12/22/2022

Predicting future outcomes

Turtle Games

Rohit Abraham Francis Giles

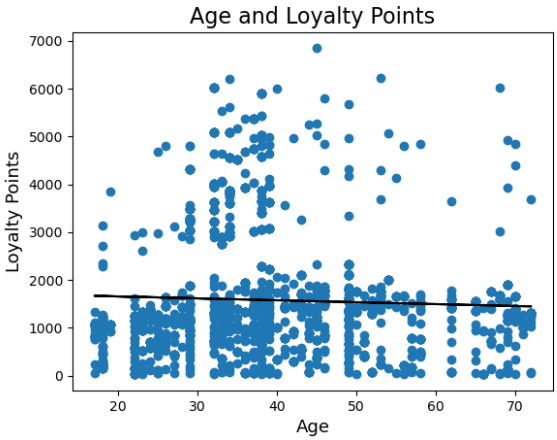
SCENARIO

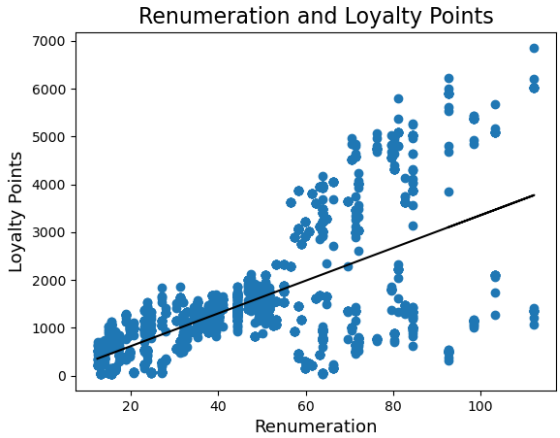
Turtle Games wants to improve overall sales performance by utilizing customer trends, so they have come up with a certain question to answer them:

- how customers accumulate loyalty points
- how groups within the customer base can be used to target specific market segments
- how social data (e.g., customer reviews) can be used to inform marketing campaigns
- the impact that each product has on sales
- how reliable the data is (e.g., normal distribution, skewness, or kurtosis)
- what the relationship(s) is/are (if any) between North American, European, and global sales?

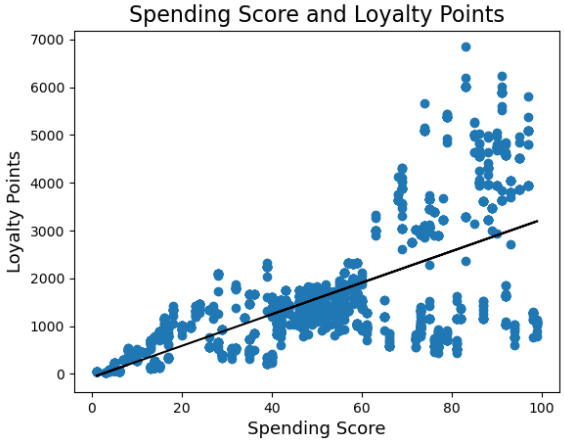
REGRESSION ANALYSIS

The turtle_reviews dataset was cleaned and explored initially. the dataset contains 2000 observations and no null values. The ‘language’ and ‘platform’ data were not considered for the regression analysis, which was used to find the relationship between the ‘loyalty points’, ‘age’, ‘remuneration’, and ‘spending scores’.





While running the linear regression function, it was found that there seems to be no relation between ‘age’ and ‘loyalty points’. If anything, there is a very small negative slope showing as age increases, the loyalty points decrease. Whereas there is comparatively a better relationship between‘ loyalty points’, ‘remuneration’, and ‘spending scores’.



To better understand the relationship between dependent and independent variables, a multiple linear regression was also done. Loyalty point was taken as the dependent variable, and the Remuneration and Spending scores were taken as the independent variable. Even though the model gave a good R^2 value of 82.8% and there was no multicollinearity, the MAE and MSE

values are very high, indicating that they are influenced heavily by the outliers. And since these values are not closer to zero, we can assume that the model is not making an accurate prediction.

