Leveraging customer trends for improved Engagement

Lara Cadel

Table of contents:

1. Scenario:

Turtle Games seeks to understand customer engagement with loyalty points, identify customer segments for targeted marketing, and leverage customer reviews to inform campaigns. Additionally, they aim to assess whether loyalty points data can be used for predictive modeling. My goal is to help Turtle Games refine their business strategies by analysing customer trends, optimising marketing efforts, and making data-driven decisions to support their growth.

2. Analytical approach

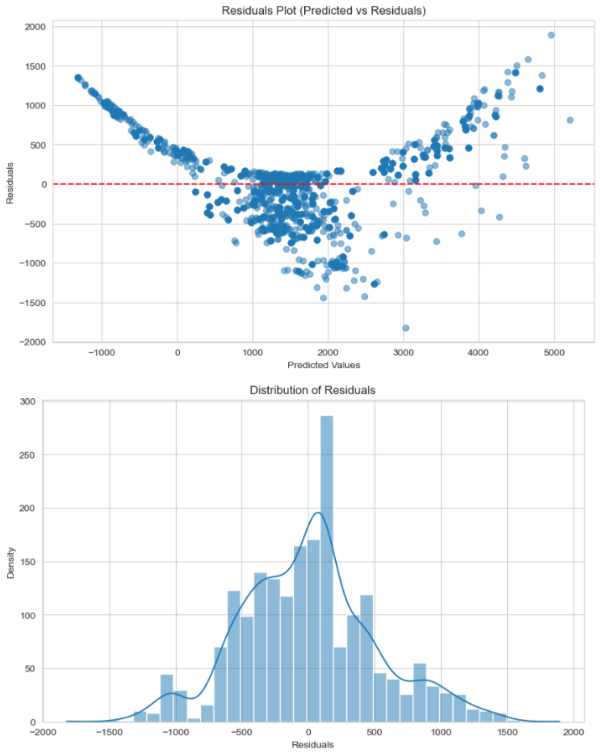
Prediction:

In my analysis, I began by exploring the DataFrame to understand its structure, followed by visualising numerical and categorical distributions. Initially, I used **simple linear regression** to investigate relationships between individual predictors (remuneration, spending score, and age) and loyalty points. Despite fitting the model with scikit-learn and generating an OLS summary with statsmodels, the results looked acceptably strong. Metrics like R-squared and P-values were significant, but normality issues were detected using Q-Q plots and the Shapiro-Wilk test.

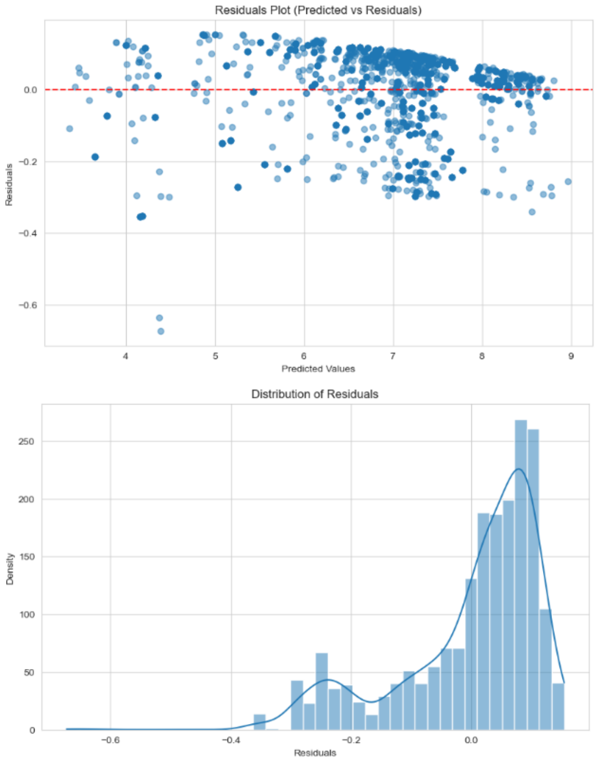
Building on model complexity, I moved to **multiple linear regression**, incorporating age, remuneration, and spending score, which considerably improved the model's accuracy. Moreover, the three factors model explains more variation in loyalty points than the one that excludes the variable 'age'. I confirmed the absence of multicollinearity using the Variance Inflation Factor (VIF).

