Data Analysis 301

Technical Report: Data Analysis for Turtle Games

Prepared by: Tuhina Srivastava

Project: Loyalty Program Analysis and Recommendations

Table of Contents

Background/Context of the Business

Turtle Games is a global manufacturer and retailer of books, board games, video games, and toys. The company is focused on improving sales performance and enhancing customer engagement by leveraging data-driven insights. Key objectives include understanding loyalty points accumulation, segmenting customers based on spending behaviour and analysing product reviews to identify sentiment trends. By addressing these goals, Turtle Games aims to optimise its loyalty program, target marketing efforts effectively, and improve overall customer satisfaction. This analysis uses exploratory data analysis, clustering, and predictive modelling to uncover actionable insights and provide strategic recommendations.

Analytical Approach

Data Import and Cleaning

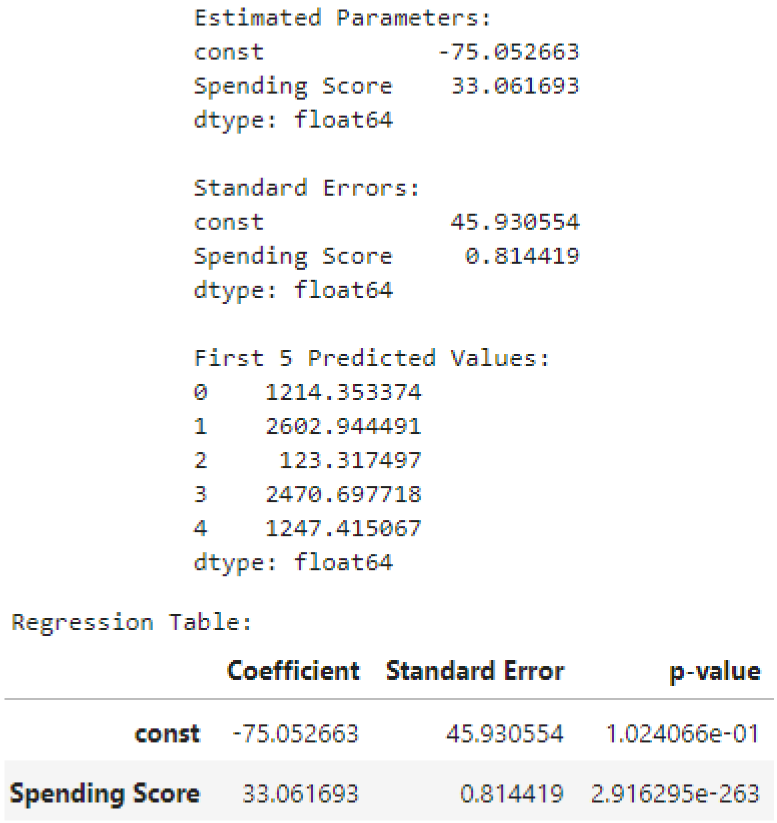
Data was imported into Python for preprocessing and analysis using libraries such as pandas, NumPy, and matplotlib. The raw data was first inspected to identify missing values, inconsistencies, and irrelevant columns. Columns like *language* and *platform* were removed due to their lack of relevance to the current analysis.

- **Renaming Columns:** Variable names were cleaned and standardised for consistency (e.g., Loyalty.Points was renamed to Loyalty Points).
- **Handling Missing Data:** Missing values in numerical variables were imputed using mean or median values, ensuring no information loss during analysis.
- **Outlier Detection:** Boxplots were created to detect outliers in critical variables like Spending Score and Loyalty Points. Outliers were retained for analysis as they provided insights into high-value customers.

Exploratory Data Analysis (EDA)

The EDA focused on understanding data distributions, relationships, and key trends:

- **Summary Statistics:** Mean, median, standard deviation, and skewness were calculated for numerical variables to evaluate central tendency and spread.

