Phil Best

Turtle Games Analysis: Technical Report

1) Problem Statement

Turtle Games would like to understand trends in its loyalty point and customer review data, and whether predictive models can help make improvements to its strategy. These insights will allow Turtle Games to better serve their customers, which ultimately will lead to increased sales and profit for the business.

2) Analytical Approach

a) Collecting the data – Python [7]

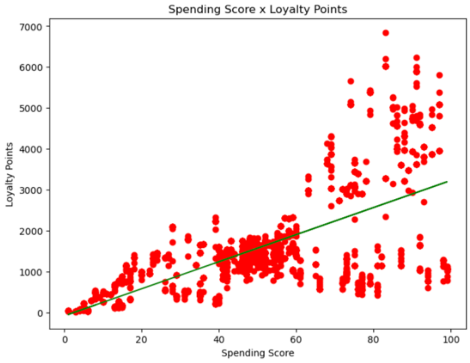
- Imported ‘'turtle_reviews.csv' using pd.read_csv(), assigning the name ‘tr’

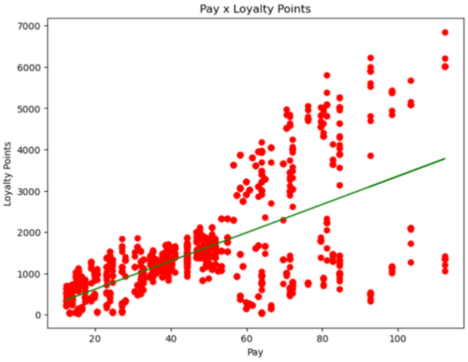
b) Basic cleaning – Python [8] to [27]

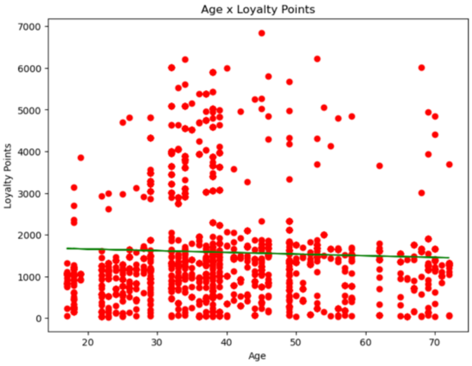
- Checked for missing values (isnull)
- Checked consistent data types (.dtypes),
- Dropped irrelevant columns Language and Platform (.drop)
- Renamed 'remuneration (k£)' : 'pay' and 'spending_score (1-100) ' : 'spending_score' (.rename)

c) Linear Regression for Loyalty Points – Python [30] to [47]

- Created linear regression models using sm.OLS().fit() and summarised performance using .summary. Visualised output using plt.scatter() **[30] to [45]**



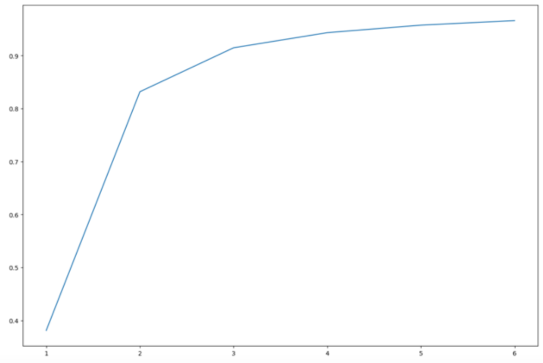




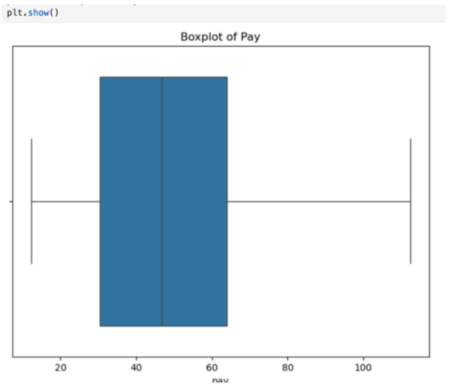
- Created multiple linear regression with spending and pay x loyalty **[45] to [47]**

c) Decision Tree for Loyalty Points – Python [53] to [116]

