TURTLE GAMES

Business Context

Turtle Games (TG) is a game manufacturer and retailer. The company sells a wide array of products, including books, board games, and video games, catering to a diverse global customer base of all ages.

This report will analyse marketing and sales data (Turtle_reviews.cvs) to improve sales performance, focusing on **four main objectives:**

1. Understanding how customers accumulate points and engage with the loyalty system[1] (see footnotes).
2. Evaluating customer segmentation and how groups can be targeted
3. Perform opinions mining about TG products.
4. Providing insights into current loyalty points system performance and recommendations on suitability of data for predictive models.

This information is intended for the marketing and sales department to support data-informed decisions on strategic planning for customer retention, engagement, satisfaction, and effective marketing strategies.

Methodological Approach:

This analysis used Python and R as analytical tools. While Python was preferred, we integrated R to align with TG' workflow systems and leverage R's visualisation capabilities[2].

Summary of Analysis

Software Tool	Analysis and ML Model Deployed	Scope
Python/R	Exploratory Data Analysis	Objectives 1 and 4
Python/R	Linear and Decision Tree Regressor	Objectives 1 and 4
Python	K-Means Clustering	Objective 2
Python	Analysis of Sentiment with VADER and TextBlob	Objective 3

ML models underwent iterative steps of adjustments. Full details can be found in the Jupyter Notebook.

[1] **The company relies on a point-based loyalty system where points are proportional to the value of purchases.** This common strategy helps driving customer retention by being simple and clear, providing tangible value to customers. (LoyaltyLion,2024)

Data Wrangling

Data validation was conducted both in Python and R to ensure data integrity, including checks for duplicate entries (see appendix 1. for duplicates classification) and null values.

Data manipulation included:

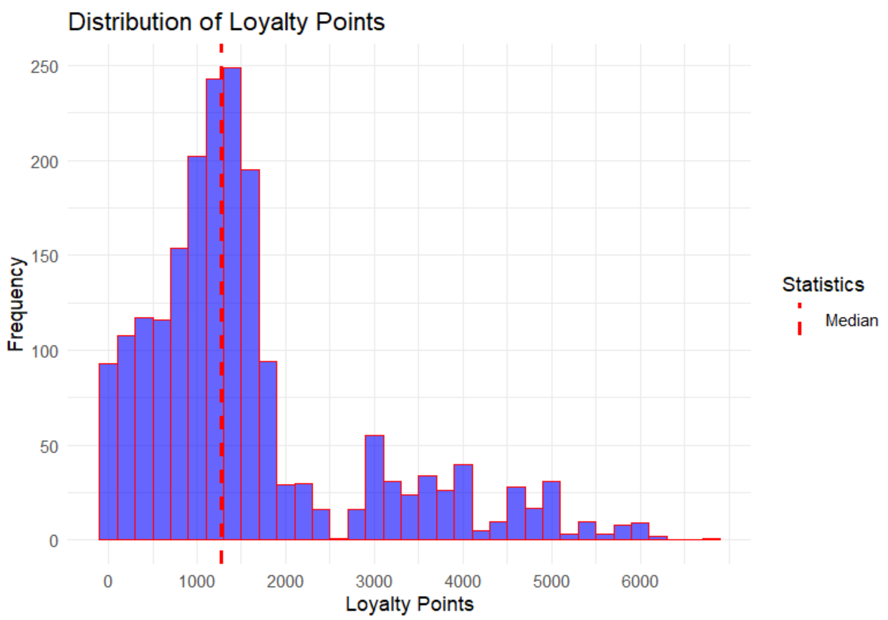
- Renaming columns for clarity (e.g "Remuneration" to "Income").
- Dropping unused features ("Language" and "Platform").

Refer to **turtle_games_final.csv** for the cleaned dataset.

Approach and Visualisation

Step 1: Exploratory Data Analysis (EDA)

We began our analysis by examining the distribution of loyalty points to evaluate customer engagement levels. To visualize the distribution, we plotted a histogram and conducted statistical tests for normality (Tab 2). Results showed not normality of distribution, highlighting the presence of high variance, outliers and many customers with low frequencies (median = 1276).



Fig(1)

Test	Value	Inference
Shapiro-Wilk	w = 0.84	p-value < 2.2e-16
Skweness	1.463	Indicates Positive Skewness
Kurtosis	4.708	Indicates Leptokurtic Distribution

This suggests different levels of customer engagements and significantly opportunities from tailored strategies.

For our next steps, high variance and outliers in loyalty points have significantly impacted LR modelling and the assumption of normal residuals, prompting consideration of alternative ML approaches.

Step 2) Analysis of factors influencing loyalty points accumulation and features importance

2.1) Explaining Loyalty Points.

We used Pearson correlation (r) to understand the linearity (strength) and direction of relationships between yearly income, spending score, age, and customer loyalty (see Table 3 for r values and the appendix for scatterplots).

	age	income	spending_score	loyalty_points
age	1.000000	-0.005708	-0.224334	-0.042445
income	-0.005708	1.000000	0.005612	0.616065
spending_score	-0.224334	0.005612	1.000000	0.672310
loyalty_points	-0.042445	0.616065	0.672310	1.000000

Tab 3.

