Technical Report: Customer Loyalty Analysis Using Regression and Clustering

PJ Stock

On Behalf of: Turtle Games

July 21, 2025

Contents

1 Introduction

The business questions driving this analysis are:

- Customer engagement with loyalty points
- Customer segmentation and marketing
- Text sentiment analysis of customer reviews
- Predictive modeling of loyalty points

This report addresses these questions through exploratory analysis, linear regression, decision tree modeling, customer segmentation via clustering, and sentiment analysis of reviews. The findings aim to provide actionable insights for enhancing customer loyalty and informing business strategies.

2 Initial Exploratory Analysis and Linear Regression

2.1 Data Cleaning and Descriptive Statistics

The dataset was cleaned, with no missing values detected. Columns were renamed for clarity, and basic descriptive statistics were computed.

Table 1: Descriptive Statistics of Key Variables

Variable	**Mean**
Age	39
Salary (k)	48
Loyalty Points	1578
Spending Score	50

2.2 Linear Regression Model

A linear regression model was developed to predict loyalty points based on age, salary, and spending score. The model achieved an R^2 score of 0.83, explaining 83% of the variation in loyalty points. All predictors are statistically significant ($p < 0.05$).

Key Insight

The linear regression model explains 83% of the variation in loyalty points, with salary and spending score as the strongest predictors ($p < 0.05$).

Table 2: Linear Regression Coefficients

Predictor	**Coefficient**	**P-Value**
Salary	34.48	< 0.05
Spending Score	33.72	< 0.05
Age	10.77	< 0.05

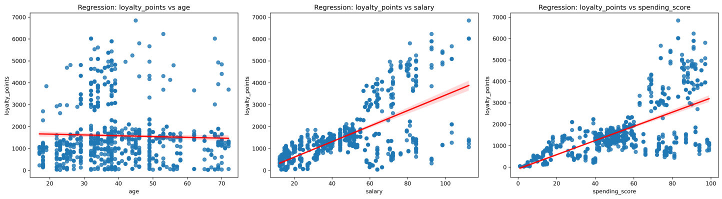

Figure 1: Regression plot of loyalty points vs. other variables, showing a strong positive relationship with spending score.

2.3 Recommendations for Future Analysis

- Explore nonlinear modeling to capture complex relationships.
- Investigate interactions between variables (e.g., spending score impact at different salary levels).

3 Creating a Decision Tree

3.1 Data Preparation

Loyalty points were right-skewed, as shown in a histogram. A log transformation was applied, resulting in a slightly left-skewed but more even distribution.

3.2 Model Development and Pruning

The dataset was split into training (70%) and test (30%) sets with a random state of 42. A Decision Tree Regressor was fitted, followed by pruning (max depth = 7, min samples leaf = 3, min samples split = 4).

Table 3: Decision Tree Performance

Model Type	**MSE**	$\mathbf{R^2}$
Unpruned	12497.48	0.9923
Pruned (Depth=7)	27839.80	0.9828

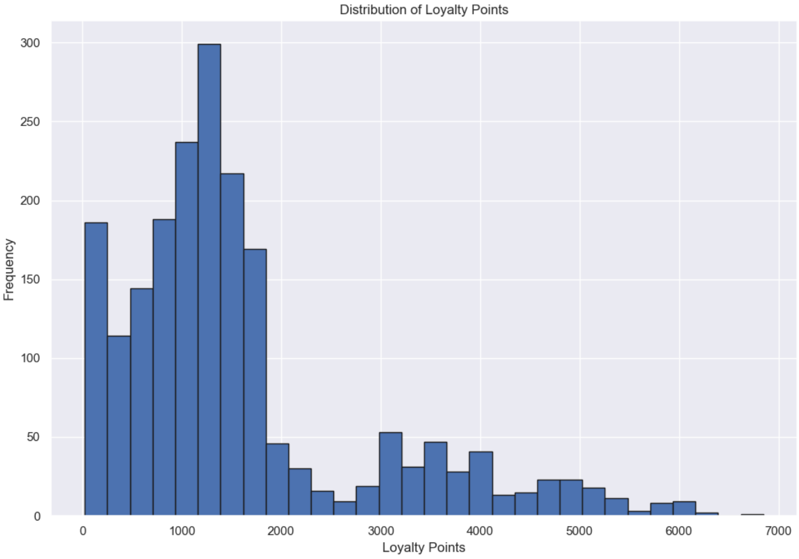

Figure 2: Original loyalty points distribution (right-skewed).

