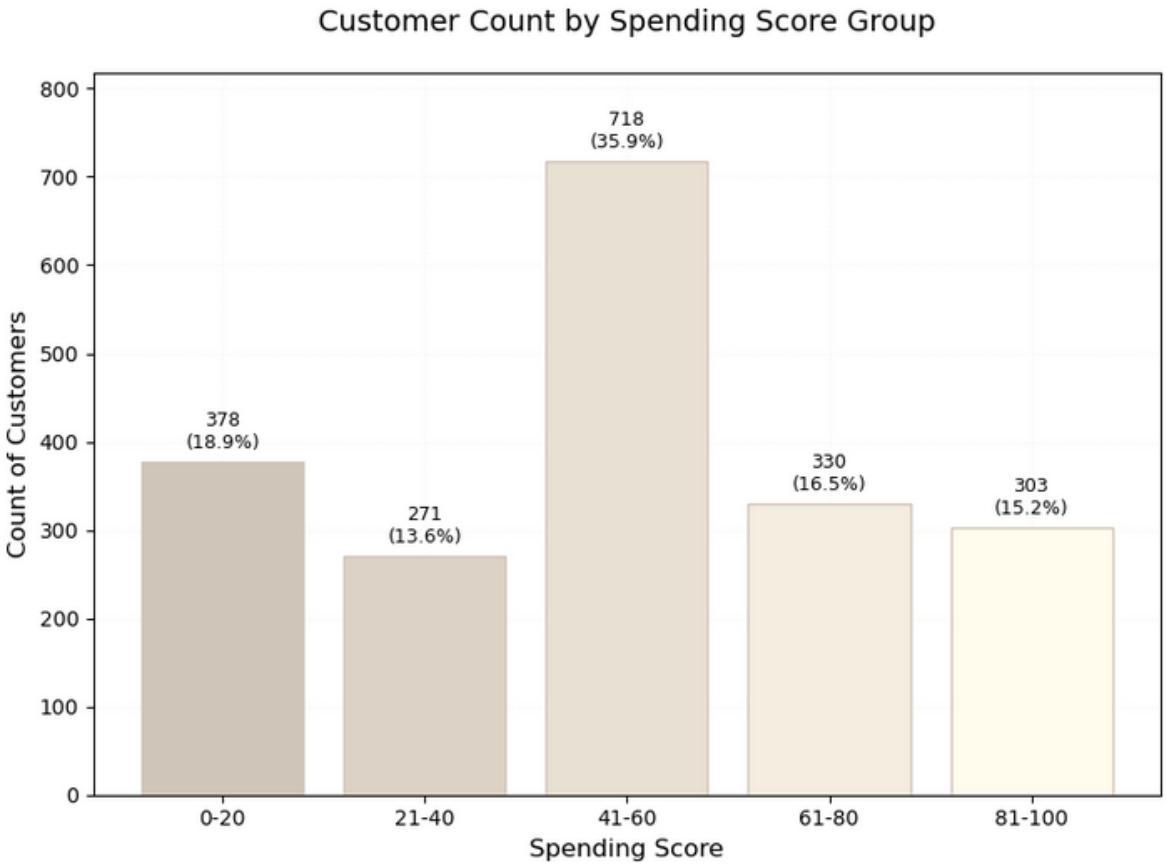
The customer demographic data reveals a diverse group:



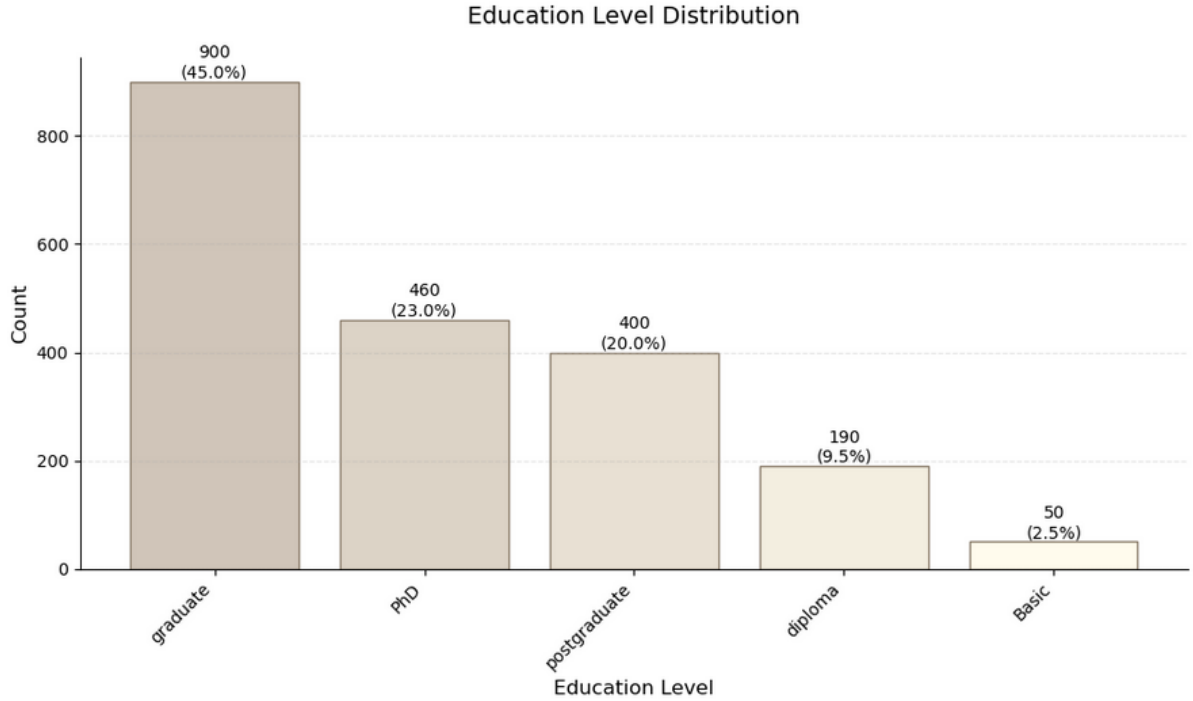
Gender: 56% female (1,120) and 44% male (879)



