Turtle Games – Predicting Future Outcomes
Technical Report

Harshdeep Kohli

Contents

1. Background context

Turtle games, a global game retailer and manufacturer tasked me to analyse their customer data to better understand customer trends, with the aim of enhancing sales performance. The **core business questions** asked by the company were:

1. ***How do customers accumulate loyalty points?***
2. ***How can customers be segmented into groups?***
3. ***How can customer reviews be used to inform marketing campaigns?***
4. ***How can descriptive statistics be used to enhance insights?***

Business objective: Improve sales performance by understanding patterns and relationships within the data.

A **problem-solving framework** can be found in a Five Why's Diagram *(Appendix 1a)* and a SWOT analysis *(Appendix 1b)*

2. Analytical approach

This project leveraged predictive modelling and machine learning to help Turtle Games. In **Python** I used advanced techniques to explore the dataset, focussing on Multiple Linear Regression, K-Means Clustering and Natural Language Processing. A Decision Tree analysis was also performed which can be seen in the appendix *(Appendix 10)*. In **R** I used statistical analysis and modelling to inform Turtle Games of any significant patterns to help enhance sales performance.

Please refer to the appendix for the detailed data import and wrangling approach *(Appendix 2)*

Data import/wrangling

In Python, the data was loaded and wrangled in Jupyter Notebook using Pandas, visualizations were prepared using Matplotlib & Seaborn, predictive modelling was performed using Scikit-learn, and natural language processing was performed using the TextBlob & Vader libraries. For multiple linear regression, I created scatterplots to satisfy the 1st assumption of a linear regression model – linearity.

This was used understand potential variables that were influencing loyalty, which revealed spend and pay as significant predictors. For customer segmentation, K-Means proved to be useful in grouping customers based on their pay and spend, using the 'Elbow' and 'Silhouette' methods to determine the optimal number of clusters. Pairplots allowed me to visualize and interpret the clusters offering the opportunity to profile customers for targeted marketing campaigns. For natural language processing I analysed 'review' and 'summary' variables to extract sentiments and quantify top positive/negative sentiments using TextBlob polarity and Vader compound scores. By preprocessing the data, I was able to create word clouds and recognize top 15 frequently mentioned words to help Turtle Games direct resources to where it was needed.

In R, the data was loaded into R Studio using the tidyverse library. Summary statistics were calculated using the dplyr package and visualized using the ggplot2 package to help understand distributions and outliers with the aid of boxplots and histograms. Predictive modelling was performed using the lm() function to conduct multiple linear regression by testing assumptions and visualizing using scatterplots.

Data validation was performed, identifying that there were 0 **duplicates** and 0 **null values.**

Outliers – 266 outliers were identified in the loyalty points variable using R. Without any business context it is hard to ascertain the reason behind these outliers and how to handle them sensitively. As such any outliers were unchanged, and this is discussion to be had with Turtle Games. *(Appendix 6)*

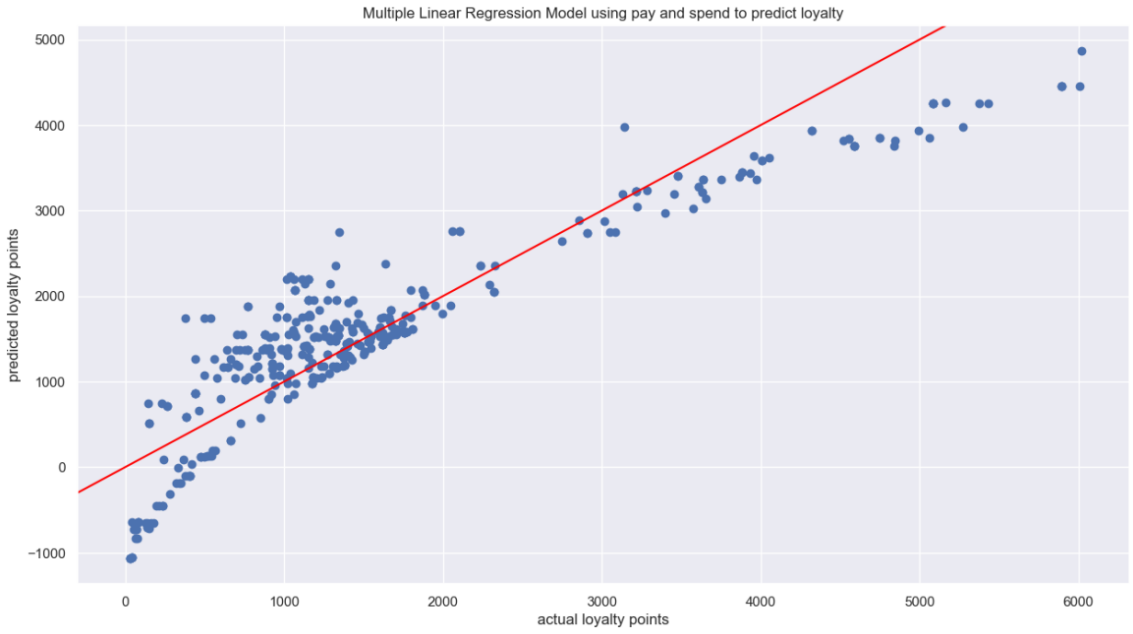
Assumptions – For a linear regression model, there are 6 assumptions that need to be satisfied. In the Jupyter Notebook you can see the evaluation of these assumptions.

Limitations – The dataset contains a column named 'product' which is a unique product code, but without any information into what the product is. This limits our understanding in how specific product categories are performing *(Appendix 11)*

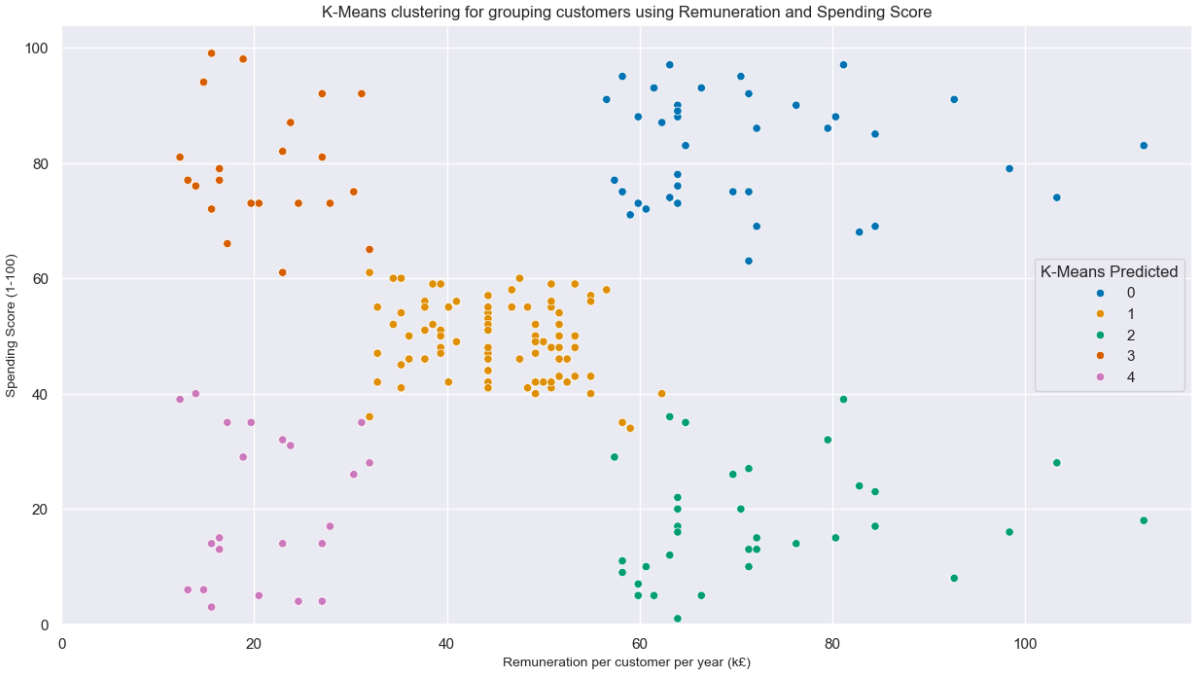
3. Visualizations & insights

In Python, my focus was on multiple linear regression, k-means clustering and natural language processing.

- To address the **1st business question**, a multiple linear regression analysis seemed pertinent. By using scatterplots, I was able to identify which independent variables demonstrated a linear relationship with our dependent variable 'loyalty points'. 'Spend and 'pay' showed the strongest positive linear relationship, as such these variables were chosen to run a multiple regression model. My model resulted in a R-Squared value of 83%, meaning that a significantly high variation in loyalty points can be explained by these variables, allowing Turtle Games to anticipate customer needs and behaviours. But this comes at the expense of the assumption of the homoscedasticity being violated. I also took steps to improve accuracy of this model by transforming the dependent variable by taking the natural log, but this did not improve accuracy and still violated MLR model assumptions. *(Appendix 3)*

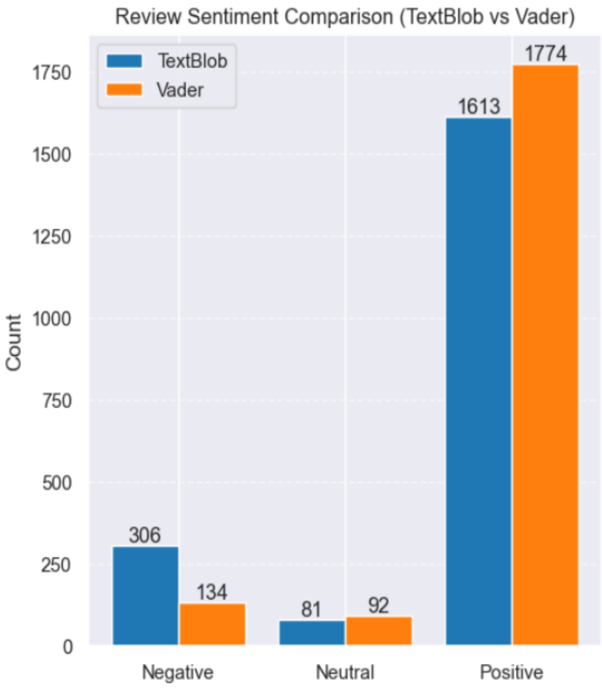


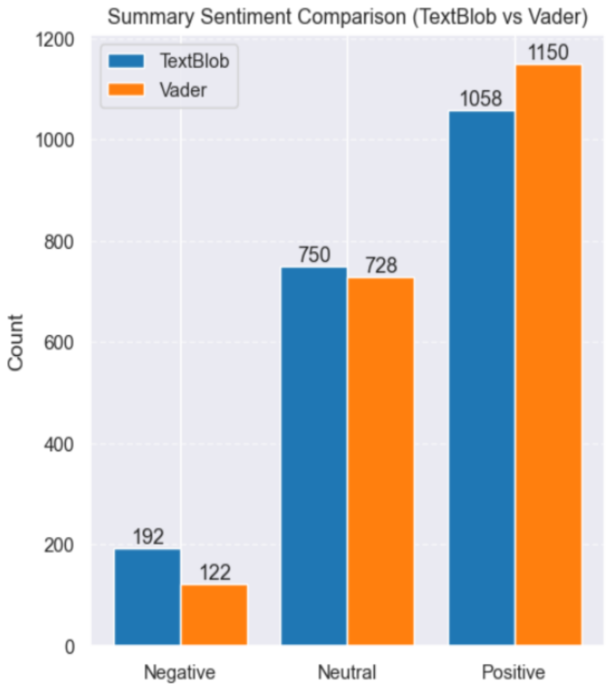
- To address the **2nd business question**, K-Means clustering was employed for customer segmentation. The first step was determining the optimal number of clusters (K) using elbow and silhouette methods. These methods determined that K=5 was the optimal number, which allowed me to assign a profile to the customers into 5 distinct groups which could then be used by Turtle Games for targeted marketing strategies *(Appendix 4)*



- To address the **3rd business question**, natural language processing was used generate word clouds to get an immediate sense of sentiment by visualising frequent positive words like ‘great’, ‘fun’, and frequent negative words like ‘difficult’, ‘disappointed’. TextBlob showed a slight positive correlation between review and summary sentiment, but with notable mismatches (positive review, neutral/negative summary). VADER showed no correlation, with frequent positive

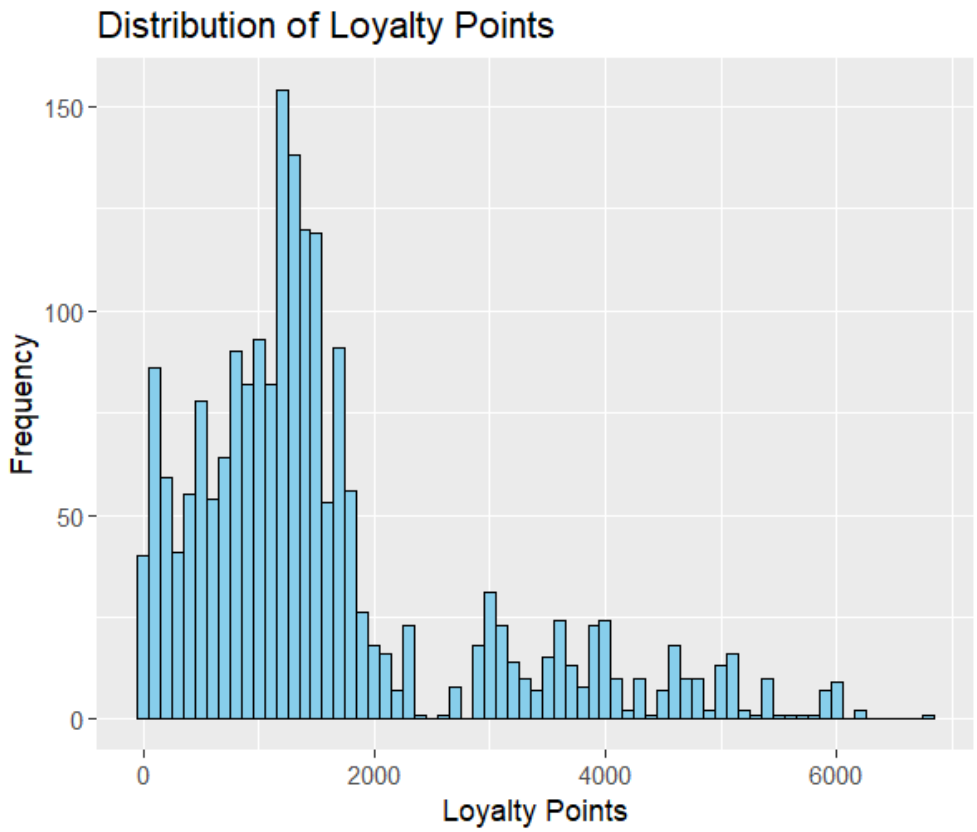
reviews paired with neutral summaries. Both tools agree on positive/neutral proportions, but TextBlob detects significantly more negative reviews and summaries than VADER *(Appendix 5)*



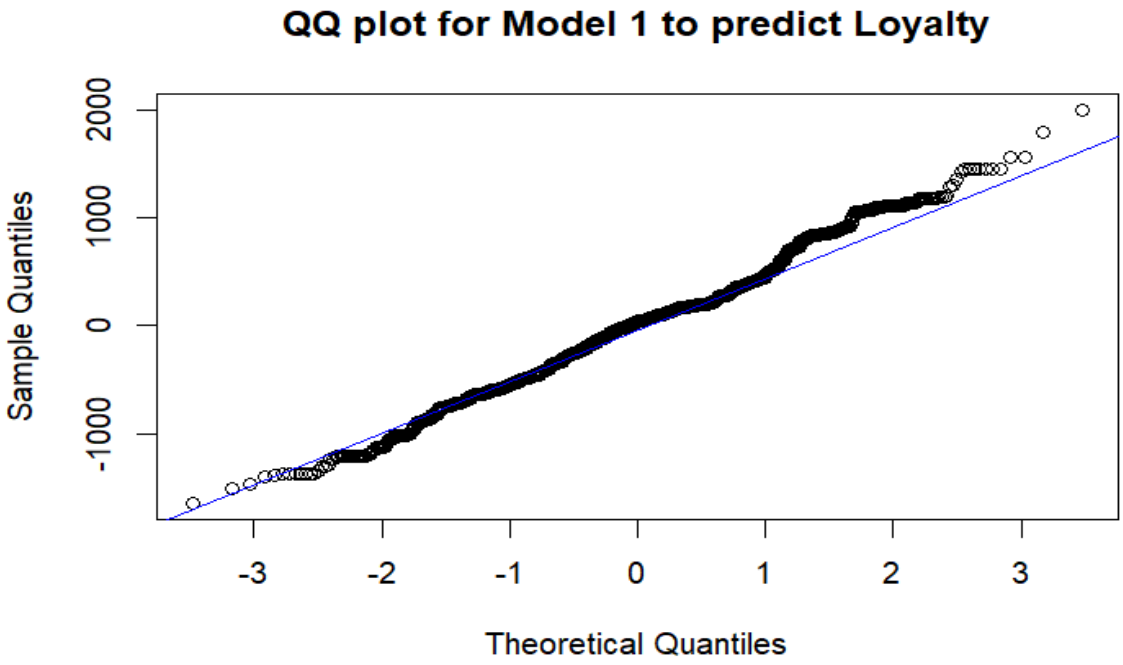


In R, my focus was on statistical analysis and multiple linear regression.

- R proved to be useful in understanding the shape and distribution of the data with the various visualisations which helps us answer our **4th business question**. In particular histograms identified that our dependent variable, loyalty, is unevenly distributed, skewed towards higher values due to a small group of highly loyal customers. This right-skewed, multimodal distribution, indicates distinct customer segments: new, regular, and highly engaged (*Appendix 6)*



- Multiple linear regression was performed again in R with the help of scatterplots to visualise relationships and errors. The model strongly predicts loyalty points, explaining 83% of the variance with high statistical significance. However, the model's predictions have an average error of 534 points and the wide range of errors suggests potential issues requiring further investigation *(Appendix 7)*



4. Recommendations

- A MLR model using pay and spend allows us to explain 83% of variation in loyalty points. The model's predictive power is questionable showing an RMSE of 548.6. However, the presence of heteroscedasticity and outliers which skew the data need to be addressed. Turtle Games should focus on improving the quality of their data, by investigating reasons for outlier presence and addressing heteroscedasticity.
- Leverage the 266 outliers as a strategic segment. On average they have about 4 times as many loyalty points suggesting that they are completely separate group and are most likely very valuable customers. Further investigation showed a strong positive correlation (79%) between pay and loyalty in the outlier data, especially for customers in the top 25% pay range exhibiting the highest average loyalty score. Turtle Games should focus marketing efforts on the high-pay customers.
- K-means clustering identified 5 distinct customer segments based of different levels of pay and spend. My recommendation is to tailor customized marketing strategies to each of these 5 segments. Please see the full details of the segments in the appendix (*Appendix 4)*
- While the data shows an overall positive sentiment, close attention needs to be paid to addressing recommendations from negative feedback to identify products that may need further attention. Turtle games should invest in customer experience by incorporating positively mentioned words in their advertising and addressing product quality and usability issues as highlighted in negative reviews.

*** Please find all insights and recommendations in the appendices**

(word count = 1259)

Appendix 1a

Problem – Turtle Games have observed that their sales performance has not been meeting quarterly targets.

Q1 – Why hasn't sales performance been meeting quarterly targets?

Answer – Because there is limited understanding of customer behaviour and engagement. Specifically, loyalty point accumulation and attitudes of customers towards Turtles Games' products.

