London School of Economics

Data Analytics for Business

Advanced Analytics for Organisational Impact

Understanding Customer Loyalty Points Accumulation at Turtle Games

Word Count: 967

Date: 26/Apr/2024

Introduction

This report outlines the analytical approach and findings aimed at identifying factors that can improve overall performance for Turtle Games. Through a comprehensive analysis of customer trends, including loyalty points accumulation, segmentation, text data analysis, sales trends, and descriptive statistics evaluation, actionable insights are provided to enhance marketing strategies and customer satisfaction.

Analytical Approach – Python

The analysis was executed using both Python and R programming languages to address the following questions:

- 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.)

EDA and (Multiple) Linear Regression Analysis

Explored relationships between loyalty points and customer demographics. Identified moderate positive correlations with remuneration/spending scores and a potential negative correlation with age.

MLR uncovered that the model built using `spending score`, `remuneration`, `age`, `education`, and `gender` explains 84.5% of the variance in loyalty points for the training data. This was further investigated in R. Spending score has the strongest positive impact on loyalty, followed by remuneration, age, and education. Gender surprisingly shows a negative association with loyalty points, requiring further investigation and consideration of industry demographics.

Decision Trees for Exploring Data Structure

Utilised decision tree regressor (DCR) to comprehend data structure and analyse performance metrics. Initial model exhibited significant overfitting and captured lots of noise. Compared pruned DCR to GridSearch DRC and RandomForest, the latter achieved lowest RMSE-MAE, suggesting it is the best choice for this dataset.

Uncovered hierarchical decision rules for loyalty points accumulation: `spending_score`, `remuneration`, `age` are the most dominant features.

Clustering with k-means

Employed *k*-means clustering to segment customers for targeted marketing. `remuneration` and `spending score` were selected as key features for clustering, based on previous analysis.

Findings for clusters (business recommendations in presentation):

- **Cluster 1 (Largest):** Moderate income, average spending (likely core customer base).
- **Cluster 0:** Potentially young adults/students with high spending score (frequent but lower value purchases).
- **Cluster 2:** Moderate income, lower spending frequency (unclear demographics due to missing age data).
- **Cluster 4:** Slightly higher income, lower spending frequency/value.
- **Cluster 3:** High income, frequent spenders (high-value segment).

