Technical Report

Background/context of the business

Turtle Games is a game manufacturer and retailer specialising in games and hobby products. The company is seeking to better understand customer behaviour in order to improve how it manages loyalty and engagement. Specifically, Turtle Games want to know how customers accumulate loyalty points, whether customer segments can be identified for more targeted marketing, and whether customer review data can provide insights into customer experience. The objective of this analysis is to identify meaningful patterns that can improve overall sales performance, using customer data such as spending score, remuneration, and loyalty points, together with review text data.

Analytical approach

The analysis was conducted using a combination of Python and R, with each language selected based on its strengths at different stages of the workflow. Python was used for data preparation, exploratory analysis, regression modelling, clustering, and text analytics, while R was used to support statistical exploration, validation, and visualisation.

Customer data were initially loaded and prepared in Python using the pandas library. Data preparation focused on ensuring quality and consistency, including inspection of data types, checks for missing values and duplicates, removal of redundant variables, and consistent renaming of columns. Descriptive statistics and grouped summaries were used to understand how loyalty value is distributed across the customer base.

Relationships between loyalty points and key explanatory variables, particularly spending score and remuneration, were explored using visualisations, correlation analysis, and simple linear regression in Python pandas and statsmodels. These models were used to examine the strength and direction of relationships and assess statistical significance.

To validate these findings, multiple linear regression was conducted in R using the lm() function. This stage focused on statistical diagnostics and model assumptions. The distribution of loyalty points was examined using measures such as skewness, kurtosis, and normality tests, while correlation matrices and Variance Inflation Factors (VIF) were used to assess multicollinearity among predictors. The resulting model explained approximately 83% of the variation in loyalty points, indicating a strong overall fit.

As an additional exploratory step, decision tree regression was conducted in Python using scikit-learn. Model performance was evaluated using training–test splits and error metrics such as MAE and RMSE across different tree depths. While decision trees provided useful structural insight, they did not materially improve predictive performance relative to simpler models.

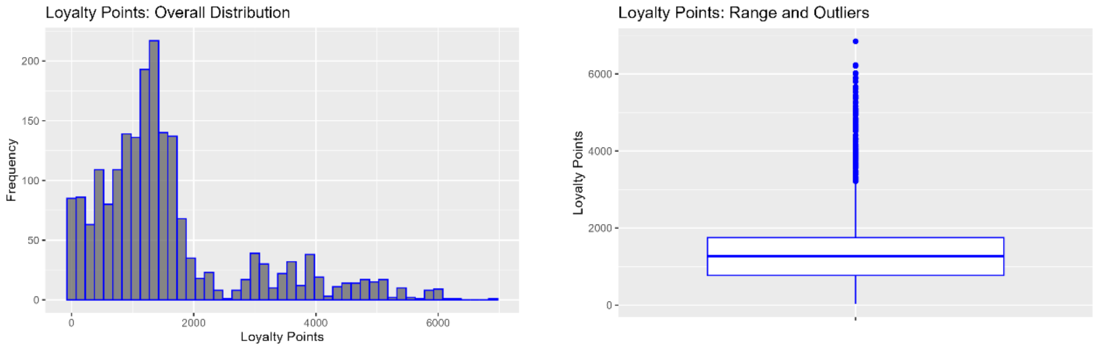
Customer segmentation was explored using k-means clustering in Python. The best number of clusters was assessed using elbow and silhouette methods. The resulting clusters identified distinct groups of customers with similar income and spending characteristics, supporting Turtle Games' aim of targeted marketing and prioritisation.

Finally, Natural Language Processing (NLP) techniques were applied to customer review text using Python libraries including NLTK. Reviews were cleaned and tokenised, and sentiment polarity scores were generated to assess whether unstructured text data provided additional insight into customer loyalty behaviour.