Q2 – Why is there limited understanding of customer behaviour and engagement?

Answer – Insufficient and untidy data prohibiting efficient customer segmentation and targeted marketing.

Q3 – Why is there insufficient and untidy data?

Answer – Obsolete data systems and poor data collection methods resulting in limited insights.

Q4 – Why are the systems obsolete and why has there been poor data collection?

Answer – Turtle Games have not prioritised data driven decision making.

Q5 – Why have Turtle Games not prioritised data driven decision making?

Answer – Because they don't understand the benefits of data analysis and machine learning, which could ultimately help them to make informed decisions and improve sales performance.

Root cause – The root cause of the inadequate sales performance is due to the lack of understanding of the benefits of analysing data, and improving their data maturity by use machine learning tools to predict drivers of sales performance.

Appendix 1b – SWOT Analysis

STRENGTHS
Product Diversity: Broad range of offerings across multiple gaming categories (books, board games, video games, toys)
Vertical Integration: Both manufacturing and retail capabilities
Global Reach: Established customer base across multiple regions
Hybrid Business Model: Combination of own products and third-party sourcing provides flexibility
WEAKNESSES
Reliance on historical marketing strategies with limited personalisation and customer segmentation
Data Integration Challenges: Potential difficulties in connecting sales data with customer reviews for analysis
Market Fragmentation: Operating across diverse product categories may dilute focus and expertise
Data Analysis Capabilities: Potential gaps in analytical tools or expertise to fully leverage collected data
OPPORTUNITIES
Data-Driven Decision Making: Leverage existing customer reviews and sales data for strategic insights
Performance Optimization: Improve operational efficiency based on sales analysis
Targeted Marketing: Use customer data to create more effective promotional strategies
Product Line Refinement: Focus investment on highest-performing categories or products
THREATS
Market Consolidation: Larger competitors acquiring smaller companies to gain market share
Economic Volatility: Regional economic fluctuations affecting discretionary spending on games and toys
Supply Chain Disruptions: Manufacturing or sourcing delays affecting product availability
Rapidly Changing Consumer Preferences: Decreased product lifecycle requiring faster innovation

Appendix 2 – Data import/wrangling

Python

Prepare workstation

- Import necessary libraries

```
# Import libraries
import os
import numpy as np
import pandas as pd
import math
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import pylab
import seaborn as sns
import statsmodels.api as sm
from statsmodels.formula.api import ols
import statsmodels.stats.diagnostic as sms
from statsmodels.stats.outliers_influence import variance_inflation_factor
import scipy.stats as stats
from scipy.spatial.distance import cdist
from scipy.stats import norm
import sklearn
from sklearn import linear_model
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from sklearn.metrics import r2_score, mean_squared_error,mean_absolute_error,accuracy_score, confusion_matrix,classification_report,silhouette_score
from sklearn.tree import DecisionTreeRegressor, plot_tree
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import OrdinalEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
import nltk
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist
from nltk.corpus import stopwords
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from textblob import TextBlob
from wordcloud import WordCloud
import re
from collections import Counter
```

Import the data

- The name of the raw data file is turtle_reviews.csv
- Using the read_csv function from the pandas library to import the file

```
# Load the CSV file(s) as reviews.
reviews = pd.read_csv('turtle_reviews.csv')
```

Validate the data

- By creating a user defined function, I was able to check the data was loaded correctly
- This function checks the number of rows & columns, null values, duplicates, column data types.

```python
# data validation function
def validate_data(df):
    """
    Performs basic data checks on a Pandas DataFrame.

    Args:
        df: The Pandas DataFrame to validate.

    Returns:
        A dictionary containing the results of the data checks to easily access individual elements
    """

    results = {}

    results['shape'] = df.shape  # determine number of rows and columns in the df
    results['null_values'] = df.isnull().sum()  # determine number of null values in the df
    results['duplicate_values'] = df.duplicated().sum()  # determine number of duplicates in the df
    results['data_types'] = df.dtypes  # determine column data types in the df

    return results
```

Data cleaning

- Due to the absence of null values, there was limited cleaning to do.
- I created a new data frame, keeping only the relevant columns. As the 'Platform' column only had 1 attribute (web), and the 'Language' column only had 1 attribute (EN), they were not needed and therefore dropped.

```python
# Drop unnecessary columns.
reviews.drop(['language', 'platform'], axis=1, inplace=True)

# View columns names.
reviews.info()
```

- I also certain **renamed** columns for simplicity.
  - **Renamed 'remuneration (k£)' to 'pay'**
  - **Renamed 'spending_score (1-100)' to 'spend'**
  - **Renamed 'loyalty_points' to 'loyalty'**

```python
# Rename the column headers.
reviews.rename(columns={'remuneration (k£)':'pay','spending_score (1-100)':'spend','loyalty_points':'loyalty'}, inplace=True)

# View column names.
reviews.columns
```

- As our data validation check returned 0 duplicate rows across all rows, no data was removed. However, there is a **limitation** in the dataset as we don't have a unique identifier to tell if the row is a true duplicate.

- However, when we look at natural language processing, we will do a duplicate check again and decide what to do.

Data export of cleaned file

- As we performed some basic data cleaning, it would be pertinent to save and export this as a new CSV file for a hard backup copy.

```
# Create a CSV file as output - rename the updated data frame as turtle and save a CSV copy to our computer
reviews.to_csv('turtle.csv', index=False)
```

R

Prepare workstation

- Import necessary libraries
- The tidyverse library is the generally the main one we need for data manipulation, analysis and visualization as it contains a variety of packages including dplyr & ggplot.

```
# import libraries
library(tidyverse)
library(skimr)
library(DataExplorer)
library(moments)
library(ggcorrplot)
library(car)
library(lmtest)
library(rstatix)
```

Import the data

- As we have already exported a 'cleaned' version of the raw file we can import that into R studio.
- Using the read.csv function from the tidyverse library to import the file.
- Our data frame in R will be called: turtle

```
# read the csv file (name of the csv file is 'turtle.csv')
turtle <- read.csv('turtle.csv',header=TRUE)
```

Validate the data

- Due to the need for only working on 1 cleaned dataset, a user defined function was not needed to validate the data. As such, we validated the data step by step, including checking for number of rows, null values, duplicates, column data types and summary statistics.

```
### DATA VALIDATION ###

# view the dimensions of the data frame
dim(turtle)

# view the column data types of the file
str(turtle)

# view the column names
colnames(turtle)

# view number of missing values in the whole data frame
sum(is.na(turtle))

# view the number of duplicate values in the whole data frame
nrow(turtle[duplicated(turtle), ])

### SUMMARY STATISTICS ###

# view the data summary using the skim package
skim(turtle)

# create a summary report using the DataExplorer library
DataExplorer::create_report(turtle)
```

Data cleaning

- As we already imported the 'cleaned' version of the file, data cleaning wasn't required, although we may manipulate the data as we go deeper in our analysis.

Appendix 3 – Linear Regression (Python)

Turtle Games 1st core business question was **How do customers engage with and accumulate loyalty points?**

'The economic implications of customer loyalty manifest in several direct, tangible benefits. These benefits not only contribute to a company's profitability but can also drive its growth and sustainability.' [1]

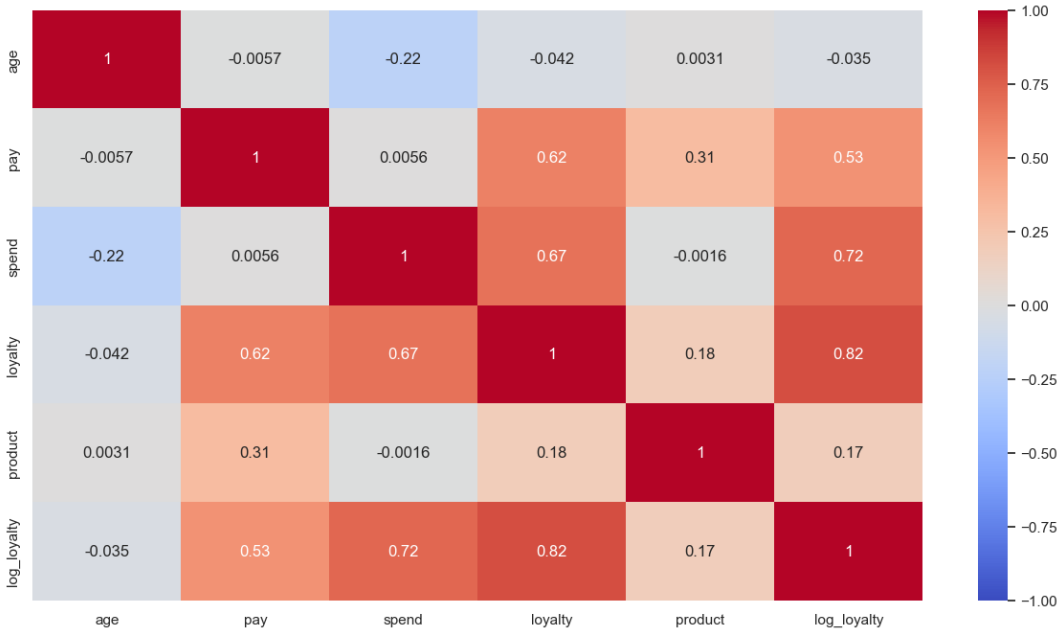
For this analysis, I examined the relationships between independent variables in our dataset to our dependent variable, 'loyalty points'.

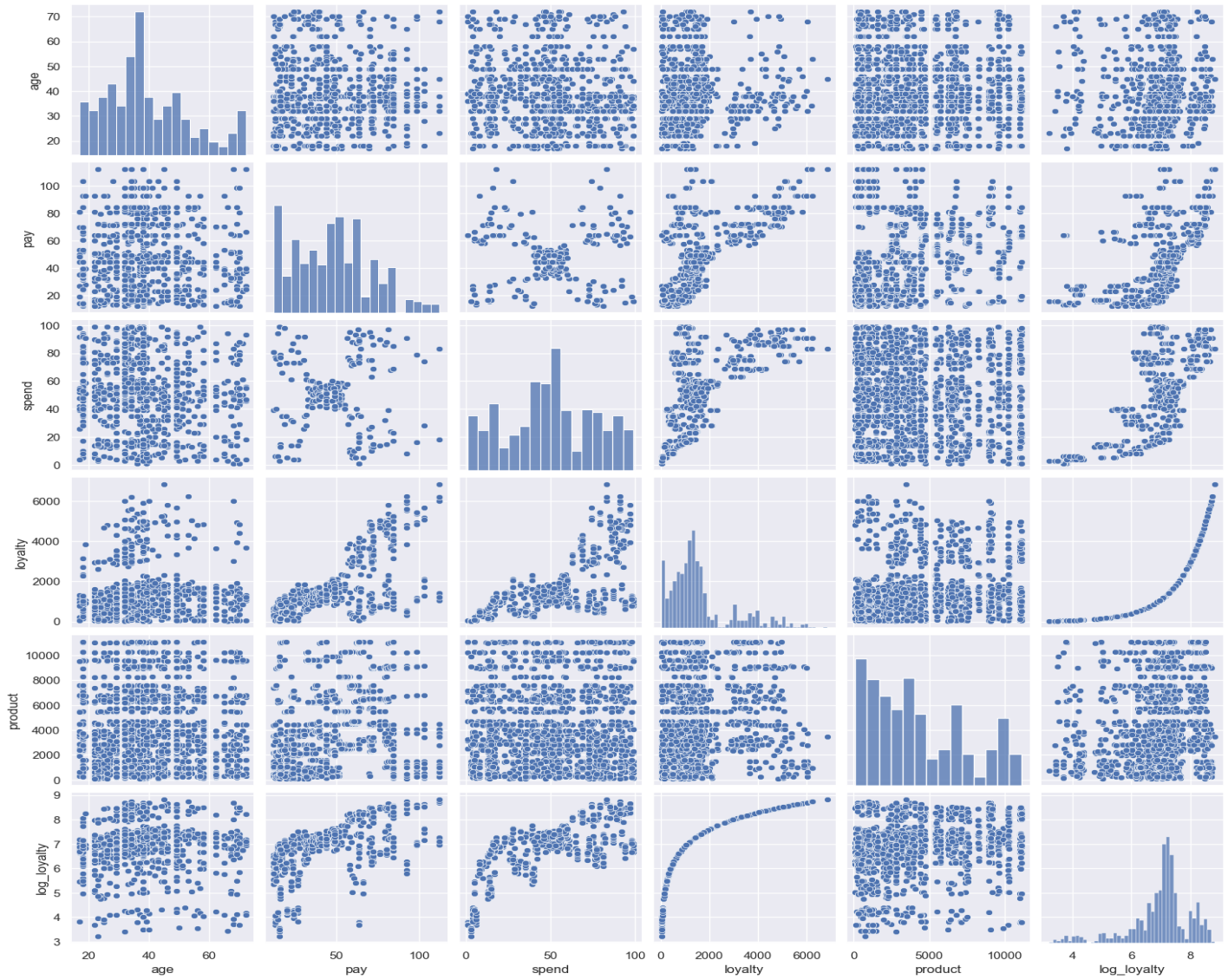
I conducted exploratory analysis and calculated descriptive statistics using NumPy, Pandas, Matplotlib, Seaborn libraries. This was used to sense check the data, understand distributions using histograms and boxplots and explore relationships using scatterplots.

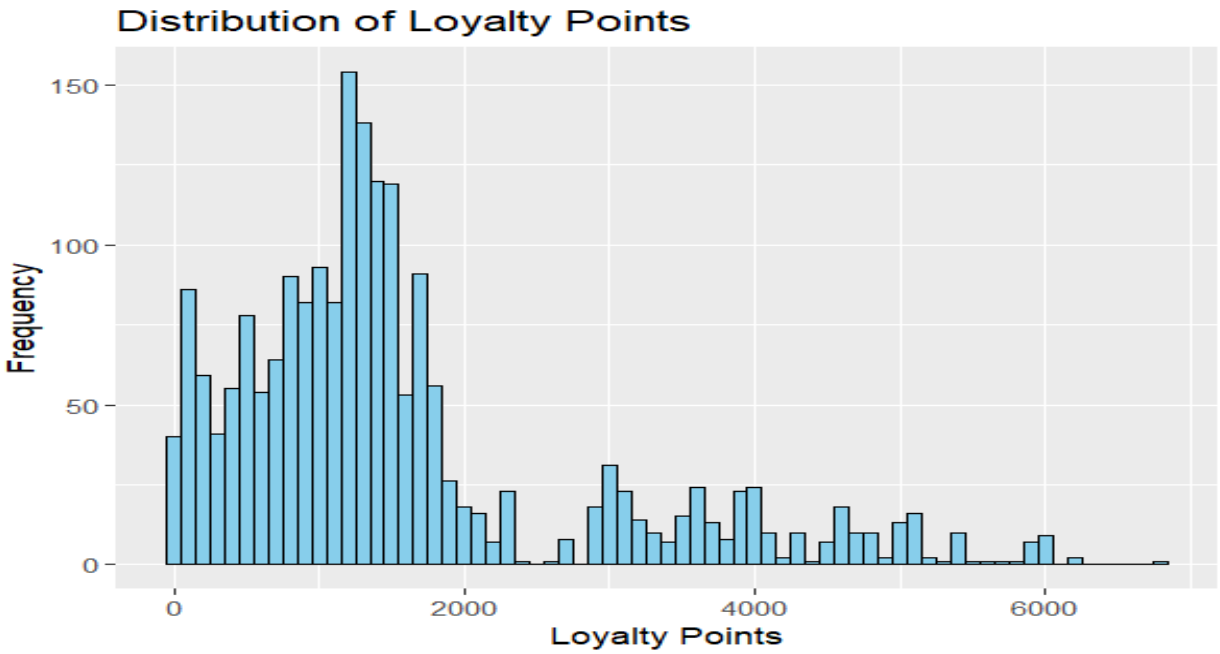
With the help of a histogram, I could see that that our dependent variable, 'loyalty points' is positively skewed. What this means is that majority of the customers generally have lower loyalty points, and there is a smaller group of customers with higher loyalty points.

The relevant numeric independent variables are: age, spend (spending score) and pay (remuneration). For the purpose of examining the relationship of these independent variables, I used a scatterplot with the X axis as each of our independent variables and the Y axis as our dependent variable, loyalty (loyalty points). Correlation analysis and scatterplots confirmed the following:

- Spend is likely to influence loyalty (correlation = 67%)
- Pay is likely to influence loyalty (correlation = 61%)
- Age is not likely to influence loyalty (correlation = -4%)




age
pay
spend
loyalty
product
log_loyalty


Distribution of Loyalty Points
Frequency
Loyalty Points
0
50
100
150
0
2000
4000
6000

Simple Linear Regression

The first step of a linear regression model is generally simple linear regression, which is using 1 independent variable to predict a dependent variable. Thus, I created a model for each of my independent variables (age, spend, pay) with my dependent variable (loyalty). I also went a step further by transforming the dependent variable (loyalty) by taking the natural LOG in an attempt to reduce heteroscedasticity.

- **Spend to predict loyalty**
  This model generated an R-Squared of 45% and statistically significant P value. This tells us that only 45% variation in loyalty points can be explained by spend. This is not enough to be reliable so perhaps we need to look at other variables or consider adding more variables. Evaluating this model with a Breusch-Pagan test also showed presence of heteroscedasticity, which is the unequal variance of residuals and this violates an assumption of a linear regression model. However, this can be addressed by transforming our dependent variable by taking the natural log. This will be explored below.

- **Spend to predict LOG loyalty**
  This model generated an R-Squared of 52% and statistically significant P value. This tells us that only 52% variation in loyalty points can be explained by spend. This is better than model 1, but still not good enough. Evaluating this model with a Breusch-Pagan test also showed presence of heteroscedasticity, which undermines the validity of my model and so taking the natural log didn't have any significant effect.

- **Pay to predict loyalty**
  This model generated an R-Squared of 38% and statistically significant P value. This tells us that only 38% variation in loyalty points can be explained by pay. This is worse than model 1 and model 2, suggesting pay isn't practical to use to predict loyalty. Evaluating this model with a Breusch-Pagan test also showed presence of heteroscedasticity. With a worse R-Squared, it is not worth transforming the dependent variable using the natural log.

- **Age to predict loyalty**
  This model generated an R-Squared of 2% and statistically significant P value. This tells us that only 2% variation in loyalty points can be explained by age. This is poor and was also confirmed by the nonlinear relationship between these 2 variables, and thus I decided no longer to explore this relationship any further.

Based on my analysis, the addition of more variables seems like the best solution.

Multiple Linear Regression (MLR)

Having seen that we can improve our model's predictive power to achieve a higher R-Squared, I decided to employ multiple linear regression. This is the process of using 2 or more independent variables to predict a dependent variable. In our case, using pay and spend to predict loyalty points.

