TURTLE GAMES

CUSTOMER SEGMENTATION & PREDICTIVE MODELING

ALEXANDRA AKAOUI

21.07.2025

Table of Contents

Introduction and Context

Project Briefing

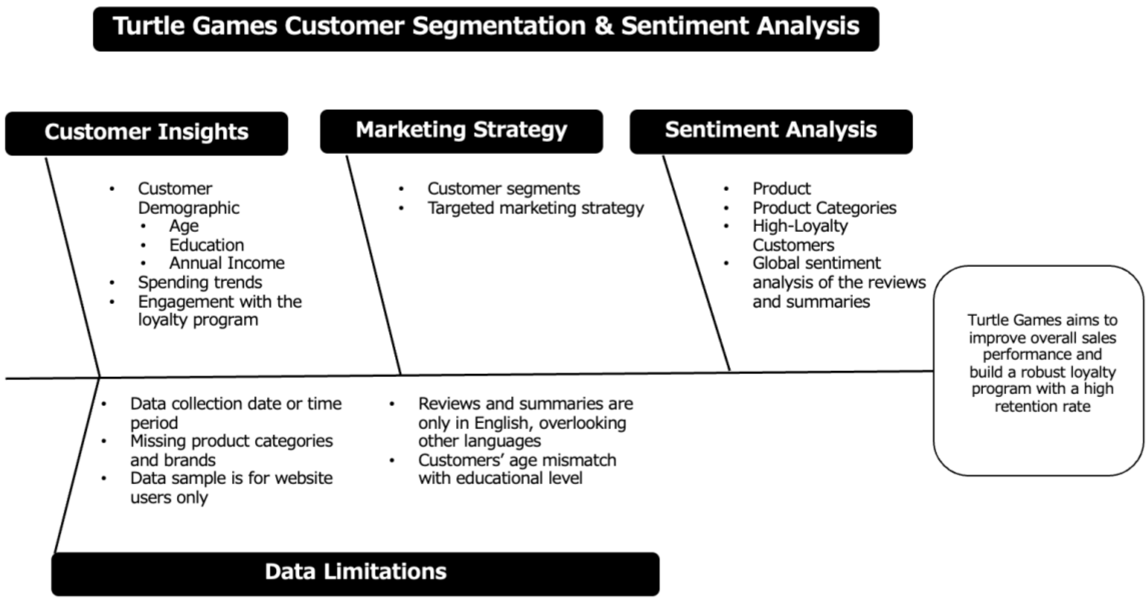
Turtle Games, a game seller and manufacturer, aims to increase its overall sales by leveraging insights drawn from the analysis of its customer base and the loyalty program scheme.

Business Problem

Turtle Games commissioned the analysis, aiming to cover the following areas:

- Exploratory analysis of the customer demographic data
- Analysis of spending and annual remuneration trends and their relation to loyalty points accumulation.
- Customer segmentation to identify high and low return customers
- Predictive Modeling including:
    - A classification model to predict customer segment membership
    - A regression model to estimate the accumulated loyalty points
- Sentiment Analysis of customer reviews and summaries, focused on:
    - Products
    - Product categories
    - Global sentiment patterns across all reviews and summaries

Analytical Framework



Limitations

- Timeframe of the data collection is not identified
- Data sample is for website users only, overlooking app users.
- Only English reviews are analyzed, overlooking reviews in other languages

Analytical Exploratory Analysis

Data Cleaning Process

- Customers were grouped both by demographic and behavioral variables – resulting in 782 unique customers.

	gender	age	remuneration_k	spending_score	loyalty_points	education	count
0	Female	17	13.94	40	233	Postgraduate	1
1	Female	17	35.26	54	797	Postgraduate	1
2	Female	17	53.30	48	1071	Postgraduate	7
3	Female	17	69.70	26	759	Postgraduate	1
4	Female	18	12.30	81	436	Graduate	1
...	...	...	...	...	...	...	...
777	Male	72	37.72	56	1264	Postgraduate	7
778	Male	72	39.36	47	1107	Phd	1
779	Male	72	40.18	55	1322	Phd	7
780	Male	72	66.42	93	3695	Phd	1
781	Male	72	112.34	18	1210	Phd	1

- Inconsistencies in the age and education of some customers were identified

```
Number of PhD/postgraduate entries with age under 22: 20
Customer IDs with age under 22 in PhD/Postgraduate:
```

	gender	age	remuneration (k£)	spending_score (1-100)	loyalty_points	education	language	platform	count	compare_key
0	female	17	13.94	40	233	postgraduate	en	web	1	female-17-13.94-40-postgraduate-en-web
1	female	17	35.26	54	797	postgraduate	en	web	1	female-17-35.26-54-postgraduate-en-web
2	female	17	53.30	48	1071	postgraduate	en	web	7	female-17-53.3-48-postgraduate-en-web
3	female	17	69.70	26	759	postgraduate	en	web	1	female-17-69.7-26-postgraduate-en-web
4	male	17	18.86	98	774	phd	en	web	1	male-17-18.86-98-phd-en-web
5	male	17	27.06	4	45	phd	en	web	1	male-17-27.06-4-phd-en-web
6	male	17	27.06	92	1042	phd	en	web	7	male-17-27.06-92-phd-en-web
7	male	17	44.28	52	964	phd	en	web	1	male-17-44.28-52-phd-en-web
8	male	17	48.38	41	830	phd	en	web	7	male-17-48.38-41-phd-en-web
9	male	17	48.38	55	1114	phd	en	web	1	male-17-48.38-55-phd-en-web
10	male	17	50.02	42	880	phd	en	web	1	male-17-50.02-42-phd-en-web
11	male	17	81.18	39	1326	phd	en	web	1	male-17-81.18-39-phd-en-web
12	male	18	24.60	73	786	postgraduate	en	web	1	male-18-24.6-73-postgraduate-en-web
13	male	18	27.06	92	1090	postgraduate	en	web	1	male-18-27.06-92-postgraduate-en-web
14	male	18	31.16	35	478	phd	en	web	1	male-18-31.16-35-phd-en-web
15	male	18	37.72	55	909	postgraduate	en	web	7	male-18-37.72-55-postgraduate-en-web
16	male	18	37.72	56	925	postgraduate	en	web	1	male-18-37.72-56-postgraduate-en-web
17	male	18	39.36	48	827	phd	en	web	1	male-18-39.36-48-phd-en-web
18	male	18	39.36	59	1017	phd	en	web	7	male-18-39.36-59-phd-en-web
19	male	18	103.32	28	1267	phd	en	web	1	male-18-103.32-28-phd-en-web

- Unique customer IDs were assigned to the identified individual customers.
- 3 Mistakenly copied reviews were identified and were excluded from the sentiment analysis.

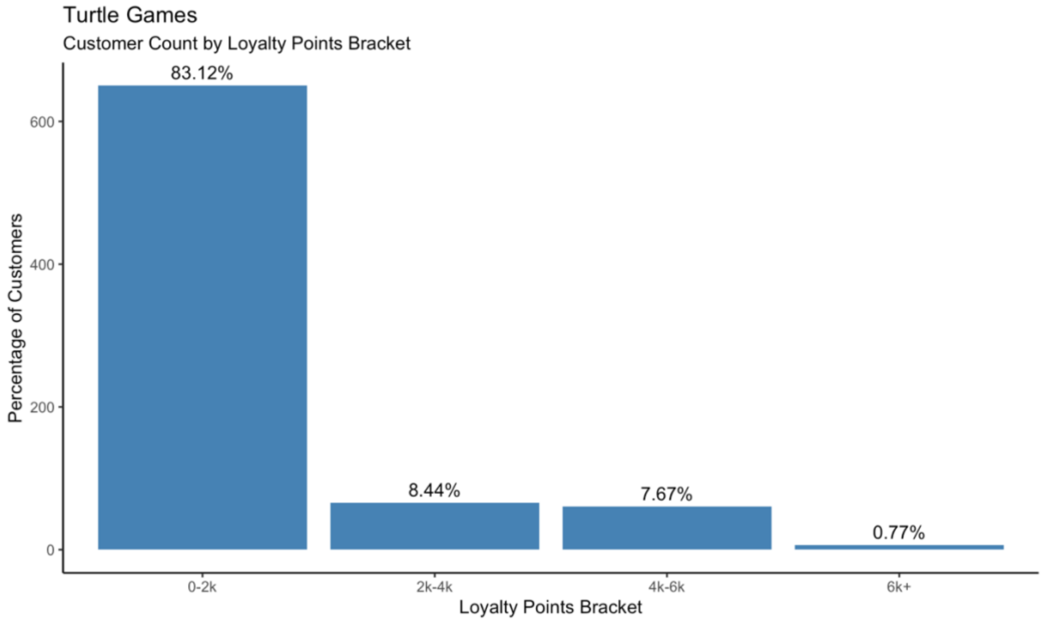
- on-time and nice item.
- this set is slightly worse than earlier ones (..
- great expansion to a great game.

Exploratory Analysis & Preliminary Insights

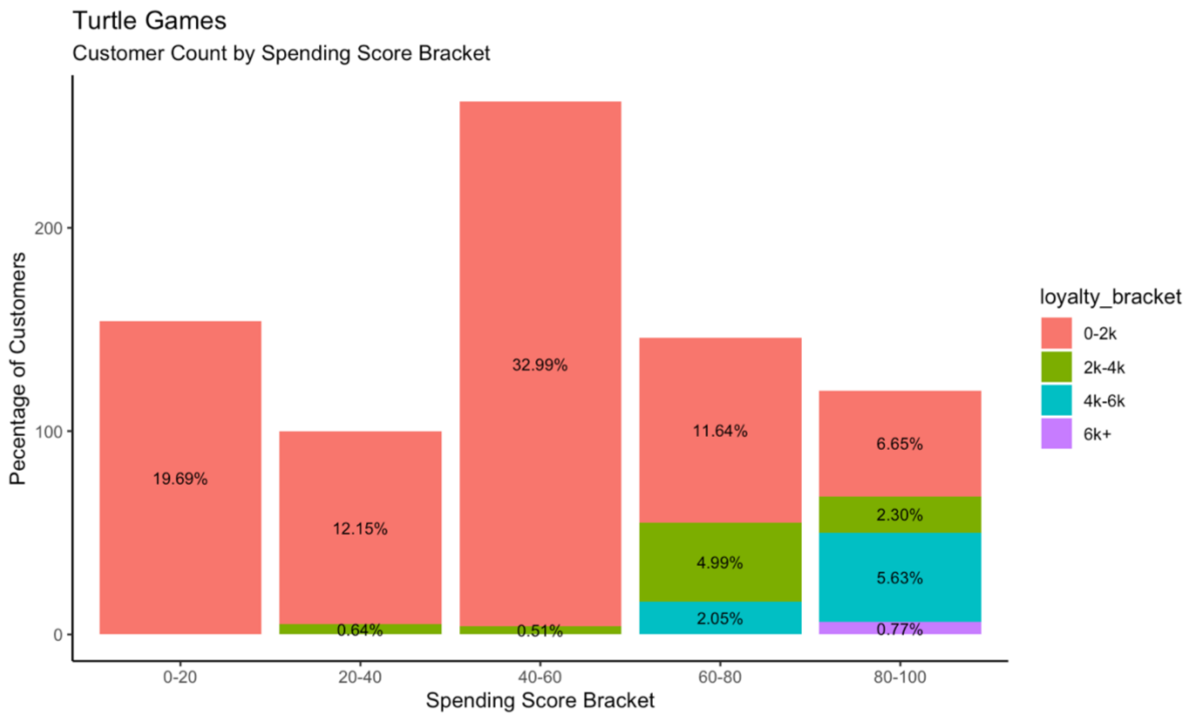
Exploratory Analysis & Trends

83.12% of the customers have between 0-2000 loyalty points, which strongly indicates low customer engagement with the current loyalty program.

Only 0.77% have above 6000 points.

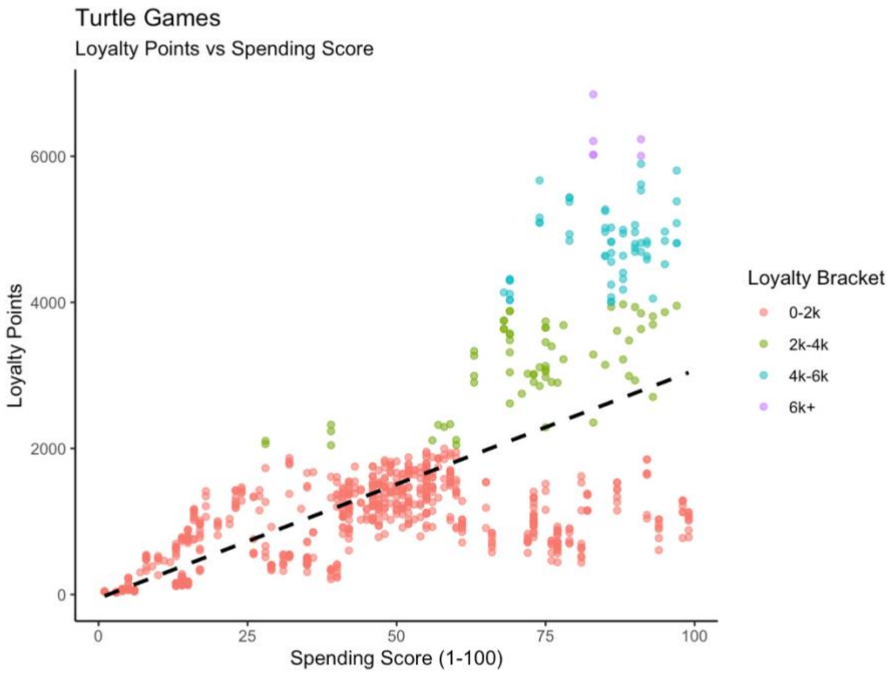


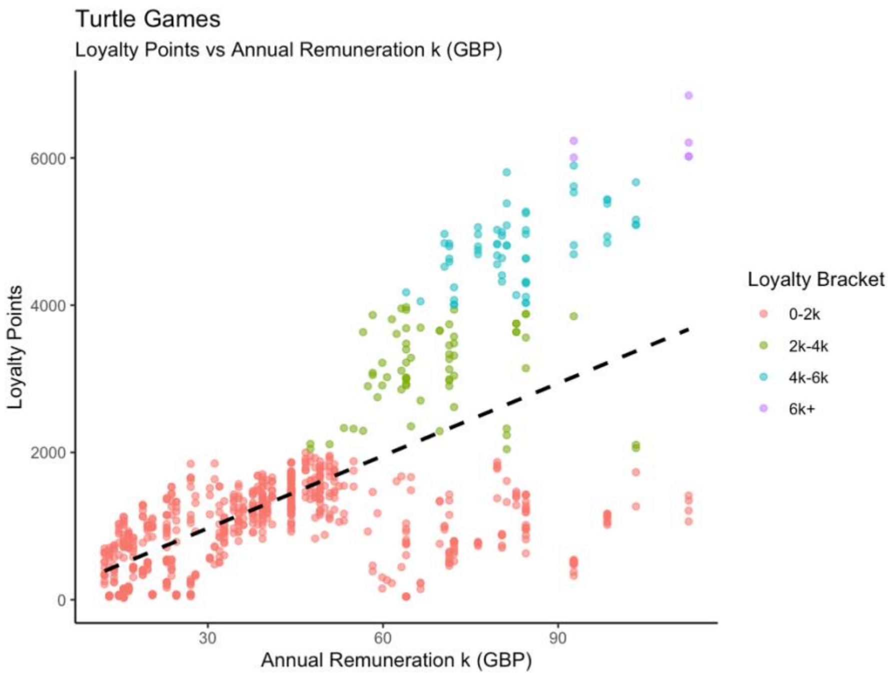
Almost 33% of the customer base has a spending score of 40-60, followed by 19.89% in the 0-20 spending score range.



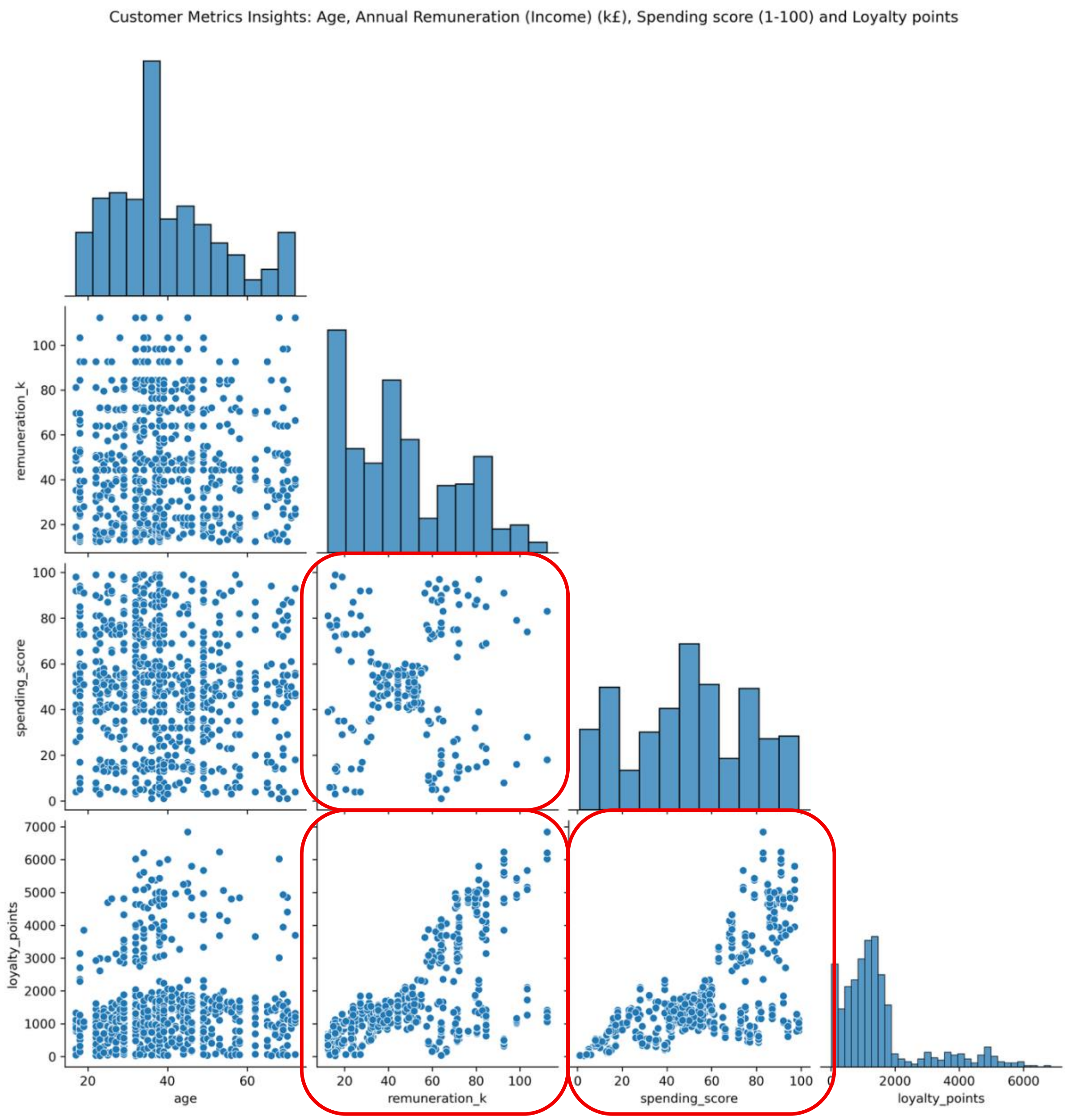
Linear Relationships & Clusters

A linear relationship between loyalty points, spending score, and annual remuneration was identified. This signifies that increases in spending score and annual remuneration are associated with an increase in loyalty points.