To reduce the influence of the outliers, a square root transformation was performed. It improved the R^2 value to 87.3%, and again there was no multicollinearity. The MAE and MSE values were now much smaller. This confirms that the model can predict with 87.3% accuracy on how the Loyalty points change with the change in Remuneration and Spending scores.

OLS Regression Results

Dep. Variable:	loyalty_points	R-squared:	0.826
Model:	OLS	Adj. R-squared:	0.826
Method:	Least Squares	F-statistic:	3322.
Date:	Sat, 31 Dec 2022	Prob (F-statistic):	0.00
Time:	14:00:42	Log-Likelihood:	-10803.
No. Observations:	1400	AIC:	2.161e+04
Df Residuals:	1397	BIC:	2.163e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1700.2975	43.150	-39.404	0.000	-1784.944	-1615.651
renumeration	34.1030	0.628	54.315	0.000	32.871	35.335
spending_score	32.8335	0.549	59.790	0.000	31.756	33.911

Omnibus:	2.692	Durbin-Watson:	2.110
Prob(Omnibus):	0.260	Jarque-Bera (JB):	2.621
Skew:	0.074	Prob(JB):	0.270
Kurtosis:	3.151	Cond. No.	220.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 2: MLR without transformation

OLS Regression Results

Dep. Variable:	loyalty_points	R-squared:	0.895
Model:	OLS	Adj. R-squared:	0.895
Method:	Least Squares	F-statistic:	5945.
Date:	Sat, 31 Dec 2022	Prob (F-statistic):	0.00
Time:	14:13:15	Log-Likelihood:	-4253.7
No. Observations:	1400	AIC:	8513.
Df Residuals:	1397	BIC:	8529.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.8571	0.395	-9.758	0.000	-4.632	-3.082
renumeration	0.4020	0.006	69.269	0.000	0.391	0.413
spending_score	0.4255	0.005	81.841	0.000	0.415	0.436

Omnibus:	240.611	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	374.071
Skew:	-1.209	Prob(JB):	5.91e-82
Kurtosis:	3.755	Cond. No.	214.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 1: Figure 2: MLR with Square root transformation

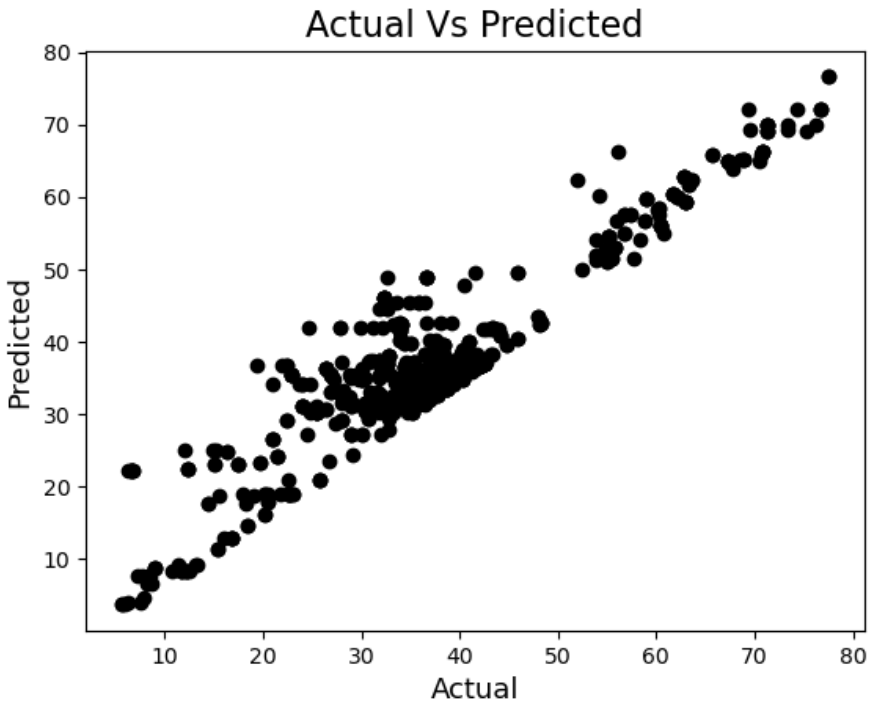
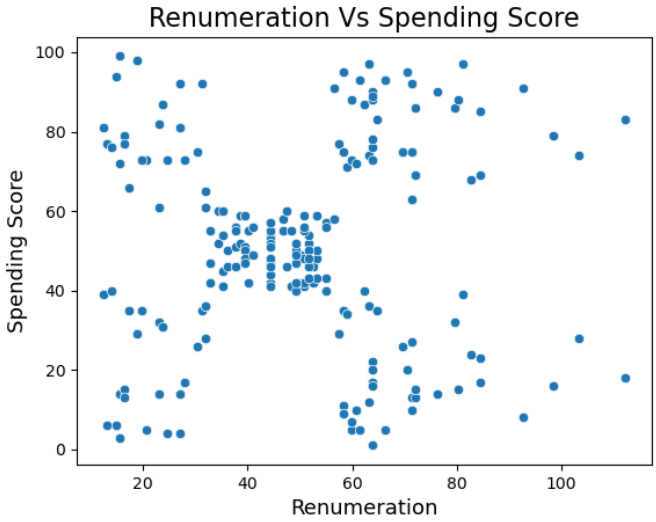
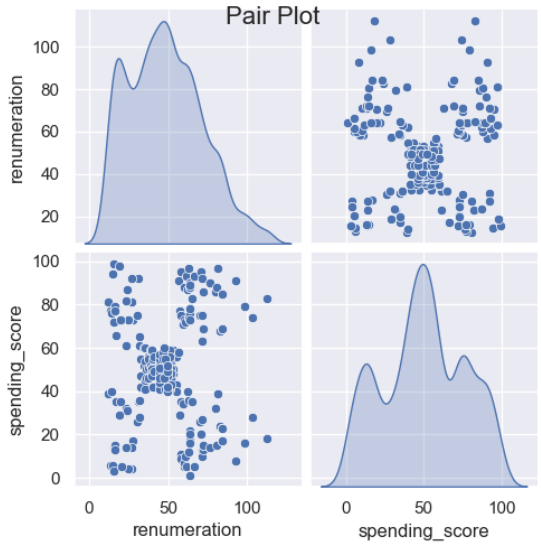