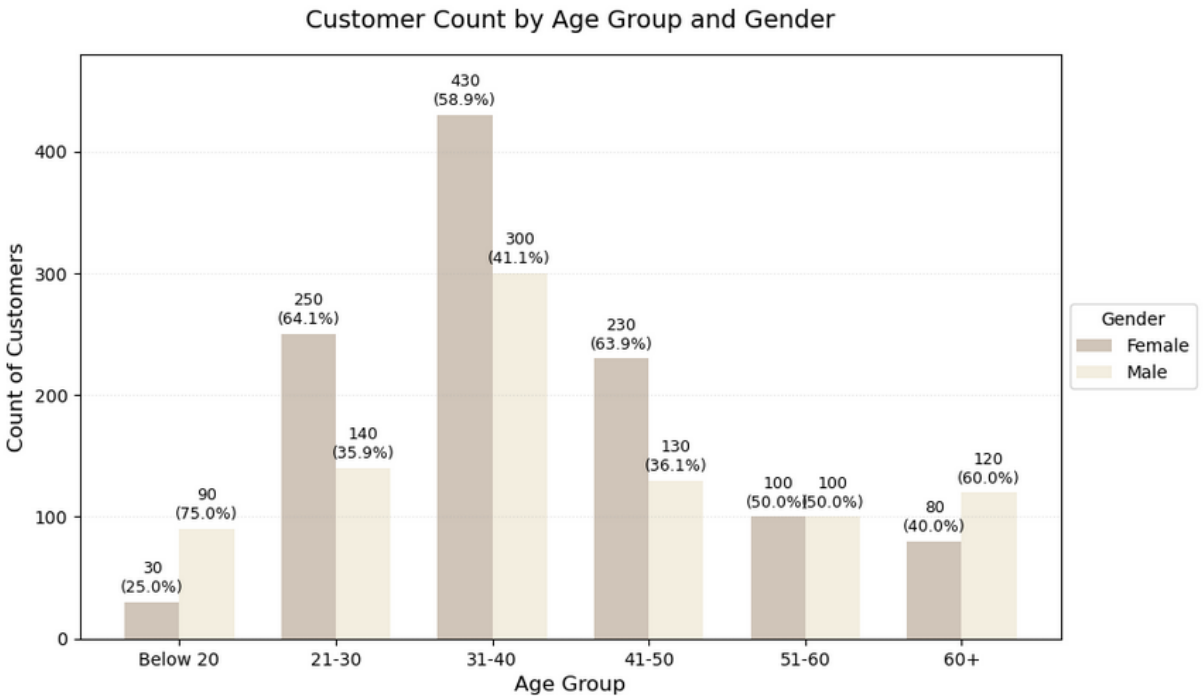
Income levels: mostly between £30K and £70K (over 1,000 customers)



Spending Scores: predominantly in the 41-60 range (around 700 customers)



Education level: most are at the graduate level (900 customers)

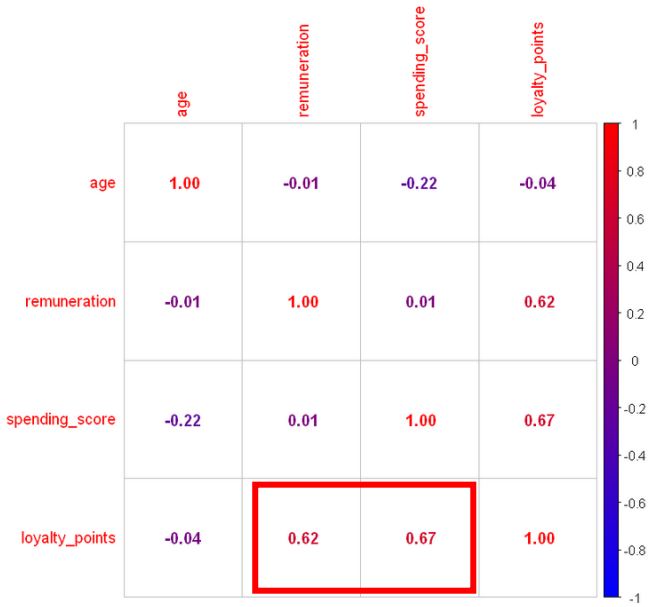


Age distribution: significant portion aged 31-40 (over 600 customers), with females slightly outnumbering males across most age groups.

A more detailed analysis of customer demographics using cluster groups will be conducted in Part 5.

Part3: Regression Model

We conducted a correlation matrix to assess the strength of linear relationships between variables and to check for multicollinearity. The results showed that loyalty points are strongly related to income and spending score. Since all correlation coefficients are below 0.7, multicollinearity is not a significant concern in this case.



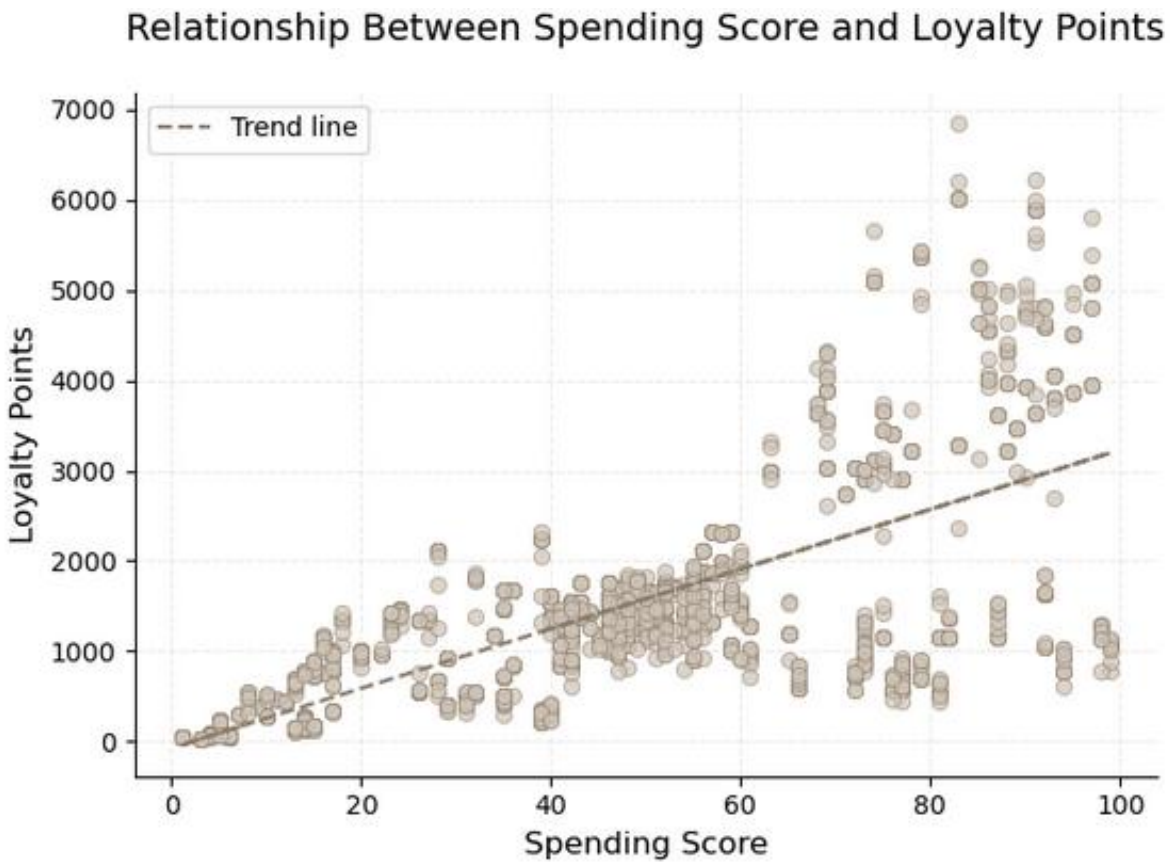
Loyalty points are mainly collected by making purchases. Points are awarded based on the amount spent or specific promotions. Customers may also earn points through app usage, referrals, product reviews, or participating in special events and campaigns offered by the brand. A high loyalty point customer is engaging frequently with Turtle Games.