This justified the implementation of Linear Regression models that we used to compute the statistical significance and explanatory power of these numerical variables on loyalty points accumulation. To ensure model accuracy, we minimized the sum of squared residuals using the Ordinary Least Squares (OLS) method.

Graphic representation of linear relationships is displayed by scatterplots with regression line.

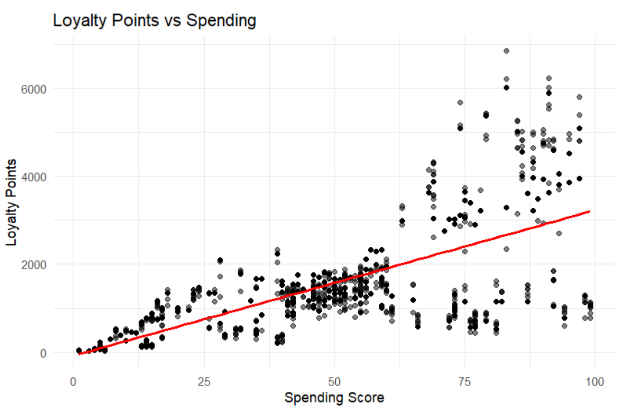


Fig 2. Regression: Spending Score vs Loyalty Points. [R^2 = 0.452, β = 33.06, p = < 0.005]

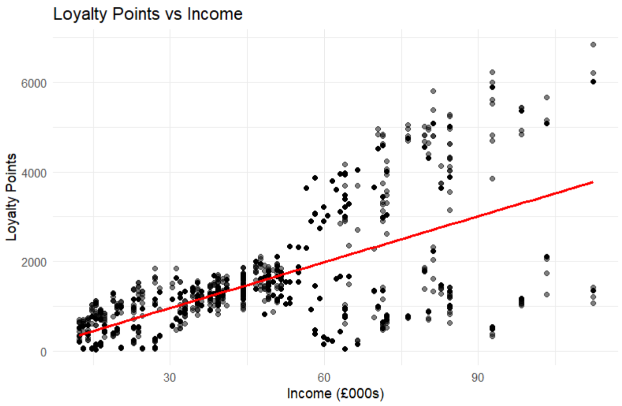


Fig 3. Regression: Income vs Loyalty Points. [R^2 = 0.380, β 34.187, p < 0.005]

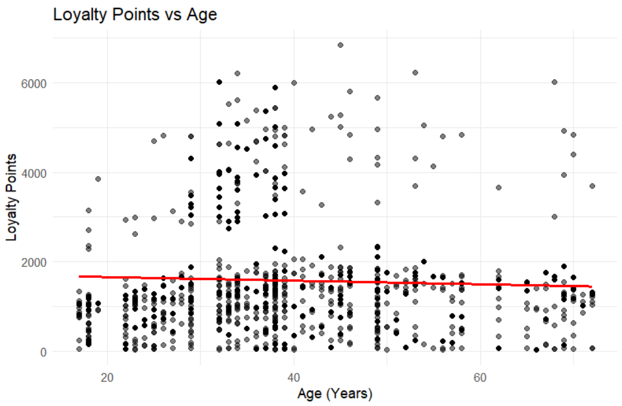
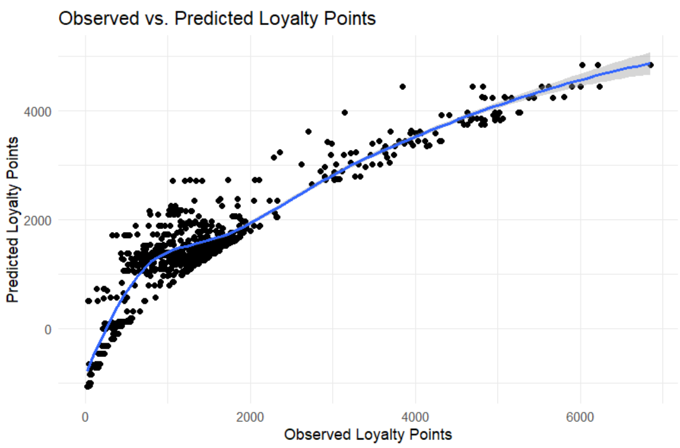


Fig 4. Regression: Age vs Loyalty Points. [R^2= 0.002, β 34.187, p < 0.005]. No linear relationship.

These variables alone did not sufficiently explain loyalty points differences. However, the model's fit notably improved when analysing subsets where both income and spending scores were below 60, indicating a stronger linear relationship within this subgroup. Appendix 2 summarizes OLS regressions.

To enhance accuracy, we used Multiple Linear Regression (MLR) with income and spending score as sole variables, prioritizing simplicity and variables that significantly explain loyalty points variance (Schneider et al., 2010). Together, these variables had a substantial impact on loyalty points accumulation, explaining 83% of the variation [Adj. R2 = 0.830, βx1 = 34.3346, βx2 = 32.6439, p < 0.005] (see appendix).