Initially, the residuals observed in the residuals vs. predicted plot displayed a distinct U-shaped pattern, indicating the model was missing key non-linear relationships. This violated the linearity assumption of the OLS model. Therefore, I transformed skewed variables into logarithms. After the log transformation, the model's R-squared improved significantly to 0.987, and the F-statistic reached 4.931e+04, showing a much better fit. However, despite this improvement, the residuals were still not completely random and their distribution remained non-normal, indicating that important variables influencing the dependent variable may be missing from the model.

Residuals before log transformation



After log transformation



To address this, I decided to explore a non-linear approach by using a **Decision Tree Regressor**. This model is may reveal decision points missed by the linear model. After splitting the data into training and testing sets, I evaluated the model's performance using metrics like MAE, MSE, RMSE, and R-squared. I also explored pruning, refining the model to avoid overfitting.

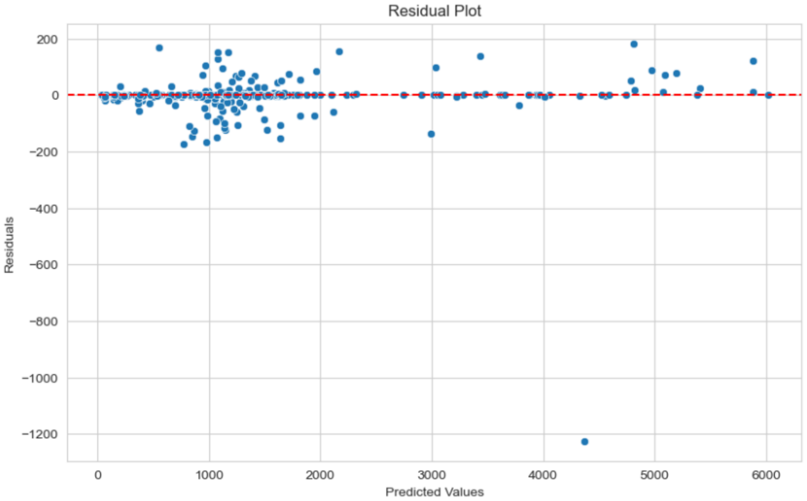
To better understand the feature relationships, I trained three decision tree models. Model 1 used 'gender_Male' and 'education' as predictors, yielding an RMSE of 653 and a very low R-squared, suggesting poor predictive power. Model 2, using `remuneration` and `spending_score`, performed better with an RMSE of 269, but R-squared remained low. Model 3 combined all four variables but showed no significant improvement over Model 2.

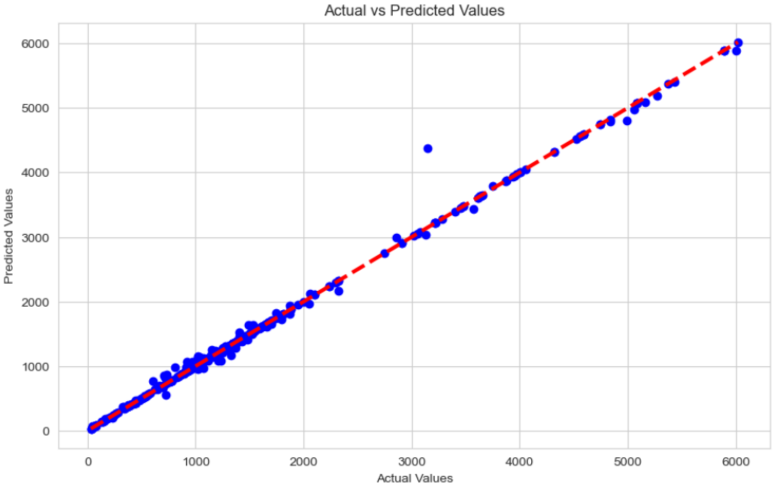
Building upon these decision tree models, I explored **Random Forest**, to capture potential interactions that individual trees may have missed. Random Forests often perform better due to averaging over multiple trees, which reduces variance.

After training the Random Forest model using age, spending, and remuneration as predictors, the results were stronger. The R-squared increased to 0.9967, which means the model explained nearly all the variability in the loyalty points. Also, the RMSE dropped to 72.79, indicating the model was much more accurate.

To further validate the model, I applied cross-validation: its RMSE was 58.82, confirming that the Random Forest model not only fits the training data well but also generalises effectively to unseen data.

Further analysis of the residual and actual vs. predicted plots supports these findings. The residuals are randomly distributed around zero, indicating that the model does not exhibit systematic bias. Although there are a few outliers, the errors remain generally small. The actual vs. predicted plot shows a strong alignment with the ideal prediction line, with most points closely clustering around it.