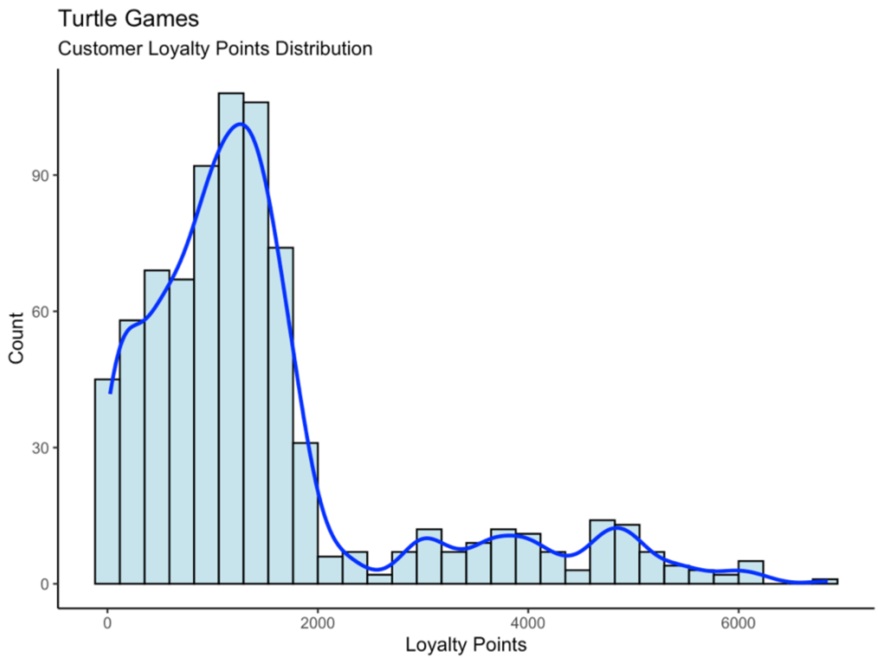
Data point clusters were identified on the spending score vs annual remuneration plot, which strongly suggested that there are customer segments in the data.



Data Distribution

The distribution of loyalty points is right-skewed, with a heavy concentration of loyalty points accumulation below the 2000 points mark. Since linear regression models assume residuals are normally distributed, a right-skewed dependent variable can lead to non-normal residuals.

The data was deemed not suitable for linear regression models, even after transforming the loyalty points and scaling the independent variables (see Appendix).



Data Modeling

Final Regression Predictive Model

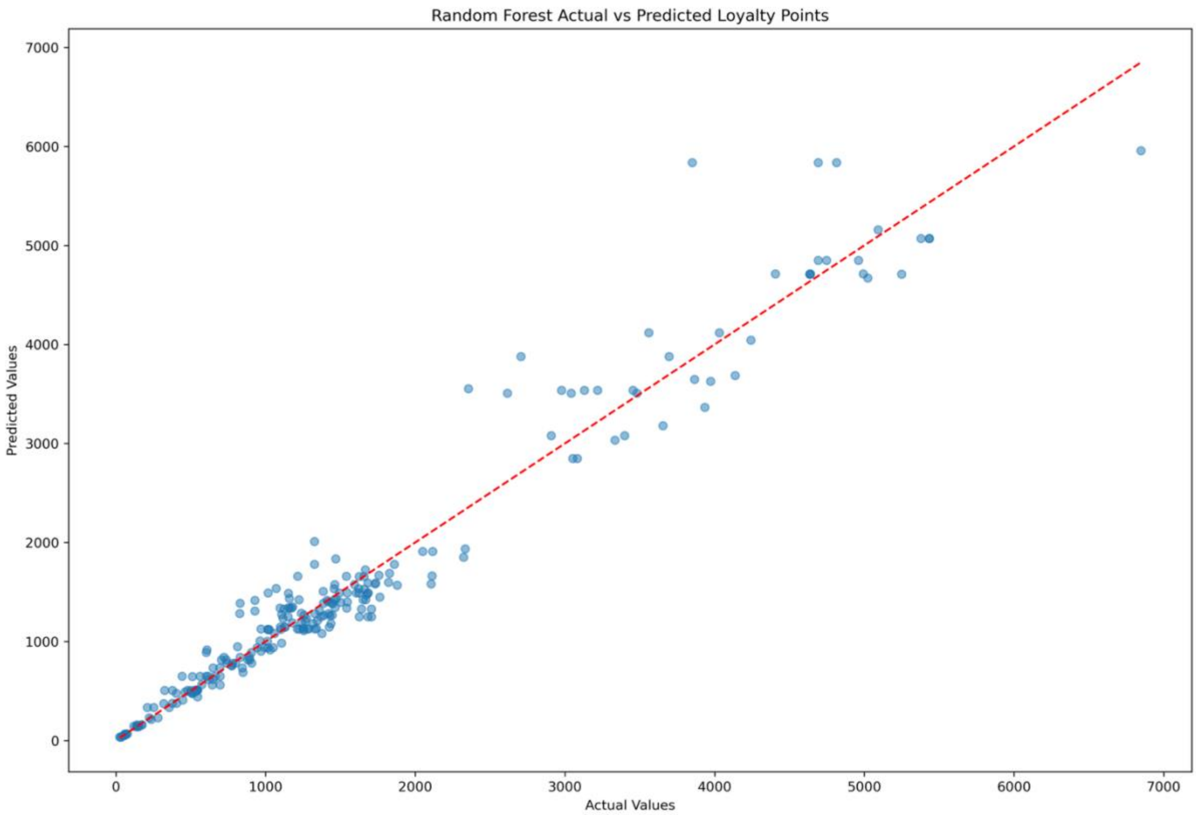
Approach

The importance of the independent variables was calculated, and factors with low importance were excluded to reduce noise and streamline the model.

The final model was built using the spending score and annual remuneration as these features showed the highest predictive power for loyalty points. No scaling was applied as it's not required for a random forest regressor using the Scikit-learn library (1).

Random Forest Predicted vs Actual Values Chart

The red line represents the perfect predictions; the closer the points are to this line, the better the model's predictions.



Random Forest Regressor Results

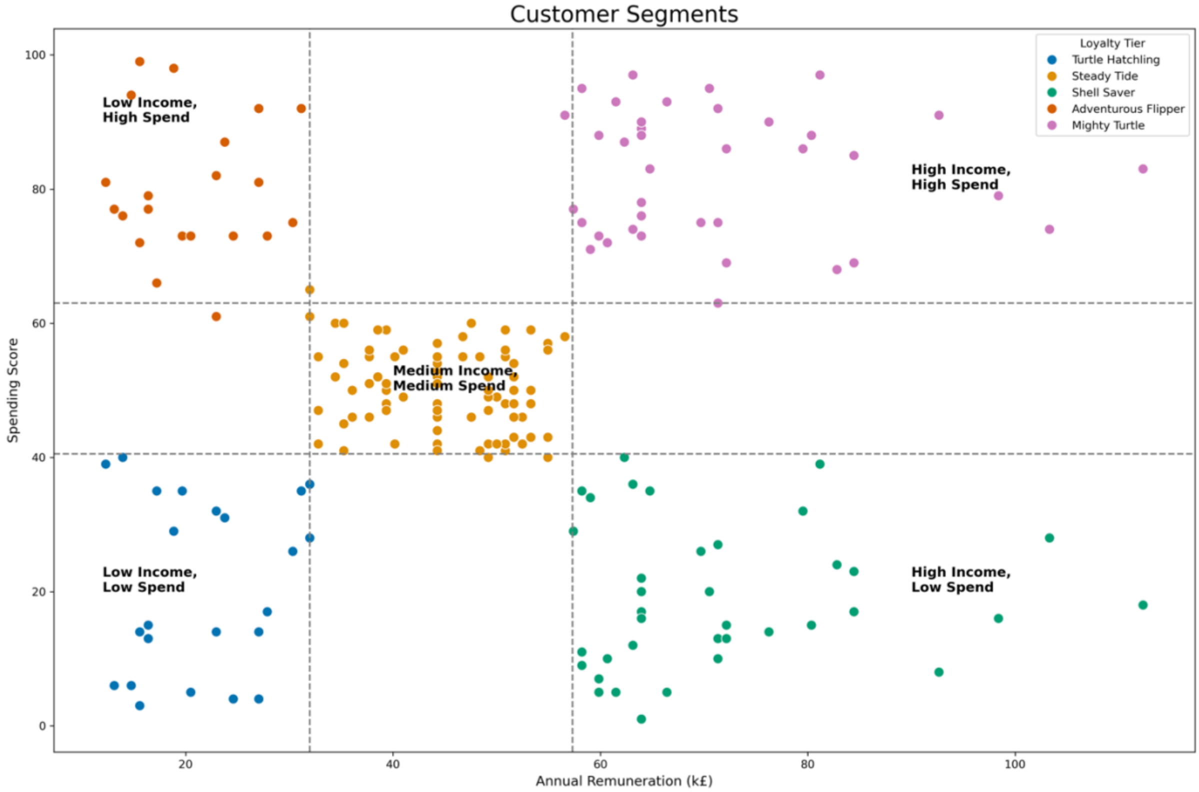
Given the spending score and annual remuneration of the customer, the model can predict 94.95% of the variability in the loyalty points.

```
Mean Absolute Error (MAE):178.90
Mean Squared Error (MSE):87995.57
Root Mean Squared Error (RMSE):296.64
R-squared: 94.95%
Avg % error: 2.62%
```

Clustering

To identify the clusters within the annual remuneration and spending scores, a hierarchical clustering model was considered, as it helps reveal the optimal number of clusters. It's, however, sensitive to unscaled data, so scaling had to be implemented to guarantee the accuracy of the model.

The model identified 5 clusters as shown below:



Medium Income/Medium Spend is the cluster with the highest percentage.

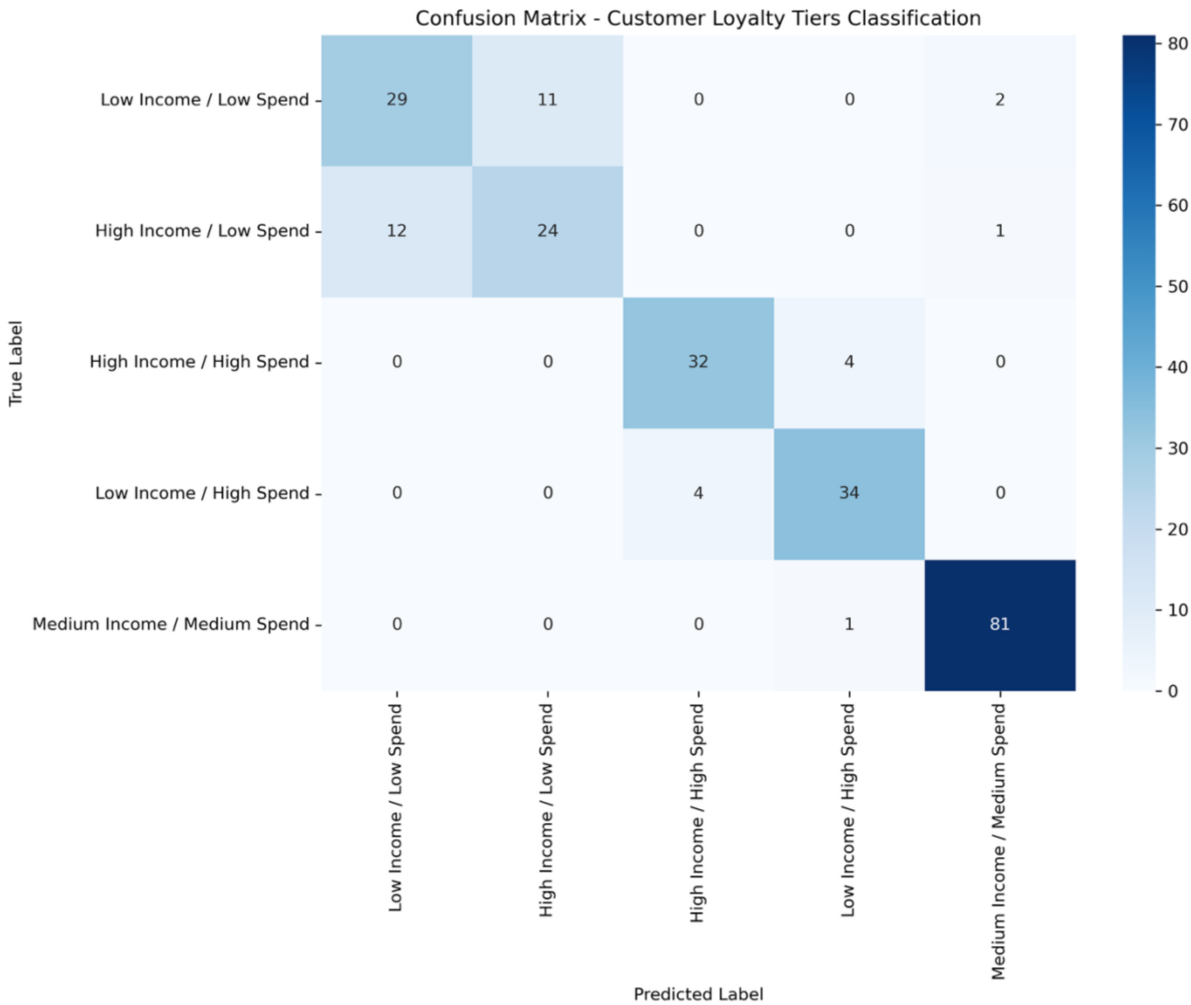
Cluster Segment	Count	Percentage
Medium Income / Medium Spend	273	34.91
Low Income / Low Spend	138	17.65
Low Income / High Spend	127	16.24
High Income / Low Spend	123	15.73
High Income / High Spend	121	15.47

This highly suggests the need for a tiered loyalty program to cater to the needs and status of the different segments.

Classification

To proactively assign a loyalty program tier to new customers and accordingly adapt marketing efforts, a classification model was built to classify new customers based on their spending score. This model has an 85% accuracy with the following precisions in predicting the different segments:

- Low Income / Low Spend → 71%
- High Income / Low Spend → 69%
- High Income / High Spend →89%
- Low Income / High Spend →87%
- Medium Income / Medium Spend → 96%

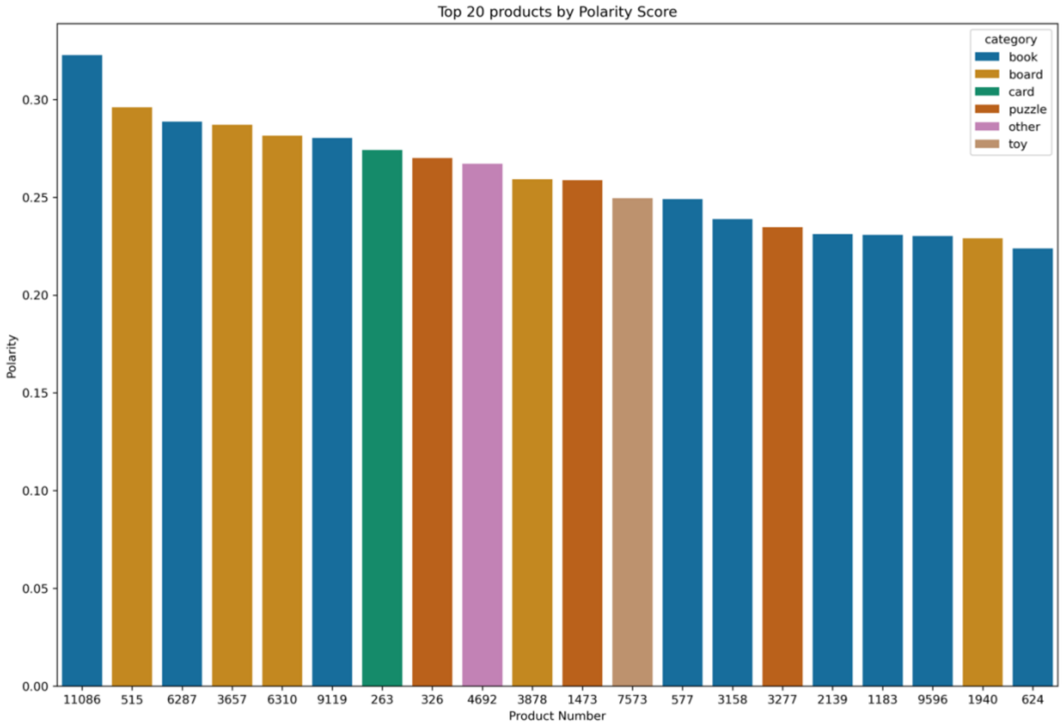


All High-Spend segments can be predicted with a precision of above 87%, which is satisfactory.

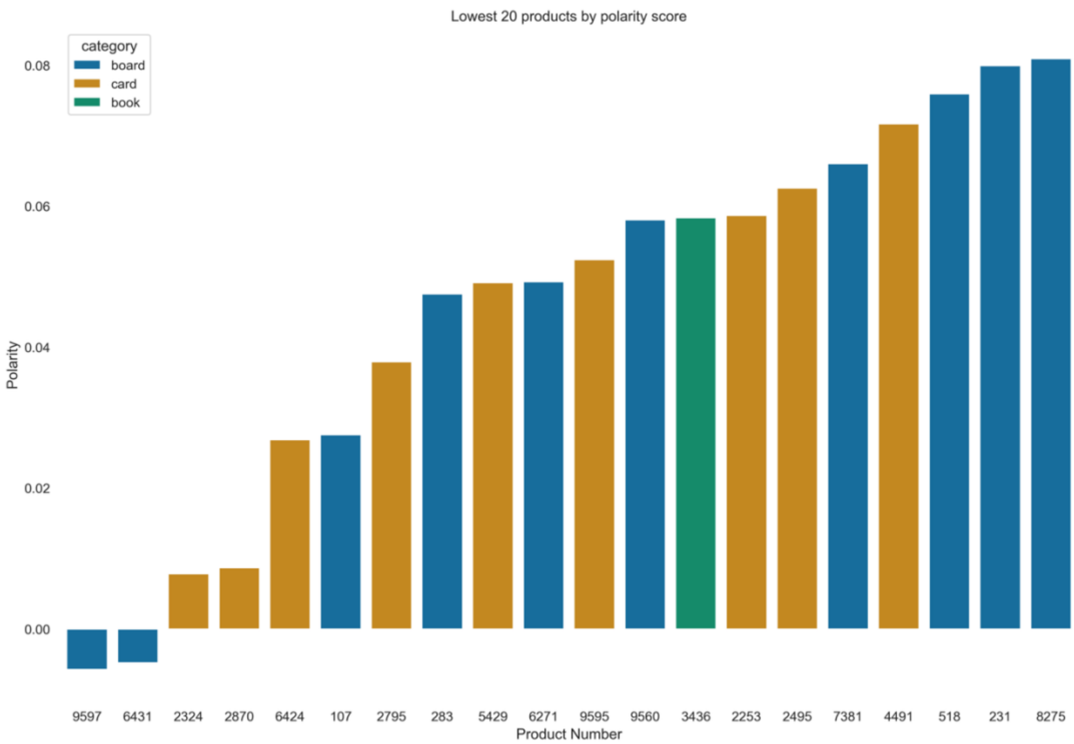
Reviews & Summaries Analysis

Products

Product reviews have a generally positive sentiment.



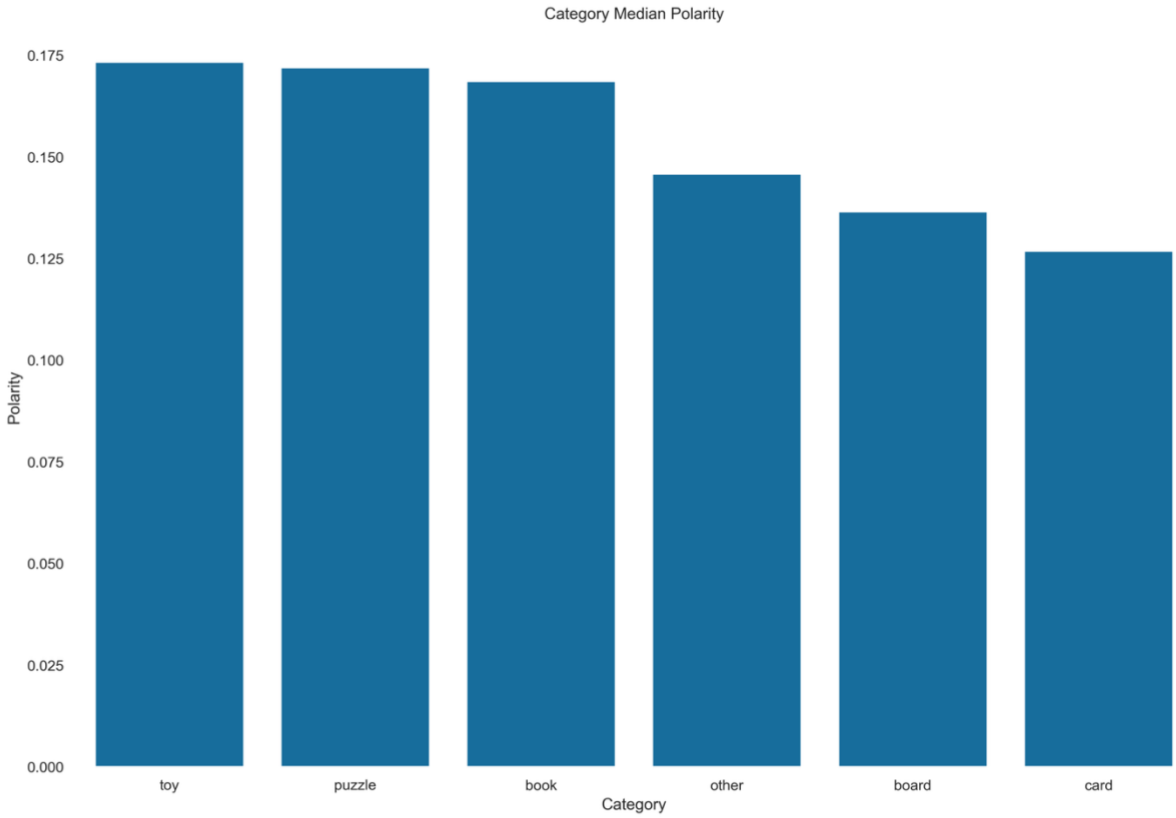
This is except for two products: **6431 & 9597**



Categories

Categories were assigned to the products using a document term matrix by assigning the category with the highest keyword count. There are limitations to this technique as it is based on the number of keywords in product reviews, but was still considered insightful.

