Turtle Games: Predictive Analytics Technical Report

Charles Garrod, Data Analyst

1.0 Project Scope

Turtle Games is a global manufacturer and retailer of games, ranging from board games to video games and toys. Turtle Games want to try and improve their sales performance. It has a robust data collection process and collects data from its customer accounts and reviews. The company wants greater clarity on how customers accumulate loyalty points and how the marketing department might use this data to inform its marketing strategy. While no data is collected at point of sale, its loyalty point system is well maintained and so this will be used as a proxy for sales. A list of key questions has been provided in appendix 1.

2.0 Analytical Approach

The metadata provided by turtle games breaks down the dataset into its component parts; gender, age, remuneration (income), spending score (on a 1-100 scale), loyalty points, education level, language, platform, product number, review and summary.

I began the analytical process by importing the necessary libraries, as detailed blow:

- Pandas & Numpy – Data ingestion, cleaning and wrangling
- Matplotlib & Seaborn – Visualisations
- Scikit Learn – Construction and deployment of ML models
- Statsmodels – Statistical Modelling
- Scipy – Computational Science
- NLTK – Natural language processing

The next step was ingestion of data (turle_reviews.csv) and reviewing the data from types to summary statistics. I then cleaned the data by ensuring no null values and dropping unnecessary columns (platform and language were all the same and therefore obsolete for the purposes of analysis). Finally, the spending score and remuneration columns were renamed making their recall easier and more efficient. I saved the new DataFrame as a CSV, and this formed my basis of the analysis.

2.1 Regression

I used both linear regression and multiple liner regression models to investigate the impact a variation in the independent variables (age, remuneration, spending score) had on the dependent variable (loyalty points). I created the formulas using the ordinary least squares (OLS) method, investigated the output checked for linearity and independence (VIF) of the residuals.

2.2 Clustering

I used K-means clustering method using a new DataFrame containing only remuneration and spending score, plotted a scatter plot to see if there were any obvious clusters from the outset. Elbow and Silhouette methods were then introduced to identify the optimal number of clusters

to use, k=5. k=6 and k=4 were also investigated but k=5 was optimal as it minimised the within-cluster sum of sum of squares.

2.3 Natural Language Processing

I performed analysis on the review and summary columns by preparing the data so as to make sure it was suitable for processing; text standardisation, removing punctuation, removing duplicates, tokenisation, and frequency analysis determining the most common words. I also investigated sentiment of the reviews using polarity and then identified the 20 most positive/negative reviews and summaries.

2.3 Further Analysis in R

2.3.1 EDA

I ended my analysis in R by, firstly, importing the necessary libraries which included:

- Tidyverse – Used to make scientific computing more efficient
- Dplyr – Streamlines data manipulation by using DataFrames
- Ggplot2 – Facilitates the creation of visualisations
- Skimr – Provides summary statistics about variables
- DataExplorer – Allows automation of most data handling and visualisation tasks
- NbClust – Determines optimal number of clusters in a given data set
- Factoextra – Helps extract and visualise the most important metrics within a dataset
- Psych – Assists in data cleaning and recording
- Moment – Helps identify statistical shape of data (skewness etc.)

I started the analysis by ingesting the cleaned data (previously cleaned in Python) then investigating the distribution of each variable in the shape of a histogram. I continued with the visualisations, using scatter plots to highlight the relationships between the various independent variables and loyalty points. I then looked at each relationship through the lens of the categorical variables (gender, education level).

2.3.2 Statistical Analysis

The final step was performing statistical analysis, in which I identified the statistical makeup of the data (mean, median, variance, standard deviation etc.) I followed this with a multiple linear regression where I tested two models, using two and three independent variables, and compared their explanatory power. I then generated new, random data for the independent variables to test the model and produce predictions about the loyalty points.

