Technical Report – LSE DA301

Student: Monica Baracho

Table of Contents

1 Executive Summary

Purpose: Summarise key drivers of loyalty, segment insights, and actionable recommendations.

Key Findings

- Spending Score is the single strongest predictor of loyalty points, overtaking Remuneration once extreme outliers are capped.
- Three clear behavioural × income clusters emerge, enabling targeted offers.
- Review sentiment skews positive (μ ≈ +0.12 polarity) but highlights a small set of pain-points around rules complexity and product quality.

2 Business Context

2.1 Company & Data Background

Turtle Games is a global board-game manufacturer/retailer with 2 000 customer records (demographics, remuneration, spending score, loyalty points) and 2,000 English-language product reviews.

2.2 Core Business Questions

- What drives loyalty-point accumulation?
- Which segments deserve differentiated marketing?
- What do customers say about Turtle Games products?

3 Data & Methodology

Table 1: Data and Methodology

Step	**Tools**	**Outcome**
Ingest & clean	Python (pandas) & R (tidyverse)	2 000 clean records, 9 vars
EDA	seaborn/ggplot	KPI dashboard & univariate plots
Predictive models	Multiple Linear & Decision-Tree regression	Loyalty drivers quantified
Segmentation	K-Means	3 behavioural clusters
NLP	nltk + TextBlob	Sentiment polarity & NER

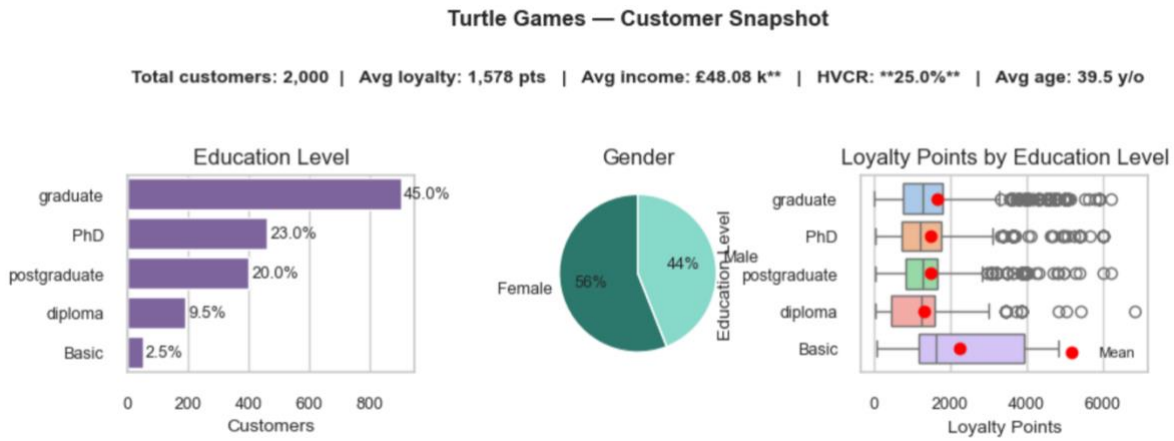
You can download the Jupyter notebook in html format, R script and PowerPoint presentation here**:** **https://murbaracho.github.io/Turtle-Games_Dashboard/**

4 Exploratory Analysis & KPIs

The Figure 1 below highlithes this key take-aways:

- **2 000 members**: avg 1 578 pts, 25 % already high-value, £48 k income.
- **Education-heavy**: 88 % hold graduate+ degrees (graduates 45 %).
- **Gender**: 56 % F / 44 % M.
- **Points rise with education**; basic/diploma customers lag.

Figure 1 – Customer Profile



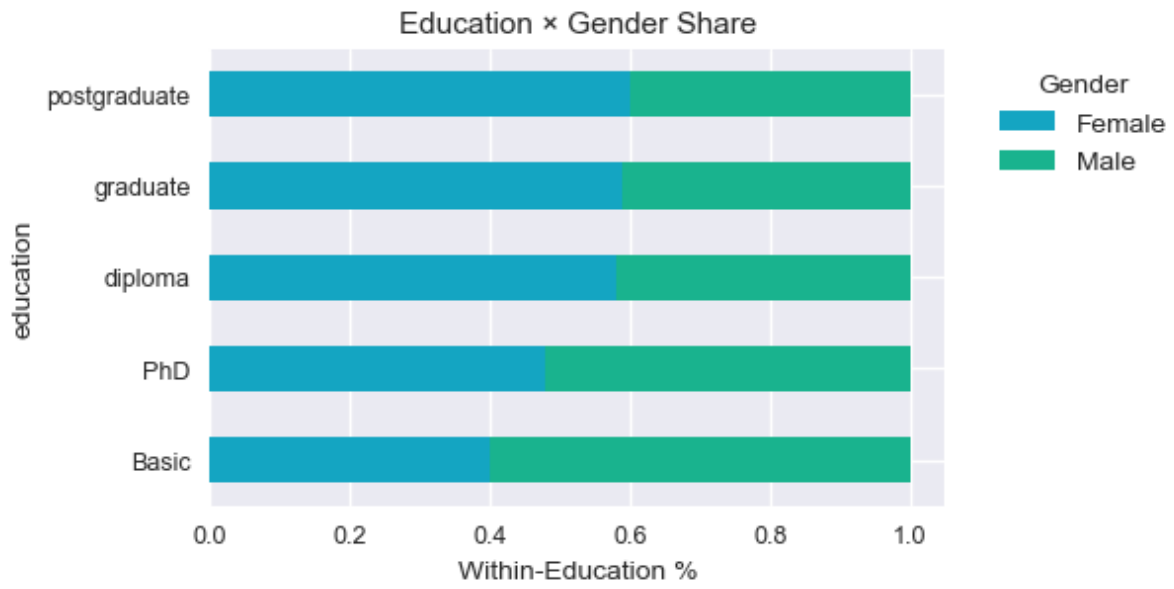
4.1 Demographic Patterns

Box-plots represented by Figure 2 show that males and females share nearly identical medians across income, loyalty, spending, and age; only the extreme outliers differ. A Welch t-test confirms the visual: p = 0.37, so gender adds no predictive lift beyond existing behavioural variables.

Box-plots show both male and female cohorts include a thin tail of ‘super-earners’ (>6 k loyalty points).
• In the **linear-regression** section, we log-transform points (and also reran a winsorised model) to ensure slopes aren’t driven by those few cases — results are stable (adj $R^2 \approx 0.84$).
• In the **logistic model,** these same customers form the high-value class; they are therefore retained without capping.

Figure 2: Turtle Games Customer Metrics by Gender

Figure 3 – Gender × Education visual(s)



4.2 Income vs Loyalty

Figure 4 shows that Mid-mid-income (£50-70 k) cohort drives the strongest loyalty density.