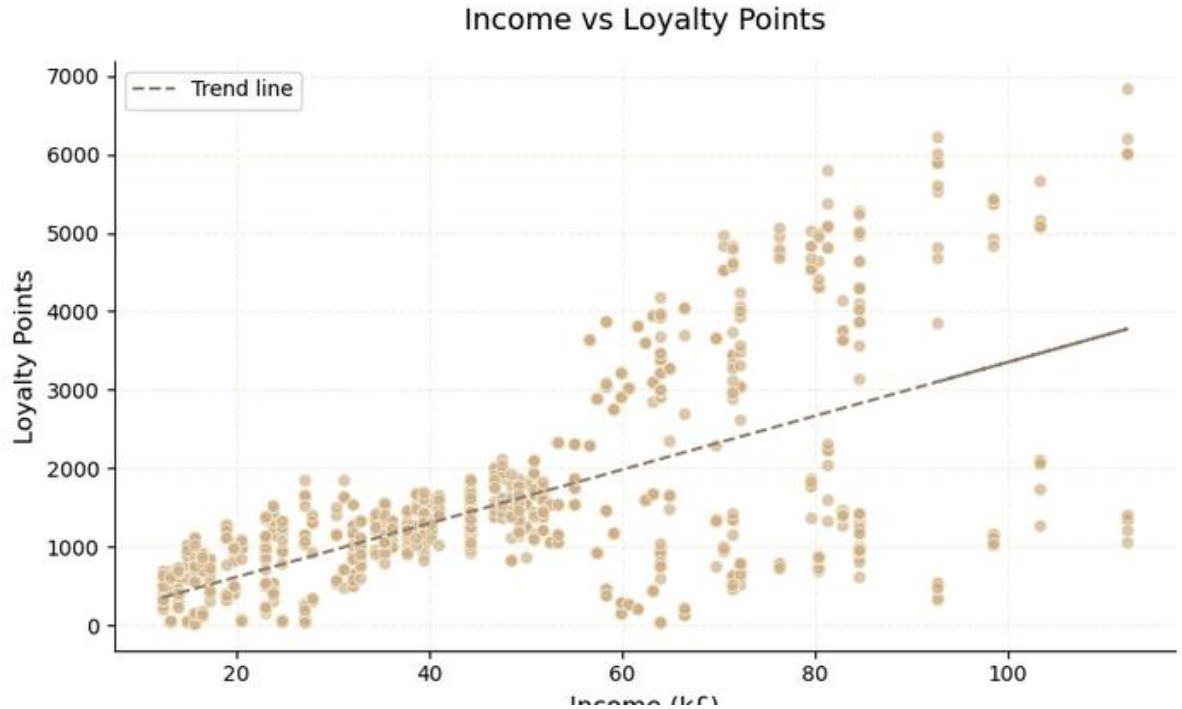
Turtle Games defines a spending score by analysing a customer's purchasing behaviour, typically combining factors such as purchase frequency, transaction amount, and recency.

Income vs. Loyalty Points and *Spending Score vs. Loyalty Points* both show OLS regression results with scatter plots.

In *Income vs. Loyalty Points*, Income (x) has a coefficient of 34.178, R-squared of 0.380, and a significant p-value (0.000), indicating a moderate positive relationship with Loyalty Points (y).

Spending Score vs. Loyalty Points with Spending Score (x), shows a coefficient of 33.0617, a higher R-squared of 0.452, and a significant p-value (0.000), suggesting a stronger positive relationship. Both models have one predictor, so multicollinearity isn't a concern. The scatter plots confirm positive trends, with the *Spending Score vs. Loyalty Points* showing a slightly tighter fit.