3.0 Observations & Insights

Linear regression models were used to explore the relationship between the independent variables (remuneration, spending score, age) and loyalty points. The below shows visuals show that there is a clear positive correlation between spending score/remuneration and loyalty points, where age shows a very weak negative correlation.

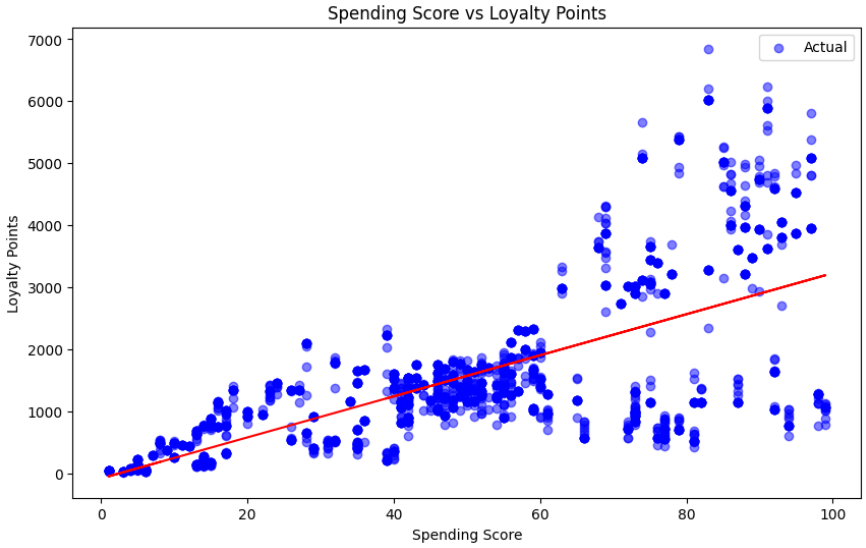


Figure 1: Spending score vs Loyalty Points

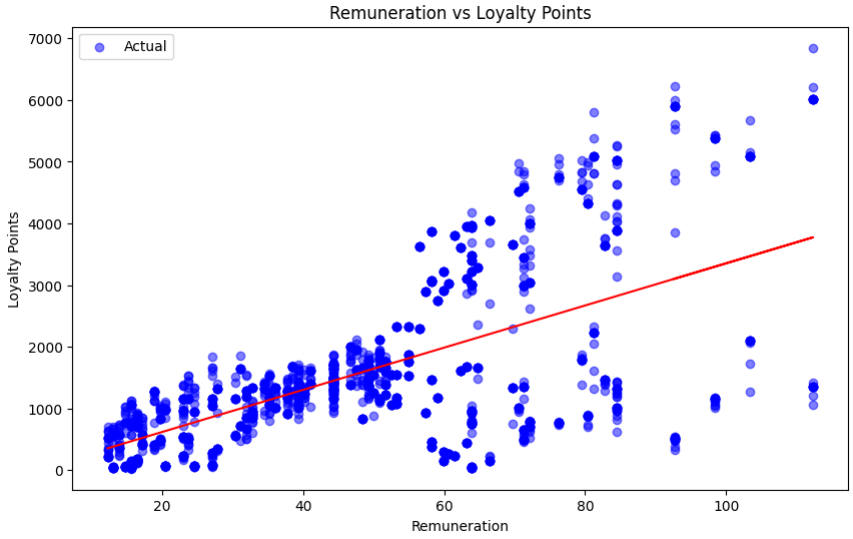


Figure 2: Remuneration vs Loyalty Points

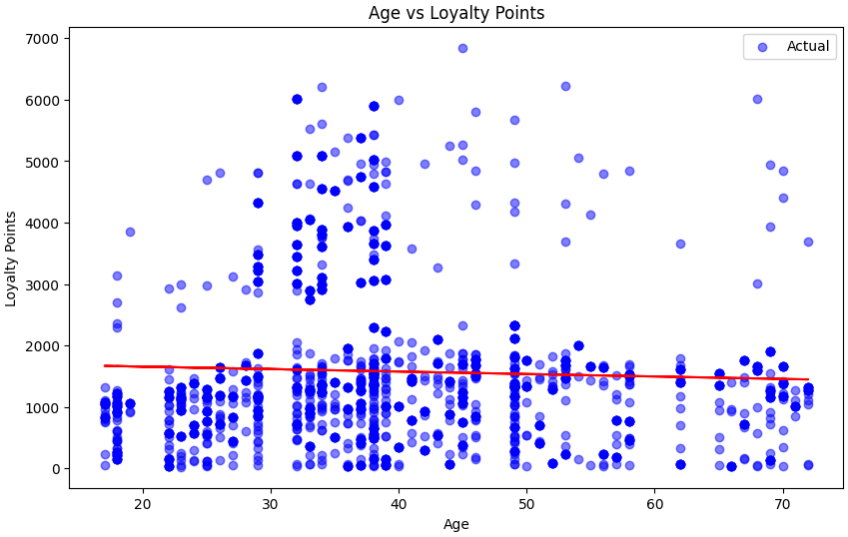


Figure 3: Age vs Loyalty Points

The models suggest that 45%, 38% and 0.02% of the variation in loyalty point accumulation can be explained by spending score, remuneration