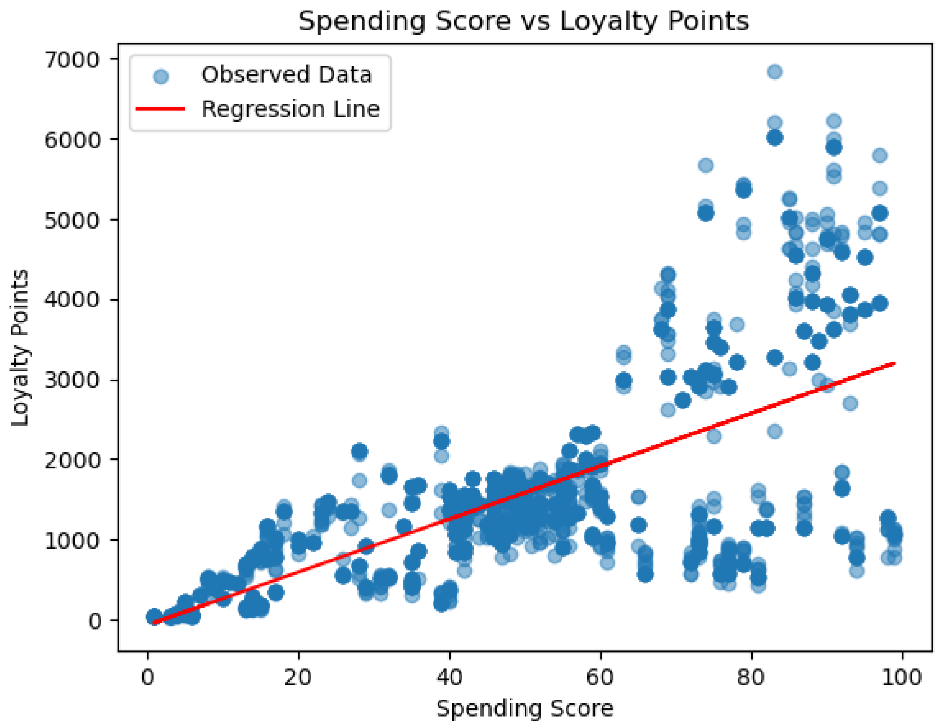


```
Estimated Parameters:
const                -75.052663
Spending Score        33.061693
dtype: float64

Standard Errors:
const                 45.930554
Spending Score         0.814419
dtype: float64

First 5 Predicted Values:
0    1214.353374
1    2602.944491
2     123.317497
3    2470.697718
4    1247.415067
dtype: float64
```

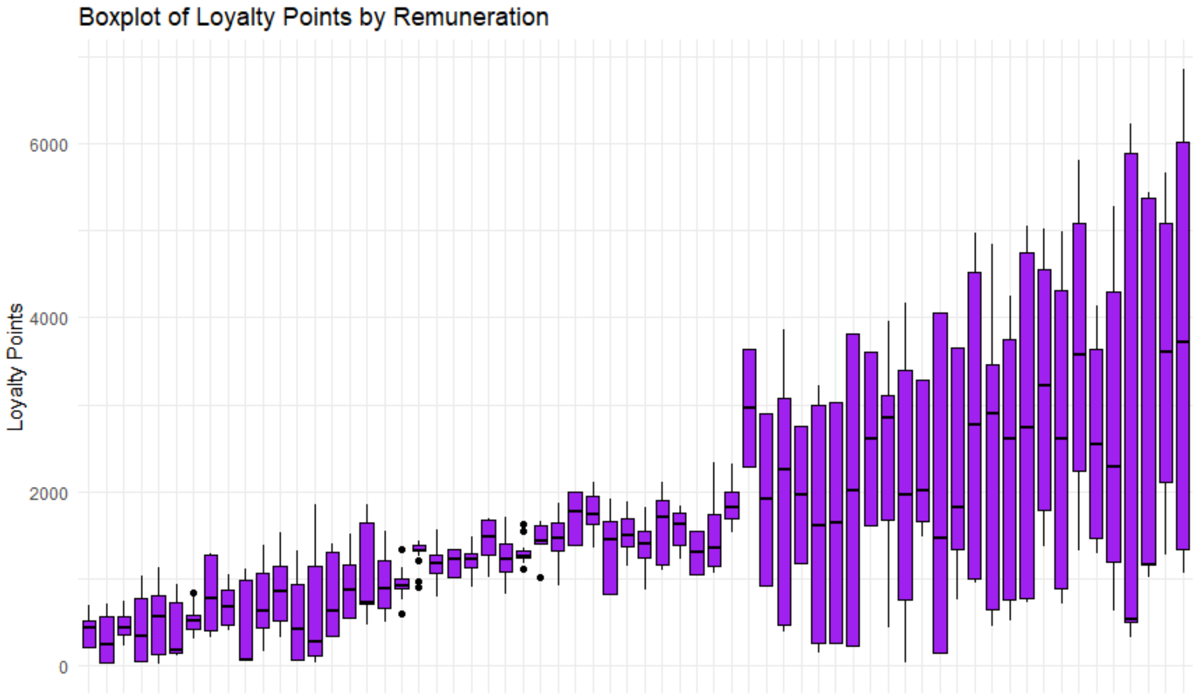
Regression Table:

	Coefficient	Standard Error	p-value
const	-75.052663	45.930554	1.024066e-01
Spending Score	33.061693	0.814419	2.916295e-263

- **Scatterplots:** A scatterplot between Spending Score and Loyalty Points revealed a strong positive correlation, suggesting customers with higher spending accumulate more loyalty points.