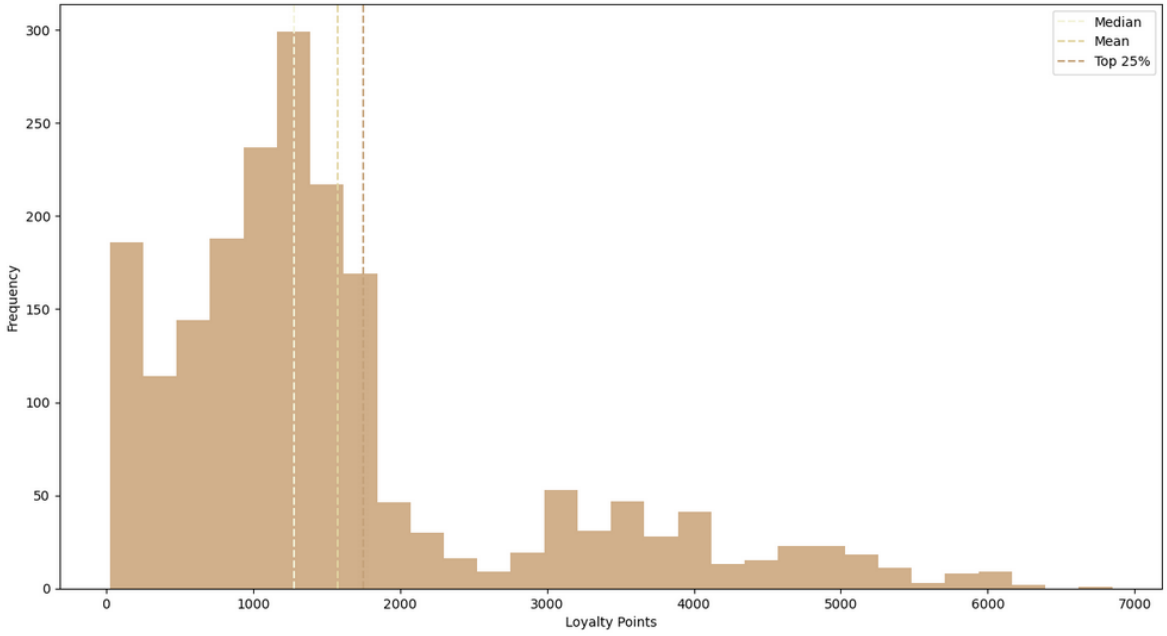


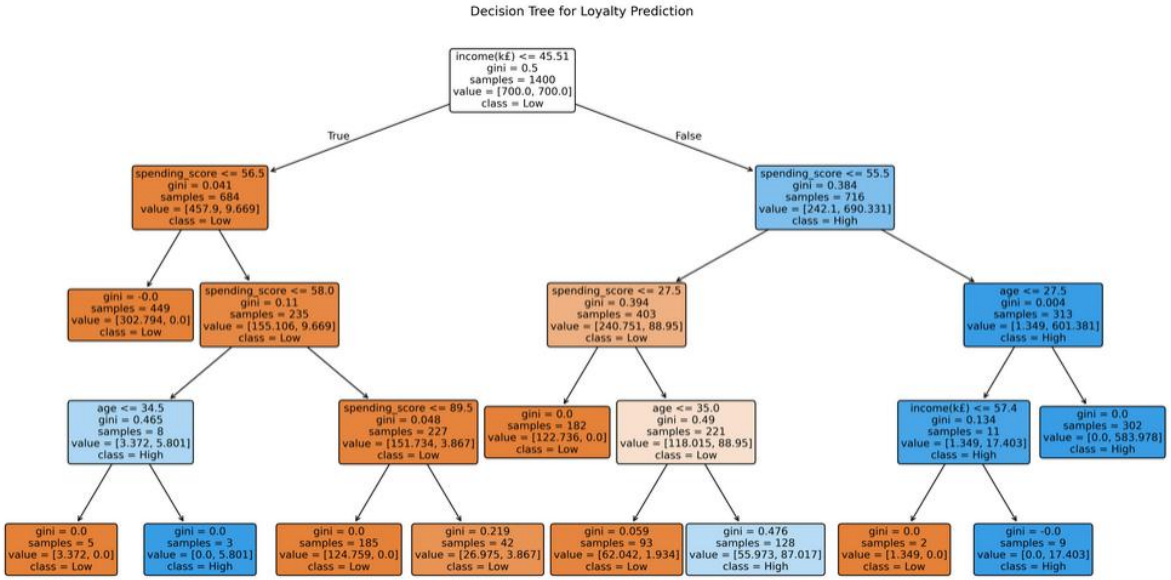


Part 4: Decision Tree

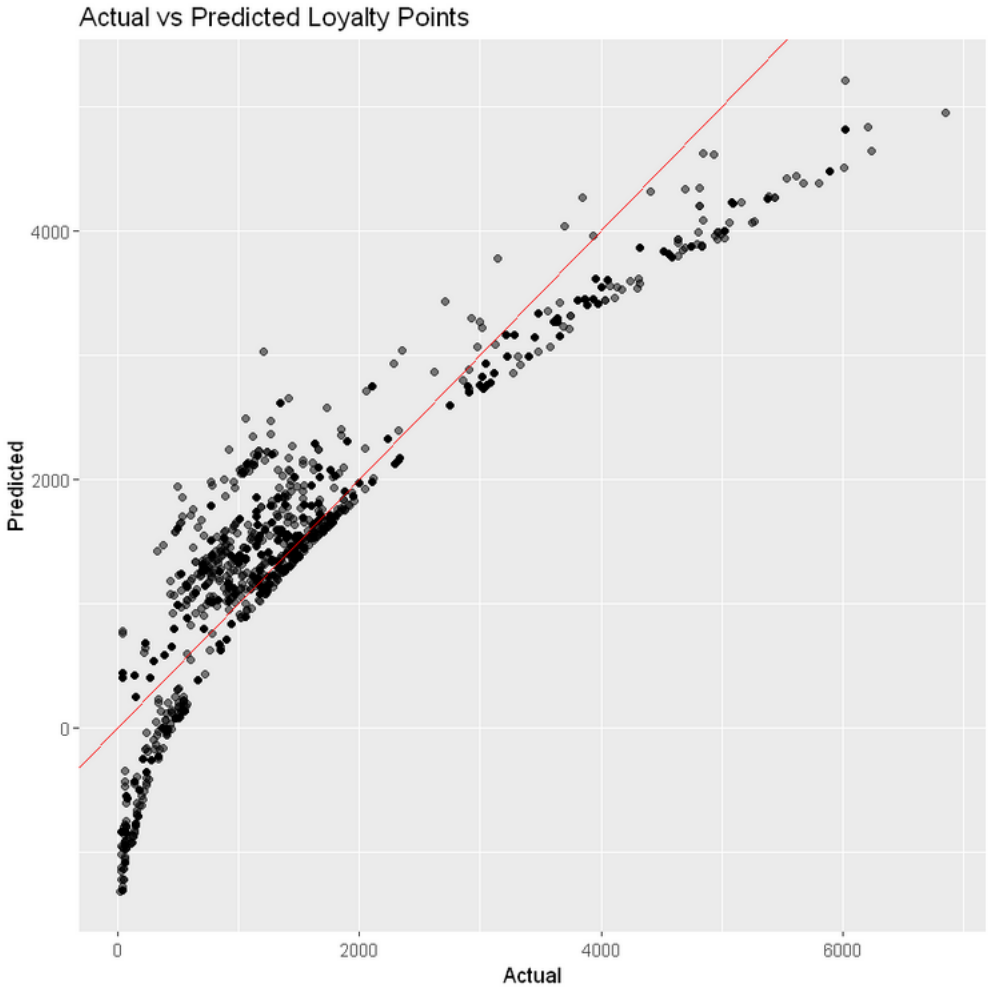
The goal of the decision tree is to identify customers with high loyalty point VS low loyalty points (Using top 25% benchmark: ≥1751 loyalty points for 'high' to identify

truly valuable customers). Model is 94.3% accurate, strong at identifying Low loyalty, slightly underpredicts High loyalty. Tree prioritizes income, then spending score.





