Advanced Analytics for Organizational Impact

Khristian Mark Alvarez*, Data Analyst*

Project Scope

Turtle Games is a global manufacturer and retailer of board games, video games, and toys. The goal of the company is to enhance its sales performance by leveraging insights from its robust data collection system, which comprises information from its customer accounts and reviews. This project looks into the customer demographics and sales-related data to uncover patterns in loyalty accumulation, identify key market segments, assess the influence of customer feedback, and evaluate reliability.

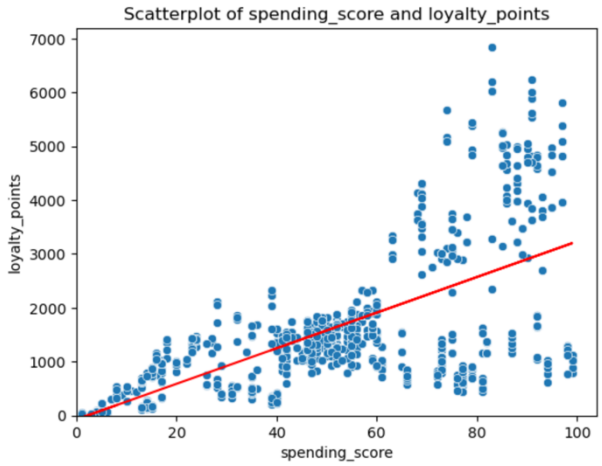
Analytical Approach

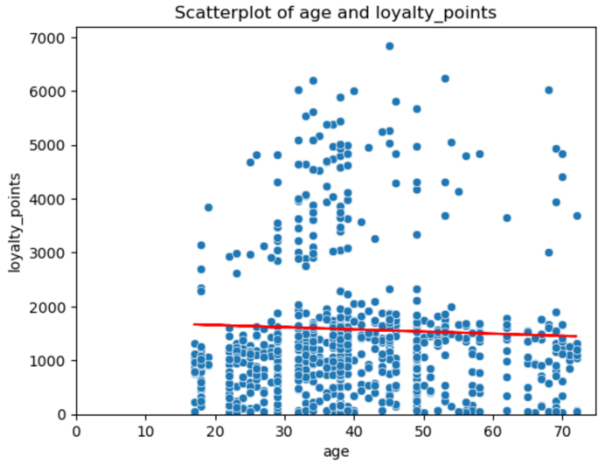
The initial phase of the analysis on the turtle_reviews.csv dataset was conducted in ***Python*** using Jupyter Notebook. All imported data sets were cleaned through sense-checking data types, finding missing values, and detecting duplicates. An examination of the descriptive statistics for the metadata was also observed to provide an overview of the dataset's structure and distribution. The analysis utilized several key libraries such as NumPy, Pandas, Matplotlib, Seaborn, SciPy, Statsmodels, and Scikit-learn in Python.

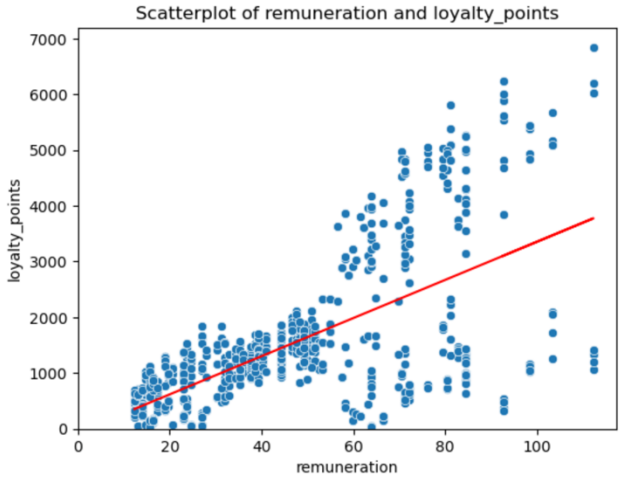
Further exploration in ***R*** was employed during the second phase of the data analysis, where Tidyverse, Dplyr, Ggplot2, Skimr, DataExplorer, NbClust, FactorExtra, Psych, and Moment were all utilized to investigate the distribution and relationship of the variables. Subsequent visualizations were then developed through histograms and scatterplots to illustrate these links. A disaggregation of data was also performed to analyze the data according to categorical values, specifically gender and education level, to uncover potential group-level variations.