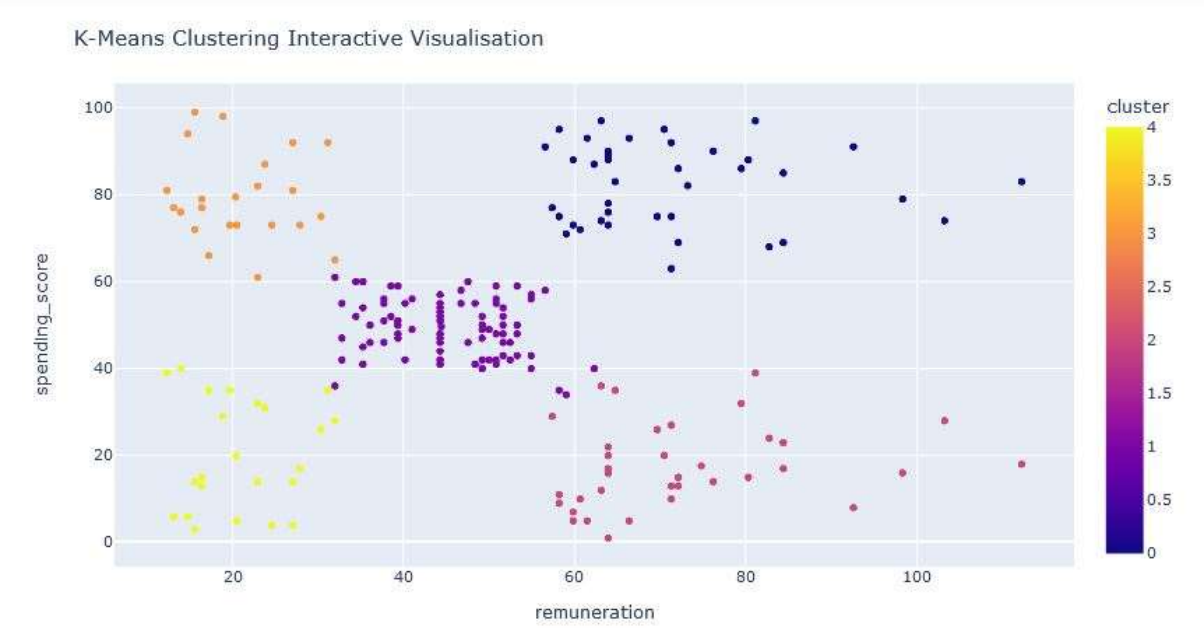


Fig.: distinct customer groups based on `remuneration` *and* `spending score`

NLP for Customer Review Analysis

Applied NLP techniques to analyse sentiment and common terms in customer `review` and `summary`. All characters were converted to lowercase and punctuation was removed to improve analysis consistency and tokenization. Duplicates were reviewed and kept to avoid losing valuable information, or skewing outcomes. Word frequency and distribution was investigated. The reviews with the most positive and negative sentiment polarity scores were identified for further analysis by the marketing team.

Conclusion and Recommendations

Targeted marketing efforts should focus on high-spending customer segments. Loyalty programme incentives should be refined based on demographic and spending behaviour analysis. Continuous monitoring of customer behaviour is crucial for ongoing improvement.

Analytical Approach – R

This report presents findings from exploratory data analysis (EDA) and the development of a multiple linear regression (MLR) model using R. The objective was to understand customer behaviour, evaluate how loyalty points are accumulated, and test if it is possible to predict them. Python findings were cross-referenced and tested further, and additional investigation focussed also on '`gender`' and '`education`', while the latter two do not appear to have a significant impact on loyalty_points accumulation, they are important metrics to understand the customer base and industry demographics.

Exploratory Data Analysis

EDA was conducted to understand the distribution, patterns, and relationships within the data, and revealed significant differences in loyalty points based on gender and education level, suggesting areas for targeted marketing efforts. Various visualisations such as scatter plots, histograms, and boxplots were created to explore relationships between variables and identify potential trends or outliers.

Boxplots were used to compare spending scores and loyalty points by gender and education level. Significant differences were observed in spending scores and loyalty points across different gender and education groups, suggesting potential areas for targeted marketing efforts.

Statistical Analysis

Statistical tests, including ANOVA, were performed to further investigate relationships between variables, which indicated significant differences in loyalty points based on education level, with post-hoc tests revealing specific group differences.

```
$education
                          diff         lwr        upr     p adj
diploma-basic         -929.0189 -1482.45675 -375.58115 0.0000478
graduate-basic        -598.9822 -1104.90071  -93.06374 0.0109254
PhD-basic             -765.2900 -1283.78691 -246.79309 0.0005543
postgraduate-basic    -765.9625 -1288.25797 -243.66703 0.0006167
graduate-diploma       330.0367    52.03967  608.03378 0.0106090
PhD-diploma            163.7289  -136.55075  464.00864 0.5699684
postgraduate-diploma   163.0564  -143.73568  469.84857 0.5945253
PhD-graduate          -166.3078  -365.87733   33.26178 0.1533410
postgraduate-graduate -166.9803  -376.22062   42.26007 0.1881057
postgraduate-PhD        -0.6725  -238.72093  237.37593 1.0000000
```

Fig.: TukeyHSD output for 'education', as ANOVA was significant.