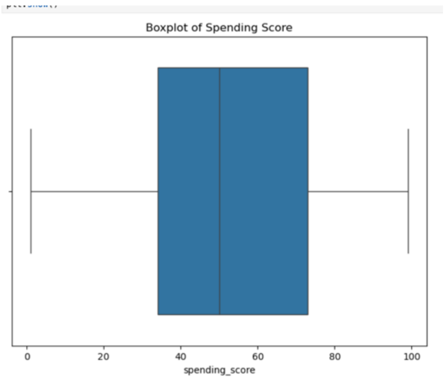
- Used one-hot encoding to convert gender using .get_dummies() **[55]**
- Used ordinal encoding to convert education using OrdinalEncoder **[56] to [59]**
- Created test and train data splitting 70:30 and used DecisionTreeRegressor **[66] to [70]**
- Evaluated the model using .mean_absolute_error() and .mean_squared_error() **[71]**
- Checked feature importance using regressor.feature_importance_. and reran model **[72] to [83]**
- Plotted R-squared values across the range of depth values evaluated

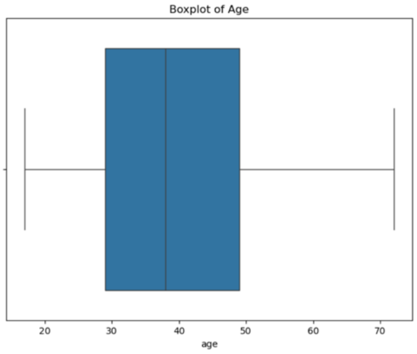


- Mean Absolute Error (26) is significantly lower than the Root Mean Squared Error (80), so tested for outliers in independent variables. No outliers found. Confirmed through boxplots **[87] to [92]**