Before we begin, we must make ourselves familiar with the **6 assumptions** of a MLR model. These assumptions need to be evaluated to test the accuracy and validity of our models. In my MLR model, I demonstrate evaluating all 6 assumptions. Let's remind ourselves of what these assumptions are:

Assumptions of MLR model

1. **Linearity: The relationship between the independent and dependent variables should be linear.**
2. **Independence of Errors: The errors (residuals) should be independent of each other, meaning the error in one observation doesn't predict the error in another.**
3. **Homoscedasticity: The variance of the errors should be constant across all levels of the independent variables.**
4. **Normality of Errors: The errors should be normally distributed.**
5. **No Multicollinearity: The independent variables should not be highly correlated with each other.**
6. **No Autocorrelation: The errors should not be correlated with each other over time or in some other sequence.**

Note – Full evaluation of these 6 assumptions can be seen in the Jupyter Notebook

- **Pay and spend to predict loyalty**

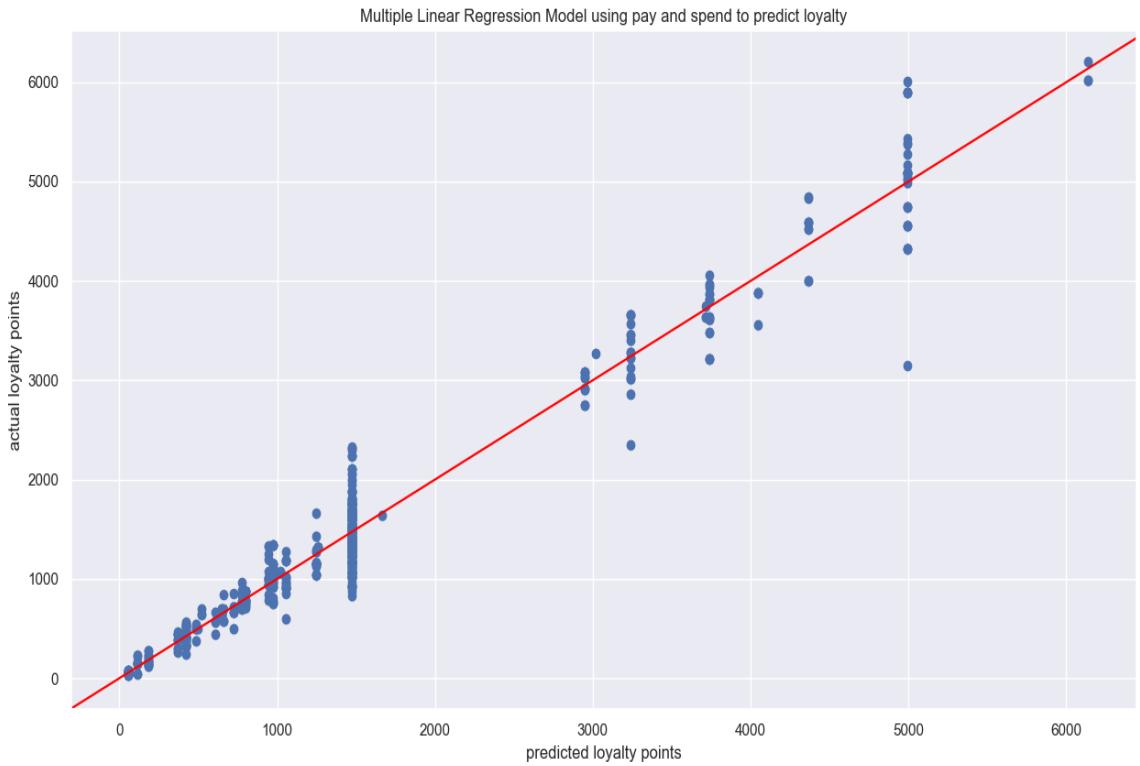
This model generated an R-Squared value of 83% and statistically significant P value. We also introduce the adjusted R-Squared metric here, which is a better metric to use when introducing more variables in a MLR model as it penalizes the inclusion of unnecessary variables. Our adjusted R-Squared value is also 83% showing that adding 1 more variable to our model didn't negatively impact the fit of our model. This tells us that 83% variation in loyalty points can be explained by pay and spend. This is extremely better than our previous models and could be a of great use to Turtle Games to predict loyalty points.

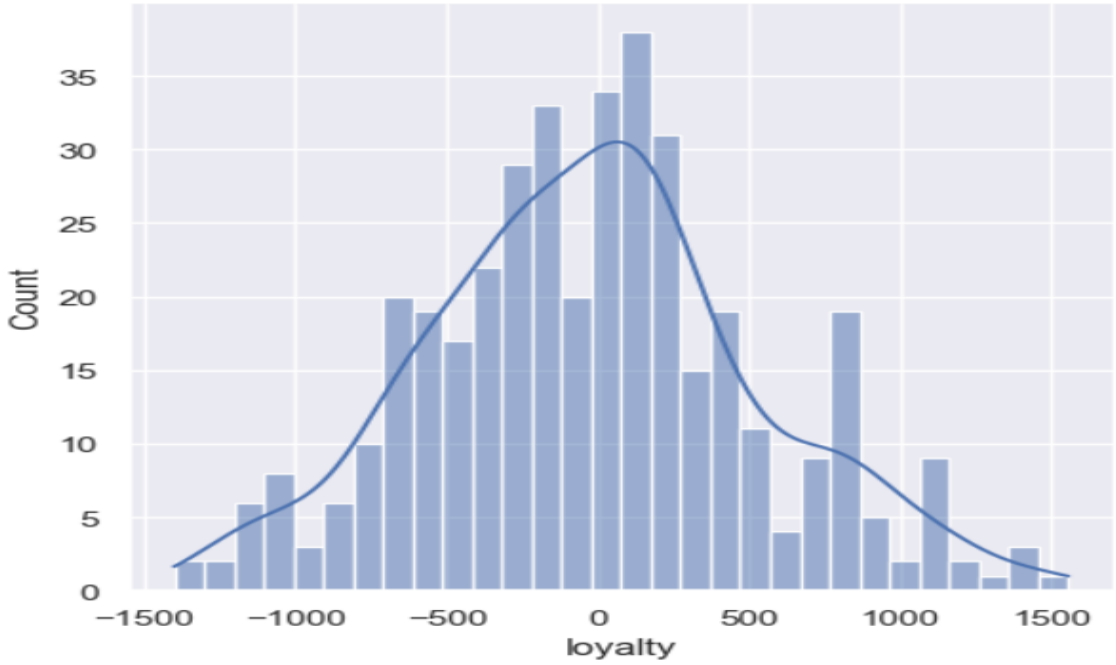
A general rule of thumb is a R-Squared value of 60% is higher is often considered a good fit in many contexts, however context matters and it depends on the field of study and the nature of the data. As we are interested in a retail business, R-Squared values above 80% is generally considered a strong fit. Thus, in our case we've got an excellent model.

Evaluating this model by testing the assumptions listed above, I found that there was no multicollinearity between pay and spend, indicated by a VIF value of 1. Comparing this result to the correlation analysis shows a 0.06% correlation between pay and spend, so this was expected. However, heteroscedasticity does seem to exist because of such a low LM test P value indicated by the Breusch-Pagan test. Checking for the normality of residuals assumption, I created a histogram and Q-Q plot to check for normality and found that whilst it is not a perfect bell curve, its close enough, suggesting that most of our predictions are on average correct. The left tail shows a decent descent where the right tail start to falter. This shows us that this model tends to overestimate loyalty points. Evaluating the 'mean of residuals' assumption, which states that the average error of

the model is 0, my model indicated a mean of -23. Whilst this is not 0, it isn't massively far off, thus we can conclude that this assumption has also been met. We cannot test for the No Autocorrelation assumption as our dataset does not have any date/time variables.

Evaluating the errors from this model, the root mean squared error (RMSE) is a popular metric to interpret the average error size in the same units as the target variable, while also penalizing larger errors. Our model has a RMSE of 548.6. What this tells us that on average, our predicted value for loyalty points is off by 548.6 points compared to our actual value. Given that the average of our loyalty points variable is 1578, being off by 548.6 points is quite high and again undermines the robustness of this model.






```
Mean Absolute Error (MAE): 429.66362016909113
Mean Squared Error (MSE): 300944.0917834269
Root Mean Squared Error (RMSE): 548.5837144715716
```

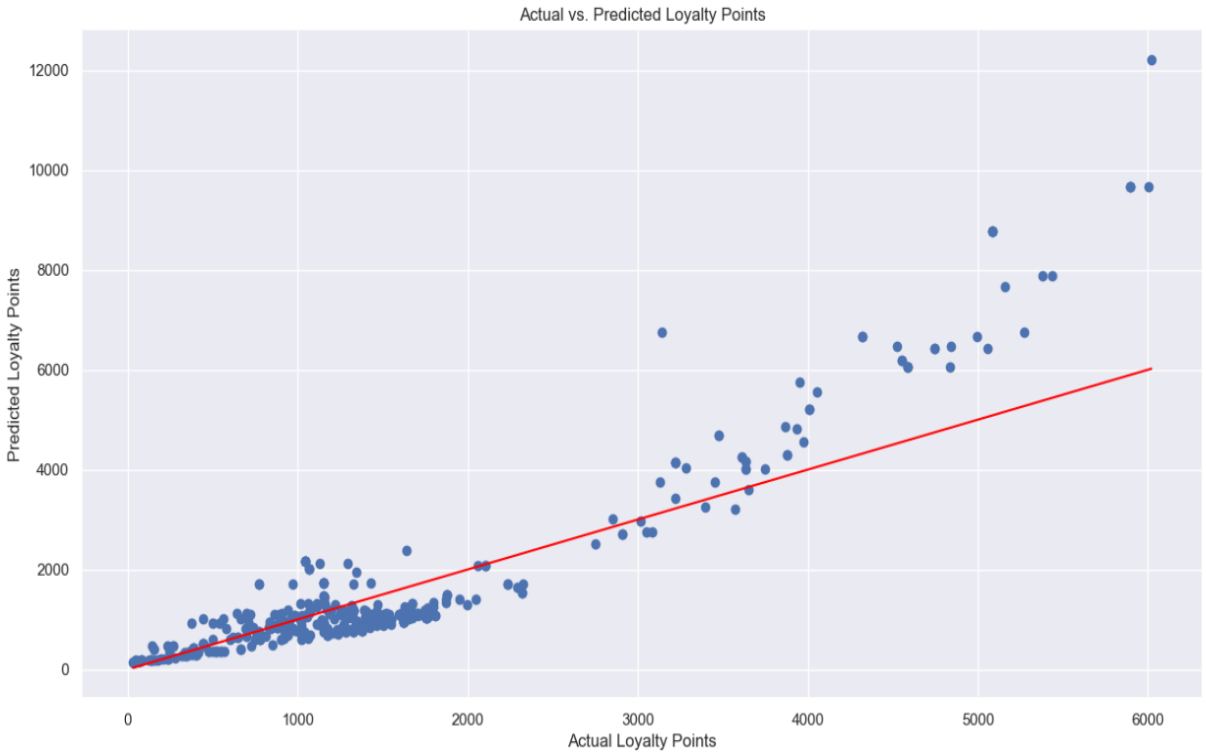
- **Pay and spend to predict LOG loyalty**

I also created another model by taking the natural LOG of loyalty in an attempt to reduce heteroscedasticity, however this model generated a lower R-Squared value of 80% which is lower than our previous model without taking the LOG of loyalty, and there was still presence of heteroscedasticity.

I then scaled by results back to the original scale by exponentiating the log transformed data for comparison against the original loyalty point data. Evaluating the residuals in this model showed that the RMSE error also increased (864). This is worse than our previous model. So, transforming the data did not help in developing a better model, most likely due to the presence of outliers that will need to be addressed.

When data is normally distributed, it simplifies analysis and interpretation because many statistical methods are designed with this assumption in mind. However, achieving a perfect normal distribution in data is not representative of a real-life business scenario. Many businesses will want some skewness in their data, such as a small group of highly engaged customers, helping increase average sales for the business.

In the case of Turtle Games, we see that there is some positive skewness in the loyalty points variable, which tells us that there does exist a small group of customers who accumulate higher loyalty points, thereby pushing up the average. This skewness will be explored later in a statistical analysis and discussions will need to be had with Turtle Games on how to improve data quality to generate better predictive models.




```
MAE (Original Scale): 503.392096858693
MSE (Original Scale): 746600.9140758816
RMSE (Original Scale): 864.06071203121
```

Key insights

- As we saw in our correlation analysis and scatter graphs right at the start of the analysis, there are only 2 meaningful variables that have a significant relationship with loyalty. These are pay and spend.
- These variables had a significant positive correlation with loyalty, simply implying that if you have higher pay (income) and spend (spending score), this will be reflected with higher loyalty points.
- In our simple linear regression model, we observed an R-Squared of 45% for spend, and 38% for pay. A model using these variables individually would not have had any real use case for Turtle Games to predict loyalty, but by combining the 2 together in an MLR model we have a much healthier R-Squared of 83%. This means that 83% of the variation in loyalty points can be explained by spend and pay.

- By LOG transforming the loyalty variable, we observed a slightly lower R-Squared (80%), and a higher RMSE (864) meaning that this transformed model didn't prove to be better.

  So, to conclude, our model with Pay and Spend to predict Loyalty with an R-Squared of 83% is currently the best one. If Turtle Games want better prediction accuracy, then they will need to invest in their data quality or try using alternative models to better capture the relationships between pay and spend in predicting loyalty.

Recommendations

- Presence of heteroscedasticity, high RMSE are issues that need to be addressed. Rather than adding more variables to the model, Turtle games should focus on increasing the accuracy of these variables, possibly by evaluating their data. Better data will lead to better models.
- By identifying factors that lead to higher loyalty points, Turtle Games can create strategies to improve customer retention. They could offer special promotions or loyalty rewards thresholds to customers who are at risk of churning but have higher spending scores or pay.
- This model can be used for predictive analytics helping Turtle Games forecast future loyalty point accumulation based on expected changes in customer pay and spend

Appendix 4 – K-Means Clustering

Turtle Games 2nd core business question was **How can customers be segmented into groups?**

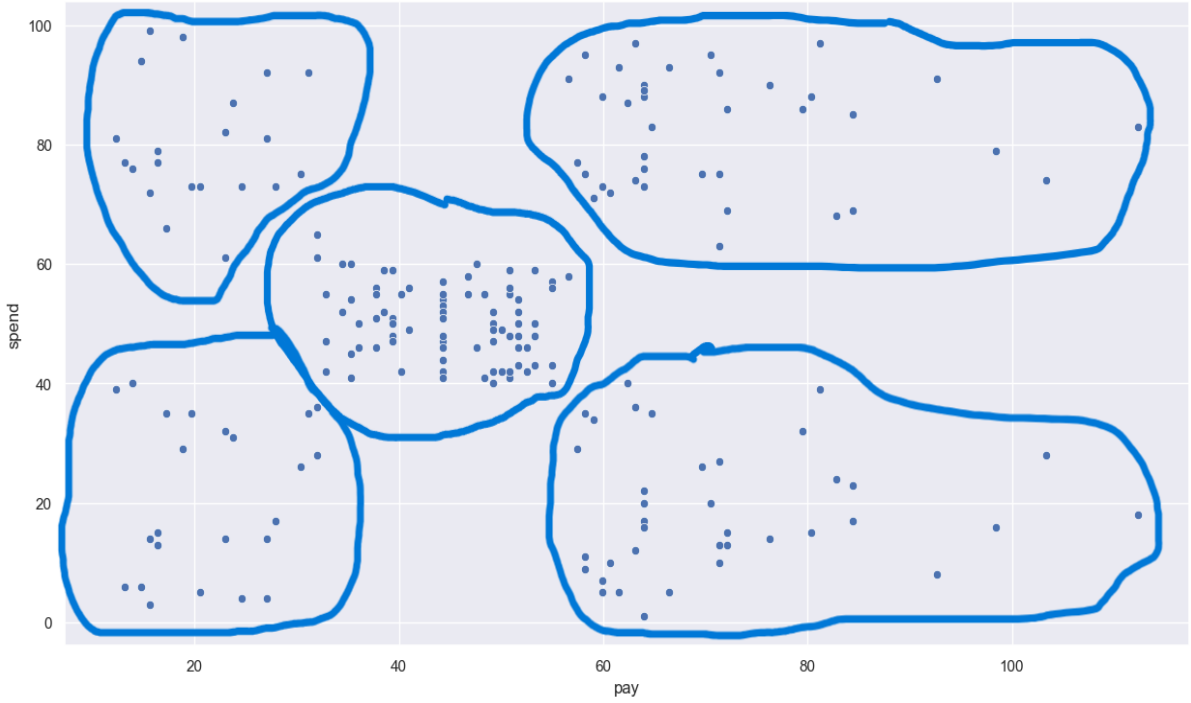
'KMeans is used across many fields; some examples of clustering use cases include customer segmentation, fraud detection, predicting account attrition, targeting client incentives, cybercrime identification, and delivery route optimization. The KMeans clustering algorithm is increasingly being used where enterprises are trying to infer patterns and optimize service offerings' [2]

In the exploratory analysis phase of this project, I created scatterplots to identify relationships between numerical variables. This revealed something very interesting for particularly 2 variables from our dataset, spend and pay. Data points for these 2 variables were grouped in separate blocks. These can be thought of as 'clusters' and we can employ an algorithm called K-Means Clustering to help us identify common data points and group them together. By grouping them together we can learn about the characteristics of each group and effectively build a profile for each group so they can be targeted using marketing strategies.

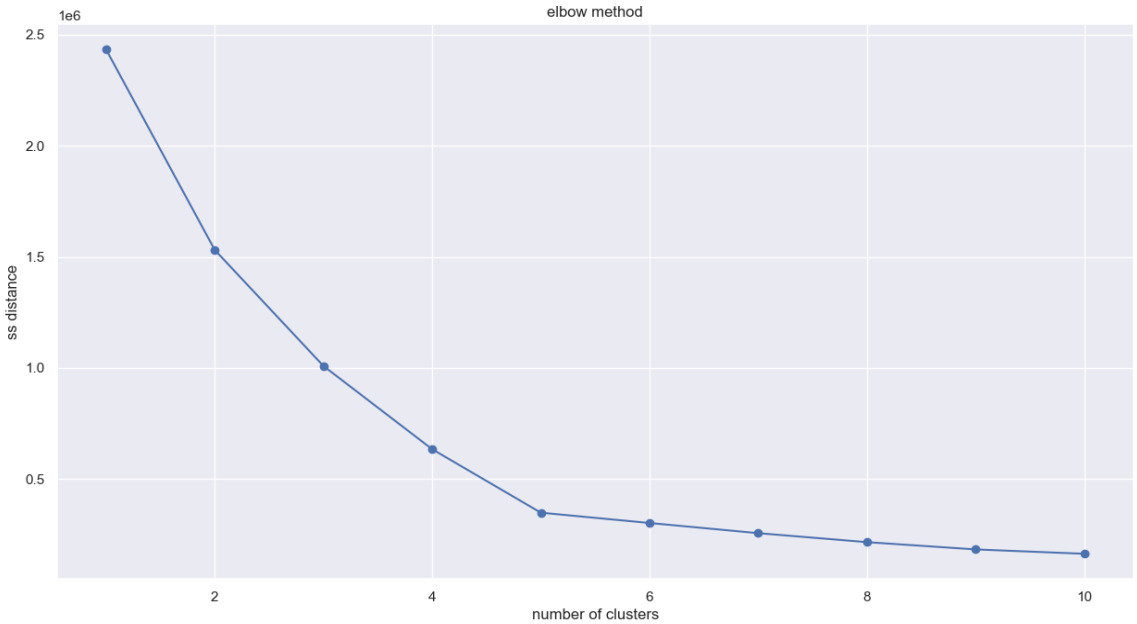
For a K-Means algorithm to work, you initially need to guess the value of K, which represents the number of clusters, and iterate by testing various levels of K until you reach a balanced conclusion for the best value. However, there are 3 tools to help us guess the value of K.

