Business Context

With a growing customer base, Turtle Games seeks to better understand how customers earn loyalty points and how to segment its market. Additionally, it aims to improve marketing using customer reviews. This report analyzes review data to identify trends, build predictive models, and make strategic recommendations aimed at improving sales performance and customer engagement.

Executive Summary

Spending Score and Income are the strongest predictors of loyalty points, with Random Forest models outperforming others. Five customer segments were identified, presenting clear opportunities for targeted marketing. Sentiment analysis shows generally positive reviews, but the review structure limits insights. Adding customer IDs, timestamps, and full purchase histories would enhance analytical capabilities. Strategic recommendations focus on improving loyalty programs and product performance.

Analytical Approach

Python was the primary tool for data cleaning, modeling, and sentiment analysis, while R was used to generate stakeholder-facing visualisations for marketing, ensuring compatibility with internal tools. Visualisations use a colour blind-friendly palette for accessibility.

- Data Cleaning & Preparation
- Linear Regression
- Decision Tree & Random Forest Modeling
- Clustering for Customer Segmentation
- Sentiment Analysis of Customer Reviews
- Strategic Business Recommendations

Code was structured for reproducibility with clear naming and sequential workflow in both Python and R scripts.

Data Cleaning and Preprocessing

- **Duplicates:** None found, ensuring unique records.
- **Categorical Encoding:** Gender and Education were numerically encoded to eliminate misinterpretation.
- **Outliers:** Present in the Loyalty Points column; retained due to the lack of timestamps, as they may indicate long-term customer behaviour.
- **Irrelevant Columns:** Language and Platform had only one unique value each and were removed.
- **Missing Data:** Customer uniqueness could not be confirmed due to lack of Customer ID.

- **Limitations:**
    - Only reviews with purchases are present—non-reviewed purchases would enhance behavioural analysis.
    - The product dataset is shallow: maximum of 13 reviews per item among 200 products, limiting predictive power for product-specific performance.

These steps ensured that our dataset was clean and optimized for analysis.

Regression Analysis

I applied simple linear regression to understand the individual effects of spending score, income, and age on loyalty points. Simple regression was used to isolate individual relationships before considering multivariate effects. Scatterplots clearly illustrate the relationship between actual and predicted values.

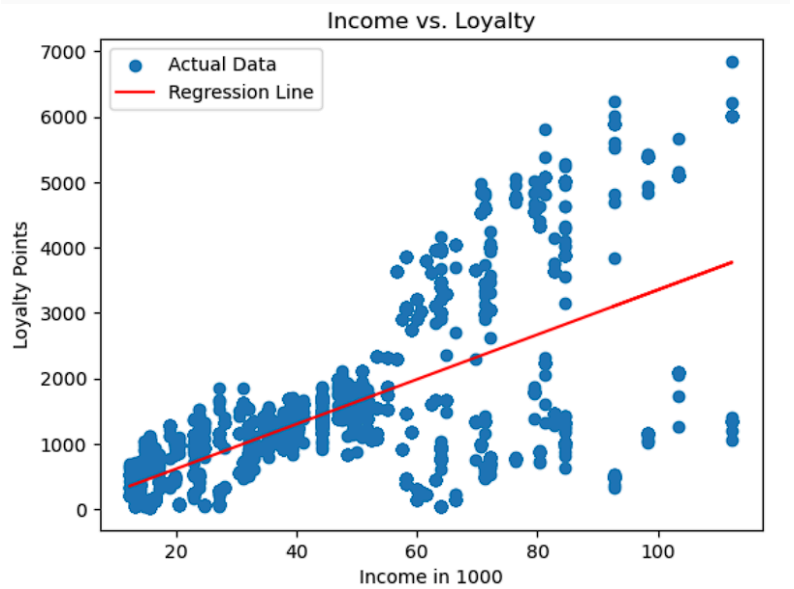
Simple Linear Regression

1. Spending Score vs. Loyalty Points

- R-squared: **0.4520** (moderate correlation)
- Coefficient: **+33.06**
- Customers who spend more tend to accumulate more loyalty points, aligning with business incentives.

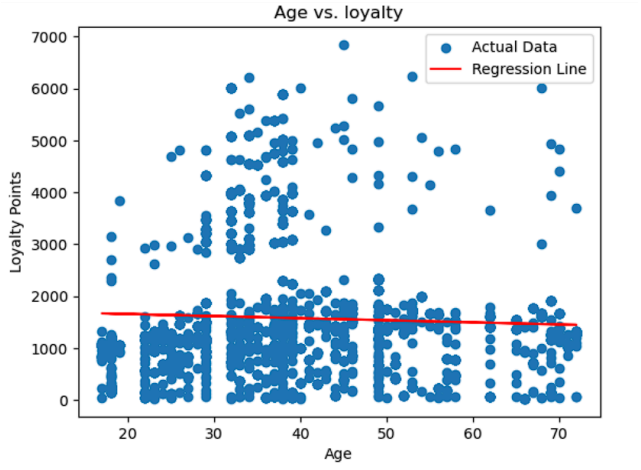
2. Income vs. Loyalty Points

- R-squared: **0.3795** (moderate correlation)
- Coefficient: **+34.19**
- While income affects loyalty point accumulation, its impact is slightly lower than spending score, suggesting spending behaviour is a stronger predictor.



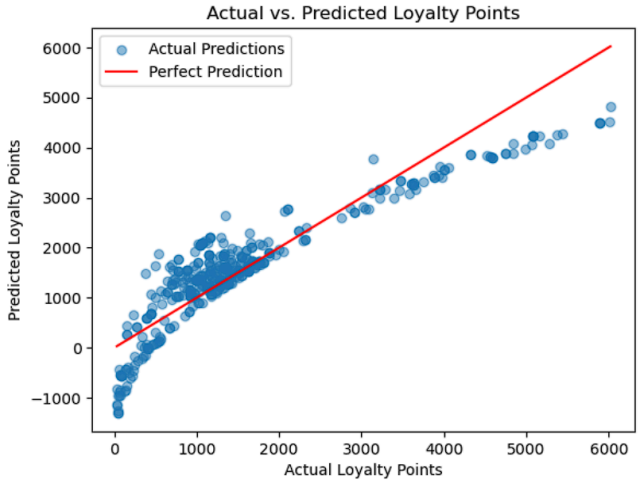
3. Age vs. Loyalty Points

- R-squared: **0.0018** (very weak correlation)
- Coefficient: **-4.01** (negligible impact)
- Age does not significantly influence loyalty points.



