# *TURTLE GAMES TECHNICAL REPORT*

Documenting and explaining approach and insights to improving overall sales performance by analysing and considering customer trends.

*David Gandary*

# Background/context of business scenario

- Turtle Games is a game manufacturer and retailer with a global customer base.
- Product range includes books, board games, video games, and toys.
- Main questions Turtle games wish to explore:
    - 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.

# Analytical Approach

Created reviews dataframe was sense checked:

- missing values
- descriptive statistics
- View of column names
- Dropping of unnecessary columns (Language, platform)
- Renaming of column headers for easier reference(remuneration, spending_score)

Linear regression models created to determine how customers accumulate loyalty points:

1. Spending vs loyalty
2. Renumeration vs loyalty
3. Age vs loyalty

- Define independent variable(spending_score)
- Define dependent variable (loyalty_points
- Split the data into training (0.7) and testing(0.3) subsets[spending vs loyalty model only to compare with OLS method]
- OLS methods created.
- Predict training and test set values.
- Plot of models **(See appendix A,B,C,D,E)**

Leverage of k-means clustering to better understand the usefulness of renumeration and spending scores.

New dataframe created and sense checked.

- Drop of unnecessary columns (age, loyalty points, education, product)
- Check for missing values.

Scatter and pair plot: renumeration vs spending score.

- Hue set to gender with 0 = female and 1 = male
- females had a higher spending score than men as the spending score increased.

Elbow and silhouette methods.

- Determines optimal number of clusters (kMeans function)
- Imported silhouette_score class from sklearn.

k-means model evaluated at different values of k

- Clusters compared: 4 and 5
- Number of observations per predicted class counted (misclassification evident)

Visualised the clusters.

- Using scatterplots based on predictions of cluster membership we can see separation of the predicted types based on the different colours and the two attributes. **(See appendix F,G,H,I)**

NLP to analyse how social data (customer reviews) can be used to inform marketing campaigns and business operations.

New dataframe created and sense checked:

- Columns dropped (Only contain review and summary columns)
- Checked for missing values.
- Data was prepared for NLP.
    - Transformed data to lowercase.
    - Replaced punctuation.
    - Dropped duplicates.
    - Tokenised data
    - Created frequency distribution with plot.
    - Removed alphanumeric characters and stopwords
- Reviewed polarity and sentiment of data in columns
    - Dataframe containing polarity and sentiment for each review row.
    - Plotted histograms (15 bins) of dataframe to visualise polarity.
- Top 20 Positive and negative reviews and summary sorted by ‘pos’ and ‘neg’ in descending order.

Leveraged R studio to load and wrangle the data and use basic visualisation techniques during exploratory data analysis.

Loading techniques

- Changing directory in R
- Imported relevant libraries.
- Imported cleaned turtle_review.csv file from python.
- Sense checked dataframe.

Wrangling techniques **(See Appendix J, K, L)**

- Scatterplot –age and spending score based on education level.
- Boxplot - How education level correlated with income.
- Histogram - Mean loyalty points based on age groups.

Computation of descriptive statistics

- Using summary () function
    - Central tendencies for mean (1578) and median (1276)
    - Extreme values (max [6847], min [25])
    - Variance (1646704)
    - Standard deviation (1283.24)

Measured the normality in loyalty points using a Q-Q plot.

- Customers with loyalty points in tails are higher than we would expect against normal distribution.
- Indicating that distribution may have heavier tails than we would expect.
- Shapiro-Wilk test
    - p-value is <0.05, so this means we can reject the null-hypothesis.
- Check for kurtosis.
    - Our kurtosis value is more than 3, suggesting our data is heavy tail distribution.

Created a multiple linear regression model.
Measured the normality in loyalty points using a Q-Q plot **(See Appendix M)**

- Adjusted R-squared = 0.8397
- Majority of the point lie close to the normal distribution line. We can suggest that model is normally distributed.

# Visualisation and insights

To compare the age and spending score based on education level a grouped scatter plot was created to visualise this relationship. The output shows a scatter plot separating each individual education group, helping to understand spread of data in more context. The visualisation can tell us that the education group with the highest concentration of spending scores in the region closer to 100 occurs for graduates in the age range of 15 – 35. We can analyse age groups further to identify more accurate insights. All education levels seem to show no strong correlations based on variables and data points.