1. **Scatterplots**
2. **Elbow method**
3. **Silhouette method**

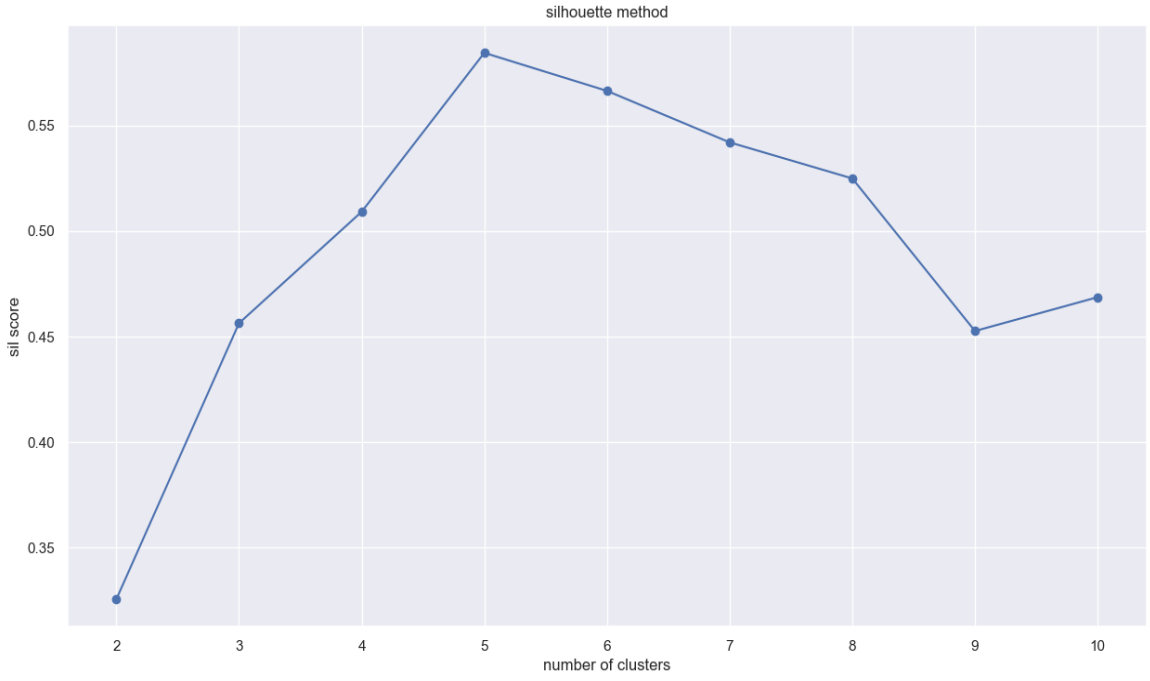
Scatterplots help us visualize our data, and have a guess at how many groups we can see in the data, as shown below by the scatterplot there seems to be 5 groups.



We can extend this further by using the elbow method, which calculates the within-cluster sum of squares (WCSS) for varying numbers of clusters (k). The "elbow" point, where the rate of decrease in WCSS sharply diminishes, is considered an estimate of the optimal k, balancing compactness and number of groups. As shown in our below graph, the ‘elbow’ seems to be at K=5.



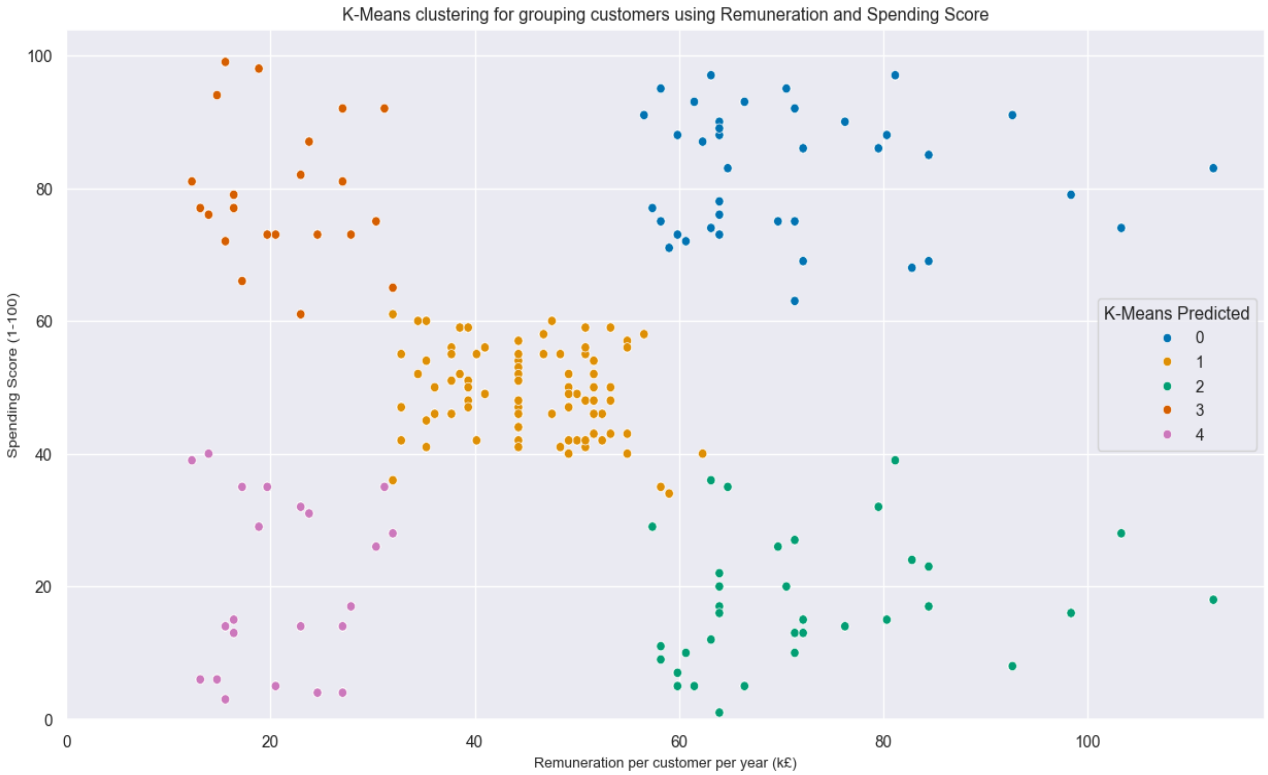
We can also use the silhouette method which evaluates the quality of clustering by measuring how similar each point is to its own cluster compared to other clusters. A silhouette score, ranging from -1 to 1, is calculated for each point. Higher average silhouette scores across all points indicate better-defined and well-separated clusters, aiding in the selection of the optimal k. As shown in our below graph, the maximum silhouette scores seem to be at K=5.



Which is the BEST number of clusters?

For the purpose of our analysis, I tested 3 different clusters (k=3, k=4, k=5), For K=3, we observed that the algorithm incorrectly classified some clusters and with K=4 we observed some overlapping. With K=5, the issues about incorrect classification and overlapping was diminished, and our methods for choosing K led us to believe that K=5 was the best which in fact it was. As all our methods pointed towards 5 as the optimal number of clusters, I was able to produce the below chart and use this to profile each customer group.

NOTE - The full evaluation of all 3 values of K can be seen in the Jupyter Notebook



Customer segmentation insights and recommendations for Turtle Games

Cluster 0 (Blue) - High Income, High Spending Score - "Luxury Enthusiasts" (18% of customers)

- Summary: Customers in this group have high remuneration (~ 60k€-105k€) and high spending scores (~ 60-100).
- Insight: Likely represents high-value customers who have significant purchasing power and spend a lot.
- Action: VIP customer service, premium loyalty programs, exclusive offers.

Cluster 1 (Orange) - Medium to Low Income, Medium Spending Score - "Value Seekers" (39% of customers)

- Summary: This cluster has remuneration ranging from ~30k€-60k€ and spending scores between ~40-60.

- Insight: Represents the general consumer group, spending moderately.
- Action: Marketing Strategy: Regular promotions, discounts, and personalized offers to encourage loyalty.

Cluster 2 (Green) - High Income, Low Spending Score - "Practical Professionals" (17% of customers)

- Summary: Customers here have high remuneration (~ 60k€-105k€) but low spending scores (~ 0-40).
- Insight: These customers can afford to spend but choose not to.
- Action: Marketing Strategy: Investigate barriers to spending (e.g., preferences, brand perception), offer tailored promotions.

Cluster 3 (Red) - Low Income, High Spending Score - "Aspirational Shoppers" (13% of customers)

- Summary: Customers in this group have low remuneration (~ 10k€-30k€) but high spending scores (~ 60-100)
- Insight: They may prioritize discretionary spending despite a lower income.
- Action: Marketing Strategy: Payment plans, installment options, or budget-friendly product lines.

Cluster 4 (Purple) - Low Income, Low Spending Score - "Budget-Conscious Consumers" (14% of customers)

- Summary: Customers in this group have low remuneration (~ 10k€-30k€) and low spending scores (~ 5-40).
- Insight: They are budget-conscious and spend very little.
- Action: Marketing Strategy: Cost-effective deals, discounts, and affordability-focused promotions.

Key Takeaways

- Cluster 0 (high earners & high spenders) are the most valuable customers and effort needs to be made to risk not losing them such as premium loyalty programs.
- Cluster 1 (moderate earners & moderate spenders) are the average customer, need to keep these happy through personalised offers to ensure sustainable business growth.
- Cluster 2 (high earners & low spenders) represents an opportunity to increase spending.
- Clusters 3 & 4 (low earners) are price-sensitive and need budget-friendly marketing.

Appendix 5 – Natural Language Processing (NLP)

Turtle Games 3rd core business question was **How can customer reviews be used to inform marketing campaigns?**

'Harnessing the power of review insights analysis empowers businesses to understand their customers more deeply. By systematically examining customer feedback, companies can uncover valuable insights into customer behaviors, preferences, and pain points. This understanding fosters better decision-making, allowing for improved products and services that align closely with customer needs.' [3]

To analyse free text data will use Natural Language Processing libraries such as TextBlob and VADER to get an idea of sentiment. NLP uses machine learning to reveal the structure and meaning of text so it is a particularly useful tool in sentiment analysis.

For this analysis we will be considering the columns 'review' and 'summary' from the data. A review being the customer's review of the product and the summary being the shortened 'summary' of the customer review.

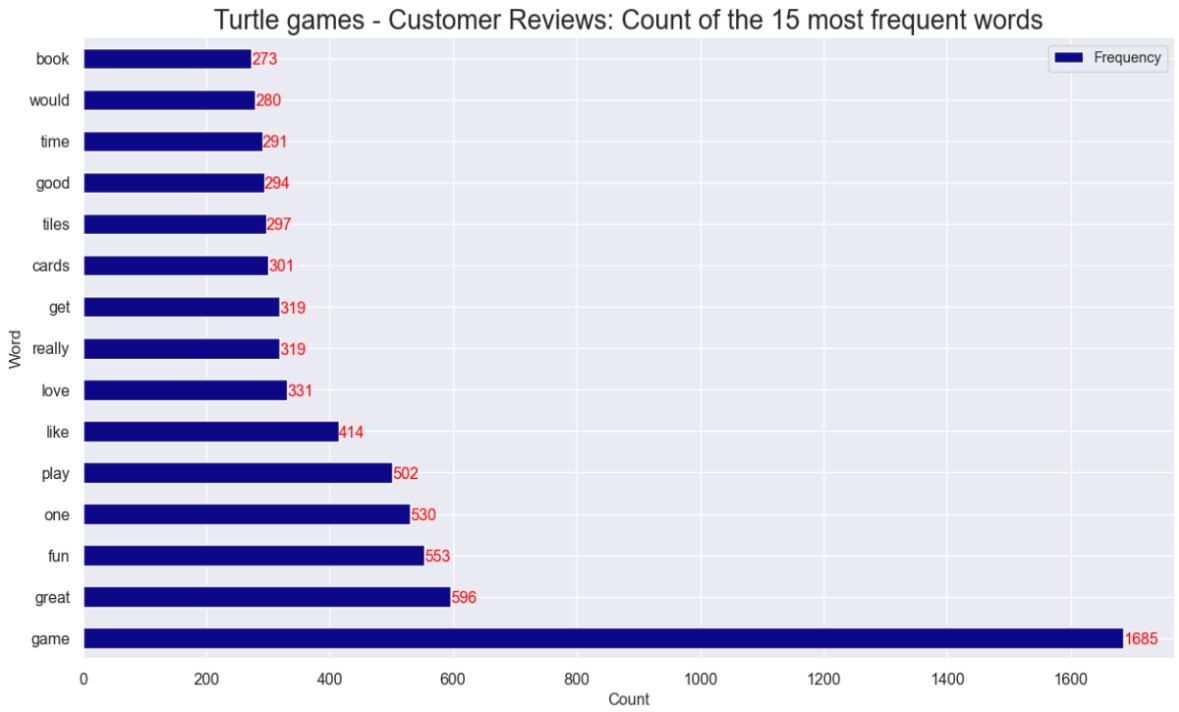
This exploration includes essential NLP data cleaning tasks such as converting words to lower case, removing punctuation, tokenizing words, removing alphanumeric characters and stop words. We can then give our models this processed data to retrieve a quantifiable metric such as TextBlob polarity/subjectivity scores and Vader compound scores to identify how positive, negative or neutral and objective or subjective a piece of text is. This was used to generate word clouds and visualize the top 15 most common words.

Note: 18 **duplicates** were identified in the review and summary columns. The decision was made not to drop them from the data, as this only represents 0.89% of the data and is highly unlikely to cause any significant changes to the skewness of the sentiment. It's also possible that the reviews were written by different people, using the same words and therefore the data was left unchanged.

Plotting word clouds of the review and summary columns gives us a quick indication of what the overall sentiment might be. From initial inspection there seems to be an overall positive sentiment.

The most frequent 15 words showed us the following insights:

• **Satisfaction**: The popularity of words like "fun," "love," "like," "good," and "great" suggests customers have a positive experience with the games.
• **Engagement Level**: The word "play" being one of the most common indicates that the act of playing is a central theme in the feedback. This reflects active engagement with the games.
• **Repetitive Quality**: The word "game" appears the most often, confirming that the feedback is focused on the gameplay experience. The high occurrence of "time" shows a repetitive acknowledgment of quality or enjoyment.

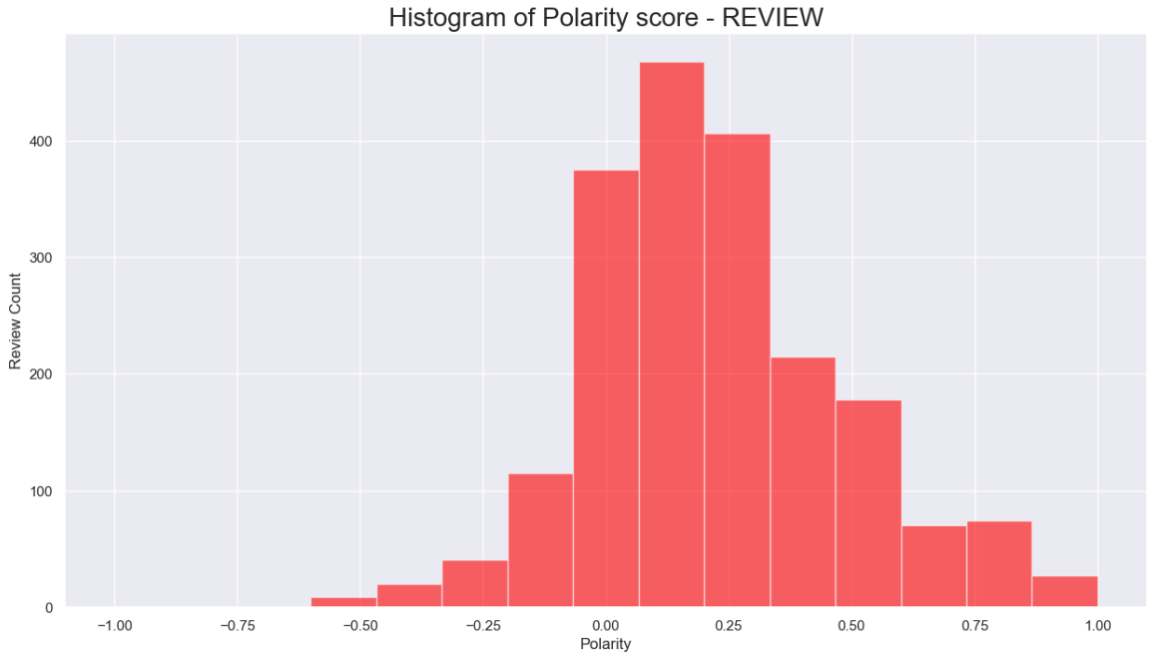


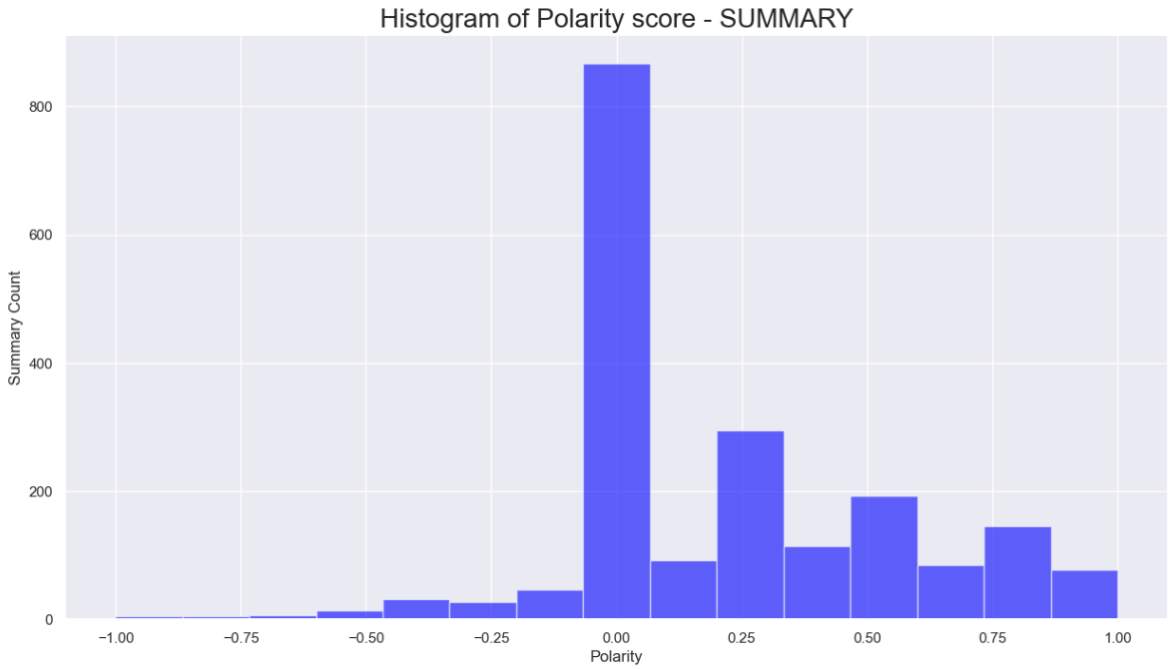
There are many useful libraries for conducting sentiment analysis, for our purposes we will consider two. TextBlob and VADER. TextBlob is a simpler rule-based system for sentiment analysis based on Naïve Bayes classifiers and can handle polarity and subjectivity. It tends to work well with clean, formal text. VADER is a rule-based model optimized for social media and short-form text. It can handle emojis, capitalization, punctuation, and negations, making it particularly strong for informal text. Using both TextBlob and VADER provides a more comprehensive sentiment analysis, as it combines TextBlob's strengths in handling general texts with VADER's expertise in detecting nuances in social media language, ensuring a broader and more accurate understanding of sentiments across different text sources.

