ABIOYE_KOLAWOLE DA301_Assignment_Report.pdf.

TUTLE GAMES DATA REPORT

Tutle game sales analysis to show sales performance across various region, relationships between sales, factors affecting the performance of sales, trends, seasonality, customer behaviours to our products, and recommendations.

Data source:

1. turtle_sales.csv
2. turtle_reviews.csv

PROBLEM STATEMENT

Tutle has a global customer base across North America and Europe. A business objective of improving overall sales performance by utilising customer trends and understanding :

1. How customers accumulate loyalty points
2. How useful are remuneration and spending scores data
3. Can social data be used in marketing
4. Impact on sales per product
5. The reliability of the data (e.g., normal distribution, Skewness, Kurtosis)
6. Relationship(s) in sales between North America, Europe, and global sales

DATA DESCRIPTION.

Top six renumeration:

age	remuneration (k£)	spending_score (1-100)
34	65.995091	61.972727
55	63.386000	34.200000
36	61.401600	43.200000
35	61.172000	52.900000

age	remuneration (k£)	spending_score (1-100)
42	61.090000	22.600000
32	59.964364	69.763636

Top six spend:

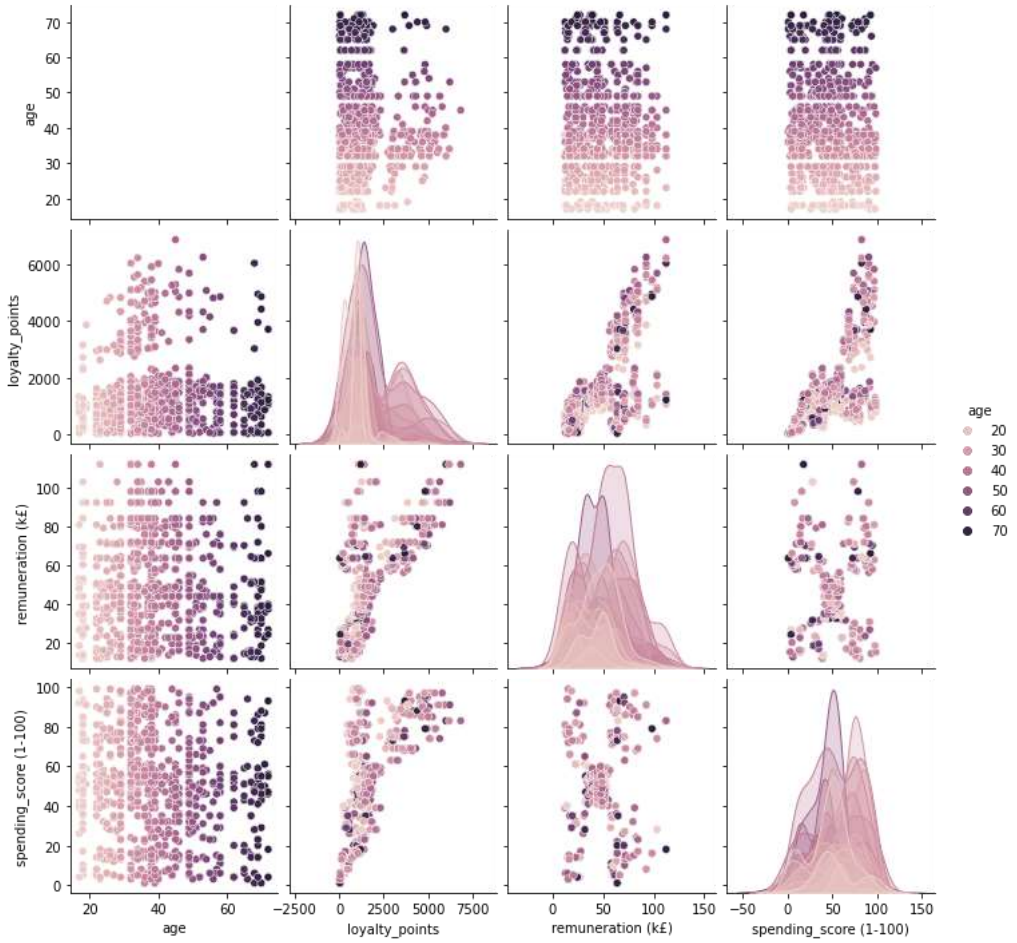
age	remuneration (k£)	spending_score (1-100)
32	59.964364	69.763636
26	33.784000	67.250000
29	49.125455	63.336364
33	38.540000	62.837500
19	43.542000	62.800000
37	38.066222	62.511111

HOW CUSTOMERS ACCUMULATE LOYALTY POINTS

Distribution of loyalty points were analysed based on gender.

```
Female    1120
Male       880
Name: gender, dtype: int64
```