To explore how education level correlates with income a boxplot was created to provide a clear visual summary of the central tendency and spread of the data, for each education level based on income to see if higher education levels correspond to higher incomes. Interesting insight is that basic education level holds the highest median remuneration which shows a more concentrated distribution of data. Statistical analysis (e.g., hypothesis testing) could be leveraged to determine if these differences are statistically significant. Another insight notes that there are significant outliers only in diploma, graduate, and postgraduate education levels.

A scatter plot using a linear regression model was used to see the relationship between loyalty points and spending score, which relates back to business objective of how customers accumulate loyalty points. A regression model was used to understand the strength the of the relationship more. Insight can be shown into how many customers who have a spending score higher than 60 don't follow line of best fit and therefore could potentially be receiving more loyalty points than they should if we follow the predictive model. This relates to Linear regression models created in python and R show similar model confirming reliability.

A histogram was plotted to show mean loyalty points based on age groups, where the histogram accurately depicts a count of individual mean loyalty points for a customer in different age groups. The grouped ages histogram is displayed to show the average loyalty points between different age groups. Insight that 25–35-year-olds have the highest average mean loyalty points and 15–25-year-olds have the lowest. Also, trend in gradual decline in mean loyalty points after age group 25-35 until age group 65-75 where there is a short increase.

# Patterns and predictions

Multiple linear regression model showed strong adjusted R-squared value = 0.8397therefore indicating a good fit to the data between age, renumeration and spending score which could be used to predict loyalty points.
For example, if age = 30, remuneration = £55000 and their spending score = 55 the model can predict that the customers loyalty points are to be 1870470. This prediction can guide personalized marketing strategies targeting high-value customers.

In conclusion, leveraging insights from the MLR model and exploratory analysis, we recommend optimizing marketing campaigns targeting specific customer segments based on predicted loyalty points. Further analysis could explore additional predictors such as education level or the actual type of products. Customers particularly in the age range of 25-35 retain more loyalty points, which can indicate targeting this demographic in marketing strategies to help raise their spending score.

# Reference


- LSE. (2023).LSE_DA301_Advanced Analytics for Organisational Impact_C4_2023. Available at: https://platform.fourthrev.com/courses/701/assignments/2431 (Accessed: April 2024).
- Towardsdatascience. (2020) Sentiment Analysis using TextBlob. Available at: https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524 (Accessed: April 2024)
- Stack overflow. (2018) ImportError: No module named 'wordcloud'. Available at: https://stackoverflow.com/questions/47298070/importerror-no-module-named-wordcloud (Accessed: April 2024)

# Appendices

## Appendix A


The output displays the regression model for the training set of data for the spending score vs loyalty points data.

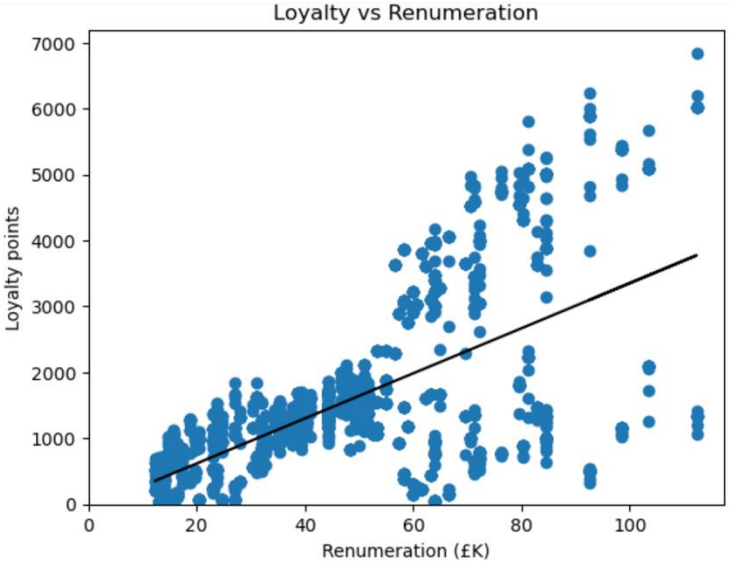
## Appendix B


The output displays the regression model for the testing set of data for the spending score vs loyalty points data.