TextBlob sentiment analysis

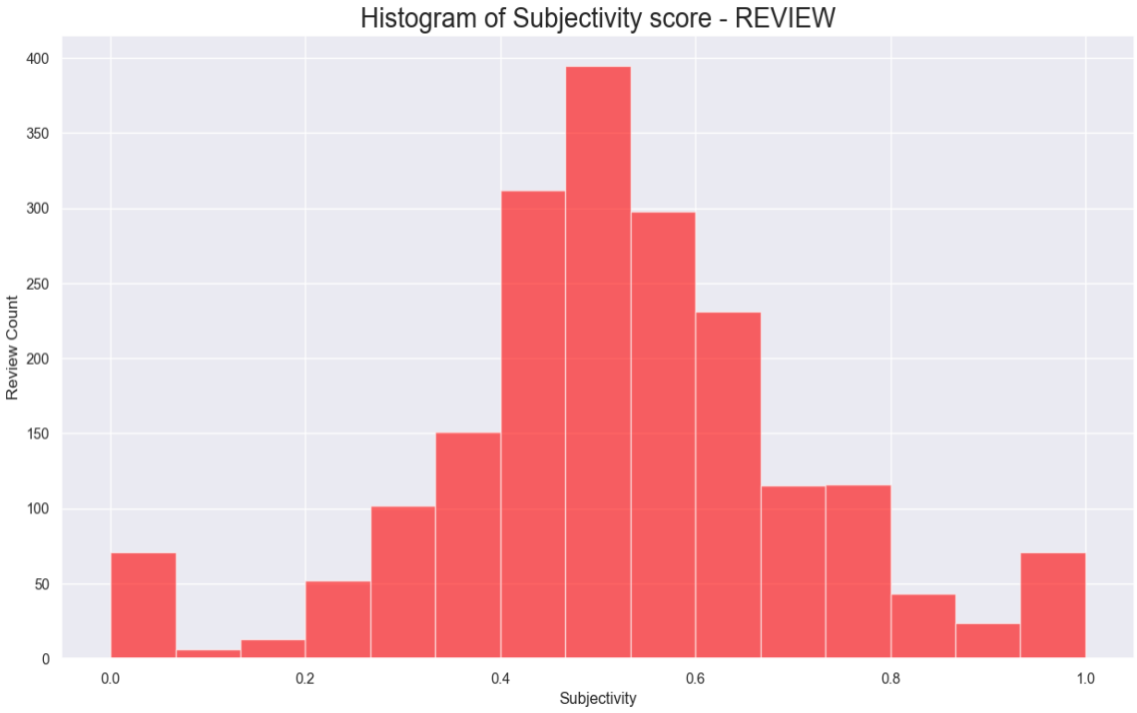
Plotting TextBlob histograms allows us to see a quick and easy visualization that again suggests that sentiment is positive for both review and summaries.

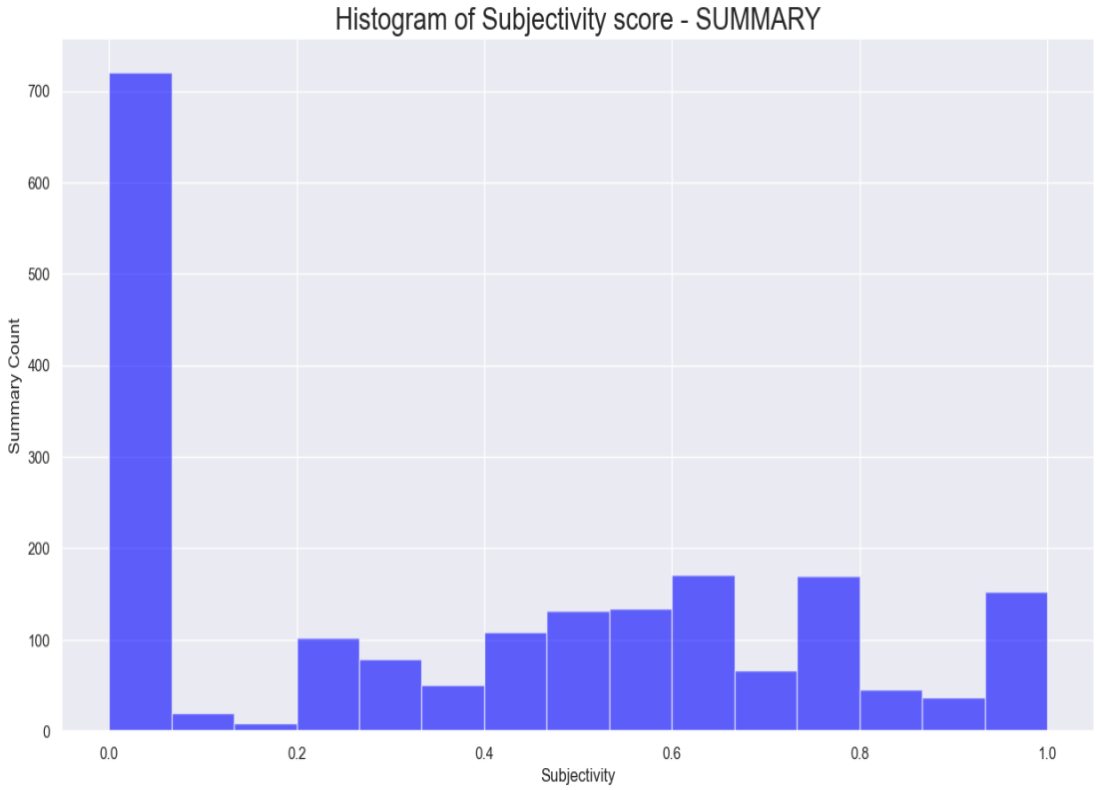
Once we had all the columns we needed, we could then consider the top/bottom 20 positive/negative reviews and summaries from the TextBlob model. This involved using the TextBlob polarity scores to find the smallest/largest 20 reviews/summaries (see Jupyter Notebook)



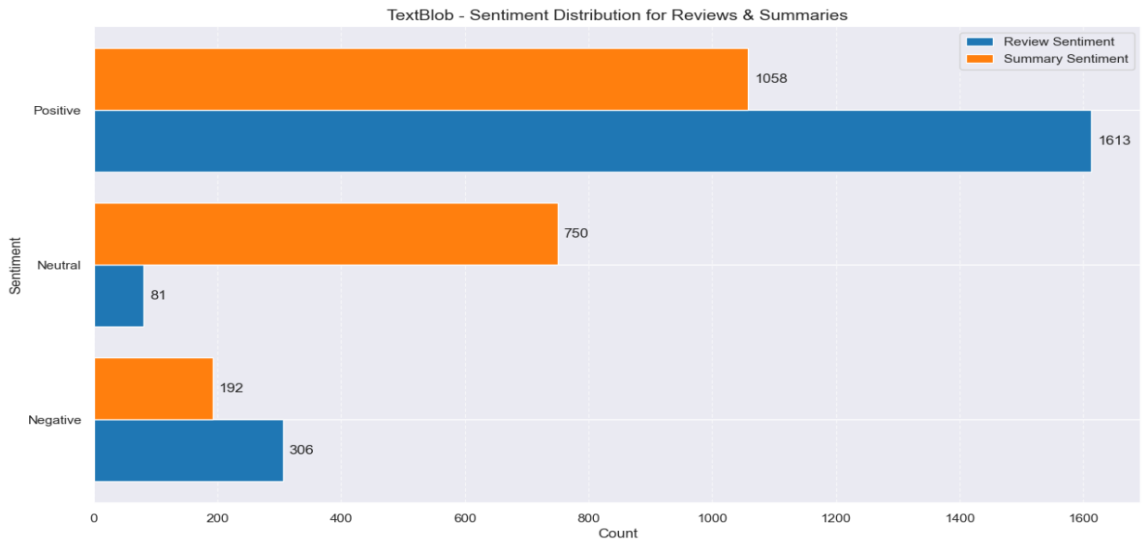


However, when we used TextBlob to determine if sentences were subjective (opinions) or objective (facts), we found something interesting. TextBlob often labeled negative summaries as objective. This likely happens because summaries are typically written in a concise, straightforward style, using shorter, more direct sentences. TextBlob, which works on basic language rules, may misinterpret this directness as a lack of opinion, even when the content is negative.





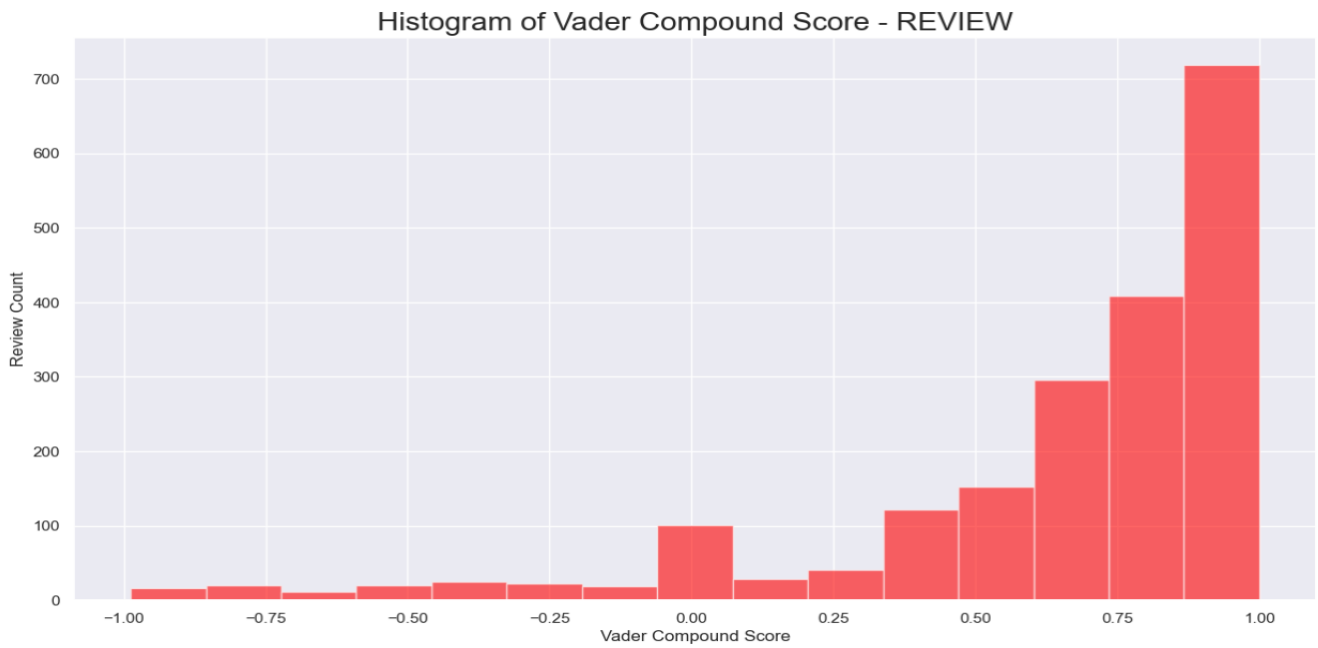
To understand the distribution of the sentiment for TextBlob reviews and summaries in a much simpler way, I have classified the polarity scores as either negative, neutral or positive based on their polarity scores.

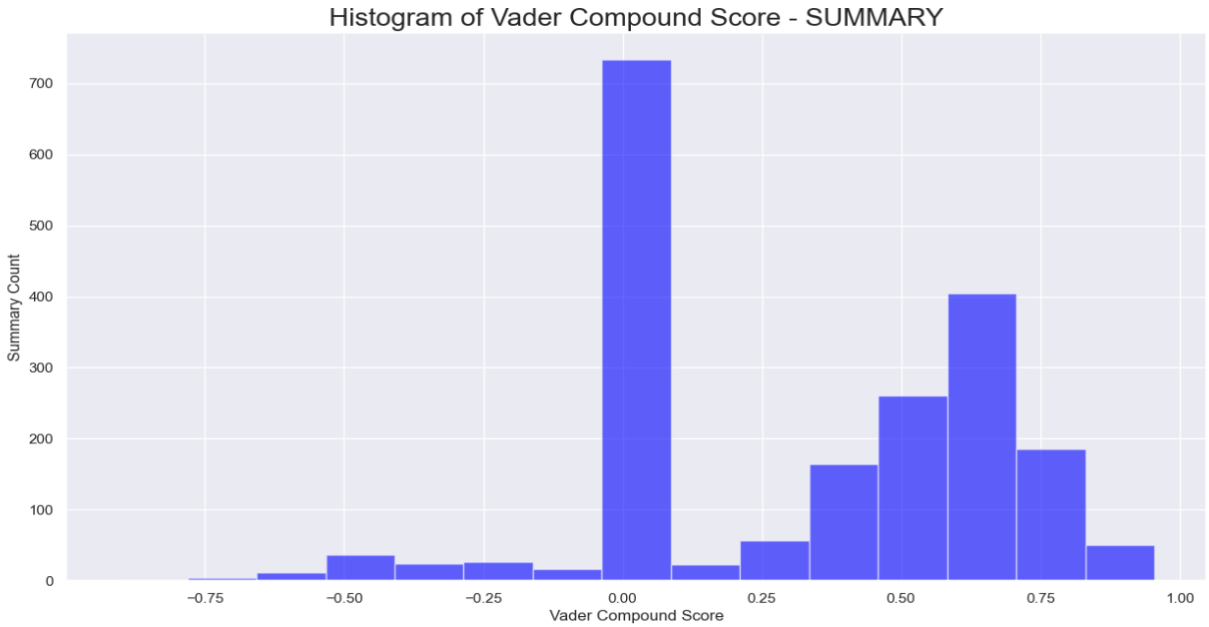


This confirms what we observed in our histograms, TextBlob scores majority of the reviews and summaries as positive, buts it's interesting to see the massive gap between reviews and summaries being classified as neutral. Given that the summary is just a shorter version of the review, you'd expect TextBlob to rank both reviews and summaries fairly equally. This suggests that shorter texts are being classed as neutral or certain words are being classified incorrectly.

Vader sentiment analysis

We then repeated the analysis using VADER. VADER is specifically designed to analyze sentiment in social media text. It excels at understanding slang, emoticons, and intensifiers (like "very" or "extremely"), which TextBlob might miss. Unlike TextBlob, VADER doesn't assess subjectivity. Instead, it provides a "compound score," which is an overall sentiment score. We used this score to create histograms, similar to those from TextBlob, for both reviews and summaries.



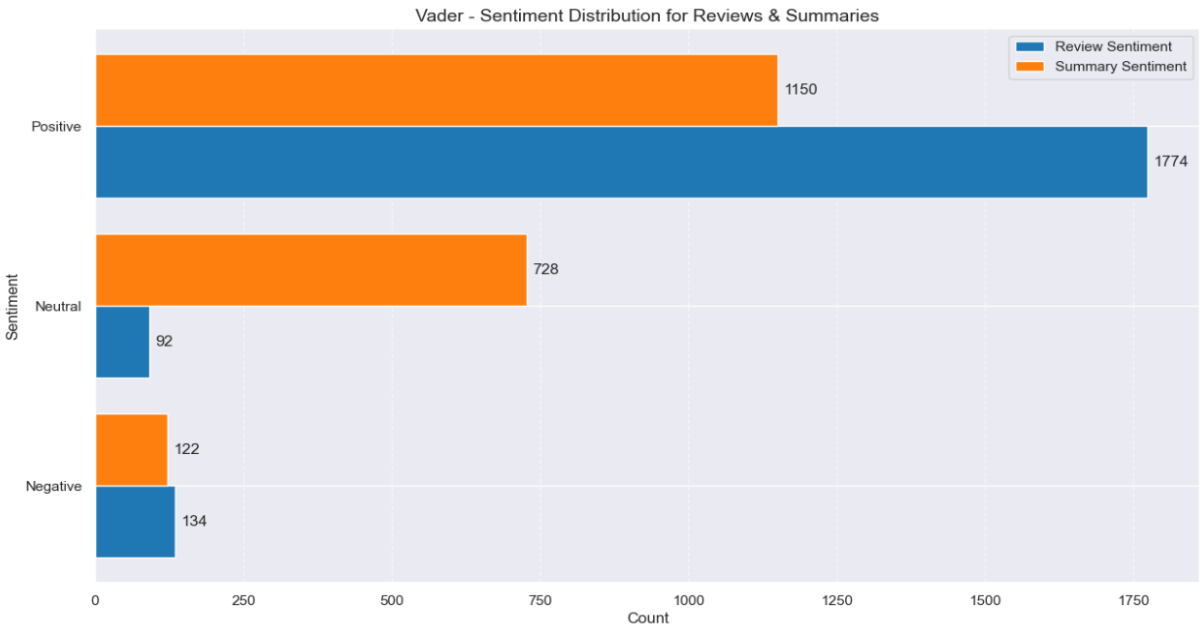


Vader scores the review as mainly positive, shown by the left skew in the histogram, but it scores the summary as mainly neutral shown by a high counts of compound scores equaling to 0. This could suggest the neutral way in which consumers tend to summarize their reviews. However, this

correlates with summaries as being seen as more objective by TextBlob, so the large spike in neutral summaries can be explained.

It would be interesting to know the summaries were obtained, were they just obtained through another model that takes the review and outputs a summary, or did the user manually have to submit a summary after submitting a review? Knowing how the summaries were obtained matters because specific words may be been included/omitted, and then whichever model we use to test the sentiment may have a different output.

We can also see the distributions for reviews and summaries together as shown by the below chart.



The Vader sentiment distribution shows reviews and summaries mainly received positive sentiment. Vader has classified reviews & summaries with similar negative counts, but again there is a much wider gap between reviews & summaries with neutral sentiment, just like we observed for TextBlob.

Insights

- **Usability Issues**: Customers find some games overly complex or poorly designed, hindering enjoyment.
- **Quality Concerns**: Product quality is inconsistent, failing to meet customer expectations.
- **Poor Value**: Customers perceive products as overpriced for their actual value.
- **Educational Success**: Products with educational value are positively received.
- **Experience Discrepancy**: Products fail to deliver the expected engaging experience, leading to disappointment.
- **Creative Products Preferred**: Customers value crafting-based products with high-quality tangible results.
- **Functionality Over Aesthetics**: Usability and playability are prioritized over visual design.
- **High playability Desired**: Games with varied and engaging content for repeated play are favored.