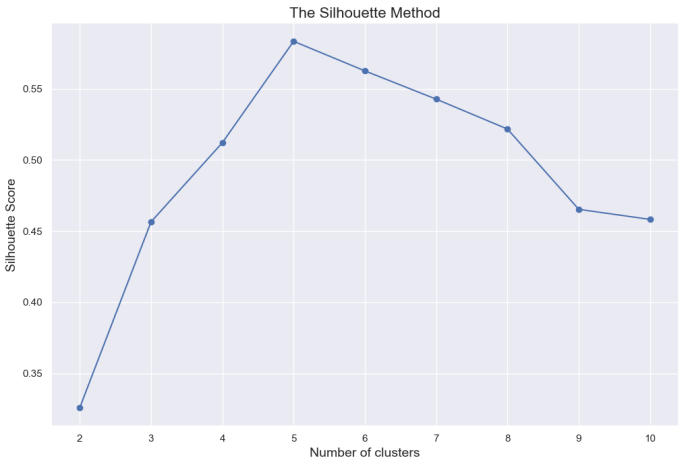
Figure 3: Fit of the model

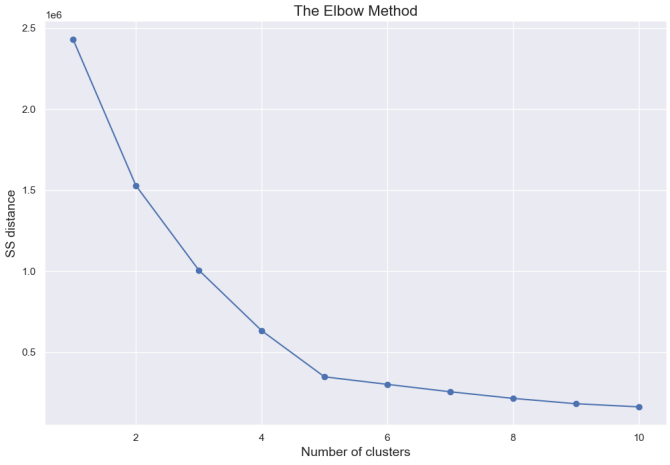
CLUSTER ANALYSIS





Cluster analysis was done to identify groups within the customer base. A cluster plot and pair plot were plotted against ‘Remuneration’ and ‘Spending scores’. We can see the formation of 5 clusters in the scatter plot, and there seems to be a relationship between the two variables. To further understand the clusters, K-values were determined to classify data points into groups (clusters) based on their similarities using the *Silhouette and Elbow* methods.