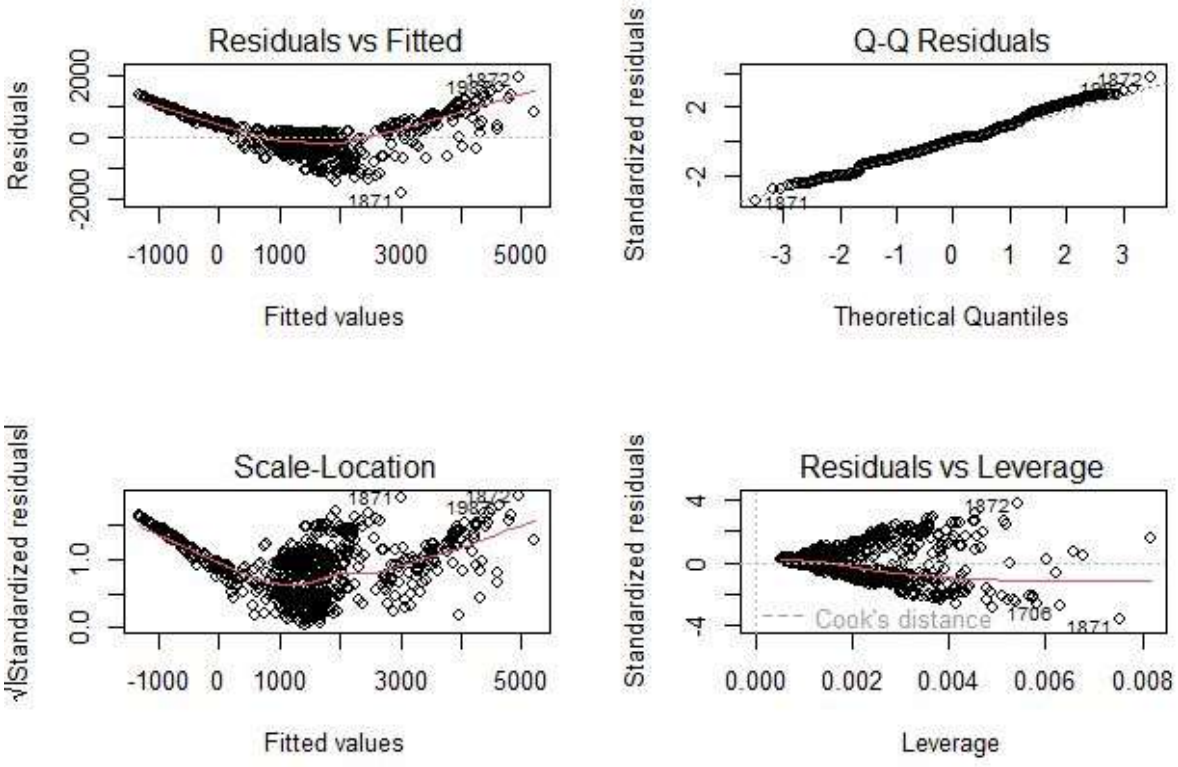
Multiple Linear Regression Modelling

A multiple linear regression model was developed to predict loyalty points based on age, remuneration, and spending score. The model's goodness of fit was evaluated using R-squared and adjusted R-squared values, indicating a strong fit to the data. Residual analysis confirmed the model's validity, and hypothetical scenarios were created to demonstrate its predictive capability. MLR for `spending_score`, `remuneration` and `age` shows 83.9% predictive capacity.

Model evaluation

The model summary `(summary(model))` provided information on coefficients, p-values, R-squared, and other statistics to assess model fit. Goodness-of-fit was evaluated using R-squared (adjusted and unadjusted), F-statistic, and p-value.

Residual plots were generated to visually inspect patterns or trends in the residuals.



Additionally, the Shapiro-Wilk test was conducted to assess the normality of residuals. The test results rejected the null hypothesis of normality, indicating non-normality.

Conclusion and Recommendations

Insights gained from the analysis can inform targeted marketing efforts and refine the loyalty program to enhance customer satisfaction and sales performance. The analysis revealed valuable insights into customer behaviour and the effectiveness of the loyalty program. Targeted marketing efforts should focus on customer segments with higher spending scores and loyalty points.

The multiple linear regression model provides a reliable method for predicting loyalty points based on available features. However, further refinement and validation is necessary for real-world application. Recommendations for improving the loyalty program include adjusting incentives based on customer demographics and spending behaviour and conducting additional analysis to identify potential areas for improvement.

Customer Loyalty Analysis Summary

Engagement and Accumulation

- Higher spending leads to more loyalty points (likely through purchases).
- Investigate programme details and customer journeys to fully understand engagement.

Segmentation and Targeting

- Segment by spending and consider income, demographics, game preferences, and engagement level.
- Target segments with personalised offers, communication, and loyalty programme adjustments.

Utilising Text Data

- Analyse customer reviews (sentiment and common themes) to inform marketing and business improvements.
- Leverage social media data for brand perception and loyalty programme insights.

