Turtle Games: Predicting Future Outcomes

By: Iulia-Diana Cristolovean

Last Updated: 12.03.2025

Turtle Games is a game manufacturer and retailer offering a wide range of products, including books, board games, and video games.

This report focuses on:

- Analysing customer interactions with the loyalty program.
- Segmenting customers for targeted marketing.
- Leveraging online reviews for business growth.
- Evaluating the loyalty points system for predictive modelling.

This report provides insights to help marketing and sales make data-driven decisions, boosting customer retention, engagement, satisfaction, and marketing strategies.

This analysis used two files: *turtle_reviews.csv* and *metadata_turtle_games.txt*. Python was primarily used with the main libraries - pandas (for data manipulation and analysis), scikit-learn (for machine learning models and preprocessing) and nltk (for natural language processing tasks). R was also integrated for alignment with TurtleGame's system.

Analytical approach

Data Wrangling

Data validation was performed both in Python and R to ensure it's correctitude. The data was checked for duplicates, null values and some data manipulation, including:

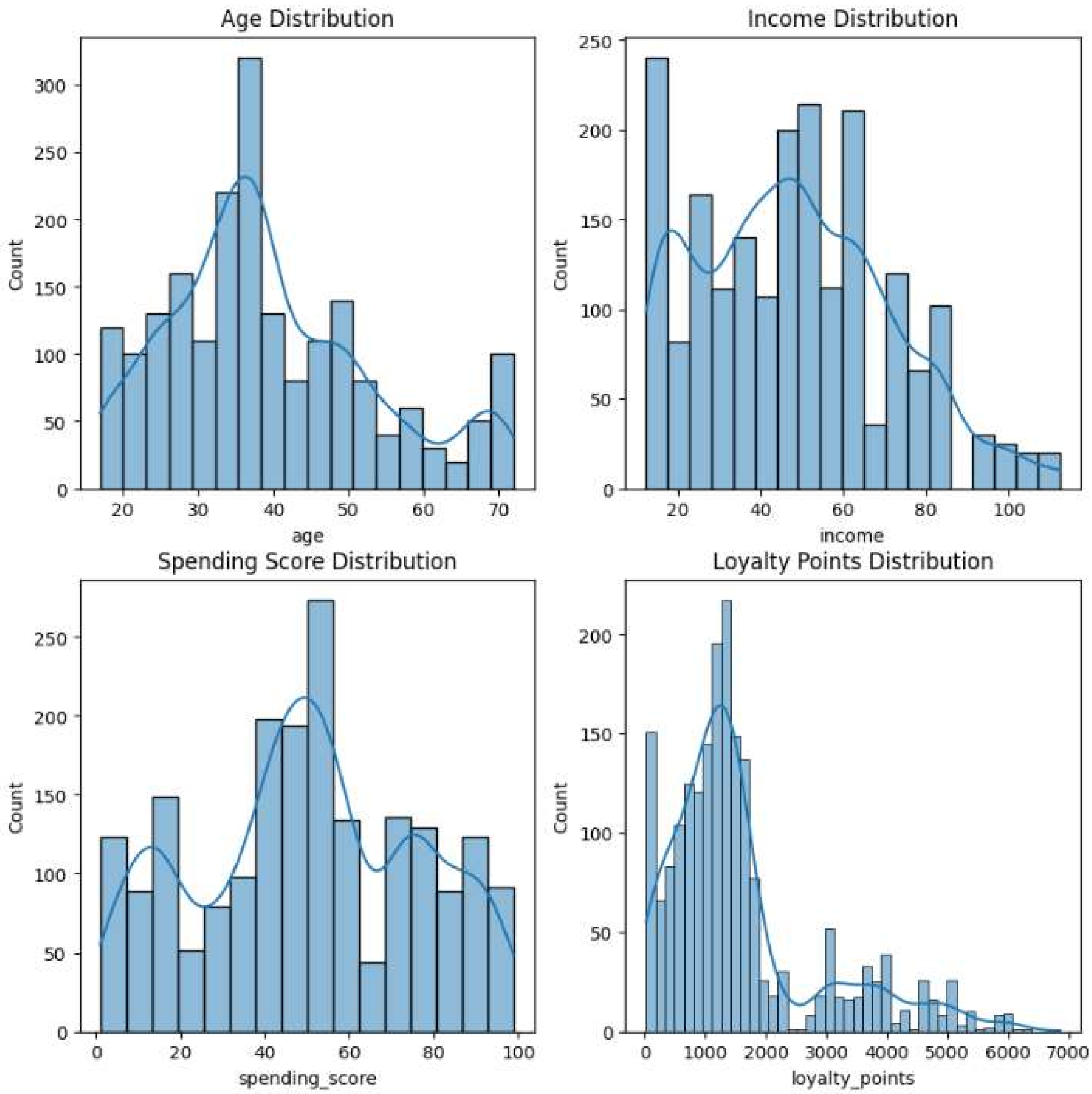
- Dropping unnecessary columns ("language", "platform")
- Renaming columns for ease of use (e.g. "renumeration" to "income")

Data Analysis and Visualization

1. Data analysis

Numerical columns were analysed with histograms and box plots, revealing no clear normal distribution. Only loyalty points had unexpected outliers, which was surprising because income typically does (*see Appendix Fig 1*). These outliers were left unchanged and flagged for clarification with TurtleGames.

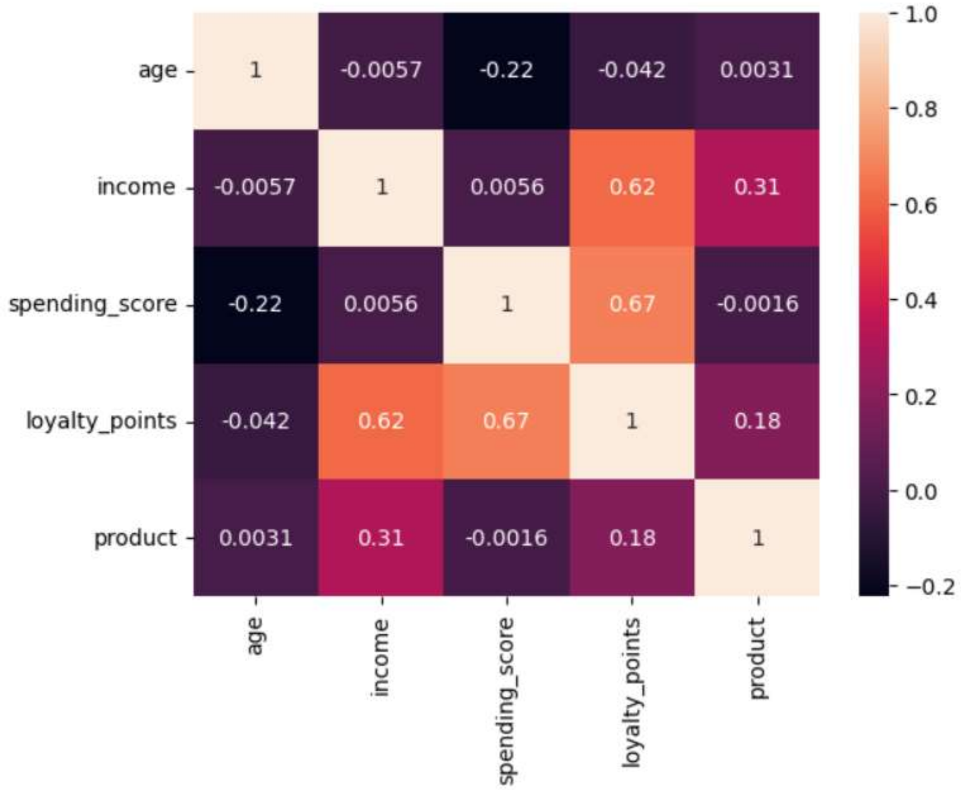
The cleaned Data Frame was exported as *reviews.csv* for future use.



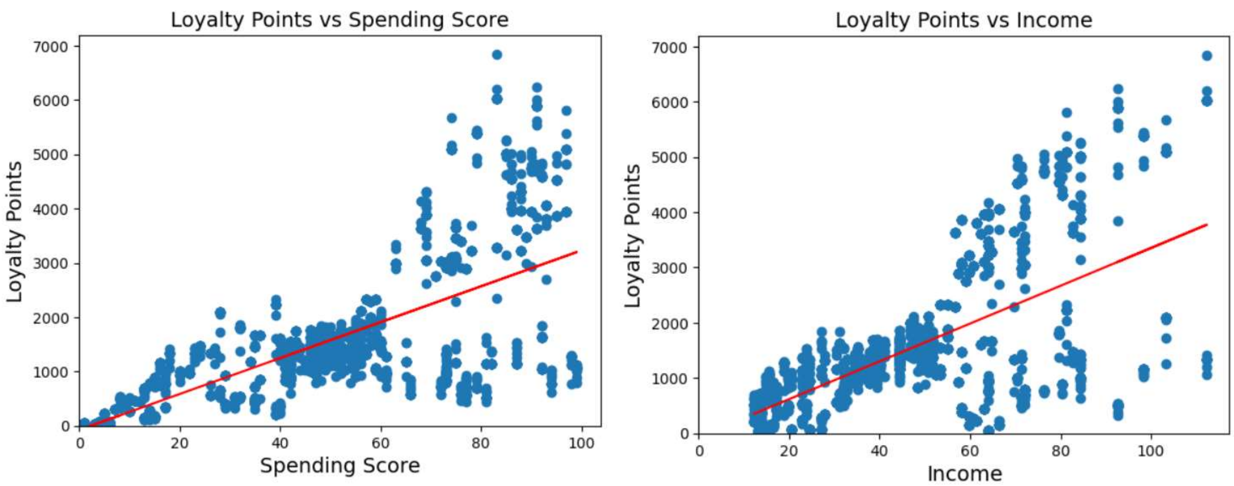
2. Factors influencing loyalty points.

- We proceeded by generating a correlation matrix on the updated DataFrame to observe the following (for the scatterplots *see Appendix Fig 2*):

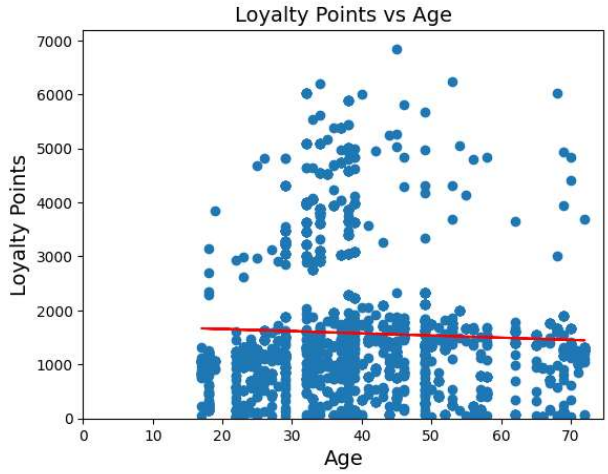
  ✓ Income shows a moderate correlation with loyalty points (0.62).
  ✓ Spending score exhibits a slightly stronger correlation with loyalty points (0.67).
  ✓ Income has a lower correlation with spending score, which is a positive insight for marketing (indicates that sales are not driven by high-income).
  ✓ Age shows no linear relationship with loyalty points,
  however, Turtle Games team would like to explore this variable further.