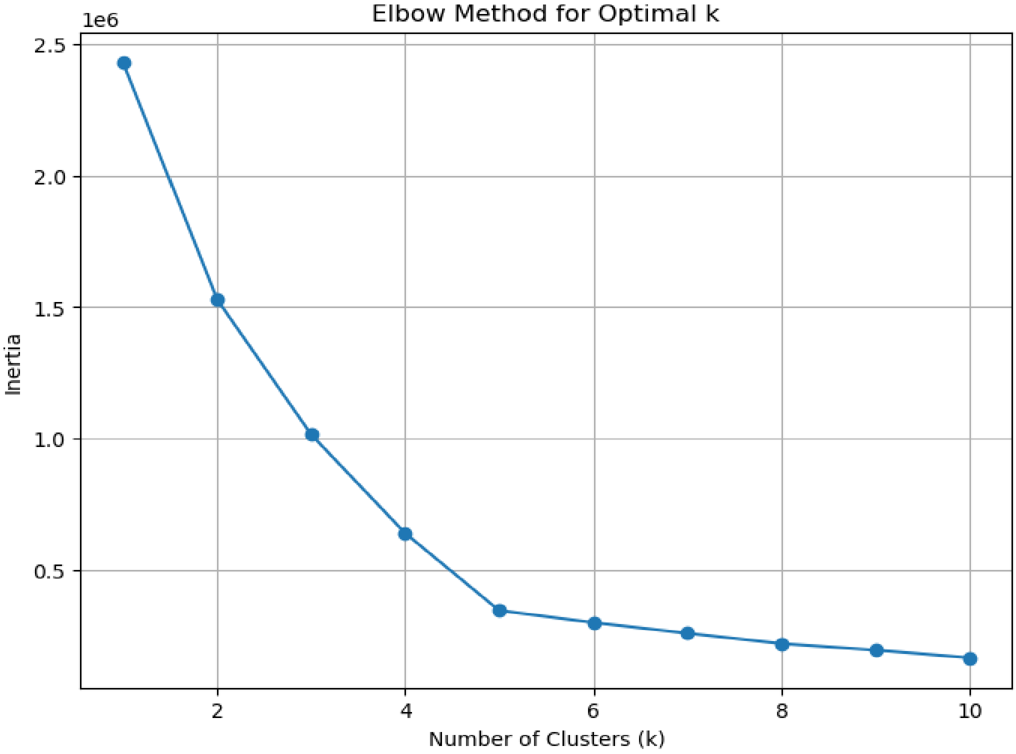


- **Histograms:** The histogram for Loyalty Points showed right-skewed data, indicating that a small number of customers earned significantly higher points.
- **Boxplots:** Boxplots for Loyalty Points by Remuneration highlighted significant variability, particularly among high-income groups.