Predictive Model Suitability

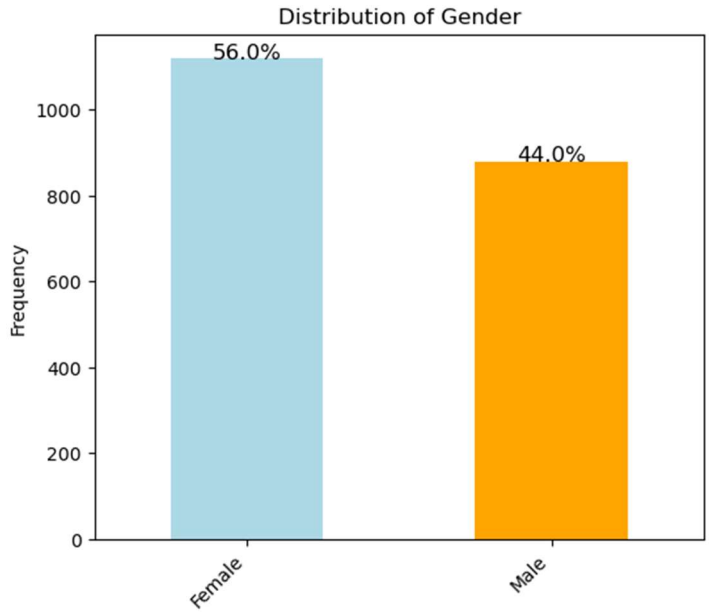
- While residuals show non-normality, the model might still predict well.
- Analyse loyalty point distribution and outlier impact for further assessment.
- Consider alternative models if necessary.

This analysis provides valuable insights for optimising Turtle Games' loyalty programme and marketing strategies.