- To compute the statistical significance and explanatory power of these numerical variables on loyalty points accumulation, we used Simple Linear Regression Models. To ensure model accuracy, and to minimize the sum of squared residuals we decided to use the Ordinary Least Squares (OLS) method.

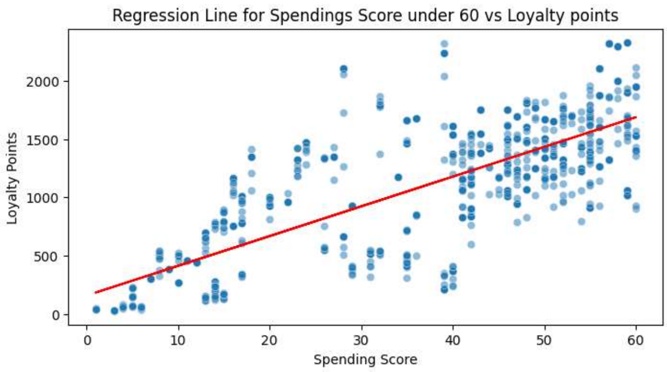


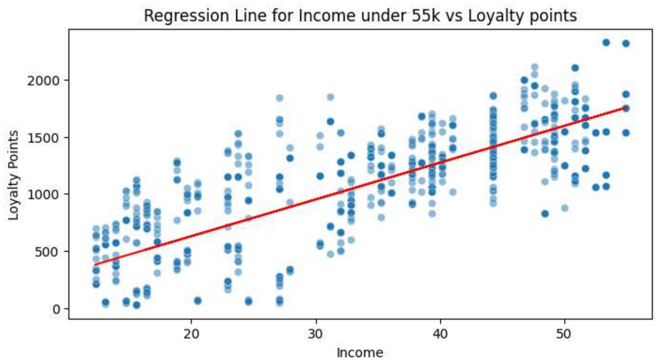
- The scatterplot has a cone-like shape which is a sign of heteroscedasticity. We therefore reject the null hypothesis (which assumes homoscedasticity) and conclude that the data are heteroscedastic.



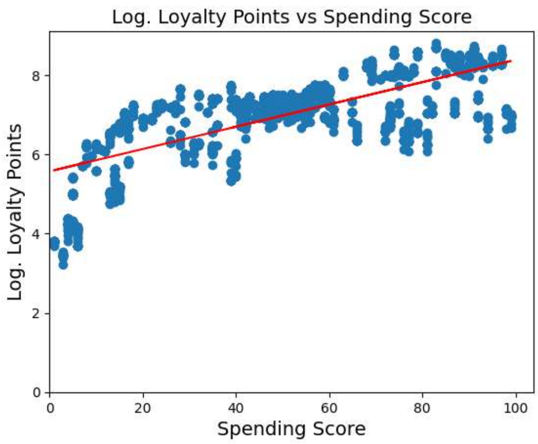
- ✓ *Spending Score vs Loyalty Points. [R2 = 0.452, β = 33.06, p = < 0.005]*
- ✓ *Income vs Loyalty Points. [R2 = 0.380, β 34.187, p < 0.005]*
- ✓ *Age vs Loyalty Points - No linear relationship.*

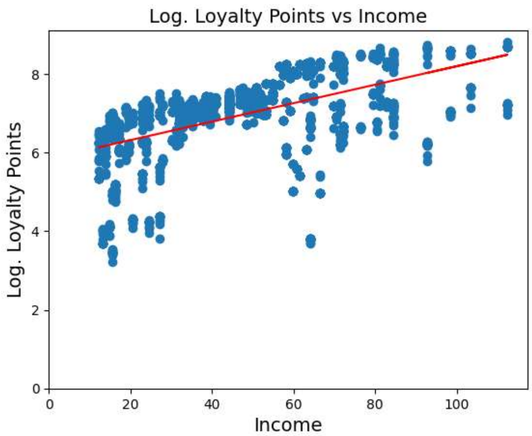
- OLS models showed weak fits, and the "cone" shape suggested heteroscedasticity (confirmed by the Breusch-Pagan test). Limiting Income to 55K and Spending Score to 60 improved R^2 values to 0.62 and 0.60 and got rid of the heteroskedasticity (*see Appendix Fig 3*).





- To reduce heteroscedasticity, we need a transformation (e.g. logarithm) as residual variance increases with the predictor (*see Appendix Fig 4*). Scatterplots with regression lines show the relationship between spending score, income, and log loyalty points.





By transforming loyalty using its natural logarithm (*see Appendix Fig 5*) we managed to:

- ✓ Reduced heteroscedasticity (used the Breusch-Pagan test).
- ✓ Obtain a more consistent variance.
- ✓ Better fit overall, with some nonlinearity and outliers remaining.

Simple Linear Regression				
	Spending Score vs Loyalty Points	Spending Score vs Loyalty Points (log)	Income vs Loyalty Points	Income vs Loyalty Points (log)
R^2	0.45	0.52	0.38	0.28
Adjusted R^2	0.45	0.52	0.379	0.28
Intercept	-75.05	5.57	-65.68	5.85
Independent Variable x	33.06	0.028	34.18	0.0235
p	0	0	0	0

- To improve accuracy, we used Multiple Linear Regression (MLR) with income and spending score, which explained 83% of the variation in loyalty points.

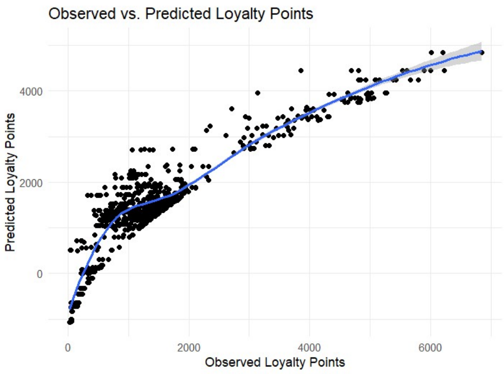
```
                            OLS Regression Results
==============================================================================
Dep. Variable:         loyalty_points   R-squared:                       0.830
Model:                            OLS   Adj. R-squared:                  0.830
Method:                 Least Squares   F-statistic:                     3895.
Date:                Mon, 31 Mar 2025   Prob (F-statistic):               0.00
Time:                        12:26:51   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
spending_score    32.6439      0.510     63.947      0.000      31.643      33.645
income            34.3346      0.574     59.838      0.000      33.209      35.460
==============================================================================
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.
==============================================================================
```

- Despite the high R^2, diagnostic tests revealed heteroskedasticity and non-linear residual patterns, which could impact standard errors and confidence intervals, questioning the reliability of MLR. A fix could be log-transforming the dependent variable (loyalty points).

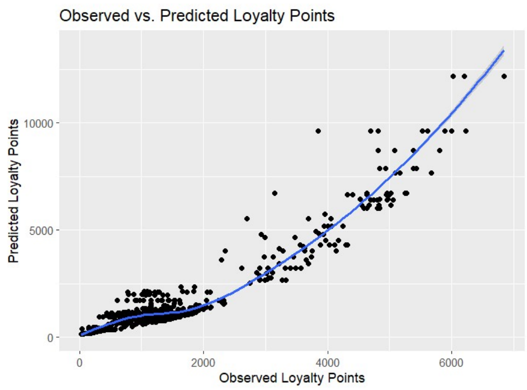
```
                            OLS Regression Results
==============================================================================
Dep. Variable:      log_loyalty_points   R-squared:                       0.795
Model:                             OLS   Adj. R-squared:                  0.795
Method:                  Least Squares   F-statistic:                     3102.
Date:                 Mon, 31 Mar 2025   Prob (F-statistic):               0.00
Time:                         12:26:51   Log-Likelihood:                -1048.6
No. Observations:                 1600   AIC:                             2103.
Df Residuals:                     1597   BIC:                             2119.
Df Model:                            2
Covariance Type:             nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              4.4423      0.035    127.649      0.000       4.374       4.511
spending_score     0.0283      0.000     63.072      0.000       0.027       0.029
income             0.0233      0.001     46.206      0.000       0.022       0.024
==============================================================================
Omnibus:                      485.405   Durbin-Watson:                   2.017
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1548.844
Skew:                          -1.511   Prob(JB):                         0.00
Kurtosis:                       6.756   Cond. No.                         220.
==============================================================================
```

- The original model has a slightly better fit, but the log-transformed model has lower absolute and squared errors, indicating that *it's better at handling outliers and producing more consistent predictions*.

Multiple Linear Regression		
	Loyalty Points	Log Loyalti Points
R^2	0.83	0.795
Adjusted R^2	0.83	0.795
Mean Squared Error (MSE)	300944.1	0.18
Mean Absolute Error (MAE)	429.66	0.34

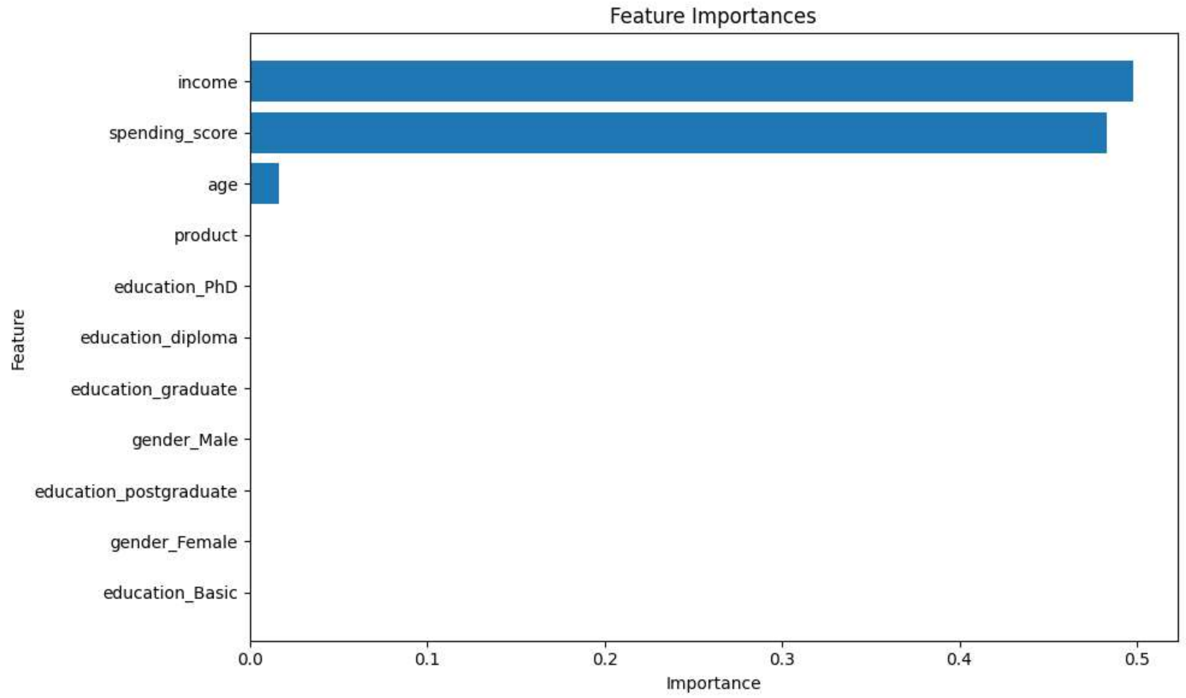


Observed vs Predicted Loyalty points before Log Transformation.