## Appendix C

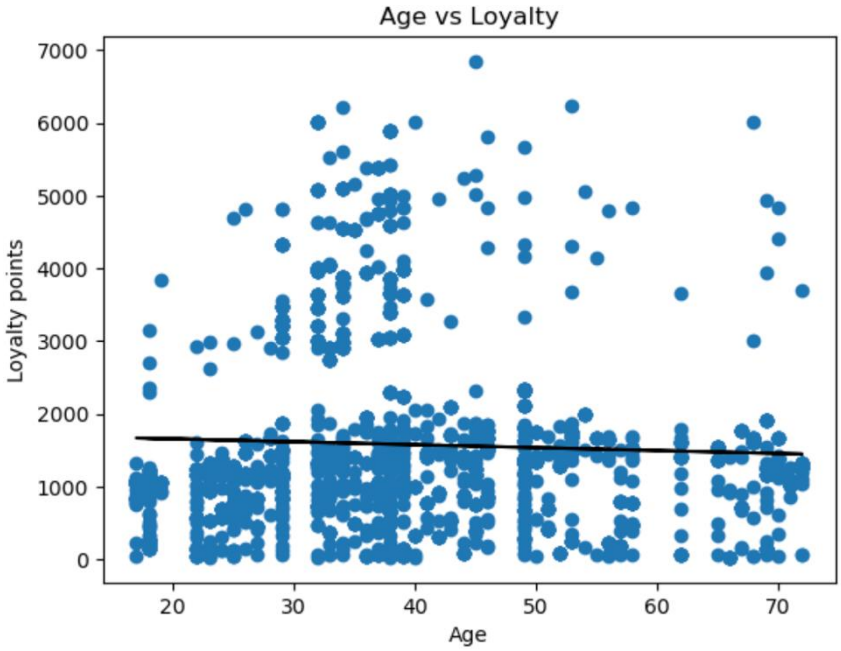

The output displays the regression model for the testing set of data for the spending score vs loyalty points data using the OLS method.

## Appendix D



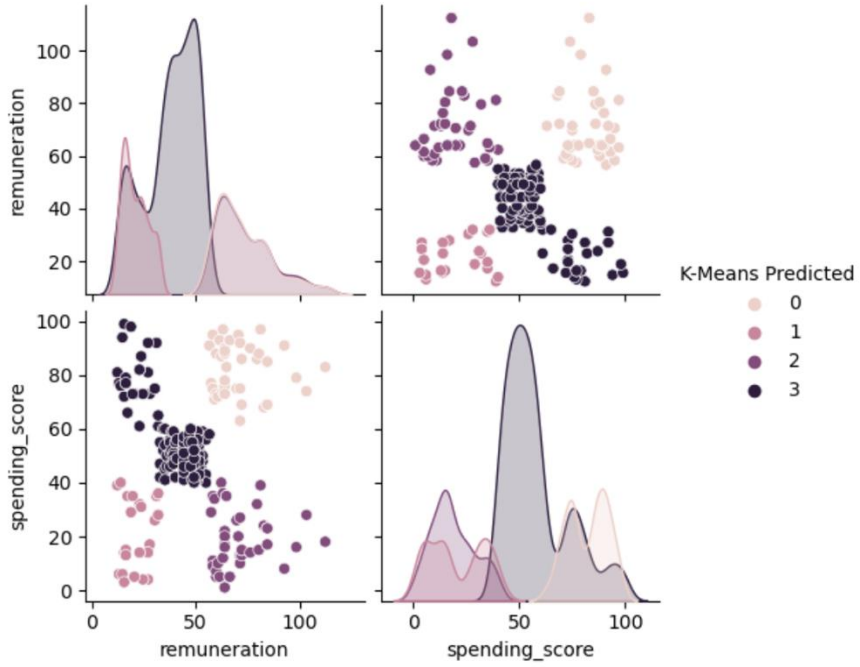

The output displays the regression model for the testing set of data for the renumeration vs loyalty points data using the OLS method.