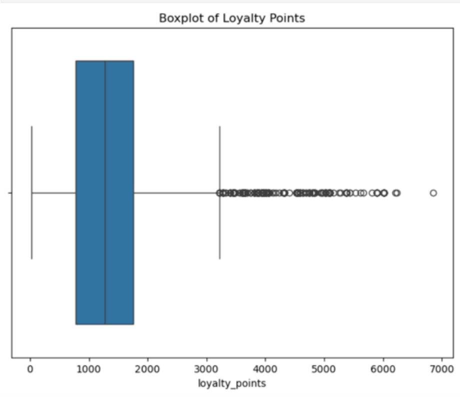


```
plt.show()
```

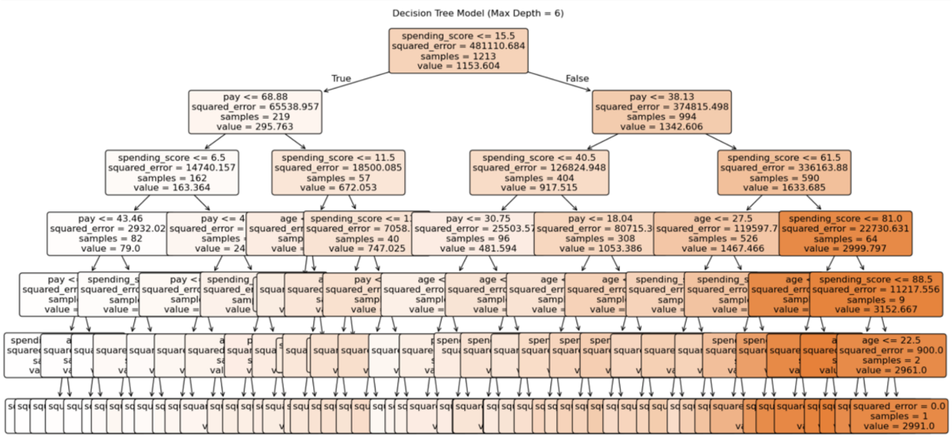




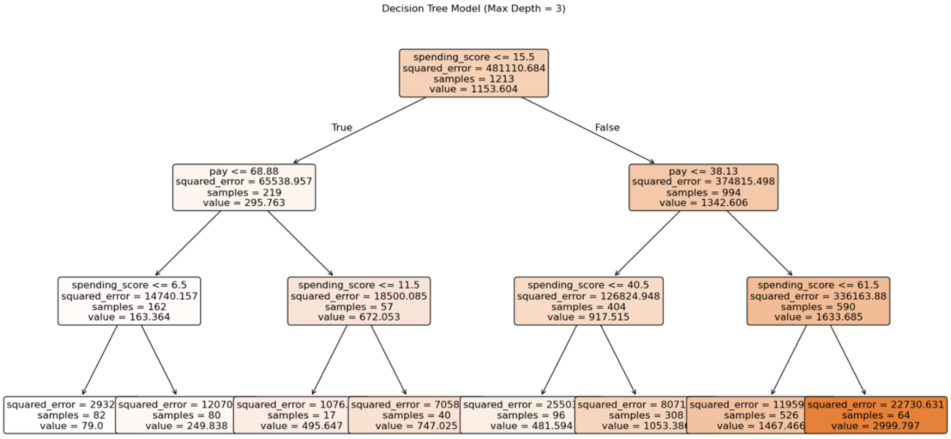
- Ran same exercise on dependent variable (loyalty points) and found many outliers **[93]**



- Reran model with outliers removed. Improved performance, but still significant variance between the two measures **[95] to [112]**
- Plotted decision tree to max depth 6 **[115]**

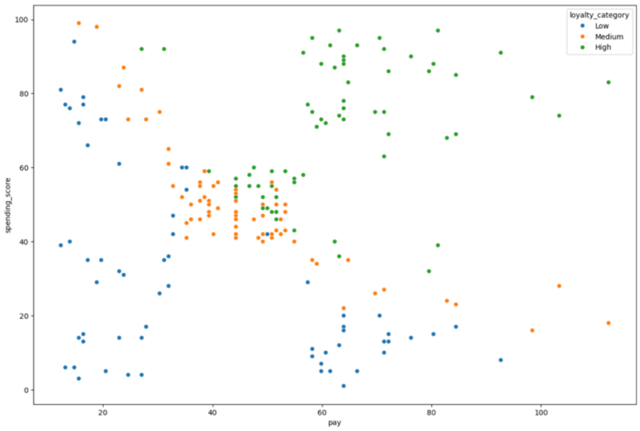


- For ease of understanding replotted to Max Depth 3 **[116]**

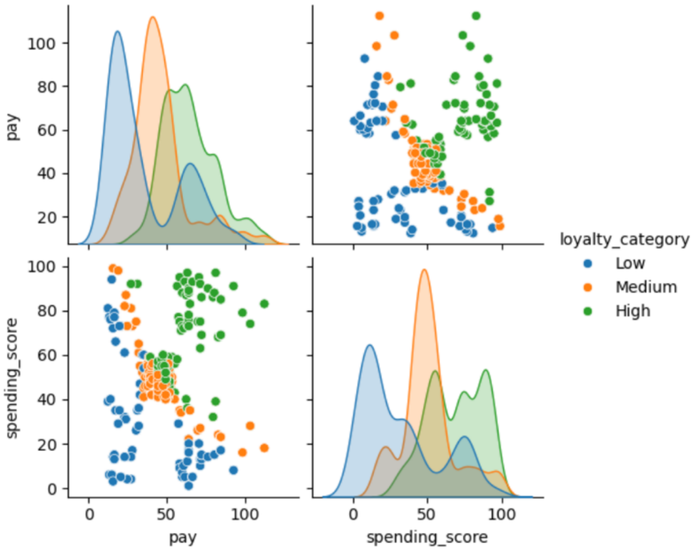


d) Clustering customers with k-means – Python [121] to [158]