Segmentation:

Since spending score and remuneration emerged as important predictor of loyalty points, I pivoted to uncovering patterns in these data, segmenting them with **k-means clustering**: first, I grouped customers based on remuneration and spending score, and then by using spending score and loyalty points. To determine the optimal number of clusters, I used both the Elbow method and the Silhouette score.

The remuneration vs. spending score clustering shows well-defined groups with clear separations, indicating distinct segments based on these features. The same does the spending score vs. loyalty points clustering even if with some overlap. I then decided to assigne each observation to its respective cluster.

3. Visualisation and insights

Prediction:

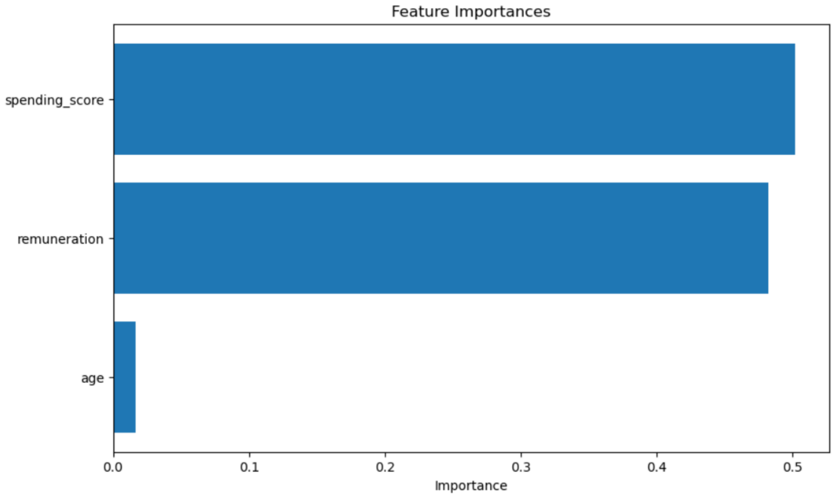
The layered approach, from prediction to segmentation, allows for incremental learning and adaptation, helping Turtle Games tailor their marketing efforts effectively.

The Random Forest model applied to predict loyalty points yielded accurate results, demonstrating strong predictive power while also revealing key drivers of customer loyalty through the analysis of feature importance.

Spending score emerged as the most significant predictor of loyalty points, with an importance score above 0.5. This underscores the critical role of customer spending behavior in driving loyalty. Marketing strategies should prioritise understanding and optimising these spending patterns to enhance customer engagement and retention.

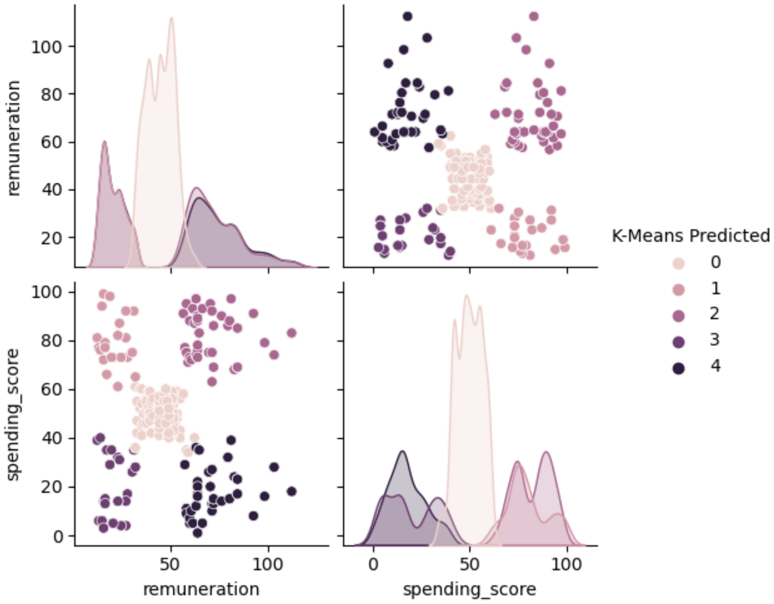
Remuneration ranked second, with an importance score slightly above 0.3. Targeting higher-income customers can still be valuable, but the focus should remain on spending-related strategies.

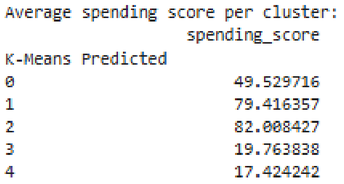
Age had the smallest impact, resulting in a negligible variable.