Despite a high R2, diagnostic tests highlighted heteroskedasticity and non-linear residual patterns (fig 5.), suggesting poor fit and the presence of unaccounted factors. This signalled issues with MLR.



(Fig. 5)

We used a square root transformation on the target variable to reduce outlier impact and address non-linearity, improving loyalty points' variation explained ($R^2 = 0.88$). However, Figure 6 indicates the model still struggles to fully grasp underlying data patterns, pointing to data limitations for precise predictions Model accuracy summary in Tab 4).

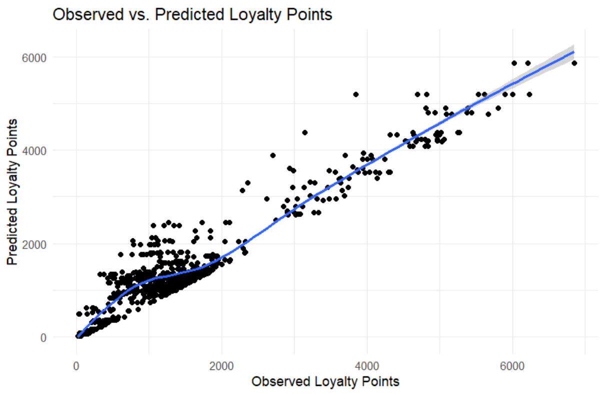


Fig 6: Observed vs Predicted Loyalty points after Squared Root Transformation.

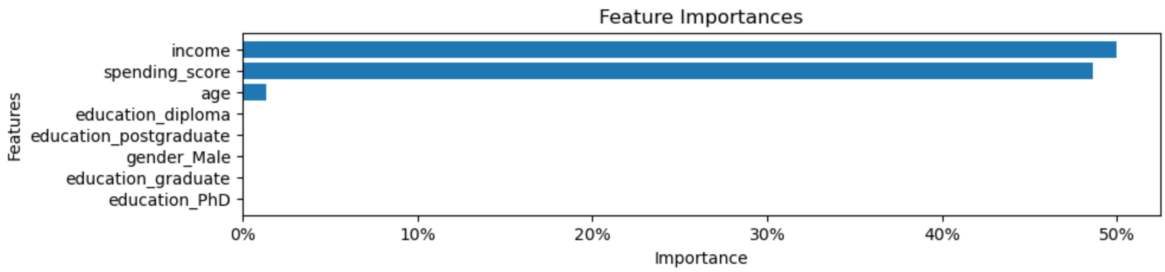
Metric	Value
R-squared	0.88
Mean Absolute Error (MAE)	1503.18
Mean Squared Error (MSE)	3562542.49
Root Mean Squared Error (RMSE)	1887.46

Tab. 4

2.2 Feature Importance

To better understand our data, we used a Decision Tree Regressor (DTR), which handles non-linear dependencies well. Employing K-fold cross-validation with 5 folds, we determined the optimal max_depth for pruning, enhancing model generalization and prediction accuracy on new data.

Feature importance analysis identified the variables with the most significant impact on loyalty points (Fig 6 below).



Age had a minimal contribution, indicating that targeting customers based on their income and spending behaviours is likely to result in different engagements and accumulation of loyalty points.

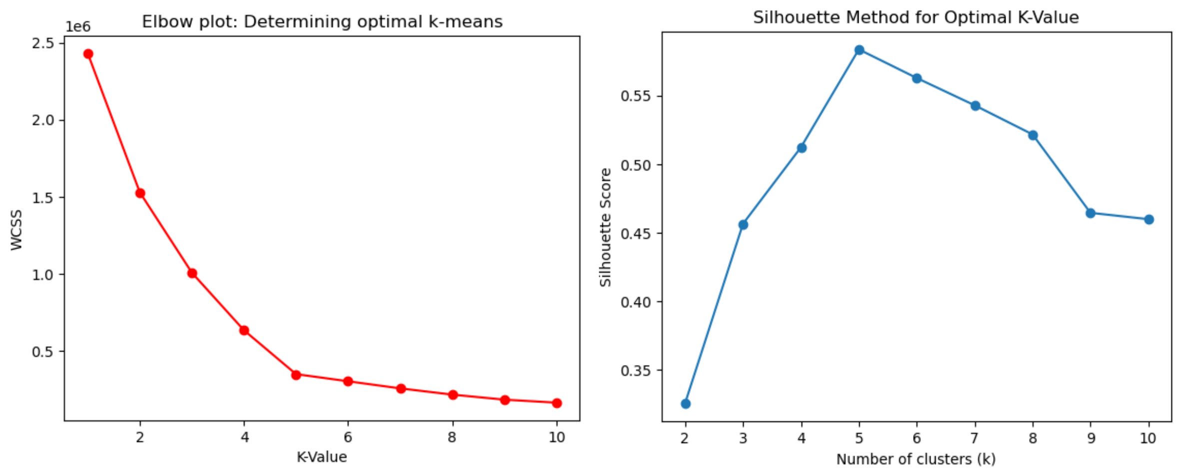
Next, we fine-tuned the model using the most important features and conducted cross-validation again. Table 5 (below) shows improved accuracy over MLR but still struggles with predicting extreme values.