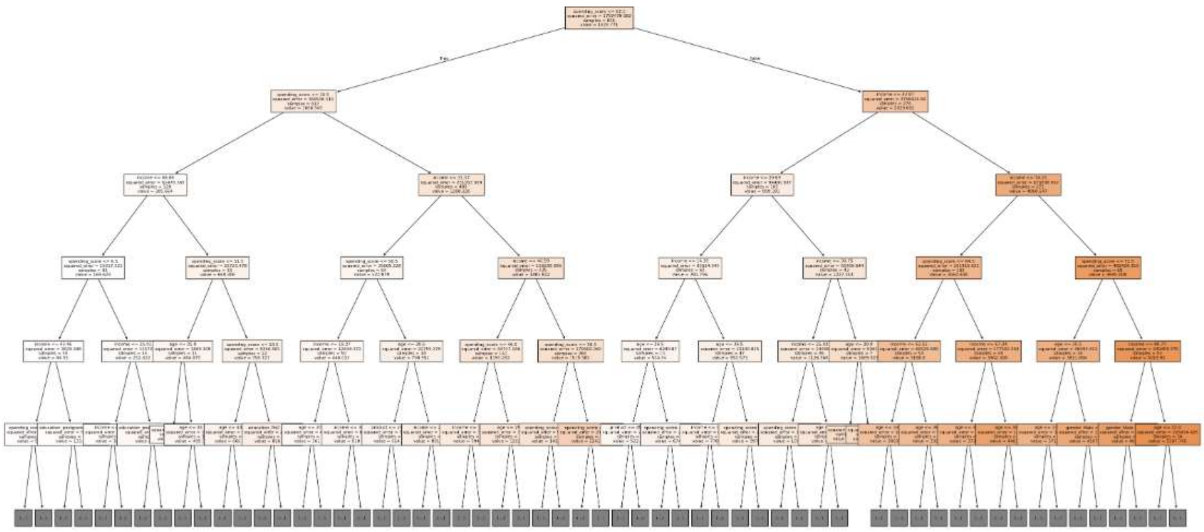
Observed vs Predicted Loyalty points after Log Transformation.

- To better understand which factor contributes to the loyalty points more, we used a Decision Tree Regressor because it manages non-linear relations well. After pruning it to the optimal depth of ten folds, we determined the variables with the most contribution (*see Appendix Fig 6*).



✓ Pruned Model Mean Squared Error: 15437.31421749669
✓ Pruned Model Mean Absolute Error: 64.48128833461625
✓ Pruned Model R-squared: 0.9904695991414618
✓ Root Mean Squared Error: 124.24698876631453

- Because DTR are greedy and prone to overfitting, I decided to test a Random Forest.



- ✓ Random Forest Model Mean Squared Error: 7421.831176681972
- ✓ Random Forest Mean Absolute Error: 41.65937560627396
- ✓ Random Forest Model R-squared: 0.9950791954139152
- ✓ Root Mean Squared Error: 86.150050357977

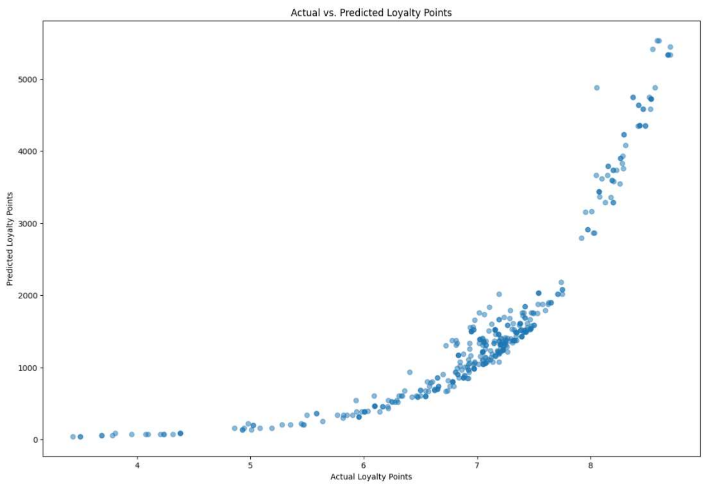
- Random Forest performed better than the pruned Decision Tree Regressor — it reduced MSE, MAE, and RMSE while slightly improving R^2. The model is now more stable and precise, with better generalization potential.

- *Key Insights on Predictors of Loyalty Points:*

    - Spending Score is the strongest predictor, appearing frequently as the root node in the decision tree.
    - Income is a secondary factor, typically appearing second in the tree.
    - Age has a modest influence, appearing in deeper nodes with a less direct impact.

- Because it works well for non-linear relationships, and it's less sensitive to outliers, I decided to try the SVR model (*see Appendix Fig. 7*). I got better results after I used the log-transformed loyalty points.

    - ✓ Mean Squared Error: 0.0222067648999956277
    - ✓ R-squared: 0.9768862100002215 (very strong predictive performance)

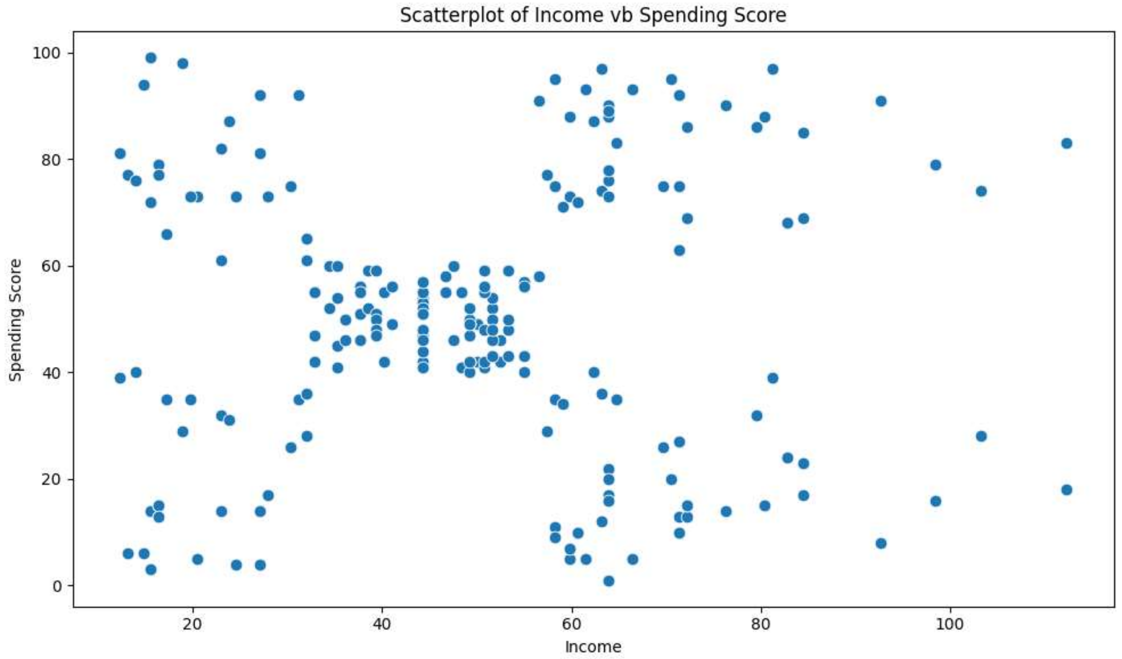


- **Random Forest** — it's clearly the best-performing model in our lineup, both in terms of predictive power (R^2) and low error (MSE). If interpretability is more important than performance, the **pruned Decision Tree** or the **Linear Regression** models might still be worth considering.

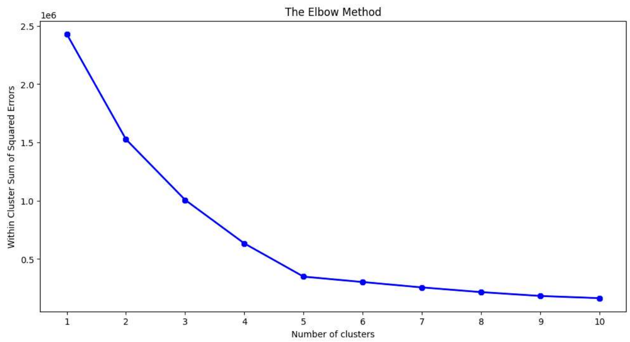
Model	MSE	R^2
Multiple Linear Regression	300,944.09	0.83
MLR with log transformation	0.178	0.795
Random Forest	7,421.83	0.995
Decision Tree Regressor (pruned)	15,437.31	0.99
SVR with log transformation	1,441,148.25	0.111

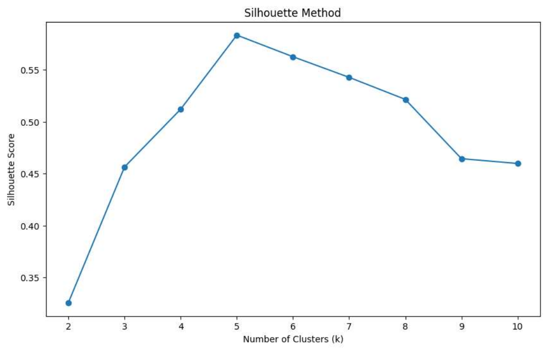
3. Customer segmentation through clustering.

- K-Means Clustering was applied to group customers based on Income and Spending Score. Before running the model, a scatterplot was generated to check for any visually identifiable clusters.