Multiple Linear Regression: Given that spending score and income had notable effects individually, I implemented a multiple linear regression to assess their combined influence along with age:

- Test R-squared: **0.8291** (minimal overfitting)
- Coefficients: Spending Score: **+33.97,** Income: **+34.25,** Age: **+11.01**
- Age, while insignificant alone, gained relevance in combination with other factors. Spending score remains the strongest predictor.



Assumption Checks:

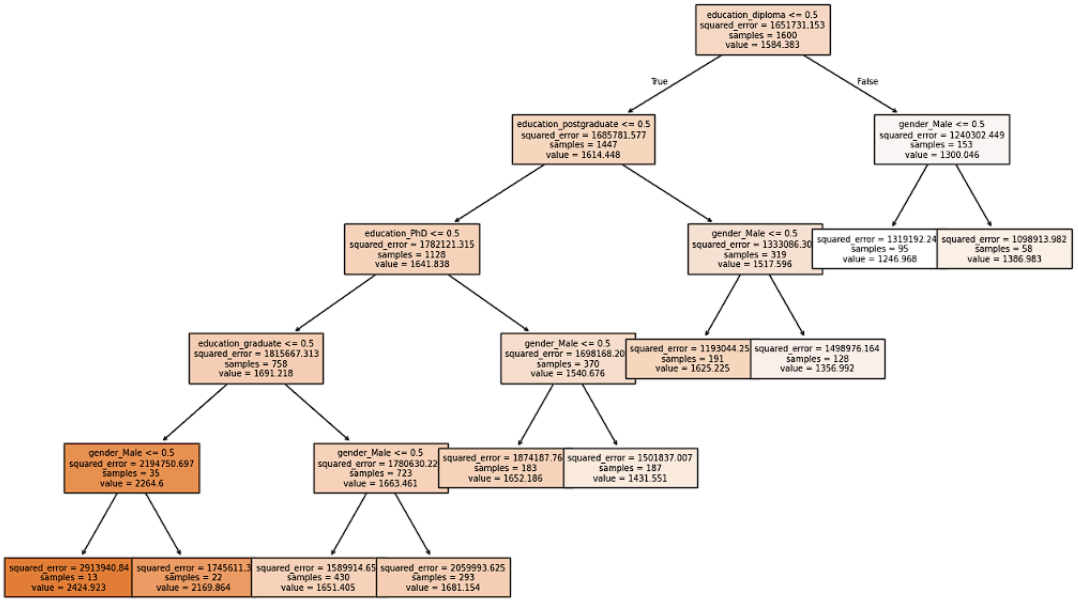
- ✅Multicollinearity: None detected (VIF ≈ 1).
- ✅Normality: Confirmed.
- ❌Homoscedasticity: Violated.
- See details in appendix 2.

Decision Tree Model

Given the moderate linear relationships, we explored non-linear relationships using a decision tree model. Tree visualisations highlight variable impact by colour intensity.

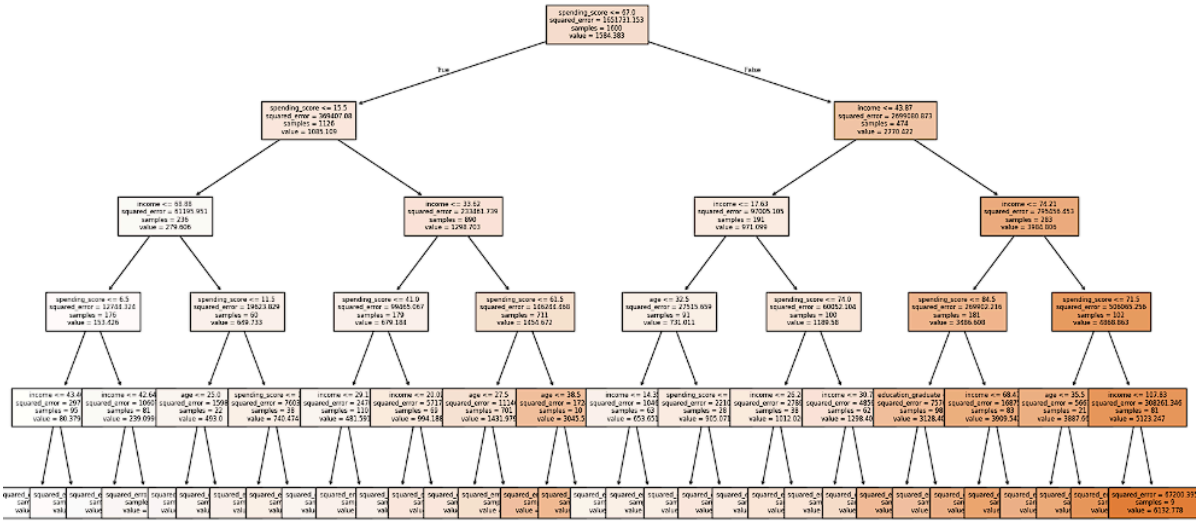
Initial Findings:

- Train R-squared: **0.0176**
- Test R-squared: **0.0151**
- *Interpretation:* Gender and education had little predictive power on their own.



Including Numerical Variables (Income, Spending Score & Age):

- Unpruned Model: R^2 = **0.9957** (high performance but overfitting)
- Pruned Model: R^2 = **0.9662** (reduced overfitting, but worsened performance)
- Interpretation: Pruning improved generalizability but confirmed that numerical variables dominate predictions.



Key Findings:

- **Spending Score** 61–84 leads to high loyalty point jumps
- **High earners:** Income 17–74, Score 15.5–67 (Avg. 3984 points)
- **Low earners:** Score ≤ 6 (Avg. 153 points)

Random Forest Model

To enhance predictive accuracy and mitigate overfitting, we applied a random forest model. See appendix 3 for details.