From the Silhouette method, we can see that the number of clusters peaks at 5 and then starts to reduce. And from the Elbow method, we can see the elbow formation starts at 5, hence K=5.

K-value was also checked if the clusters are balanced, so K-values of 3 & 4 were also checked. It was found that the K=5 was the most balanced, with the data points more or less evenly distributed. The only issue was that cluster 0 was larger, but when considering cluster 0 when K=3 & K =4, it can be assumed insignificant.

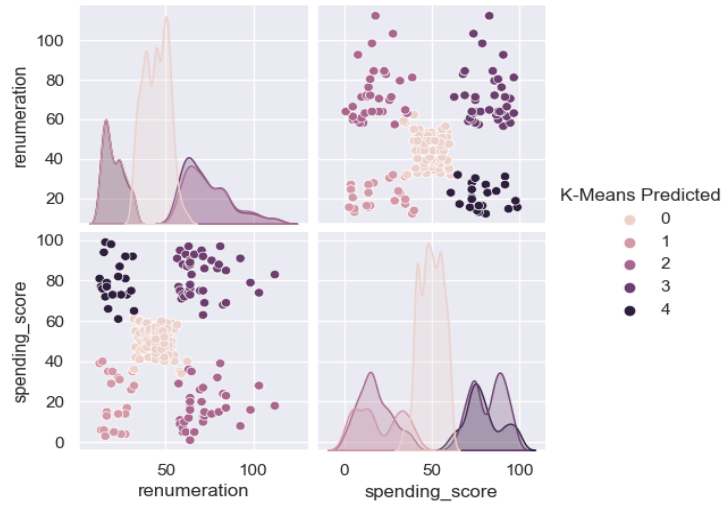
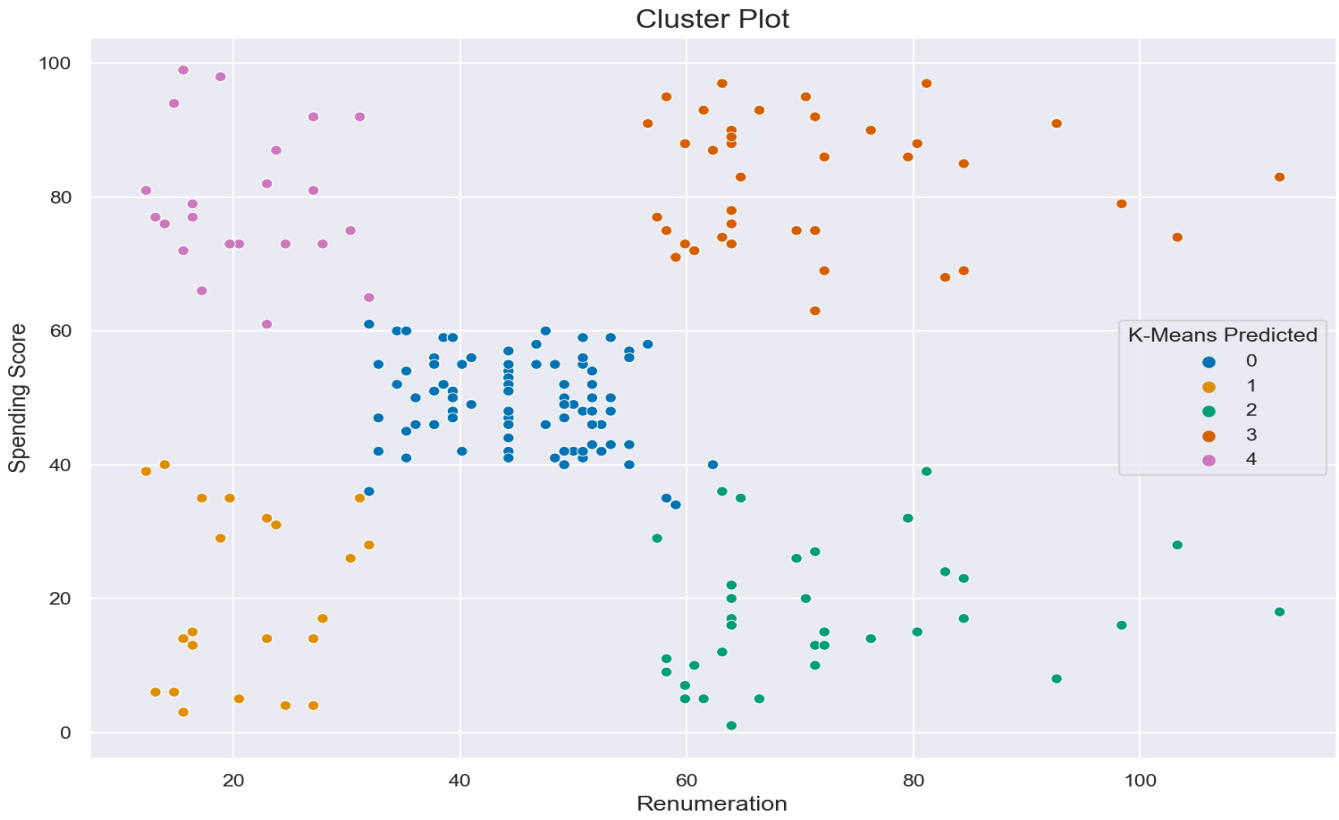