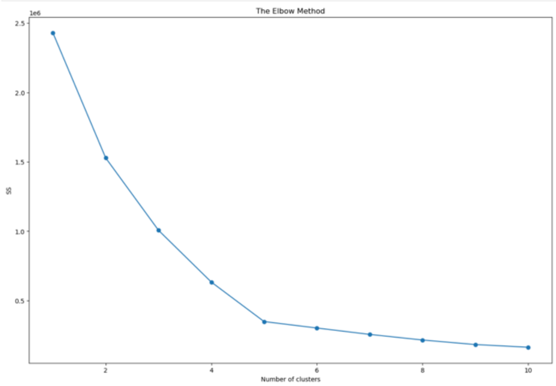
- Created a scatterplot using sns.scatterplot() for pay and spending score. Created loyalty categories – low, medium, and high - represented by hue on plot **[124] to [130]**

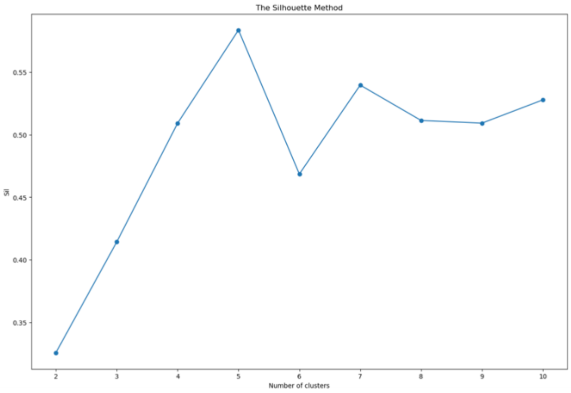


- ○ Created a .pairplot() to further demonstrate the relationship **[132]**

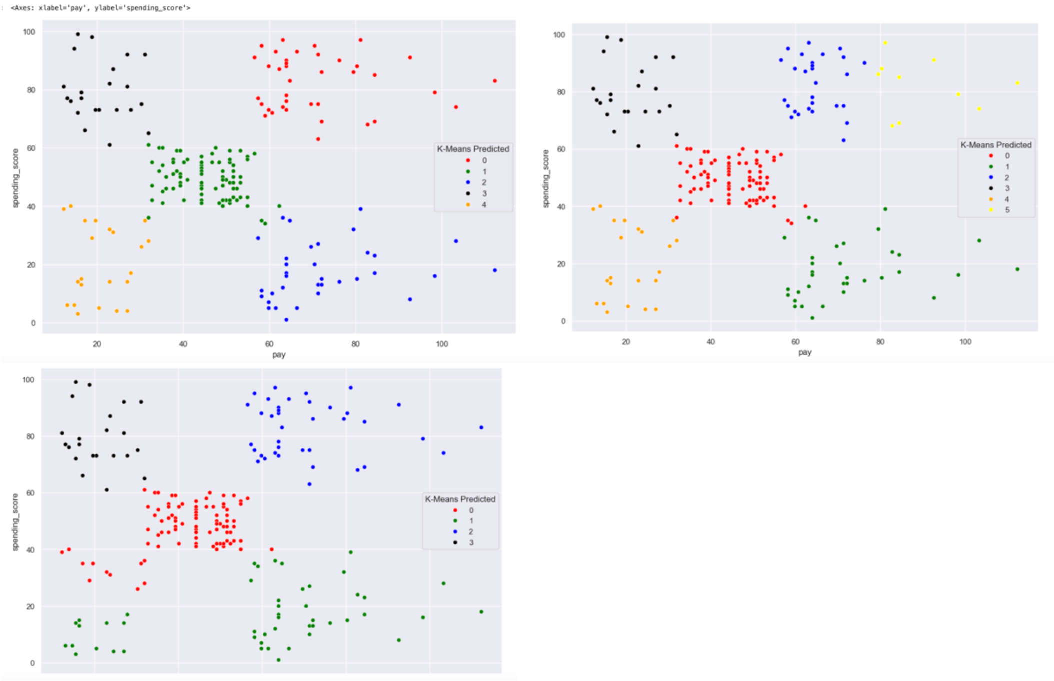


- ○ Evaluated the number of potential clusters through Elbow and Silhouette models **[134] to [135]**