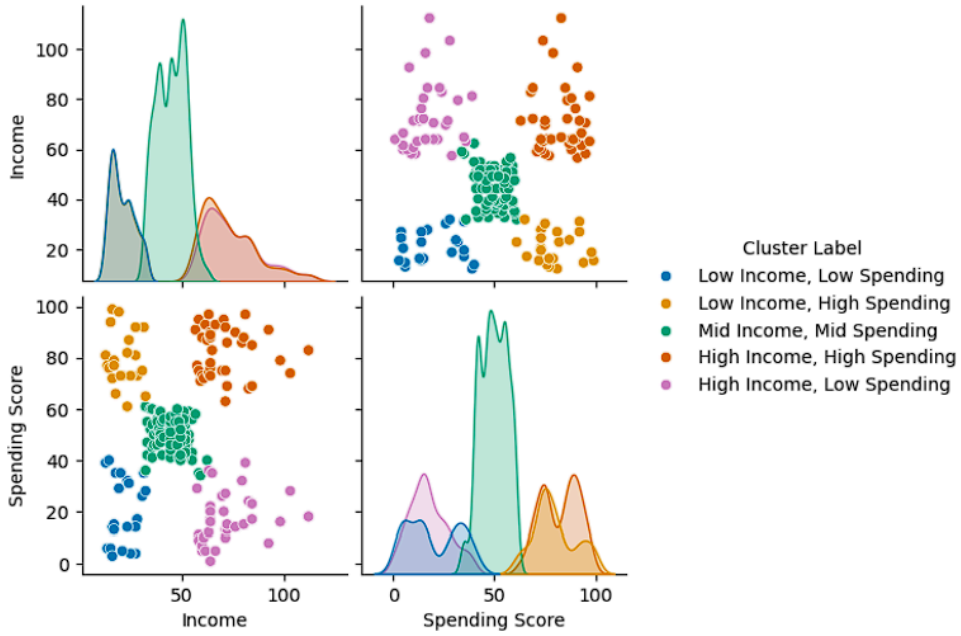
Conclusion: Best model for prediction. Reinforces the dominance of Spending Score and Income. Gender and Education deemed irrelevant.

Customer Segmentation (Clustering)

K-Means with 5 Clusters (Silhouette Method Approved):

1. High-Income, High-Spending (Avg. 3988 loyalty points) → VIP strategy
2. High-Income, Low-Spending (Avg. 912 loyalty points) → Incentivize purchases
3. Low-Income, High-Spending (Avg. 972 loyalty points) → Value deals
4. Low-Income, Low-Spending (Avg. 275 loyalty points) → Entry-level rewards
5. Mid-Income, Mid-Spending (Avg. 1420 loyalty points) → Loyalty maintenance

Pairplot to visualise the cluster distribution in different colours, to see the grouping better and also in the same graph see the customer count (mid income & spending as the biggest cluster)

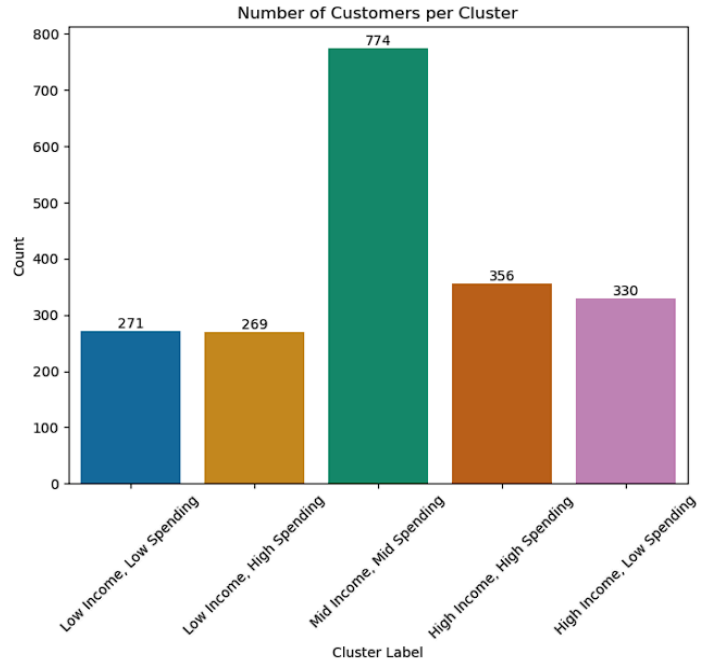


Best-selling products per cluster are identified, and promoting similar items to specific clusters can be explored.

cluster_label best_selling_products	High Income, High Spending	High Income, Low Spending	Low Income, High Spending	Low Income, Low Spending	Mid Income, Mid Spending
1	3403	3112	2162	1031	978
2	4399	4390	2253	2139	999
3	4415	4405	195	2173	4459
4	6504	6507	231	2261	6215
5	8923	8933	624	107	6233
6	8962	9064	811	760	9507
7	9080	3158	1940	1945	9529
8	2457	3277	1970	2793	9530
9	2518	3711	2829	2814	9596
10	2795	4047	3667	3657	249

Hierarchical Clustering suggested 4 clusters, but visually and quantitatively, 5 clusters are more distinct and practical. Analysis in appendix 4.