Visualisation and insights

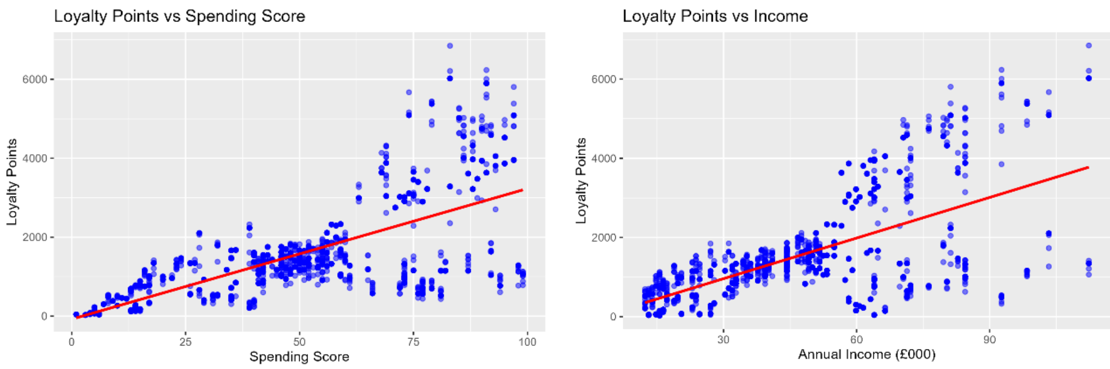
The distribution of loyalty points, the dependent or outcome variable used in the analysis, was examined using a histogram and a boxplot (Figure 1). These visualisations were chosen to assess both the overall shape of loyalty accumulation and the presence of extreme values. The results show that most customers accumulate a moderate number of loyalty points, while a small proportion earn an exceptionally high number of points. This right-skewed distribution indicates that customer value is unevenly distributed and that average loyalty metrics alone don't show the whole picture.

Fig. 1: Loyalty point distribution



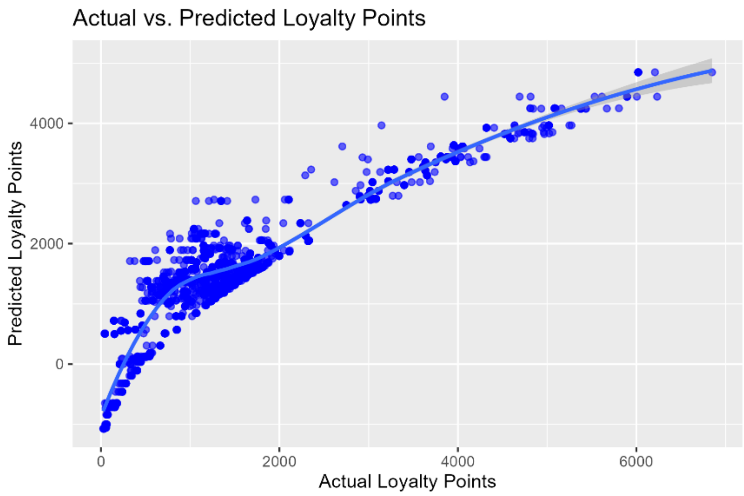
Scatterplots were then used to explore the relationship between loyalty points and two key customer attributes, spending score and income (Figure 2). Scatterplots were selected because they allow both the direction and strength of relationships to be assessed visually. Both variables show a clear positive association with loyalty points (with correlations of approximately 0.62 and 0.67 respectively). Importantly, spending score and income capture different aspects of customer behaviour and together provide complementary insight.

Fig. 2: Drivers of loyalty points



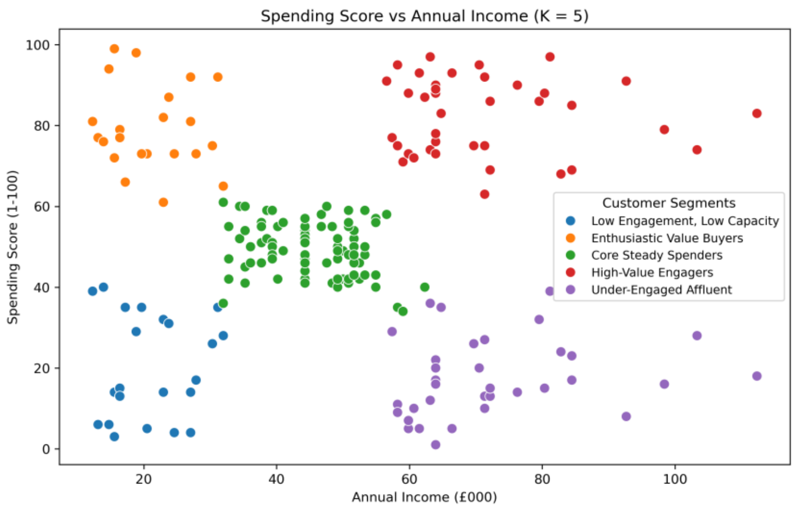
A multiple regression model was then run with an actual versus predicted plot (Figure 3). This visualisation was chosen to assess overall model performance rather than individual coefficients. The close alignment between predicted and actual values indicates that approximately 83% of the variation in loyalty points can be explained using income and spending behaviour. While prediction accuracy is lower for customers with very low engagement, the model performs well overall.