- ○ Evaluated having 4, 5 or 6 clusters, visually plotting them in scatterplots **[137] to [151]**



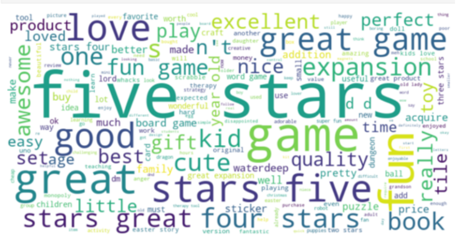
<Axes: xlabel='pay', ylabel='spending_score'>

- Concluded that 5 clusters made the most sense with 5 easily explained segments:
  - high pay & high ss
  - medium pay & medium ss
  - high pay & low ss
  - low pay & high ss
  - low pay & low ss
- Add Clusters as a column on the dataframe for future analysis **[161]**

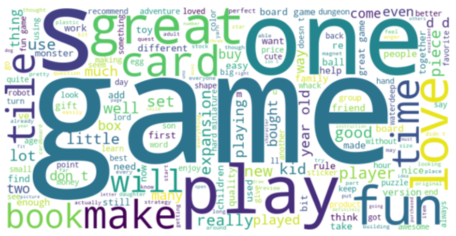
e) NLP to assess review sentiment – Python [173] to [267]

- Edited Review and Summary columns into a readable format
  - converted to lower case using str.lower() **[181] to [182]**
  - replaced all punctuation using str.replace(f[{string.punctuation}]) **[186] to [187]**
  - dropped duplicates using drop_duplicates **[189] to [190]**
  - tokenised each column using .apply(word_tokenize) **[192] to [198]**
- Plotted most frequent words and create wordclouds using .imshow(wordcloud) **[199] to [203]**

SUMMARY

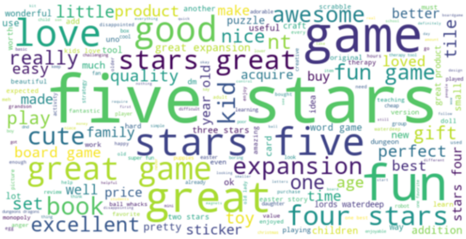


REVIEW

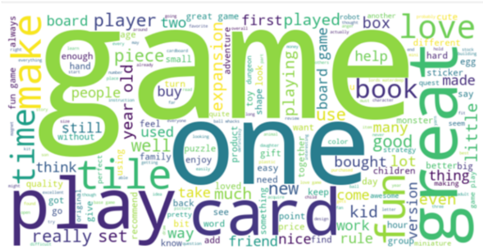


- Removed alphanumeric characters and stopwords, and created new wordclouds **[208] to [217]**