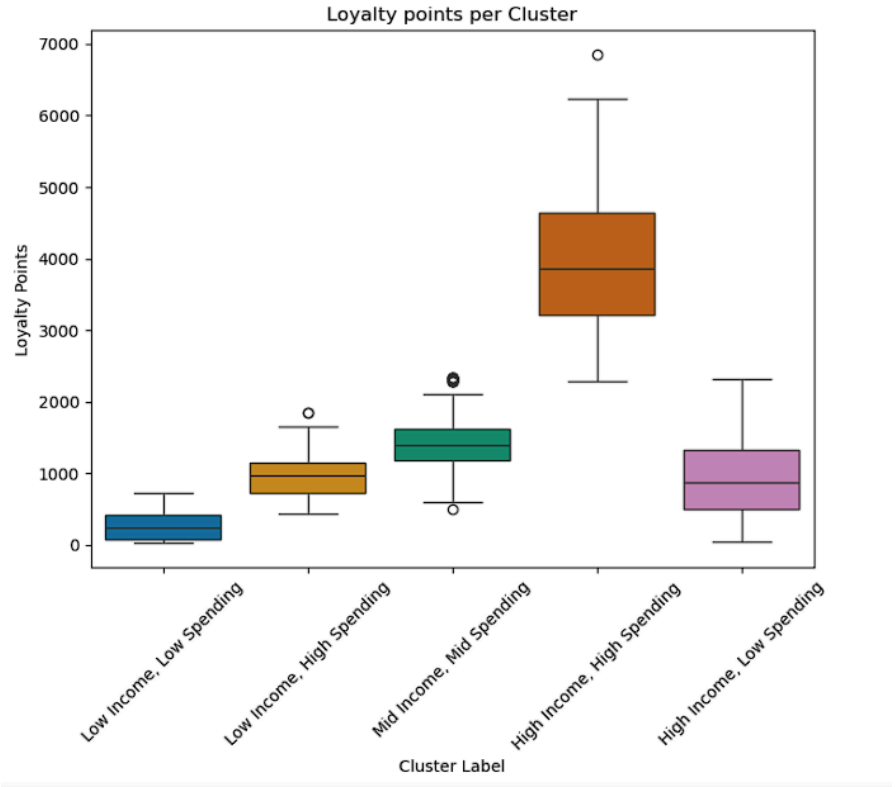
An additional bar plot shows cluster size.



Cluster averages for age, income, spending, and loyalty illustrate key differences. Age is not a good indicator, but average income and spending score is.

cluster_label	avg_age	avg_income	avg_spending_score	avg_loyalty_points
High Income, High Spending	36.0	73.0	82.0	3988.0
High Income, Low Spending	41.0	75.0	17.0	912.0
Low Income, High Spending	32.0	20.0	79.0	972.0
Low Income, Low Spending	44.0	20.0	20.0	275.0
Mid Income, Mid Spending	42.0	44.0	50.0	1420.0

Loyalty points per cluster as box plot, for further visualisation.



Other cluster combinations explored in appendix 5.
Conclusion: No further distinct groups detected that could support specific clustered marketing strategies.

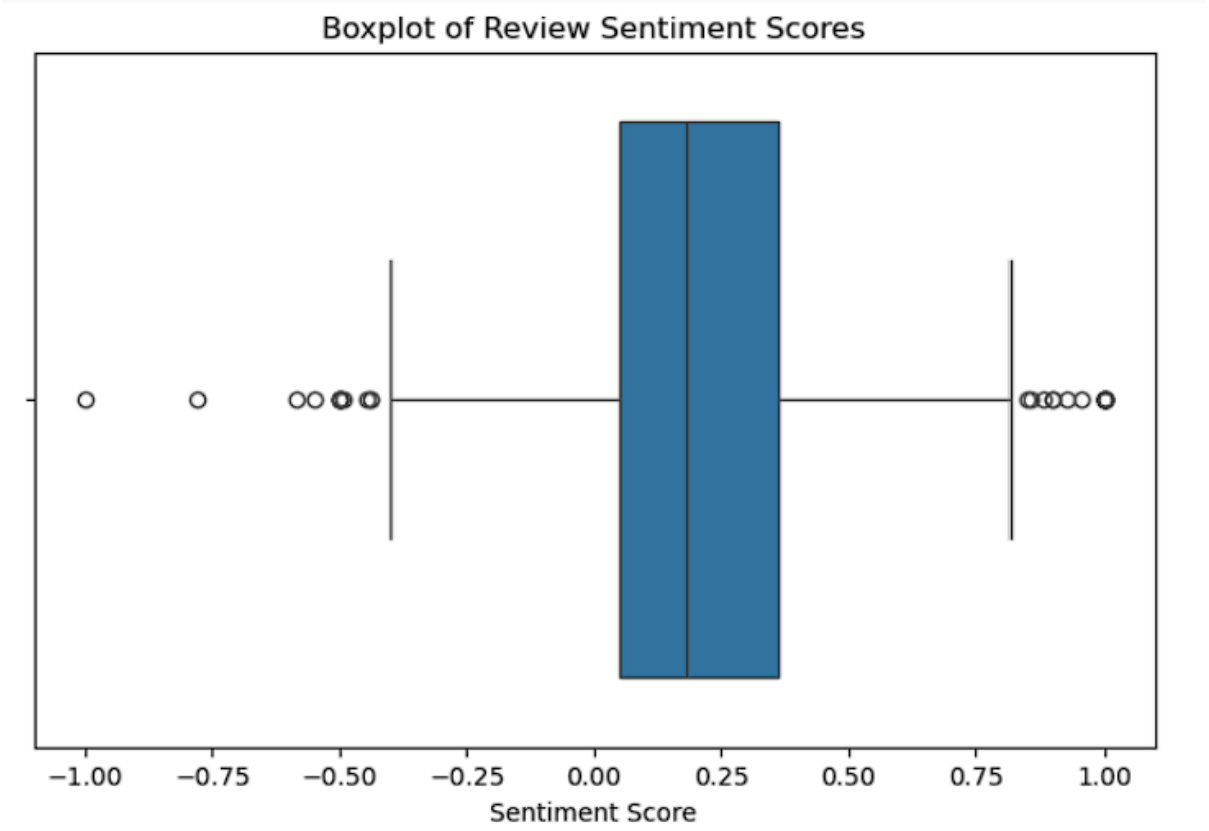
Sentiment analysis

Overall positive sentiment. 2nd quartile starts above 0. Very few extremely positive and extremely negative reviews.

