

TURTLE GAMES TECHNICAL REPORT

By Dhru Mistry

Introduction

In this report, I will be talking about Turtle Games, which is a game manufacturer and retailer with a global customer base. They assigned the task to a team of data analysts which was to unpack data to answer multiple different business questions which will ultimately help them with their task of wanting to increase their overall sales performance, we hope to help them do this by looking at many categories within the provided data such as age, gender, income, spending score and others.

Analytical Approach

The first step taken was to import the data using `read.csv` within both programs, however python was used first but steps were very similar in both programs, then next step was to clean the data, this was done by first removing unnecessary columns which in this case was 'language' & 'platform' this was because both columns only had 1 value and so would not be any use in identifying patterns & trends for this cause. After this, looking at the column headers, 'remuneration' and 'spending_score (1-100)' had to be renamed as they could be made simpler to understand but also quicker to use within coding, along with this I multiplied the whole income column by 1000 to get full £ values easier to understanding within visualisations. They were renamed to 'income' & 'spending_score' respectively. Then exploring the data was next, looking for null values, which there was none, understanding the data types for each column and the structure/dimension of the dataset. The data has now been cleaned and was ready for exploratory data analysis but first, the clean dataset needed to be saved using `'turtle_reviews.to_csv()'` (Appendix 1) so I would not have to reclean the data within R. Within python many libraries/packages (Appendix 2) were used to help firstly, to help make predictions with regression to determine how customers accumulate loyalty points. Next was to be able to explore data with decision trees to further gain insights to how loyalty points are accumulated. Thirdly, it enabled being able to make predictions with clustering which helped segment customers which will help allow the marketing team to target those segments specifically. Lastly, it allowed to analyse customer sentiments observed in reviews by being able to use NLP to determine how the social data could influence & determine marketing campaigns and business operations. Within R, this was much simpler as packages are merged into one library, the main one used was 'tidyverse' which contains most of the packages needed, further ones used were 'ggplot2' to be able to create visualisations and 'patchwork' to show plots within the same grid. All together these libraries help clean & wrangle the data to perform EDA to communicate basic inform business decisions.

Visualisation & Insights

To start with the graphs used to show the regression ([Appendix 3](#)), from the first graph it highlights that there is no correlation between age and the accumulation of loyalty points. However, when it comes to the other 2 graphs it highlights a strong correlation between loyalty points and income/spending score but similarly in both as income and spending score get higher (60,000 income & 80 spending score) the markers become erratic on both sides of the trend line with higher highs for loyalty points but lower lows also.

Next, looking at the decision tree ([Appendix 4](#)), we can identify the most important influences within the predictor. At the top is income as it is the root node, the dark orange boxes on the right highlights that the wealthier, active customers accumulate the most loyalty points and lower income people have the lowest loyalty points which could link to affordability, the 2nd most influential categories are spending score & age.

Now onto the next graph ([Appendix 5](#)) which uses k clustering to segment customers into groups, in this case the groups are based on income-to-spending rate, the reason behind using 3 clusters is due to using the elbow method ([Appendix 6](#)). Cluster 1 (orange) shows the low income/ low spenders but hints towards that being the status of the average customer as it has the most populated cluster. Cluster 2 (green) shows the high income / low spenders and lastly, Cluster 3 (blue) showcases the high income / high spenders.

The next visualisations relate to reviews and summary columns. First, looking at the wordclouds for both columns ([Appendix 7&8](#)) the bigger/most frequent words such as 'quality', 'price', 'service' are most in-depth in comparison to summary which the bigger/ most frequent words are quite broad examples are 'great', 'poor', 'disappointed' which also have a more negative connotation. Lastly, looking at the top 20 positive & negative reviews, the top 10 positive reviews & summary they have a mean polarity of 0.88 & 1 respectively which highlights that for reviews although not perfect, most customers are happy but have also given small amounts of criticism. However, for summary, having a score of 1 show that all comments are purely positive. Now for the top 10 negative for both, they have a mean polarity of -0.54 & -0.82 respectively which are on the higher side then turtle games would like to showcase in reviews they are moderate whereas summary is strongly negative.

Lastly, on to using R to clean and wrangle the data to be able to perform visual EDA. From this I was able to use a wide range of charts. To start using a bar chart to identify the most popular product IDs ([Appendix 9](#)) to allow turtle games to potentially focus product strategy on bestselling items. Next, the patchworks library was used to show trends in how an increase in age affects loyalty & spending ([Appendix 10](#)) and they both show the same result of a decrease as age increases and so turtle games could begin personalising marketing towards older age groups.

Patterns & Predictions

The main patterns that can be identified that in regards to overall sales performance is the lack of repurchases within the low income & older segments of people as the main stand out pattern across all forms or analysis but mainly potent in the decision tree is that income drives loyalty and so a way for turtle games to go around this is promotional campaign such as buy an item and get a cash back voucher to use on the next purchase etc to help increase affordability for the lower income band to encourage a repurchase.