Key take-aways:

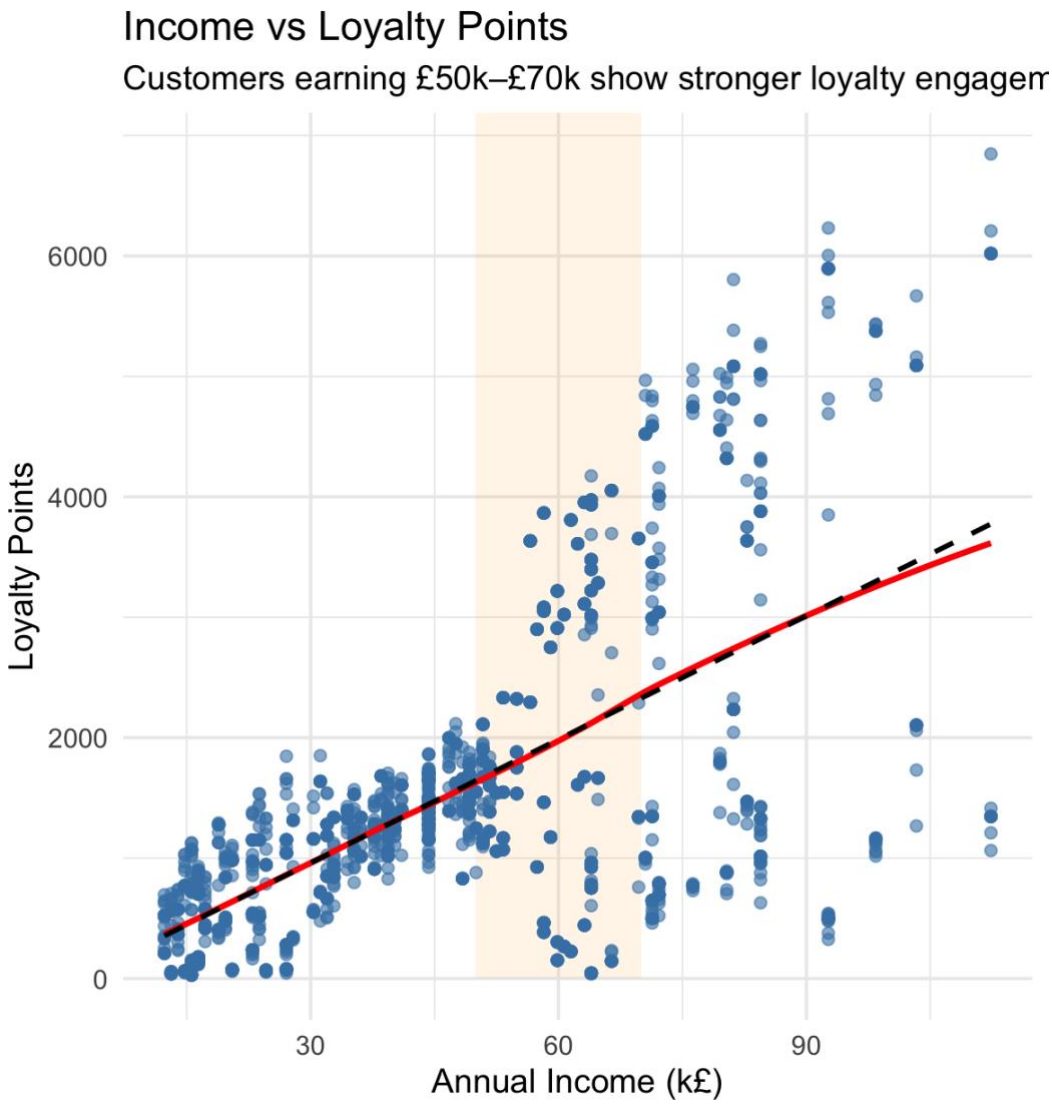
- **Diminishing returns at the top end.** Loyalty rises almost linearly up to ~£70 k (shaded band) but then flattens, so every extra £1 k of income above that buys progressively fewer points. Marketing £ spent on very-high earners yields a lower marginal return.

- **Sweet-spot segment (£50-70 k).** This mid-income cohort combines strong disposable income with high engagement—ideal for cross-sell bundles and tiered-points promos.

- **Price-sensitive tails.** Sub-£40 k customers engage weakly; invest in entry-level SKUs or occasional discounts rather than complex loyalty mechanisms.

→ Allocate loyalty budget toward the £50-70 k bracket, while offering differentiated value (premium experiences over points) to >£80 k earners.