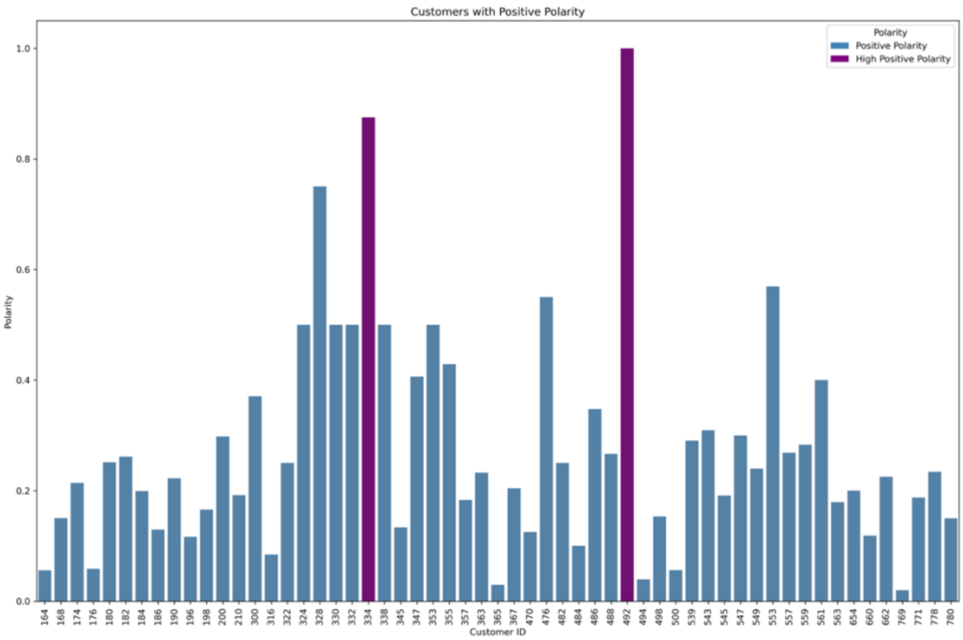
The median sentiment score was calculated and aggregated for each category. The toy category showed slightly higher sentiment scores, suggesting that toy reviews can be more emotionally expressive, due to their association with children.



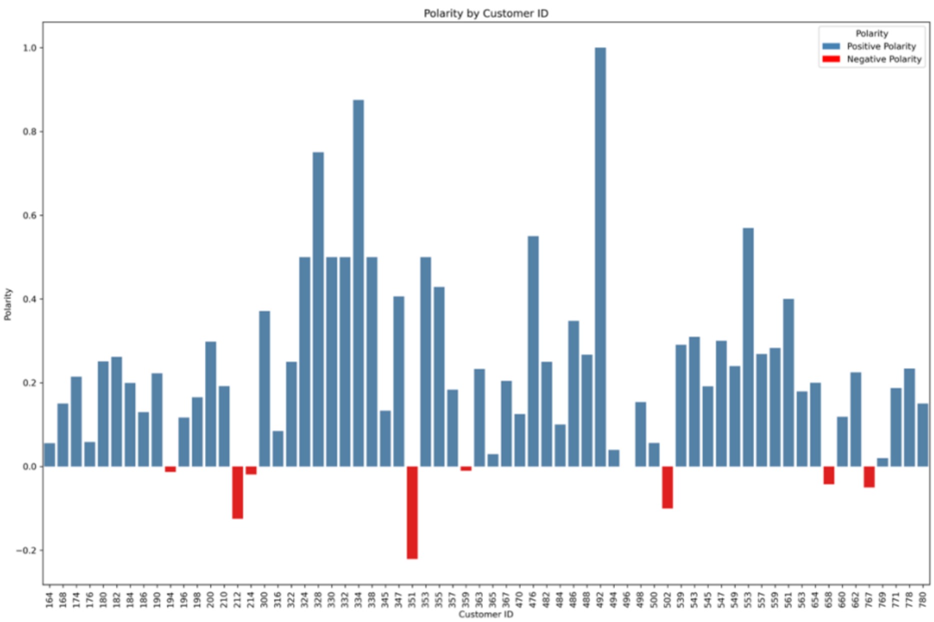
High-Loyalty Customers

Filtering the data to only include customers with >= 4000 loyalty points, two insights were uncovered.

There are 2 customers (334 & 492) whose reviews scored high positive polarity, who could be approached to become brand ambassadors or further engaged with exclusive events and perks.



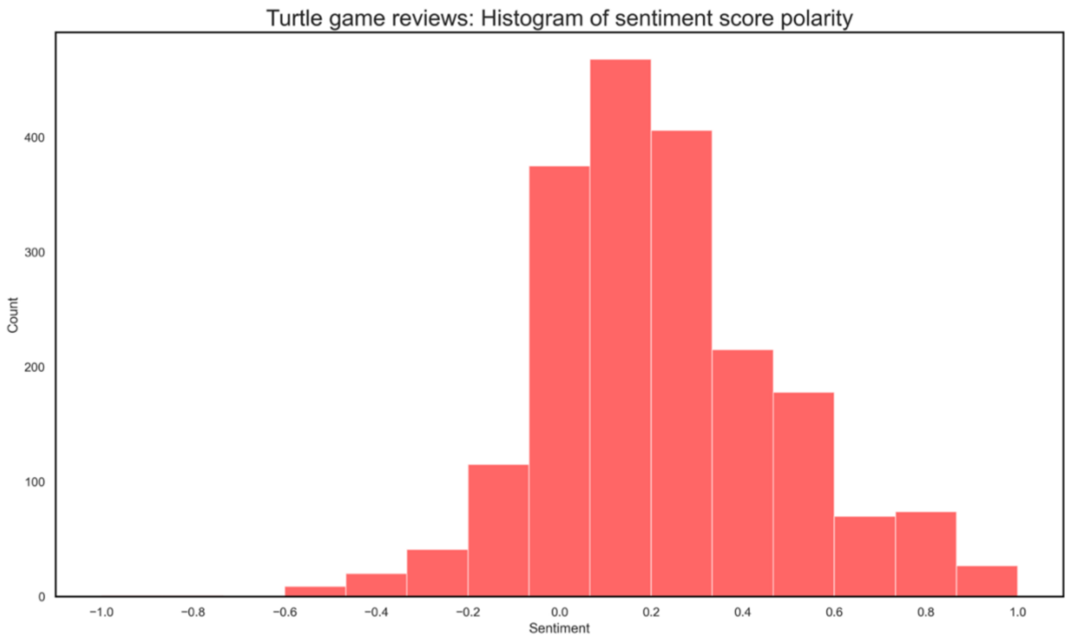
There are, however, 8 customers in this high loyalty bracket, whose reviews have negative polarity and who are at risk of churning. Their concerns should be proactively addressed to minimize that risk.



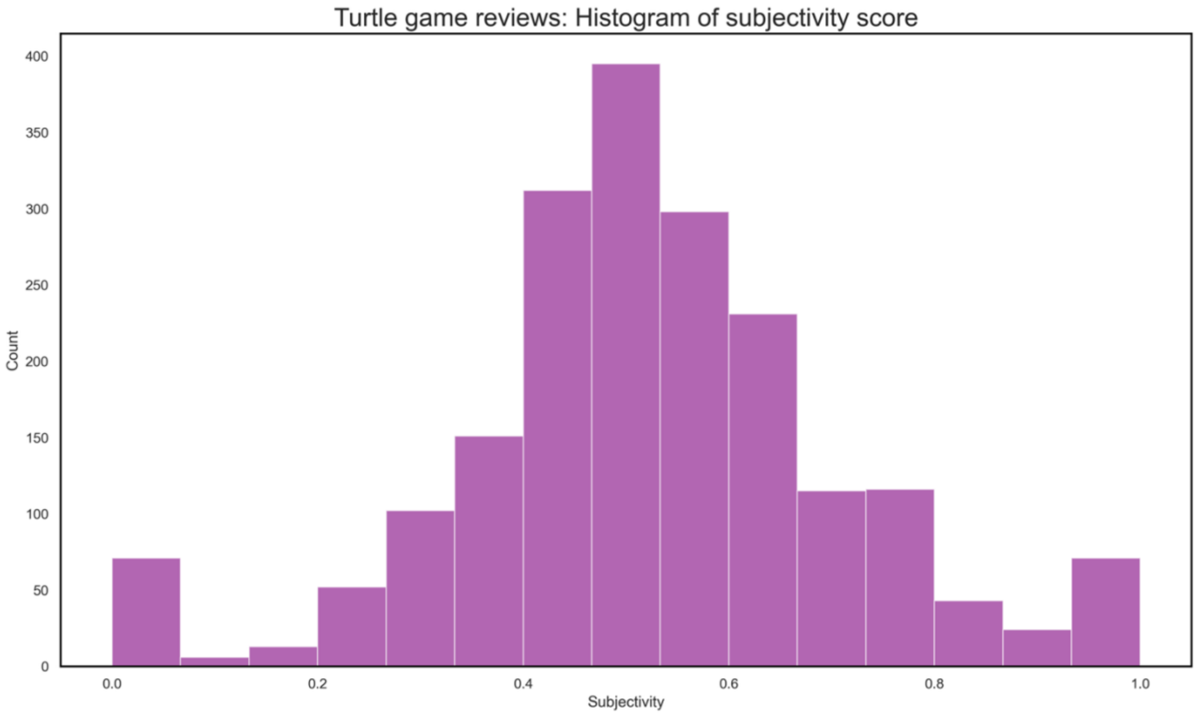
Global Sentiment Analysis

The reviews' word cloud was mostly dominated by subjective words such as game, card, play… etc.

This was also confirmed through the sentiment score histogram, which shows the highest density for 0.2 sentiment scores. This shows that reviews are not dominated by strong opinions or emotions.



Checking the subjectivity score, we can also confirm that most reviews are a mix of facts and opinions.


Turtle game reviews: Histogram of subjectivity score
400
350
300
250
200
150
100
50
0
Count
0.0
0.2
0.4
0.6
0.8
1.0
Subjectivity

Recommendations & Next Steps

Recommendations

- Build a tiered loyalty program that rewards high-value customers (2)
- Proactively assign new customers to their corresponding segments using the classification model with an accuracy of 85% for better target marketing.
- Provide loyalty-based discounts for the medium income / medium spend (~34% of the customers) to encourage retention and higher sales (3).
- Proactively address high-loyalty customers with negative sentiment to reduce the risk of losing them, as 45% of customers are likely to churn when their concerns are not addressed promptly (4)
- Leverage High-loyalty customers with high positive sentiment as brand ambassadors.
- Investigate the reason behind negative reviews on products 6431 and 9597.

Loyalty Program Future Health Check

- Check the sales and loyalty program performance (5) quarterly
- Recheck the Customer segments a year after the analysis to account for changes

Appendix

Five Whys Framework:

Turtle Games aims to understand its customer base and its loyalty program scheme.

Why does Turtle Games want to understand its customer base and its loyalty program scheme?

To be able to invest in a targeted marketing strategy to reach their high-return customers

Why is Turtle Games considering a targeted marketing strategy?

To avoid overspending on low-return customers and focus on better engaging high-return customers through exclusive events and valuable perks.

Why does Turtle Games want to focus on its high-return customers?

Satisfied high-return customers can function as brand ambassadors and can also be drivers for increased positive sentiment around the brand's products.

Why is it useful to leverage high-return customers as brand ambassadors?

This would help maximize the benefits of the marketing strategy and improve the overall sales performance

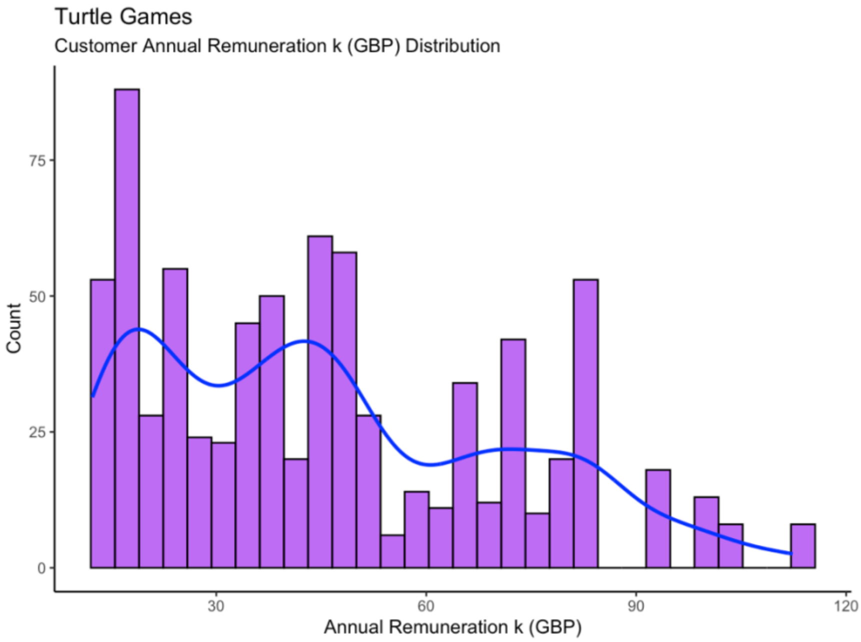
Why is it important for Turtle Games to improve its overall sales performance?

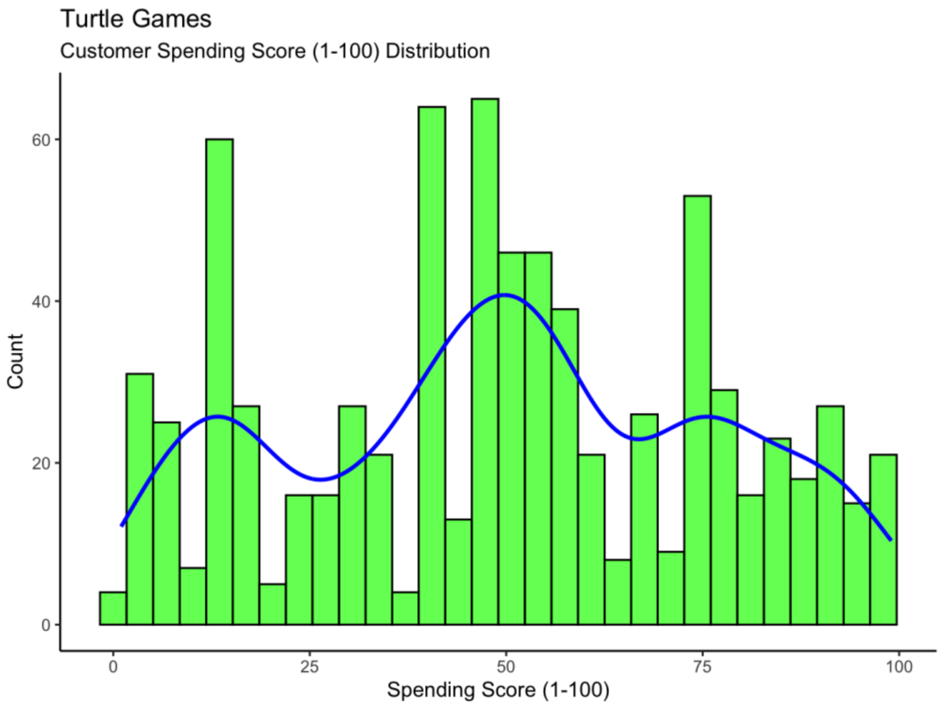
Because, as a retailer of other brands, Turtle Games aims to build a reputable portfolio of products and brands that would make it stand out against the competition as a market leader in that segment.

Data Distribution

Histograms

Both the annual remuneration and spending score distributions are multimodal. The multiple peaks in the distribution strongly suggest that there are subgroups within the data that have different trends and behaviors.



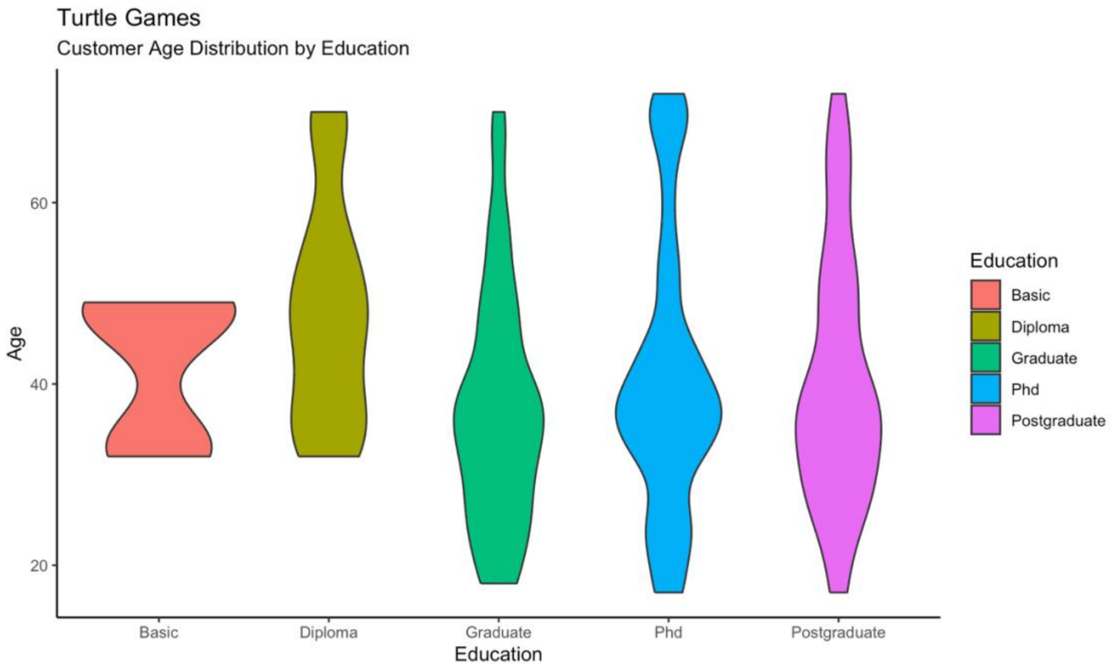


Violin Plots – Customer Demographic

Age Distribution

Male customers dominate the younger and older age groups, while female customers are more concentrated in the age bracket between 30 – 50 years of age.

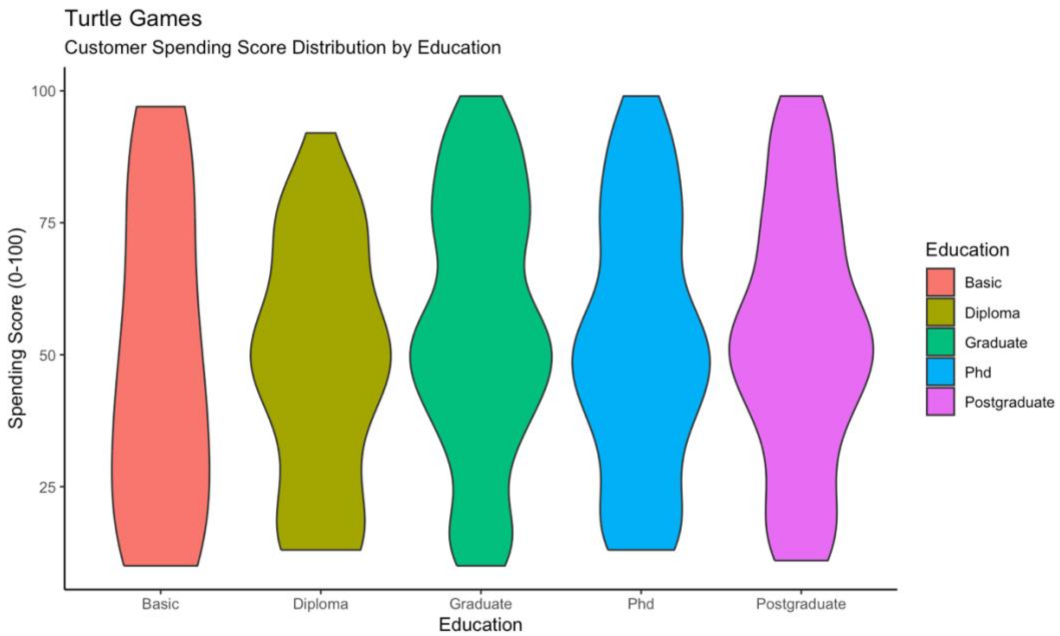
The data is showing very young customers with Graduate, Postgraduate, and Phd degrees. These will be highlighted to Turtle Games as a data quality issue. Although they seem highly illogical, these data points are not statistical outliers as confirmed by the boxplot and the IQR calculation.