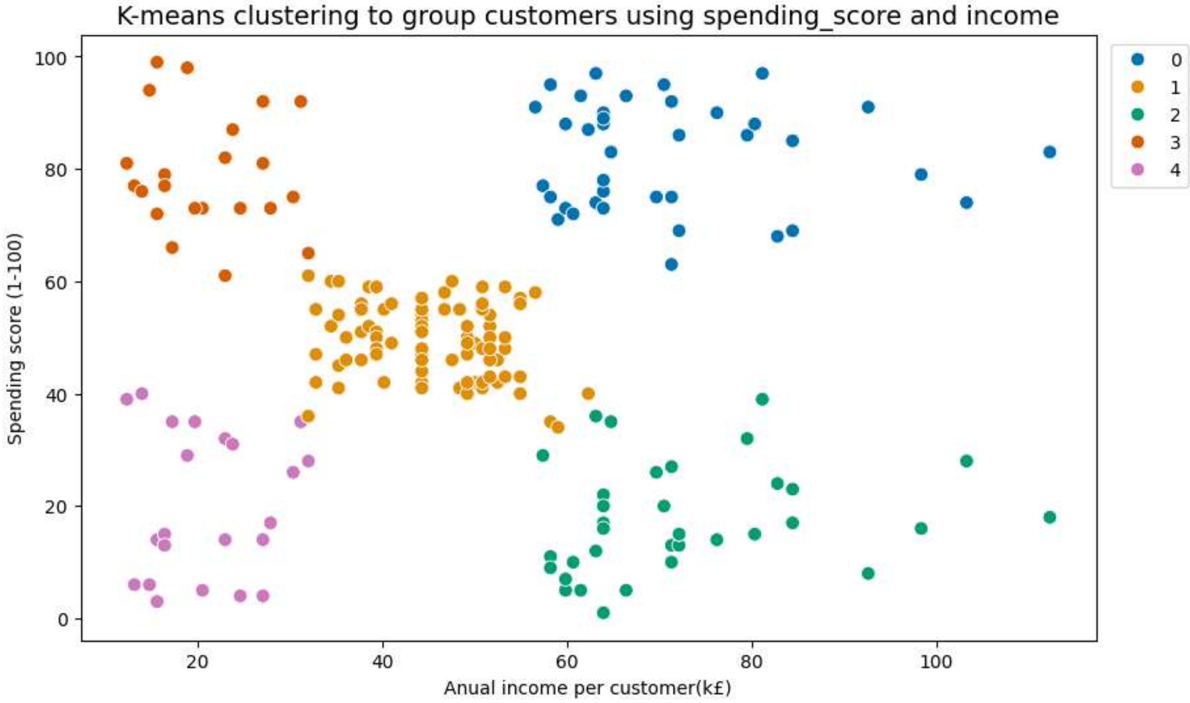


- Initial visualization suggested five clusters, confirmed by the Elbow and Silhouette plots. (we tested K=4, 5, and 6 using scatterplots, *see Appendix Fig 8*)



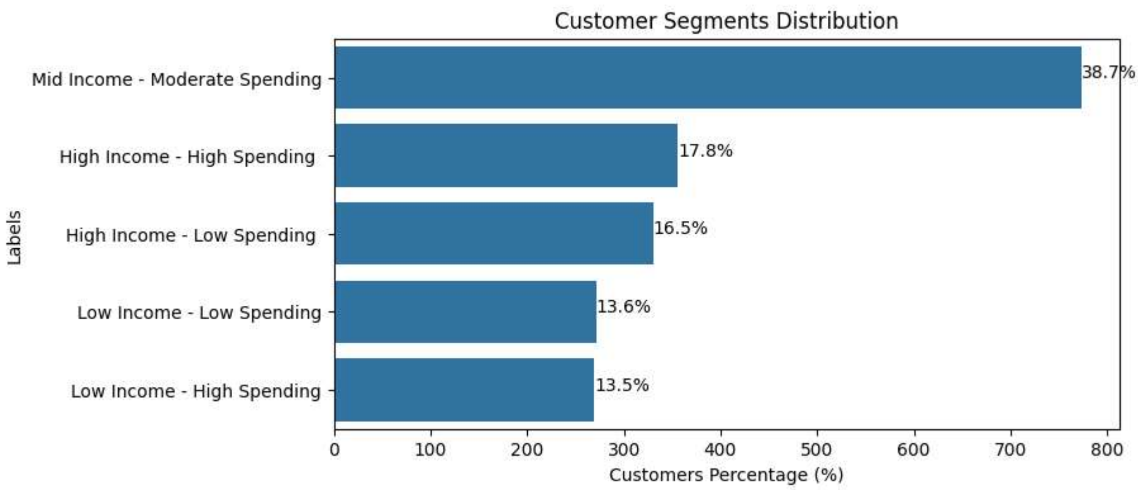


- These clusters that fell into the five categories with similar characteristics are:



0	High Income - High Spending (Blue)
1	Mid Income - Moderate Spending (Orange)
2	High Income - Low Spending (Green)
3	Low Income - High Spending (Red)
4	Low Income - Low Spending (Purple)

- The next step was to analyse segment distribution to better understand the customer base. This revealed a concentration of Mid Income - Moderate Spending, along with a considerable number of high earners.



- *See Appendix Fig.9* for gender and education interactions with loyalty points.

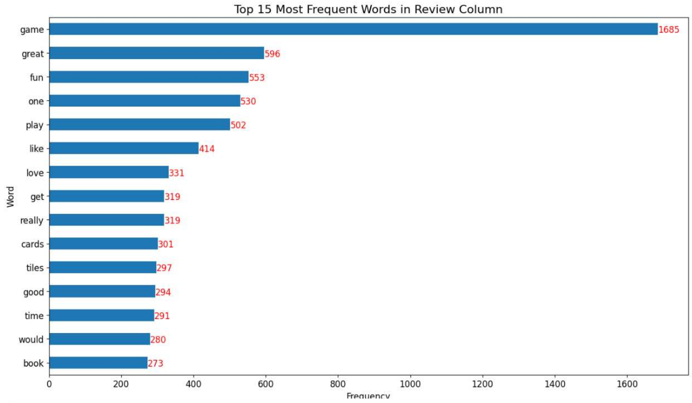
4. **Sentiment Analysis toward Products.**

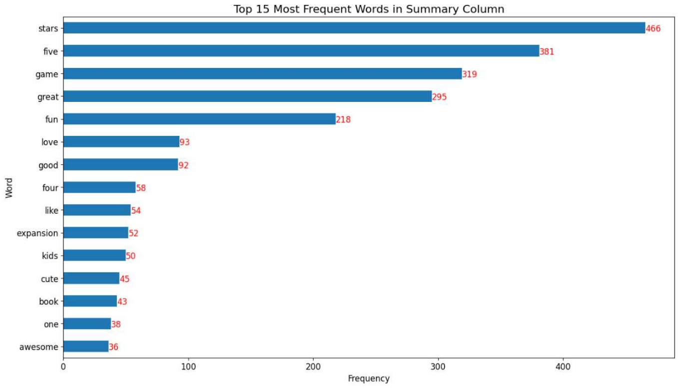
- To consider language analysis we used Natural Language Processing libraries such as TextBlob and Vader to get a balanced view of sentiment analysis. We began with the original clean DataFrame, with 2000 entries (I have decided to keep the duplicates, as they are generic and could have been written by different customers). We will focus mainly on the *review* and *summary* columns, dropping the rest.

- WordClouds were produced to visualise common words.

Word Cloud for ***review*** colmn

Word Cloud for ***summary*** colmn



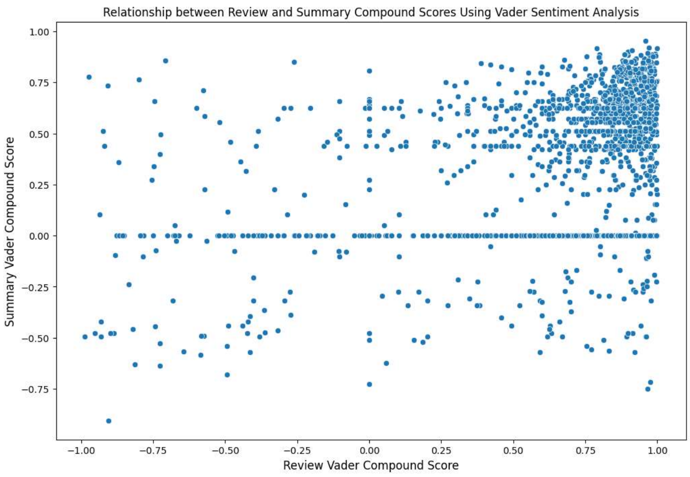


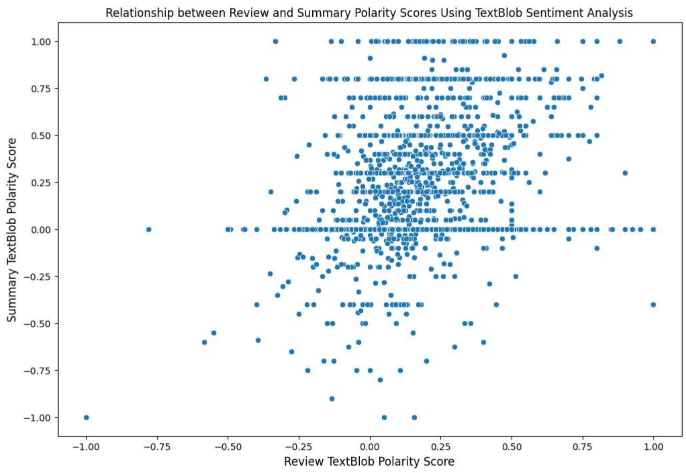
- The most frequent words in both summary and reviews are all positive - this suggests that customers feel very positive about Turtle Game's products.