Metric	Value
R-squared	0.97
Mean Absolute Error (MAE)	32.89
Mean Squared Error (MSE)	8120.35
Root Mean Squared Error (RMSE)	219.82
RMSE - MAE	186.94

Step 3. Customer Segmentation Through Clustering Algorithm

To find segments based on income and spending score, we applied unsupervised K-Means clustering. This ML algorithm was chosen for its simplicity and interpretability and implemented using the sklearn library (See Documentation[3]).

The optimal number of clusters (K value) was determined using the elbow and silhouette methods, by calculating the WCSS[4]and Silhouette Coefficient respectively for a more robust decision (Fig 4 below).



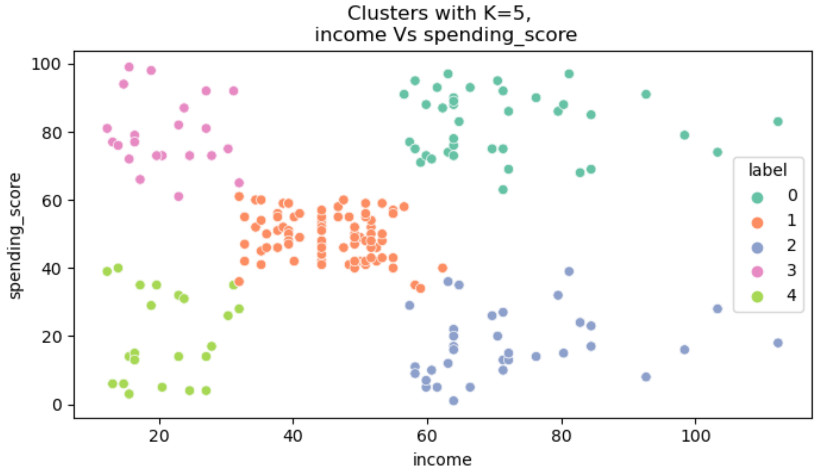
Both methods suggested K = 5 as the optimal number of clusters.

To address the manual decision caveat inherent in K-Means clustering, we evaluated the model clustering prediction with different K values (K=4, 5, 6) via scatterplots.

Fig 5. (below) displays the 5 clusters based on similar spending scores and income that best partition the data. This allowed us to identify the customer segments classified using conventional marketing naming (Tab 2. Full summary can be found in Appendix)

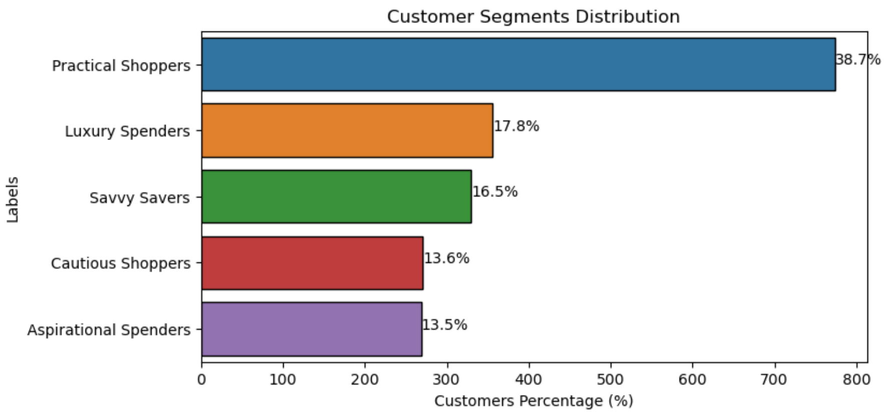
[3] https://scikit-learn.org/stable/modules/clustering.html

[4] Within-Cluster Sum of Squares. Also called Inertia.



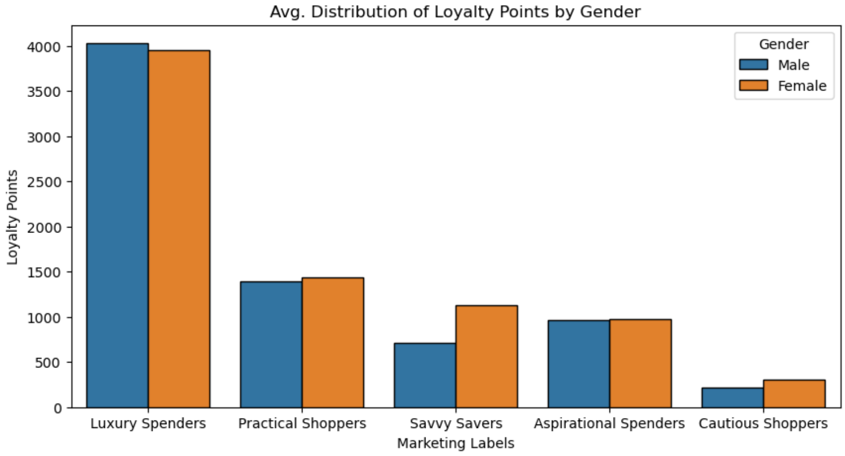
Customer Segments	Characteristics
Luxury Spenders	Affluent individuals with High Spendind Behaviours
Practical Shoppers	Moderate income with Moderate Spending
Savvy Saver	High Income with Low spending
Aspirational Shoppers	Low Income with high Spending
Cautious Shoppers	Low income individuals with Low Spending