Figure 4: Pair Plot with K=5

K-Means Predicted	
0	774
3	356
2	330
1	271
4	269

Table 1: Observations per predicted class



The clusters are well separated. They don't overlap each other. The cluster shape is roughly circular. But they are not densely populated.

We can see that cluster 2 has highly remunerated people, but they are not spending much. Cluster 1 is the least spending group due to their less remuneration. Cluster 0 is right in the middle. Clusters 4 & 3 are the high spenders. With further analysis, we can find that clusters 0, 1, and 2 are segregated and give them attractive offers to buy more products.

SENTIMENT ANALYSIS

The sentiment analysis was conducted to find out how the customers felt about the brand and the products.

Figure 6: Word Cloud with no Stop Words - Summary

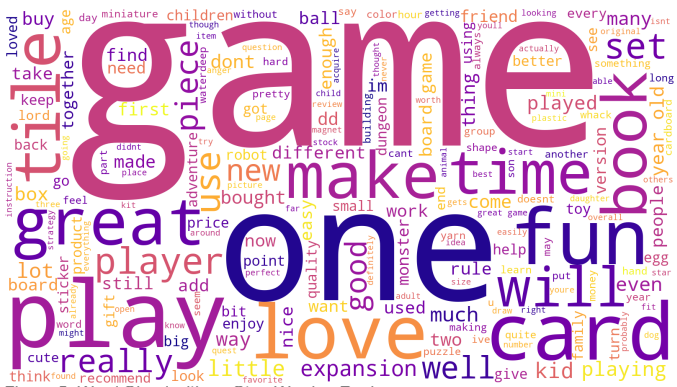
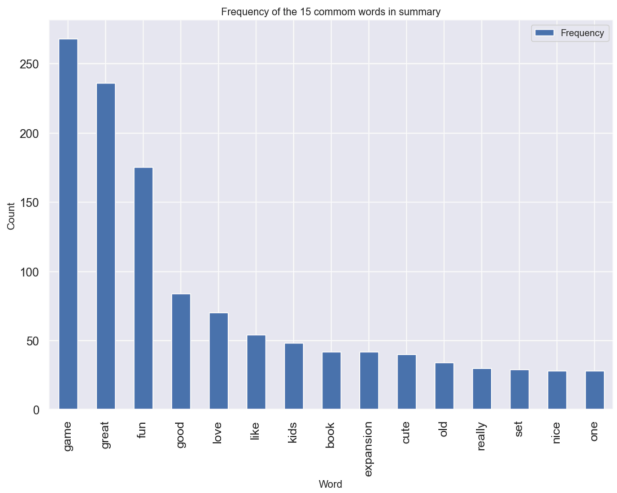
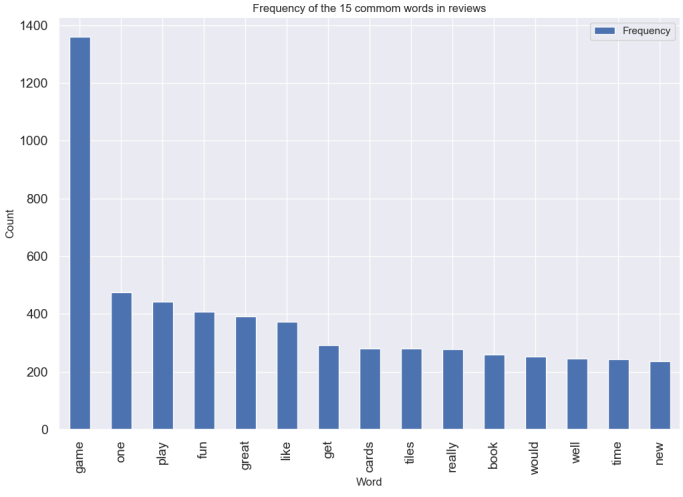
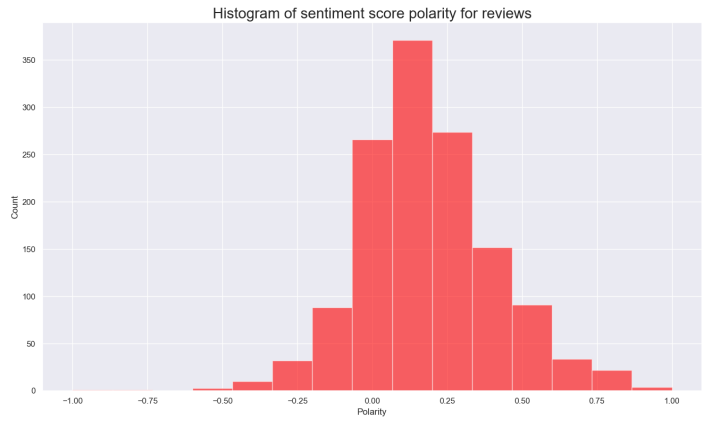


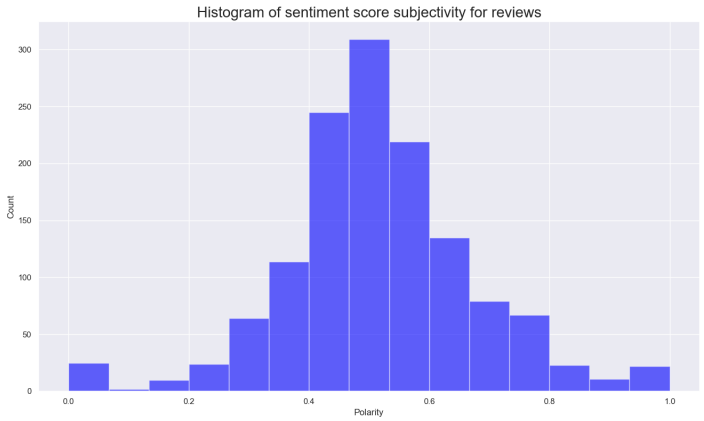
Figure 5: Word Cloud with no Stop Words - Review

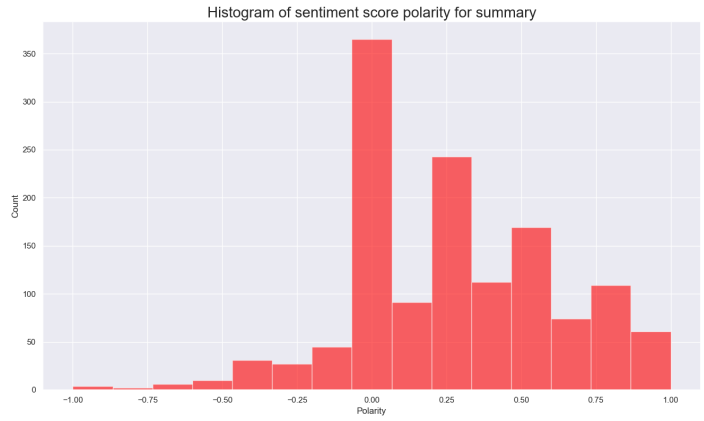
The word cloud was plotted to find the most common words in the reviews and summary dataset, with the size of the word in the word cloud determining the word which is most frequent.