*Data Analysis in **Python***

Prior to using the Multiple Linear Regression model, individual simple linear regressions were first conducted for each independent variable against Loyalty Points.







These regression models indicate that spending score, remuneration, and age explain approximately 45%, 38%, and 0.02% of the variation in the loyalty point accumulation, respectively. Based on these results, a one-unit increase in spending score, or remuneration corresponds to an estimated increase of 33 and 34 loyalty points, respectively, while a one-unit increase in age is associated with a decrease of about 4

loyalty points. Hence, the visuals show that there is a clear positive correlation between spending score / remuneration and loyalty points, whereas age shows a very weak correlation.

Consequently, a multiple linear regression was built using statsmodels function to explore these similar linear associations. The correlation coefficients for the variables are displayed below:

	age	remuneration	spending_score	loyalty_points
age	1.000000	-0.005708	-0.224334	-0.042445
remuneration	-0.005708	1.000000	0.005612	0.616065
spending_score	-0.224334	0.005612	1.000000	0.672310
loyalty_points	-0.042445	0.616065	0.672310	1.000000

The correlation results therefore show that:

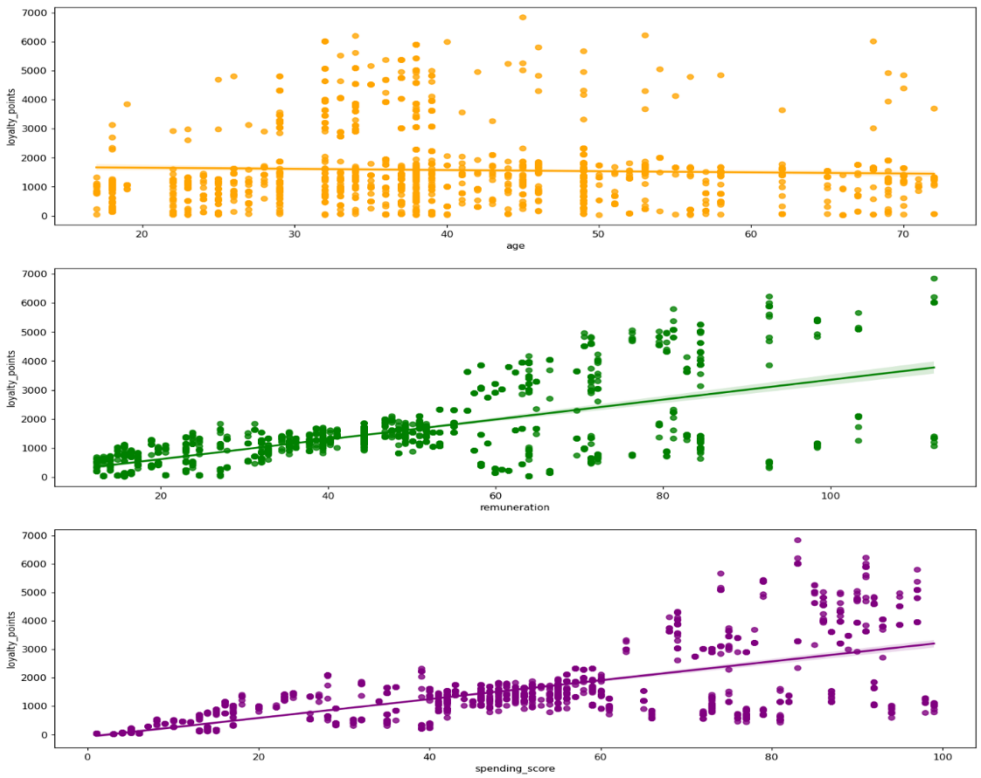
Age and Remuneration (-0.0057): Very weak, almost no relationship — age has little effect on earnings.

Remuneration and Spending Score (0.0056): Very weak positive link — income doesn't strongly relate to spending habits.

Age and Loyalty Points (-0.0424): Very weak negative link — age barely affects loyalty points.