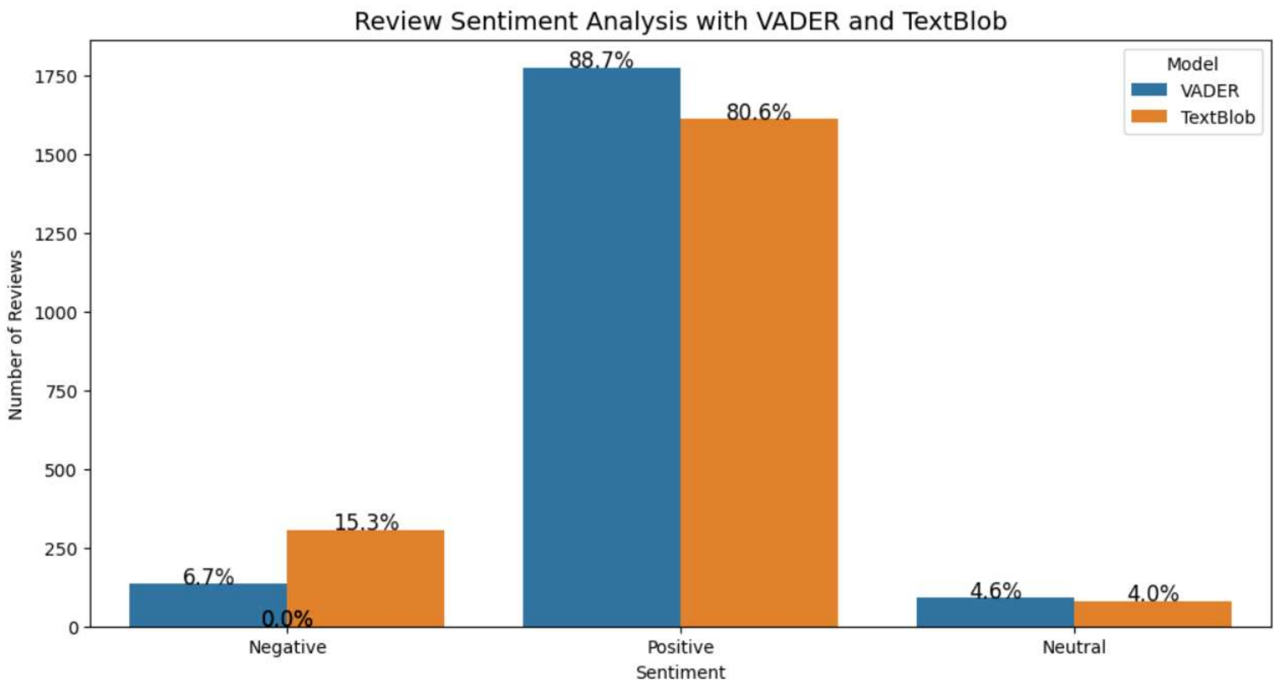
- The polarity of the top 15 words in the review column was analysed using TextBlob's *.sentiment.polarity* method. "Game" had a negative polarity of -0.4, while positive words like "great" (0.8), "good" (0.7), and "love" (0.5) stood out. Over half of the top 15 words (9/15) had a neutral polarity of 0.0.
- In the summary, "game" also had a negative polarity of -0.4. Positive words included "awesome" (1.0), "great" (0.8), "good" (0.7), "cute" (0.5), and "love" (0.5). Over half of the top 15 words (8/15) were neutral, including "stars" and "five," indicating that neutral reviews are over-represented.

- For the impact of removing "Five Stars" reviews on sentiment polarity distribution, *see Appendix Fig 10.*

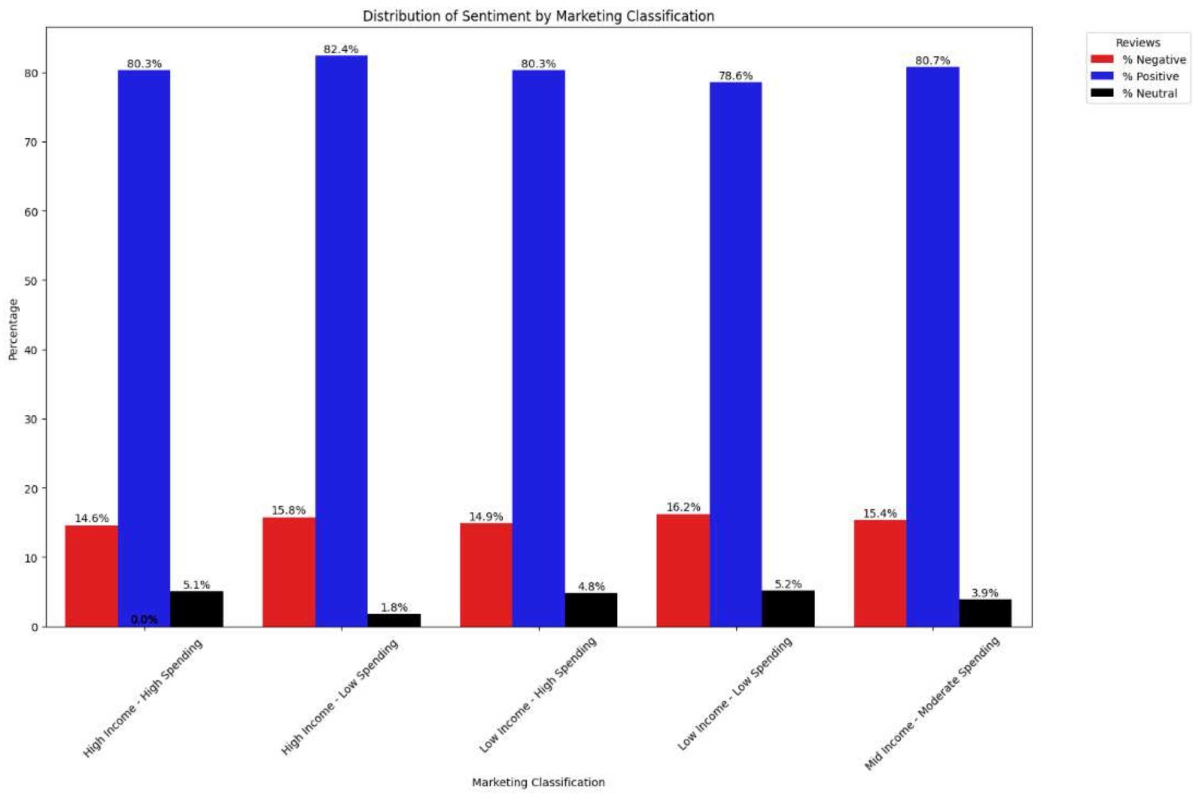
- Scatter plots reveal sentiment analysis limitations, for both Vader and TextBlob. If sentiments were accurately assigned, a strong positive correlation would be expected, but this is not the case. The relationship between the two variables is more noticeable in the TextBlob analysis.





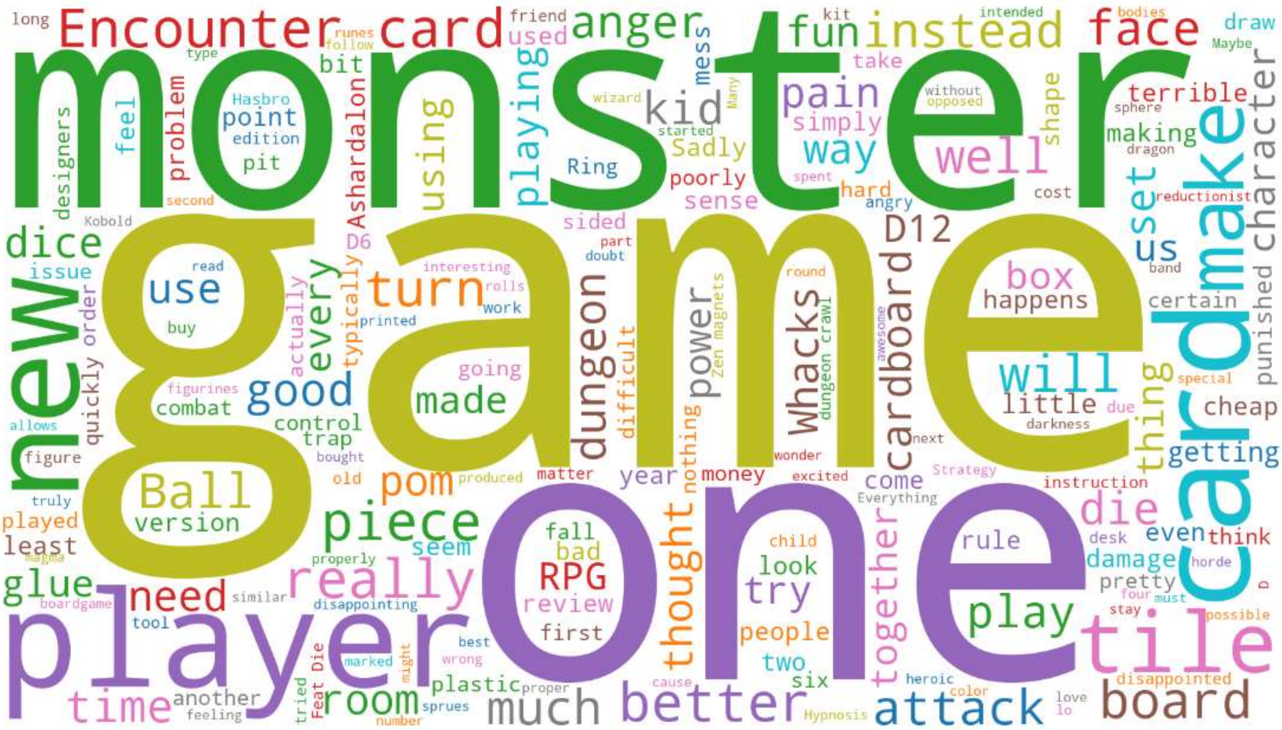


- Full reviews outperform summaries in sentiment analysis due to frequent misclassification, so only full reviews are used in further analysis.



- There is a high percentage of postive reviews across all groups.
- The group with high income but low spendings (conservative spenders) and mid income - moderate spenders (practical spenders) are the most satisfied.
- The mos dissatified custoemers are low income - low spenders (occasional shoppers) and high income - low spendings (conservative spenders). The last segment is probably more cristical and have higher expectations , so I think it is important to target them specifically and increase their satisfaction.

- Given the higher risk posed by negative reviews for Turtle Games, we created a WordCloud, revealing frequent complaints about board games, cards, quality, and usefulness. However, it doesn’t pinpoint specific products.



- To guide marketing strategies and product development, I made a list of products with the highest number of negative reviews. Using spaCy’s Matcher, I could automate adjective extraction from negative and positive reviews, to highlight common descriptors per product for better development and customer insights. However, I was unable to install spaCy’s Matcher due to hardware limitation.