Appendix

APPENDIX 1

```
#save cleaned version of data  
turtlereviews.to_csv("turtlereviews_clean.csv", index=False)
```

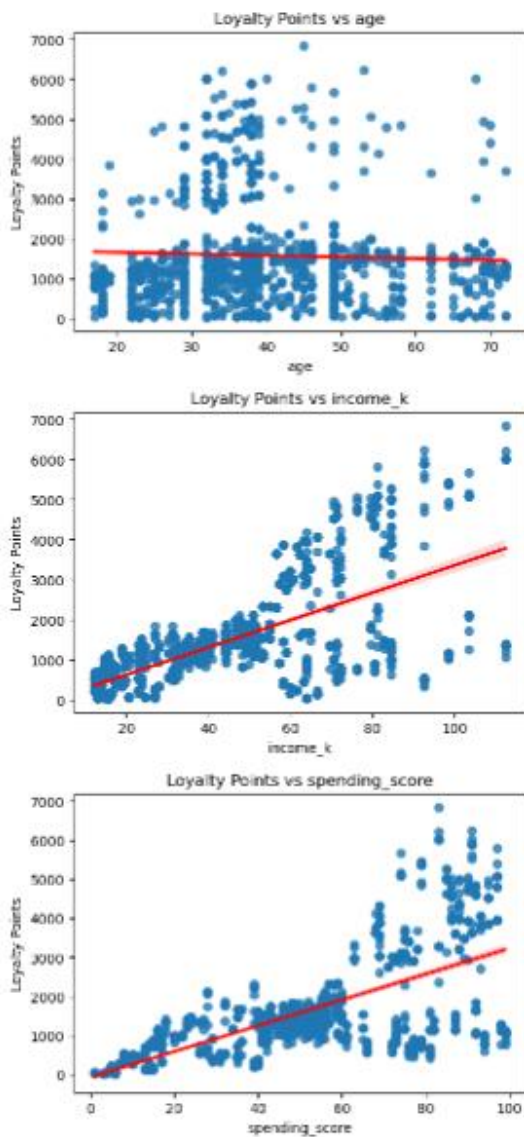
APPENDIX 2

```
# Import necessary libraries.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import seaborn as sns

from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.metrics import accuracy_score
from scipy.spatial.distance import cdist

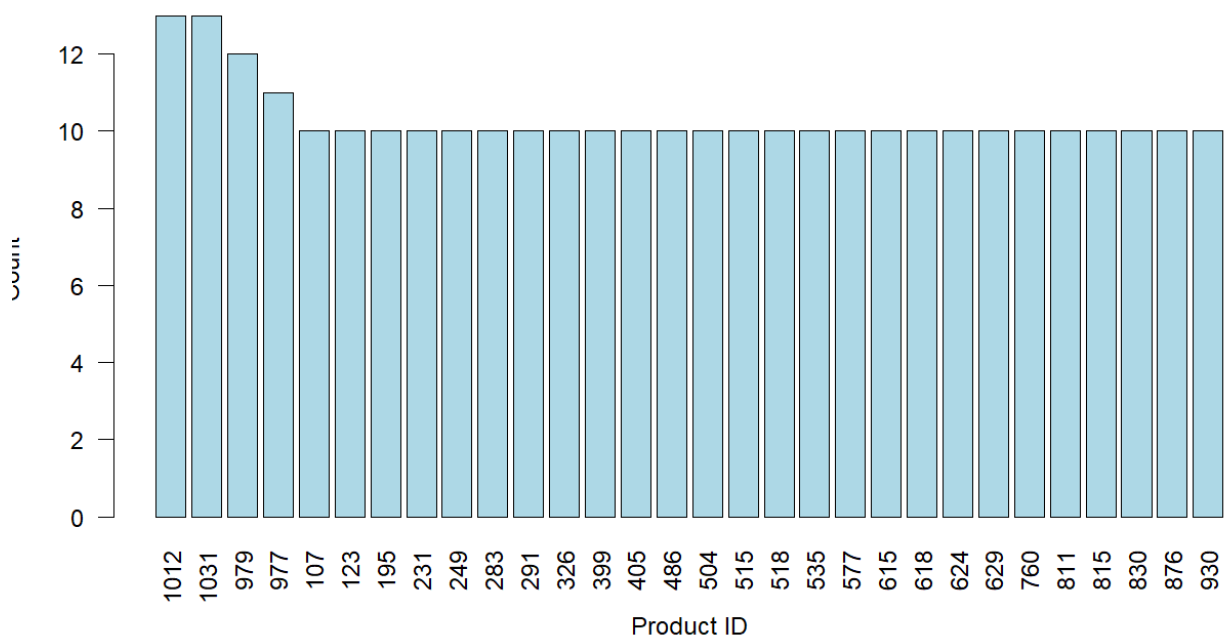
import warnings
warnings.filterwarnings('ignore')
```

APPENDIX 3



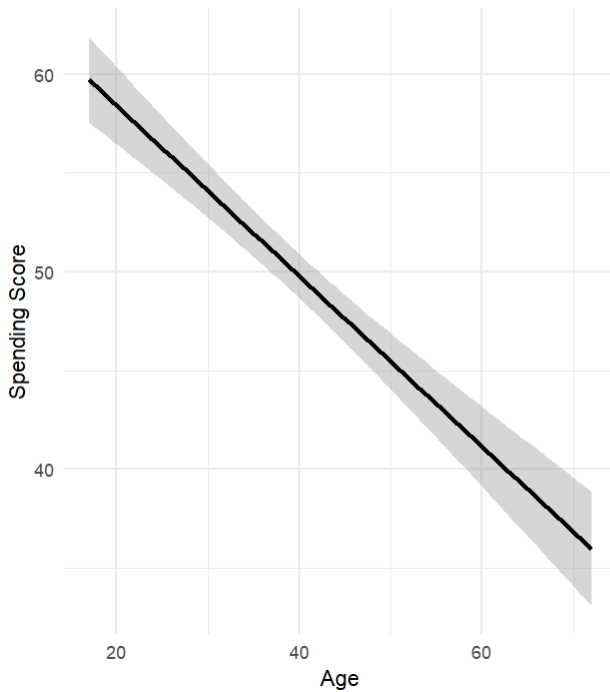
APPENDIX 9

Top 30 Most Popular Products



APPENDIX 10

Trend: Age vs Spending Score



Trend: Age vs Loyalty Points