Part 5: Loyalty Point Prediction

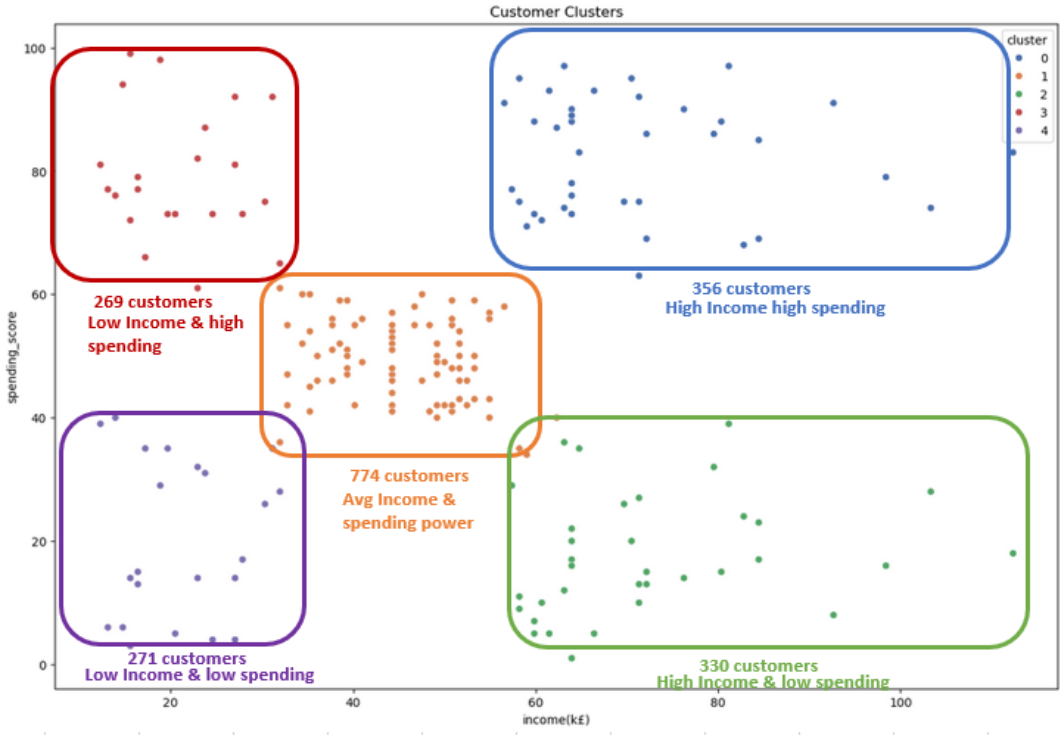


The spread is tighter at lower values (e.g., 0 to 2000 loyalty points) and widens as the actual loyalty points increase (e.g., beyond 4000). This indicates that the model performs better for customers with lower loyalty points but struggles with higher values.

The points are generally clustered around the red linesuggests the model is capturing the overall scale of the data, but the spread indicates room for improvement in precision.

Part 6: Clustering

The goal of this analysis is to segment customers into distinct groups using clustering based on their income and spending score. The optimal number of clusters was determined to be 5 using the Elbow and Silhouette methods.



Cluster 0 (356 people – 17.8%) includes **high-income, high-spending customers**—ideal for premium product promotions and loyalty programs.

Cluster 1 (774 people – 38%) represents **average income and average spending**—suitable for mainstream campaigns and bundled offers.

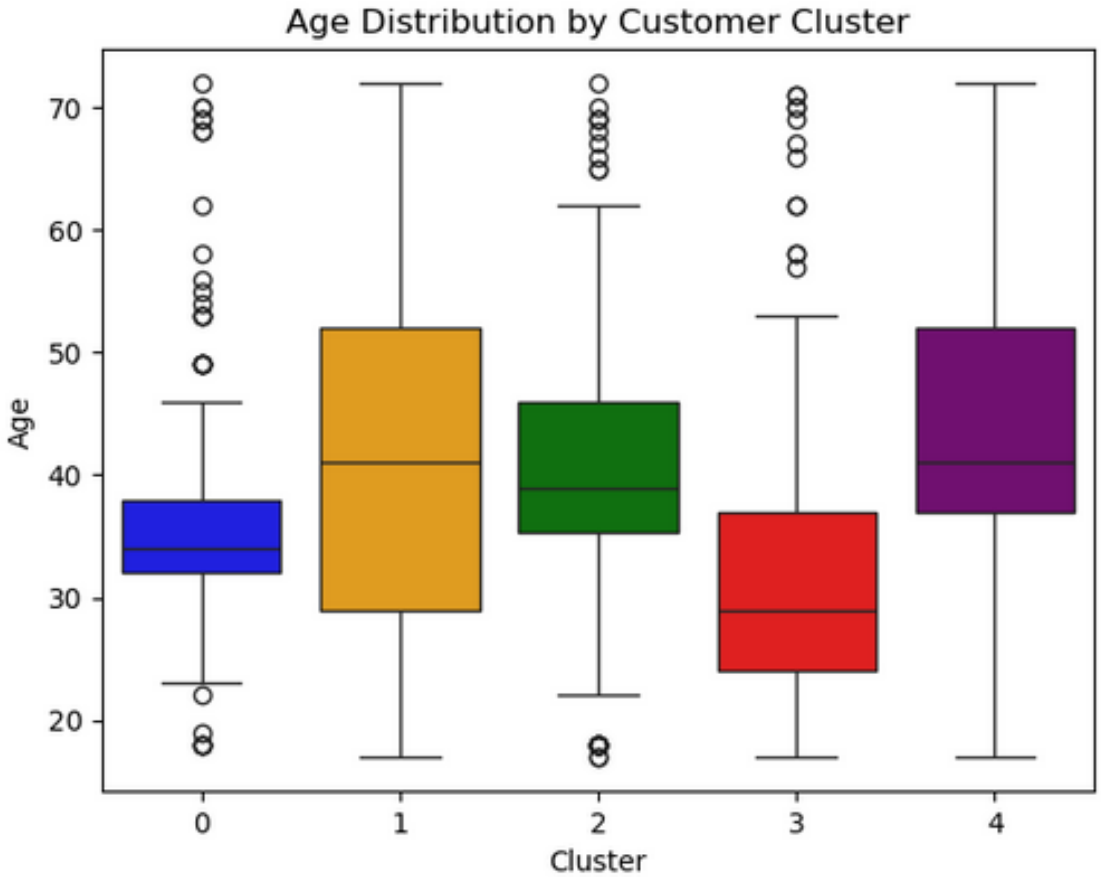
Cluster 2 (330 people – 16.5%) with **high income but low spending**, presents an opportunity for targeted upselling and value-driven incentives.