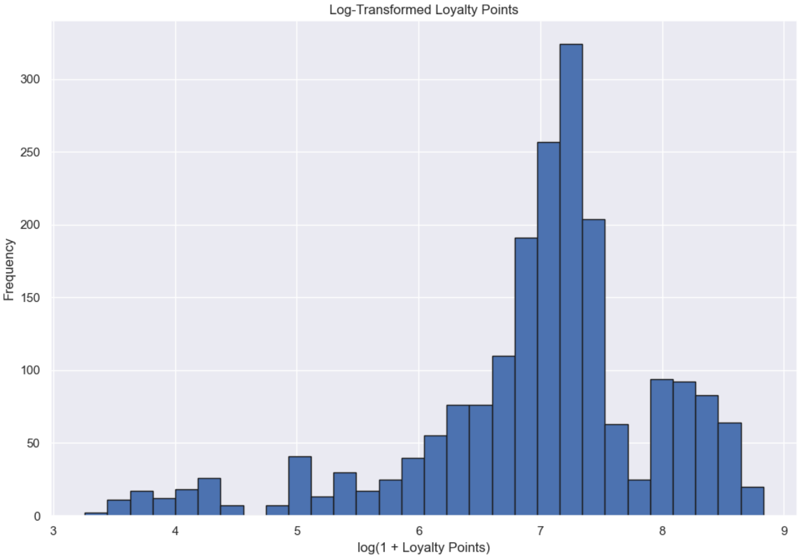

Figure 3: Loyalty points distribution after log transformation.

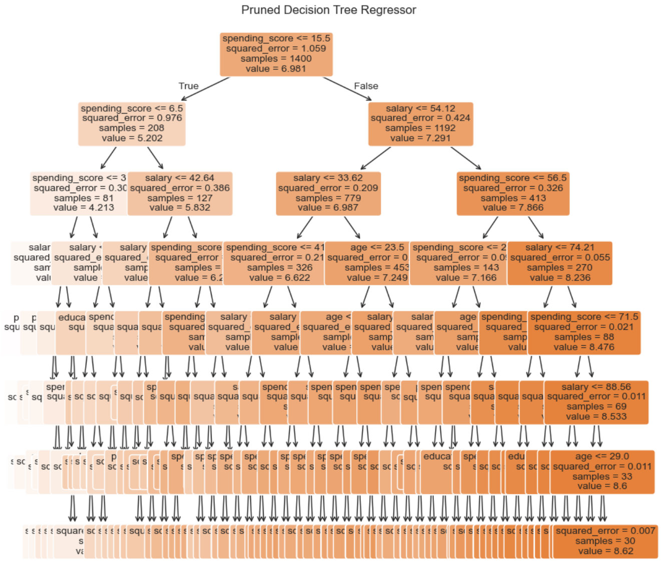

Figure 4: Simplified decision tree, with spending score as the primary split.

3.3 Key Findings

- The first split is on spending score (> 71.5), indicating its dominance in driving loyalty.
- High spenders (score > 71.5, salary > 74.21) fall in the top loyalty quartile.
- Low spenders (score ≤ 6.5, salary ≤ 42.64) have the lowest loyalty points.

Key Insight

High-spending customers (score > 71.5, salary > 74.21) form the top loyalty quartile, while low spenders (score ≤ 6.5, salary ≤ 42.64) are at the bottom.