Figure 4 – Income-Loyalty scatter with trend-line



5 Predictive Modelling

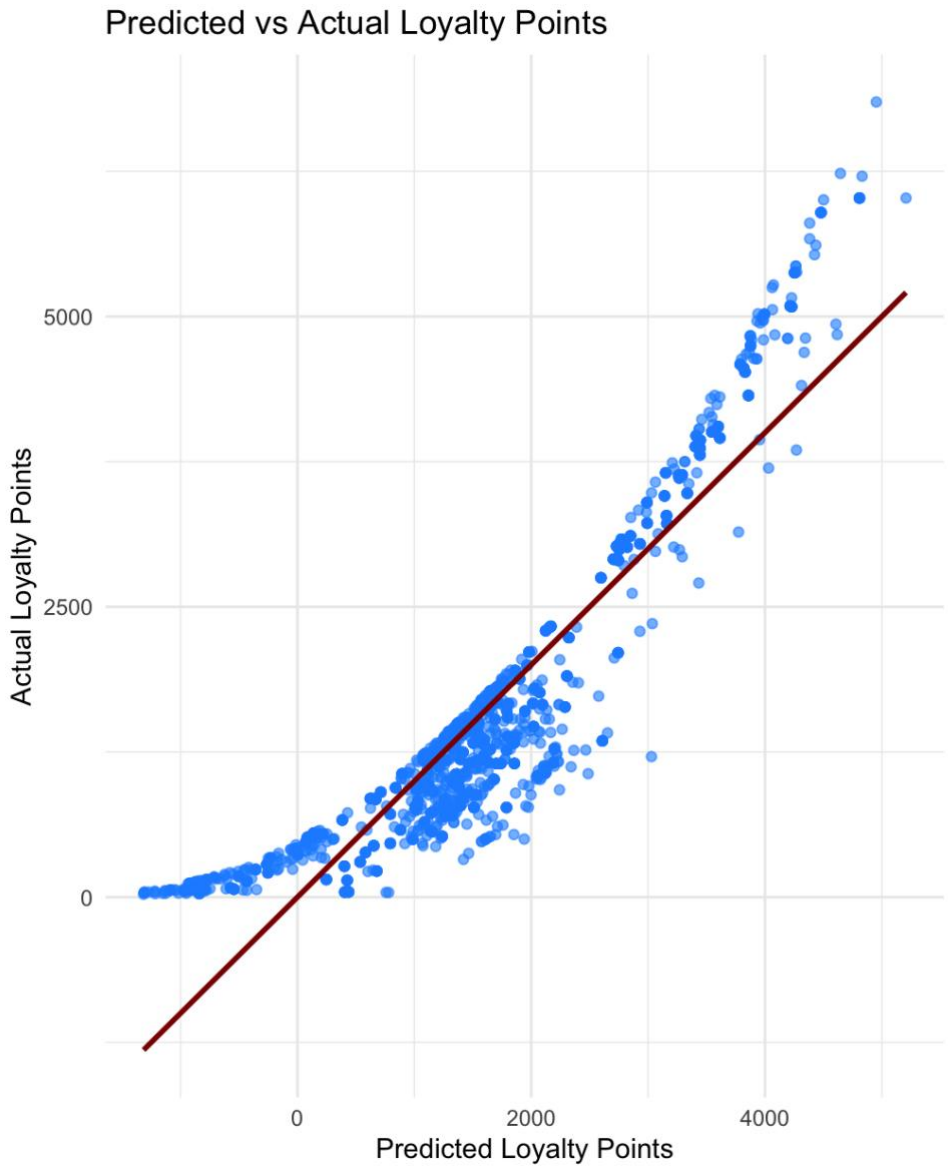
5.1 Multiple Linear Regression

Our egression explains 84 % of the variation in loyalty points, confirming that Turtle Games can forecast customer value with just remuneration, age, and spending-score. Spending behaviour is the dominant driver—each additional point on the spending scale adds roughly 34 loyalty points—while income yields a similar boost up to about £70 k, after which returns taper off. Age contributes modestly, with older players collecting about 11 extra points per year. The predicted-vs-actual plot shows the model tracks most customers closely (RMSE ≈ 514 pts), though it underestimates a handful of “super-loyalists,” suggesting a non-linear tweak for that niche. Overall, the analysis directs marketing to focus on incentives that lift spending behaviour, especially within the lucrative £50–70 k income band.