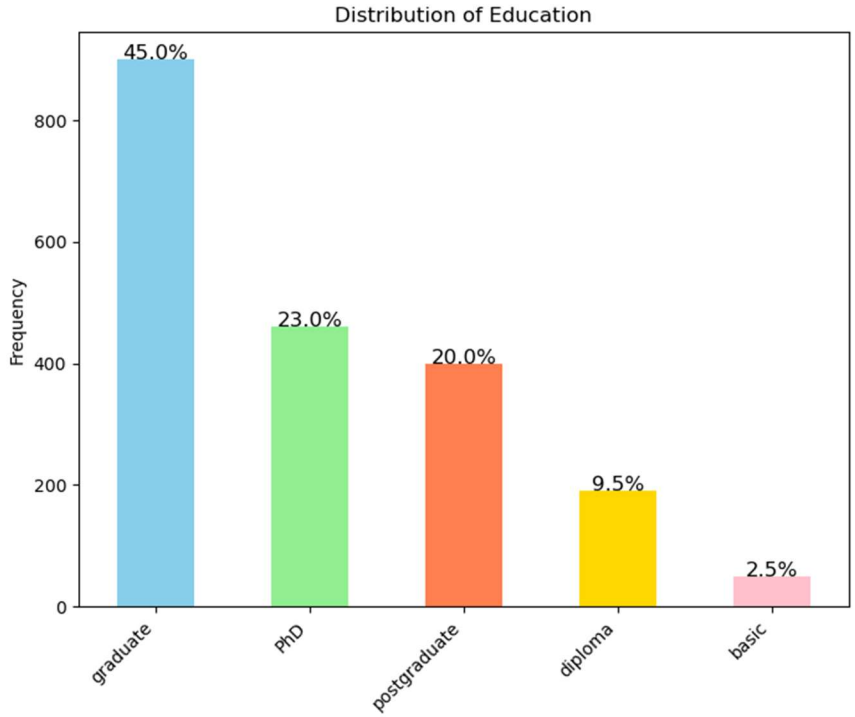
Appendix

Python visualisations

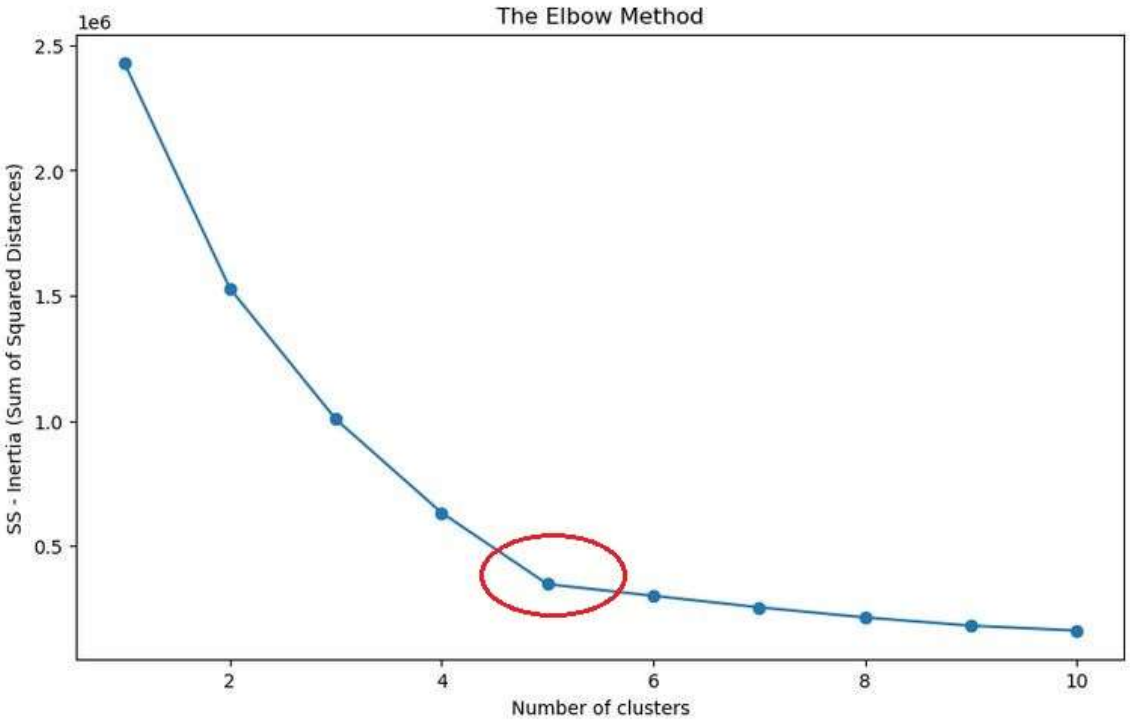
Distribution of Gender



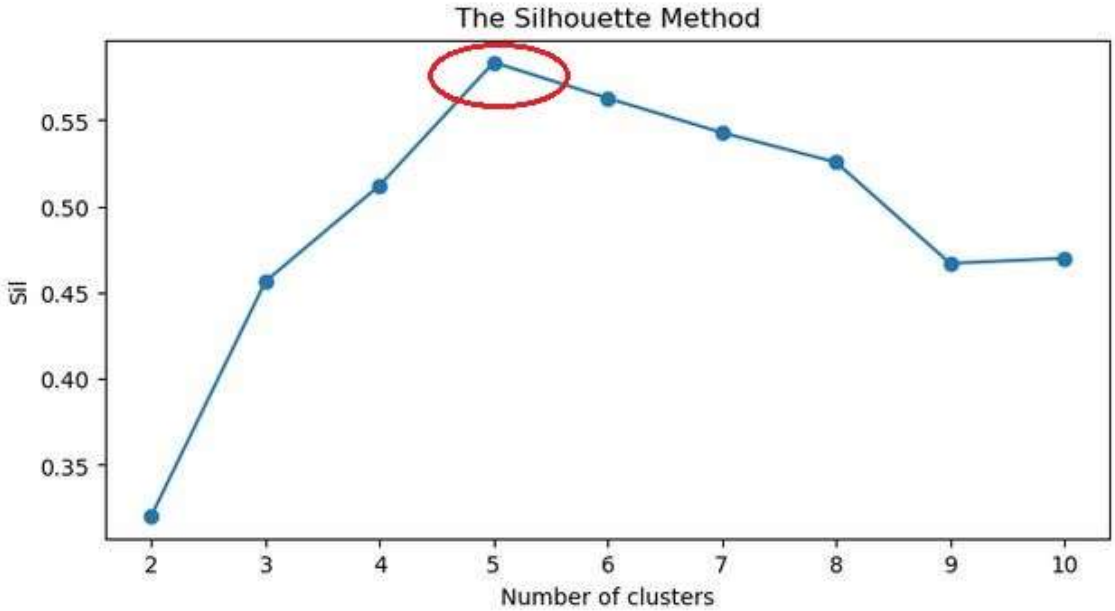
Distribution of Education



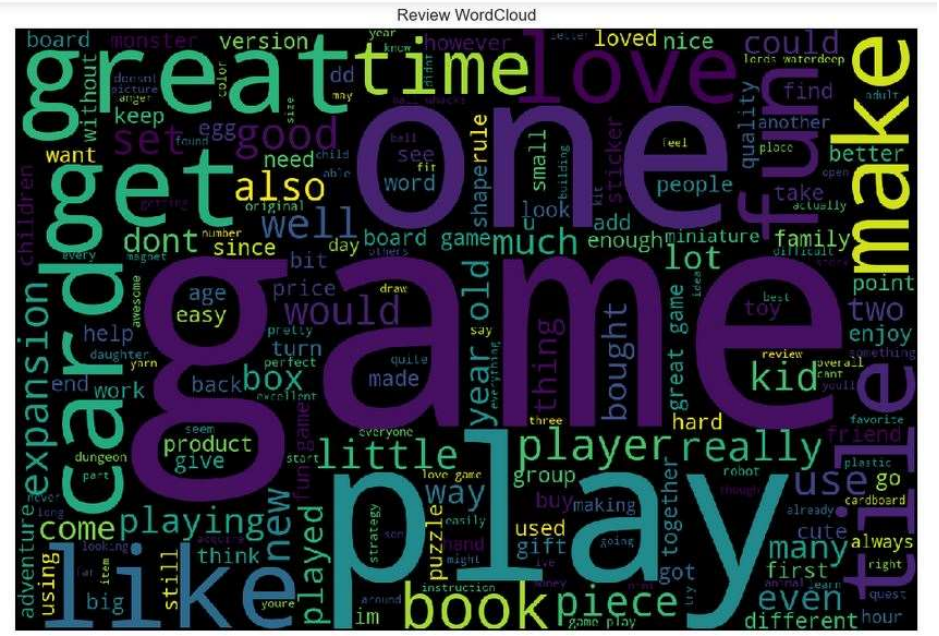
Elbow method



Silhouette method



WordCloud for Review



R visualisations

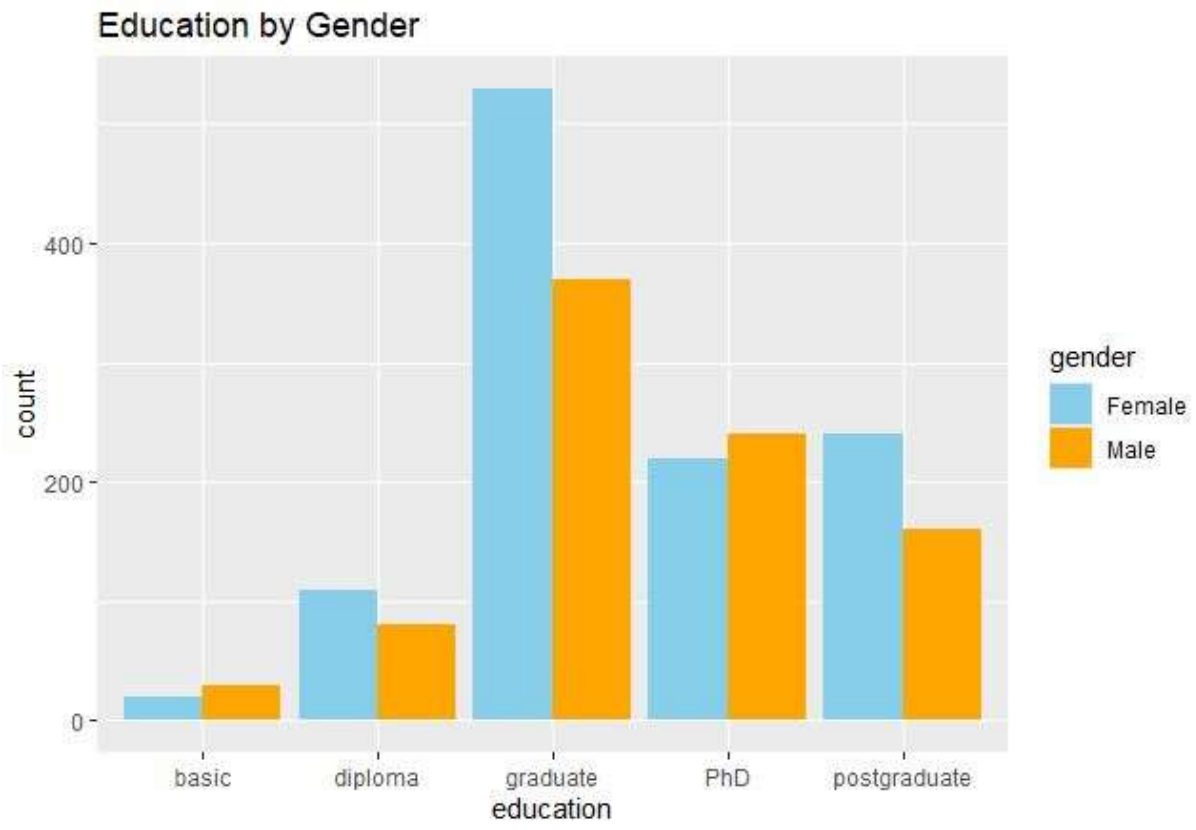
Summary Statistics

```
    gender               age          remuneration    spending_score loyalty_points
 Length:2000        Min.   :17.00   Min.   : 12.30   Min.   : 1     Min.   :  25
 Class :character   1st Qu.:29.00   1st Qu.: 30.34   1st Qu.:32     1st Qu.: 772
 Mode  :character   Median :38.00   Median : 47.15   Median :50     Median :1276
                    Mean   :39.49   Mean   : 48.08   Mean   :50     Mean   :1578
                    3rd Qu.:49.00   3rd Qu.: 63.96   3rd Qu.:73     3rd Qu.:1751
                    Max.   :72.00   Max.   :112.34   Max.   :99     Max.   :6847
  education            product
 Length:2000        Min.   :  107
 Class :character   1st Qu.: 1589
 Mode  :character   Median : 3624
                    Mean   : 4321
                    3rd Qu.: 6654
                    Max.   :11086
```