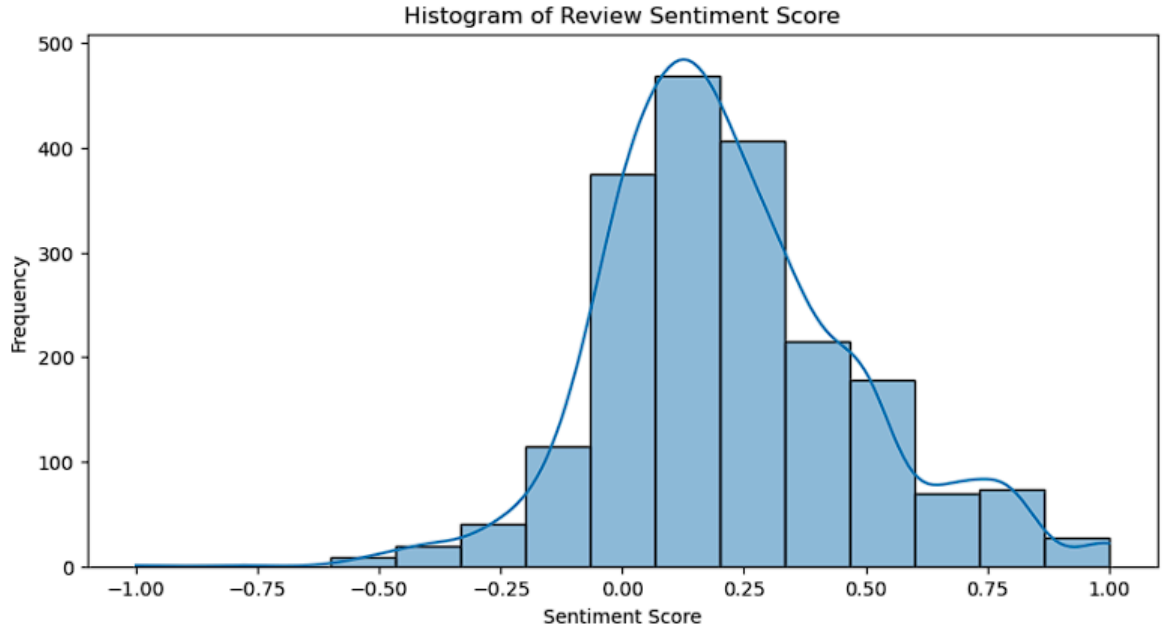
Most popular words in reviews

Most popular words in summaries

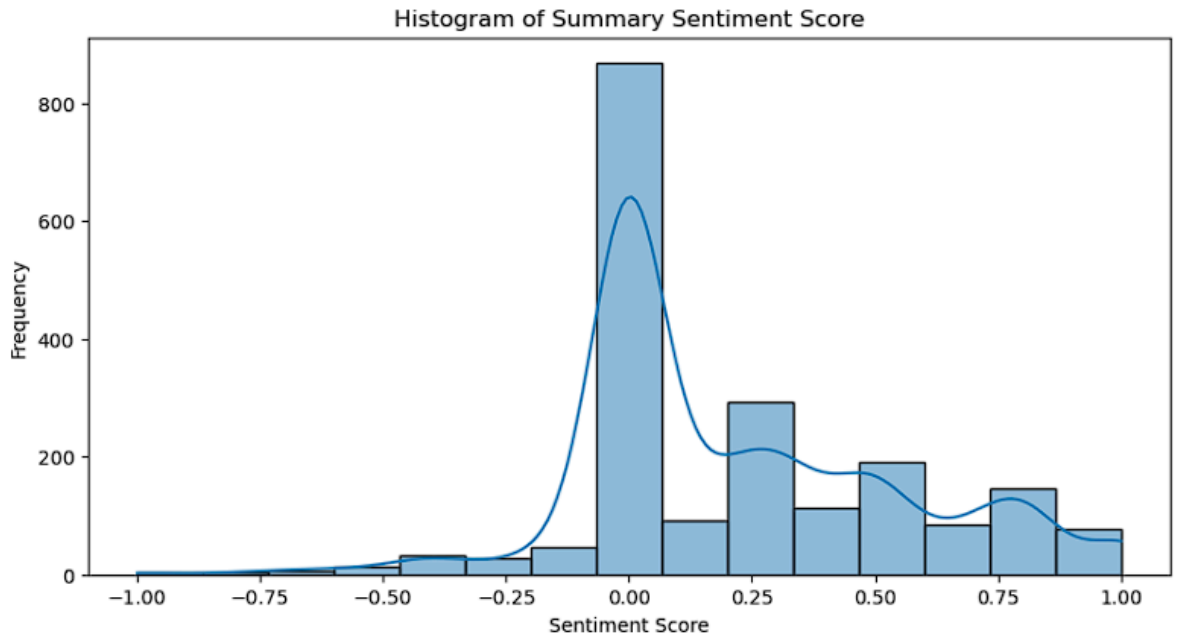
The boxplot of review sentiment scores reveals a positive skew, with the median above zero and minimal extreme outliers. This confirms the overall positive sentiment.



As boxplots can be difficult to interpret for some, a histogram offers an alternative view of sentiment scores.



Summary fields lack utility—frequent pseudo-star ratings suggest a 5-star system would be more actionable and engaging.



Product-Level Insights

- 200 products listed, most with fewer than 13 reviews
- Hard to draw definitive conclusions on product performance
- Use (positive) review popularity to identify potential top-performing products

Products with the most negative and positive reviews are visualised in the tables below.

Products with the most negative reviews:

sentiment	negative	neutral	positive
product			
2253	4	0	6
3165	4	0	6
486	4	0	6
9597	3	2	5
4047	3	0	7
2795	3	1	6
876	3	1	6
1212	2	0	8
979	2	1	9
6504	2	0	8

Products with the most positive reviews:

sentiment	negative	neutral	positive
product			
1012	1	1	11
3629	0	0	10
6770	0	0	10
2371	0	0	10
2286	0	0	10
6271	0	0	10
6424	0	0	10
2139	0	0	10
6715	0	0	10
1618	0	0	10

Further analysis on segmentation of review sentiment in appendix 6.

Strategic Recommendations

Spending Score and Income show the strongest positive correlations with loyalty points, confirmed by regression and decision tree models. Random Forest achieved the highest predictive power with minimal overfitting. Cluster analysis uncovered five actionable customer segments, offering clear paths for targeted marketing. Sentiment analysis revealed largely positive reviews, though the review system lacks structure. These patterns support predictive strategies for loyalty and retention.