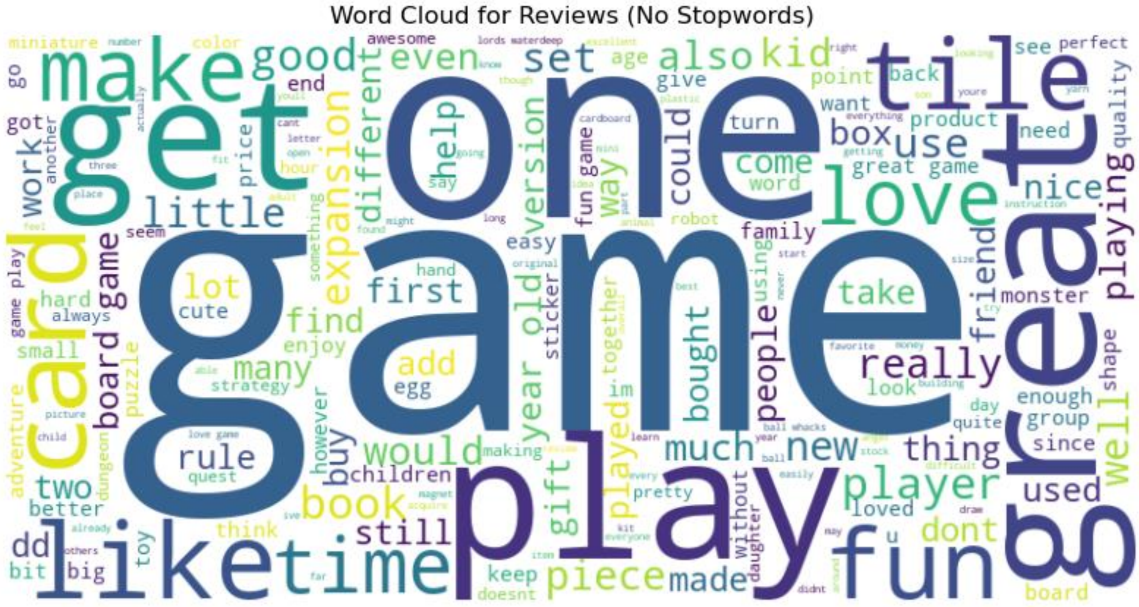
Statistical Modeling and Clustering

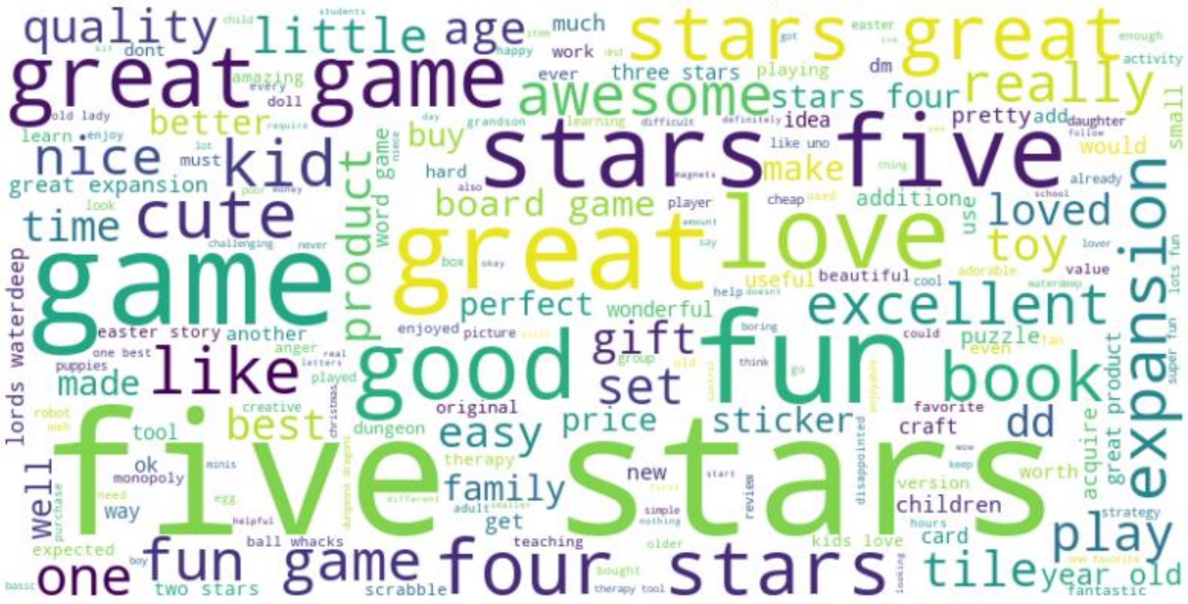
1. **Linear Regression:** A multiple linear regression (MLR) model was created in R using Spending Score, Remuneration, and Age as predictors. The model's goodness of fit was validated using R-squared (0.85) and Adjusted R-squared (0.84), demonstrating high predictive accuracy.
2. **K-Means Clustering:** Optimal clusters were determined using the elbow method, resulting in three customer segments:
   - Cluster 1: High spenders with high remuneration.
   - Cluster 2: Moderate spenders with medium remuneration.
   - Cluster 3: Low spenders with low remuneration.



Sentiment Analysis

NLP techniques using TextBlob were applied to product reviews and summaries. Polarity scores were calculated to measure sentiment, and word clouds were created to visualise frequent terms. The results revealed slightly positive sentiment overall.