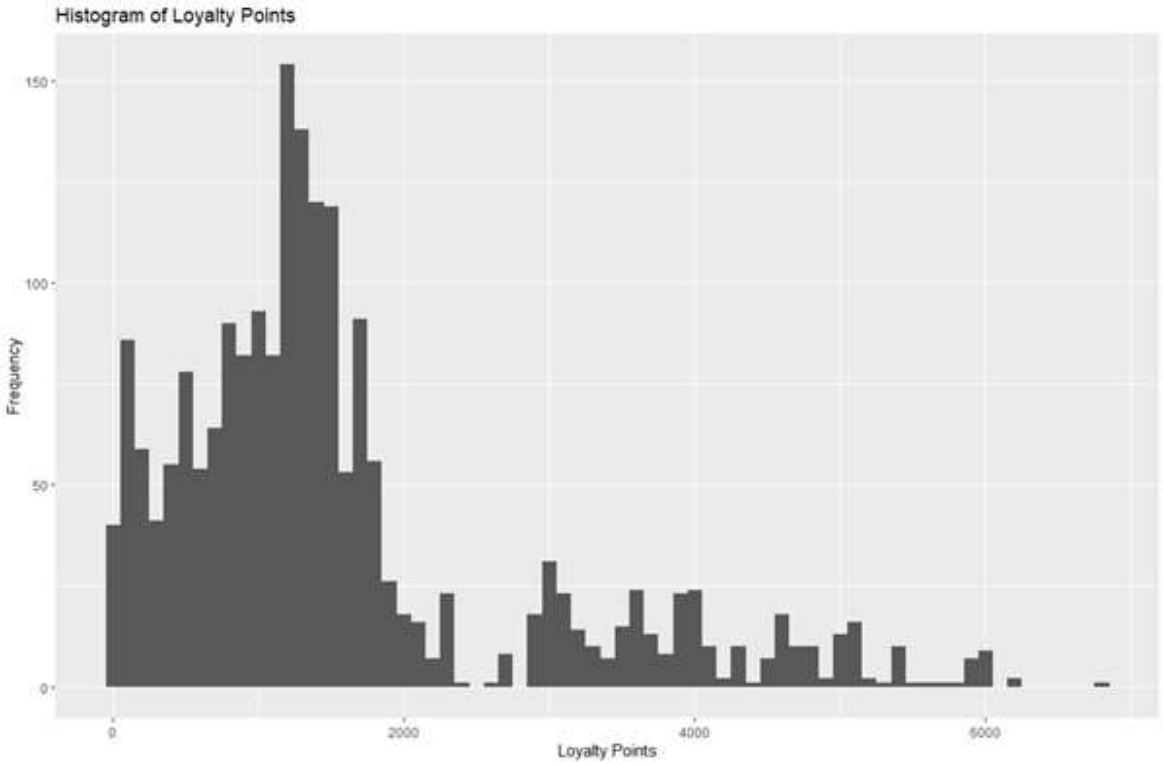
Summary Statistics – SKIMR

```
> # Summary statistics using skimr
> skim(turtle_new_r)
── Data Summary ────────────────────────
                           Values
Name                       turtle_new_r
Number of rows             2000
Number of columns          7
_________________________
Column type frequency:
  character                2
  numeric                  5
_________________________
Group variables            None

── Variable type: character ────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 gender                0             1   4   6     0        2          0
2 education             0             1   3  12     0        5          0

── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable  n_missing complete_rate   mean     sd    p0    p25    p50    p75   p100 hist
1 age                    0             1   39.5   13.6  17     29     38     49      72  ▃▇▆▃▂
2 remuneration           0             1   48.1   23.1  12.3   30.3   47.2   64.0   112. ▃▇▇▃▁
3 spending_score         0             1   50     26.1   1     32     50     73      99  ▅▃▆▅▇
4 loyalty_points         0             1 1578.  1283.   25    772   1276   1751.   6847 ▇▃▂▁▁
5 product                0             1 4321.  3149.  107   1589.  3624   6654   11086 ▇▆▂▂▃
> |
```