4 Clustering Salary and Spending Scores

4.1 Data Visualization

Salary and spending score distributions were plotted to understand customer behavior.

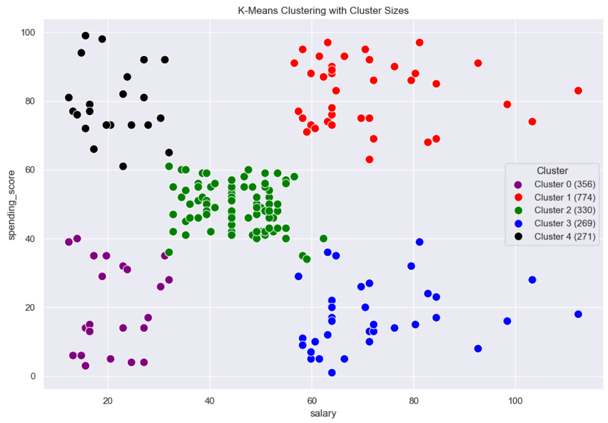

Figure 5: Scatter plot of salary vs. spending score, color-coded by clusters.

4.2 Clustering Methodology

The elbow and silhouette methods determined an optimal 5 clusters.

4.3 Cluster Insights

Five distinct customer segments were identified.

Table 4: Customer Segments from Clustering

Cluster	Salary Range	Spending Score Range
1	Low	Low
2	Medium	High
3	High	Low
4	High	High
5	High	Medium

Marketing Opportunity

Cluster 3 (high salary, high spending score) represents premium customers ideal for targeted loyalty campaigns.

5 Sentiment Analysis of Customer Reviews

5.1 Data Preparation

The customer review dataset was cleaned and preprocessed. The following steps were applied:

- Dropped unnecessary columns to focus only on reviews and summaries.
- Checked for missing values.
- Converted all words to lowercase and removed punctuation.
- Removed duplicates.
- Tokenized words to prepare the text for generating word clouds and frequency analysis.
- Removed stop words and generated frequency distribution.

5.2 Word Cloud and Frequency Analysis

To visualize the most common terms in the reviews and summaries, word clouds were created with the tokenized data.

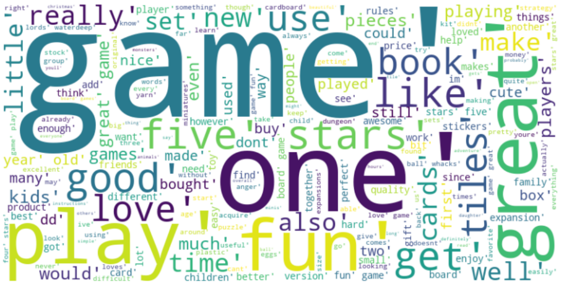

Figure 6: Word cloud of customer reviews and summaries combined.

5.3 Sentiment and Polarity Analysis

Sentiment analysis was conducted using TextBlob for polarity scores and VADER for sentiment scores to evaluate the emotional tone of customer reviews. Polarity and sentiment distributions were visualized using histograms.

- **Full Review Polarity (TextBlob)**:
  - Polarity scores range from -1 (negative) to +1 (positive), with most scores between -0.25 and +0.75.
  - The distribution peaks at +0.1 to +0.2, indicating slightly positive reviews.
  - Few reviews are extremely negative (below -0.5), suggesting moderate dissatisfaction rather than strong negativity.
- **Sentiment Scores (VADER)**:

  - Scores are strongly positively skewed, with most values clustered near +1.0.
  - This suggests VADER detects stronger positive sentiment compared to TextBlob.
  - Negative reviews are infrequent, indicating generally favorable customer reception.

- **Review Summaries**:
  - There is bimodal clustering, with scores concentrated around 0.0 (neutral) and 0.6 to 0.7 (positive).
  - This could suggest summaries are more likely to be either neutral or highly positive.
  - Extremely negative summaries are rare, possibly due to their short nature.