Spending Score

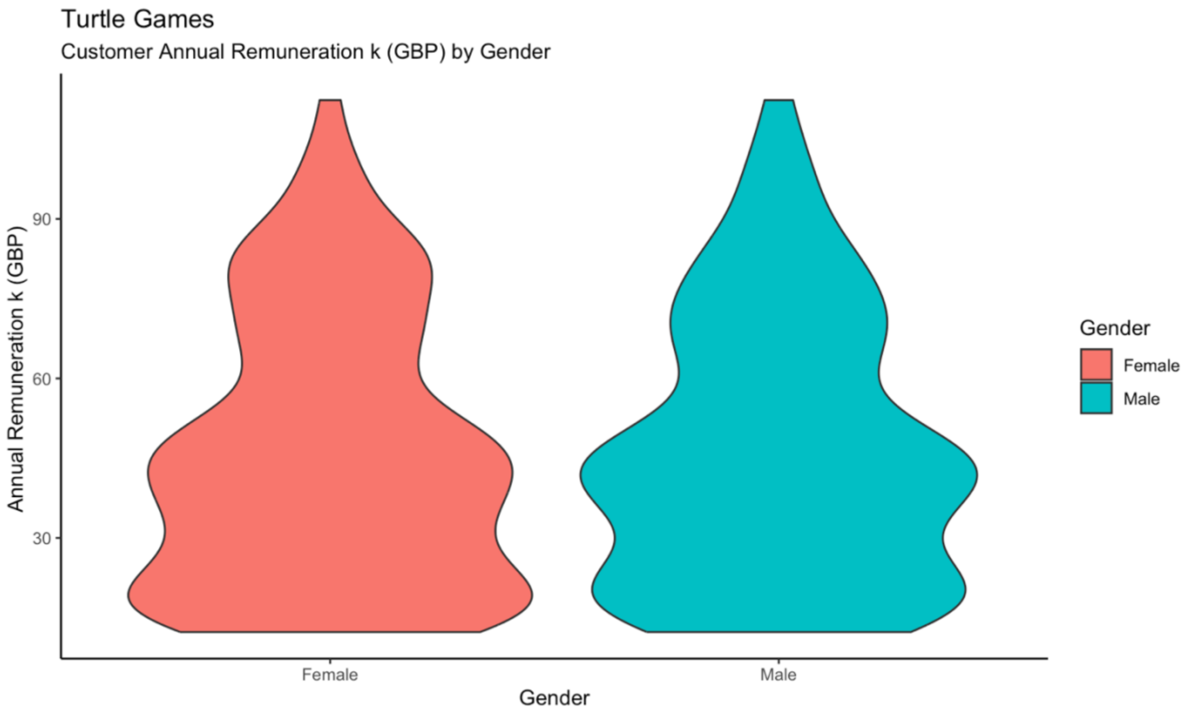
No major visible difference between the distribution of Male and Female customers, they more or less follow the same pattern.

A bigger proportion of the higher-education customers have higher spending scores compared to customers with basic educational levels.

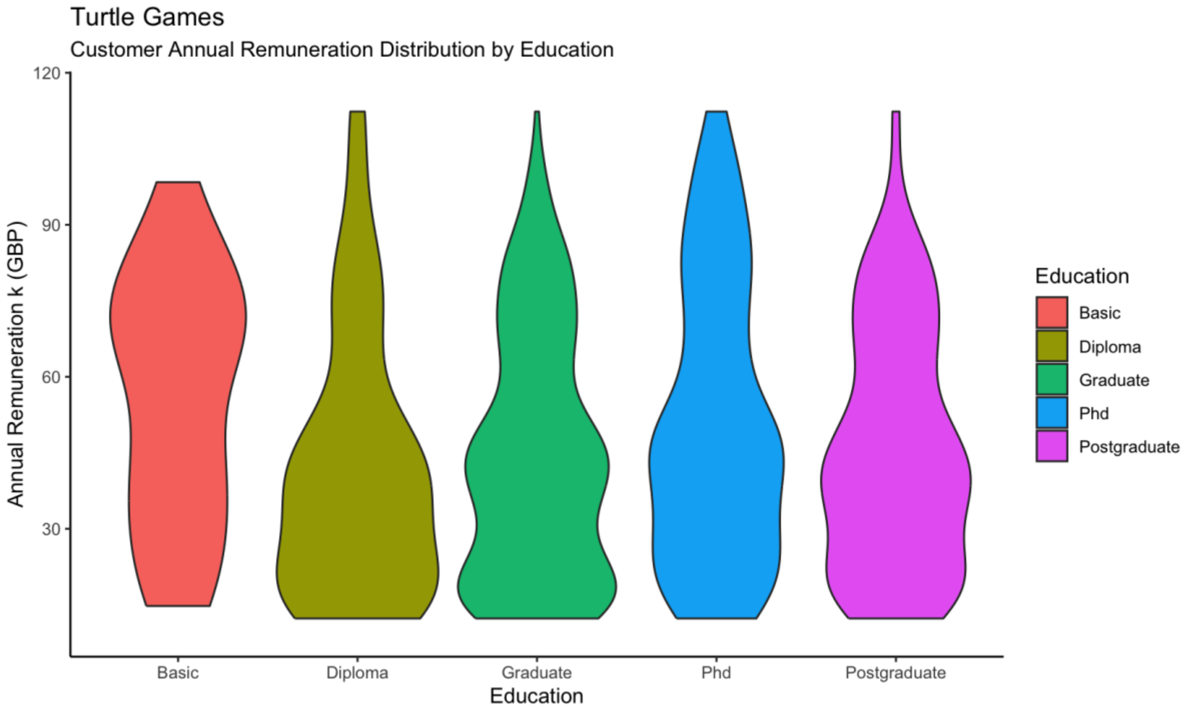


Annual Remuneration

The variation in annual remuneration doesn't differ significantly between genders; they more or less follow the same patterns.



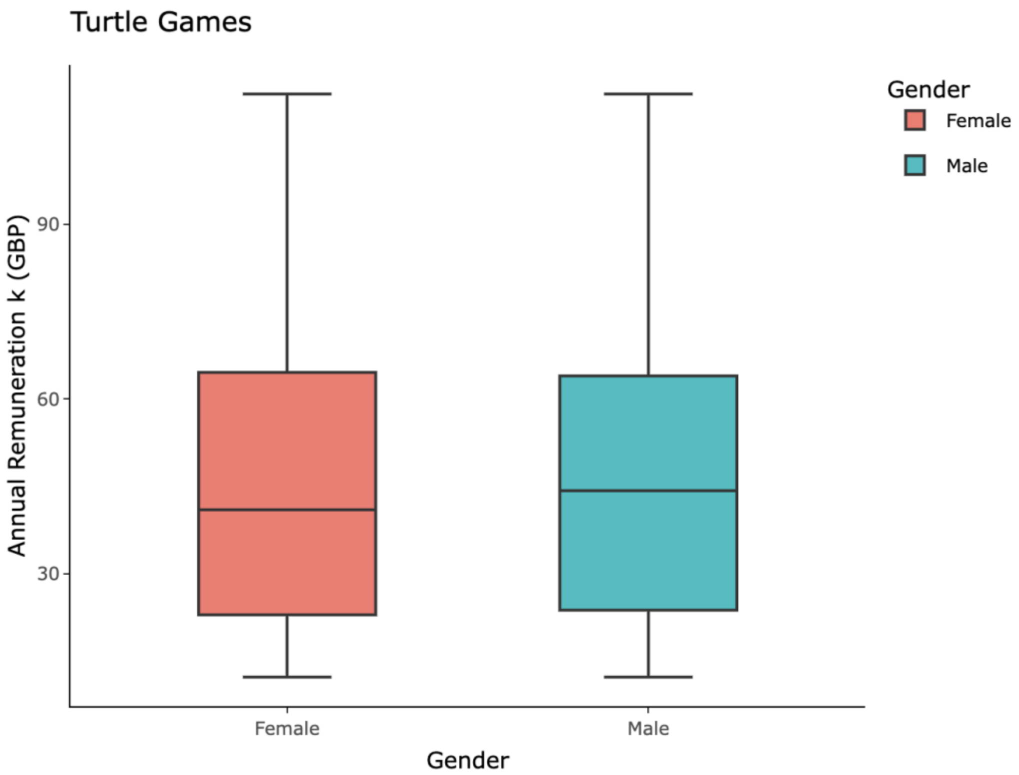
Customers with a basic education make up the largest portion of those earning above 60k annually; other customers with other educational levels are concentrated in lower income brackets.

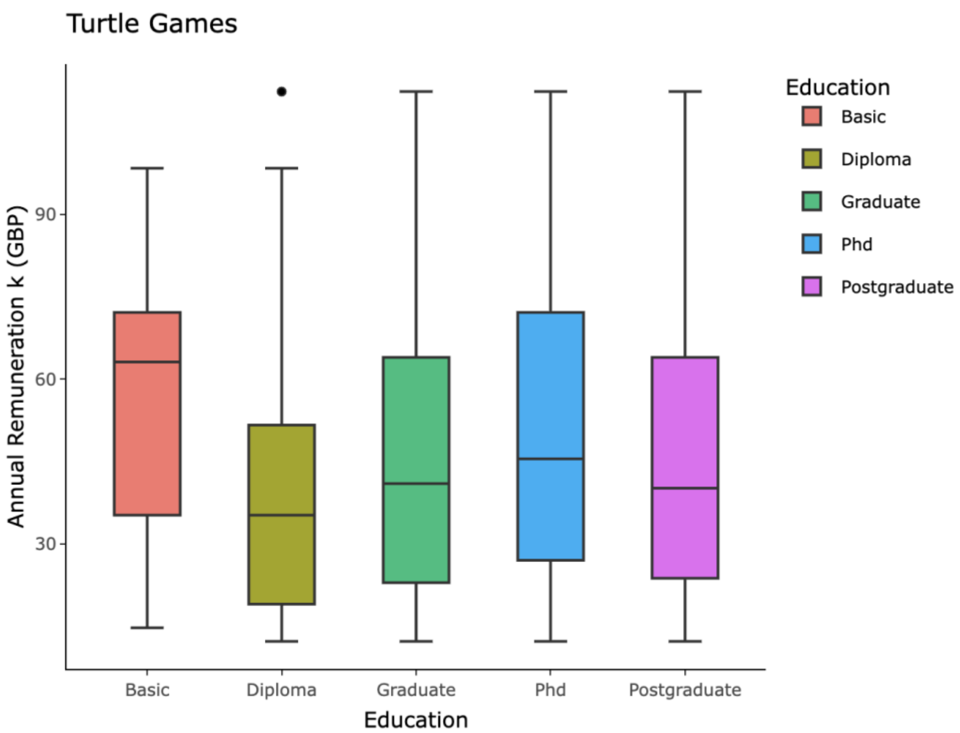


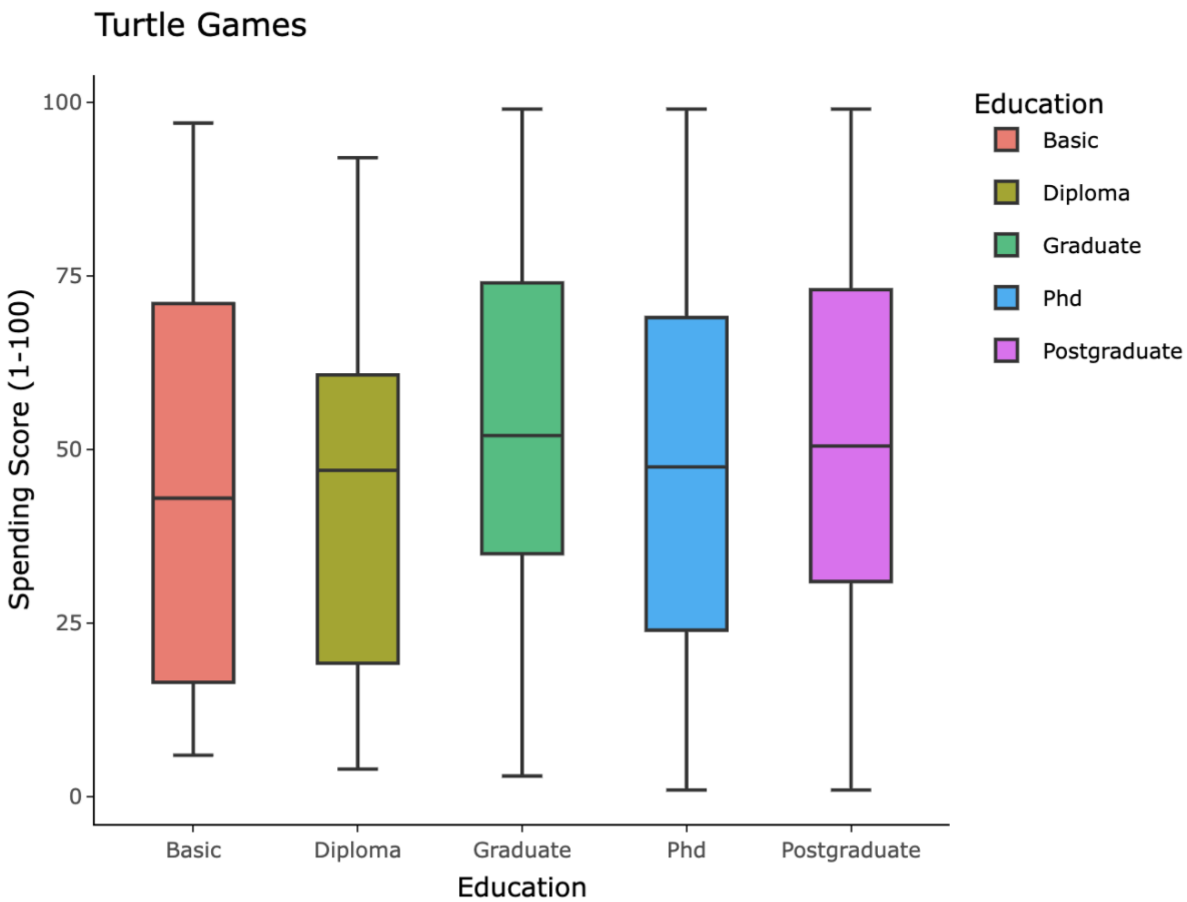
Outlier Analysis

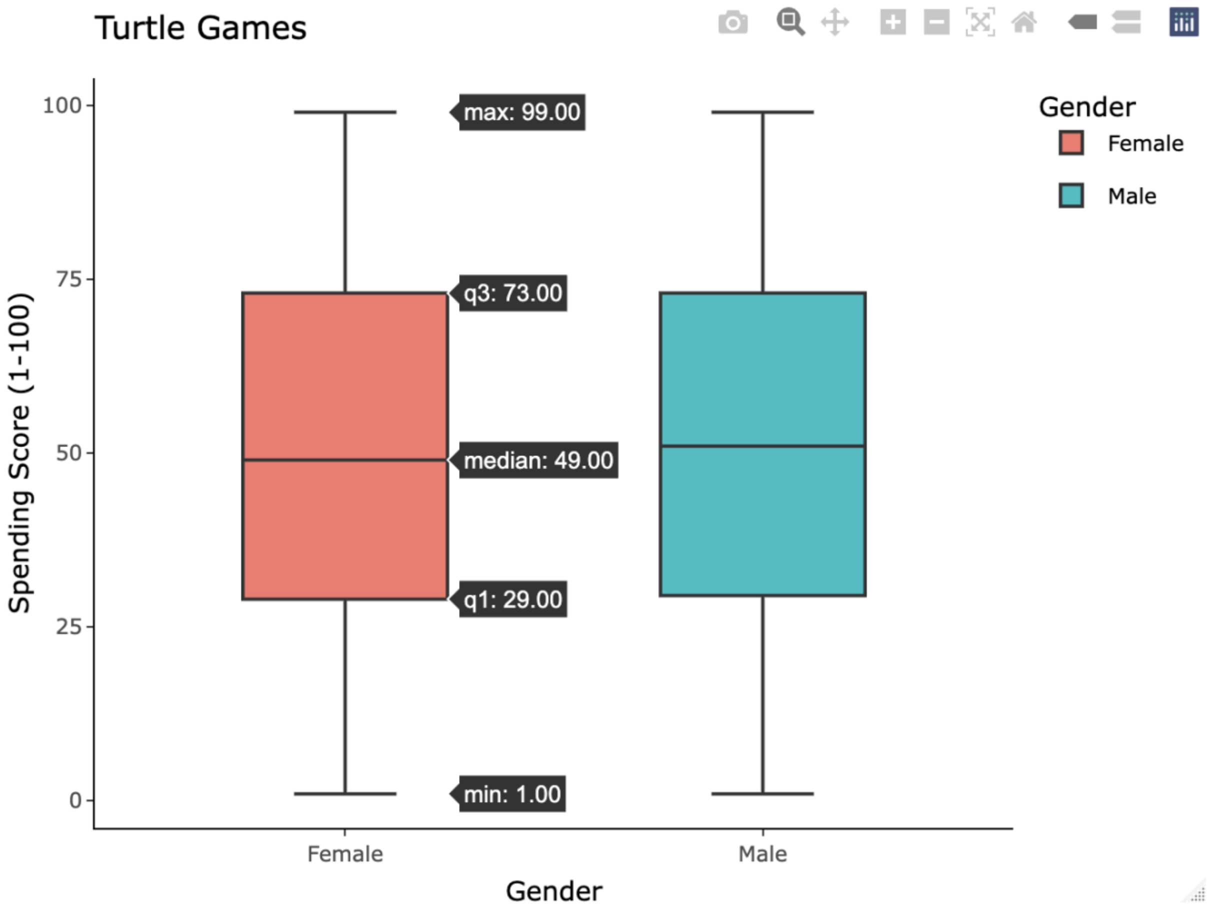
No outliers identified

Annual Remuneration and Spending Score Data exhibited no statistical outliers, with only one exception: a single high remuneration value in the remuneration chart. It was retained in the data as it didn't appear erroneous.