Segmentation:

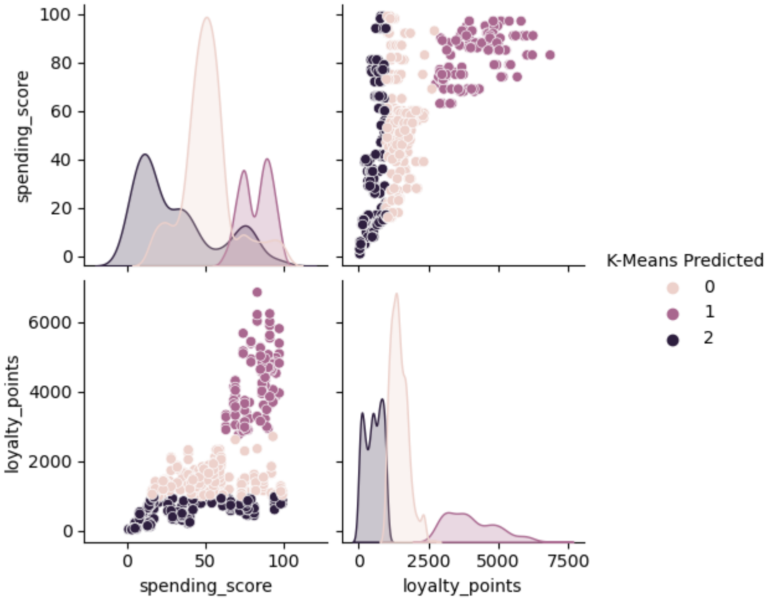
The identification of clusters that highlight variations in customer engagement and spending patterns, provides valuable opportunities for targeted marketing strategies.

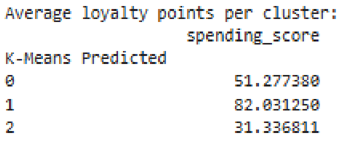




```
Average spending score per cluster:
                         spending_score
K-Means Predicted
0                             49.529716
1                             79.416357
2                             82.008427
3                             19.763838
4                             17.424242
```

- Cluster 0: Low remuneration and spending, likely lower spenders.
- Cluster 1: Moderate remuneration and spending, representing an average customer base.
- Cluster 2: High spenders with low to moderate remuneration, possibly spending beyond their means.
- Cluster 3: High remuneration and spending, representing premium, high-value customers.
- Cluster 4: High remuneration but lower spending, indicating untapped spending potential.



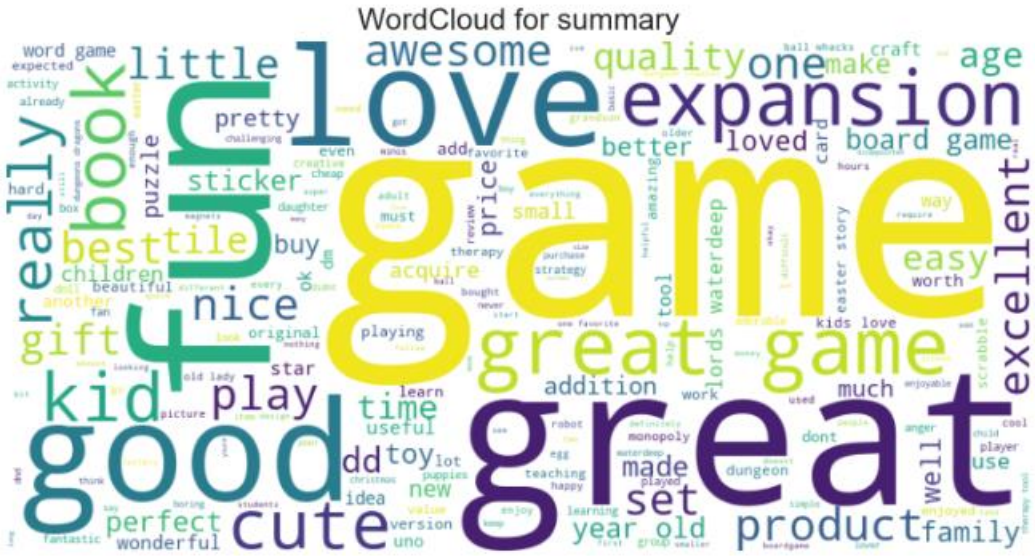


```
Average loyalty points per cluster:
                         spending_score
K-Means Predicted
0                             51.277380
1                             82.031250
2                             31.336811
```