and age respectively. According to these models, a 1 unit change in either spending score, remuneration or age equates to a 33, 34 and -4 unit change in loyalty points respectively.

The relationship between spending score and remuneration allowed me to identify specific clusters in Turtle Games' customer base. Using the elbow and silhouette method, its clear that 5 clusters was optimal.

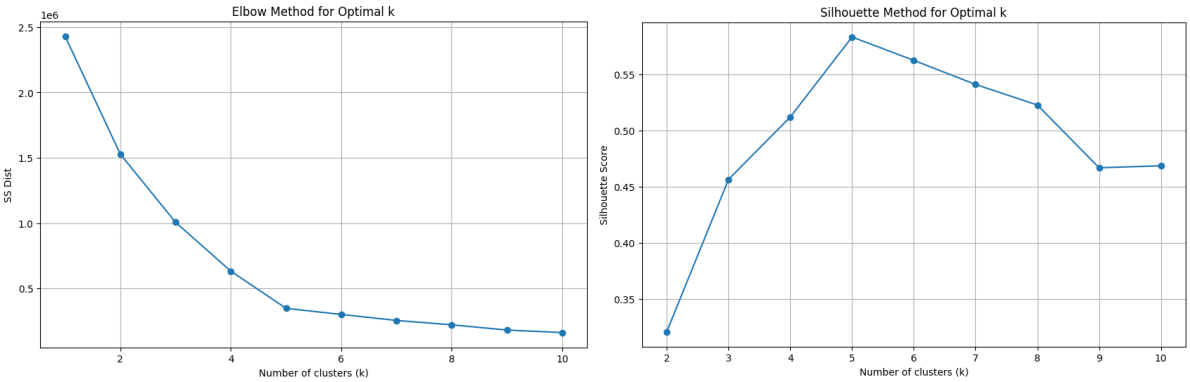


Figure 4: Elbow and Silhouette method to identify optimal k

The segmentation of customers is extremely important for the marketing department to focus their marketing strategies and not use a blanket strategy that may be expensive but yield poor results, see appendix 2 for more information.