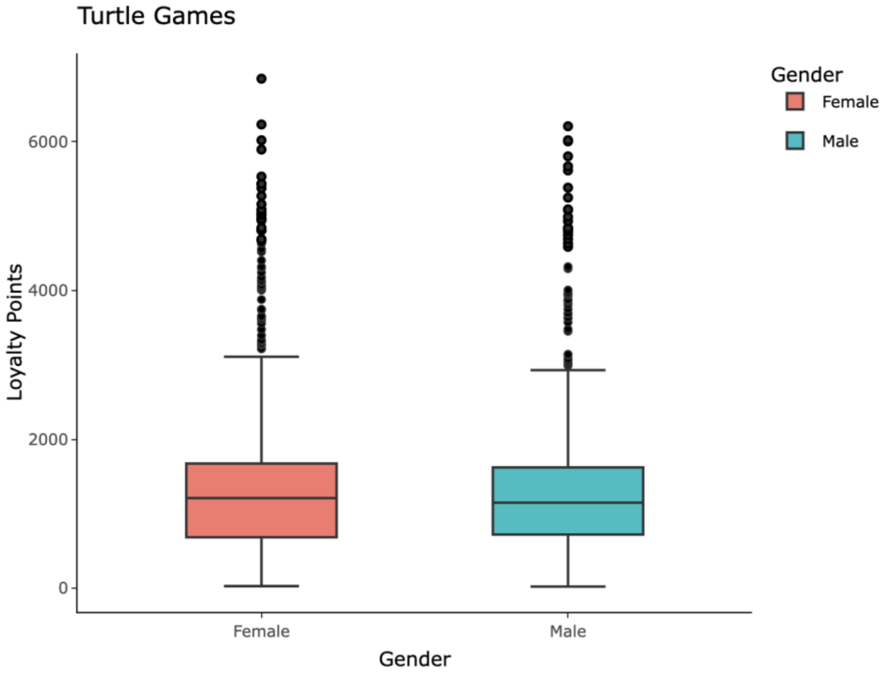


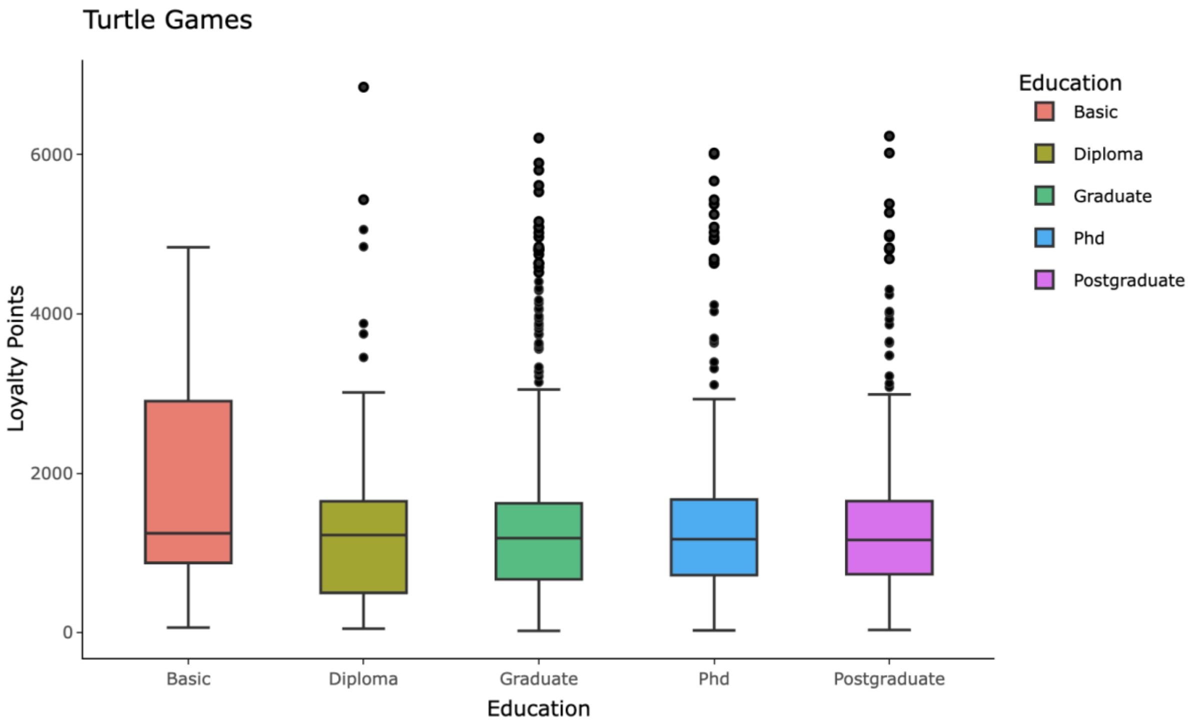

Turtle Games
Spending Score (1-100)
100
75
50
25
0
Basic
Diploma
Graduate
Phd
Postgraduate
Education
Education
Basic
Diploma
Graduate
Phd
Postgraduate


Turtle Games
Spending Score (1-100)
100
75
50
25
0
max: 99.00
q3: 73.00
median: 49.00
q1: 29.00
min: 1.00
Female
Male
Gender
Gender
Female
Male

Identified outliers

Almost 12% of the customers have loyalty points that are considered statistical outliers. They were all, however, retained as they constitute a big portion of the data, and their presence reflects natural behavior common in loyalty programs.





Simple Linear Regression

Linear Regression was considered as a preliminary model to predict loyalty points.

Simple linear regression models were built using spending score, annual remuneration, and age as independent variables.

All three models violated a key linear regression assumption, as heteroscedasticity is visible on all residuals' charts.

In addition to heteroscedasticity, Loyalty vs Age violates the linearity assumption. R-squared for all models was also below 40%.

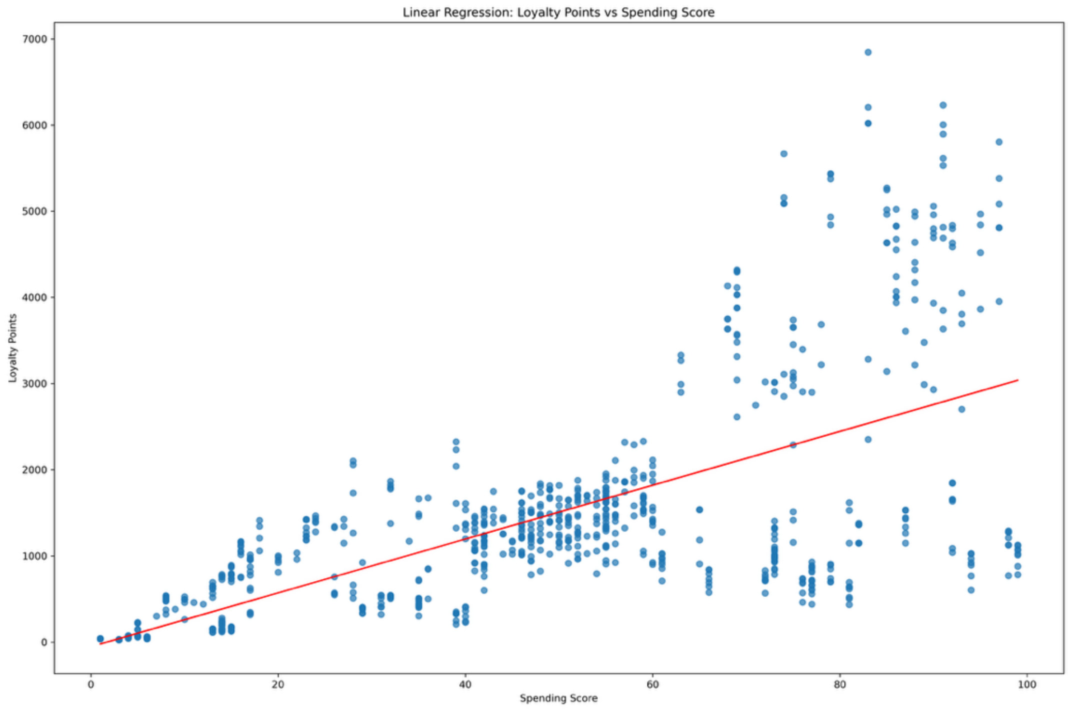
Spending Score

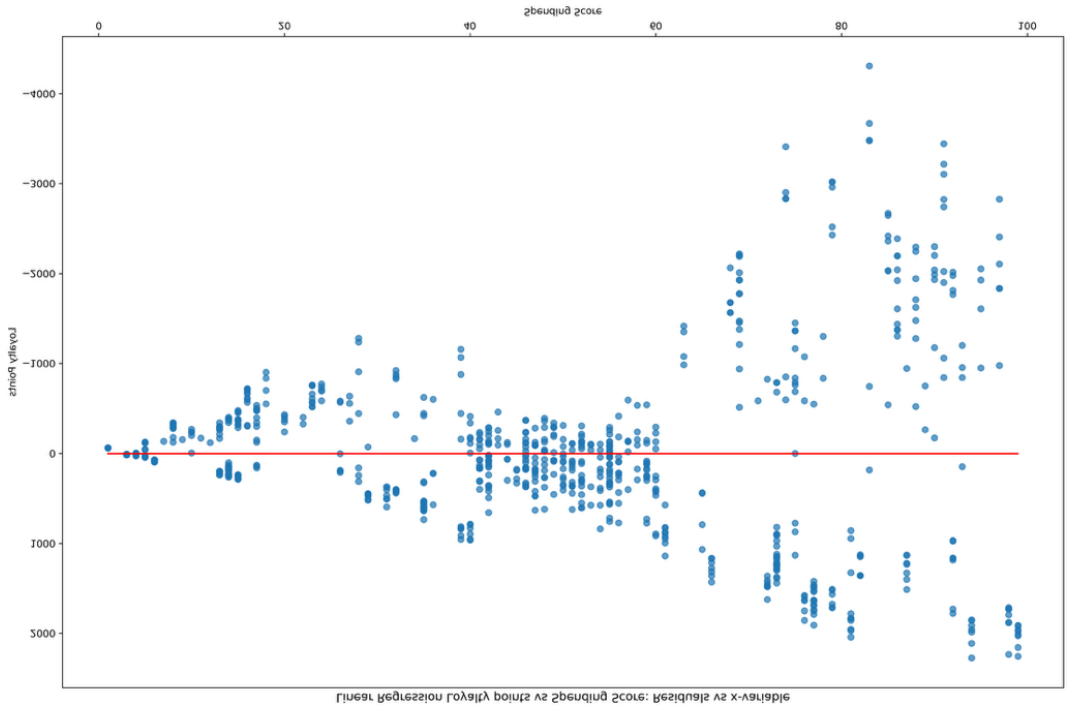
OLS Regression Results

Dep. Variable:	y_sp	**R-squared:**	0.400
Model:	OLS	**Adj. R-squared:**	0.399
Method:	Least Squares	**F-statistic:**	519.1
Date:	Fri, 18 Jul 2025	**Prob (F-statistic):**	1.80e-88
Time:	19:05:10	**Log-Likelihood:**	-6524.6
No. Observations:	782	**AIC:**	1.305e+04
Df Residuals:	780	**BIC:**	1.306e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-51.0144	77.103	-0.662	0.508	-202.369	100.340
x_sp	31.2151	1.370	22.783	0.000	28.526	33.905

Omnibus:	88.013	**Durbin-Watson:**	1.125
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	148.229
Skew:	0.739	**Prob(JB):**	6.49e-33
Kurtosis:	4.537	**Cond. No.**	119.


Linear Regression: Loyalty Points vs Spending Score
Loyalty Points
Spending Score
0
1000
2000
3000
4000
5000
6000
7000
0
20
40
60
80
100


Linear Regression Loyalty points vs Spending Score: Residuals vs x-variable
Loyalty Points
Spending Score
−4000
−3000
−2000
−1000
0
1000
2000
0
20
40
60
80
100

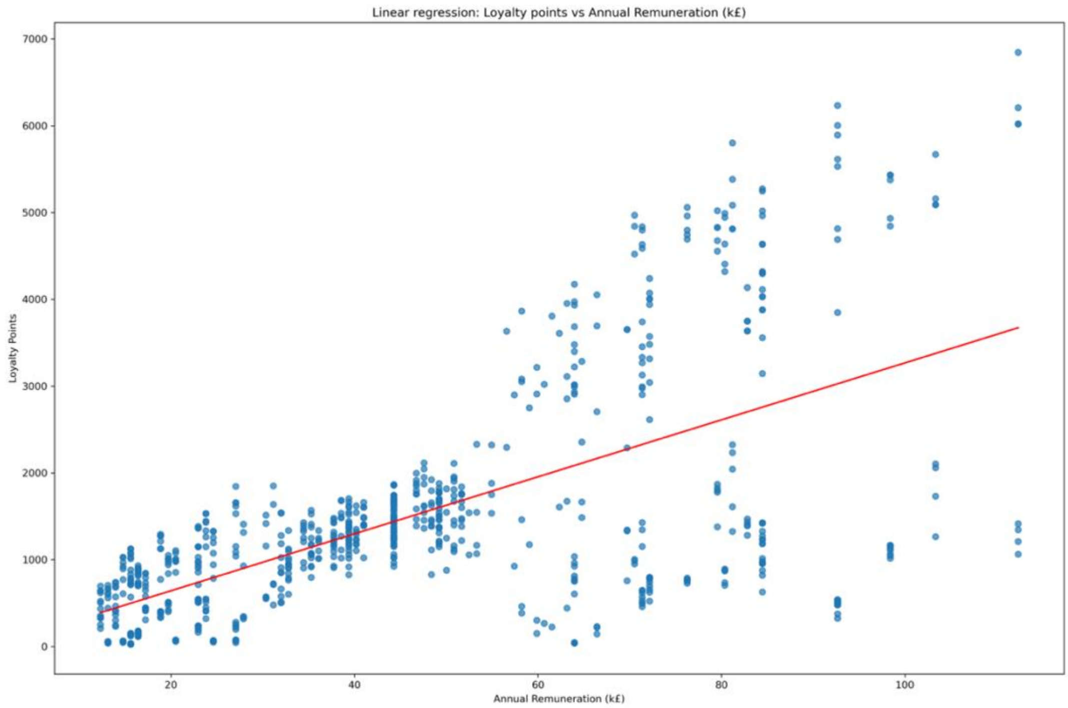
Annual Remuneration

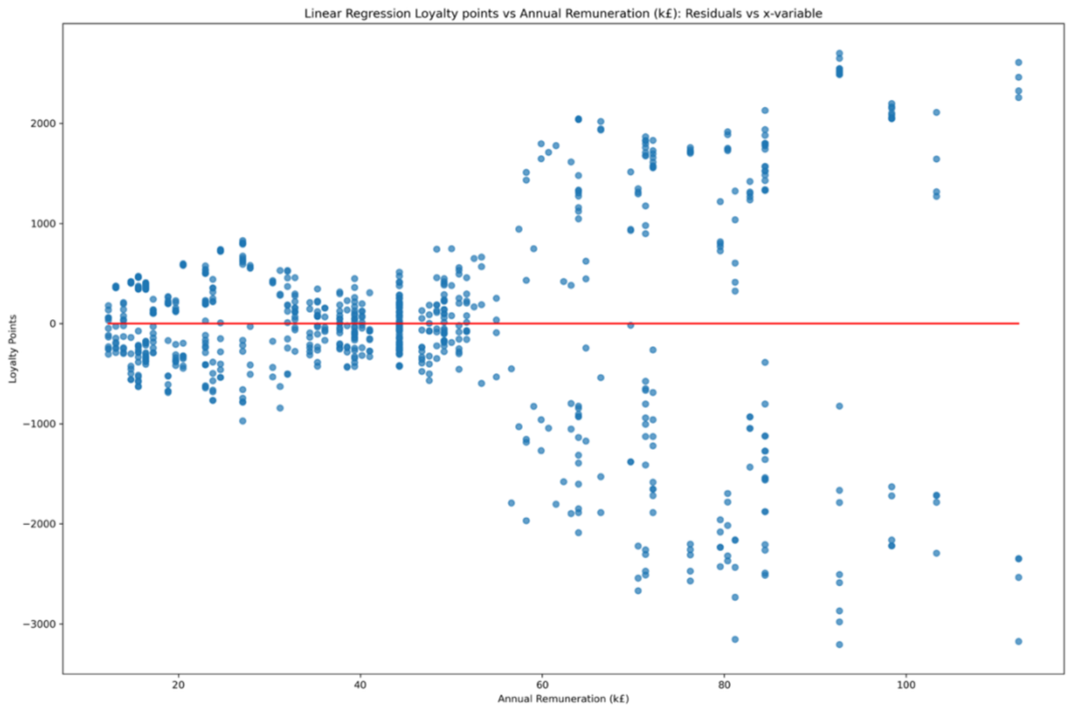
OLS Regression Results

Dep. Variable:	y_rn	**R-squared:**	0.398
Model:	OLS	**Adj. R-squared:**	0.397
Method:	Least Squares	**F-statistic:**	515.8
Date:	Fri, 18 Jul 2025	**Prob (F-statistic):**	4.87e-88
Time:	19:05:11	**Log-Likelihood:**	-6525.6
No. Observations:	782	**AIC:**	1.306e+04
Df Residuals:	780	**BIC:**	1.306e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-13.9200	75.881	-0.183	0.854	-162.875	135.035
x_rn	32.8132	1.445	22.710	0.000	29.977	35.649

Omnibus:	21.752	**Durbin-Watson:**	2.184
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	36.593
Skew:	0.205	**Prob(JB):**	1.13e-08
Kurtosis:	3.977	**Cond. No.**	109.


Linear regression: Loyalty points vs Annual Remuneration (k£)
Loyalty Points
Annual Remuneration (k£)
0
1000
2000
3000
4000
5000
6000
7000
20
40
60
80
100


Linear Regression Loyalty points vs Annual Remuneration (k£): Residuals vs x-variable
Loyalty Points
Annual Remuneration (k£)
−3000
−2000
−1000
0
1000
2000
20
40
60
80
100

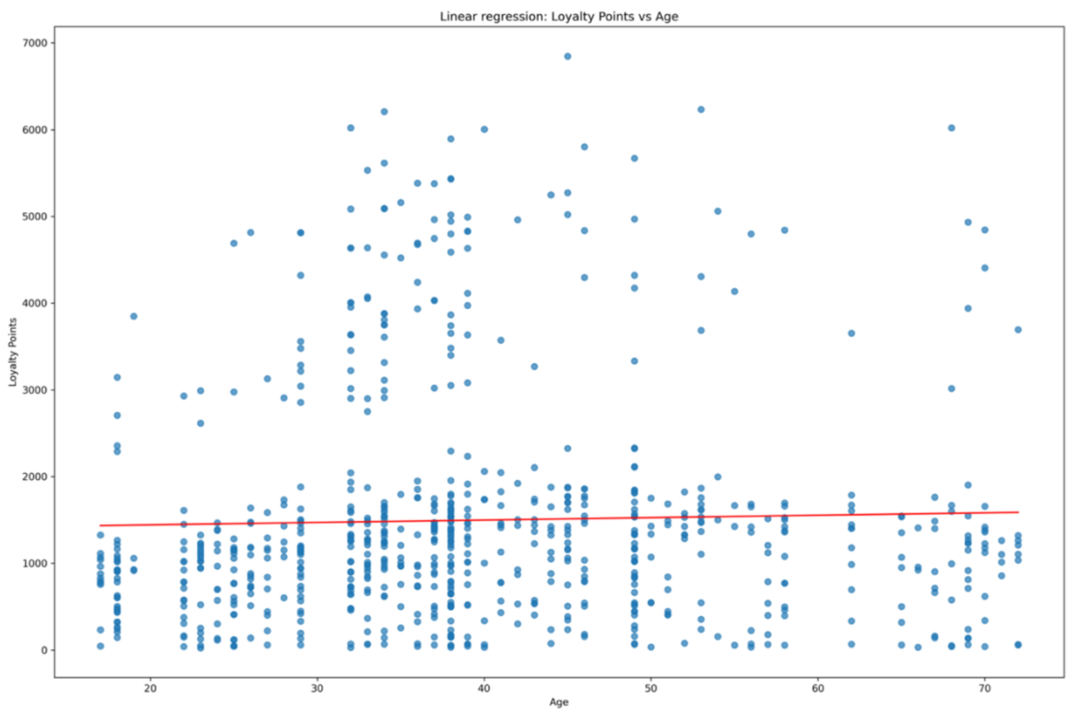
Age

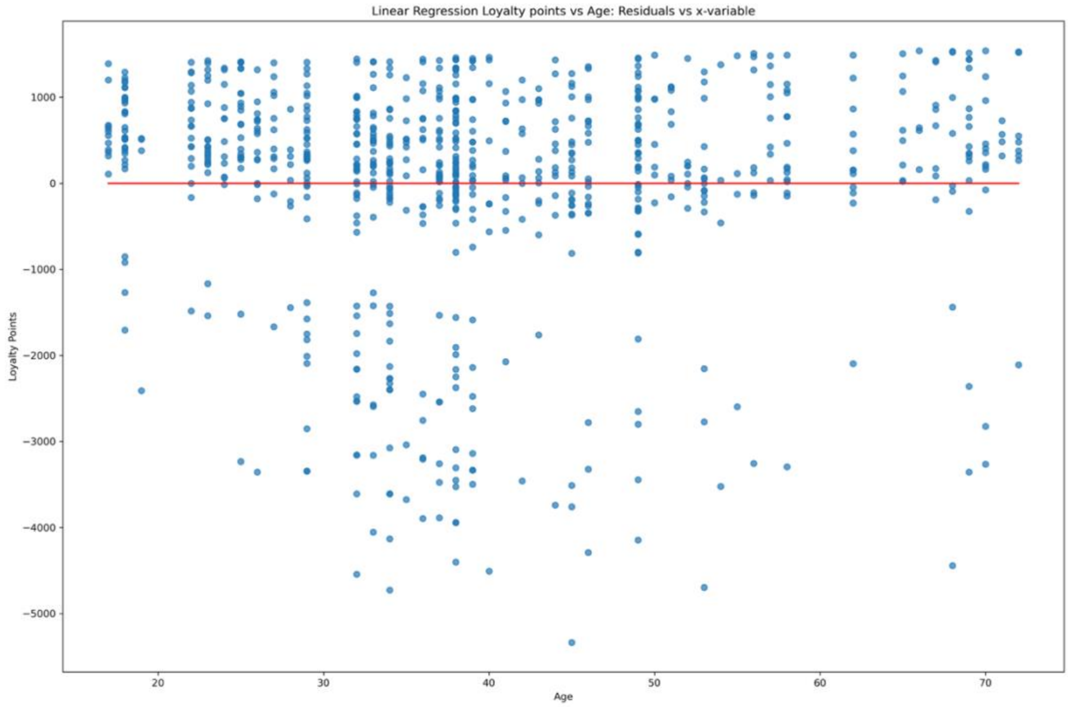
OLS Regression Results

Dep. Variable:	y_ag	**R-squared:**	0.001
Model:	OLS	**Adj. R-squared:**	-0.000
Method:	Least Squares	**F-statistic:**	0.6522
Date:	Fri, 18 Jul 2025	**Prob (F-statistic):**	0.420
Time:	19:05:12	**Log-Likelihood:**	-6723.7
No. Observations:	782	**AIC:**	1.345e+04
Df Residuals:	780	**BIC:**	1.346e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1387.5427	143.917	9.641	0.000	1105.033	1670.052
x_ag	2.7774	3.439	0.808	0.420	-3.974	9.529

Omnibus:	236.296	**Durbin-Watson:**	1.607
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	531.155
Skew:	1.659	**Prob(JB):**	4.58e-116
Kurtosis:	5.302	**Cond. No.**	128.