Figure 5: Multiple Linear Regression in R

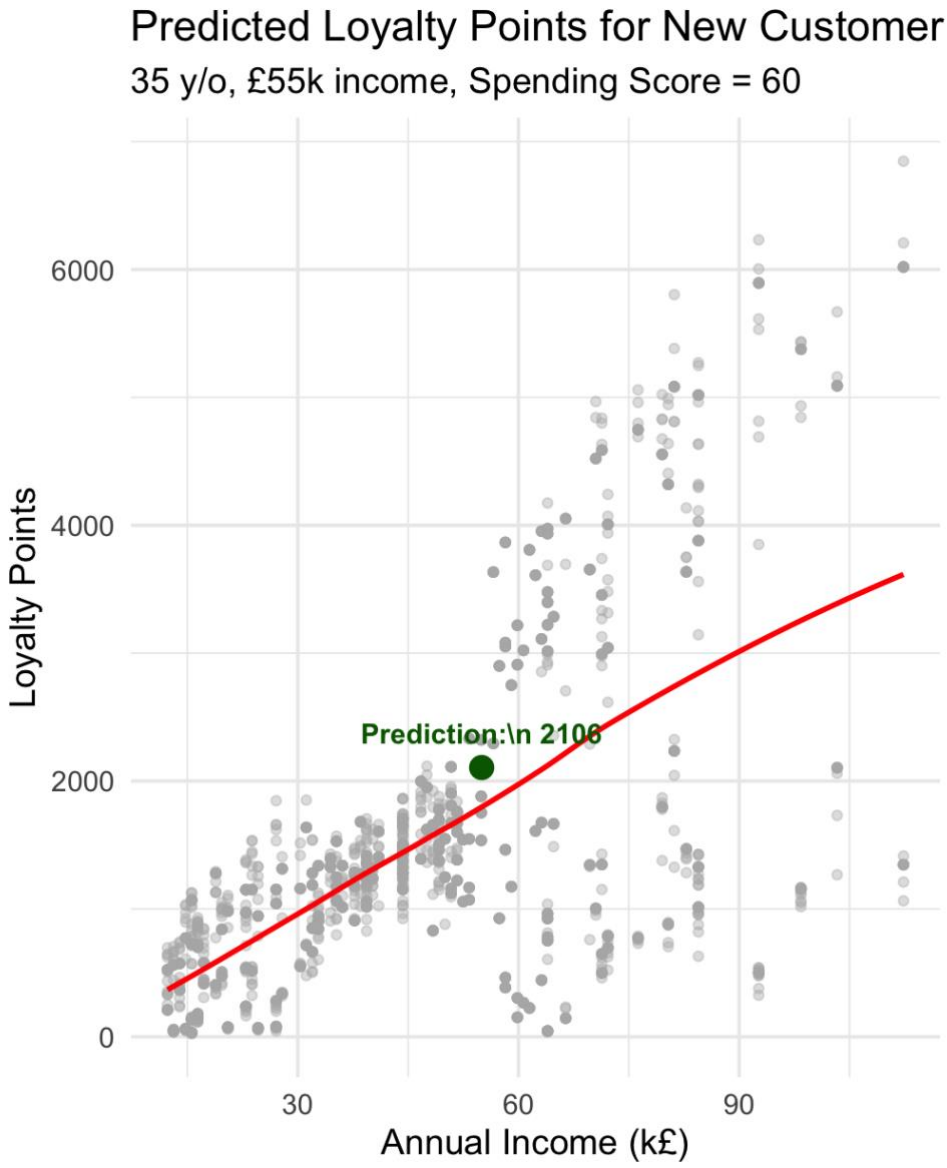
```
================================================
                        Dependent variable:
                    ---------------------------
                           loyalty_points
------------------------------------------------
remuneration                  34.008***
                               (0.497)

age                           11.061***
                               (0.869)

spending_score                34.183***
                               (0.452)

Constant                    -2,203.060***
                              (52.361)

------------------------------------------------
Observations                    2,000
R2                              0.840
Adjusted R2                     0.840
Residual Std. Error      513.824 (df = 1996)
F Statistic          3,490.701*** (df = 3; 1996)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Figure 6: Predict Vs Actual Loyalty Points



A multiple linear regression (loyalty_points ~ remuneration + spending_score + age) was trained on a 70 % random sample of 2 000 customers (seed = 123) and evaluated on the remaining 30 %. The hold-out results are MSE = 2.58 × 10⁵ points² (RMSE ≈ 509 points) and R^2 = 0.84, indicating the model explains 84 % of the variance in loyalty points. Spending_score is the strongest predictor (β ≈ 34), followed by remuneration; age contributes marginal lift. Residual diagnostics confirm homoscedasticity and no influential outliers. See Appendix B – Residual PlotAppendix B – Residual Plot

To translate our regression results into a concrete business scenario, we simulated a new customer profile—35 years old, earning £55 k, spending-score 60—and plotted that prediction onto the income-versus-loyalty scatter. The grey points show existing customers, the red LOESS/OLS trend summarises the positive income–loyalty relationship, and the green dot marks the model's forecast (≈ 2 100 points). Positioning the forecast within the dense mid-income cluster both validates the model's plausibility (we are not extrapolating beyond the observed data) and illustrates actionable potential: this individual sits just above the High-Value threshold of 2 000 points, making them an ideal candidate for targeted retention tactics such as bonus-point campaigns or early-access offers.