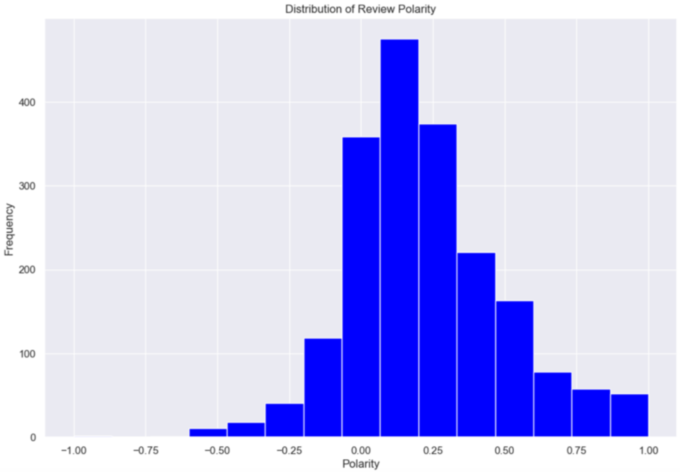
SUMMARY



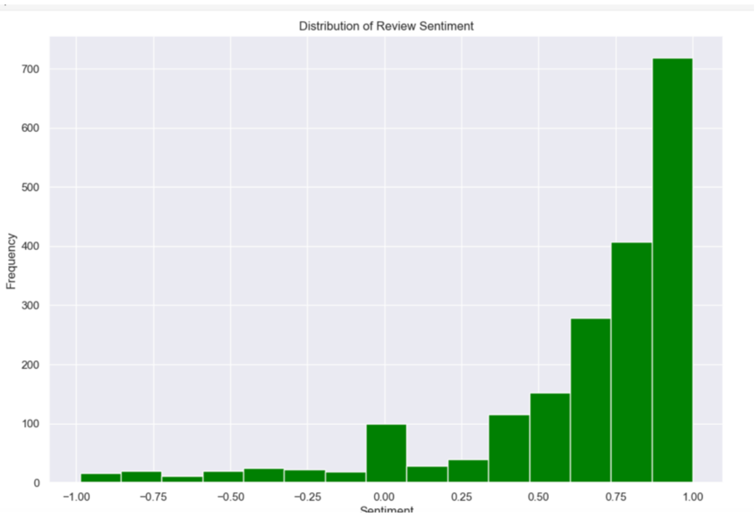
REVIEW



- Assessed Review Polarity using TextBlob(), visualising the distribution **[222] to [231]**



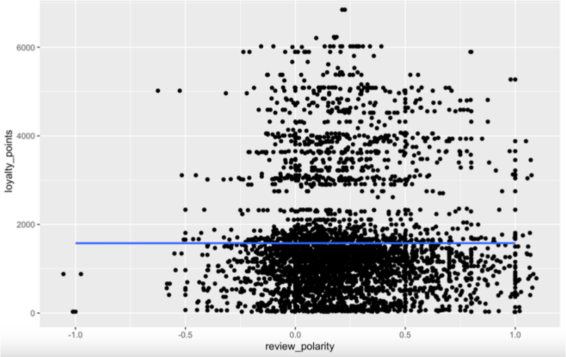
- Assessed Review Sentiment using Vadar, visualising the distribution **[232] to [234]**

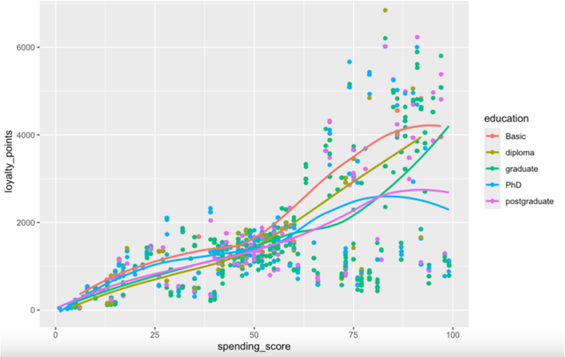


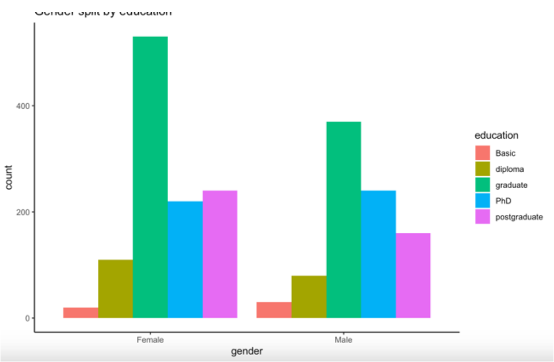
- Identified top 20 reviews for both Vadar and TextBlob. Manually assessed whether positive or negative. **[243] to [255]**
- Created Confusion Matrix and Classification report to compare accuracy vs. my assessment **[256] to [262]**
- Vadar was slightly more accurate (86% vs 85%), but TextBlob handled both positive and negatives in a more balanced way. Used TextBlob as preferred method going forward.