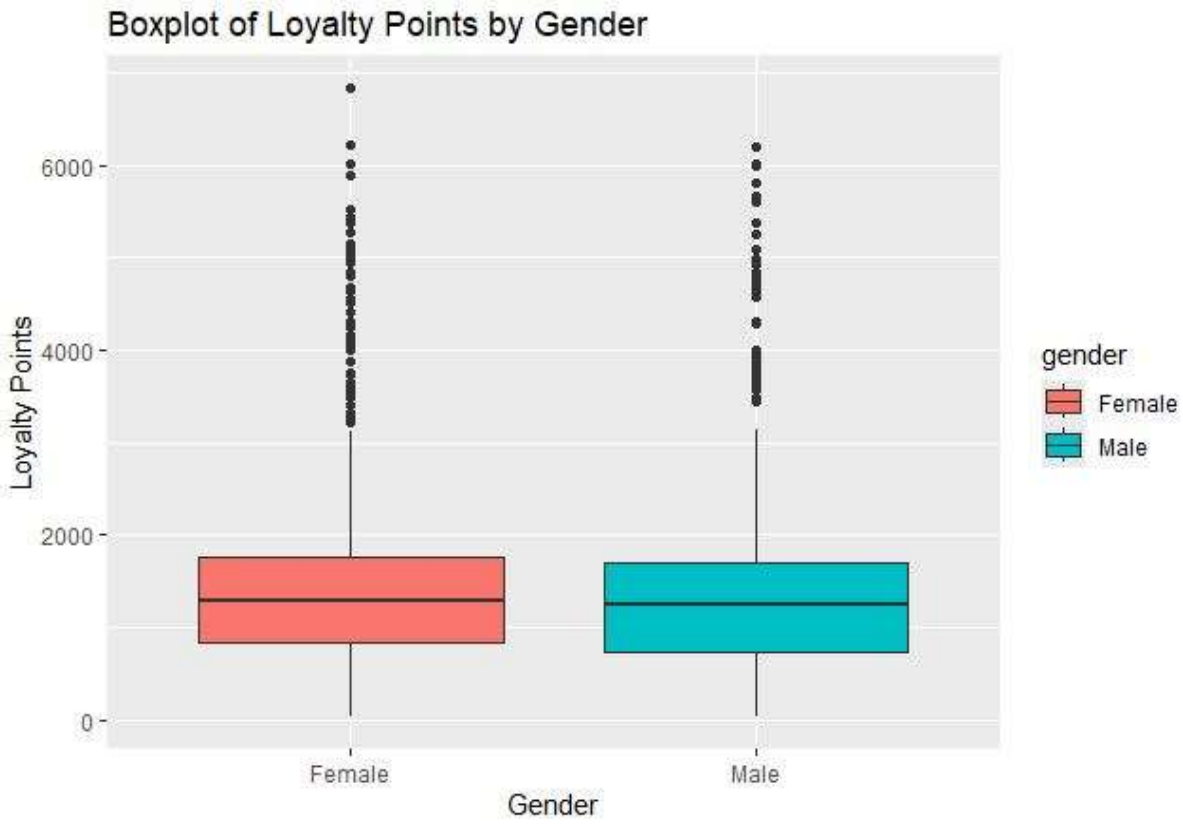
Education by Gender



Loyalty Points – Histogram



Loyalty by Gender



Loyalty by Education