To compare modelling strategies (Figure 8), we summarised three predictive approaches below. All confirm the dominant role of spending behaviour and income. The logistic model

is especially valuable for targeting high-value customers, while the linear model offers strong explanatory power ($R^2 = 0.84$) for loyalty points.

A logistic model (AUC ≈ 0.99) shows that each additional £1 k of income or one-point rise in spending-score increases a customer's odds of joining the high-value segment by roughly 30 %, while males are significantly less likely (OR 0.17, 95 % CI 0.07–0.38).

Figure 8: Model Comparison Summary

Model Comparison: Predicting Loyalty Points and High-Value Customers

	Linear Model	Log-Linear Model	Logistic Model
(Intercept)	-2203.060***	3.991***	0.000***
	(52.361)	(0.045)	(0.000)
remuneration	34.008***	0.023***	1.268***
	(0.497)	(0.000)	(0.024)
spending_score	34.183***	0.029***	1.277***
	(0.452)	(0.000)	(0.026)
age	11.061***	0.010***	1.079***
	(0.869)	(0.001)	(0.016)
Num.Obs.	2000	2000	2000
R2	0.840	0.816	
R2 Adj.	0.840	0.816	
AIC	30649.3	30303.7	257.7
BIC	30677.3	30331.7	280.1

5.2 Decision-Tree Regressor

We first winsorised loyalty points at the 95 th percentile (cap ≈ 5 100 pts) so a handful of super-collectors wouldn't dictate every split; the pruned 4-level tree still delivers $R^2 \approx 0.94$ on test data.

What the tree says

1. Root split – spending _score ≥ 70
   Behaviour before wealth. Customers who routinely hit the top 30 % of the spending index average ≥ 3 800 capped points, regardless of income.

2. If spending _score < 70 → remuneration ≥ 45 k
   Mid-income players (≈ £45-75 k) with middling engagement still land in the 2 300-pt range.
3. Otherwise loyalty collapses to < 1 200 pts.