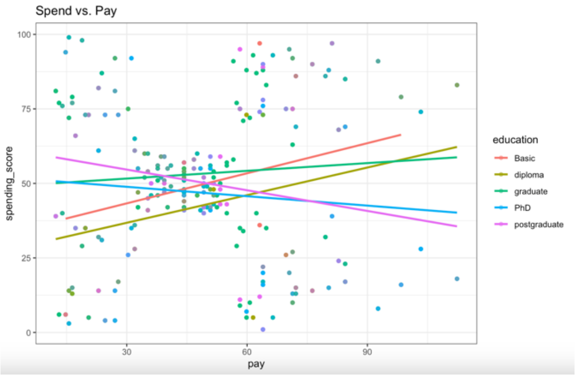
f) Exploratory Analysis and Multiple Linear Regression using R – R [71] to [593]

- Convert gender, education and cluster columns to categorical data types using mutate() **[98] to [116]**
- Created Histograms and Density plots for loyalty points, spending score and pay, concluding that the loyalty points data was very skewed, with a few very high outliers **[126] to [167]**
- Explored relationships between the different variables using scatterplots to visualise **[171] to [267]**





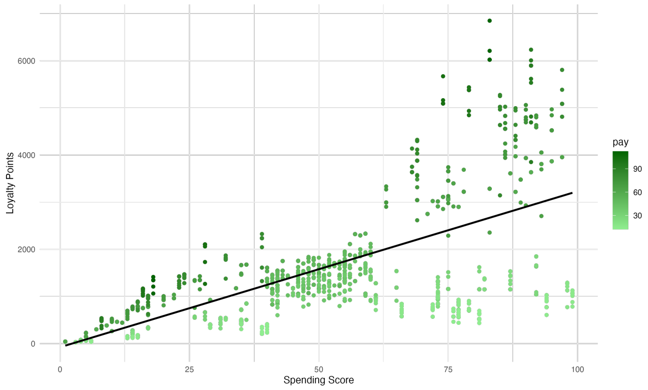




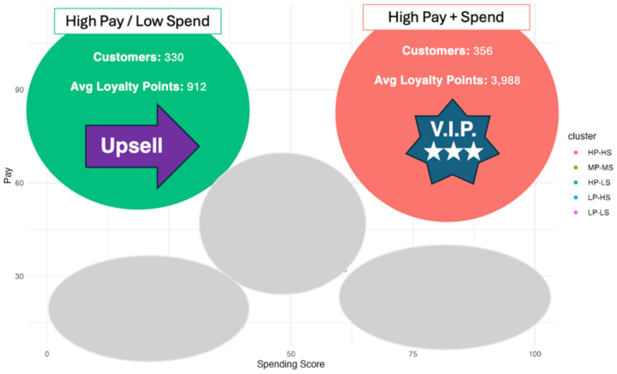
- Found different education levels have different correlation between Pay and Spending Score
  - Basic & Diploma: Positive correlation between Pay & Spending Score
  - Graduate: Mild positive correlation between Pay & Spending Score
  - PHD & PostGrad: negative correlation between Pay & Spending Score
- Filtered dataframe to explore high loyalty points users (>2000 points) **[270] to [304]**
- Explored descriptive statistics and visuals for the Clusters created through k-means **[306] to [355]**
- Grouped data by product and analysed the Review sentiment per product **[357] to [420]**
- Created a multiple linear regression for loyalty points, starting with pay and spending score, then adding in variables. Optimal model performance included pay, spending score, age and education. **[513] to [550]**
- Due to loyalty points being skewed (skewness: 1.5, kurtosis: 4.7), transformed the model to reduce impact of outliers **[575] to [593]**