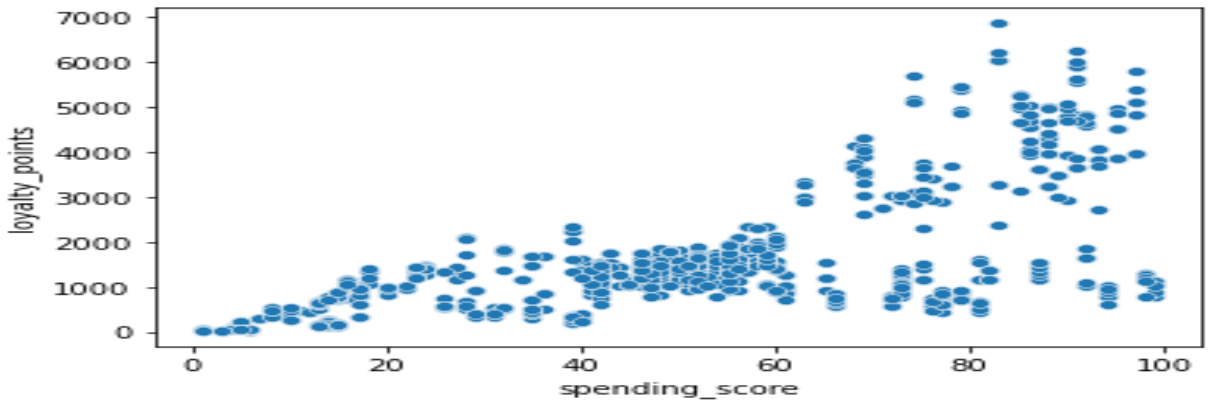
Further analysis was carried out on the existing data to know the age distribution on loyalty and spending power.



<u>The distribution of loyalty, spending score and renumeration based on age is high. A balanced spread within age 20 to mid 30s</u>.

Further analysis to compare the relationship between two variables such as SPENDING VS LOYALITY, RENUMERATION V LOYALITY, AGE V LOYALITY.

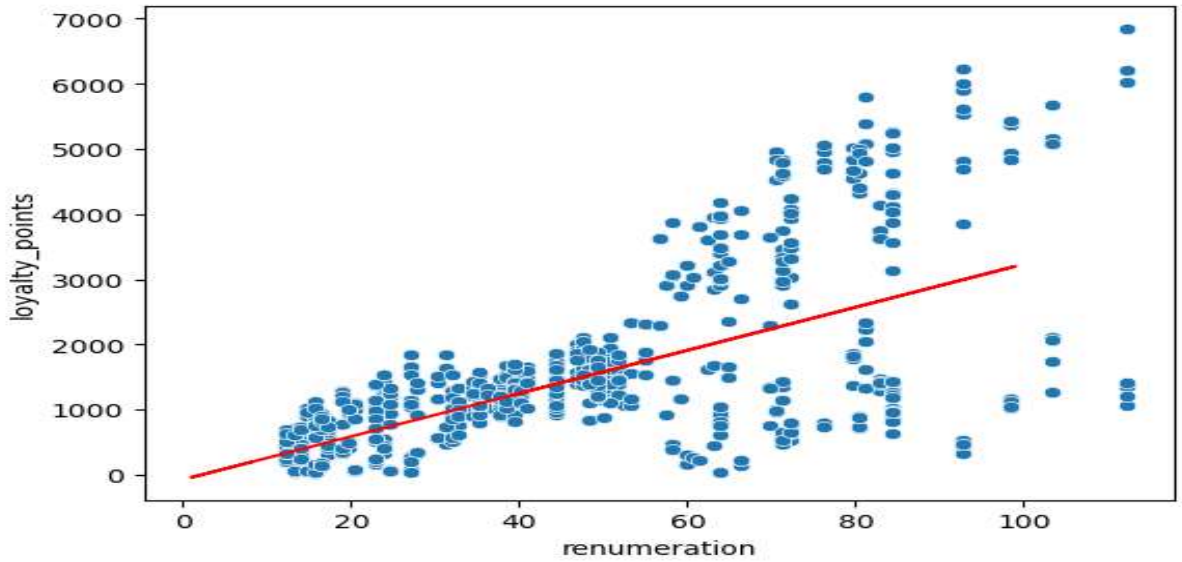
1. **<u>SPENDING v LOYALITY</u>**



OLS was done to compare the relationship between loyalty and spending.