Word Cloud for Summaries (No Stopwords)

Visualisations and Insights

Key Visualisations

1. **Scatterplot: Spending Score vs Loyalty Points**
   - The scatterplot showed a strong positive correlation between Spending Score and Loyalty Points, confirming that customers who spend more tend to accumulate higher loyalty points.
2. **Histogram: Loyalty Points Distribution**
   - The histogram revealed that most customers accumulate fewer than 2,000 loyalty points, while a small group earned significantly higher points, contributing to the skewness.
3. **Boxplot: Loyalty Points by Remuneration**
   - The boxplot highlighted variability in loyalty points across income levels. Higher earners consistently achieved greater loyalty points, indicating income as a significant factor.
4. **Regression Line: Spending Score vs Loyalty Points**
   - A scatterplot with a regression line reinforced the linear relationship between Spending Score and Loyalty Points. The positive slope validates the importance of spending as a predictor.
5. **Predicted Loyalty Points**
   - Bar charts illustrated loyalty points predictions for two customer scenarios, demonstrating the model's practical application.

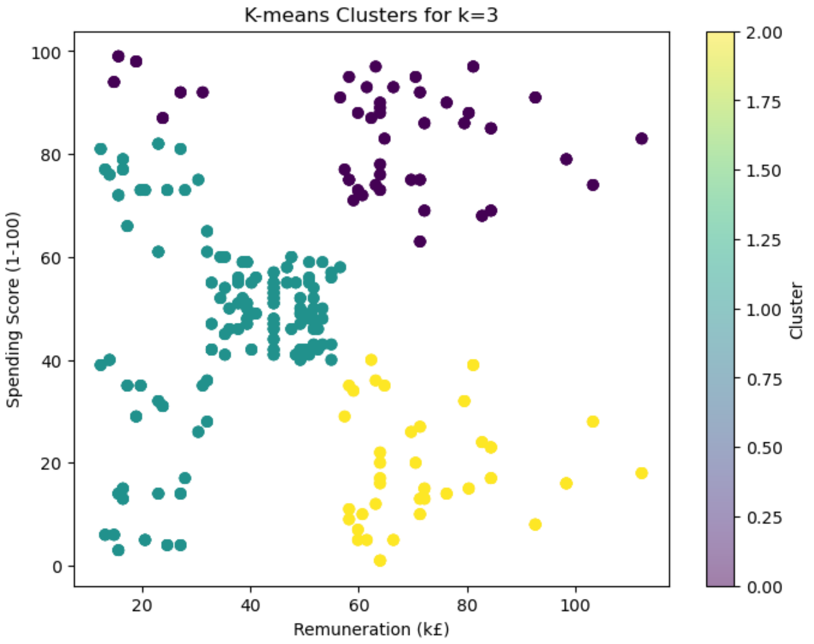
Insights

- **Spending behaviour:** Customers with higher spending scores earn significantly more loyalty points, confirming spending as the strongest predictor.
- **Income segmentation:** High-income customers (Cluster 1) are key contributors to loyalty points, suggesting targeted premium offers.
- **Sentiment trends:** Product reviews were slightly positive, with common themes including satisfaction with product quality and minor dissatisfaction with utility.

Patterns and Predictions

Discovered Patterns

1. **Spending Score Correlation:** Spending Score showed the strongest correlation ($R^2 = 0.452$) with loyalty points, reinforcing its role as the primary driver.
2. **Income Impact:** Remuneration moderately correlated with loyalty points ($R^2 = 0.38$), with higher earners contributing disproportionately to loyalty accumulation.
3. **Age Negligibility:** Age had minimal impact on loyalty points ($R^2 = 0.002$), suggesting it can be excluded from predictive models.
4. **Customer Segments:** K-means clustering identified three groups:
   - **Cluster 1:** High spenders and high earners (prime targets for premium campaigns).
   - **Cluster 2:** Moderate spenders (opportunities for incentives).
   - **Cluster 3:** Low spenders (potential engagement through discounts).



Predictions

The multiple linear regression model achieved an R-squared value of 0.85, confirming its reliability in predicting loyalty points. Spending Score emerged as the most significant predictor, followed by Remuneration.

Business Relevance

1. **Loyalty program:** Introduce tiered rewards based on spending and remuneration to drive engagement.
2. **Marketing focus:** Prioritise high-value customers (Cluster 1) and incentivise moderate spenders (Cluster 2).
3. **Product strategy:** Address areas of dissatisfaction highlighted in negative reviews to improve customer experience.