3) Insights and Visualisations

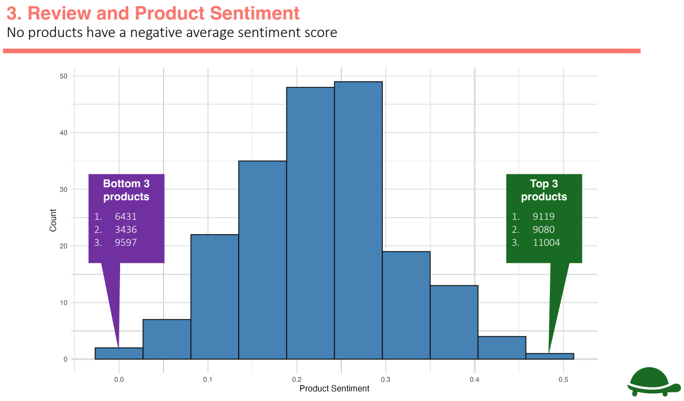
a) Pay and Spending Score account for over 80% of variation in loyalty points



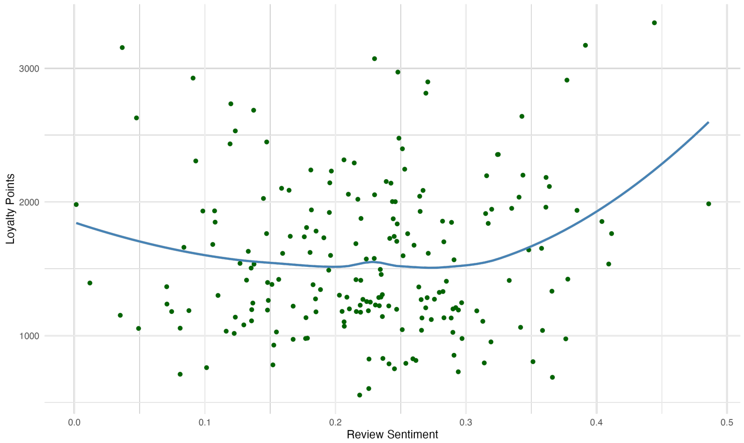
b) Marketing strategy should focus on maintaining High Pay / Spend Score Segment and upselling High Pay/Low Spend Score Segment



c) Top sentiment rated products should be promoted over lesser products, potentially using loyalty points to incentivise



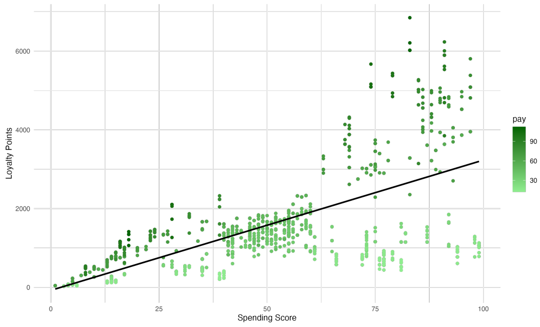
d) Currently there is no consistent relationship between loyalty points accrued and product purchased

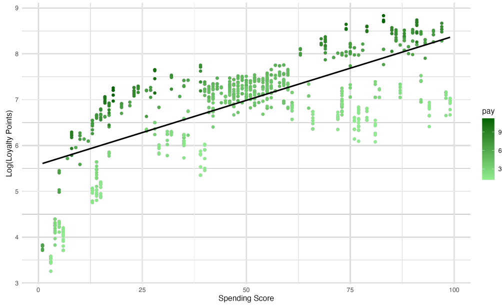


e) Given available variables can explain over 80% of variation in loyalty points, using multiple regression models can help inform strategy. However, its recommended data is transformed to remove the impact of the significant outliers in loyalty data

<u>Pre-transformation</u>

<u>Post-transformation</u>





4) Predictions and applications of model

- Best performing loyalty point linear regression model (with transformed data) includes Pay, Spending Score, Age and Education
- Pay and Spending Score remain the biggest impact on loyalty points. Tested model with several scenarios **[608] to [696]**
- Model could be used to identify potential high earners of loyalty points after their first purchase (presuming spending score is assigned early in a customer's life). This could help:
    - Marketing put in place tailored strategies to ensure customers reach their potential with the business
    - Finance assess the future cost of loyalty points
- Questions:
    - How is spending score assigned, and when is it assigned to a customer?
    - Do loyalty points translate to customer spend?
    - What is the overall aim of the loyalty scheme?
    - How are loyalty points spent?