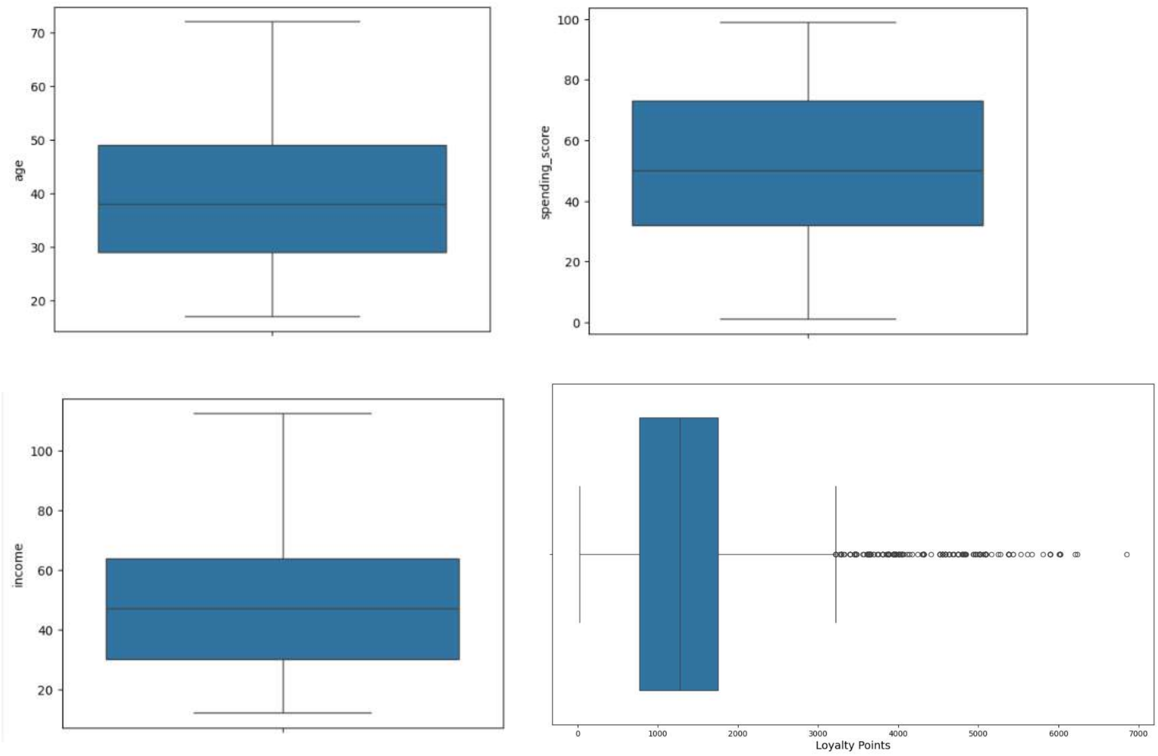
	product	negative_review_count	total_review_count	negative_review_proportion
116	6431	5	10	0.5
90	4399	5	10	0.5
145	9597	5	10	0.5
106	5512	5	10	0.5
53	2387	5	10	0.5
10	486	4	10	0.4
3	231	4	10	0.4
58	2795	4	10	0.4
64	2870	4	10	0.4
72	3436	4	10	0.4

5. Insights and recommendations

- Income and Spending Score together provide a strong explanation (83%) for Loyalty Points. These variables enable TurtleGames to better predict its most valuable customers.
- There is a wide variation in the loyalty points distribution, which suggests that customers aren't effectively engaging.
- A multiple linear regression model predicts how income and spending score affect loyalty points. Marketing can segment customers for targeted campaigns - boosting retention for high-loyalty customers with exclusive offers and attracting low-loyalty ones with enhanced rewards.
- Leverage clustering to tailor marketing efforts and personalized strategies, focusing on High Income - High Spending, Mid Income - Moderate Spending. Females consistently demonstrate higher engagement with loyalty points than males.
- Sentiment analysis can inform marketing and SEO by integrating common words from positive reviews. Turtle Games can engage loyal customers by thanking them for positive feedback.
- Based on sentiment analysis of customer's negative reviews, it's clear that the main concerns are product quality, age appropriateness, and the usefulness of board games. To address these issues, I recommend the following:
    - Improve Product Quality Control: tighten QC processes, offer easy returns, and highlight improvements in listings.
    - Redesign Board Games: conduct user testing, simplify instructions, and provide video tutorials.
    - Refocus Product Descriptions: emphasize problem-solving, add learning points, and use testimonials.
    - Clarify Age Appropriateness: reassess age ratings, clarify on packaging, and add age-specific tags.
    - Address Feedback Proactively: publicly respond to reviews and track sentiment trends quarterly.

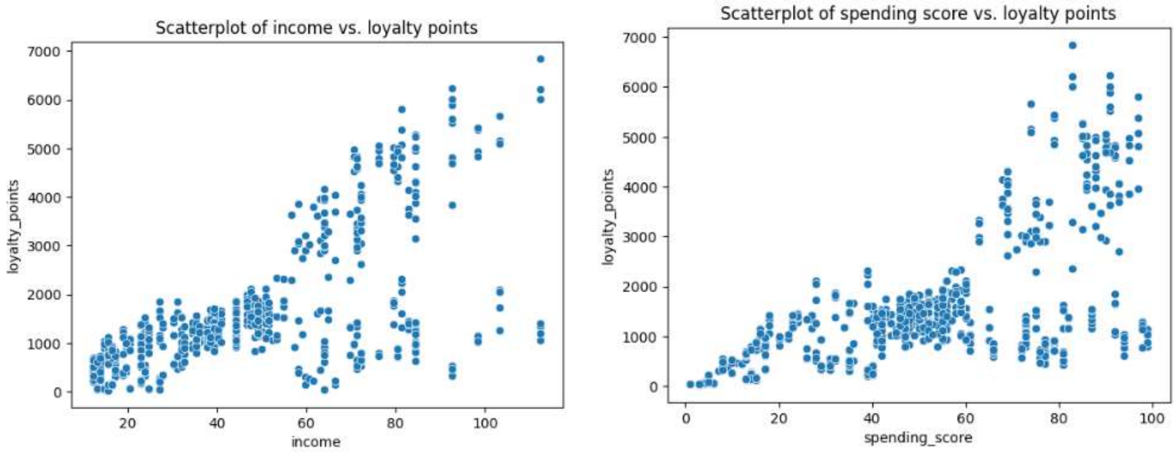
APPENDIX

Fig 1. Loyalty points present several outliers on the right whisker, suggesting that some customers have significantly higher loyalty points compared to others (right skewed distribution). Given the objective is to understand factor contributing to this variation outliers won't be removed.



.

Fig 2. We used Pearson correlation to understand the linearity (strength) and direction of relationships between income, spending score, age, and customer loyalty points. Scatterplots were chosen as the best way to quickly visualize this, to show the relationship between two variables and quickly identify a relationship. Immediately it was obvious no reliable correlation between age and loyalty points but there was some positive correlation for the other variables.



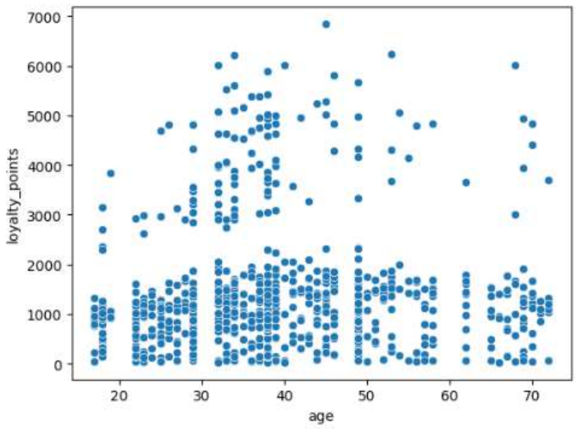


Fig 3. The model's fit significantly improved when analysing subsets with income below 55, and spending scores below 60, suggesting a stronger linear relationship in this group (the values were chosen based on the scatterplots observation).

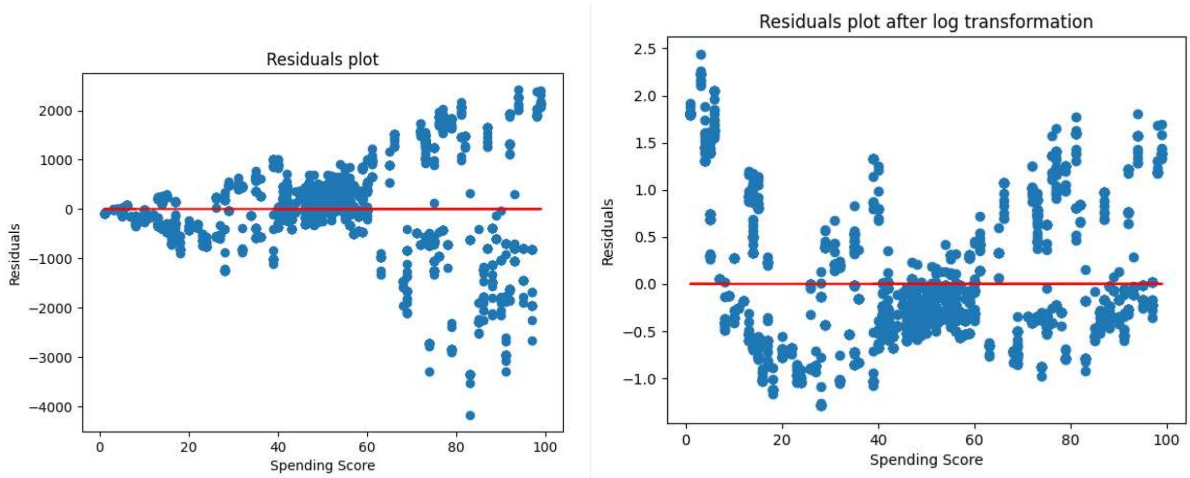
- OLS regressions for spending_score under 60 vs loyalty points (the fit of the model has increased significantly, and heteroskedasticity is not present).

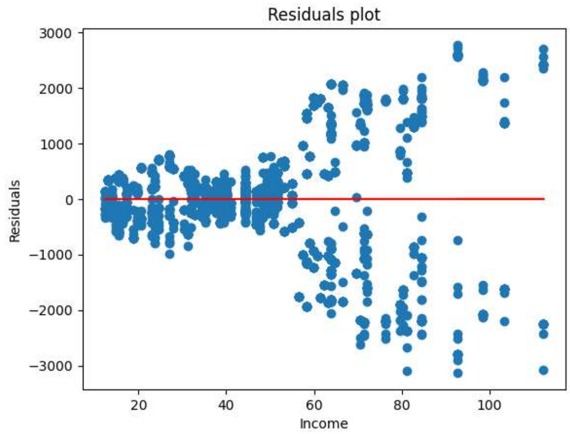
```
                            OLS Regression Results
==============================================================================
Dep. Variable:         loyalty_points   R-squared:                       0.601
Model:                            OLS   Adj. R-squared:                  0.601
Method:                 Least Squares   F-statistic:                     2055.
Date:                Wed, 02 Apr 2025   Prob (F-statistic):          1.74e-274
Time:                        20:48:22   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
==============================================================================
```

- OLS regressions for income under 55k vs loyalty points (the model fit this subset best, indicating a stronger linear relationship between income and loyalty points within this subgroup).