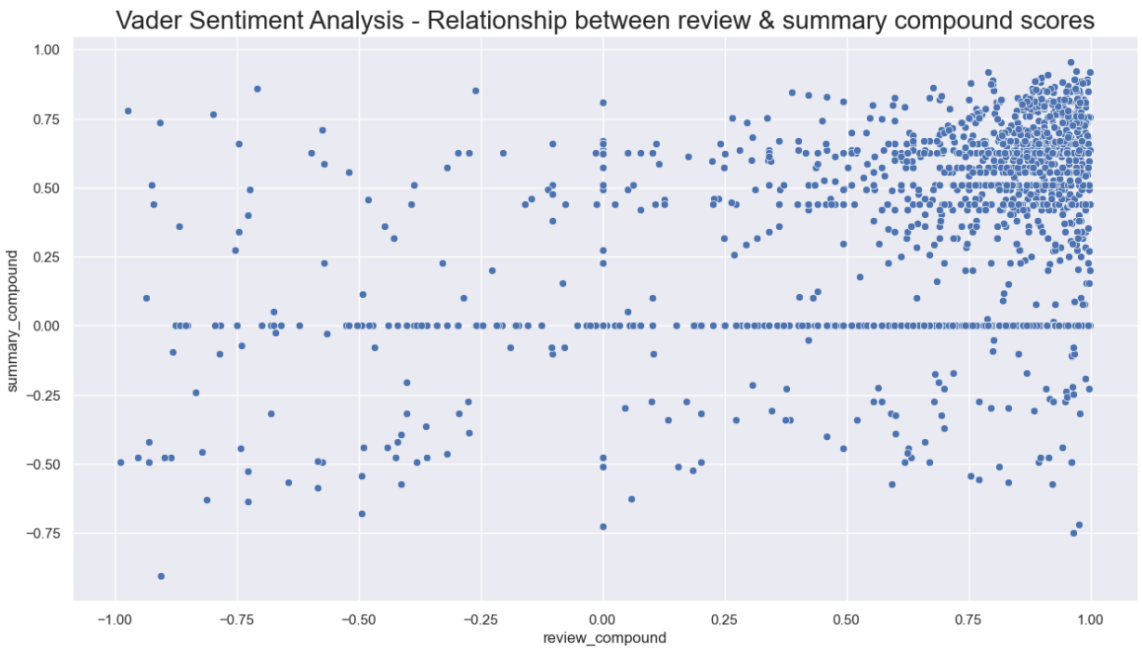
Recommendations

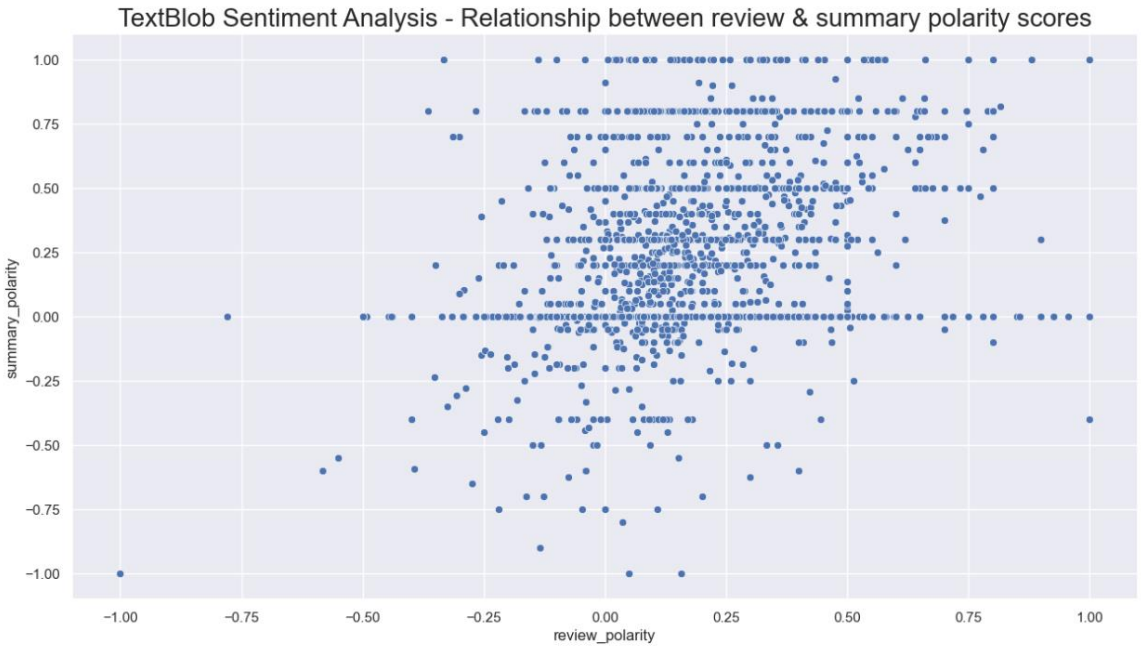
- **Optimize Product Design & Experience -** Simplify game mechanics and adjust complexity for target audiences to enhance usability. Prioritize functional design and interactive elements over purely aesthetic features. Ensure products deliver engaging experiences that match marketing promises.
- **Improve Product Quality & Value -** Implement rigorous quality control to meet customer expectations. Reassess pricing to align with perceived value and quality.
- **Expand & Innovate Product Offerings -** Capitalize on the success of educational products by expanding this range. Continue developing creative, crafting-based products with clear instructions and quality materials. Innovate for increase playability through dynamic content, expansions, or updates.

Evaluating TextBlob vs Vader

Sentiment analysis using TextBlob, allows us to plot the below scatter graph showing polarity scores for 'review' against 'summary'. We can see that there is a slight positive correlation between 'review' and 'summary', with a majority of the points concentrated around the positive polarity area. This in itself is a good thing for Turtle games. However, we do see that is classifies many reviews as positive, but their summary as neutral, as well as some reviews and positive but summary as negative.

Sentiment analysis using Vader, allows us to plot the below scatter graph showing compound scores for 'review' against 'summary'. We can see that there is no correlation between 'review' and 'summary', where we would expect them to be. As we can see, there is a high concentration of data points on the top right which indicates that there are many reviews and summaries being classified as a positive sentiment. There are also a large number of data points whose review is classified as positive, but their summary compound scores are 0, therefore they will be classified as neutral, indicating a miss match between these 2 variables using the Vader sentiment algorithm. Review Vader being so positive perhaps shows the difference in technique compared to TextBlob when it comes to language analysis. With VADER specializing in analysing sentiments in social media perhaps we are seeing it understand the subtleties of the review better than TextBlob.

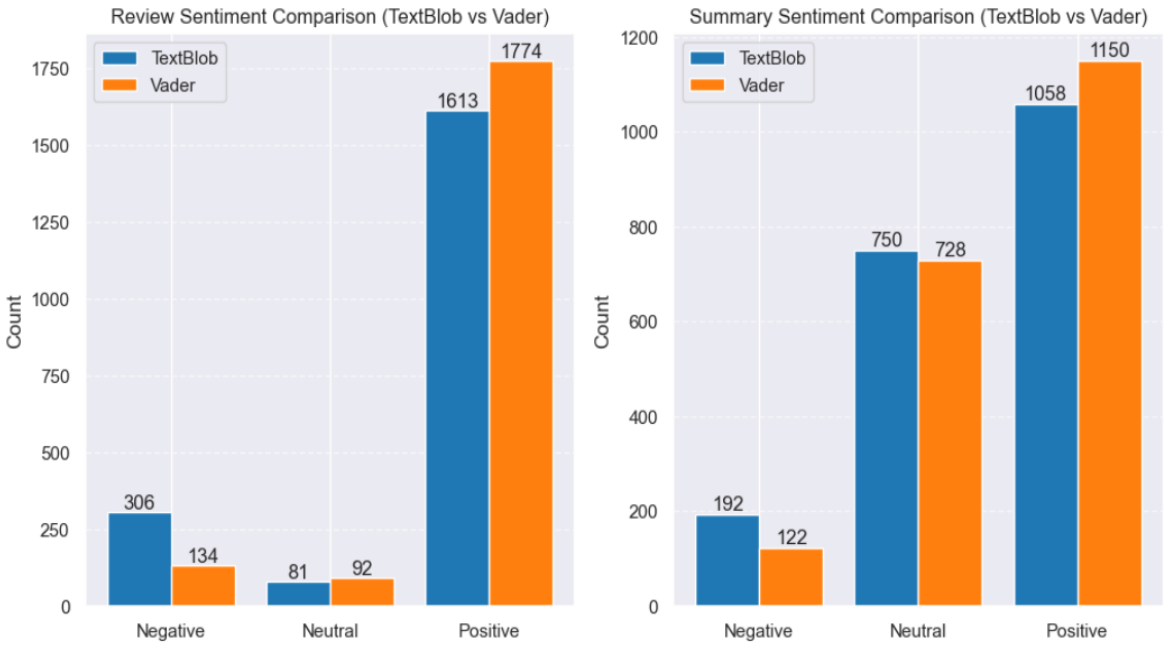




Review Sentiment Comparison Analysis - TextBlob and Vader scores roughly similar proportions for neutral and positive reviews, but there is a much wider gap for negative reviews, where TextBlob is much more dominant in scoring negative reviews, possibly due to lacking the complexity that Vader has.

Summary Sentiment Comparison Analysis - TextBlob and Vader scores roughly similar proportions for neutral and positive summaries, but there is a much wider gap for negative summaries. TextBlob is much more dominant in scoring negative summaries.

Both NLP algorithms work in different ways, with Vader accommodating five heuristics that would not be picked up by TextBlob, including punctuation and capitalization.



Main Insights from NLP analysis

- **Strong Positive Feedback**: Customer reviews and summaries predominantly express positive sentiment, with words like "fun," "love," "good," "great," and "play" appearing frequently.
- **Data Preparation is Key**: Accurate sentiment analysis requires thorough text data cleaning and normalization.
- **Comprehensive Analysis with NLP Tools**: Combining TextBlob and VADER provides a well-rounded sentiment analysis, addressing both general language and review-specific nuances.

Recommendations

- **Amplify Positive Language**: Incorporate frequently used positive words (e.g., "fun," "love," "great") into marketing to reinforce key product strengths.
- **Contextualize Positive/Neutral Feedback**: Analyze the context of positive and neutral comments to identify potential areas for product or marketing improvement.
- **Proactively Engage Negative Feedback**: Initiate direct conversations with customers who left negative reviews or summaries, as longer reviews tend to be more positive, these conversations can lead to constructive solutions.
- **Expand Multilingual NLP Capabilities**: Invest in or develop NLP tools that accurately analyze sentiment across multiple languages to better serve a global customer base.

Appendix 6 – Statistical Analysis (R)

Turtle Games 4th core business question was **How can descriptive statistics be used to enhance insights?**

'A business's success is determined by its relationship with its consumers. Statistical analysis helps business owners understand consumer buying patterns and their usage of products or services. This information can be used to determine which products and services should be offered to meet consumer demands.' [4]

Following on from appendix 2 where I loaded and wrangled the data, I calculated summary statistics, distributions, skewness, kurtosis. Through the help of the ggplot package I was able to create informative visualisations like scatterplots, histograms and boxplots to interpret the data more easily.

Loyalty points

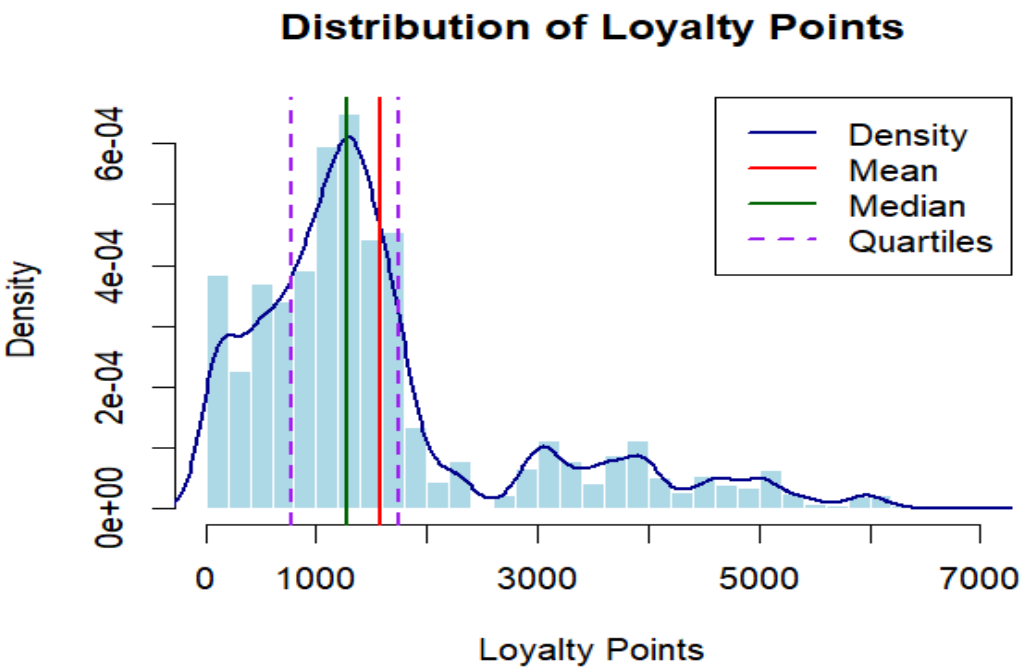
The loyalty points distribution is significantly right-skewed, indicating a small segment of customers with substantially higher balances. This is supported by:

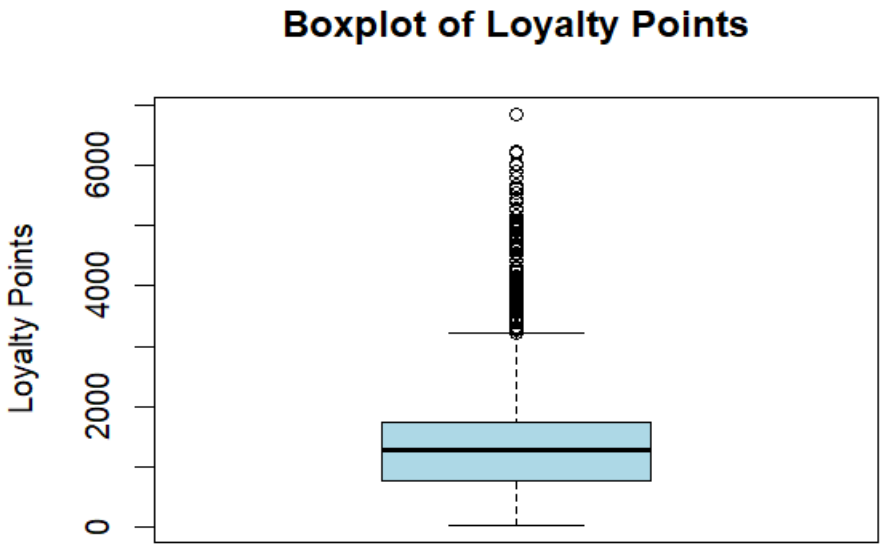
Mean (1578) > Median (1276): Demonstrating the right-skew.
Range (25 - 6847): Suggesting diverse customer segments (new, regular, highly loyal).
Multimodal Density Plot: Further indicating distinct customer groups.

Statistical tests confirm non-normality:
Shapiro-Wilk (W = 0.84307, p < 2.2e-16): Strongly rejects normality.
Skewness (1.46): Confirms positive skew.
Kurtosis (4.71): Indicates heavy tails, sharp peak, and outliers.

Summary Statistics:
Variance (1646704) & Standard Deviation (1283.24): High values indicate significant data spread, consistent with skewness and outliers.
Quartiles (Q1 = 772, Q3 = 1751) & IQR (979)

Overall, these summary statistics demonstrate the observed skewness and kurtosis.
They provide numerical evidence for the presence of significant variability and outliers within the loyalty points data.





Pay

Customer pay displays a near-symmetrical distribution with a slight positive skew, as indicated by:

Mean (48.08) ≈ Median (47.15): Suggesting near-symmetry.
Range (12.30 - 112.34): Showing moderate variability and potential high-income outliers.

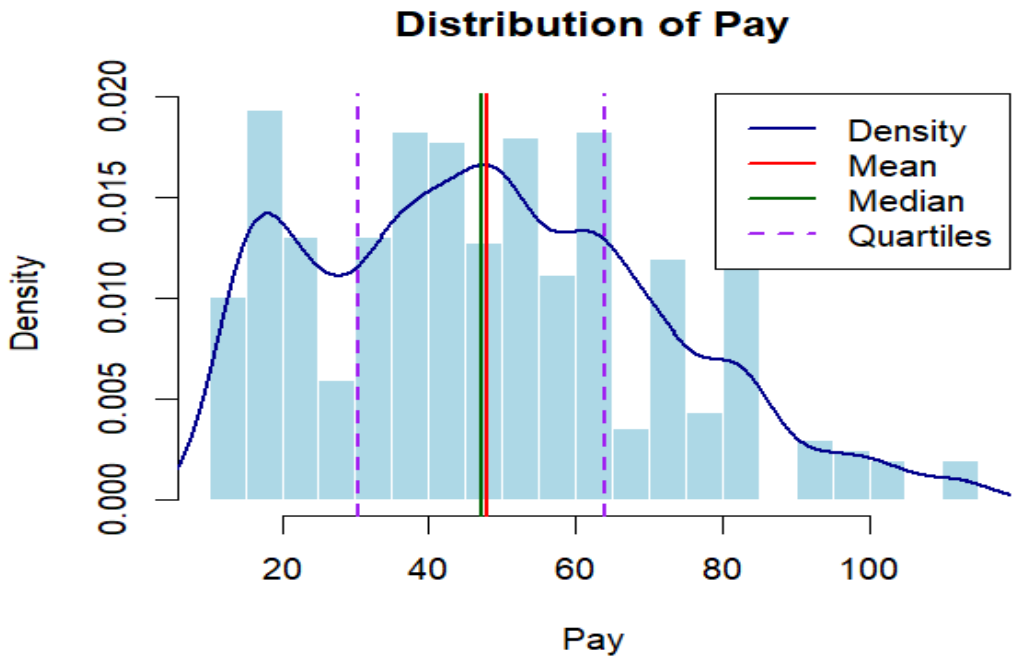
IQR (33.62): Demonstrating moderate spread in the middle 50% of incomes.
Concentration (30.34 - 63.96): Indicating where most incomes fall.

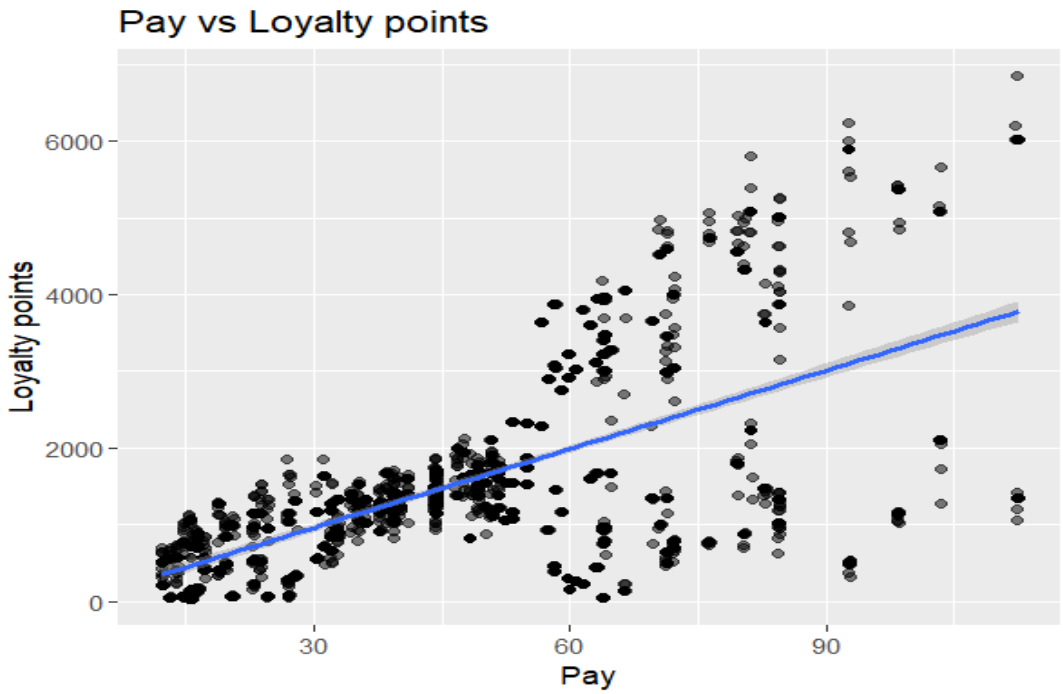
A strong positive linear relationship exists between income and loyalty points, with:
Clustering (15-50): Most customers at lower income levels.
Outliers (>50): High-income customers showing deviation.