Linear Regression Loyalty points vs Age: Residuals vs x-variable
Loyalty Points
Age
1000
0
−1000
−2000
−3000
−4000
−5000
20
30
40
50
60
70

Variance Inflation Factor

In preparation for the multiple linear regression, the variance inflation factor was calculated for all three features to check the level of collinearity. Values for all 3 features and for the spending score vs annual remuneration are within acceptable limits <5, indicating no significant multicollinearity.

	VIF Factor	features
0	4.2	age
1	3.5	remuneration_k
2	3.3	spending_score

	VIF Factor	features
0	2.5	remuneration_k
1	2.5	spending_score

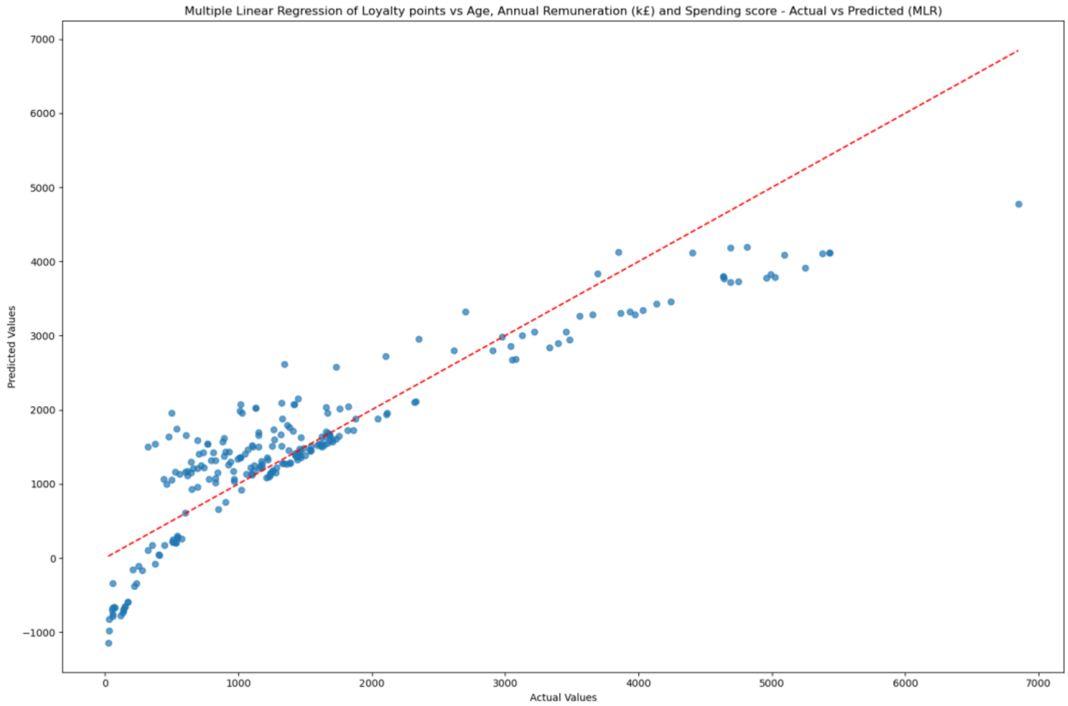
Multiple Linear Regression (MLR)

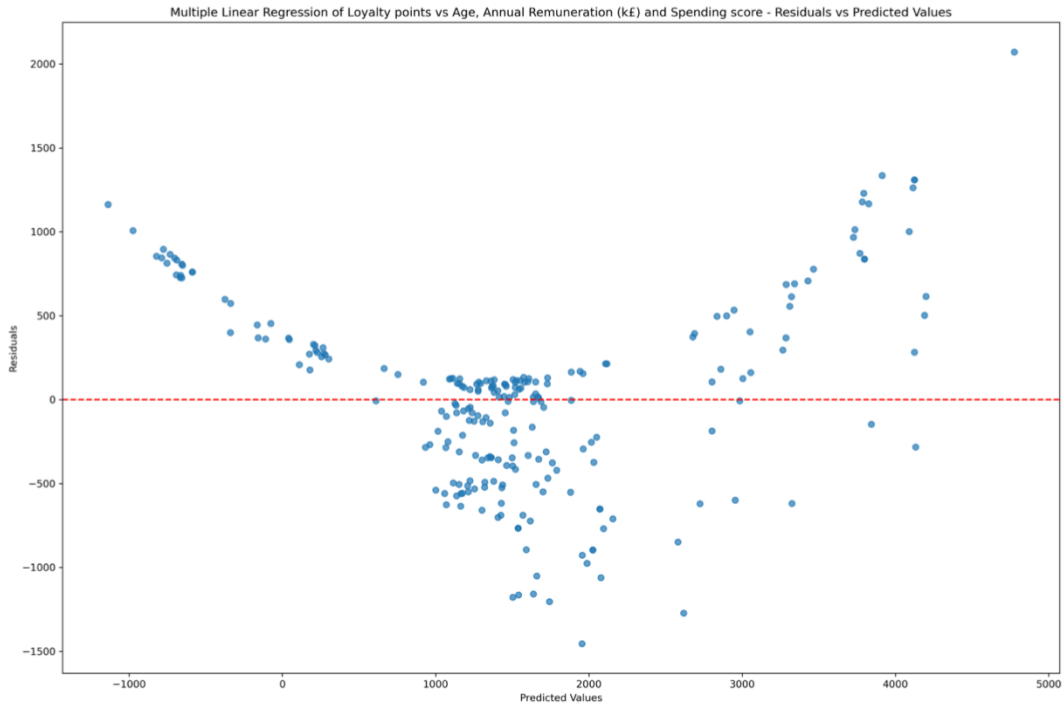
Five different multiple linear regression models were considered. One set including the age as a feature and one excluding it. Another set transforms the loyalty points by taking the square root.

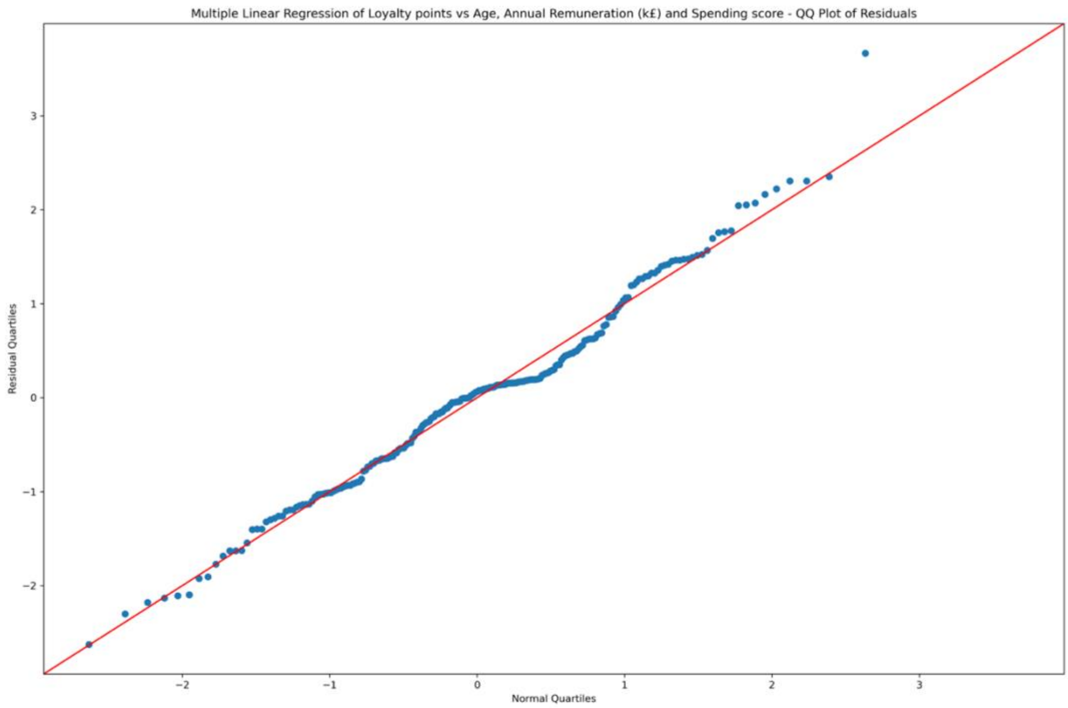
Since linear regression is also sensitive to unscaled data, the data was scaled in R, and a fifth model was built for the loyalty points vs Annual Remuneration and Spending Score. All models showed improvement in the R-squared values compared to the simple linear models.

They all, however, violated key linear regression assumptions, such as normality of the residuals and homoscedasticity, suggesting that the models' results are unreliable in predicting the loyalty points.

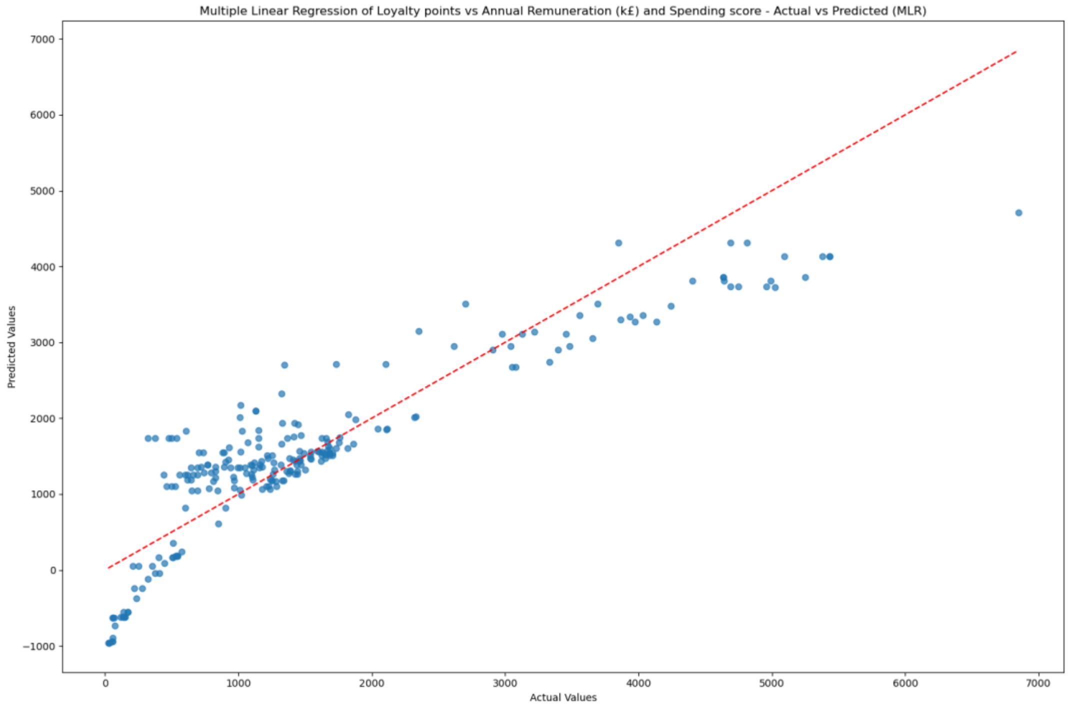
MLR (Age, Annual Remuneration, Spending Score)

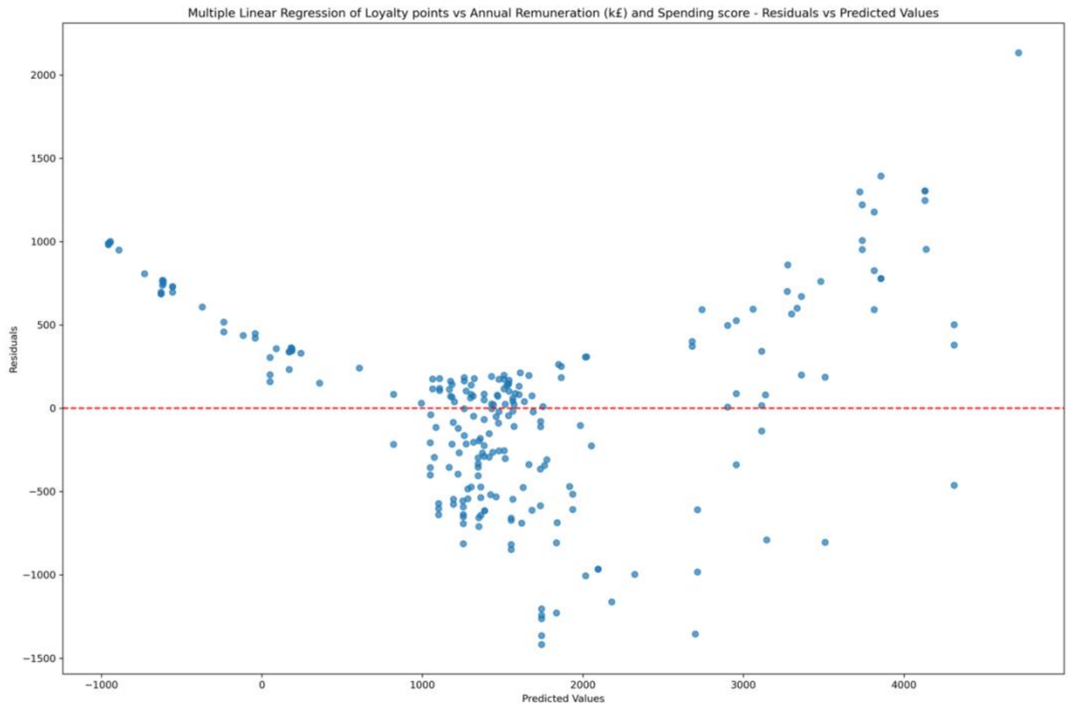


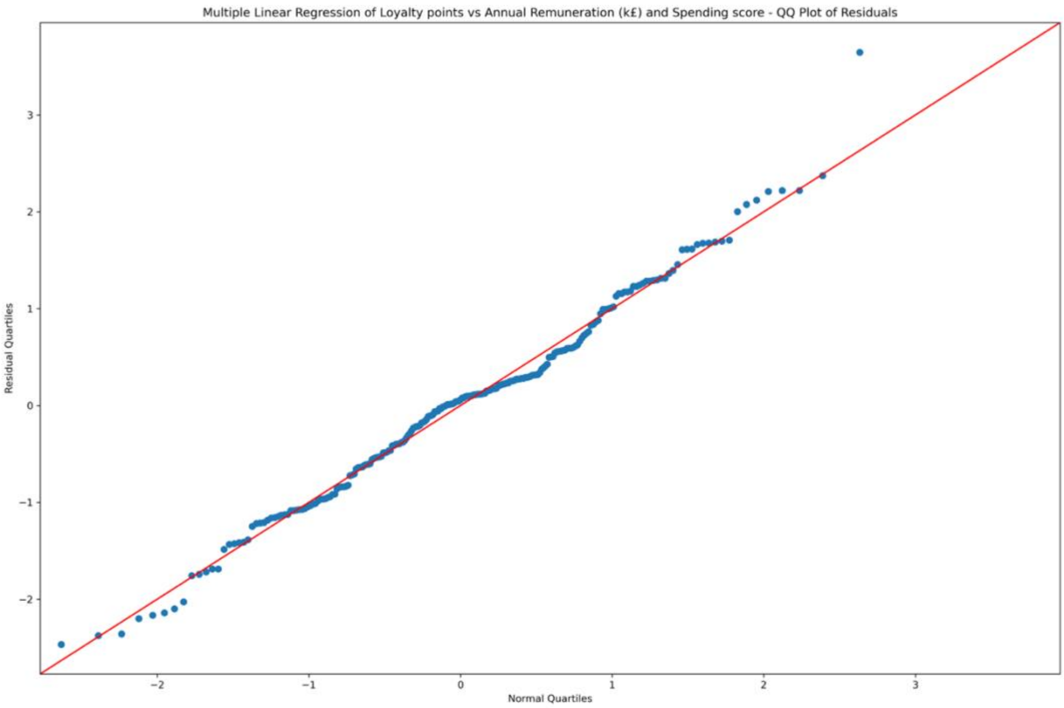

Multiple Linear Regression of Loyalty points vs Age, Annual Remuneration (k£) and Spending score - Residuals vs Predicted Values
Residuals
Predicted Values
2000
1500
1000
500
0
−500
−1000
−1500
−1000
0
1000
2000
3000
4000
5000


Multiple Linear Regression of Loyalty points vs Age, Annual Remuneration (k£) and Spending score - QQ Plot of Residuals
Residual Quartiles
Normal Quartiles
3
2
1
0
−1
−2
−2
−1
0
1
2
3

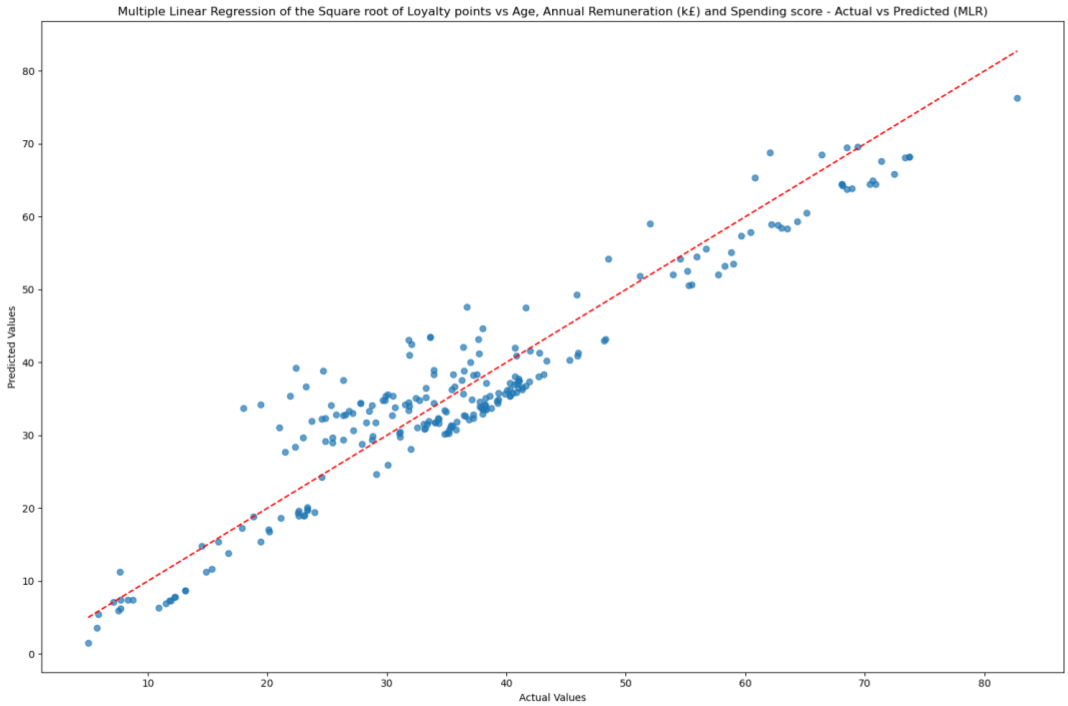
MLR (Annual Remuneration, Spending Score)