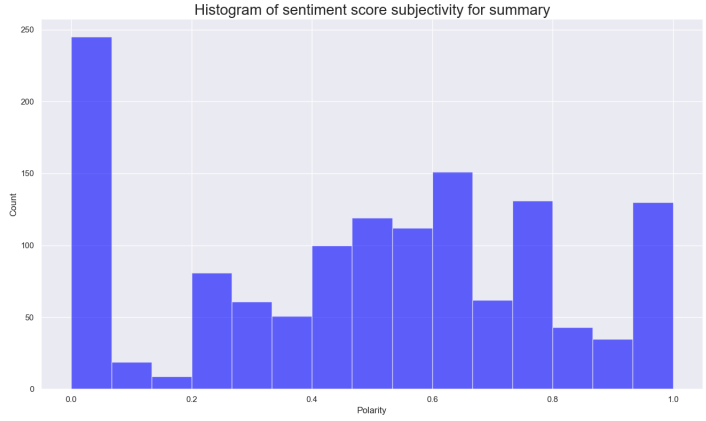










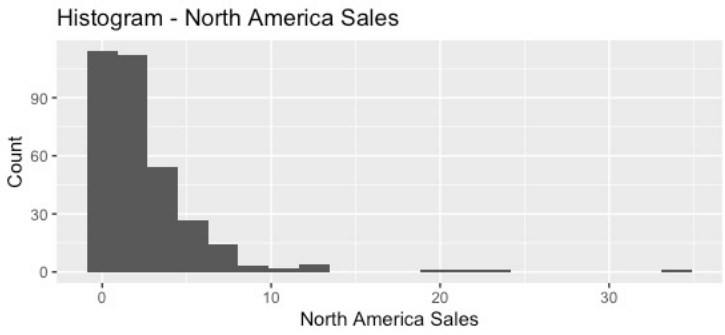


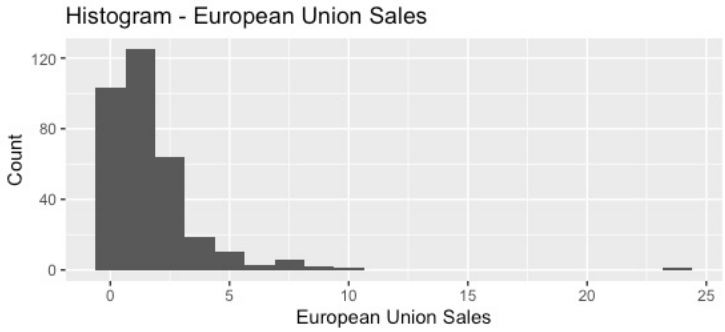
The polarity score and subjectivity of both review and summary are positive. The top 20 positive reviews and summaries are strongly positive as values are 1 or closer to 1. The top 20 negative reviews and summaries are closer to -0.5. Both of these show that people have stronger positive sentiments towards the products. The brand name Turtle appears to be not popular as in the whole data frame, there is only one reference in the review.

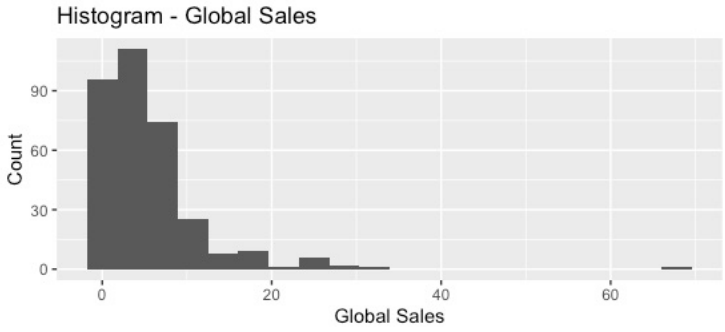
VISUALIZING THE SALES DATA

The turtle_sales.csv contains the observations from 3 regions. It has 352 rows. The NA_sales and EU_sales have sales with a value of 0. Eliminating those values, we are now left with 334 rows of data.