Spending Score and Loyalty Points (0.6723): Moderately strong positive link — higher spenders tend to have more loyalty points, suggesting frequent or loyal customers.

Remuneration and Loyalty Points (0.6161): Moderately strong positive link — higher earners often have more loyalty points.

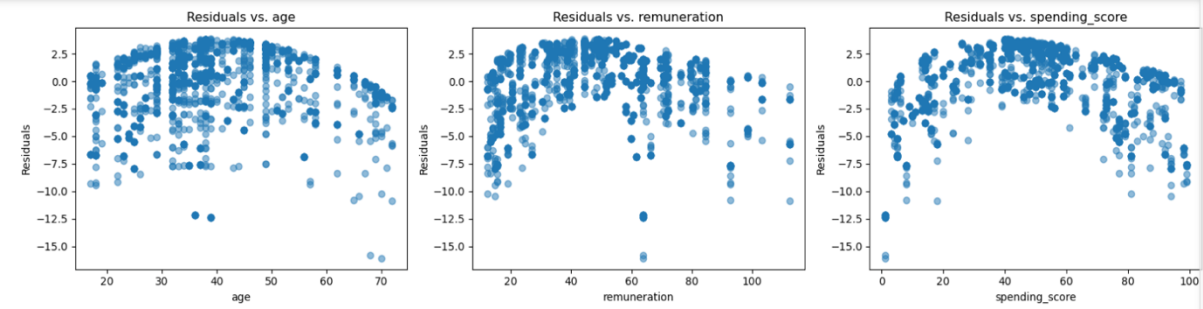


These results indicate that remuneration and spending score exhibit a significantly positive relationship with loyalty points as compared to age which shows a weaker correlation with it, much like of what the simple linear regression also outlined.

Furthermore, after an examination of the distribution of the original loyalty points using a histogram, Q-Q plot, and the Shapiro-Wilk test, a significant deviation from normality was observed. Although normality of residuals is one of the regression assumptions, it is generally less critical than others, and such deviations can be acceptable when the sample size is sufficiently large.

An OLS regression was also performed, where key assumptions such as linearity, homoscedasticity, and normality of residuals were evaluated. This resulted to both the original and log-transformed loyalty models showing violations of linearity and homoscedasticity. To mend these issues, a box-cox transformation was applied.

Despite these improvements, the Breusch-Pagan test indicated heteroscedasticity in the residuals (as shown below), meaning that residual variance varied across different values of the independent variables, thereby possibly affecting the reliability and interpretation of the regression outcomes.

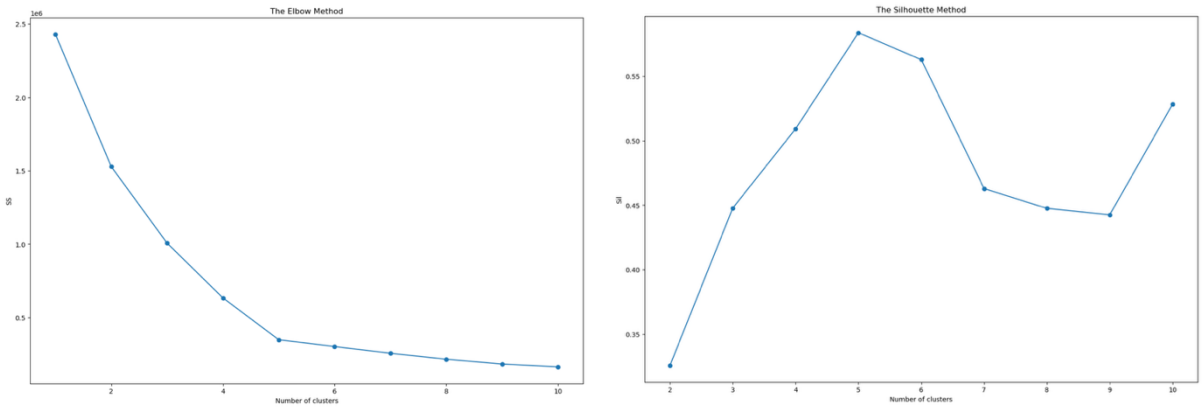


Nevertheless, The Box-Cox transformed OLS model showed improved R-squared values and performance, but heteroscedasticity remains a concern, limiting its predictive reliability for the turtle game's objectives. Outlier analysis was also excluded due to unclear stakeholder definitions.