## Appendix E



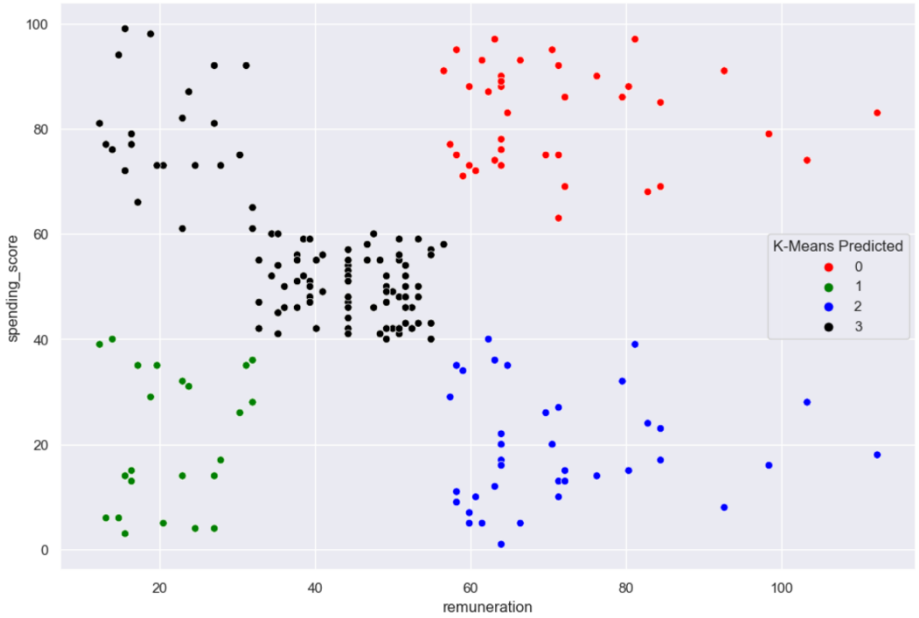

The output displays the regression model for the testing set of data for the age vs loyalty points data using the OLS method.

## Appendix F



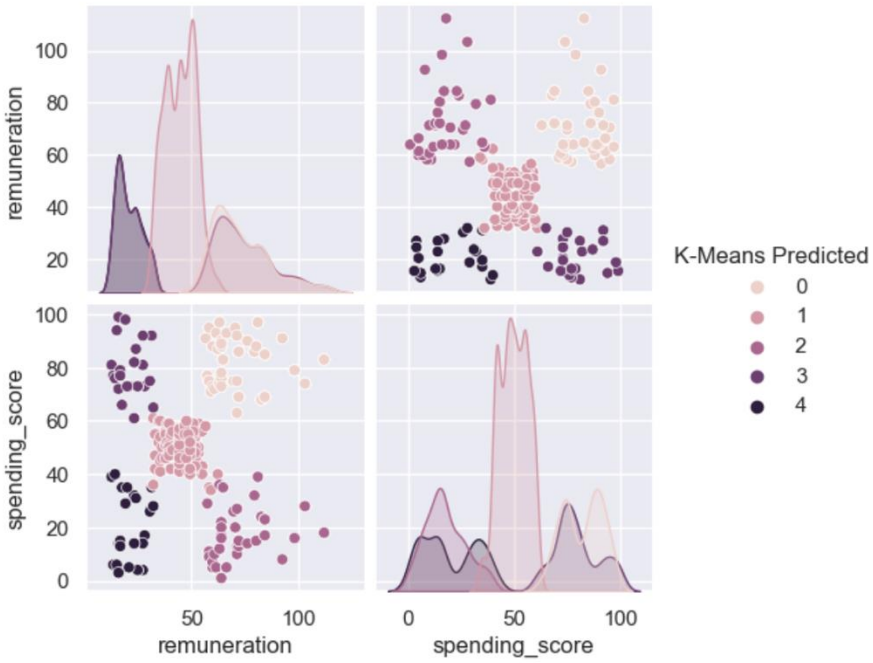

Pair plot based on all the objects assigned to these 4 cluster solutions we can see what the distribution of the spending score and renumeration are.