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         loyalty_points   R-squared:                       0.626
Model:                            OLS   Adj. R-squared:                  0.626
Method:                 Least Squares   F-statistic:                     2153.
Date:                Wed, 02 Apr 2025   Prob (F-statistic):         8.88e-277
Time:                        20:49:17   Log-Likelihood:                -9273.3
No. Observations:                1286   AIC:                         1.855e+04
Df Residuals:                    1284   BIC:                         1.856e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -15.5503     25.333     -0.614      0.539     -65.248      34.148
income        32.2286      0.695     46.397      0.000      30.866      33.591
==============================================================================
Omnibus:                       15.755   Durbin-Watson:                   2.905
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               11.473
Skew:                          -0.120   Prob(JB):                      0.00323
Kurtosis:                       2.604   Cond. No.                         101.
==============================================================================
```

Fig 4. Residuals plot before and after the application of the logarithmical transformation of the dependent variable for spending score and income.





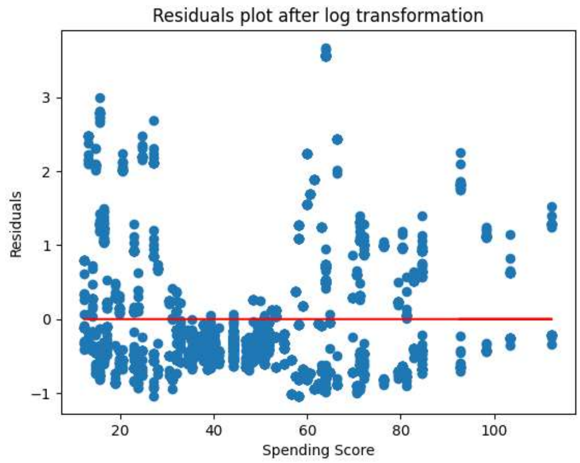


Fig 5. OLS models

- Spending Score vs Loyalty Points before and after log transformation.

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.
Date:	Mon, 31 Mar 2025	**Prob (F-statistic):**	2.92e-263
Time:	12:26:39	**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]**
Intercept	-75.0527	45.931	-1.634	0.102	-165.129	15.024
x	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.

OLS Regression Results

Dep. Variable:	y1	**R-squared:**	0.519
Model:	OLS	**Adj. R-squared:**	0.518
Method:	Least Squares	**F-statistic:**	2153.
Date:	Wed, 02 Apr 2025	**Prob (F-statistic):**	1.44e-319
Time:	17:54:16	**Log-Likelihood:**	-2146.7
No. Observations:	2000	**AIC:**	4297.
Df Residuals:	1998	**BIC:**	4309.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	5.5740	0.034	162.833	0.000	5.507	5.641
x	0.0282	0.001	46.400	0.000	0.027	0.029

Omnibus:	247.764	**Durbin-Watson:**	0.562
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	344.804
Skew:	-1.000	**Prob(JB):**	1.34e-75
Kurtosis:	3.366	**Cond. No.**	122.

  - After log transformation, the model explains 51.9% of the variation (vs. 45.2% before).
  - The increase in R-squared suggests that the log transformation improved the model's explanatory power.
  - The intercept is now statistically significant.
  - The coefficient for spending_score is now smaller because of the log scale, but $p<0.05$ so it remains highly significant.

- Income vs Loyalty Points before and after log transformation.

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.380
Model:	OLS	**Adj. R-squared:**	0.379
Method:	Least Squares	**F-statistic:**	1222.
Date:	Mon, 31 Mar 2025	**Prob (F-statistic):**	2.43e-209
Time:	12:26:42	**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]**
Intercept	-65.6865	52.171	-1.259	0.208	-168.001	36.628
x	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.

OLS Regression Results

Dep. Variable:	y1	**R-squared:**	0.284
Model:	OLS	**Adj. R-squared:**	0.284
Method:	Least Squares	**F-statistic:**	794.3
Date:	Wed, 02 Apr 2025	**Prob (F-statistic):**	1.98e-147
Time:	17:54:18	**Log-Likelihood:**	-2543.1
No. Observations:	2000	**AIC:**	5090.
Df Residuals:	1998	**BIC:**	5101.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	5.8505	0.045	131.318	0.000	5.763	5.938
x	0.0235	0.001	28.184	0.000	0.022	0.025

Omnibus:	610.463	**Durbin-Watson:**	2.844
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1512.287
Skew:	-1.669	**Prob(JB):**	0.00
Kurtosis:	5.647	**Cond. No.**	123.

  - After log transformation, the model explains 28.4% of the variation (vs. 38% before).
  - Adj. R-squared = 0.284 – Since it's equal to the R-squared, it confirms that the model is not overfitting.
  - The model is statistically significant.
  - The independent variable x and the interceptor are strong predictors (significant because p<0.05).
  - The R-squared value is small, suggesting that the model might not be the best for predicting loyalty_points.
  - Condition Number = 123 – Tests for multicollinearity (>30).
  - The x coefficient shows that a 1-unit increase in income increases loyalty points by 2.35%.

- Age vs Loyalty Points.

OLS Regression Results

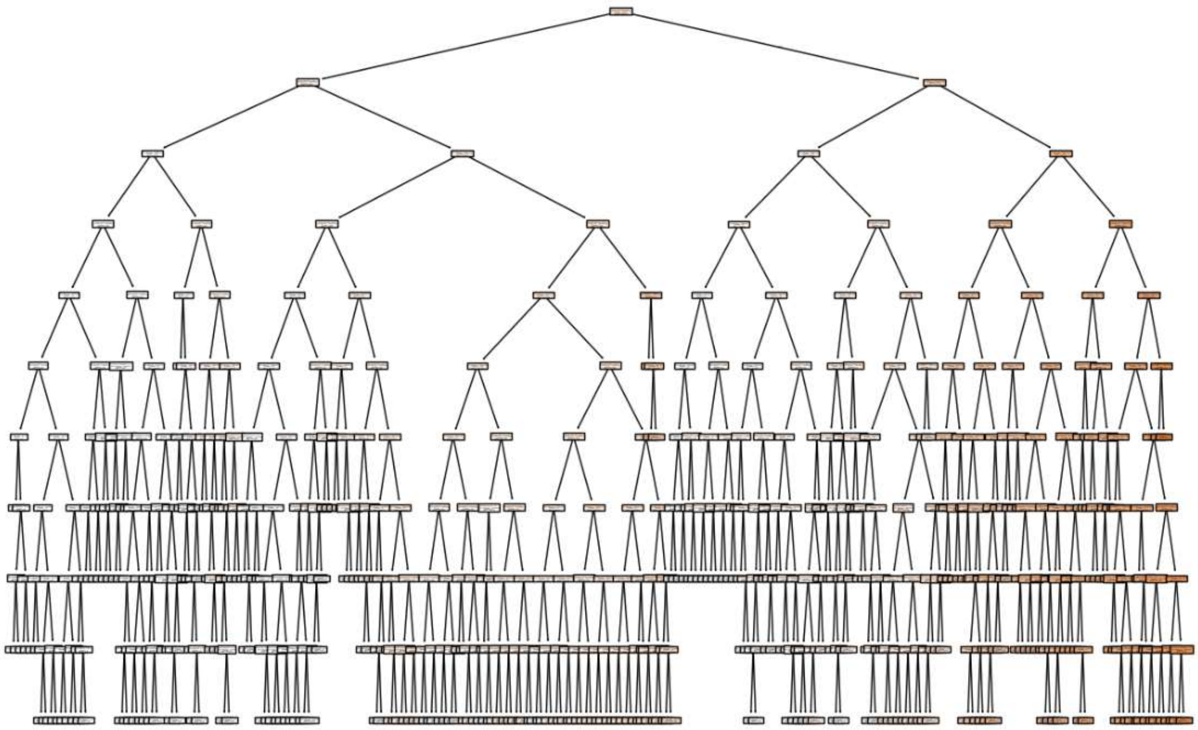
Dep. Variable:	y	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.606
Date:	Mon, 31 Mar 2025	**Prob (F-statistic):**	0.0577
Time:	12:26:44	**Log-Likelihood:**	-17150.
No. Observations:	2000	**AIC:**	3.430e+04
Df Residuals:	1998	**BIC:**	3.431e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1736.5177	88.249	19.678	0.000	1563.449	1909.587
x	-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.

  - Only 0.2% of the variance in loyalty points is explained by age. This is very low, indicating that age has almost no explanatory power for predicting loyalty points.
  - Adjusted R-squared is also very low, confirming that adding more variables would likely not improve the model's predictive ability.
  - p-value for is slightly above the typical significance threshold of 0.05, so the relationship is not statistically significant at the 5% level (though it's borderline).
  - F-statistic = 3.606, with a p-value of 0.0577 shows that the overall model is not statistically significant at the 5% level, but it's close.
  - The model has very poor explanatory power and lacks statistical significance, suggesting that age is not a meaningful predictor of loyalty points.
  - Checking residuals could help diagnose other issues (e.g., heteroscedasticity), but based on the weak relationship, a different modelling approach might be more useful.

Fig 6. The high depth and large number of leaves of the original tree (depth is 23, no. of Leaves are 563) suggest that the tree is complex, which could lead to overfitting . We want to try reducing the depth to simplify it by pruning it. The pruned model looks like this:



- The MSE increased from 10,468.88 to 15,437.31 after pruning, which is expected after pruning since the tree is less complex.
- The MAE increased from 38.47 to 64.48 — indicating that the model's average absolute prediction error increased by about 26 units. A small increase in MAE is expected when reducing model complexity — the goal is better generalization, not perfect training accuracy.
- R^2 dropped slightly from 0.9935 → 0.9905 — but it's still very high! A drop of 0.003% in R^2 is totally acceptable if it improves stability and reduces overfitting.
- RMSE increased from 102.32 → 124.25 — so the average size of prediction errors increased by about 22 units. A slight increase in RMSE is a sign that the model is no longer overfitting the training data — which is a positive trade-off if it improves performance on new data.

Fig 7. Support Vector Regression – SVR (because the outliers are a big problem the results were not satisfactory):

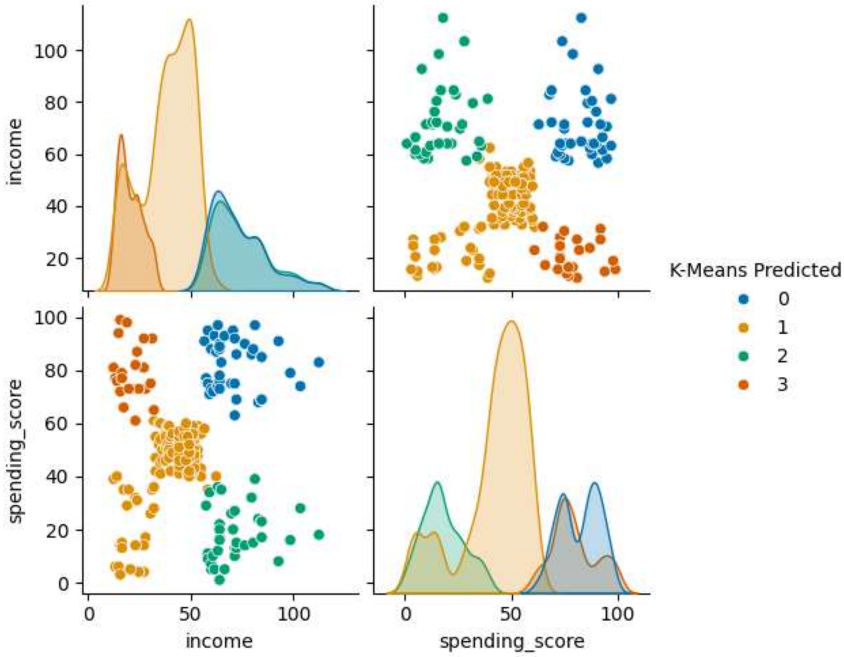
Mean Squared Error: 1441148.2471936087

R-squared: 0.11131984794377492

- SVR model's $R^2 = 0.11$, meaning it explains only 11% of the variation in loyalty points.
- Compared to our Decision Tree Regressor ($R^2 = 0.99$), SVR is underperforming.
- SVR may work if the relationship is continuous & smooth, but it's failing on structured/tabular data.

Fig 8. Evaluate k-means model at different values of *k* (k=4, k=5, k=6).

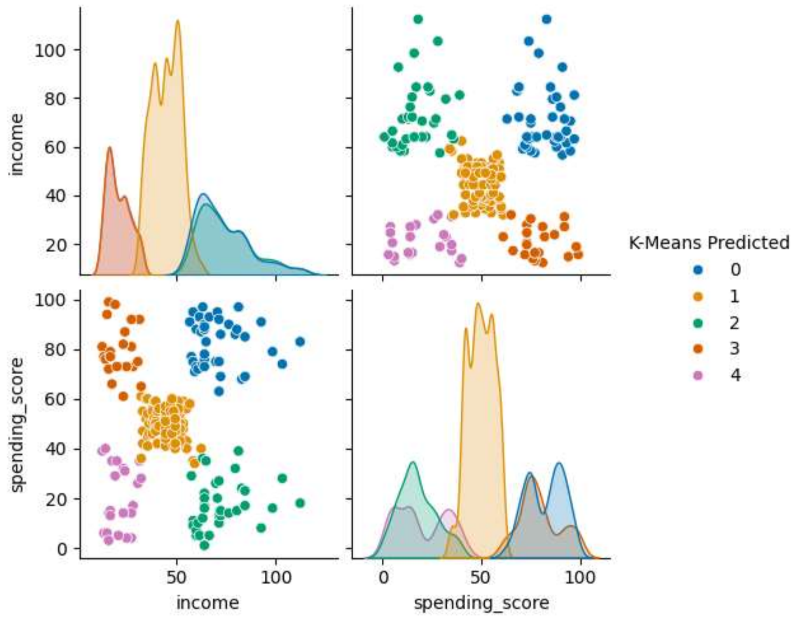
- k=4