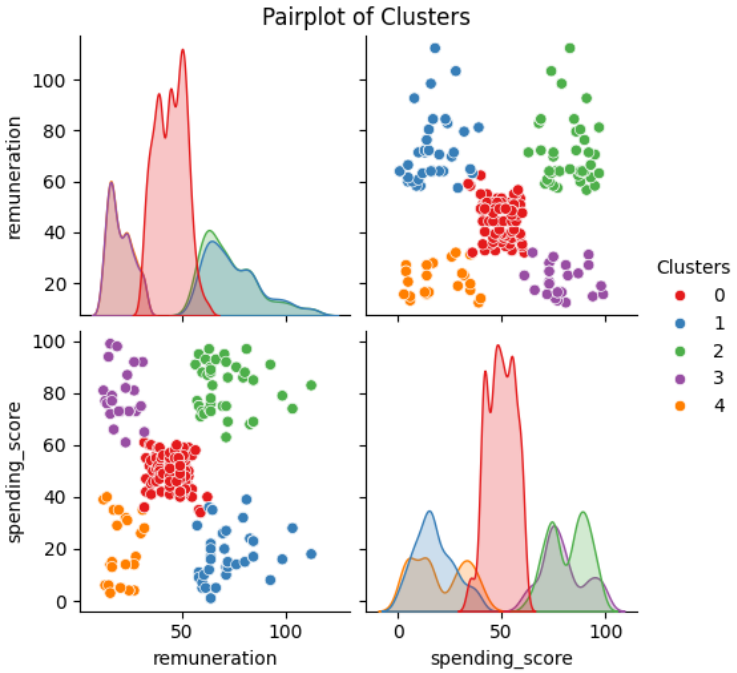


Figure 5: Pairplot highlighting key customer clusters

NLP allowed the analysis of customer reviews and summaries. Using a polarity score, its clear to see those reviews and summaries have an overall positive trend.

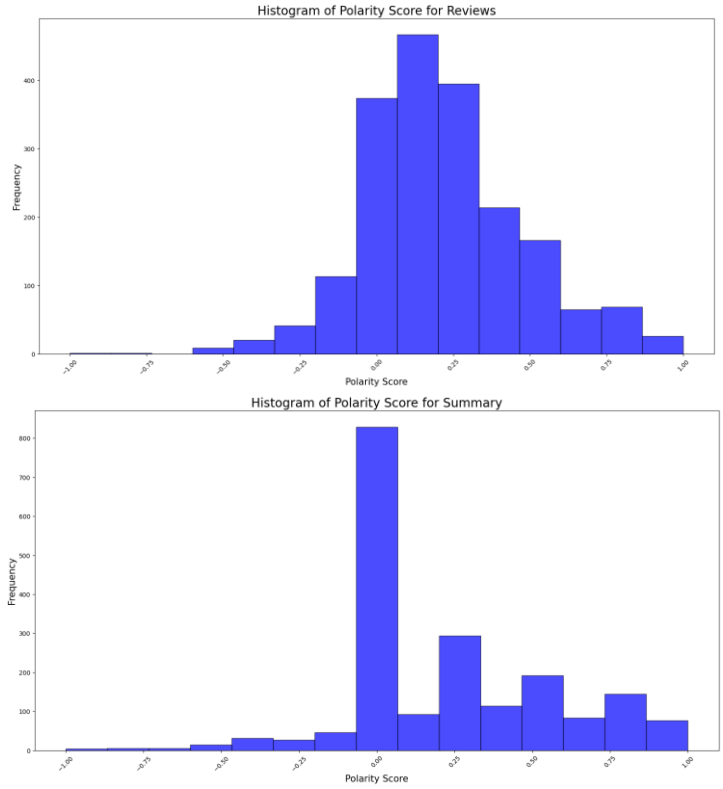
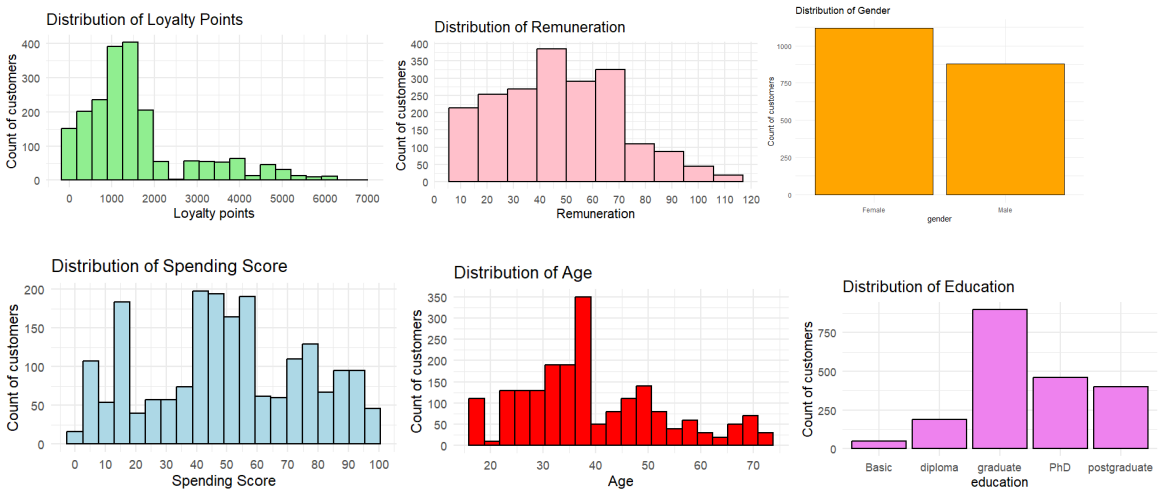
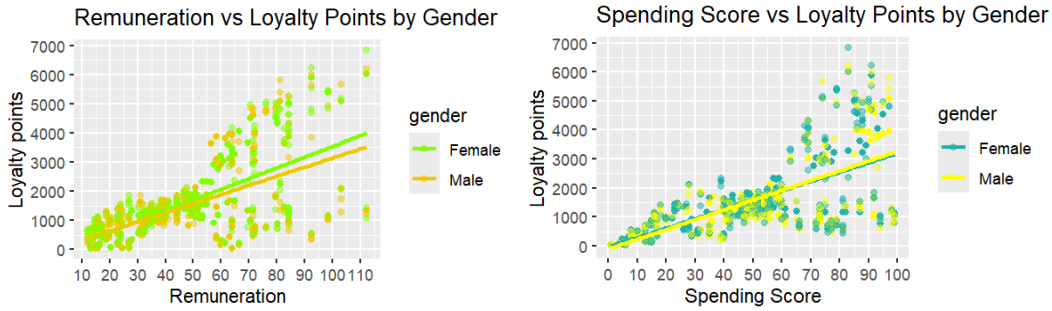