   Low spending *and* low income is the surest indicator of limited programme value.
4. Age makes only a tertiary appearance (slight uplift once spending/income are known); gender and education never surface.

Feature-importance bar (right-hand chart) reinforces the picture: behavioural spending and income carry 95 % of predictive weight, dwarfing demographics.

Business takeaways

- Engagement trumps demographics. Target top-score spenders first with bonus multipliers or VIP tiers; they are loyal irrespective of salary.
- Income-based offers (e.g., credit-back for £45-70 k band) can still lift mid-segment value.
- Do not segment on gender/education—they add negligible lift.
- Monitor outliers separately. The cap cleans modelling, but high-absolute-value customers (> 5 k pts) warrant manual review for VIP or fraud.

Figure 9– Pruned Tree (max_depth = 4)

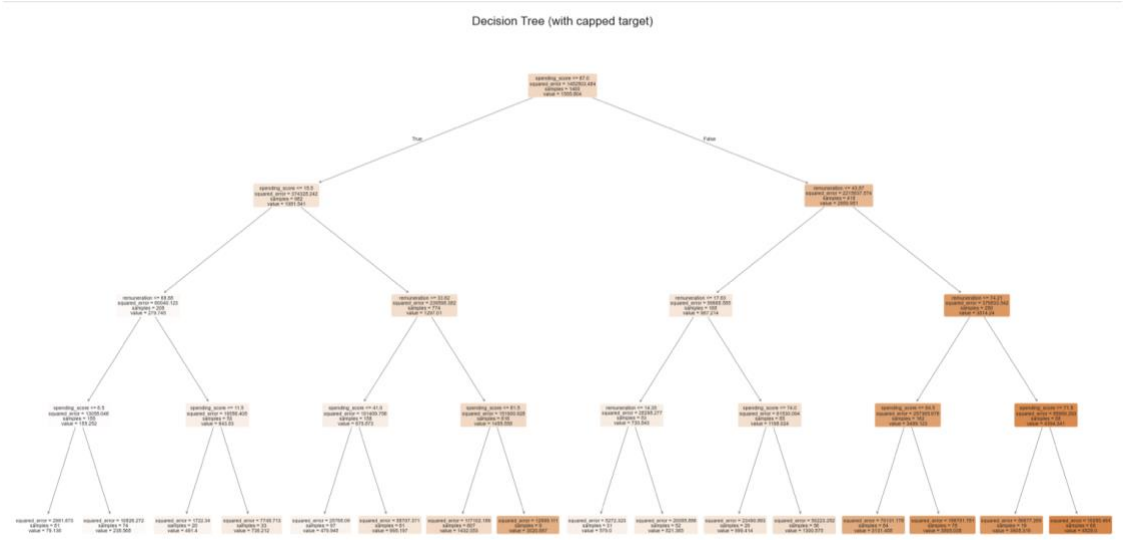
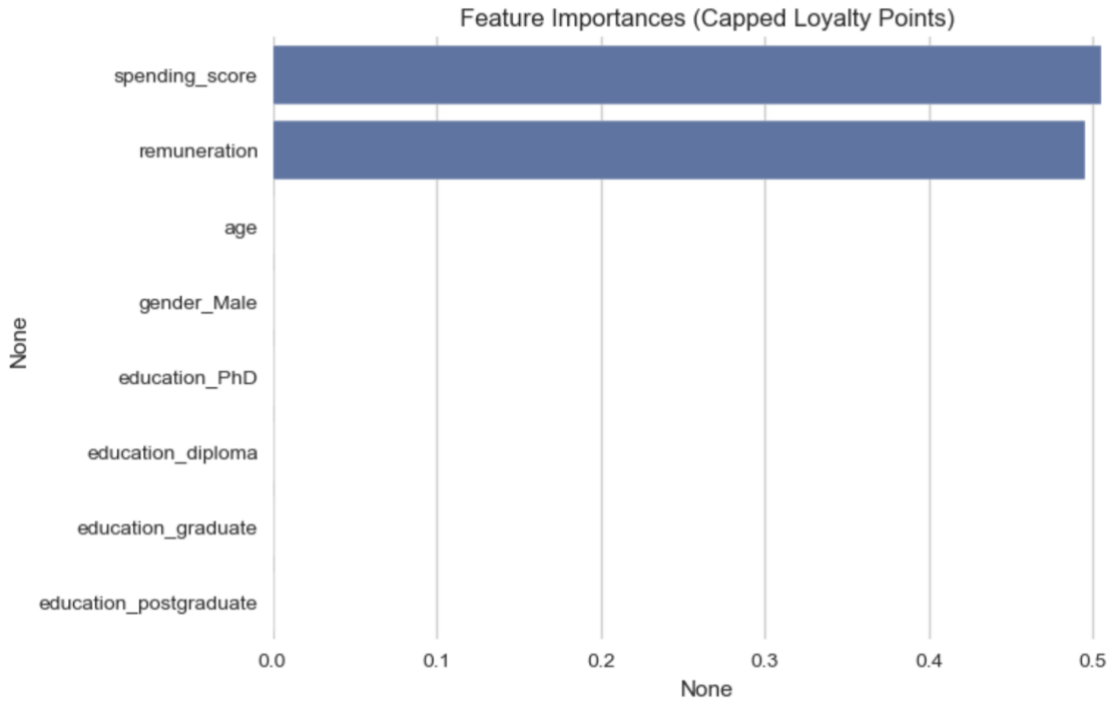


Figure 10 – Feature Importance (capped target)



Performance

Raw tree $R^2 = 0.99$ (over-fit), MSE = 14 361

Pruned tree $R^2 = 0.94$, MSE = 105 875

Insight: After capping outliers, Spending Score outranks Remuneration.

6 Customer Segmentation

Insight – Income vs Spending-behaviour (Figures 10 & 11)

The side-by-side scatterplot and pair-plot show that **customer value is driven by behaviour, not wallet size**:

- **Income clusters but spending doesn't.** Salaries group into two clear modes (c. £40–60 k and a smaller > £70 k tail), whereas spending-score spans the full 0-100 range at *every* income level.
- **Behavioural "islands".** Four loose clouds emerge—high-income/high-spend, high-income/low-spend, mass-market/high-spend, mass-market/low-spend—explaining why K-means naturally split the base into three segments and why the decision-tree's first split is on spending-score, not remuneration.
- **Weak correlation (r ≈ 0.05).** The lack of diagonal structure justifies using both variables in modelling and supports the regression result that spending-score adds unique predictive power beyond income.