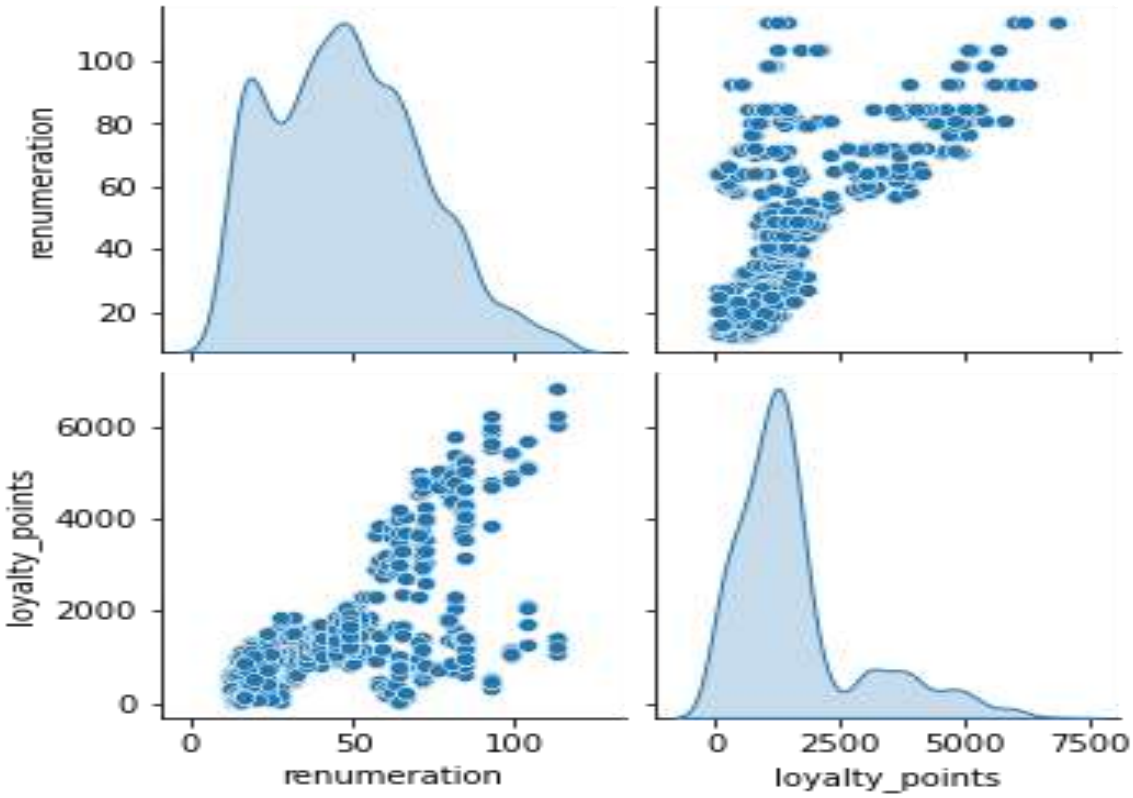
R-Squared is 45.2% which showed loyalty points has a moderate dependence on spending score but there is still a significant unexplained variance such as skewness and kurtosis.