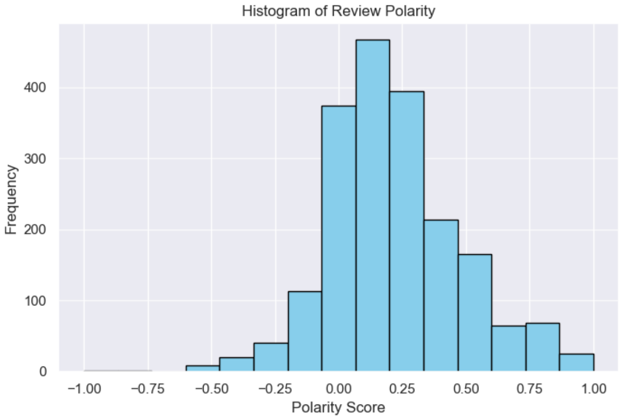

Figure 7: TextBlob polarity scores of full reviews (peak at +0.1 to +0.2).

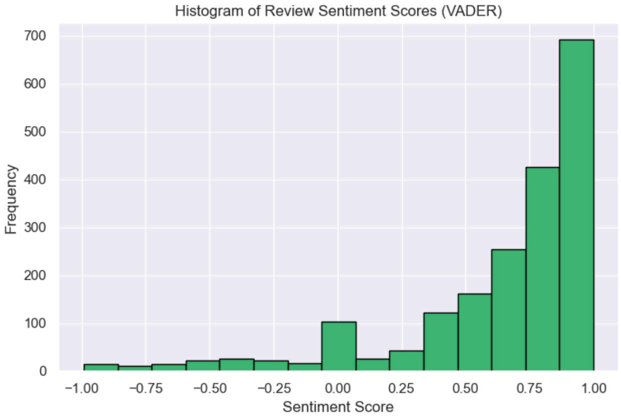

Figure 8: VADER sentiment scores (strong positive skew near +1.0).

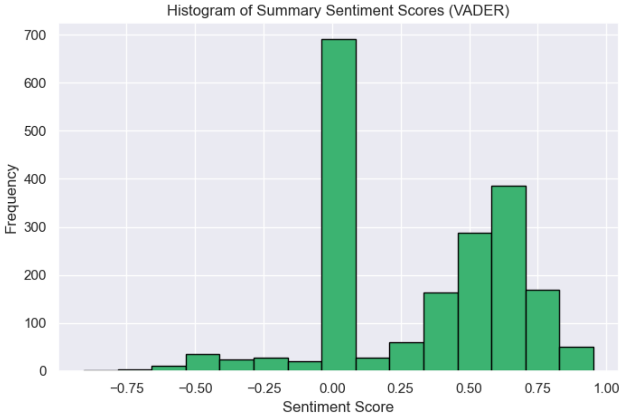

Figure 9: Summary scores (bimodal at 0.0 and 0.6–0.7).

Key Insight

Customer reviews are generally positive, with TextBlob showing slightly positive polarity (peak at +0.1 to +0.2) and VADER indicating stronger positive sentiment (near +1.0).

6 Further Linear Regression

6.1 Data Preparation and Normality Testing

Linear regression was performed excluding categorical variables (gender and education) to focus on numerical predictors: age, salary, and spending score. The distributions of loyalty points and spending score were assessed for normality using the Shapiro-Wilk test:

- **Loyalty Points**: Shapiro-Wilk score = 0.84307, p-value < 0.05, indicating a non-normal distribution with strong deviation.
- **Spending Score**: Shapiro-Wilk score = 0.96835, p-value < 0.05, non-normal.
- **Age**: Shapiro-Wilk score = 0.95242, p-value $< 2.2 \times 10^{-16}$, non-normal.
- **Salary**: Shapiro-Wilk score = 0.96768, p-value $< 2.2 \times 10^{-16}$, non-normal.

Loyalty points exhibit the strongest deviation from normality, suggesting potential need for data transformation (e.g., log transformation).