- **Data Improvements:**
  - Add Customer ID
  - Include purchases without reviews
  - Integrate product metadata (names, descriptions)
  - Add timestamps for all transactions

- **Loyalty Point Incentives:** Create gold, silver, and bronze tiers, offering exclusive items or early novelty access.

- **Cluster-Specific Marketing:**
  - High-Spending, High-Income: Reward with exclusives & early novelty access
  - Low-Spending, High-Income: → Upsell premium items & give bonus loyalty points on high value purchases
  - High-Spending, Low-Income: Encourage repeat buys with small next-purchase discounts.
  - Low-Spending, Small-value vouchers to boost engagement.
  - Mid-Spending, Mid-Income: Maintain loyalty with next-purchase loyalty point boost

- **Review System Overhaul:**
  - Implement a 5-star rating system and incentivize reviews with additional loyalty points to increase feedback volume.
  - Suggest similar items to those rated 4 or 5 stars by the customer.

Appendix 1 – Data Cleaning and Preprocessing

Given that high loyalty points likely indicate longer participation, these were retained as they reflect legitimate customer behaviour. Some examples below.

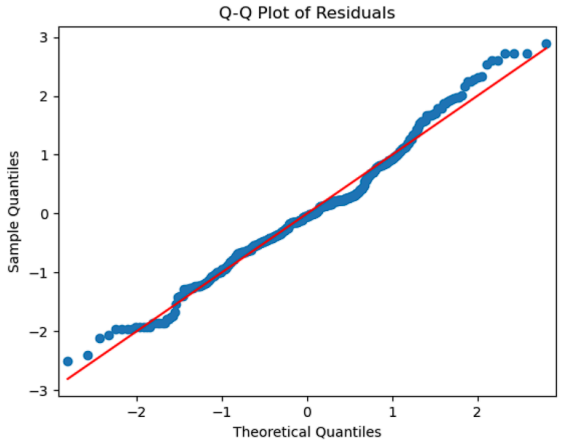
	gender	age	remuneration (k£)	spending_score (1-100)	loyalty_points	education	language	platform	product	review	summary
123	Male	39	56.58	91	3634	graduate	EN	Web	5510	This book is small in size and probably best f...	Small sized...
127	Male	38	58.22	95	3866	postgraduate	EN	Web	2849	... In a little package. My 6 year old loved m...	Lots of fun...
141	Male	34	61.50	93	3808	graduate	EN	Web	4477	My son is 5 and LOVES robots. He likes to chan...	robot lover
143	Female	34	62.32	87	3610	graduate	EN	Web	2795	My one-year-old got this as a birthday present...	Doesn't hold up well
145	Male	32	63.14	97	3954	Basic	EN	Web	2811	This is a poorly put together piece of junk. W...	Junk
...	...	...	...	...	...	...	...	...	...	...	...

Appendix 2 – Regression Modeling and Evaluation

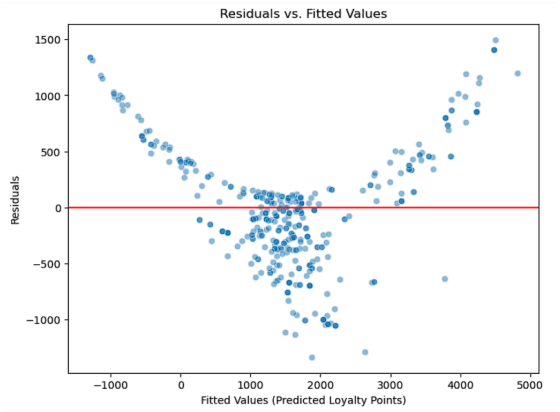
- **Assumption Checks:**
  - Multicollinearity: VIF values (~1) confirmed variable independence.

	Feature	VIF
0	const	20.77
1	spending_score	1.05
2	income	1.00
3	age	1.05

  - Normality: Residuals followed a normal distribution.



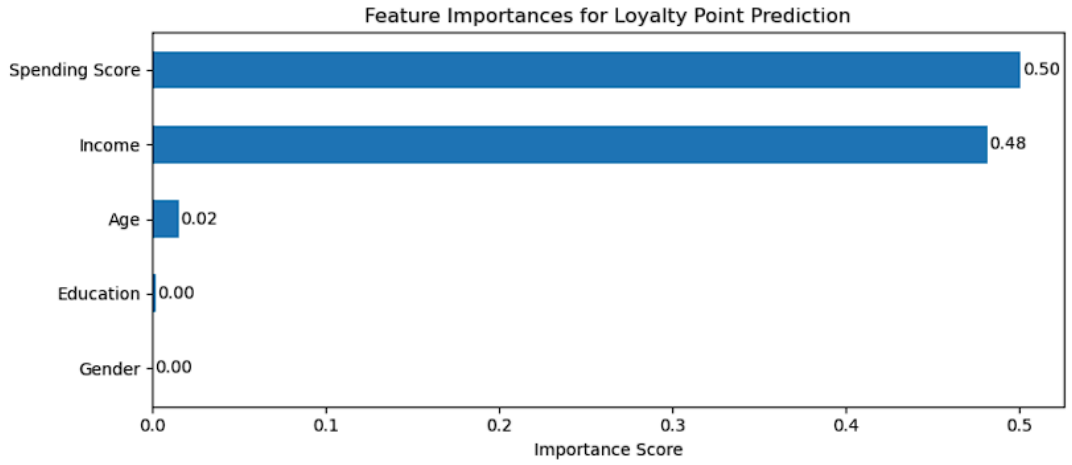
- Homoscedasticity: A U-shaped pattern in residuals indicated violation, affecting predictive reliability.