2. RENUMERATION v LOYALITY

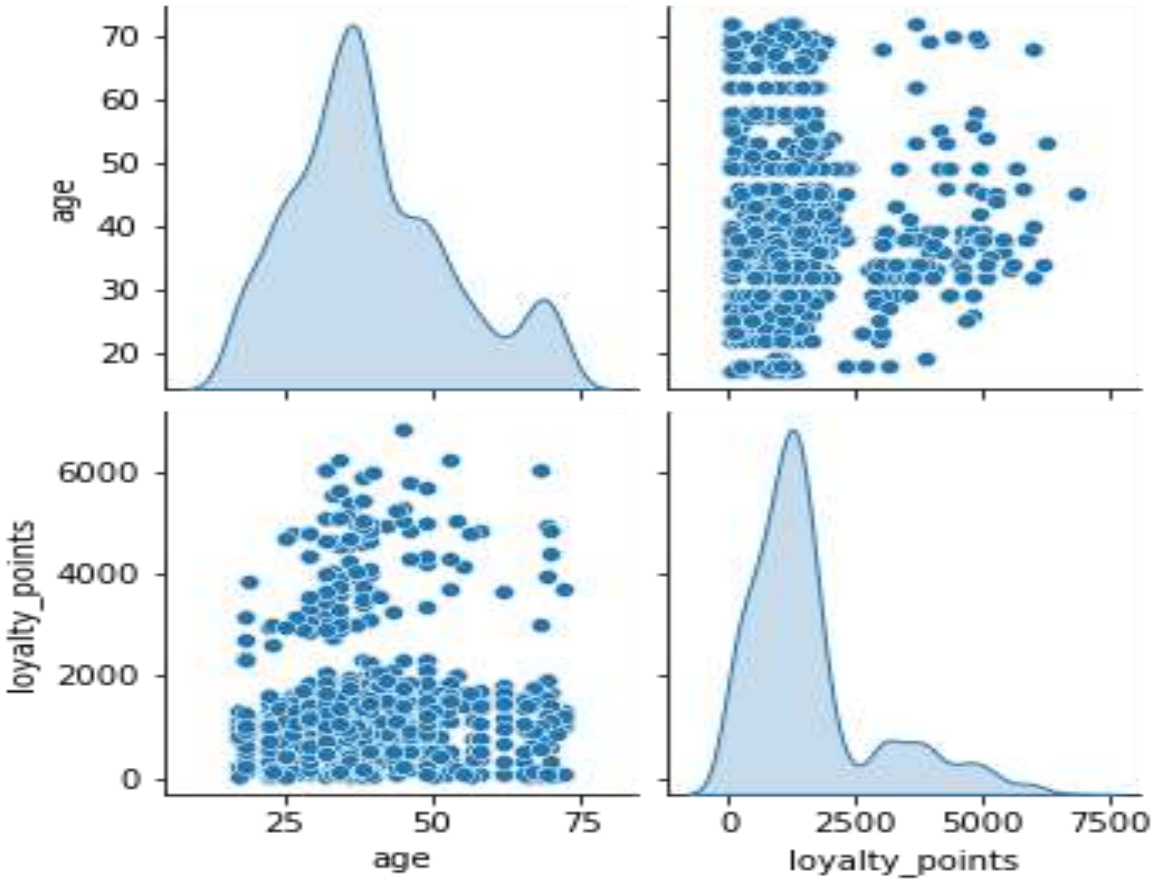


Renumeration is the total income of a customer in a year. Despite the R-Squared indicates 38%, the impact of customers with a higher salary to loyalty point is high.
The graph shows there is a considerably high uniform correlation between the two indicators.
More marketing and services need to be done to customers with a high net spend

Prediction between these two variables showed there is a slight split between level of loyalty and renumeration and loyalty as it progressed with a coefficient of -65.686513 and constant of 34.187825



3. **AGE v LOYALTY**

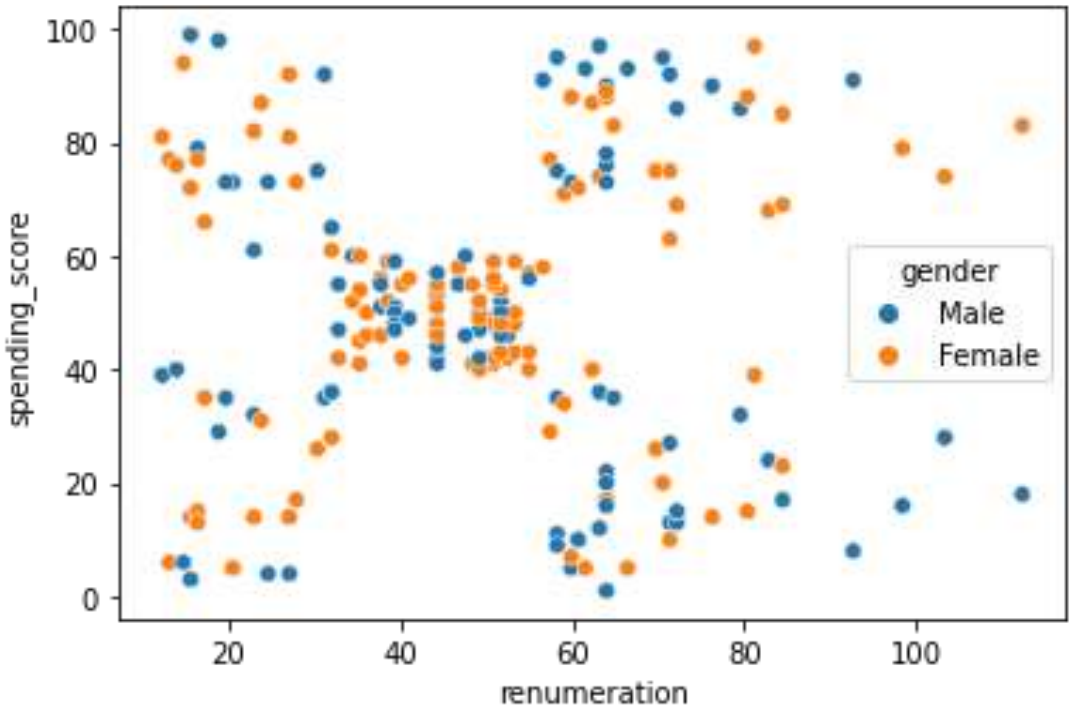


The bivariate distribution is highly skewed to the right. This indicates a poor relationship or unbalanced data between the two variables.