6.2 Correlation Analysis

Correlation analysis was conducted to assess relationships between variables. The correlation matrix is shown below:

Table 5: Correlation Matrix for Key Variables

	Loyalty Points	Age	Salary	Spending Score
Loyalty Points	1.0000	-0.0424	0.6161	0.6723
Age	-0.0424	1.0000	-0.0057	-0.2243
Salary	0.6161	-0.0057	1.0000	0.0056
Spending Score	0.6723	-0.2243	0.0056	1.0000

Key Insight

Spending score (0.6723) and salary (0.6161) show medium to high correlation with loyalty points, while age has negligible correlation (-0.0424).

6.3 Initial Linear Regression Model

The linear regression model was fitted using age, salary, and spending score as predictors. All variables were statistically significant (p-value $< 2 \times 10^{-16}$), as shown below:

Table 6: Linear Regression Coefficients

Predictor	Coefficient	Std. Error	t-value	p-value
Intercept	-2203.0598	52.3609	-42.08	$< 2 \times 10^{-16}$
Age	11.0607	0.8688	12.73	$< 2 \times 10^{-16}$
Salary	34.0084	0.4970	68.43	$< 2 \times 10^{-16}$
Spending Score	34.1832	0.4519	75.64	$< 2 \times 10^{-16}$

The model shows a strong fit with $R^2 = 0.8399$ and adjusted $R^2 = 0.8397$. Residuals were analyzed for normality:

Multicollinearity was assessed using Variance Inflation Factor (VIF), with low values indicating no significant issues:

Table 7: VIF for Predictors

Predictor	VIF
Age	1.053015
Salary	1.000052
Spending Score	1.053014

The actual vs. predicted plot shows a positive trend but reveals scatter at higher predicted values, suggesting potential improvements via log transformation or nonlinear methods like Random Forests.

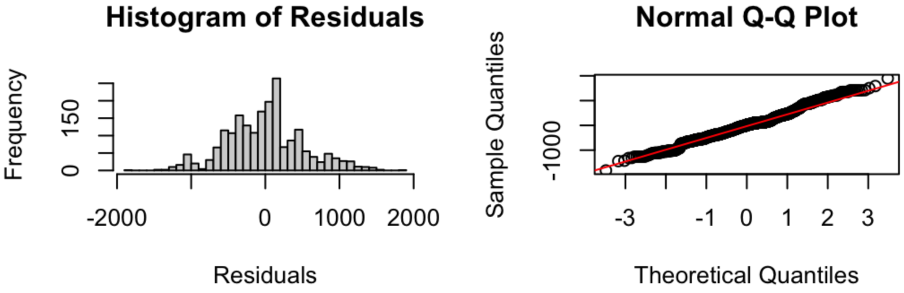

Figure 10: Histogram (left) and Q-Q plot (right) of residuals, indicating approximate normality.

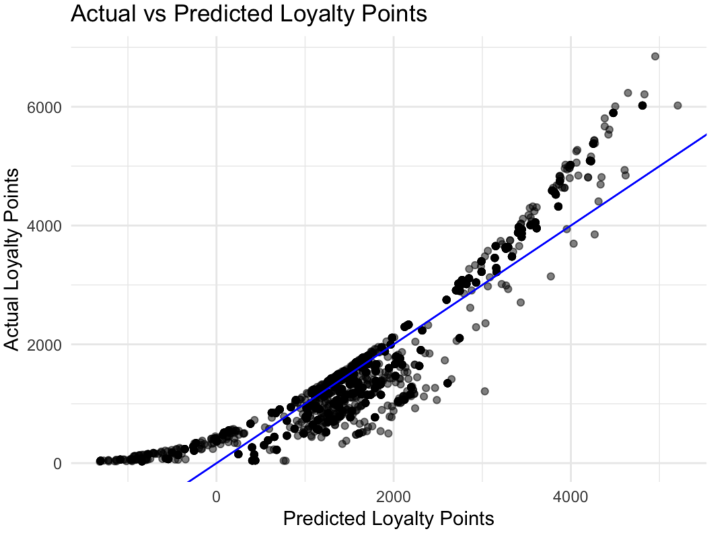

Figure 11: Actual vs. predicted loyalty points, showing customers overpeforming the trend line at higher values.