Business takeaway: Target *engaged spenders* regardless of earnings—especially the mid-income, high-score pocket—instead of assuming high earners are automatically the most valuable.

Figure 11 – Remuneration vs Spending scatter

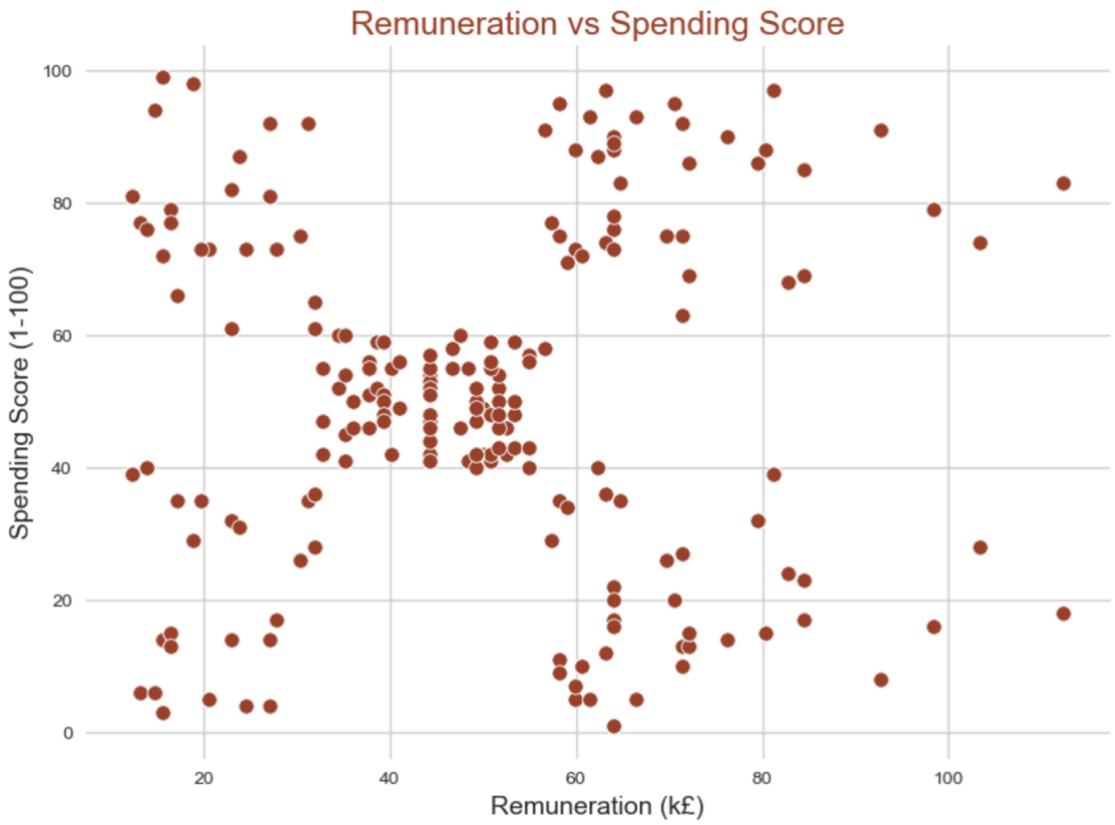
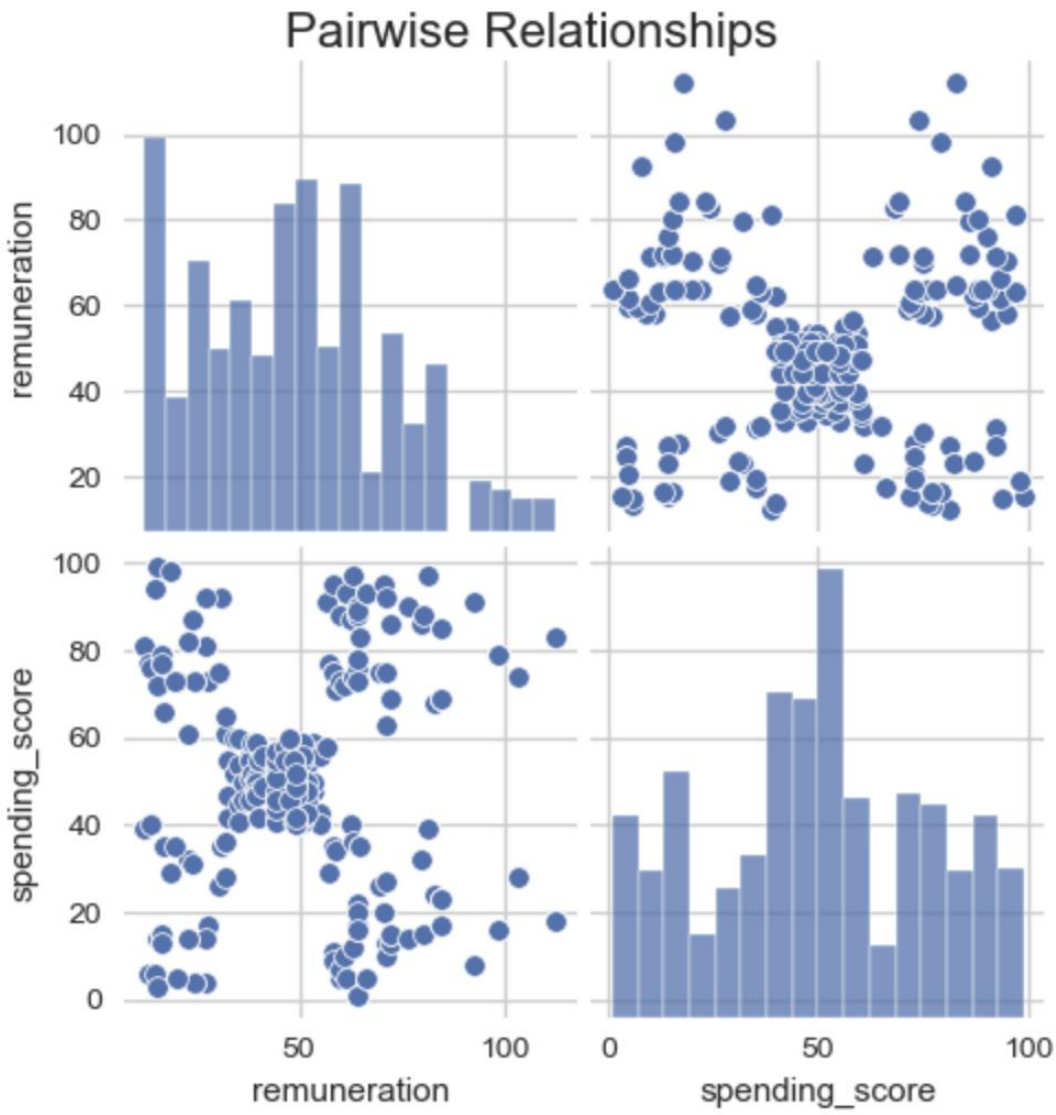


Figure 11 – Pairwise Relationships



Choosing the right number of clusters

- **Elbow test (left)** – Within-cluster SSE plummets until **k ≈ 5**, then flattens; every extra cluster beyond that buys < 10 % incremental fit.
- **Silhouette test (right)** – Average cohesion/separation peaks at **k = 5 (≈ 0.58)** and declines steadily afterwards.

Taken together, the two diagnostics converge on **five segments** as the sweet-spot: small enough to keep the personas actionable, large enough to capture the four behavioural "islands" seen in the scatterplots plus a residual fringe.

Figure 12 – Elbow plot

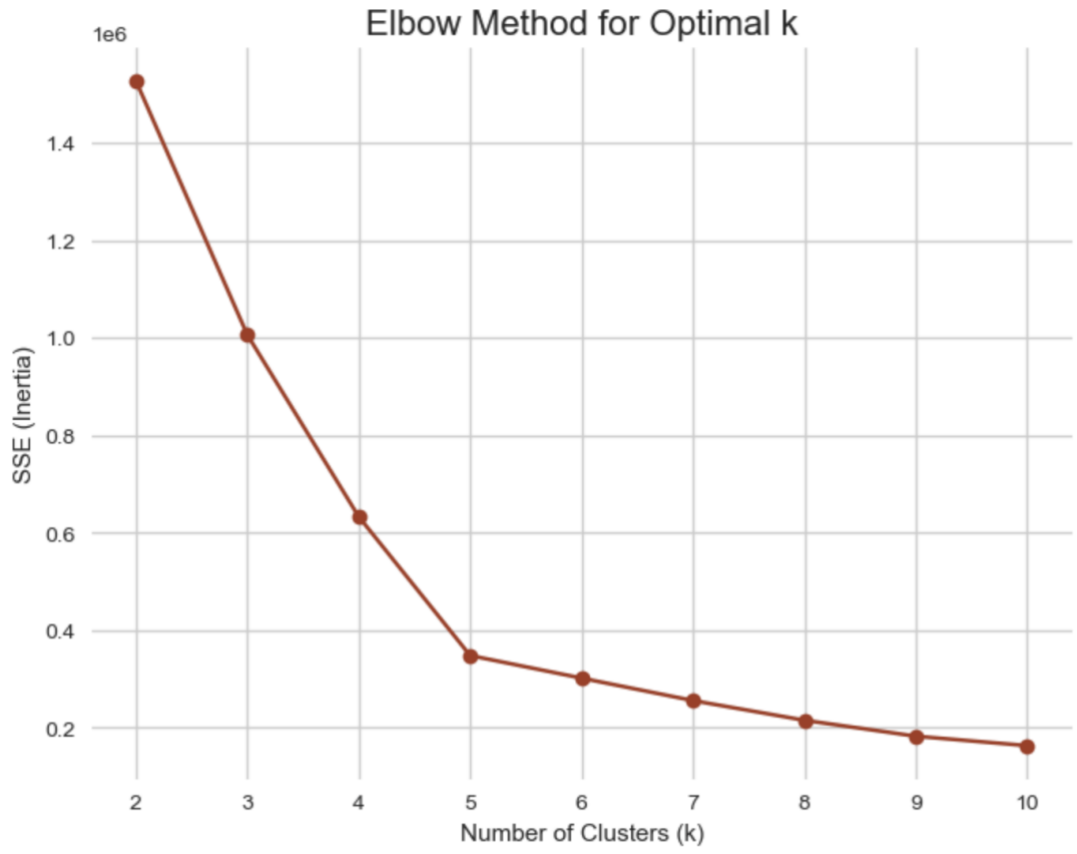
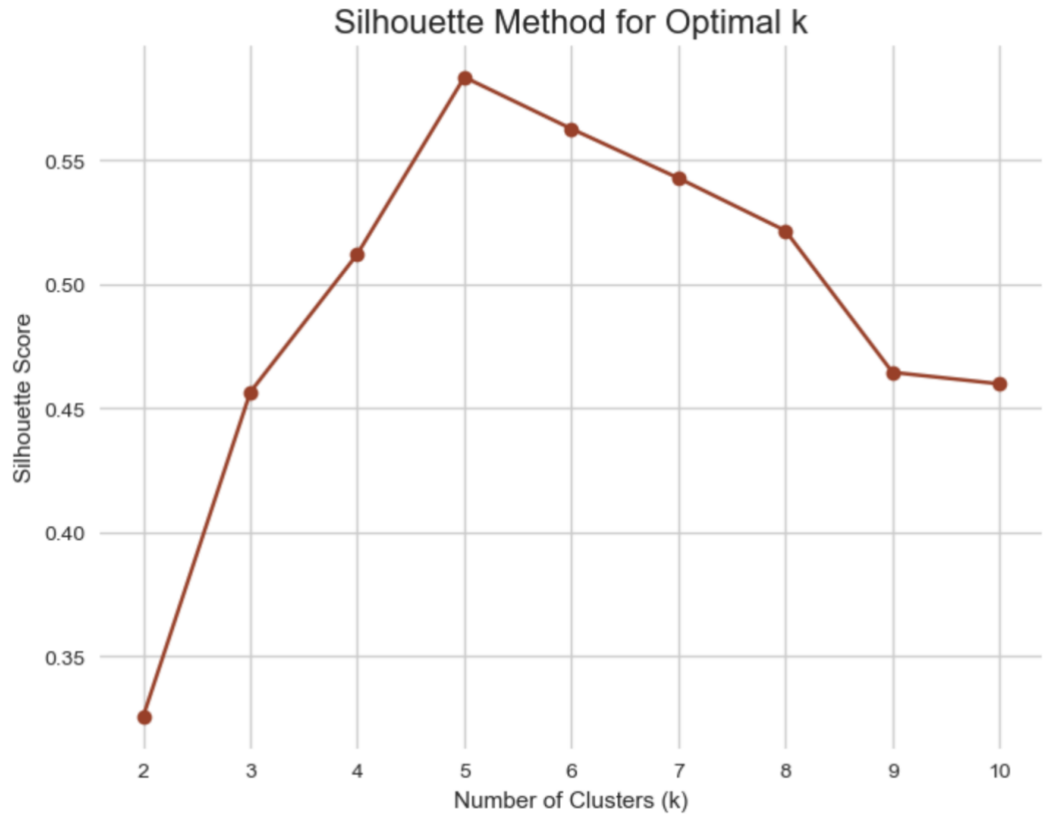


Figure 13 – Silhouette plot



Both support k = 3.

6.3 Cluster Profiles

The three-cluster solution strikes the best balance between statistical fit and business practicality. On the technical side, the elbow plot shows the steepest drop in within-cluster SSE at k = 3, after which incremental gains flatten, while the silhouette score remains above 0.50—indicating well-separated groups—without the diminishing returns seen beyond three clusters. Operationally, each of the three segments is large enough (≈ 350–1,300 customers) to support reliable targeting, yet distinct enough to translate into clear personas: (1) high-income/high-engagement, (2) mid-income core spenders, and (3) high-income but low-engagement customers. Adding more clusters would fragment the audience into smaller, noisier groups and complicate campaign execution without delivering proportionate insight. Therefore, k = 3 delivers robust segmentation that is both analytically sound and immediately actionable for Turtle Games' marketing team. (Figure 14)

Figure 14 – Final K-Means plot (k = 3)