Cluster 3 (269 people -13.4%) includes **low-income, high-spending** individuals—promotions and discounts can drive loyalty here.

Cluster 4 (271 people – 13.5%) consists of **low-income, minimal spenders**, best approached with essential value offerings and budget-friendly deals.

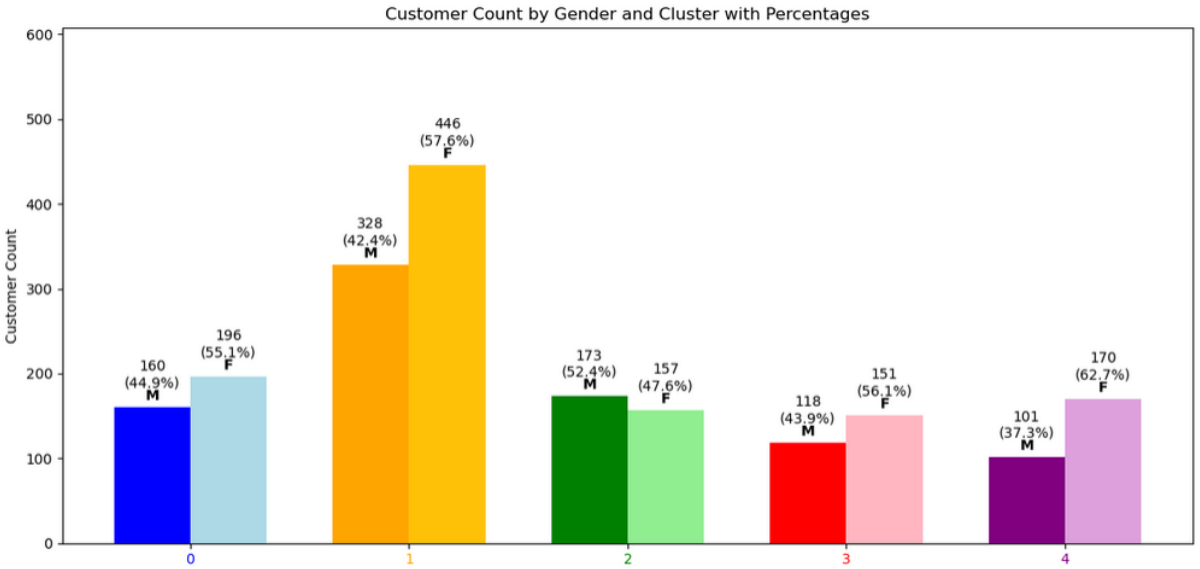
Detailed below is additional analysis targeting specific demographics:

5.1 Age Cluster 1 (average income and average spending) and 2 (high income but low spending) has the widest age range 25–48. Cluster 4 (low-income, minimal spenders) is the oldest, ranging 40–55, median near 50. Cluster 3 has the youngest age.