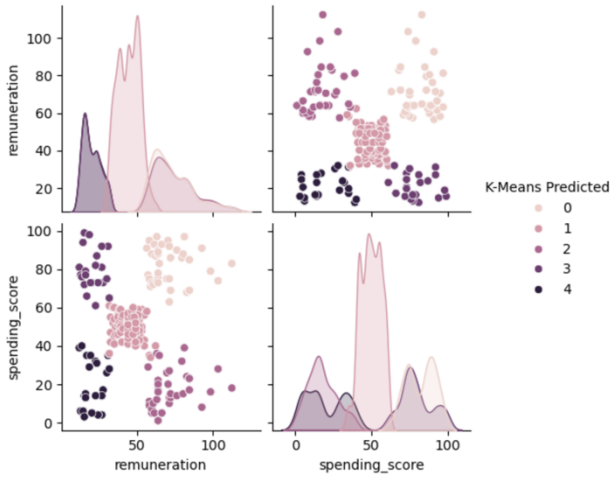
*Clustering with **k-means***

A structured K-means clustering process was also applied, beginning with data exploration and visualization, followed by the Elbow and Silhouette methods to identify the optimal cluster count. The analysis suggests that five (5) clusters offer a balanced and interpretable segmentation, enabling targeted marketing strategies—such as

discounts for low earner–high spender groups and personalized campaigns for high earner–low spender segments.



Customer segmentation is crucial for the marketing team to design targeted strategies instead of broad, costly campaigns that may deliver limited results.

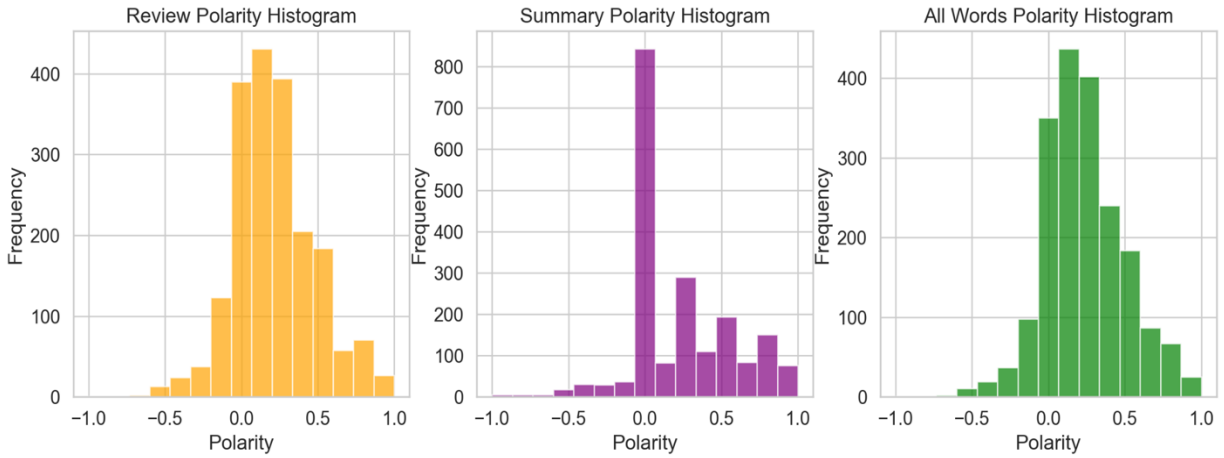


Sentiment Analysis for Customer Reviews

Customer reviews were processed using NLTK, including text cleaning, tokenization, and sentiment analysis. Word clouds, frequency distributions, and sentiment histograms were generated, revealing overall positive sentiment with common words