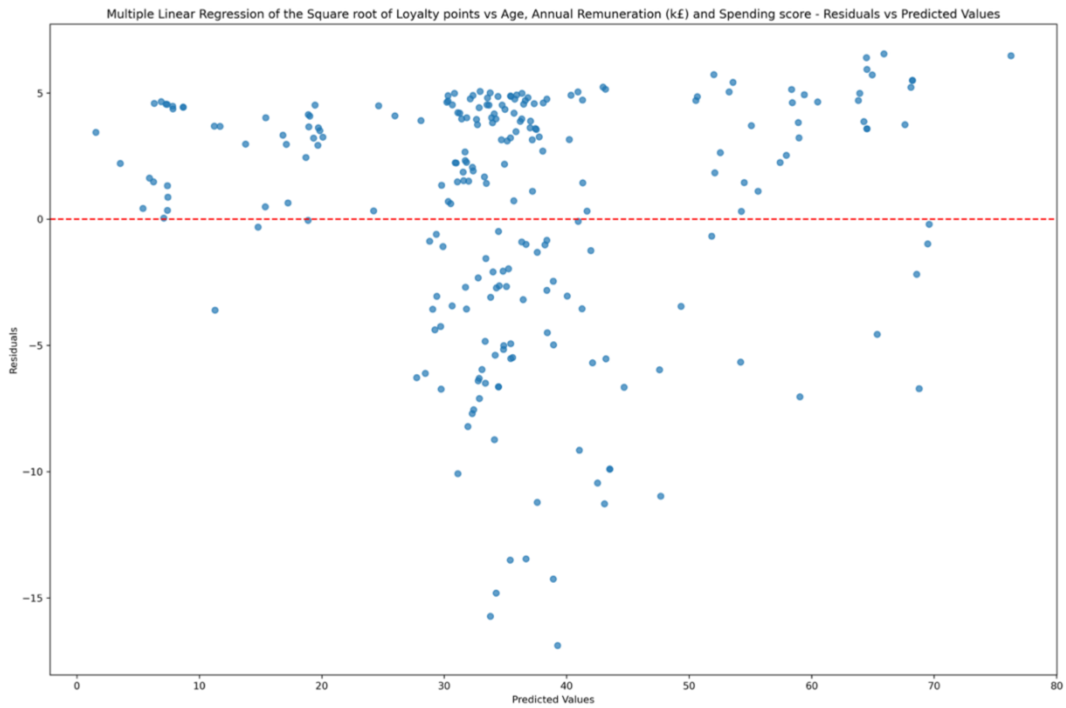


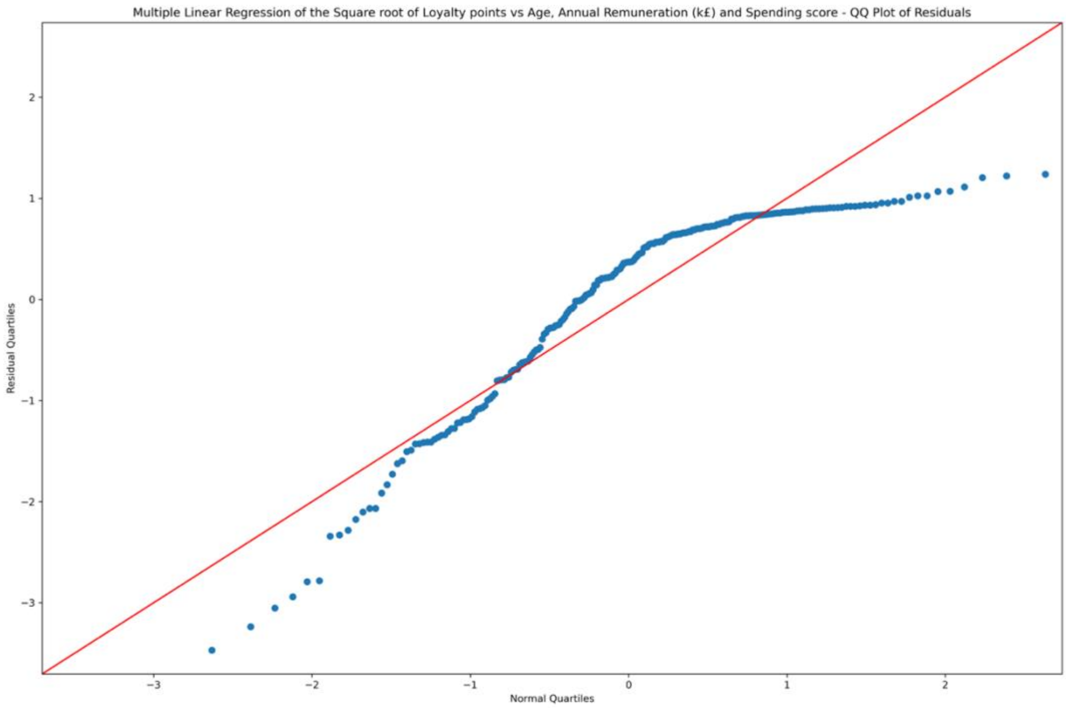



Multiple Linear Regression of Loyalty points vs Annual Remuneration (k£) and Spending score - QQ Plot of Residuals
Residual Quartiles
Normal Quartiles
3
2
1
0
−1
−2
−2
−1
0
1
2
3

MLR (Transformed Loyalty points (Square Root) vs Age, Annual Remuneration, Spending Score)

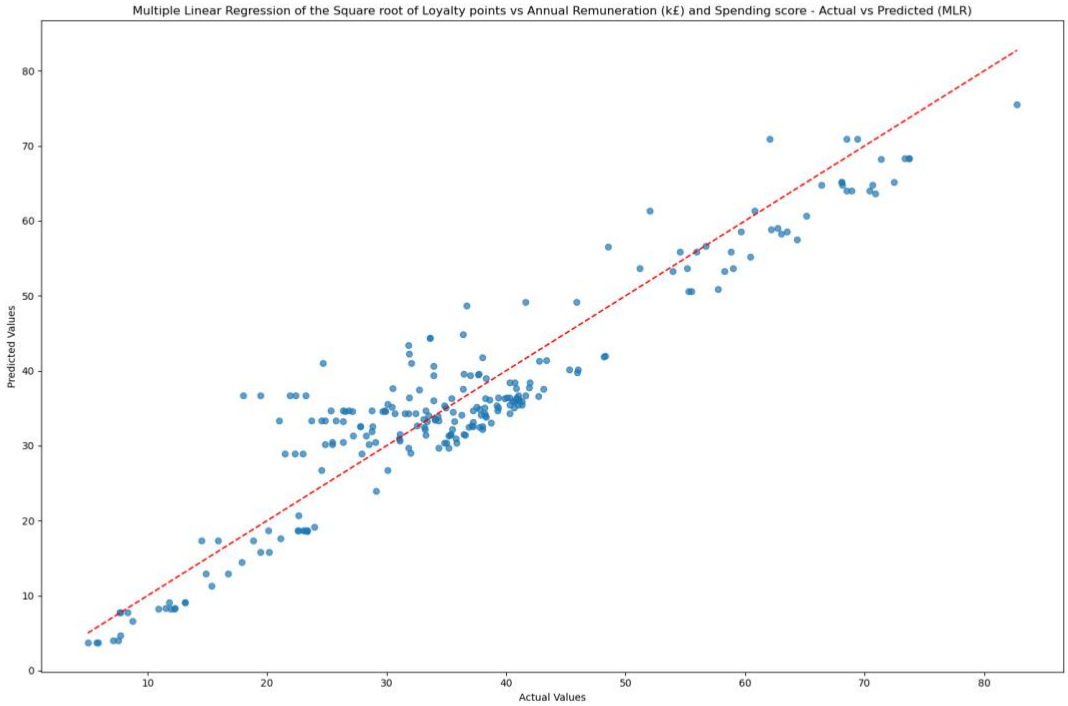


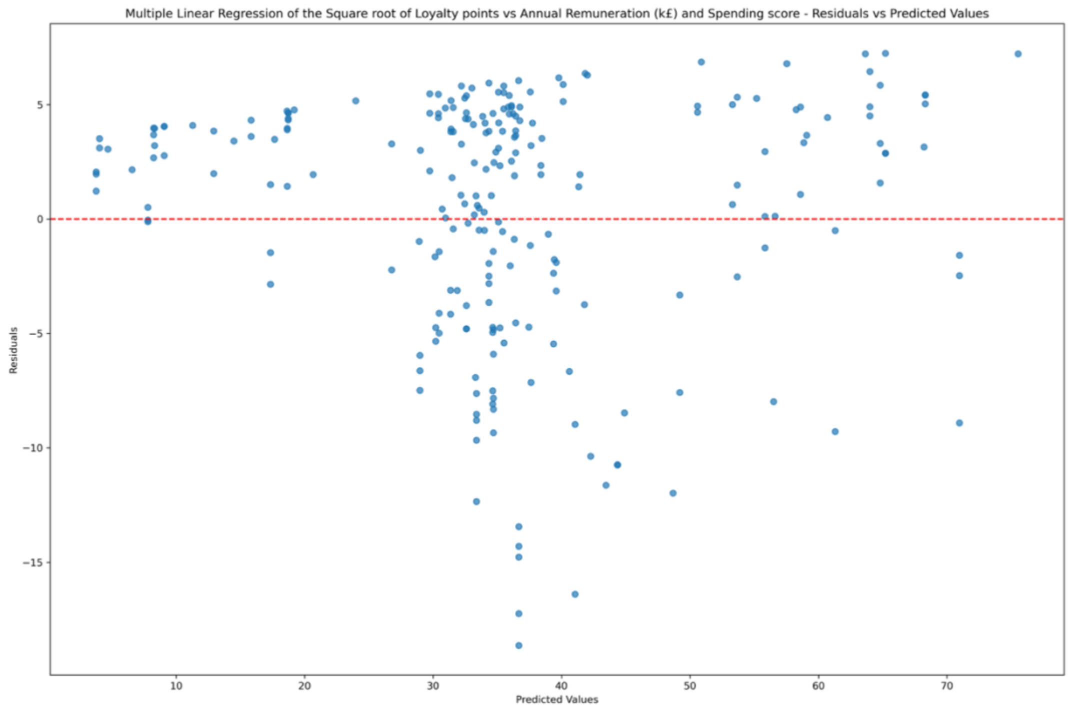


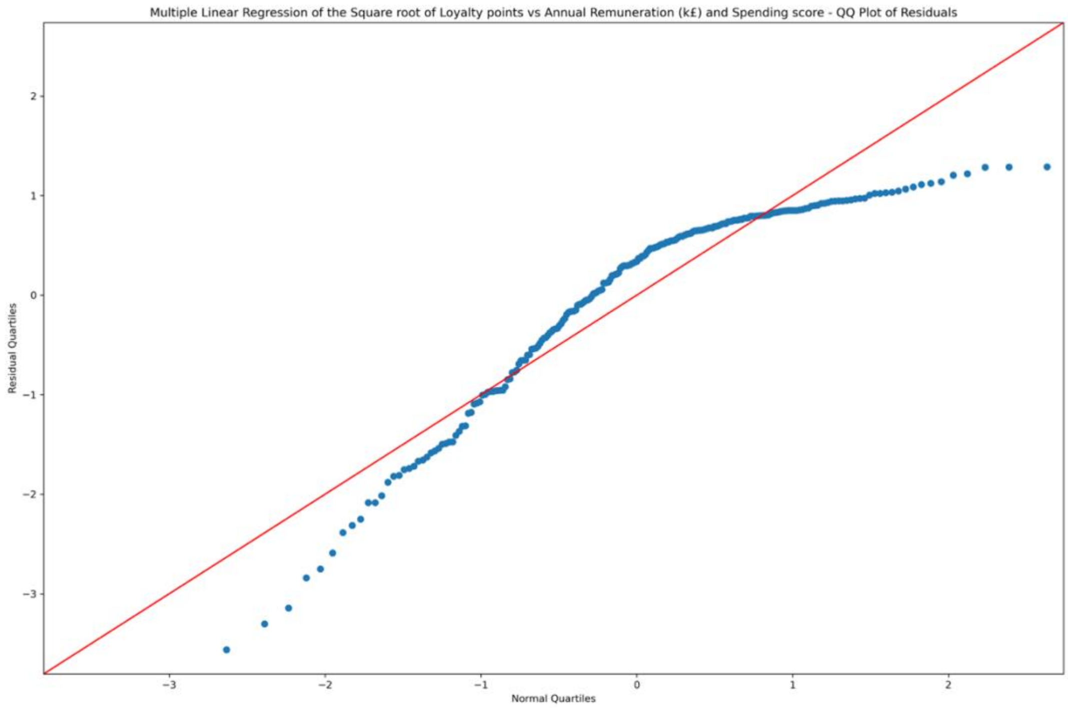


Multiple Linear Regression of the Square root of Loyalty points vs Age, Annual Remuneration (k£) and Spending score - QQ Plot of Residuals
Residual Quartiles
Normal Quartiles
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2

MLR (Transformed Loyalty points (Square Root) vs Annual Remuneration, Spending Score)

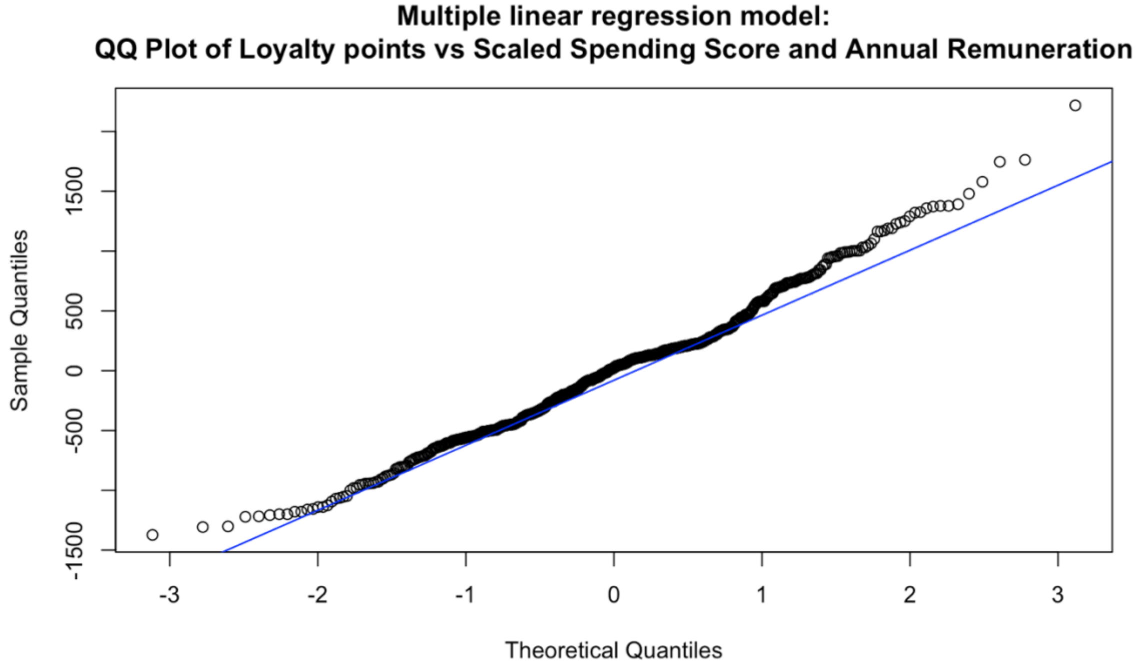


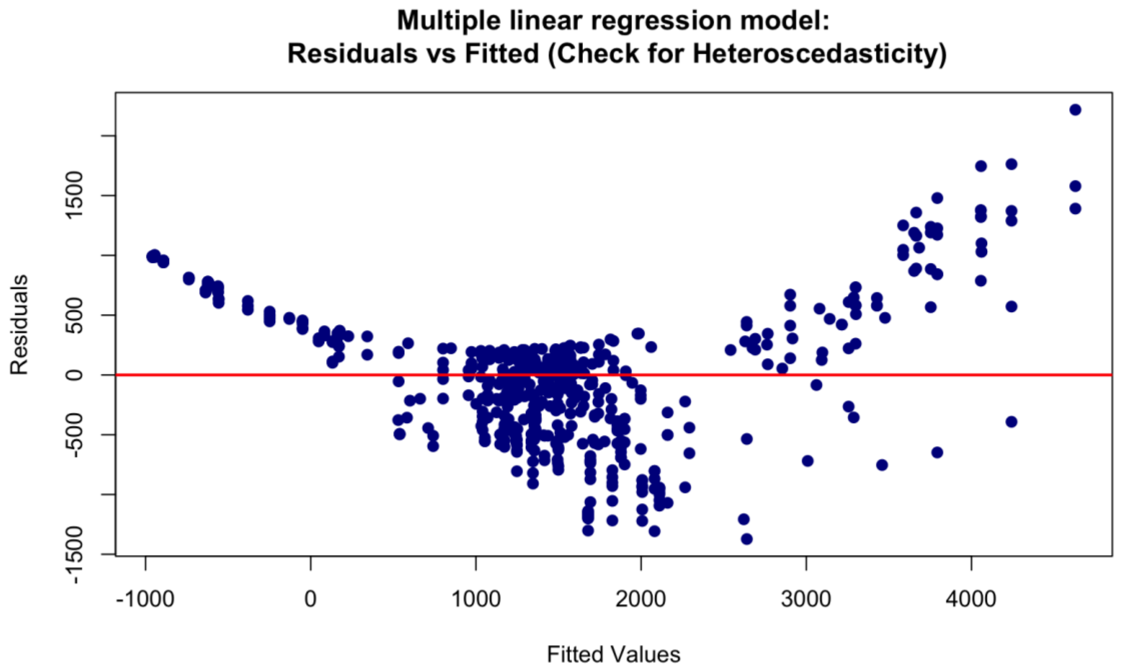




Multiple Linear Regression of the Square root of Loyalty points vs Annual Remuneration (k£) and Spending score - QQ Plot of Residuals
Residual Quartiles
Normal Quartiles
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2

Scaled Multiple Linear Regression Model





Decision Tree / Random Forest Regressor

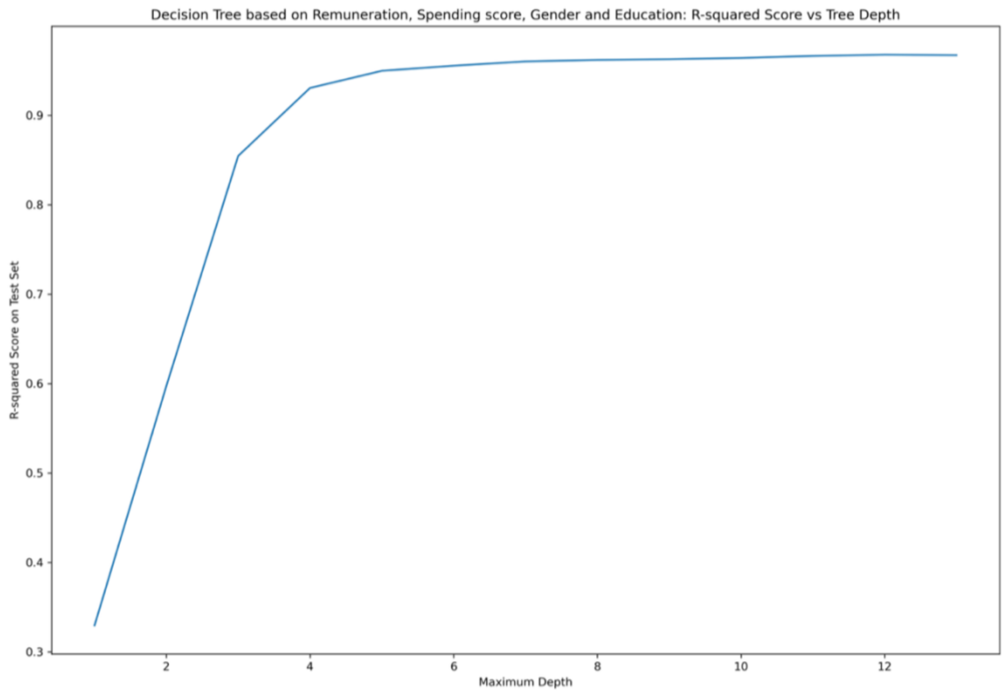
Decision Trees and Random Forests are stronger, more advanced models and are not sensitive to the same assumptions of the linear regression model.