Appendix 3 – Random Forest Model

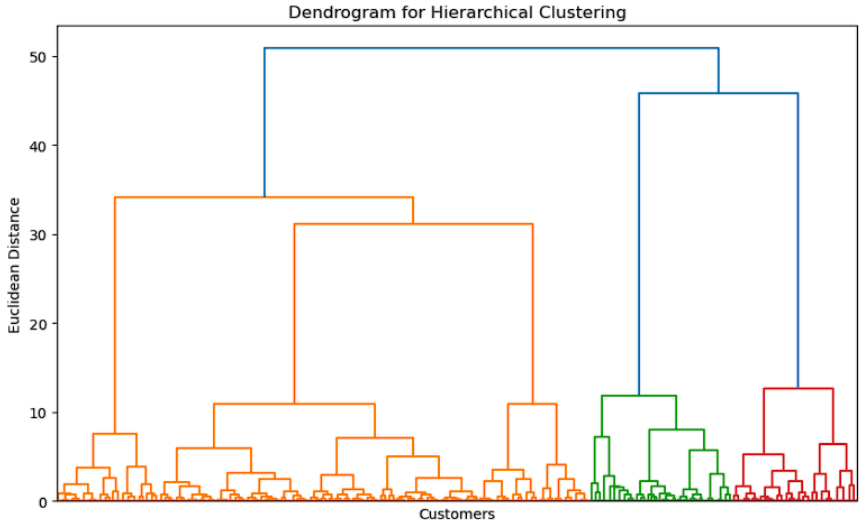
- *Train* R^2: 0.9778
- *Test* R^2: 0.9707
- *Cross-Validation* R^2: 0.9732
- Outperformed all other models while avoiding overfitting

Barplot was chosen to visualise feature importance, confirming clearly the insignificance of education and gender.



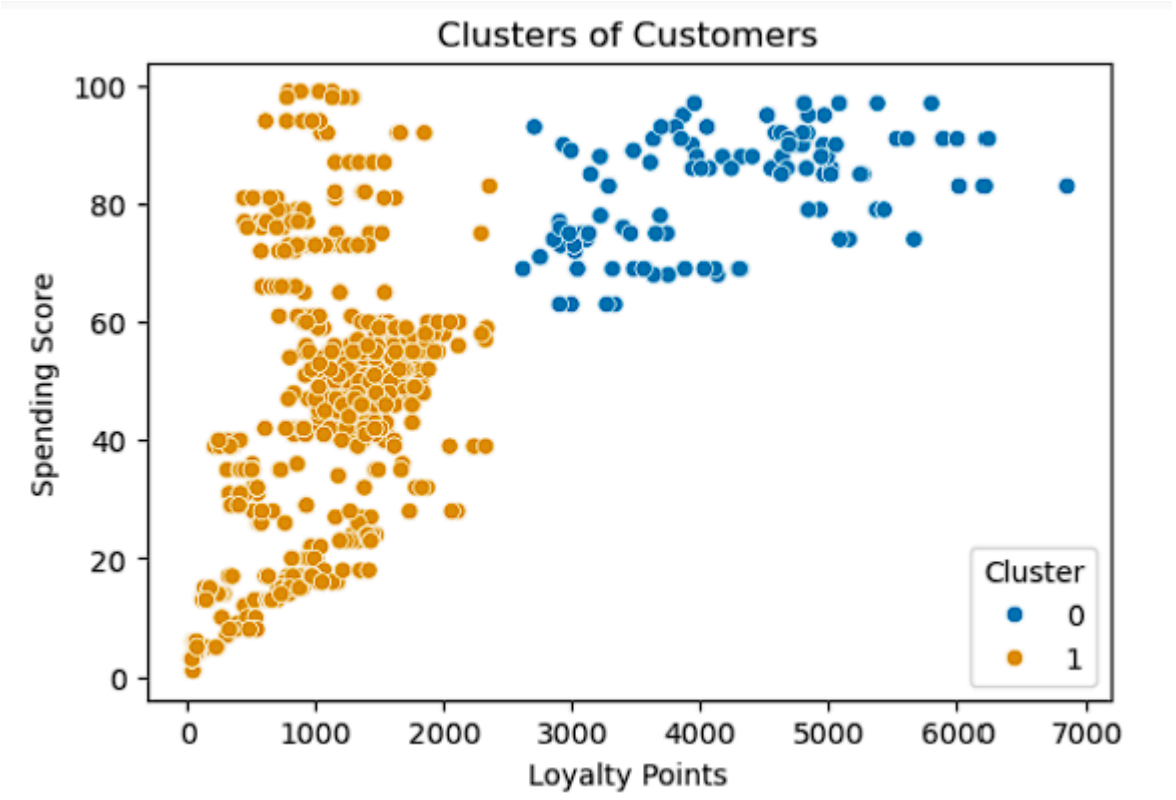
Appendix 4 – Additional Clustering Analysis

Hierarchical clustering explored to compare to k means clustering. Results suggested 4 clusters to be best but looking at the big amount of midrange, this does not seem intuitive since the mid range group contains a big group and is clearly inbetween the other clusters..