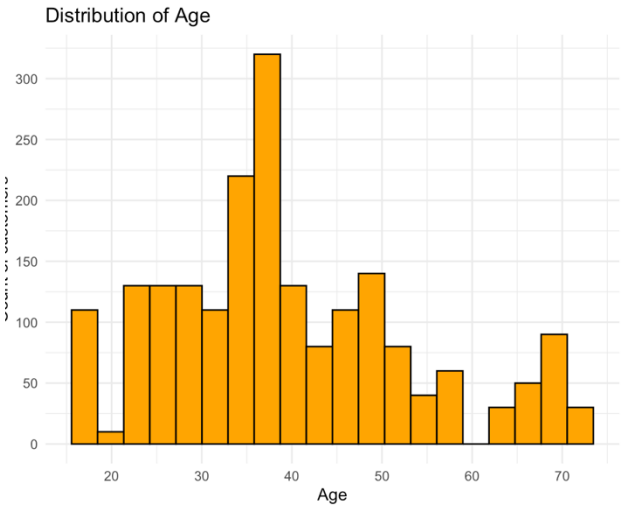
like *great*, *fun*, *excellent*, and *five stars*. The top 20 positive and negative reviews provided further insight into customer perceptions across products and platforms.

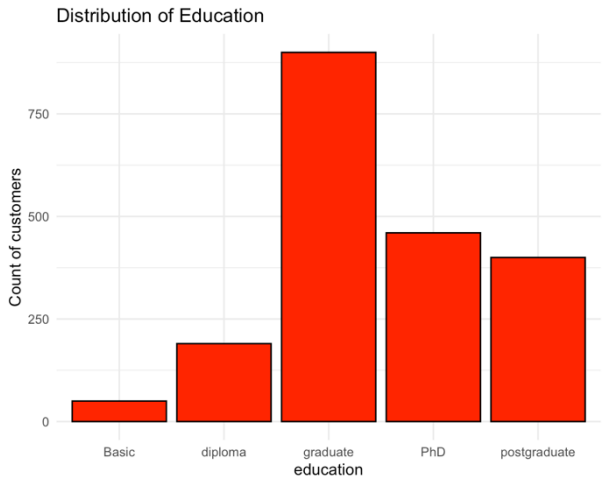
Sentiment analysis showed mean polarity scores of 0.21 (reviews) and 0.22 (summaries), indicating generally positive sentiment with moderate variability. Quartile values confirmed predominantly positive tone. Common negative themes included “unclear instructions” and “faulty components”, suggesting opportunities for Turtle Games to enhance product quality and personalize support for dissatisfied customers.