Next steps involved analysing the segment distribution to assess the customer base, which revealed a concentration of practical shoppers and a high proportion of high earners with diverse spending behaviours, from conservative to lavish. Gender and education-based engagement with the loyalty point system was analysed to refine customer profiling and enhance marketing strategy. These comparisons were visualised with different bar plots (fig 6 and 7).



Fig(6) .

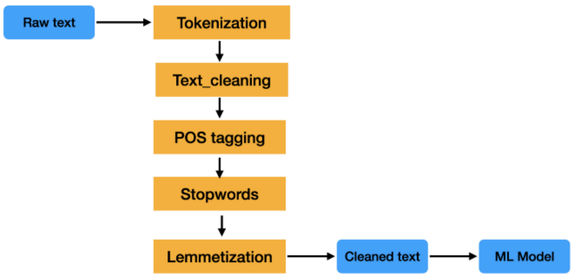
Females were shown with overall avg. higher loyalty points than male customers which need further investigation. Customer segments by Education did not show specific patterns in loyalty engagements.



Fig(7)

Step 4. Sentiment Analysis toward Product Purchased.

The review features[5] underwent pre- processing tasks (Fig 8) which prepared the text data for NLP (Natural Language Processing).



Fig(8)

We analysed sentiment testing VADER[6] and TextBlob Python libraries[7], recognizing that model performance can vary based on dataset characteristics and inherent properties of each model.

Our focus was on the 'Reviews' variable to detect sentiment comprehensively. We calculate Vader compound score[8] and TextBlob Polarity metrics for intensity of sentiments, categorizing them into sentiment class. This allowed to conduct comparative analysis though bar chart (fig 8.)

[5] ‘review’ and ‘summary_review’ columns.
[6]Valence Aware Dictionary and sEntiment Reasoner : Lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and other text domains. See documentation https://vadersentiment.readthedocs.io/en/latest/
[7] Also Lexicon Based, identify different entities based on its entities library, en-entities.txt and tag phrases by Parts of Speech (POS). (see documentation: https://textblob.readthedocs.io/en/dev/).
[8] **VADER compound Score**: weighted composite score that summarizes the overall sentiment of a text. It ranges from -1 to 1, where -1 indicates the most negative sentiment possible, 1 the most positive sentiment possible, 0 a neutral sentiment.
Textblob Polaraty Score: float that lies between [-1,1], -1 indicates negative sentiment and +1 indicates positive sentiments.

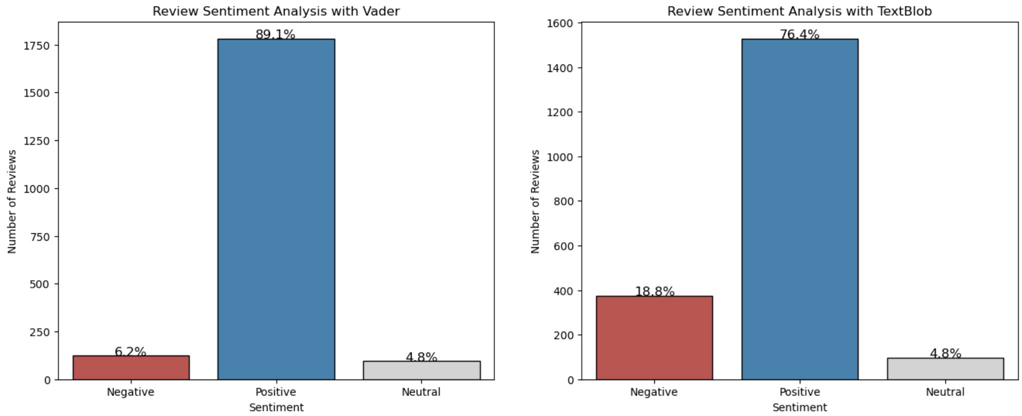
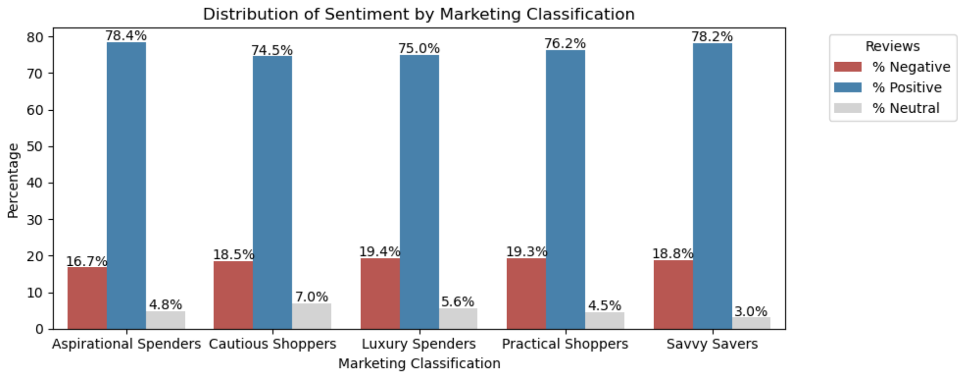


Fig 8.