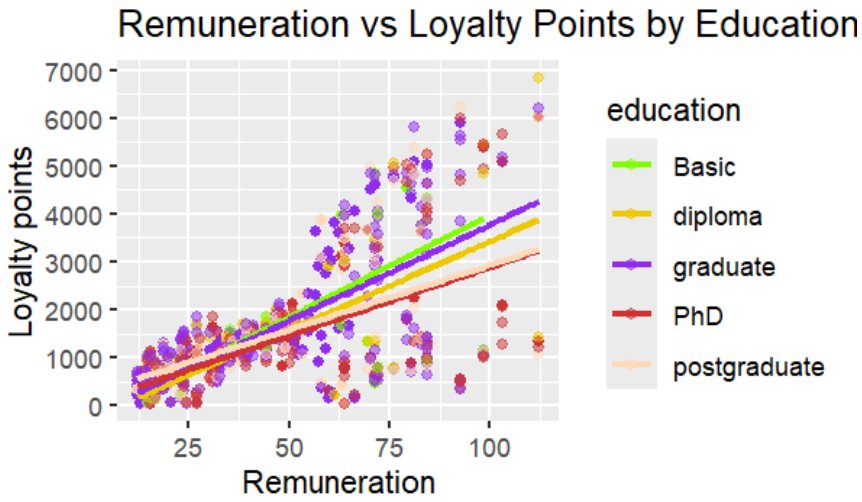
Figure 6: Polarity score for reviews and summaries.