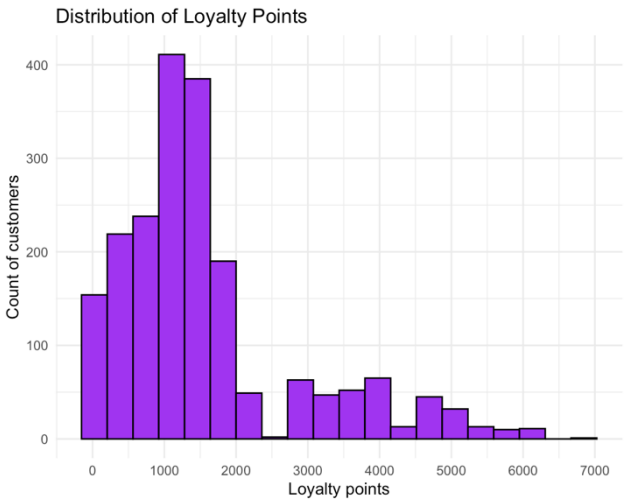


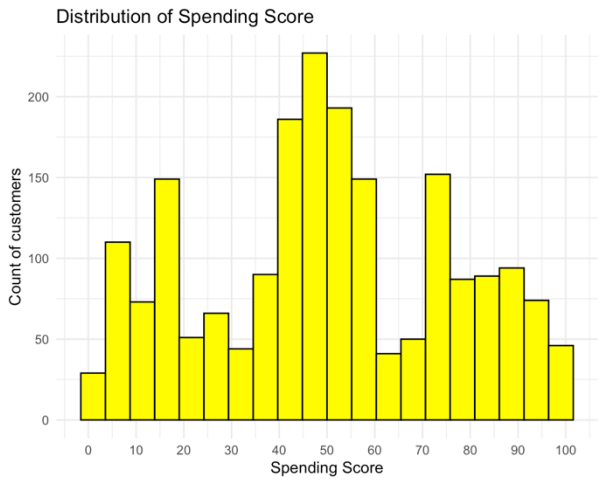
*Data Analysis in **R***

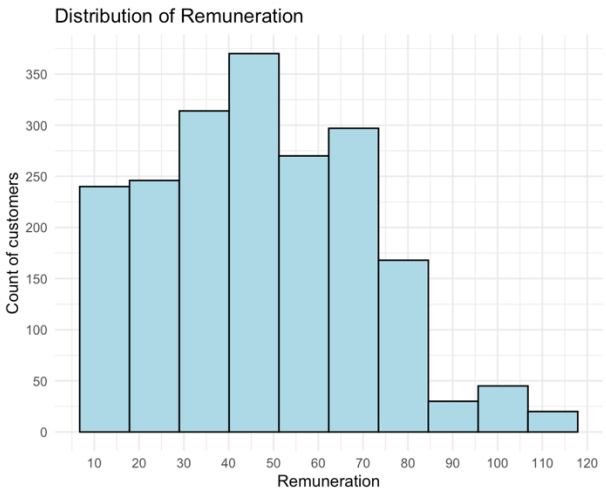
Using R, an exploration of the distribution of numerical variables was done with histograms and bar charts, then examined how loyalty point accumulation varied across demographic groups.