Fig. 3: Predicting loyalty points



Customer behaviour was further explored using clustering based on income and spending behaviour (Figure 4). The visualisation, taken together with supporting analysis (see Analytical Approach section) indicated that a five-cluster solution best represented customer groups. This segmentation provides a framework for differentiated customer strategies.

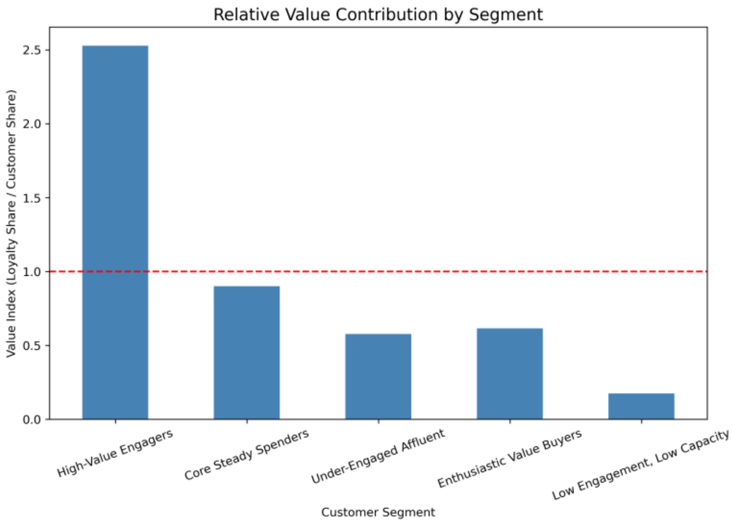
Fig. 4: Customer segmentation



Finally, a Customer Value Index was introduced to compare segments on a like-for-like basis by relating loyalty contribution to segment size (Figure 5). This visualisation highlights that High-Value Engagers contribute disproportionately to total loyalty value, while other segments under-contribute relative to their size.

This shows the importance of prioritising segments based on value contribution rather than customer count alone.

Fig. 5: Customer Value Index



Patterns and predictions

Several patterns emerged from the analysis that address Turtle Games' objective of understanding and strengthening customer loyalty. First, loyalty point accumulation is highly uneven across the customer base, with a relatively small proportion of customers contributing a disproportionately large share of total loyalty value.

Second, loyalty behaviour shows clear relationships with both spending and income. Customers who are more actively engaged in purchasing, as well as those with greater spending capacity, tend to accumulate substantially higher loyalty points. These patterns underpin the segmentation results, which show that customers fall into distinct and interpretable groups based on these characteristics.

Analysis of customer review text provided additional insights. Sentiment and keyword analysis showed that positive reviews frequently emphasised enjoyment, fun, gifting, and shared family experiences. Negative feedback often related to unclear instructions, quality concerns, and expectations not being met. While no correlation was found between review sentiment and loyalty points, the NLP analysis provided insights which guided recommendations for Turtle Games' marketing and operations team to consider.

With regard to prediction, the regression analysis demonstrates that loyalty behaviour is largely predictable using income and spending score alone, explaining over 80% of the variation in loyalty points. This means that Turtle Games can reasonably estimate customer value using data it already collects. While prediction accuracy is lower for low spending customers, the model performs well overall and is particularly effective for identifying medium- and high-value customers.

Together, these findings suggest a clear strategic focus: protecting and developing relationships with high-value customers, increasing engagement among under-engaged but affluent customers, and managing lower-value segments efficiently. Insights from customer review text further suggest that factors such as enjoyment, gifting, and shared family use play an important role in positive customer experiences, while there is also an opportunity to address usability issues and quality concerns. Some potential areas for further analysis include exploring which products or categories drive loyalty most strongly, and also evaluating the impact of specific marketing or loyalty initiatives.