To assess model accuracy, we took a balanced approach, evaluating overall performance with a focus on negative reviews through oversampling. Manual testing was conducted on three separate data frames[9]. In this dataset, lengthy reviews (see appendix 9) affected VADER's accuracy, while TextBlob exhibited high false negatives, notably misclassifying true positive reviews.

Due to the higher risk associated with negative reviews for Turtle Games, we opted for TextBlob despite its drawbacks. These outputs were used to calculate sentiment across marketing segments and visualized using grouped chart (Fig 9).Findings should be interpreted cautiously but provide a foundation for automating sentiment analysis.



Finally, we proposed a method to pinpoint top positive and negative product reviews and flag products with the most negative feedback. Using spaCy's Matcher class, we automated adjective extraction from negative reviews (e.g., product 2795, Tab below), identifying common negative descriptors per product to aid in product development and customer satisfaction strategies.

---

[9] For general accuracy, summary of outputs can be found in ‘sample_test1.xlsx’. ‘sample_test_neg_review_vader.xlsx’: including 10 sample negatives reviews classified by Vader vs Textblob .
'sample_test_neg_review_blob.xlsx': including 10 sampled negative reviews classified by Texblob Vs Vader.

	product	adjective	count
288	2795	different	2
289	2795	many	2
290	2795	young	1
291	2795	best	1
292	2795	difficult	1
293	2795	helpful	1
294	2795	much	1
295	2795	old	1
296	2795	open	1
297	2795	poor	1

Insight and Key Recommendations

- **Loyalty Points Variation**: Significant variation suggests some TG customers aren't effectively engaging. Tailored strategies are needed to balance engagement and inform program adjustments

- **Influencing Factors**: Differences in loyalty points accumulation are primarily influenced by yearly income and spending behaviors, explaining 83% of the variation: A £1000 income increase adds 34.33 points, and a one-level spending score rise adds 32.64 points.

- **Non-Linear Engagement**: Loyalty points do not always increase proportionally with higher income and spending scores. We recommend developing strategies based on different income and spending levels to maximize customer retention and sales. Implementing a tiered loyalty program could potentially address these variations more effectively by providing tailored rewards and incentives, encouraging higher engagement across different customer segments.

- **Customer Clustering**: Use clustering to guide targeted marketing efforts and personalized strategies (Detailed cluster characteristics in Appendix). Efforts should focus on Luxury Spenders, Practical Shoppers, and Savvy Savers. Females consistently show higher engagement with loyalty points compared to males. Further analysis should always investigate whether this demographic is statistically associated with higher engagement with the loyalty system and refine the marketing strategy accordingly.

- **Predictive Accuracy**: Middle data predictions were moderately accurate (MAE = 32.885). For outliers, consider: A) Incorporating sales and customer engagement data through different feature engineering. B) Using models like SVMs that handle non-linear relationships better than DTR.

- **Sentiment Analysis:** Turtle Games products generally received positive sentiment, reflecting strong brand reputation and customer satisfaction. Based on accuracy test suggests negative reviews may realistically be around 10%. The marketing team should leverage these insights to promote positive feedback and address issues for strategic and product development guidance.

To improve model accuracy, we recommend:

- Using VADER for short texts and TextBlob for longer reviews.
- Employing advanced pretrained BERT models through PyTorch for contextual understanding.

Ongoing opinion mining is crucial, especially monitoring luxury spenders with higher negative reviews.

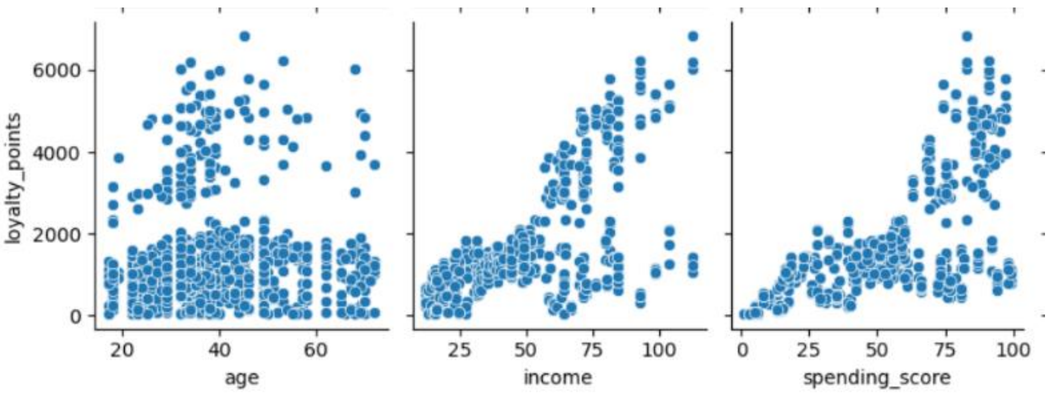
Appendix

1. Duplicate check and classification.

Although we found identical reviews, each row corresponded to product reviews left by different customers. In instances where reviews were from the same customer, they were related to different products, as such these entries were not treated as real duplicates.