## Appendix G



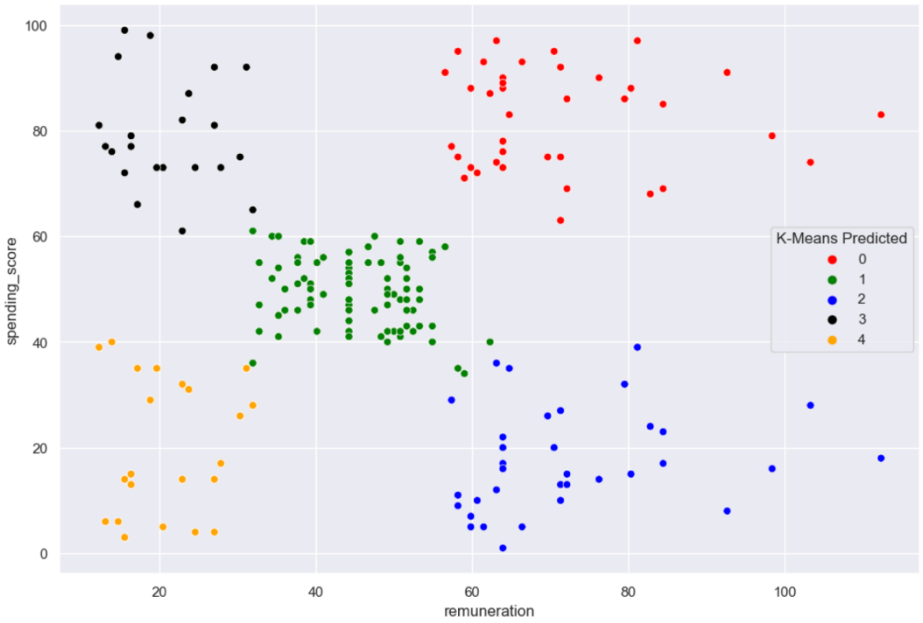

Scatterplot based on predictions of 4 cluster membership between spending score and remuneration, we can see separation of the predicted types based on the different colours and the two attributes.