HOW USEFUL ARE RENUMERATION AND SPENDING SCORE DATA

renumeration	spending_score	
0	12.30	39
1	12.30	81
2	13.12	6

The clustered distribution between renumeration, spending score across gender:

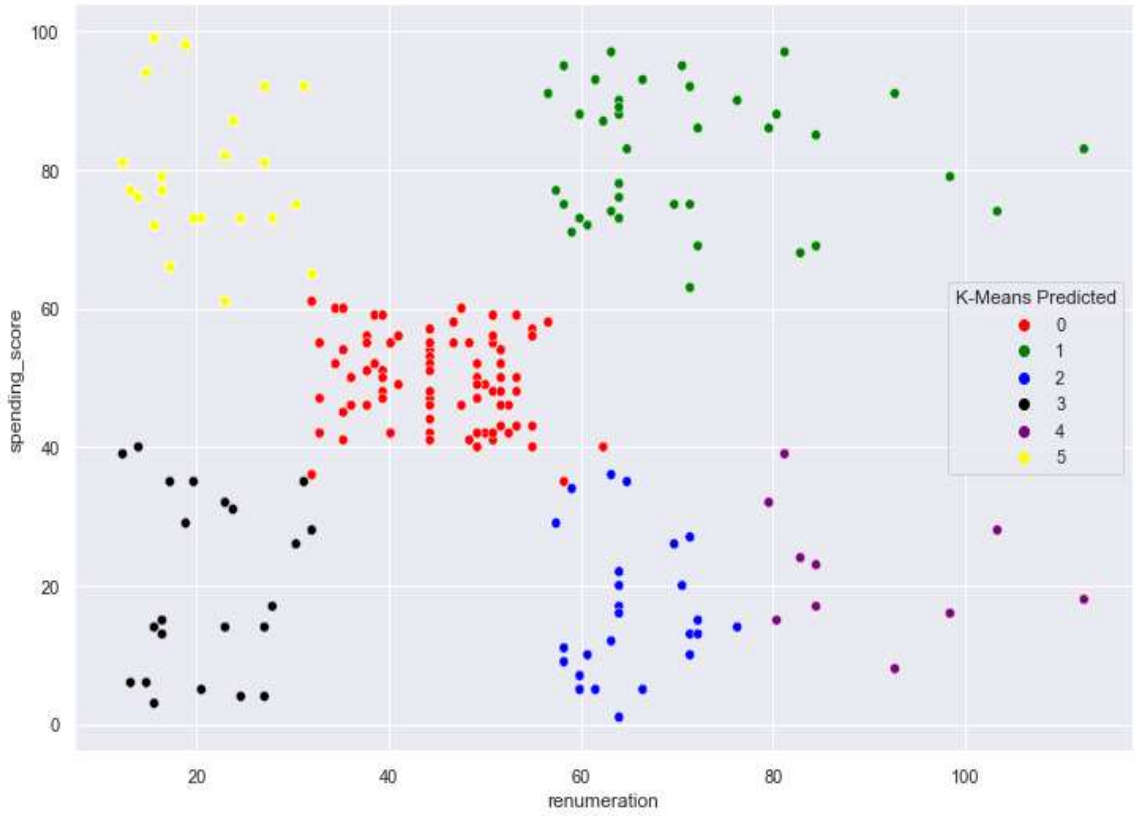


An elbow and Silhouette method was used to determine the best clustered fit to our data. A split between five and six clusters.

A prediction was carried out on renumeration and spending score data to analyse the impact they have on the business.
SVM model was used to analyse the prediction of the two variables with an accuracy level of 49%.

This is a low accuracy level, but it signifies the disparity of the spending power across gender.

Final predicted using six clustering:



This shows a high earn compared to a low earn with more distribution to the female gender.

predicted_High_Earn	predicted_low_Earn	
High Earn	325	0
Low Earn	275	0

Gender classification report:

```
precision    recall  f1-score   support

      Female       0.54      1.00      0.70       325
        Male       0.00      0.00      0.00       275

    accuracy                           0.54       600
   macro avg       0.27      0.50      0.35       600
weighted avg       0.29      0.54      0.38       600
```

conclusion: Other models will need to be analysed after refining the variables analysis. Other variable relationships need to be evaluated.

CAN SOCIAL DATA BE USED IN MARKETING CAMPAIGNS.

Analysis of positive and negative customer reviews on goods and services. As a result, we plan remedial solutions to displeased customers, manage our churn rate and improve on Net Promoter Score (NPS).

Sample Review:

	review	summary
0	When it comes to a DM's screen, the space on t...	The fact that 50% of this space is wasted on a...