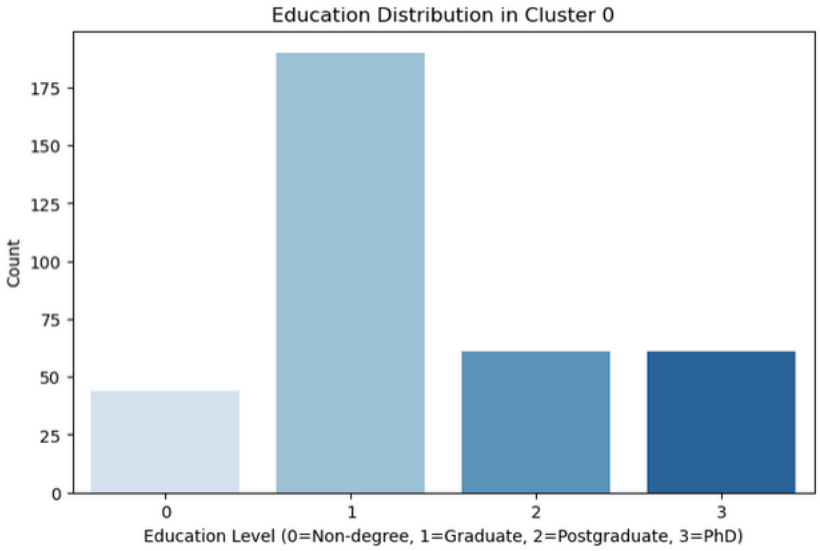
5.2 Gender

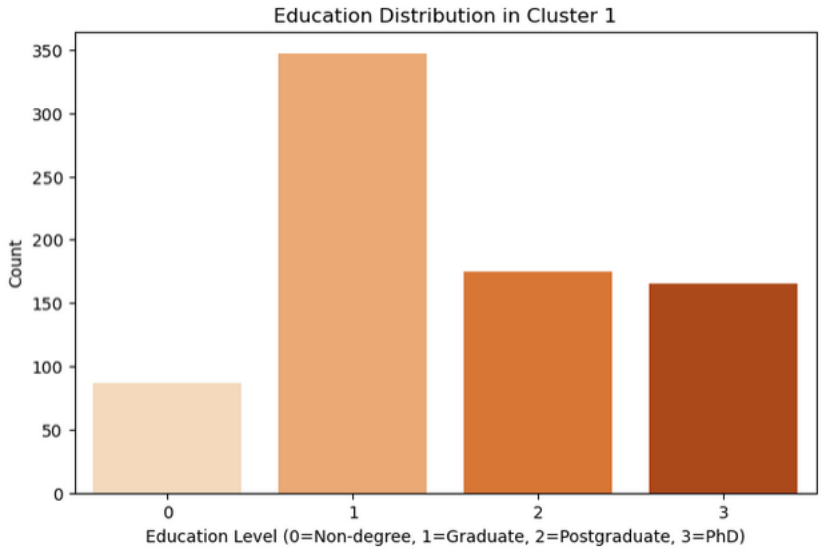
On average, 56% of customers are women, with a consistent trend across clusters. Although Cluster 2 (high income, low spending) shows a slight gender shift (0.3%), the difference is minimal.

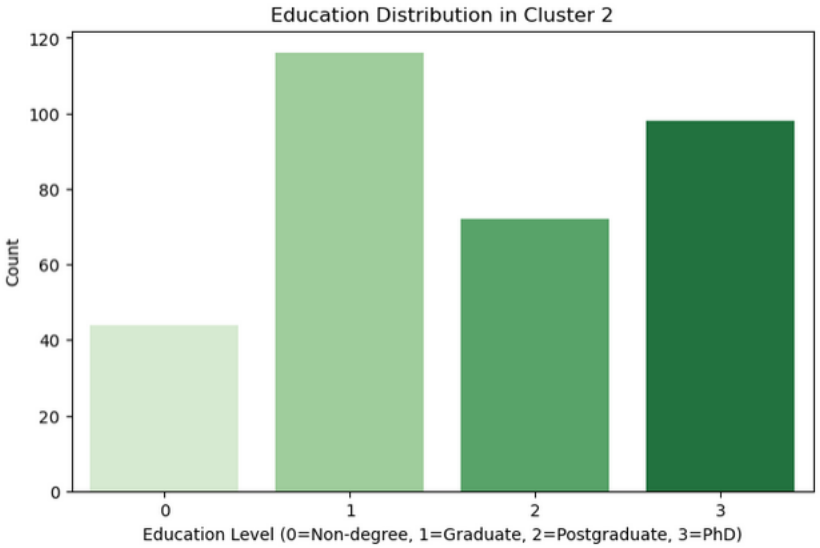


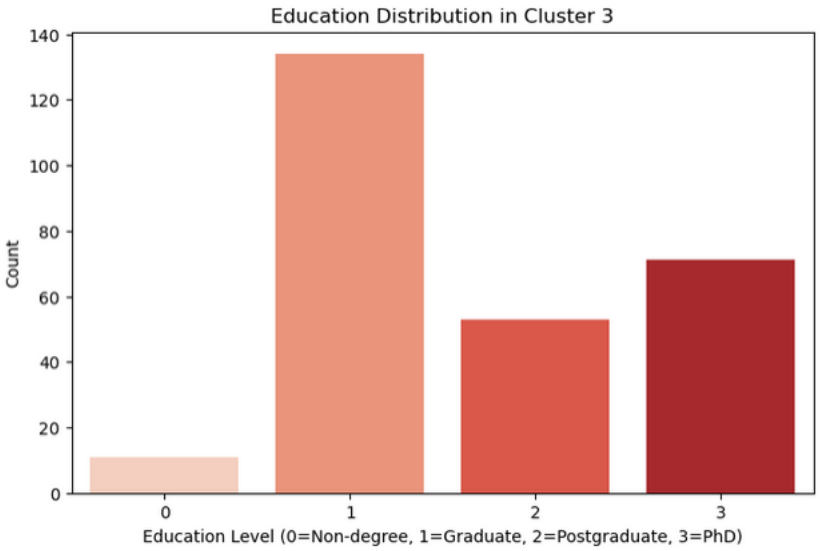
5.3 Education

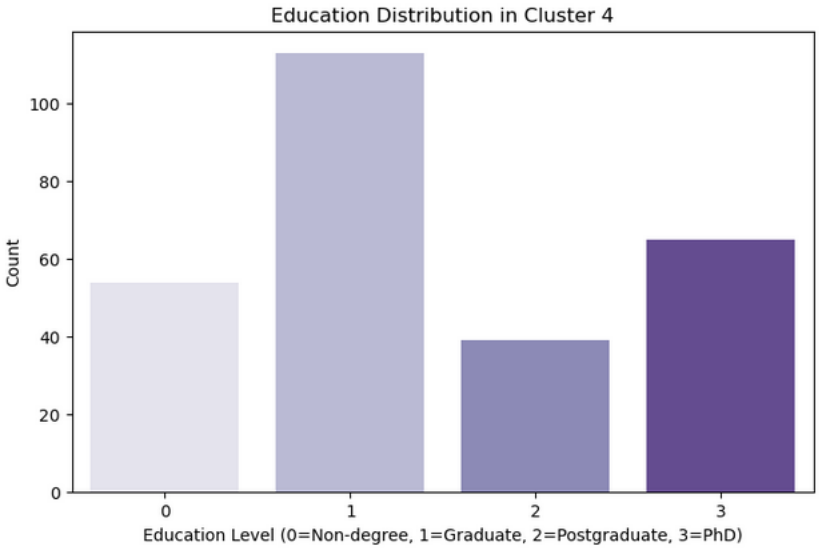
The majority of our customers hold a degree. Cluster 3 (low income, high spending) has the highest share of non-degree holders, while Cluster 2 (high income, low spending) includes a significantly higher percentage of postgraduate and PhD holders.