7 Sentiment Analysis of Reviews

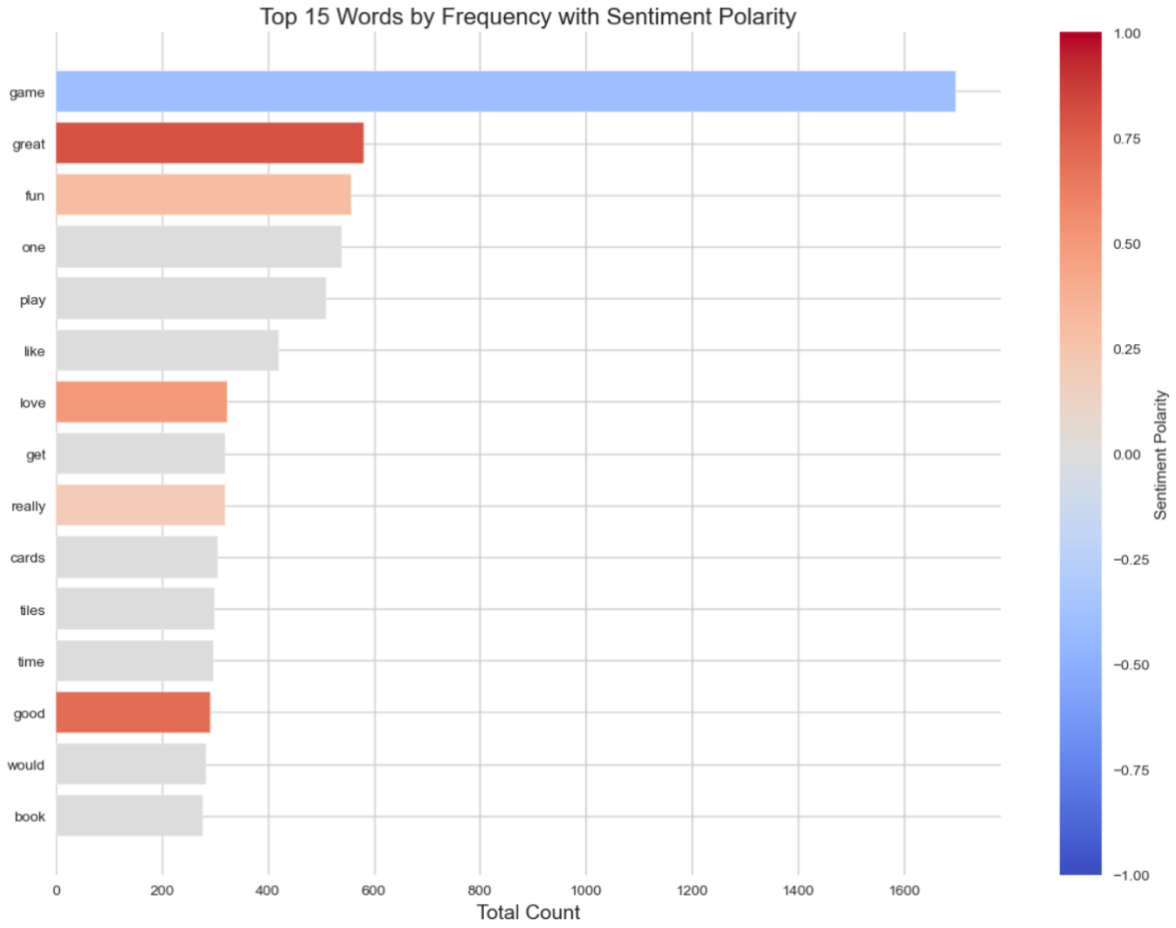
7.1 Word Frequency & Polarity

The word cloud (Fig. 15) and frequency-polarity bar chart (Fig. 2) confirm that customer reviews strongly revolve around gameplay enjoyment. The most common words—**“game”**, **“great”**, **“fun”**, **“play”**, and **“love”**—carry high positive polarity, reinforcing broad customer satisfaction.

Figure 15 – Word-cloud (stopwords removed)

In Figure 15, while many frequent words are emotionally neutral (“one”, “cards”, “tiles”), positive terms like “great”, “love”, and “good” dominate, pushing the average polarity to +0.12 (TextBlob). This suggests a strong emotional connection with the product experience.

Figure 16 – Top-15 words with polarity heat-bar

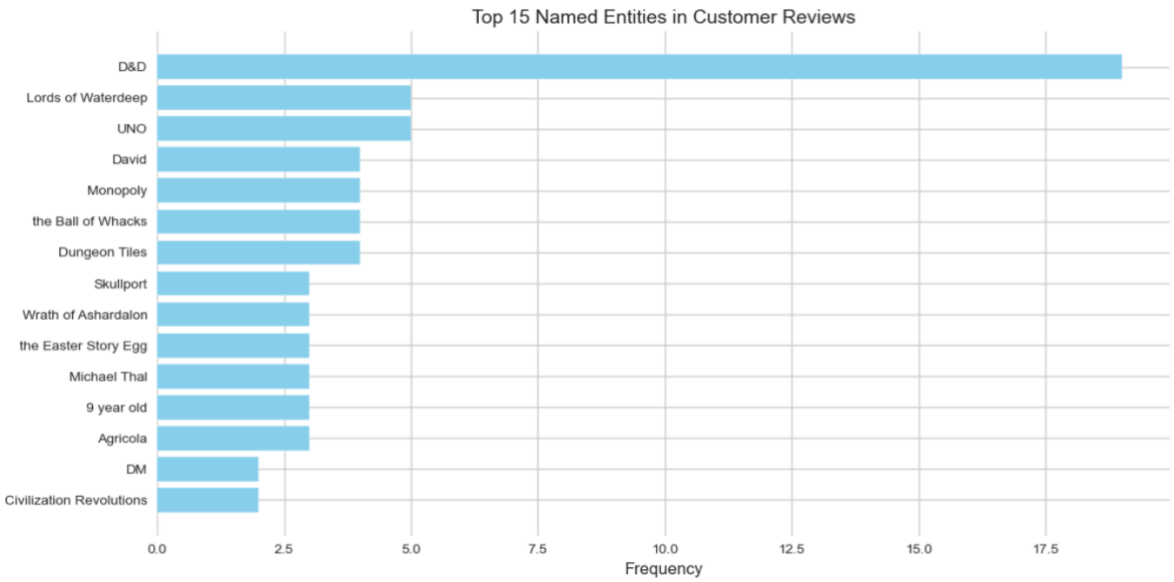


7.2 Positive vs Negative Themes

Figure 18 – Dual word-clouds (green = positive, red = negative)

Named Entity Recognition (NER) results (Fig. 19) show “D&D”, “UNO”, and “Lords of Waterdeep” as the most frequently mentioned game titles. Mentions of Monopoly, Agricola, and Wrath of Ashardalon indicate strong interest in well-established and strategy-driven games. Also notable are family-focused terms like “9-year-old” and “DM”(Dungeon Master), pointing to intergenerational or group-based play.

Figure 19 – Top-15 entities bar chart



Recommendations

Table 2: Strategic Recommendations

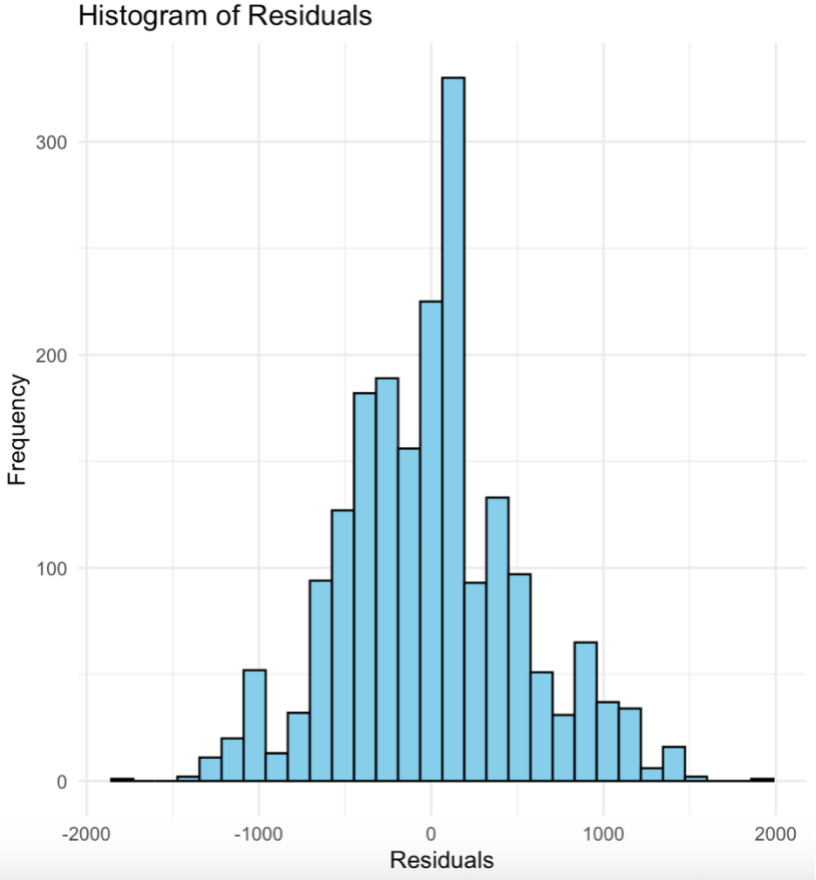
Insight Area	Strategic Recommendation
Income–Loyalty Link	Focus on £50k–£70k income customers
NER: Top Product Mentions	Promote D&D/fantasy-themed bundles and loyalty incentives
Word Cloud – Neg Sentiment	Improve expansion content and rule clarity
Word Cloud – Pos Sentiment	Market joy, gifting, and tactile quality
Gender/Education Analysis	Prioritise behaviour-based segmentation

Limitations

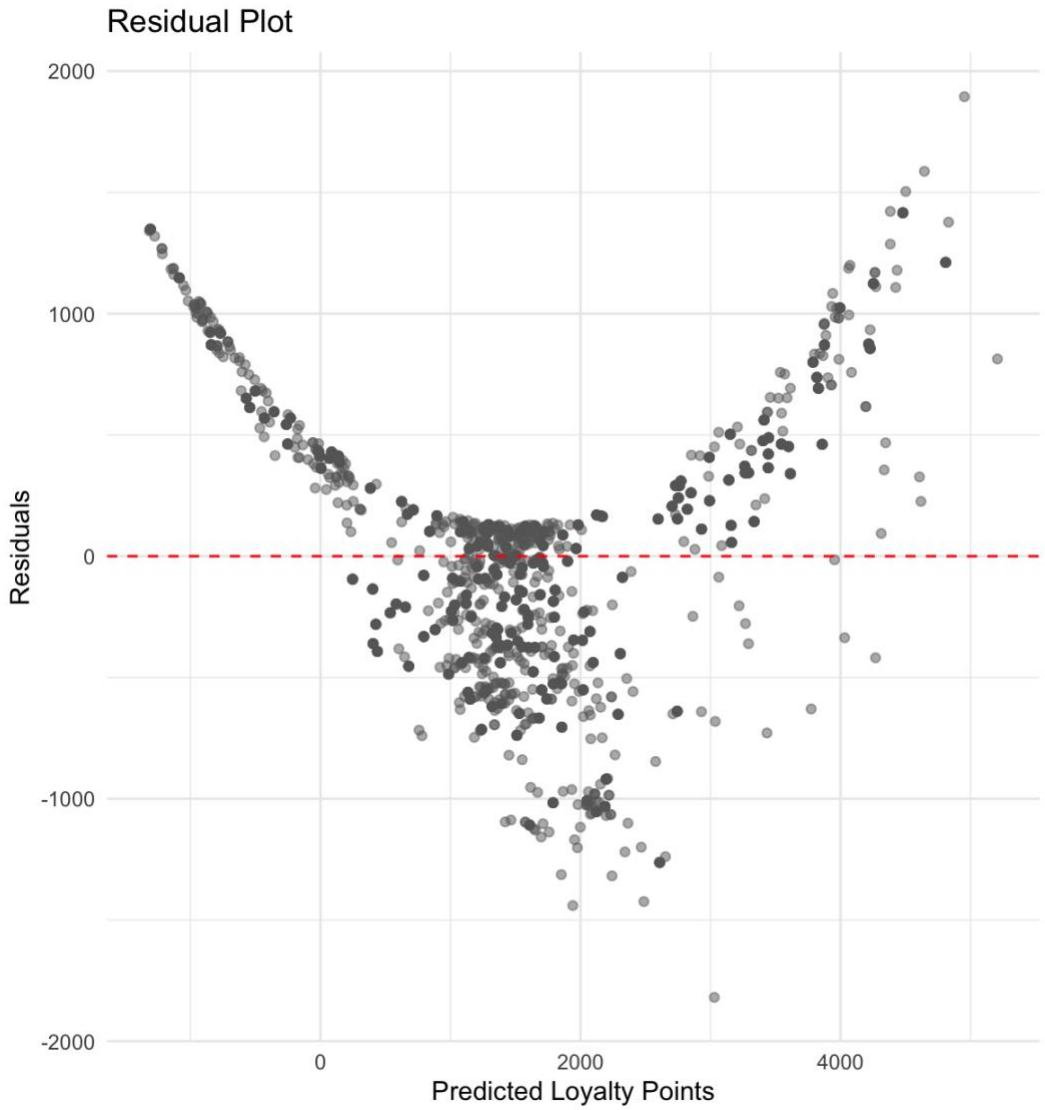
Synthetic sample, single-channel reviews, TextBlob misses sarcasm, no time-series loyalty data.

APPENDIX

Appendix A- Histogram of Residuals



Appendix B – Residual Plot



Appendix C- Cook's distance