Using R I investigated the distribution of the numerical variables in the data using histograms and bar charts. I then dove further into how the accumulation of loyalty points differed amongst different demographic groups.





It's clear to see that, while men and women both accumulate more loyalty points the more they earn, it is women that generally accumulate the most loyalty points. This may well be virtue of the fact that there are more women in the data. The trend reverses when looking at spending score, however.



When considering education, customers across all education levels see a greater accumulation of loyalty points when their remuneration increases. What's interesting is that the higher the level of education, the less pronounced this trend is. This presents an opportunity for Turtle Games' marketing department.

As we can see from the above plots none of the variables exhibit a normal distribution with most of them exhibiting a medium to heavy positive skew. The key observation here is the distribution of loyalty points, the dependent variable. Loyalty points exhibit a skew of 1.46 and a kurtosis of 4.71, as a result the performance of linear regression models may be impacted. The best approach to ensure a more accurate result in the future is to apply a transformation to the data or to use a non-linear model.

Conclusion

This analysis suggests that both spending score and remuneration levels have significant explanatory power when looking at predicting the accumulation of loyalty points. Based on this model, Turtle Games are accurately able to predict the number of loyalty points a customer will accumulate and highlight specific groups for targeted marketing campaigns increasing the likelihood of customer engagement.

Recommendations

Utilise customer feedback to improve product offering: The NLP model employed showed that there was some dissatisfaction with Turtle Games' products. By isolating and analysing the negative reviews, Turtle Games could address these issues and improve overall customer satisfaction.

Reform the review process: Many of the reviews returned a negative polarity score. By adding an additional layer in the review process (a likert scale for example) would enable Turtle Games to collect more quantifiable data that the company can lean on to inform future product offerings.

Customer loyalty: The predictive models employed suggest that remuneration significantly influences the number of loyalty points accumulated. This is further influenced by education level. By broadening their product offering, and targeting high income, highly educated individuals, Turtle Games can bring their customer loyalty in line with their other customers.

Targeted marketing campaigns: By employing cluster analysis, we can see that Turtle Games' customer base can be grouped into 5 clear segments. By curating and deploying targeted marketing campaigns based on each cluster, Turtle Games can more effectively convert leads into sales.

Appendices

Appendix 1

- How do customers engage with and accumulate loyalty points?
- How can customers be segmented into groups, and which groups can be targeted by the marketing department?
- How can text data (e.g. social data such as customer reviews) be used to inform marketing campaigns and make improvements to the business?
- Can we use descriptive statistics to provide insights into the suitability of the loyalty points data to create predictive models (e.g. normal distribution, skewness, or kurtosis) to justify the answer.)

Appendix 2

Cluster Characteristics:

- **Cluster 1**:
  - Average Remuneration: £40.12k
  - Average Spending Score: 45.67
  - Represents customers with moderate remuneration and spending scores.
- **Cluster 2**:
  - Average Remuneration: £75.32k
  - Average Spending Score: 20.45
  - Represents customers with high remuneration but low spending scores.
- **Cluster 3**:
  - Average Remuneration: £73.45k
  - Average Spending Score: 82.12
  - Represents high-value customers with both high remuneration and high spending scores.
- **Cluster 4**:
  - Average Remuneration: £20.34k
  - Average Spending Score: 79.45
  - Represents customers with low remuneration but high spending scores.
- **Cluster 5**:
  - Average Remuneration: £19.87k
  - Average Spending Score: 18.23
  - Represents customers with both low remuneration and low spending scores.

Business Implications:

High-Value Customers (Cluster 3)

- Customers in this cluster should be targeted with premium offers, loyalty programs, and personalized marketing to maximize revenue.

Low-Value Customers (Cluster 5)

- Customers in this cluster may require cost-effective strategies to improve engagement or may not be worth significant investment.

Intermediate Customers (Clusters 1, 2, and 4)

- These clusters represent potential growth opportunities. Tailored marketing strategies can help move these customers into higher-value segments.