Statistical tests reveal:
Shapiro-Wilk (W = 0.96768, p < 2.2e-16): Rejects normality despite W being close to 1.
Skewness (0.41): Confirms slight positive skew.
Kurtosis (2.59): Indicates a platykurtic distribution (thinner tails, flatter peak).

Summary Statistics:
Variance (534.72) & Standard Deviation (23.12): Moderate spread, aligning with the observed variability.
Quartiles (Q1 = 30.34, Q3 = 63.96) & IQR (33.62)
The range of the overall data (12.30 to 112.34) is much larger than the IQR, showing the potential presence of outliers.

Overall, these summary statistics support the near-symmetrical distribution and the presence of potential high-income outliers. The wide range of the data compared to the IQR, shows that the data does have a small number of values that are quite far from the mean. The low skewness value shows that the data is close to symmetrical.





Spend

Spending scores show near-perfect symmetry, indicated by:

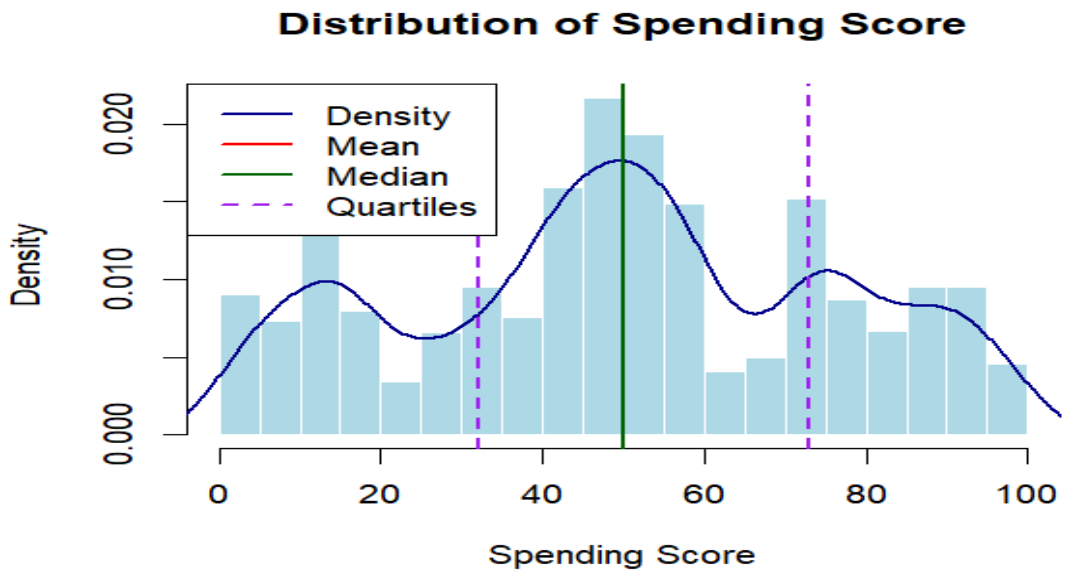
Mean (50) = Median (50): Suggesting a balanced distribution.
Range (1-99): Spending scores are between 1-100 as mentioned in the metadata
Similar Quartile Distances (min-Q1: 31, Q3-max: 26): Reinforcing symmetry.
IQR (41): Indicating significant spread in spending behavior.

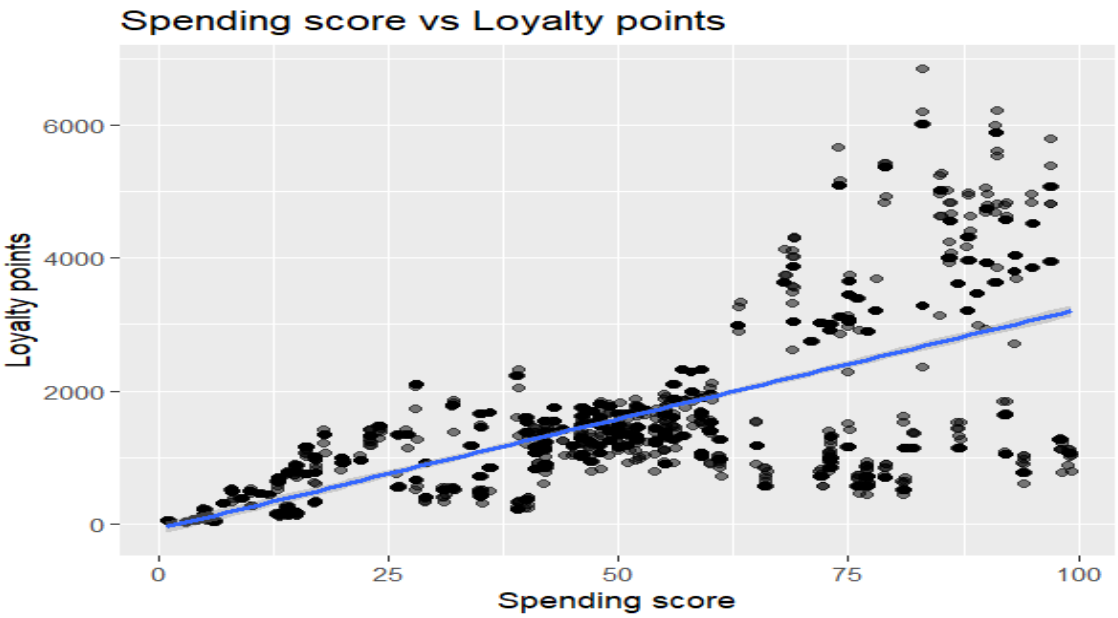
A strong positive linear relationship exists between spending score and loyalty points, with:
Clustering (around 50): Most customers in the middle spending score range.
Outliers (>60): High spending scores showing deviation and potential outliers.

Statistical tests reveal:
Shapiro-Wilk ($W = 0.96835$, $p < 2.2e-16$): Rejects normality despite W being close to 1.
Skewness (-0.04): Confirms near-symmetrical distribution.
Kurtosis (2.11): Indicates a platykurtic distribution (thinner tails, flatter peak).

Summary Statistics:
Variance (680.93) & Standard Deviation (26.09): Indicates significant spread, consistent with the large IQR.
Quartiles (Q1 = 32, Q3 = 73) & IQR (41)
The range of the overall data (1 to 99) is much larger than the IQR, showing the presence of possible outliers.

Overall statistics support the near-symmetrical distribution and the possible presence of outliers at higher spending scores. The large IQR and standard deviation show a wide range of spending habits.





Outlier Analysis

Using R with boxplots and the identify_outlier function, only the 'loyalty points' variable exhibited genuine outliers. 266 out of 2000 records (13.3%) were identified as outliers — a significant portion. A histogram of loyalty points shows a right-skewed distribution, indicating that most loyalty outliers are significantly on the higher end.

Analysing further also helps us understand loyalty point outliers by gender, with males accounting for 55% and females 45%. Analysing by education levels shows 48% of outliers are graduates.

Analysing average loyalty points in the outlier and non-outlier data shows that the loyalty points for outliers are nearly 4 times higher, suggesting this group may represent exceptionally a distinct group of high-value customers. These customers are not just a little above average; they are a completely separate group with much higher loyalty than everyone else. This tells us Turtle Games need to pay special attention to these customers. They are likely their most valuable customers.

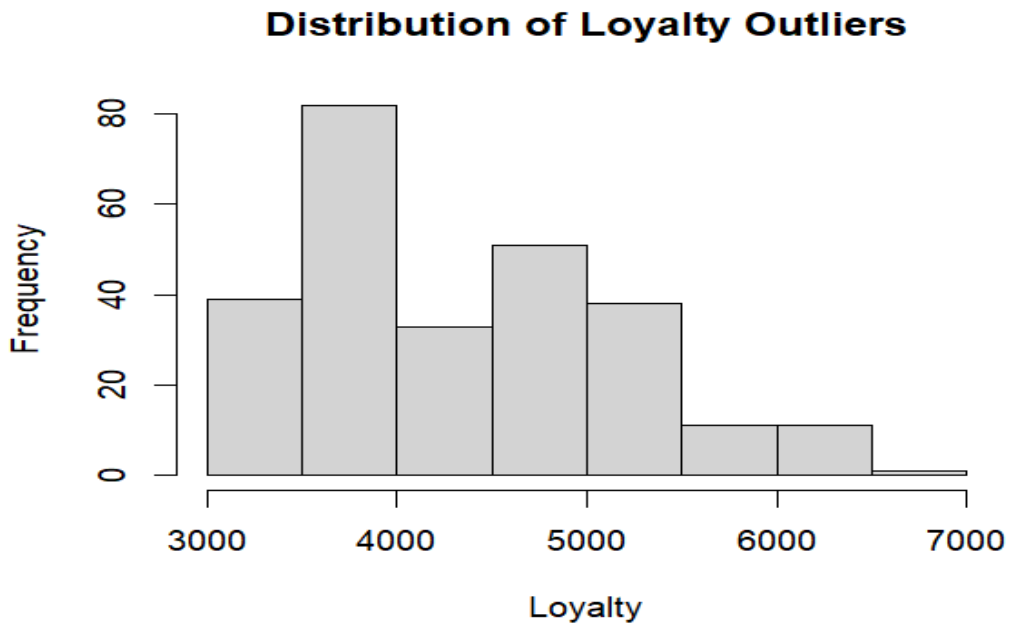
Average loyalty points (Outliers) = 4,328
Average loyalty points (non-outliers) = 1,156

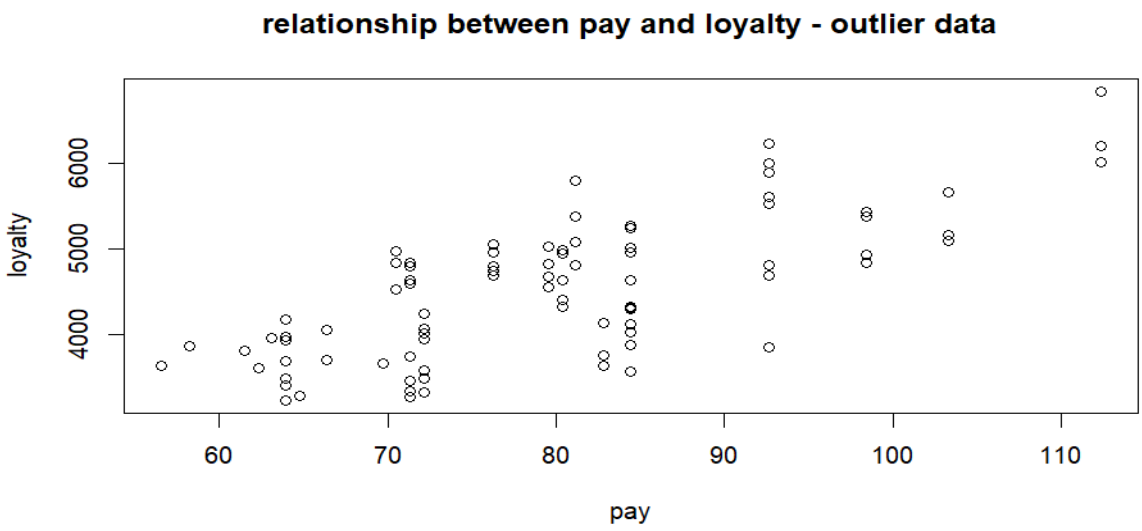
Correlation in outlier subset shows a strong positive correlation (79%) between Pay vs Loyalty. A weak correlation (24%) between Spend vs Loyalty and also a weak correlation (19%) between Age vs Loyalty.

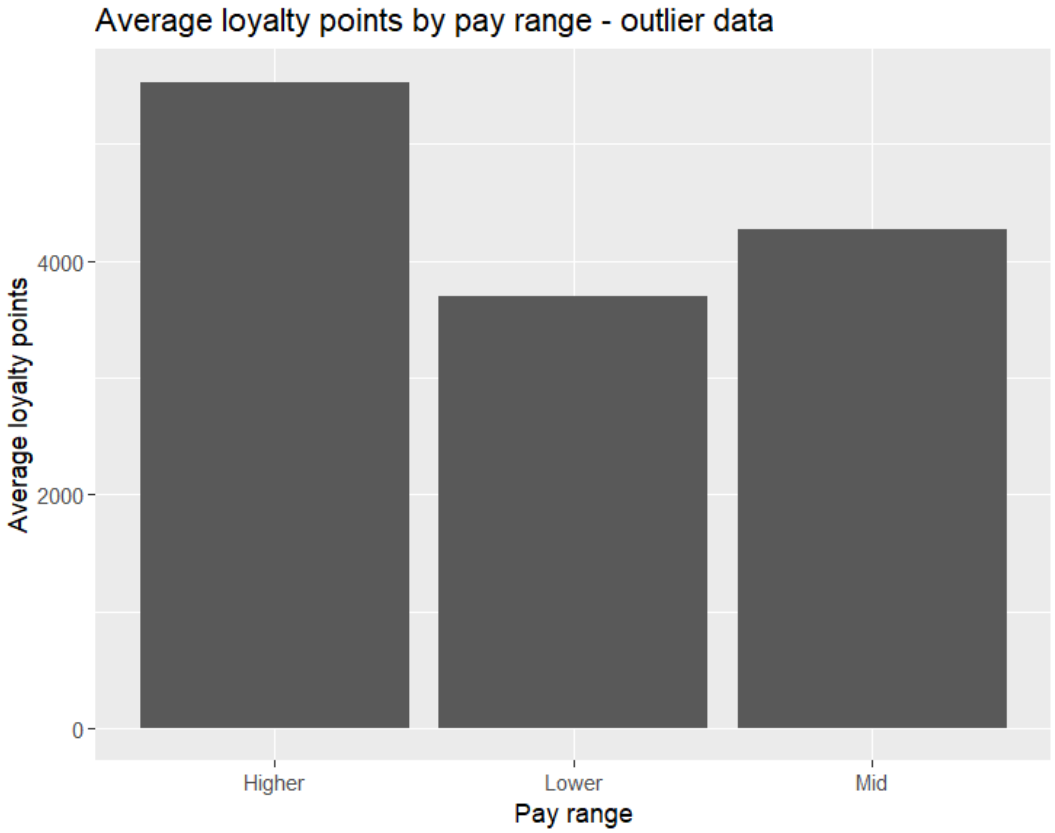
Customers in the higher pay range (top 25%) showed to accumulate the highest average loyalty points (5,535) suggesting Turtle Games should target these through VIP customer service and exclusive loyalty programs.

Recommendations

- Leverage the strong pay-loyalty correlation to design income-informed loyalty segments and targeted promotions or perks for higher-income customers
- Prioritize the 266 outliers as a strategic segment: Conduct deeper analysis to identify patterns in their purchase frequency, product preferences, and channel use.







Note: Higher represents top 25% pay customers, lower represents bottom 25% pay customers and mid represents middle 50% pay customers.

Appendix 7 – Multiple Linear Regression (R)

Following on from our statistical analysis and exploratory analysis, I was asked to create a multiple linear regression model in R.

'Regression allows us to gain insights into the structure of that relationship and provides measures of how well the data fit that relationship. Such insights can prove extremely valuable for analyzing historical trends and developing forecasts.' [5]

I rechecked my initial data frame and its properties i.e. missing value checks, duplicates, data types and computed summary statistics. Multiple linear regression in R was performed using the lm() function.

As previously confirmed, pay and spend were the variables with the highest correlation with loyalty and they also showed a significant linear relationship, more so than other variables. As such pay and spend became our dependent variable and loyalty was our dependent variable in this MLR model.

Based on our statistical analysis, there are some concerns we must make ourself aware of:

Potential concerns

- **Non-Normality of Loyalty Points -** Loyalty points are significantly right-skewed and non-normally distributed (Shapiro-Wilk $p < 2.2e-16$). MLR assumes normality of residuals, which is likely violated. This can affect the reliability of our model's coefficients and p-values.
- **Outliers in Loyalty variable -** These outliers can exert unwarranted influence on the regression line and affect the model's overall fit.
- **Heteroscedasticity -** The right-skewed distribution of loyalty points suggests potential heteroscedasticity (unequal variance of residuals). This violates a key assumption of MLR and can lead to unreliable standard errors and hypothesis tests.
- **Potential for Non-Linear Relationships -** While there seems to be a linear relationship between pay/spend and loyalty points, the presence of outliers and non-normality suggests that a linear model might not capture the full complexity of the relationships.
- **Multimodal Distribution of Loyalty Points -** The multimodal distribution of loyalty points may suggest that there are underlying groupings within the data, that a MLR model may not be able to account for.

Corrective Actions

- **Transform Loyalty Points -** Apply a logarithmic transformation (natural log) to the loyalty points variable to reduce skewness and potentially address heteroscedasticity.
- **Investigate Multimodal Distribution -** We observed that our loyalty points variable has a multimodal distribution. We could investigate if this is because of groupings within the data and consider adding grouping variables to the model, or splitting the data into groups, and running separate MLR models.
- **Outlier management -** Consider trimming the extreme values (e.g., cap at 95th percentile).
- **Model Validation -** Split the data into training and test sets, to properly validate the model and prevent overfitting.

- **Pay and spend to predict loyalty**

As discovered from our correlation analysis, pay has a strong correlation with loyalty (0.62), and spend also has a strong correlation with loyalty (0.67). Plotting these variables in a scatterplot showed a significant linear relationship and therefore we conducted a MLR model in R using the lm() function.

R-Squared = 83%. This is a very high R-squared value, indicating that our model explains 83% of the variance in loyalty points. P-value: < 2.2e-16 (statistically significant result):
This extremely low p-value indicates that the overall model is highly statistically significant.

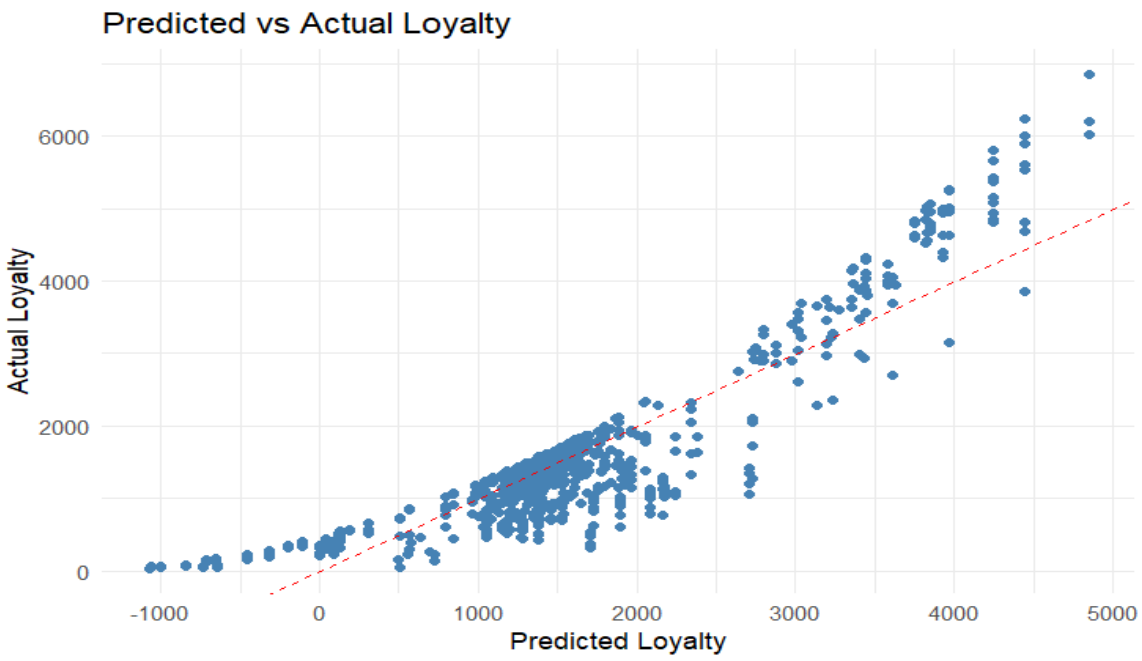
RMSE for this model is 533.7. In the context of the data, this means that on average, the model's prediction of loyalty points will be off by about 534 points. The average of loyalty points is 1578, so being off by 534 points is quite significant.