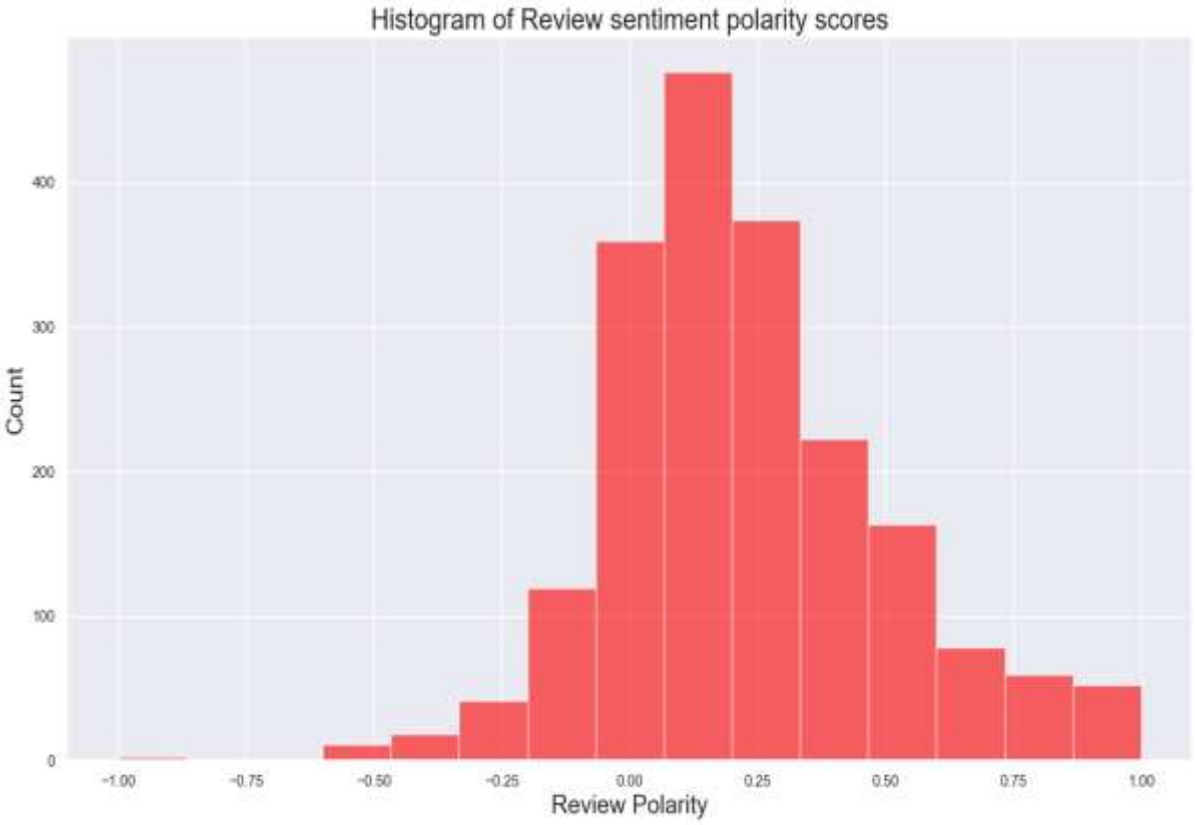
Further analysis showed high positive words such as:

15 most common review words:

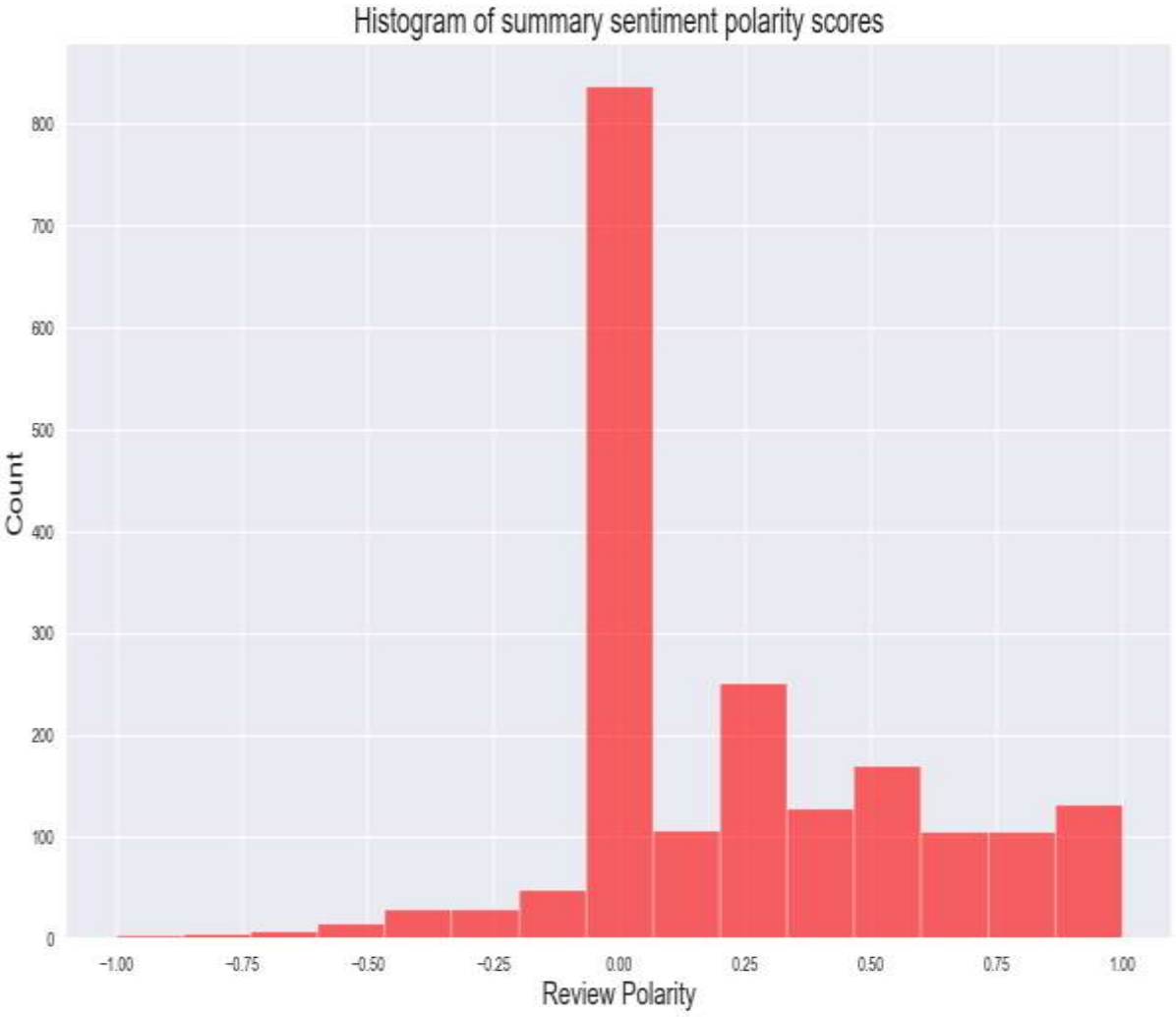
Word	Frequency
game	1693
great	587
fun	555
one	532
play	504
like	414
love	325
get	320