```
K-Means Predicted
1    1038
0     356
2     337
3     269
```

- The output shows that the first cluster (n=1038) is still significantly larger than the rest.

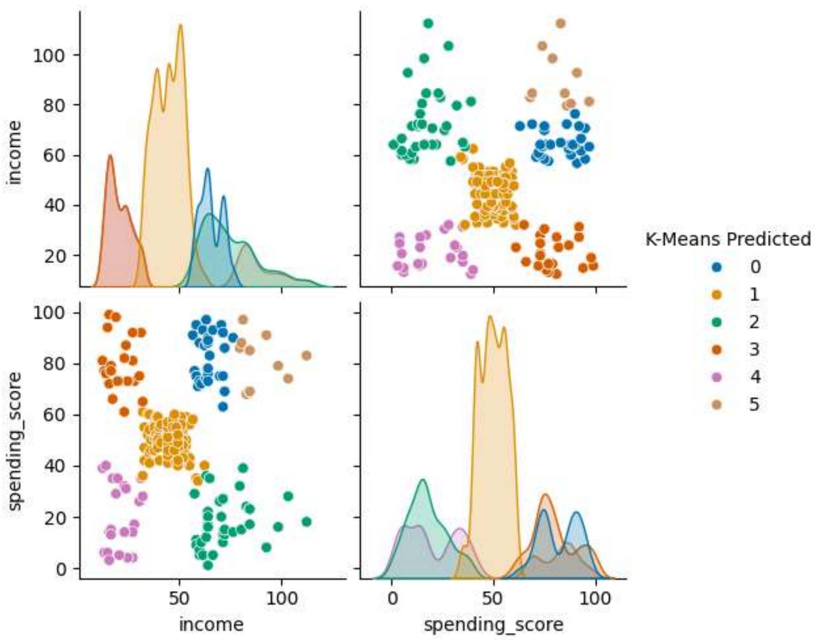
- k=5



K-Means Predicted	
1	774
0	356
2	330
4	271
3	269

  - The output shows that the clusters are grouped reasonably well.
  - Most customers sit in the average of income and spending score.

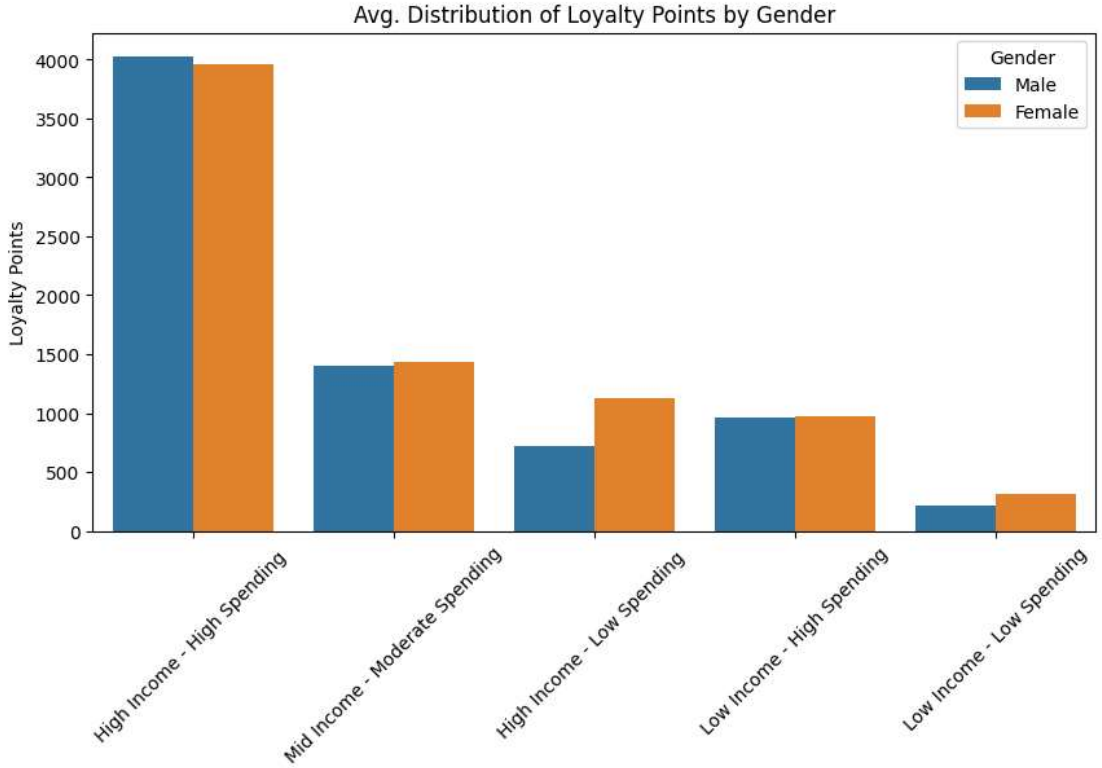
- k=6



K-Means Predicted	
1	774
2	330
4	271
3	269
0	238
5	118

  - The output shows that the last cluster (n=118) is significantly smaller than the rest.

Fig 9. Gender and education interactions with the loyalty program were analysed. Education showed no clear pattern, but females engaged slightly more with loyalty points, warranting further study.



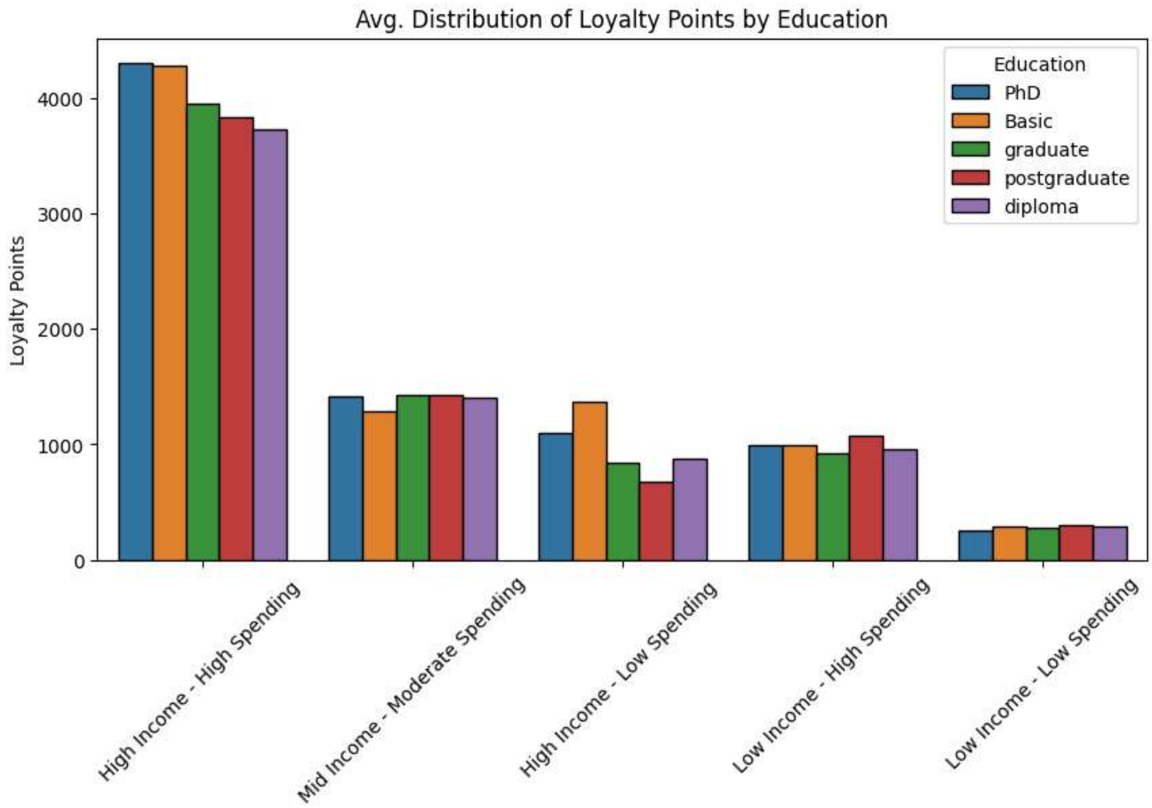


Fig 10. I noticed that the ‘Five Stars’ reviews are classified as neutral by the sentimental analysis, so I did a new histogram to see how much it affects the polarity. Removing it exposes more varied sentiment distributions and reduces the neutral peak.