## Appendix H



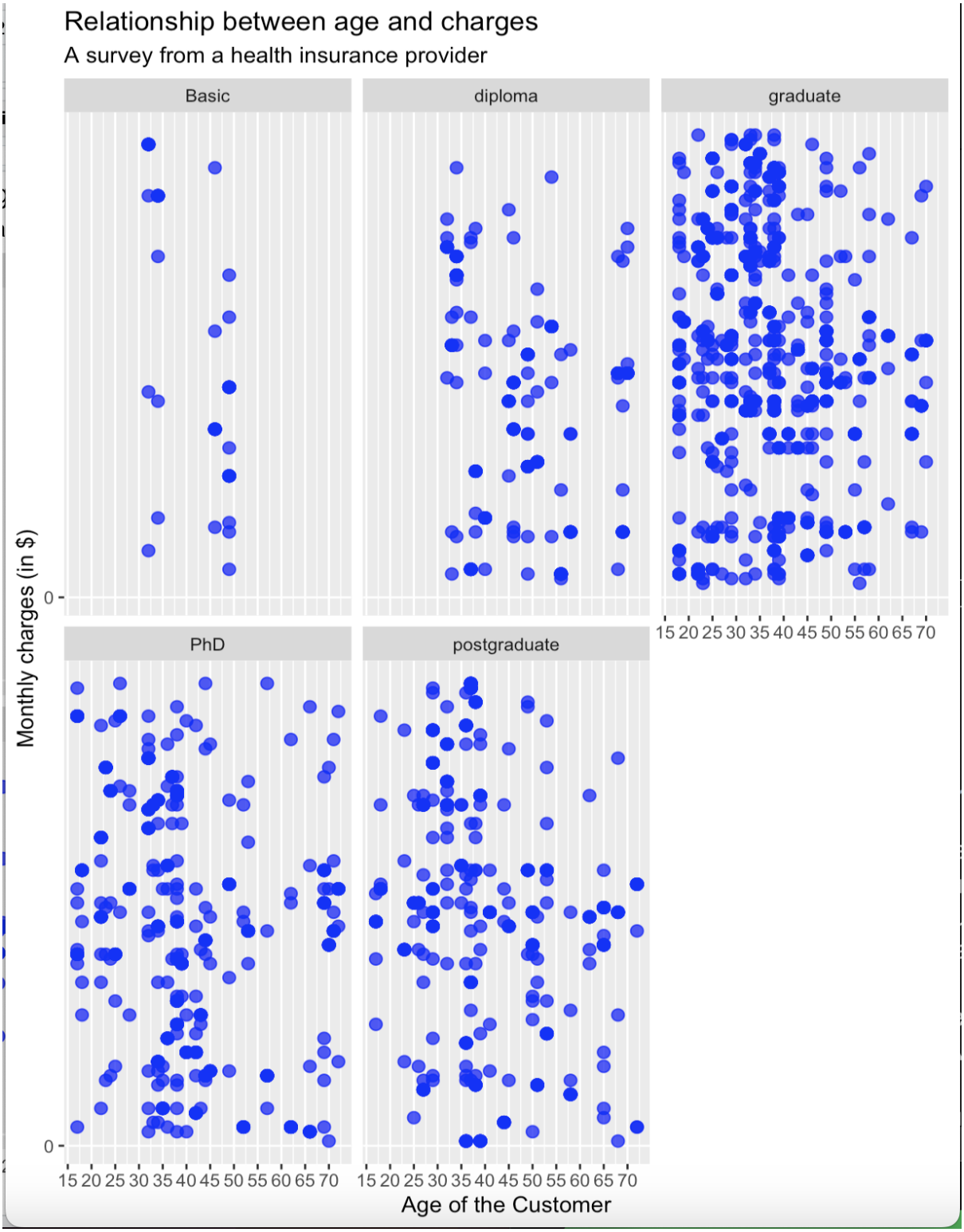

Pair plot based on all the objects assigned to these 4 cluster solutions we can see what the distribution of the spending score and renumeration are.

# Appendix I



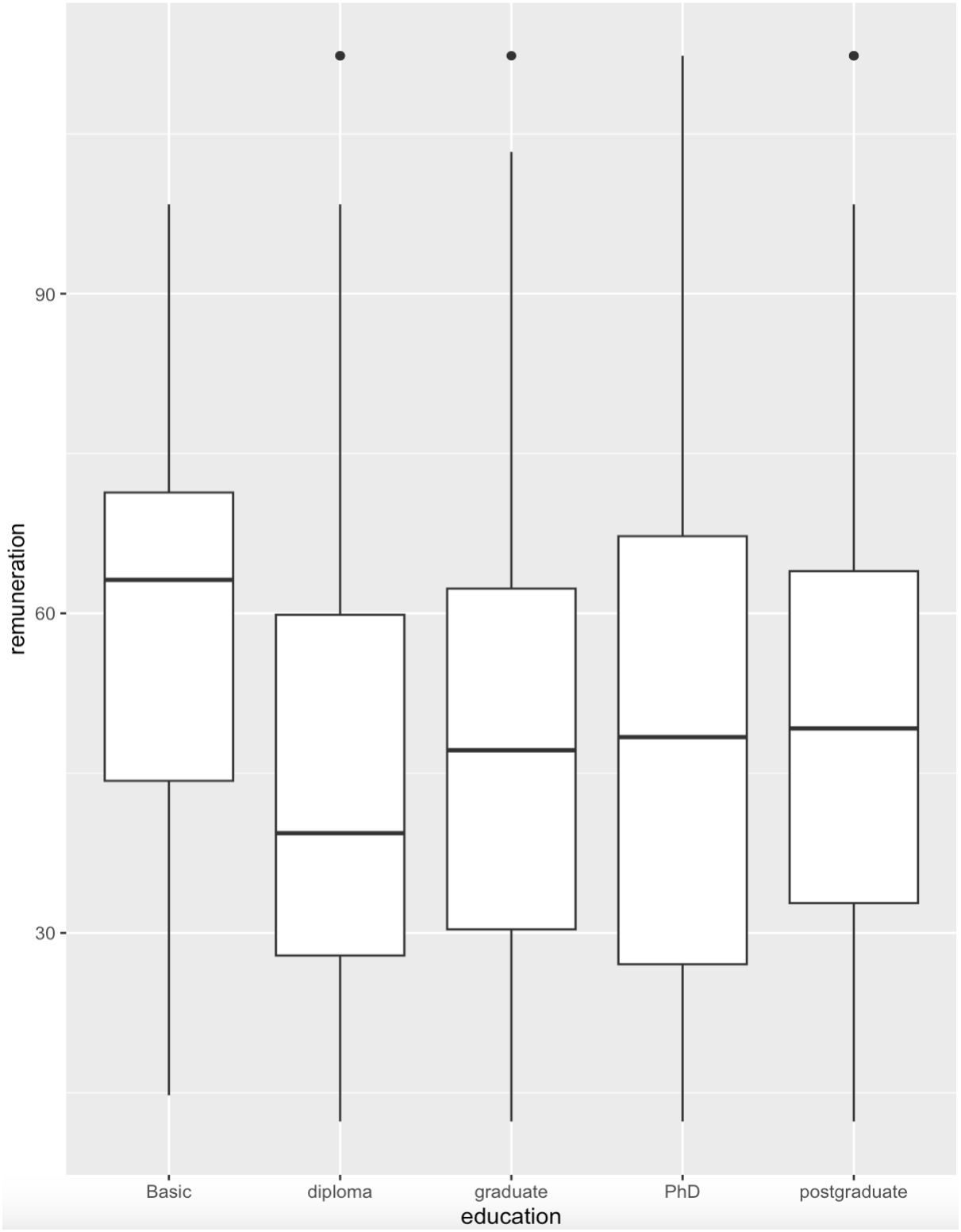

Scatterplot based on predictions of 5 cluster membership between spending score and remuneration, we can see separation of the predicted types based on the different colours and the two attributes.