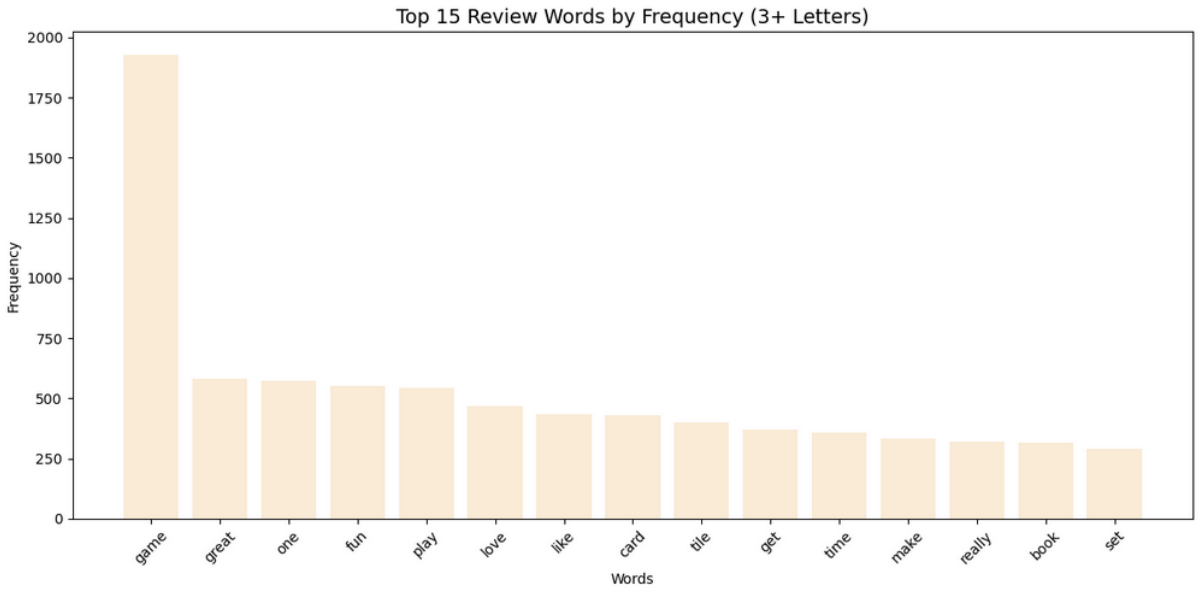


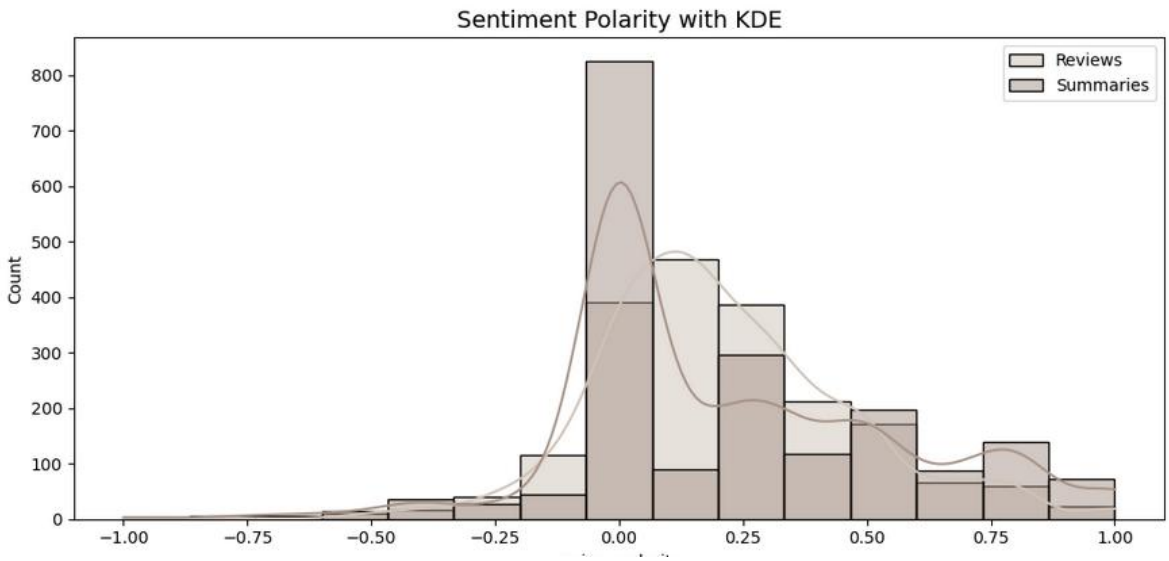


Part 6: Natural Language Processing (NLP) and Sentiment Analysis

The NLP analysis used tokenization with NLTK's word tokenize, frequency distribution with FreqDist, and stopword removal (including custom terms). Word clouds with sentiment-based colouring were created, and nouns and adjectives were filtered via POS tagging. Sentiment polarity was assessed using TextBlob, and bigrams were created with NLTK. Th length of words have also been limited to 3 or above. Visualizations, including word clouds and bar charts, were made using matplotlib. These tools analysed and visualized text data, focusing on frequency, sentiment, and key terms.

Both reviews and summaries are mostly neutral to slightly positive. Summaries are more neutral, likely due to concise language.





D. Data quality Improvement:

i. Product Information: To improve data quality, it's essential to enhance product information by including product names and subcategories alongside product codes, enabling more meaningful analysis.

ii. Spending Score and Loyalty Points: Clarifying how spending scores and loyalty points are calculated through detailed metadata would increase transparency and model reliability.

iii. Marital Status: Incorporating customer marital status can provide deeper demographic insights, allowing for more targeted segmentation and marketing strategies.

These improvements will strengthen the dataset's usability and support more informed decision-making.

E. Marketing Strategy Recommendation

Cluster 0 - High-Income, High-Spending (17.8%)

- **VIP programs (early access, exclusive in-game content, concierge support).**
- **Beta tester invites for new releases to deepen brand loyalty.**

Cluster 1 – Average Income, Average Spending(38%)

- **Our largest customer segment, and it's important to retain it**
- **Introducing a referral program can help expand this group by encouraging existing customers to bring in new ones.**

Cluster 2 - High-Income, Low-Spending (16.5%)

- **High proportion of PhD and postgraduate degree holders.**
- **Organize webinars with game designers discussing "The Science of Immersive Narratives" or "How Games Challenge Critical Thinking."**

Cluster 3 - Low-Income, High-Spending (13.4%)

- **Rewarded Video Ads:** Specifically involves watching short videos to earn
- **Flexible Payment Plans**: Offer instalment options for big purchases
- **Student Discount**: Target cluster 3 as the youngest age group

Cluster 4 - Low-Income, Minimal Spending (13.5%)

- **Launch a free-to-play mobile game (or promote an existing one) with a "Play Free, Win Big" campaign that emphasizes no-cost fun, rewarding social engagement, and micro-incentives**