	gender	age	income	spending_score	loyalty_points	education	product	review	summary
48	Female	29	32.80	42	842	graduate	2079	love it	Five Stars
55	Male	45	35.26	41	1062	graduate	3896	Great!	Five Stars
94	Female	34	49.20	42	1376	graduate	6721	great	Five Stars
294	Female	34	49.20	42	1376	graduate	6770	Good	Five Stars
326	Male	41	58.22	35	1463	graduate	2849	love it	Five Stars
371	Male	32	71.34	75	3455	diploma	5726	Great!	Five Stars
408	Male	66	15.58	3	31	PhD	1459	great	Five Stars
416	Female	37	17.22	35	417	graduate	830	love it	Five Stars

2. Income, Spending Score, Age vs Loyalty Points Scatterplots



3. OLS Regression Table: Spending Score vs Loyalty Points

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:         loyalty_points   R-squared:                       0.452
Model:                            OLS   Adj. R-squared:                  0.452
Method:                 Least Squares   F-statistic:                     1648.
Date:                Tue, 02 Jul 2024   Prob (F-statistic):          2.92e-263
Time:                        22:59:36   Log-Likelihood:                -16550.
No. Observations:                2000   AIC:                         3.310e+04
Df Residuals:                    1998   BIC:                         3.312e+04
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const            -75.0527     45.931     -1.634      0.102    -165.129      15.024
spending_score    33.0617      0.814     40.595      0.000      31.464      34.659
==============================================================================
Omnibus:                      126.554   Durbin-Watson:                   1.191
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              260.528
Skew:                           0.422   Prob(JB):                     2.67e-57
Kurtosis:                       4.554   Cond. No.                         122.
==============================================================================
```

4. OLS Regression Table: Income vs Loyalty Points

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:         loyalty_points   R-squared:                       0.380
Model:                            OLS   Adj. R-squared:                  0.379
Method:                 Least Squares   F-statistic:                     1222.
Date:                Tue, 02 Jul 2024   Prob (F-statistic):          2.43e-209
Time:                        22:59:38   Log-Likelihood:                -16674.
No. Observations:                2000   AIC:                         3.335e+04
Df Residuals:                    1998   BIC:                         3.336e+04
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -65.6865     52.171     -1.259      0.208    -168.001      36.628
income        34.1878      0.978     34.960      0.000      32.270      36.106
==============================================================================
Omnibus:                       21.285   Durbin-Watson:                   3.622
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               31.715
Skew:                           0.089   Prob(JB):                     1.30e-07
Kurtosis:                       3.590   Cond. No.                         123.
==============================================================================

Notes:
```

5. OLS Regression Table: Age vs Loyalty Points

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:         loyalty_points   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     3.606
Date:                Wed, 03 Jul 2024   Prob (F-statistic):             0.0577
Time:                        07:51:04   Log-Likelihood:                -17150.
No. Observations:                2000   AIC:                         3.430e+04
Df Residuals:                    1998   BIC:                         3.431e+04
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1736.5177     88.249     19.678      0.000    1563.449    1909.587
age           -4.0128      2.113     -1.899      0.058      -8.157       0.131
==============================================================================
Omnibus:                      481.477   Durbin-Watson:                   2.277
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              937.734
Skew:                           1.449   Prob(JB):                    2.36e-204
Kurtosis:                       4.688   Cond. No.                         129.
==============================================================================
```

6. OLS Regression Table: Spending Score < 60 vs Loyalty Points

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         loyalty_points   R-squared:                       0.601
Model:                            OLS   Adj. R-squared:                  0.601
Method:                 Least Squares   F-statistic:                     2055.
Date:                Tue, 02 Jul 2024   Prob (F-statistic):          1.74e-274
Time:                        22:59:37   Log-Likelihood:                -10015.
No. Observations:                1367   AIC:                         2.003e+04
Df Residuals:                    1365   BIC:                         2.005e+04
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const            157.8290     22.483      7.020      0.000     113.725     201.933
spending_score    25.5150      0.563     45.327      0.000      24.411      26.619
==============================================================================
Omnibus:                       29.904   Durbin-Watson:                   0.918
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               32.134
Skew:                           0.335   Prob(JB):                     1.05e-07
Kurtosis:                       3.340   Cond. No.                         90.2
==============================================================================
```