6.4 Model with Categorical Variables

Gender and education were converted to factors and included in a second linear regression model. Results indicate:

- **Gender**: Males accumulate more loyalty points than females.
- **Education**: Customers with graduate degrees accumulate more loyalty points than the baseline, while PhD holders accumulate fewer.

Key Insight

Including gender and education improves the model, with males and graduate-degree holders showing higher loyalty points.

6.5 Customer Personas and Predictions

Three customer personas were defined to predict their loyalty points based on the linear regression model:

Table 8: Defined Customer Personas

Persona	Characteristics
Persona 1	Young, low salary, low spending score
Persona 2	Middle age, high salary, high spending score
Persona 3	Older age, middle salary, middle spending score

The predicted loyalty points for these personas are shown below:

Table 9: Results: Predicted Loyalty Points for Personas

Persona	Predicted Loyalty Points
Persona 1	168
Persona 2	939
Persona 3	320

Higher salary and spending score correlate with higher predicted loyalty points, consistent with prior findings.

6.6 Segmenting Underperforming Customers

A model was used to identify customers whose actual loyalty points are lower than predicted, indicating potential for targeted interventions. The plot below highlights these underperforming segments:

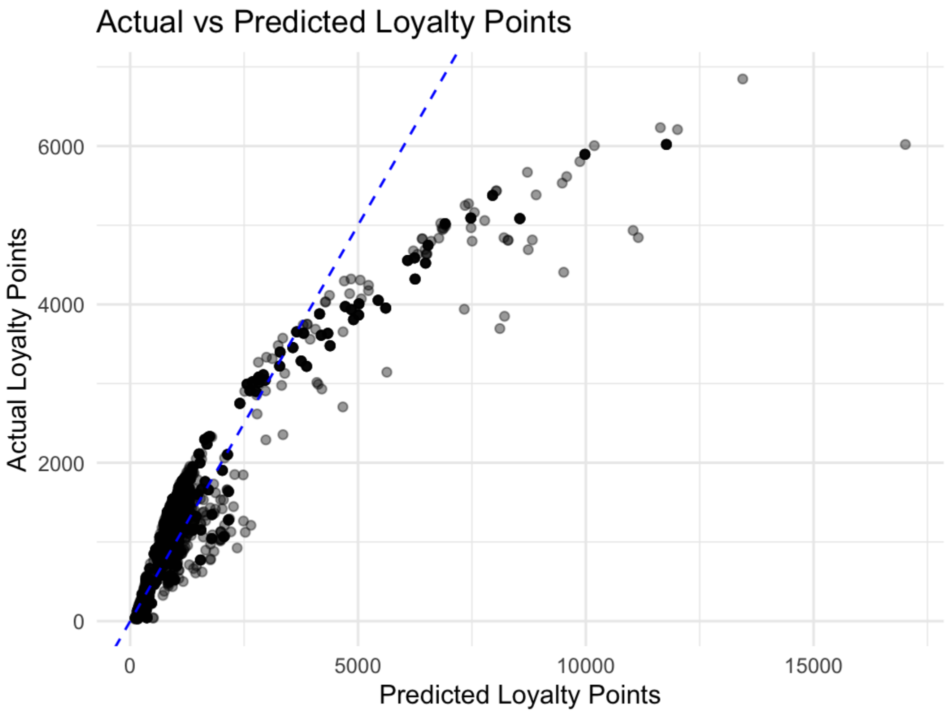

Figure 12: Customers at the higher end fall below the expected trend line.

Marketing Opportunity

Customers with high engagement or income but lower-than-predicted loyalty points represent a key segment for targeted loyalty campaigns to boost point accumulation.