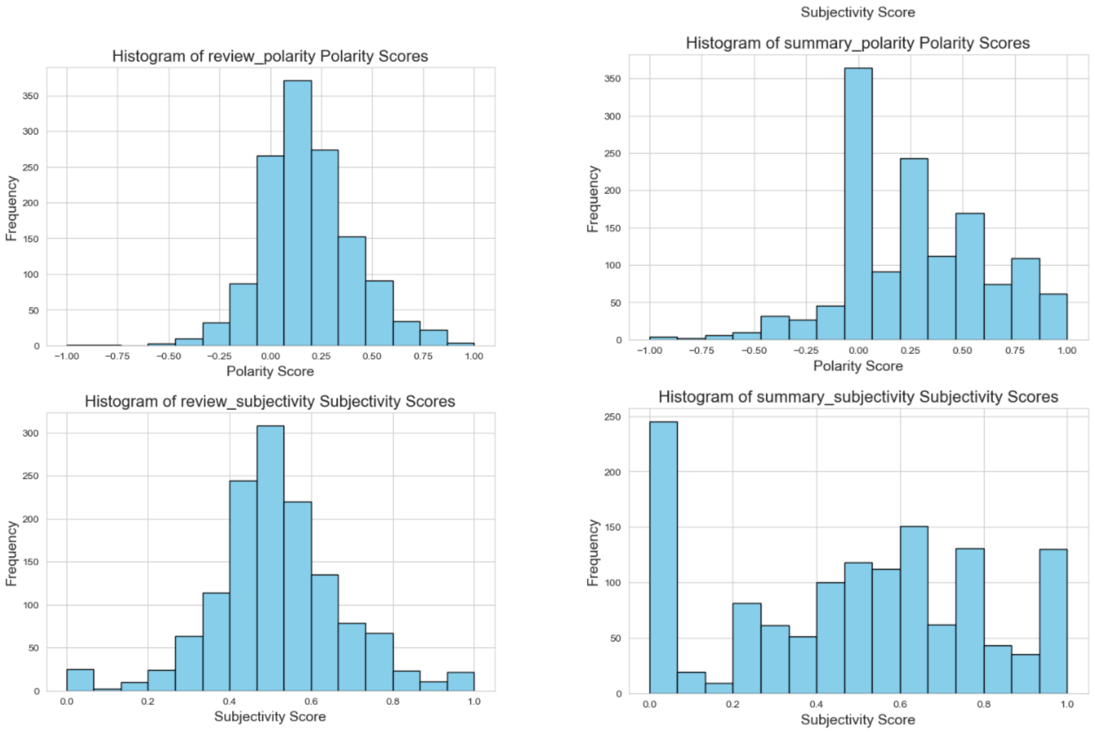
- Cluster 0: Low spending and loyalty points, likely inactive customers.
- Cluster 1: High spending and loyalty points, representing highly engaged, loyal customers.
- Cluster 2: Moderate spending but low loyalty points, suggesting potential for increased engagement through targeted marketing.

4. NLP for Customer Feedback Analysis:

The aim of this analysis is to examine customer feedback through text analysis, focusing on reviews and summaries data, in order to identify key patterns and sentiment. I prepared the text columns for NLP by cleaning, tokenising, and removing punctuation and stopwords. Next, I generated word clouds and frequency distributions to visualise the most common terms.



I then calculated polarity and sentiment scores to assess the emotional tone of the feedback.



The results of my analysis show that customer reviews generally have a neutral sentiment, as seen in the polarity histogram, with most scores clustering around zero. The reviews also tend to be moderately subjective. In contrast, the summaries show a wider range of sentiments, with more positive scores and a noticeable split in subjectivity, indicating that they are either highly subjective or objective, unlike the more balanced reviews. This suggests that reviews offer balanced feedback, whereas summaries tend to emphasise stronger opinions.

5. References:

- Brownlee, J 2020a, Classification And Regression Trees for Machine Learning, machinelearningmastery.com, https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/ Links to an external site. .

- Harmouch, M 2021, 17 Clustering Algorithms Used in Data Science and Mining, towardsdatascience.com, https://towardsdatascience.com/17-clustering-algorithms-used-in-data-science-mining-49dbfa5bf69a Links to an external site. .

- Cran.R-Project 2022, Introduction to stringr, cran.r-project.org, https://cran.r-project.org/web/packages/stringr/vignettes/stringr.html Links to an external site. .