To prep the data for them, dummy variables were created for the following categorical variables:

- Gender
- Education

The following sequence was followed:

- An original tree was created with no restrictions
- Results were evaluated, and over-fitting was suspected
- Feature importance was evaluated to determine the features to be included
- R-squared score vs Tree depth was plotted to identify the ideal tree depth for pruning
- The pruned tree was created based on a maximum depth of 9



- Features with low importance were removed, and the results of the pruned tree were re-evaluated
- Evaluation metrics were assessed for all the models. This includes R-squared, MAE, and RMSE

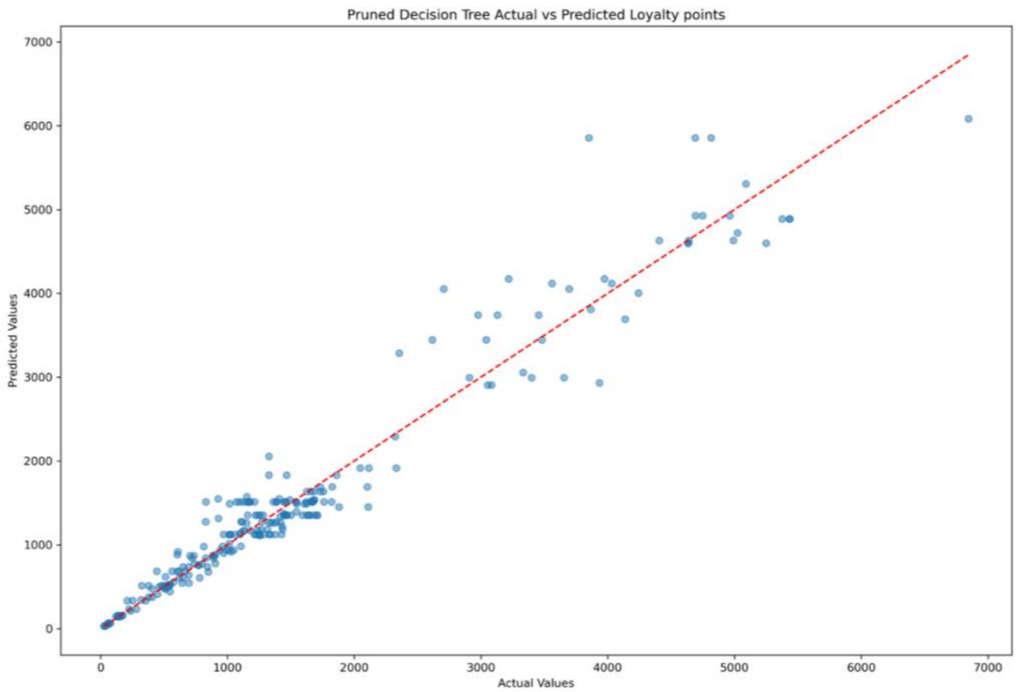
Pruned Tree Results – Before removing features with low importance

```
Depth = 9
Leaves = 231
Mean Absolute Error (MAE):137.99
Mean Squared Error (MSE):64730.04
R-squared: 96.29%
Avg % error: 2.02%
                feature  importance
          spending_score       0.575
          remuneration_k       0.407
                     age       0.017
                   count       0.001
             gender_Male       0.000
       education_Diploma       0.000
      education_Graduate       0.000
           education_Phd       0.000
  education_Postgraduate       0.000
```

Pruned Decision Tree Results – After removing features with low importance

```
Depth = 9
Leaves = 121
Mean Absolute Error (MAE):193.35
Mean Squared Error (MSE):101972.11
R-squared: 94.15%
Avg % error: 2.83%
       feature  importance
spending_score        0.59
remuneration_k        0.41
```

Pruned Decision Tree Predicted vs Actual Values chart after removing features with low



Future Models – To Investigate

Hybrid Model – Cluster as feature regression – in R

Since the data has 5 visible clusters, an experimental model was built in R to test the power of adding clusters as features, along with the spending score and remuneration, to a random forest regressor.

The model produced satisfactory results and is deemed to better model the different behavior of the subsets in the data.

The limitation is, however, that clustering is an unsupervised model and is modeled on the whole dataset.

For real-world applications, where predictions for new customers have to be made, the following steps:

- The data has to be split into train and test datasets using stratified sampling to maintain the distribution of the data (7)
- The classification model will need to be leveraged to predict the clusters of the new data
- The predicted clusters will then be used as features in the regression model along with spending score and remuneration.

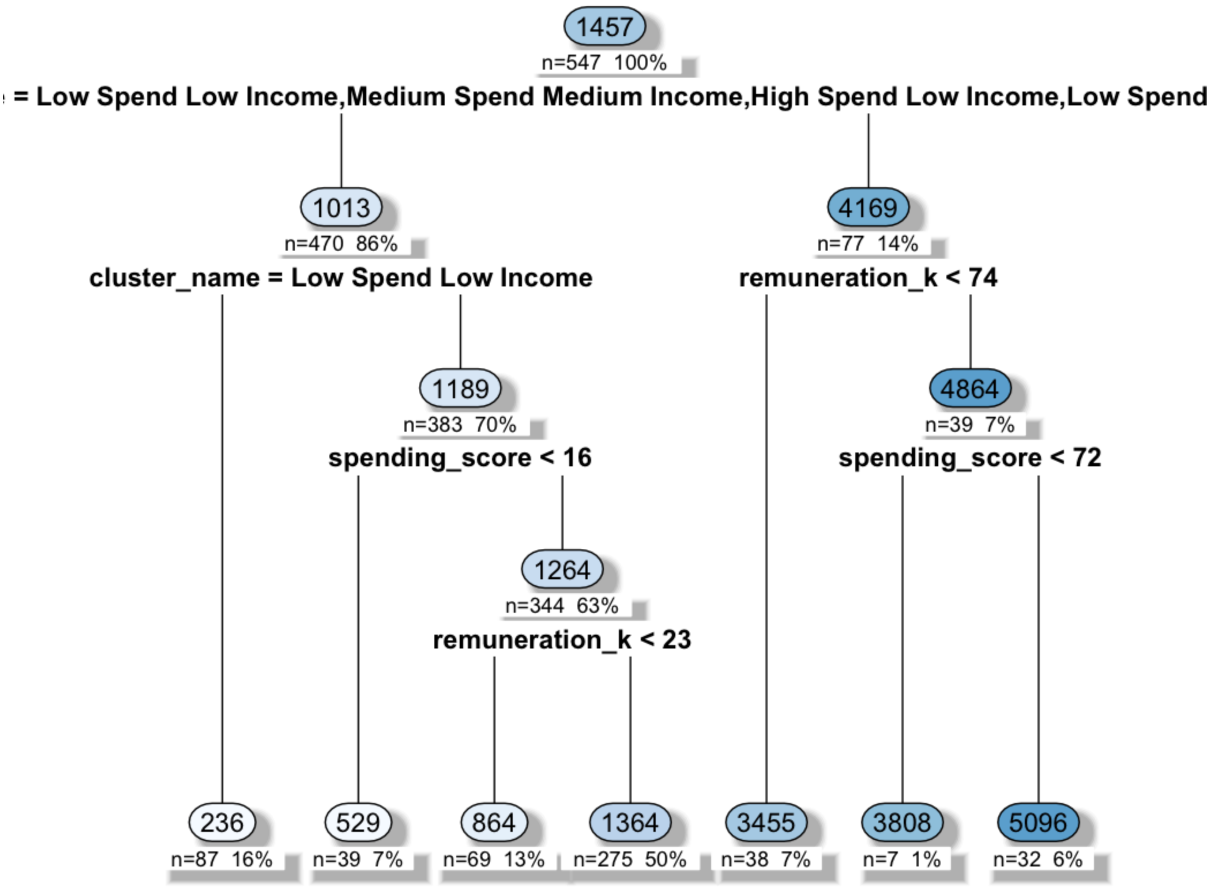
Given the time constraint, the model was considered but not built.

It can be done in the second phase of this project to explore the power of having such an advanced model.

A drawback of such a model is, however, the risk of error propagation from the clustering into the regression model.

It is, however, still interesting to investigate as it takes the different attributes of the clusters into account.

Cluster as Feature Regression



Natural Language Processing

Methodology

The logic behind the natural language processing followed the following sequence:

A. Product and Product Categories

First reviews and summaries were grouped by product number. This means that all reviews and summaries belonging to the same product number

were joined together. They were then tokenized, lemmatized, and stop words were removed along with extra keywords that were found to be redundant, and a document term matrix (DTM) was then built.

The most common categories were then collected and categories were assigned to the products based on the keywords in the reviews. While not 100% accurate, assigned categories still give a sense of the categories that dominated the reviews, and where Turtle games need to collect more reviews to have a more representative sample.

Polarity and subjectivity were then calculated per product, and the median of both scores was calculated to assign a polarity and subjectivity score per product category

Charts were then created to show the top and lowest 20 products by polarity score. This is particularly useful for Turtle Games to identify problematic products.

Plots were then created for the top and worst 20 products based on polarity scores.

B. High Loyalty Customers' Sentiment Analysis

This analysis is extremely valuable in identifying customers with negative sentiment to proactively address their concerns. It is also valuable to identify the customers with positive sentiment to further engage them through the company's marketing efforts.

C. Reviews and Summary Analysis

Other than the analysis per product, product category, and the sentiment analysis of high-loyalty customers, a global assessment of the reviews and summaries was also created.

Word Clouds:

- Word clouds were first plotted for summaries and reviews including the stop words

- Stop words were then removed, and updated word clouds were replotted

Top 15 Keywords

- Top 15 keywords were plotted for the reviews and the summaries on a horizontal bar plot

- While Word clouds are interesting representation of the most frequently used keywords, they may not be as clear as the bar plot with the top keywords.

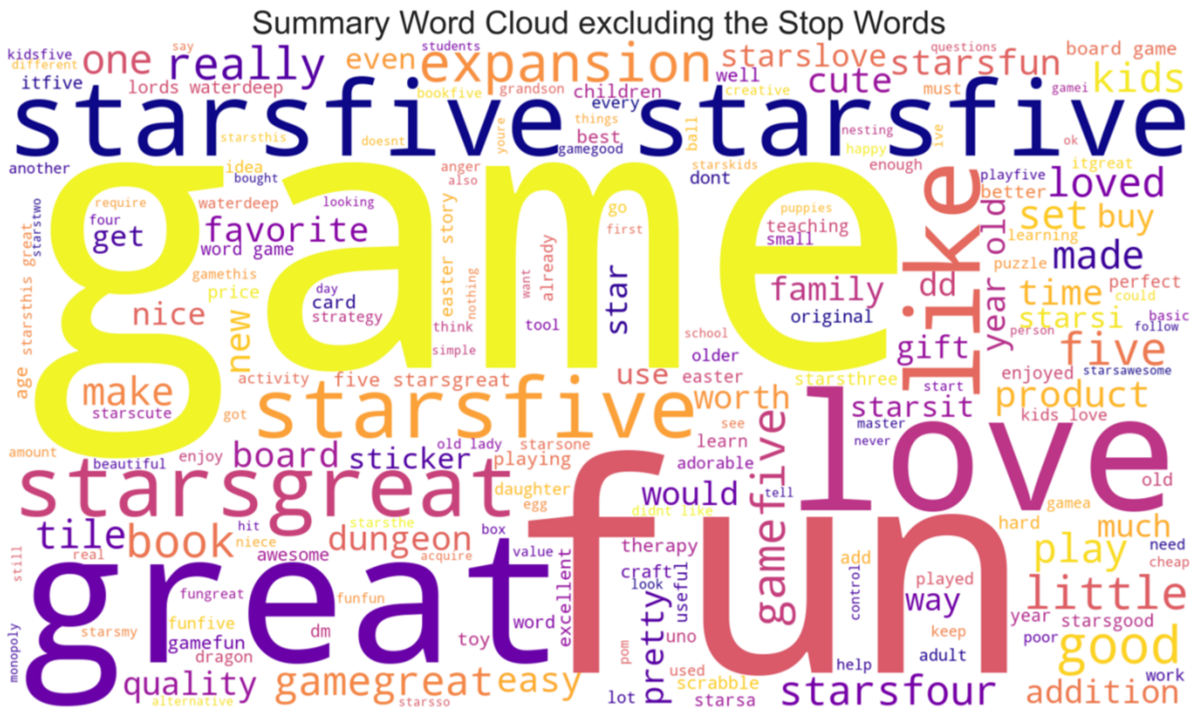
Top 20 Positive And Negative Summaries and Reviews

A list of the top 20 positive and negative summaries and reviews was created to carefully identify the main concerns in the negative reviews and maintain the positive points in the positive ones.

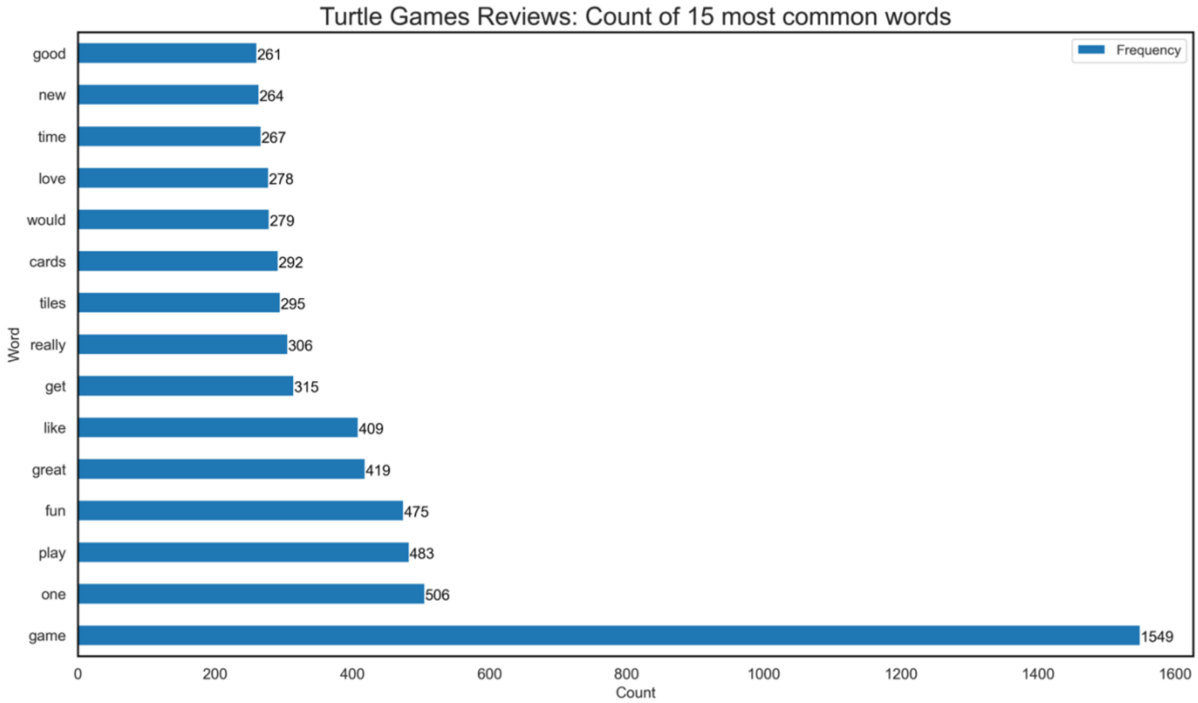
Summary Word Cloud

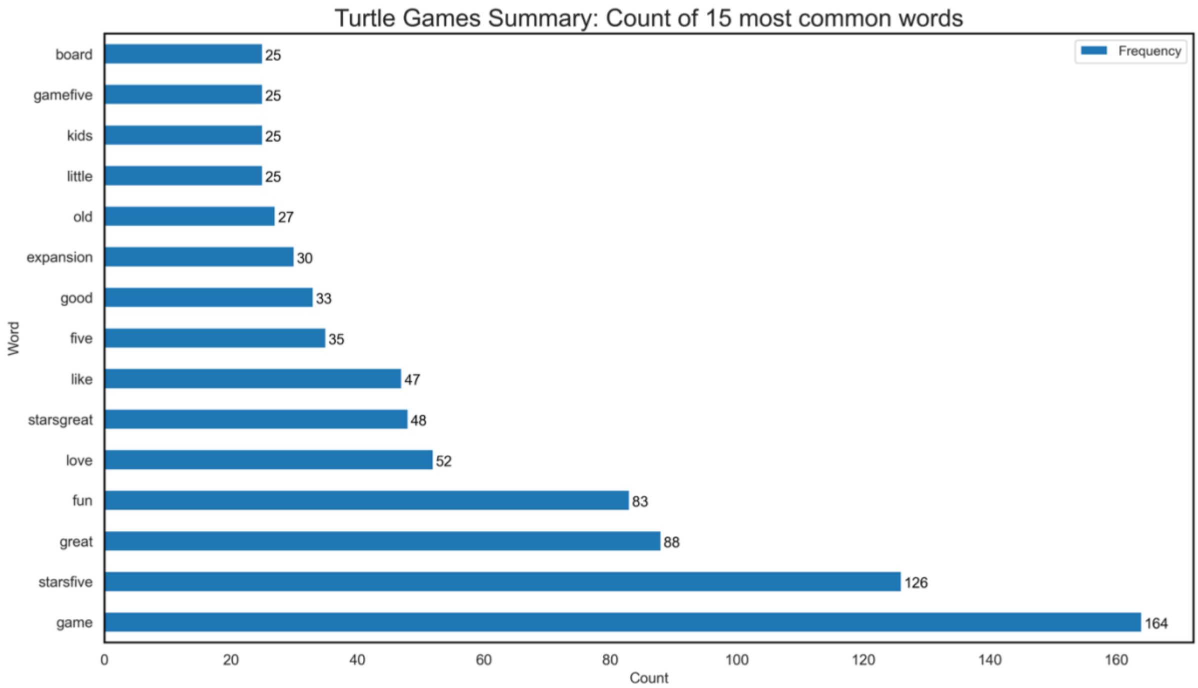
Duplicates were not removed in the case of the summaries as they didn't seem erroneous and was deemed required to see the pattern of common keywords.

Five Sars, fun love, and great are among the top keywords.



The most common keywords give a deeper insight into the pattern of keywords. Games is the most dominant keyword in both reviews and summaries, clearly identifying Turtle Games as a games producer and retailer.





Final Conclusions

- Overall, turtle games have positive sentiment reviews and summaries. This indicates the general satisfaction of the customers with the company's product portfolio. A proactive approach towards negative sentiment products and customer reviews would help the company's image and would deepen the customers' loyalty and engagement with the brand.

- The analysis of the reviews and summaries was, in general, insightful as it clarified the reason behind the presence of senior (age-wise) customers. Many of the reviews included the words gift, grandkids, granddaughter…etc clearly explaining the presence of this age bracket.
- Senior customers can be targeted with seasonal offers and discounts around holidays to allow them to keep buying gifts for their loved ones at Turtle Games.

review	polarity_rv
came in perfect condition	1.000000
awesome book	1.000000
awesome gift	1.000000
excellent activity for teaching selfmanagement skills	1.000000
perfect just what i ordered	1.000000
wonderful product	1.000000
delightful product	1.000000
wonderful for my grandson to learn the resurrection story	1.000000
perfect	1.000000
awesome	1.000000
awesome	1.000000
awesome set	1.000000
best set buy 2 if you have the means	1.000000
awesome addition to my rpg gm system	1.000000
its awesome	1.000000
one of the best board games i played in along time	1.000000
my daughter loves her stickers awesome seller thank you	1.000000
this was perfect to go with the 7 bean bags i just wish they were not separate orders	1.000000
awesome toy	1.000000
it is the best thing to play with and also mind blowing in some ways	1.000000

summary	polarity_sm
best gm screen ever	1.000000
wonderful designs	1.000000
perfect	1.000000
theyre the perfect size to keep in the car or a diaper	1.000000
perfect for preschooler	1.000000
awesome sticker activity for the price	1.000000
awesome book	1.000000
he was very happy with his gift	1.000000
awesome	1.000000
awesome and welldesigned for 9 year olds	1.000000
perfect	1.000000
excellent	1.000000
excellent	1.000000
excellent therapy tool	1.000000
the pigeon is the perfect addition to a school library	1.000000
best easter teaching tool	1.000000
wonderful	1.000000
all f the mudpuppy toys are wonderful	1.000000
awesome puzzle	1.000000
not the best quality	1.000000

Works Cited

1. ***Scikit Learn*. [Online] https://scikit-learn.org/stable/modules/ensemble.html#random-forests.**

2. *helika.io*. [Online] Helika. https://www.helika.io/top-3-best-customer-loyalty-programs-for-games/.

3. Perkins, Carina. Emarketer. [Online] June 2023. https://www.emarketer.com/content/uk-loyalty-programs-2023.

4. Emarsys. *emasys.com*. [Online] October 2024. https://emarsys.com/learn/blog/customer-loyalty-statistics/.

5. Zinrelo. *zinrelo.com*. [Online] Zinrelo, July 2023. https://www.zinrelo.com/blog/customer-loyalty-program-metrics/.

6. Scikit Leanr. *scikit-learn.org/*. [Online] 2025. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html.

7. Ed Crouch, Ben Eppler, Lauren Taylor, Christina Mühlenbein, and Elizabeth Hearne. Boston Consulting Group. *bcg.com*. [Online] December 2024 . https://www.bcg.com/publications/2024/loyalty-programs-customer-expectations-growing.

8. —. BCG. *bcg.com*. [Online] Boston Consulting Group, December 2024. https://www.bcg.com/publications/2024/loyalty-programs-customer-expectations-growing.

9. Reinchheld, Fred. Bain. *bain.com*. [Online] Bain , September 2001. https://media.bain.com/Images/BB_Prescription_cutting_costs.pdf.

Cover Image

Karl Ibri [Online] Unsplash.

https://unsplash.com/photos/grayscale-photography-of-turtle-CQqZ2BWeoAU?utm_content=creditShareLink&utm_medium=referral&utm_source=unsplash