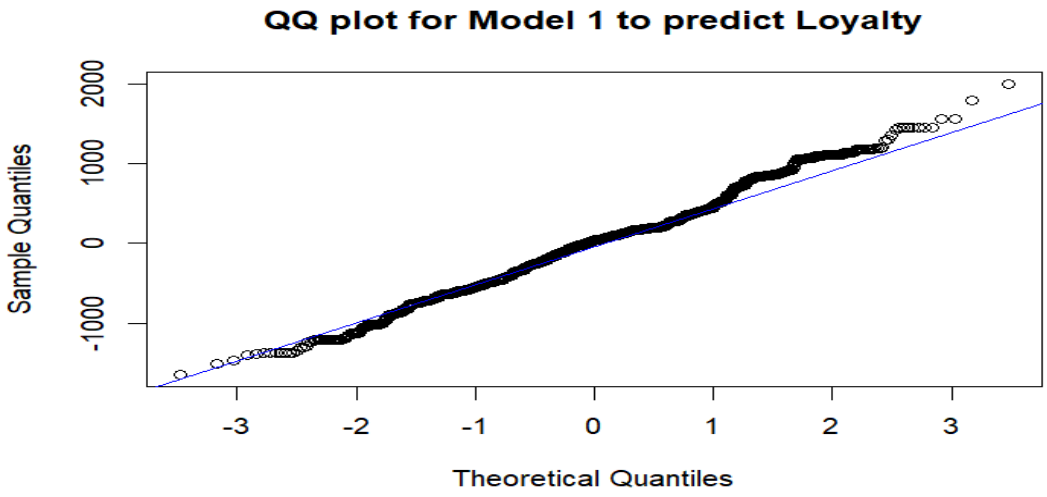
The wide range of residuals (-1646 to 1999) suggests that some predictions have substantial errors. The fact that the median is 40, shows that the residuals are relatively centered around 0. The wide range of the residuals, shows that even though the model has a very good R squared value, that there are still some large errors within the predictions of the model. This is something that would need to be investigated further.

Checking for multicollinearity using the VIF showed values of 1 for both pay and spend confirming no correlation between these 2 variables.

Checking for heteroscedasticity using the BP test resulted in a LM test P value less than our significance level of 0.05, therefore we reject the null hypothesis and assume there is evidence of heteroscedasticity.

Extension – I have also provided the option for a stakeholder to input their data for spend and pay to use my model to predict loyalty.





- **Pay and spend to predict LOG loyalty**

As demonstrated in model 1 there was evidence of heteroscedasticity, we can transform our dependent variable by taking the natural log.

R-Squared = 80% (GOOD, but less than model 1). This is still a good R-squared value, meaning our model explains 80% of the variance in the natural logarithm of loyalty points.
However, it's slightly lower than our previous model (83%), indicating a slightly weaker fit after the log transformation. This shows that the first model, fit the data slightly better.

P-value: < 2.2e-16 (statistically significant result). This extremely low p-value signifies that the model is highly statistically significant. The relationship between pay, spend, and the logarithm of loyalty points is very unlikely to be due to chance.

Again, I had to transform the log transformed data back the original scale by exponentiating it to compare the models' predictions in the original loyalty point scale.

RMSE for this model is 866. 1. In the context of the data, this means that on average, the model's prediction of loyalty points will be off by about 866 points. This is worse than our previous model. The average of loyalty points is 1578, so being off by 866 points is quite significant.

Residuals vary from -6175 (min) to 783 (max) with a median of 71.65. The range of residuals is much larger compared to the previous model. The median is also much further away from zero. This suggests that the log transformation has not helped reduce heteroscedasticity, and that there are outliers in the data which are having a negative impact on the model's predictive power.

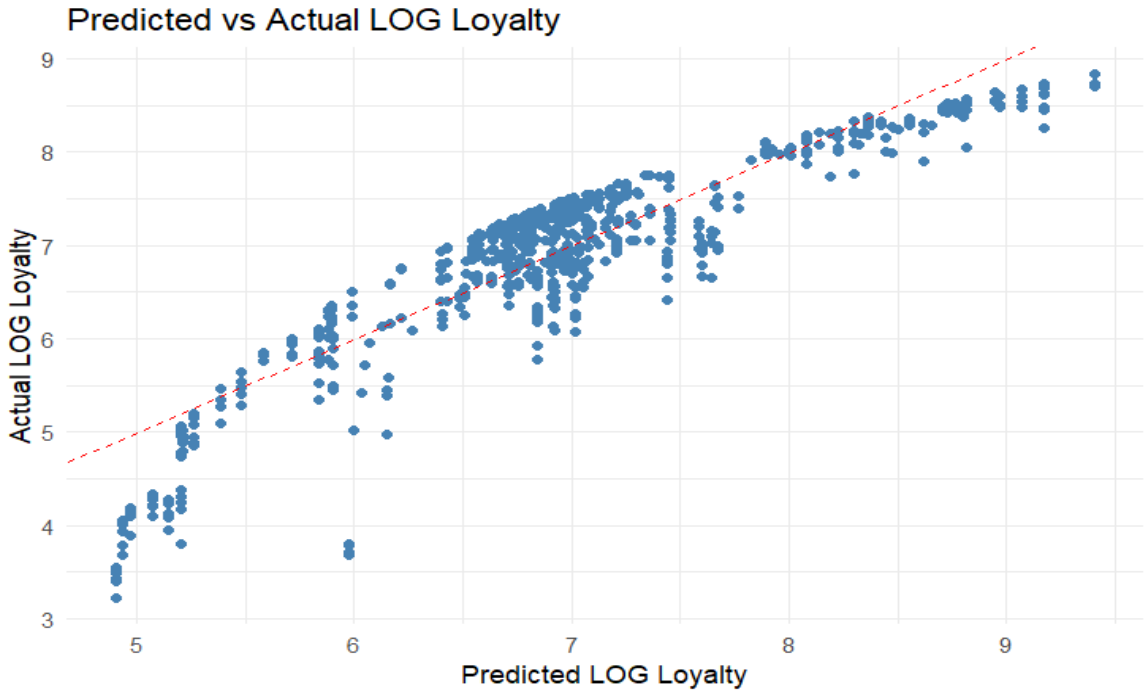
Checking for multicollinearity using the VIF showed values of 1 for both pay and spend confirming no correlation between these 2 variables.

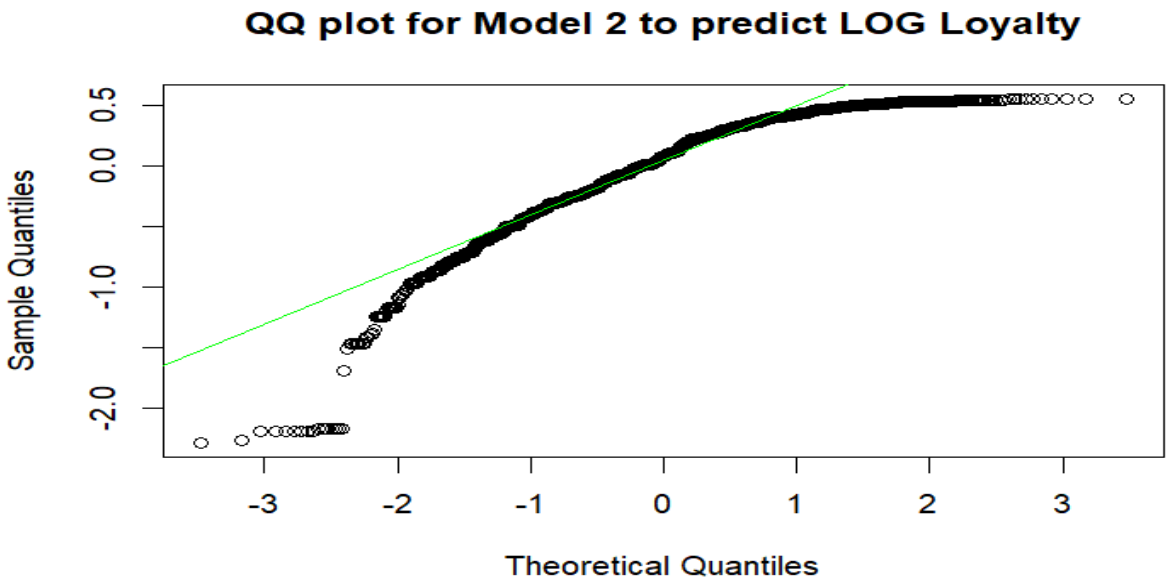
Checking for heteroscedasticity using the BP test resulted in a LM test P value less than our significance level of 0.05, therefore we reject the null hypothesis and assume there is evidence of heteroscedasticity.

To summarize, model 1 performed the best due to the higher R-Squared value of 83%, compared to model 2 R-Squared value of 80%. Model 1 also had a slightly lower RMSE compared to model 2, showing its predictive power is better than model 2. However, both models' validity was undermined

by presence of heteroscedasticity., due to the presence of outliers which means the model could not capture the linear relationship as well.

It's also important to note that both of these models were not properly validated by using the standard 'train and test' split, which is usually a better way to test a model and prevent it from overfitting. If we did apply a train and test split, we may have observed a better R-Squared value and a lower RMSE to improve the models fit and predictive power.





Appendix 10 – Decision trees

An analysis using Decision Trees was also done in Python using the Decision Tree Regressor library, but I decided to focus only on multiple linear regression. Nevertheless, the below are some insights found from my decision tree analysis. The full decision tree process can be seen in the Jupyter Notebook.

'The importance of Decision Trees stems from their ability to simplify the decision-making process by visualizing the options and potential outcomes, thus helping businesses to evaluate the probable impacts of various choices before committing to a course of action.' [6]

As with any analysis we need to perform some data cleaning steps:

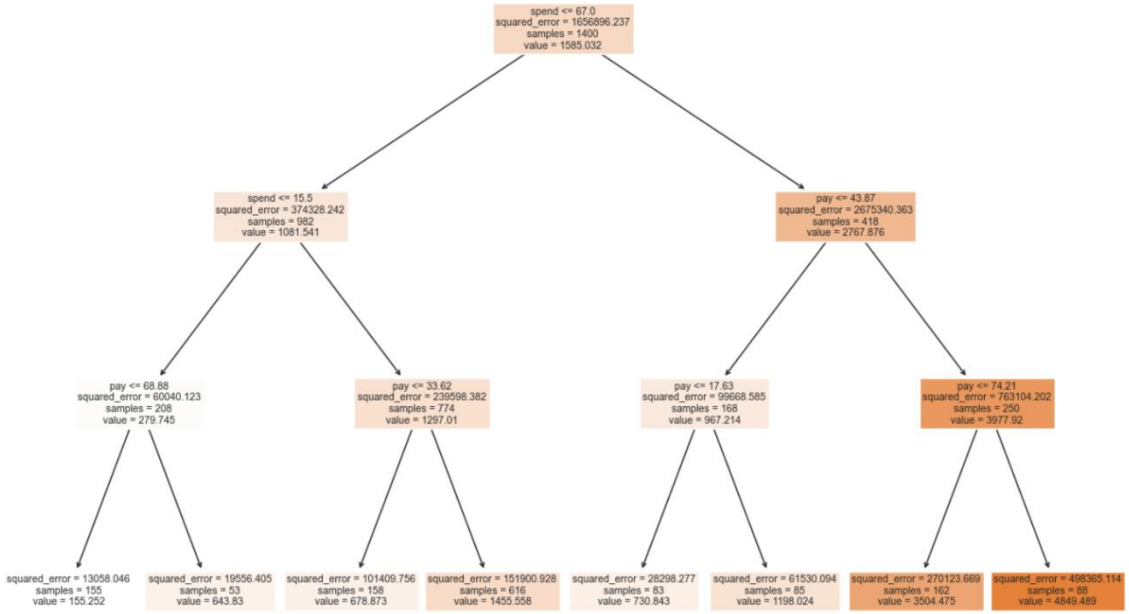
- Applied One Hot Encoding for 'gender' (gender has no inherent order)
- Applied Ordinal Encoder for 'education' (education has a hierarchical order)

I first decided to use all the numerical variables in my model with no limit on max_depth to see how the model performed. This could then be our baseline for analysis and evaluation.
Created training (80%) and (20%) testing splits to prevent model overfitting and using a random state=42 for reproducibility. Evaluated training and testing results using R-Squared and RMSE

Conducted feature importance check which then led me to pruning the decision tree for more accuracy. I realized that only pay (50%) and spend (48%) were the most important features in the model which led me to pruning the decision tree, and creating a new decision tree algorithm with only these 2 variables. I then tested different levels of max_depth for 3,4,5.

max_depth	train_r2	test_r2	train_rmse	test_rmse
none	1.00	0.99	0.00	98.36
3	0.91	0.91	385.31	371.43
4	0.95	0.94	298.87	302.93
5	0.96	0.96	255.49	262.24

Max Depth = 5 is the best model as it achieves the best trade-off between accuracy and generalization while keeping test RMSE low. Max Depth = 5 is the best choice. Test R^2 is high (0.96), close to train R^2 (0.96). Test RMSE (262.24) is the lowest among pruned models. Not overfitting (train vs. test R^2 is similar). Max Depth = None (fully grown tree) is poor. Train R^2 = 1.00 (perfect fit) which means its most likely overfitting. Test RMSE (98.36) is much lower than others, but train RMSE = 0.00, which is unrealistic. Max Depth = 3 underfits slightly (lower R^2 and higher RMSE). Max Depth = 4 is also a good option, but Depth 5 performs slightly better. Based on the above we can conclude that our pruned decision tree with max_depth=5 is the best model.



The above figure shows the finalised decision tree with max_depth=5. However, I have only shown max_depth=3 as the output becomes too cluttered to interpret. Please refer to the Jupyter Notebook to see the full decision tree and analysis.

Insights & recommendations from our final model

A decision tree regression model was developed to predict customer loyalty points for Turtle Games using spend and pay as predictor variables. The model, constrained to a maximum depth of five, revealed that spend is the most influential predictor of loyalty, as it appeared at the root of the tree. Specifically, customers with a spend of 67 or less were separated from those with higher spend, indicating that lower-spending customers tend to exhibit different loyalty behaviors than high spenders. This segmentation underscores the importance of spending habits in loyalty dynamics.

Within the group of lower spenders (spend ≤ 67), a further split at spend ≤ 15.5 identified a particularly disengaged group. These very low spenders were then divided based on pay. Those with higher pay (pay ≤ 68.88) but low spending levels may represent a missed opportunity—customers who can afford to spend but are not currently engaged with Turtle Games' offerings. Conversely, the segment with both low pay and low spend (pay ≤ 33.62) likely reflects more budget-conscious consumers who require value-driven propositions to drive loyalty.

On the other side of the tree, among higher spenders (spend > 67), pay again played a differentiating role. A split at pay ≤ 43.87 highlighted differences between lower-pay and higher-pay high spenders. The segment with low pay but high spending (pay ≤ 17.63) may be highly engaged but financially limited—these customers show strong loyalty but could be sensitive to pricing. Those with both high pay and high spending (pay ≤ 74.21) represent the most valuable customer group, combining purchasing power with strong engagement.

From a business perspective, the model provides several key insights. Firstly, spending behavior is the most reliable indicator of loyalty, supporting its use in segmentation and targeted marketing strategies. Secondly, pay plays a supporting role, helping to refine customer understanding within spending categories. Notably, the presence of high-pay but low-spending customers presents a significant growth opportunity; these individuals could be re-engaged through premium marketing

campaigns or personalized outreach. Meanwhile, high-spending, low-pay customers, while loyal, may require more affordable loyalty incentives or flexible pricing options.

Based on these insights, several strategic recommendations are proposed for Turtle Games. Implementing a tiered loyalty program can help cater to different customer groups—for instance, offering exclusive rewards to high-pay, high-spending customers, while designing value-based incentives for price-sensitive but loyal segments. Additionally, targeted campaigns should be deployed to convert high-pay, low-spending customers by showcasing the unique value of Turtle Games products and emphasizing aspirational or exclusive experiences. Finally, careful attention should be given to pricing and promotional structures to ensure that loyal but budget-constrained customers are not overlooked.

Although we have used the decision tree algorithm to help us predict loyalty points, they aren't necessarily the best algorithm to use as they are still prone to overfitting. Using ensemble methods like random forests can help reduce overfitting and increase testing accuracy. Also, there is difficulty in capturing linear relationships thereby affecting the use of decision tree models in this analysis.

Appendix 11 – Extra comments

Challenges/limitations/analytical recommendations

- Product data not at a granular level – Product data was only given at a unique product code level and not what product category this related to. As such we cannot infer the performance of product categories. Turtle games should provide product data at a more granular level.
- Mismatch between reviews & summaries – when I was reviewing certain reviews and their corresponding summaries, I noticed that that there was a discrepancy. A product mentioned in a customer review, mentioned a different product in the same customer summary. Clearly there is a data issue here that needs to be addressed.
- Focusing on only MLR – Although I did do a decision tree analysis, I only focused on MLR in my presentation. With the results of my decision tree being quite lengthy and the actual tree having max_depth=5, I find that this would be difficult to explain to a non-technical stakeholder for the presentation. Also, there is a high chance of overfitting in decision tree models which also needs to be noted. Options such as random forests also help reduce overfitting but they are harder to interpret.
- Not having any sales data – Having sales data would have helped understand the relationship between pay, spend, loyalty points better and may have generated a better model to help understand what kind of customers are loyal and if theyre contributing to sales too.
- As Turtle Games have a global customer base, it would have been useful to have some regional data too to provide regional specific insights.
- Investigate outliers and right skewed/multimodal distribution of loyalty points for better predictive accuracy for linear regression models.
- Understand how the 'summary' variable was generated because this affects model interpretation and performance. Why not just use 1 column (reviews) as the single source of truth?

---

6. References

1. *Piggy. (n.d.). Economic value of customer loyalty. https://www.piggy.eu/en/glossary/economic-value-of-customer-loyalty*

2. *Qlik. (2024). KMeansCentroid example. https://help.qlik.com/en-US/sense/November2024/Subsystems/Hub/Content/Sense_Hub/ChartFunctions/RelationalFunctions/KmeansCentroidExample.htm*

3. *Insight7. (n.d.). How to analyze insights from customer reviews. https://insight7.io/how-to-analyze-insights-from-customer-reviews*

4. *Bvarta. (n.d.). 5 reasons why statistical analysis is important for business growth. https://bvarta.com/5-reasons-why-statistical-analysis-is-important-for-business-growth*

5. *Harvard Business School. (n.d.). What is regression analysis? https://online.hbs.edu/blog/post/what-is-regression-analysis*

6. *International Institute of Executive Careers. (n.d.). Decision tree analysis: Strategic decision making. https://www.iienstitu.com/en/blog/decision-tree-analysis-strategic-decision-making*