## Appendix J



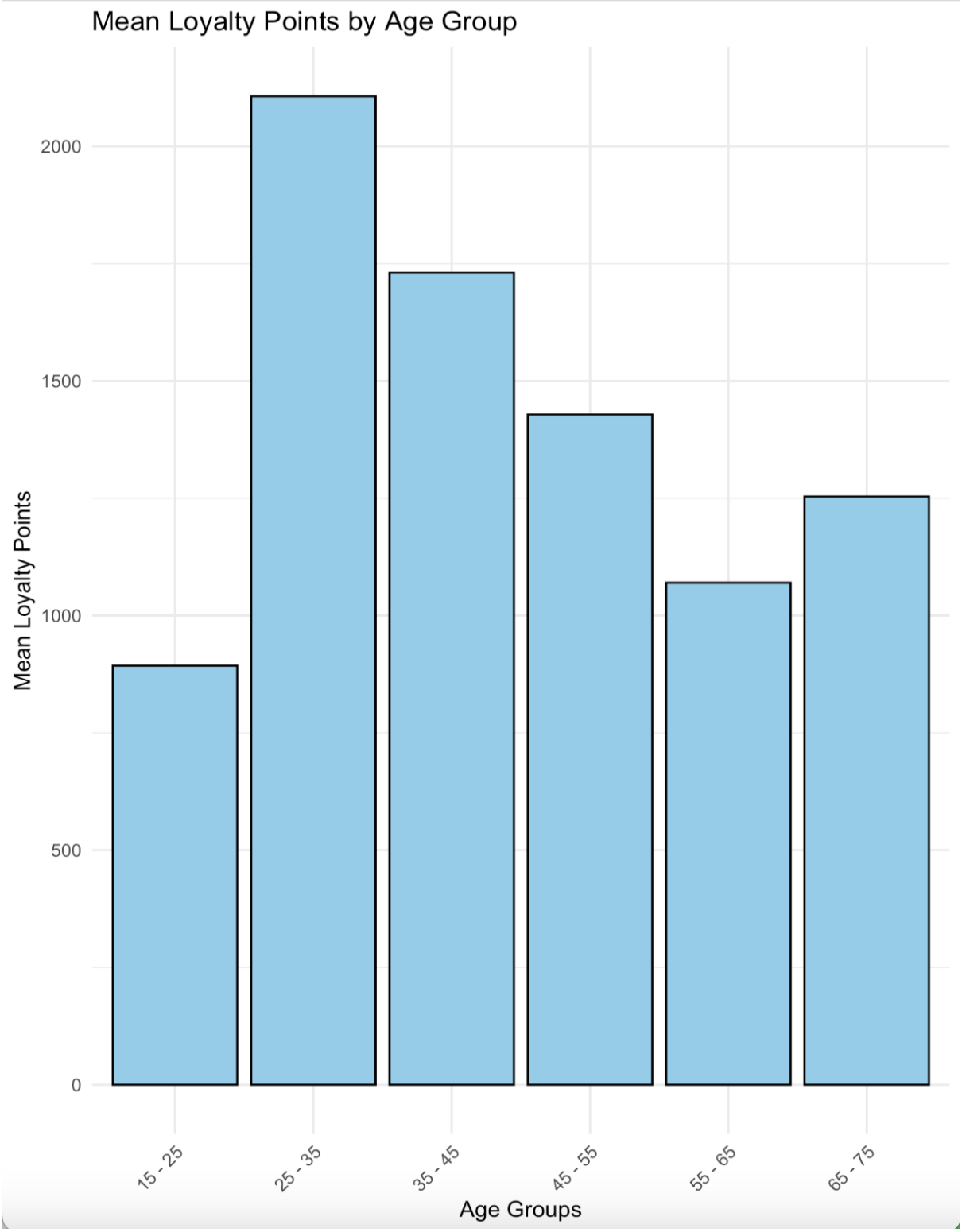

Output shows scatter plot separating each individual education group, helping to understand spread of data in more context.