Appendix 5 – Further Clustering Analyses

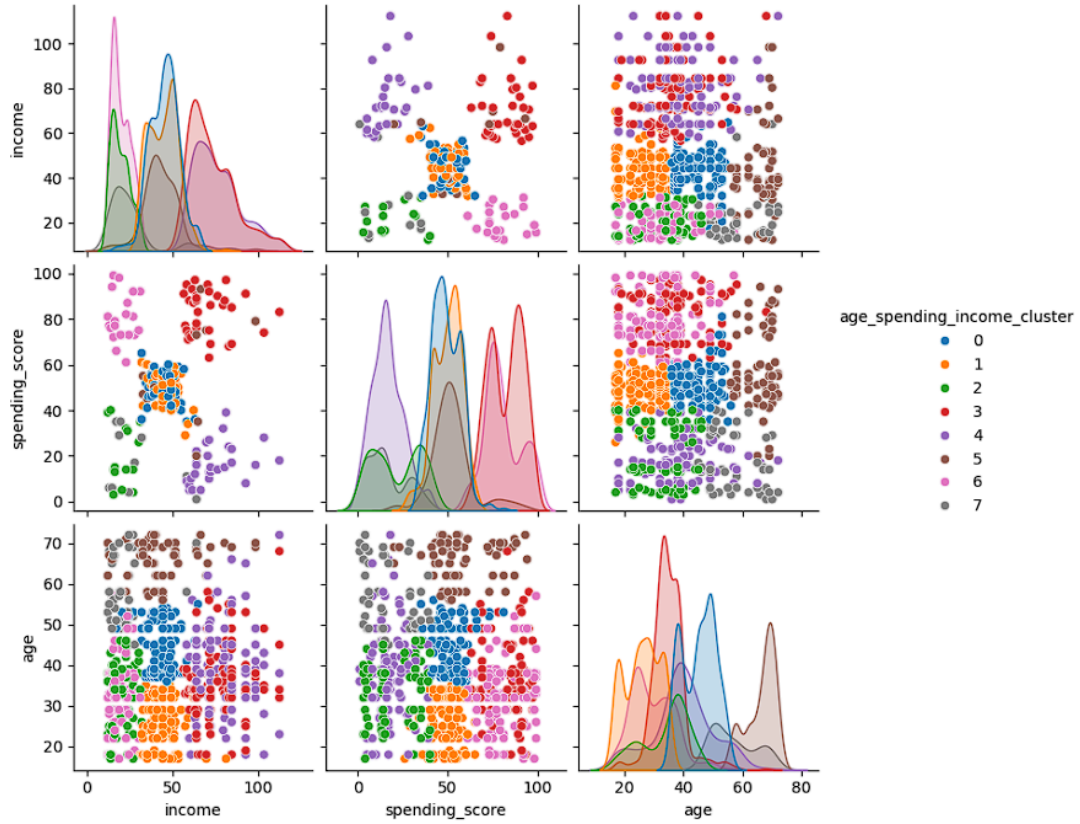
Spending score vs loyalty points, suggested 2 clusters, visualised in a scatterplot.



Income vs loyalty points also 2 clusters suggested, visualised in a scatter plot.

Spending score vs age, 2 clusters suggested visualised in a scatterplot.

Clustered age, income and spending score together. There is too much overlap, the 7 clusters that were suggested are not really interpretable. Visualised in a pairplot to detect how logical the clusters are, too complex.



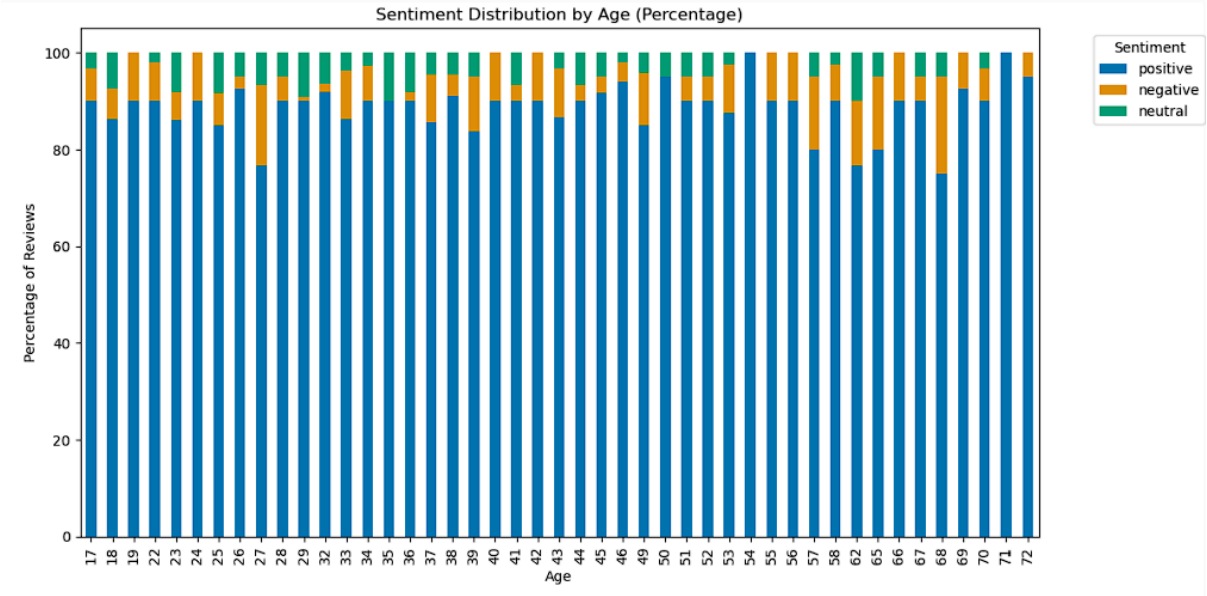
Appendix 6 – Segmentation of Review Sentiment

I have done some more calculations to find out if there are any other segmentations showing with the review sentiment. A table is displaying how many reviews males and females left and the percentage of positive, neutral and negative to the total.

Review stats per Gender:

gender	Female	Male
sentiment		
positive	987.0	783.0
negative	78.0	57.0
neutral	55.0	40.0
total_reviews	1120.0	880.0
positive_percent	88.1	89.0
negative_percent	7.0	6.5
neutral_percent	4.9	4.5

Visualised in a stacked bar plot percentages of positive, negative and neutral reviews. The stacked barplot lets us compare it way better than a bar plot, because the amount of reviews varies significantly.



Loyalty points have been grouped into 100 buckets, loyalty point buckets with most positive reviews are displayed in a table.

Loyalty Point Buckets with most Positive Reviews as Percentage to total Reviews:

	Loyalty Points Bucket	Positive Review Percentage	Total Reviews
0	1300	92.1	151
1	1100	90.3	134
2	1600	88.8	116
3	1200	91.7	109
4	1000	87.5	104
5	900	88.8	98
6	1400	90.8	98
7	700	94.7	95
8	0	87.1	93
9	500	85.5	83

Income has been grouped into 1k buckets, the income buckets with most positive reviews are displayed in a table.

Income Buckets with most Positive Reviews as Percentage to total Reviews:

	Income Bucket	Positive Review Percentage	Total Reviews
0	44.0	92.5	134
1	63.0	84.8	125
2	50.0	82.8	64
3	39.0	88.9	63
4	49.0	93.5	62
5	27.0	95.0	60
6	71.0	91.5	59
7	16.0	87.9	58
8	84.0	98.3	58
9	15.0	92.7	55

Overall, no valuable segmentation could be found regarding review sentiment.