Word	Frequency
really	318
cards	301
would	301
tiles	298
time	292
good	290
book	278

We have a high neutral sensitive review and positive reviews compared to negative reviews.



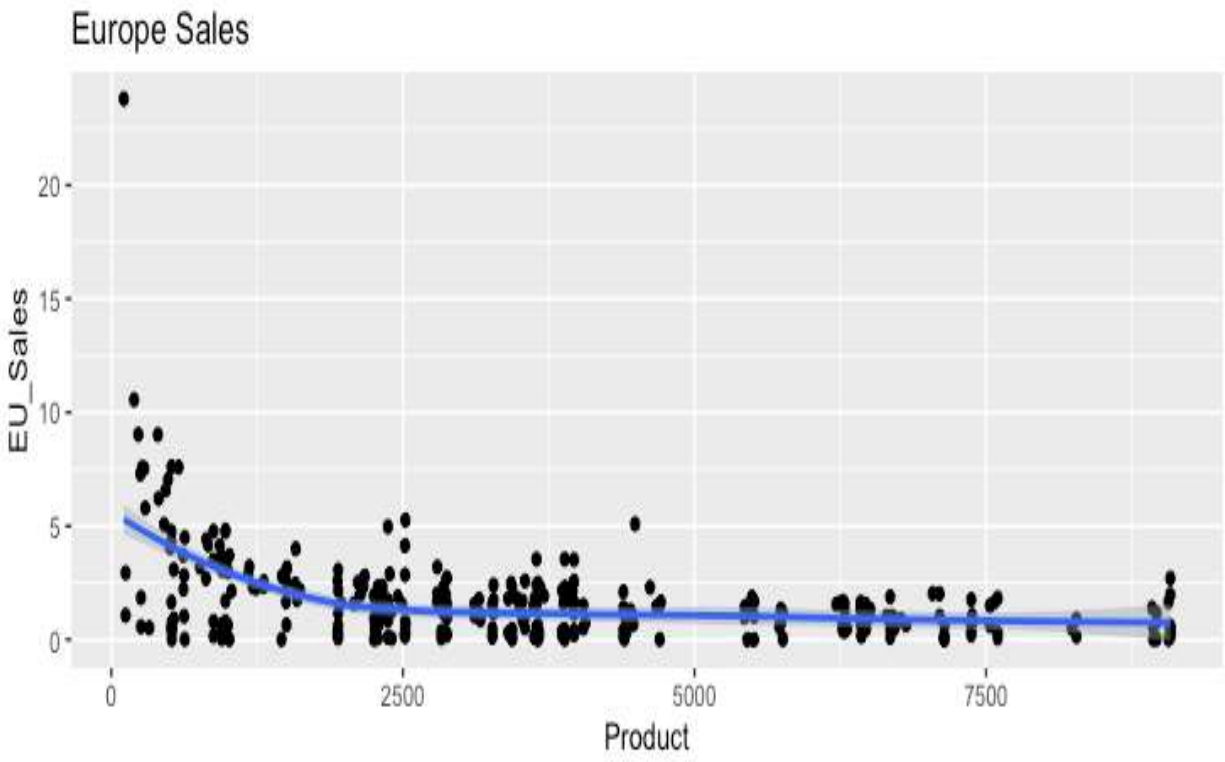
Review polarity shows how customers feel towards our products and services.