## Appendix K



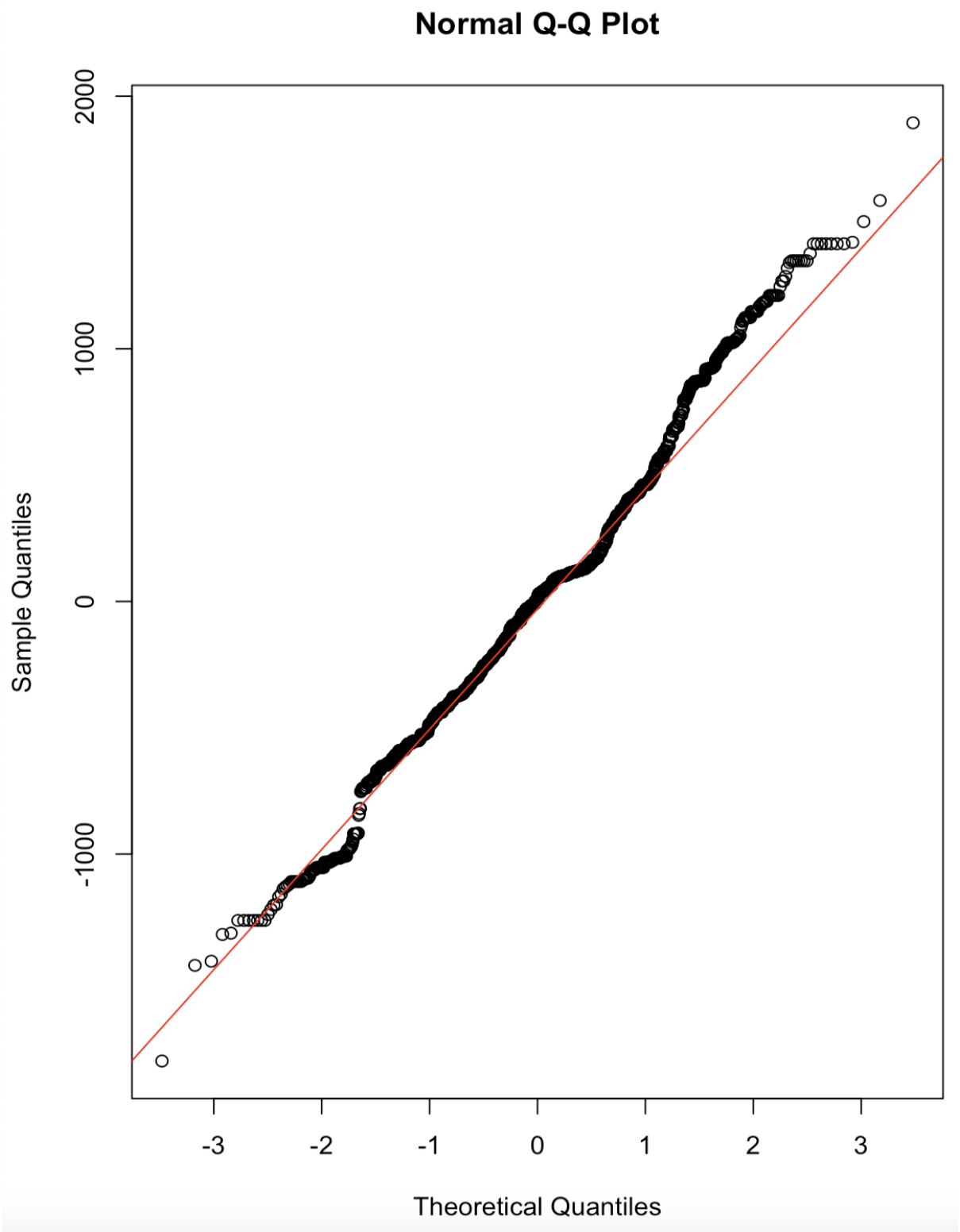

Output shows boxplot to provide a clear visual summary of the central tendency and spread of the data, for each education level based on income to see if higher education levels correspond to higher incomes.

## Appendix L




The grouped ages histogram is displayed to show the average loyalty points between different age groups.

## Appendix M




This plot assesses model assumptions such as normality of residuals, linearity, and homoscedasticity. Majority of the point lie close to the normal distribution line. We can suggest that model is normally distributed. Slightly higher tails at point1 and 2 and A slight lower tail at point between -1 and -2