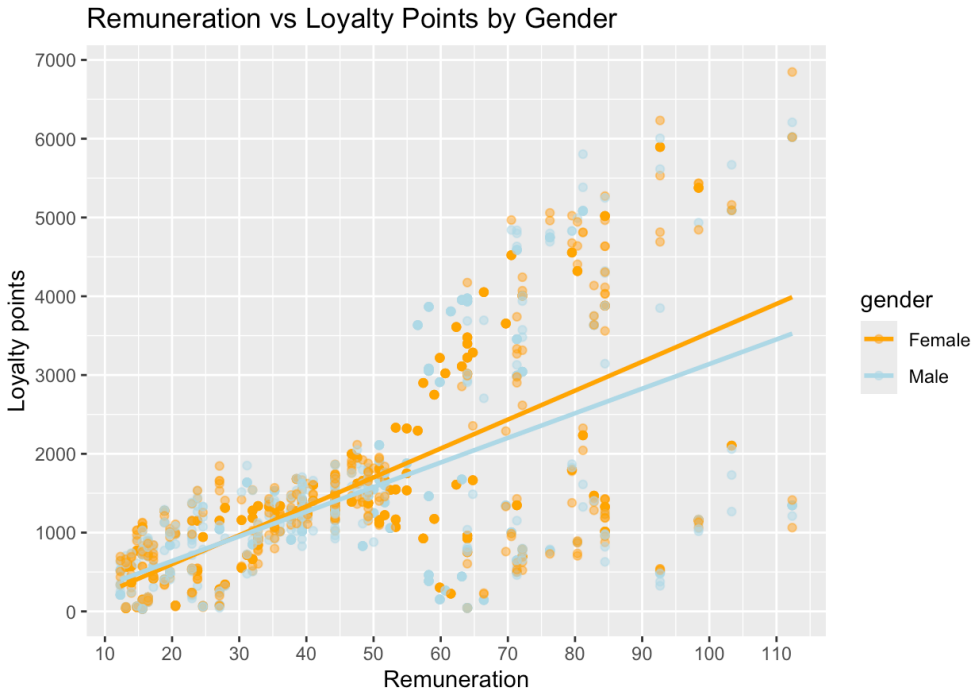




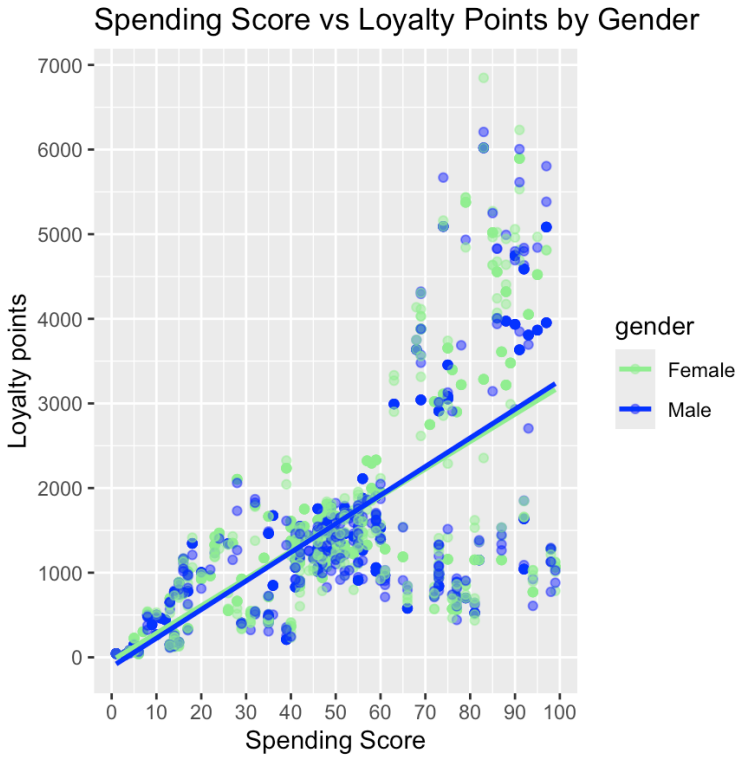




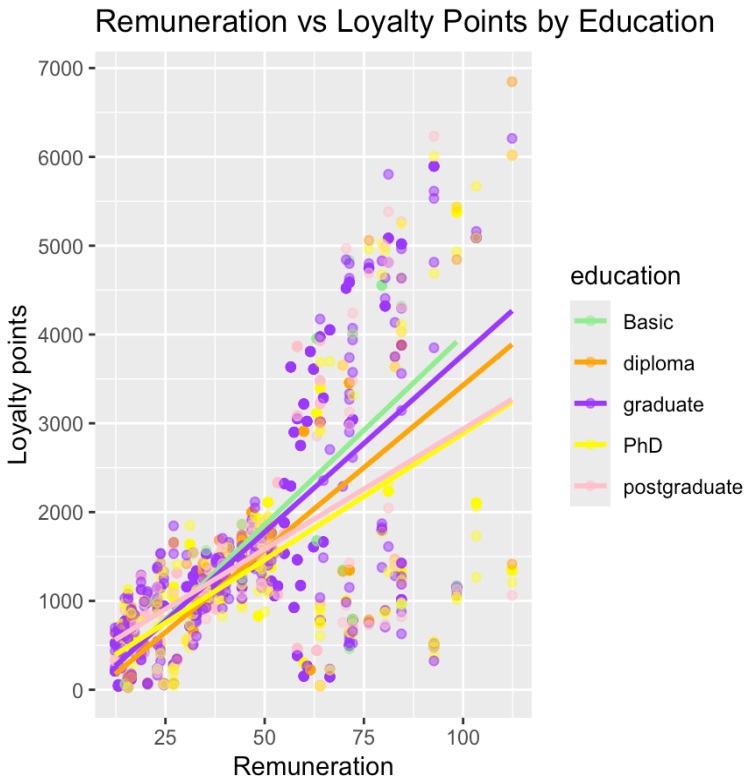
Both genders showed higher loyalty points with increasing income, though women generally accumulated more, likely due to greater representation in the data.



The trend, however, is reversed for spending scores.



Higher income increased loyalty points across education levels, though the effect lessened with higher education—suggesting a marketing opportunity for Turtle Games. Loyalty points were positively skewed (skew = 1.46, kurtosis = 4.71), indicating that transformations or non-linear models could improve regression accuracy.



Conclusion

The analysis demonstrates that both spending score and remuneration possess significant explanatory power in predicting loyalty points accumulation. This model enables Turtle Games to generate more accurate forecasts of customer loyalty behavior and to identify distinct customer segments for targeted marketing initiatives, thereby enhancing overall engagement effectiveness.

Recommendations

NLTK sentiment analysis revealed generally positive feedback but highlighted several issues such as “unclear instructions” and “faulty components”. Thus, structured reviews (e.g., Likert scales) could improve feedback quality. Predictive models showed income and education influence loyalty, while cluster analysis identified five segments, therefore presenting an opportunity for Turtle Games to refine products, target marketing, and boost customer engagement.