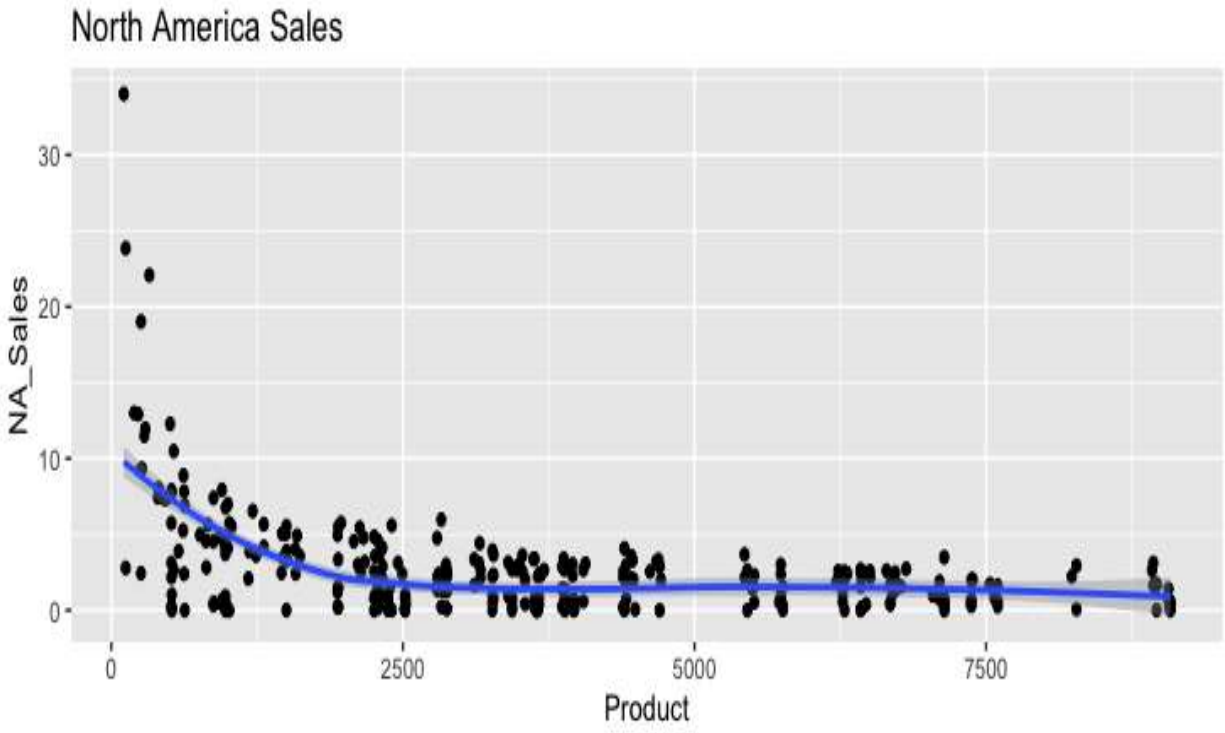
The estimated NPS score is 75%. This can be improved by further analysis into the data for trends and patterns.

IMPACT ON SALES PER PRODUCTS

A distribution of product count to sales for European region



North American sales:



There are a few spikes in sales performance across both regions, but sale performance is very similar. Further analysis needs to be carried out to fully understand the covariance between products, age, seasonality to name a few.

HOW RELIABLE IS THE DATA

This is to determine the percentage reliability of our dataset to determine accuracy and high percentage rate of predictions.

Data distribution:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.200	5.515	8.090	10.730	12.785	67.850

Shapiro normality was carried out on the datasets:

Shapiro-Wilk normality test

data: tut$EU_Sales
W = 0.64687, p-value < 2.2e-16

data: tut$NA_Sales
W = 0.6293, p-value < 2.2e-16
We have a low p value below 5 which shows an abnormal distribution.

Skewness:
A skewness value of 4.818688 and 4.30921 suggests a highly positive skew. This means that the distribution has an extended right tail with few extreme values to the left side.

Correlation:

Recommendations and conclusions:

- .There is a high correlation between sales in each region through the year.
- Other factors need to be considered such as political factor, demography and more to fully understand purchase trends.
- A more detailed data needs to be analysed so a time series analysis can be carried and a balanced data set for a better forecast.
- A more dataset needed to map reviews to customer’s account.