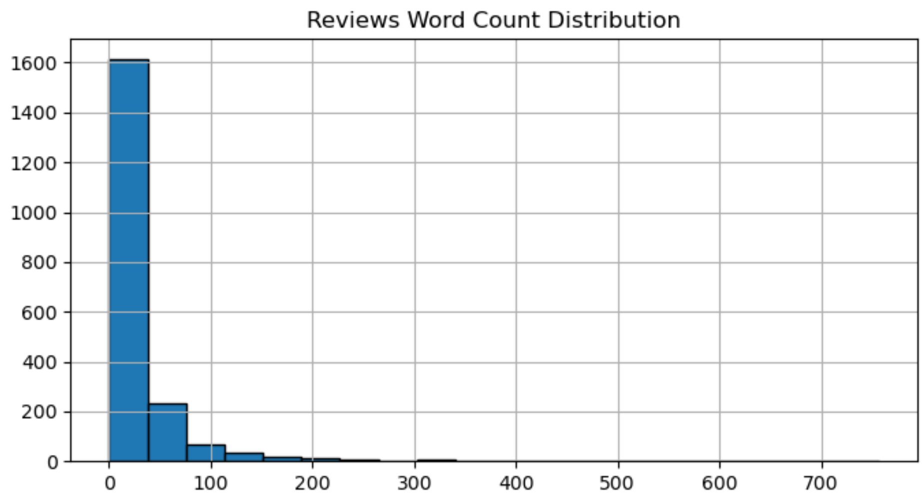
7. OLS Regression Table: Income < 60 vs Loyalty Points

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         loyalty_points   R-squared:                       0.499
Model:                            OLS   Adj. R-squared:                  0.498
Method:                 Least Squares   F-statistic:                     1389.
Date:                Tue, 02 Jul 2024   Prob (F-statistic):          1.36e-211
Time:                        22:59:39   Log-Likelihood:                -10615.
No. Observations:                1398   AIC:                         2.123e+04
Df Residuals:                    1396   BIC:                         2.124e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -50.2057     34.849     -1.441      0.150    -118.568      18.156
income        33.5671      0.901     37.268      0.000      31.800      35.334
==============================================================================
Omnibus:                      118.345   Durbin-Watson:                   3.272
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              686.547
Skew:                           0.063   Prob(JB):                    8.28e-150
Kurtosis:                       6.431   Cond. No.                         105.
==============================================================================
```

8. OLS Multi Linear Regression Table

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         loyalty_points   R-squared:                       0.830
Model:                            OLS   Adj. R-squared:                  0.830
Method:                 Least Squares   F-statistic:                     3895.
Date:                Sun, 07 Jul 2024   Prob (F-statistic):               0.00
Time:                        08:04:32   Log-Likelihood:                -12307.
No. Observations:                1600   AIC:                         2.462e+04
Df Residuals:                    1597   BIC:                         2.464e+04
Df Model:                           2
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const          -1700.3237     39.588    -42.950      0.000   -1777.974   -1622.674
income            34.3346      0.574     59.838      0.000      33.209      35.460
spending_score    32.6439      0.510     63.947      0.000      31.643      33.645
==============================================================================
Omnibus:                        2.977   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.226   Jarque-Bera (JB):                2.923
Skew:                           0.075   Prob(JB):                        0.232
Kurtosis:                       3.147   Cond. No.                         220.
==============================================================================
```

9. Reviews words count distribution



10. Customers Segmentation Frameworks

Customer Type	Characteristics	Opportunities	Recommended Strategies
Luxury Spenders	Affluent individuals, high spending on premium items	High transaction values, opportunities for high-margin sales, brand ambassadors through exclusive experiences.	Consider tiered membership levels with escalating rewards based on spending thresholds. Leverage psychology of benefits
Practical Shoppers	Middle-income, moderate but consistent spending	Steady revenue stream, broad customer base appeal, potential for loyalty through value-for-money offerings	Emphasize value-for-money and quality. Provide rewards for consistent purchases, bundle deals on frequently bought items. Consider tiered rewards based on cumulative spending over time, with incentives like cashback
Savvy Savers	High disposable income, low current spending	Potential for increased future spending, long-term loyalty through value and quality emphasis, financially stable customer base	Promote value-for-money deals, provide information on how savings can accumulate to future benefits, and introduce long-term saving schemes
Practical Shoppers	Low-income individuals who prioritize spending	Receptive to upselling, present market growth opportunities as their income grows	Create loyalty programs with attainable milestones, and emphasize quality and brand value through targeted marketing campaigns
Cautious Shoppers	Budget-conscious, low-risk spending behavior	Stable, though low, revenue stream, loyal customer base if value is consistently delivered.	Emphasize budget-friendly options, offer clear and tangible benefits for loyalty (es. discount on necessities) and provide education on how to maximize value from purchases

Reference

Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010 Nov;107(44):776-82. doi: 10.3238/arztebl.2010.0776. Epub 2010 Nov 5. PMID: 21116397; PMCID: PMC2992018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992018/

Unveiling Decision Tree Regression: Exploring its Principles, Implementation, Medium, 2023. https://medium.com/@vk.viswa/unveiling-decision-tree-regression-exploring-its-principles-implementation-beb882d756c6