HISTOGRAM

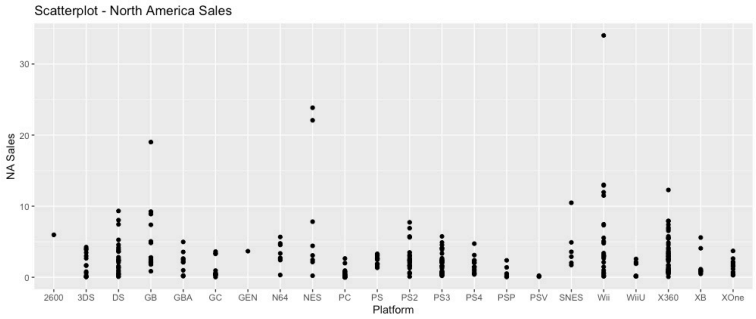


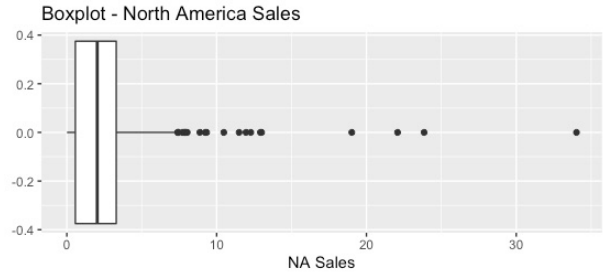


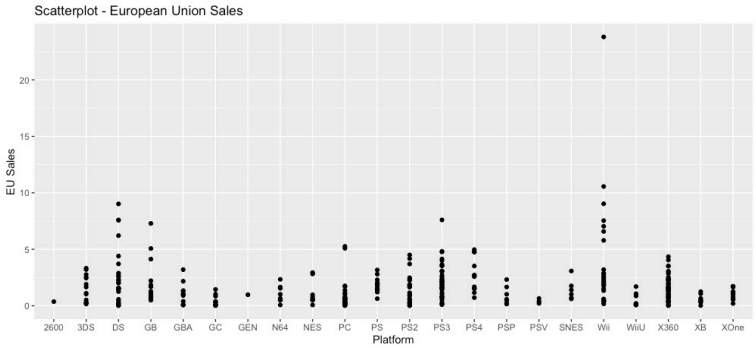


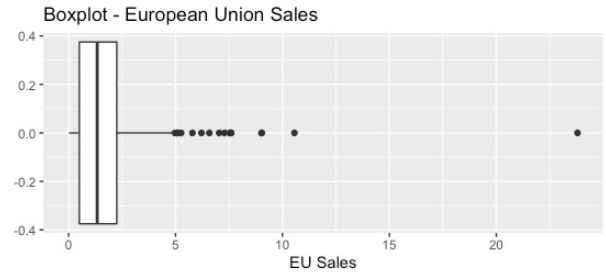
The histogram of all three regions shows that they are right skewed or positively skewed, many values are near the lower end of the range, and higher values are infrequent.

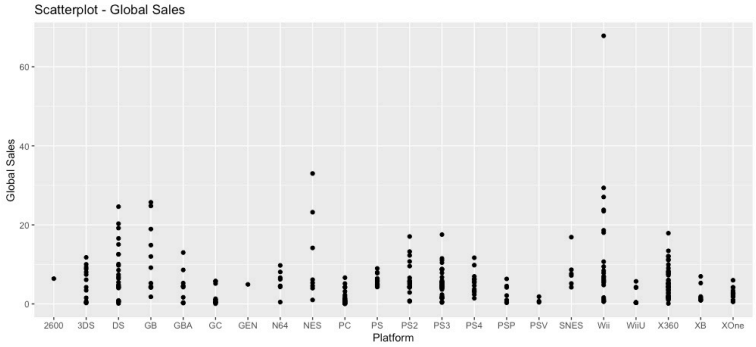
SCATTER AND BOX PLOT

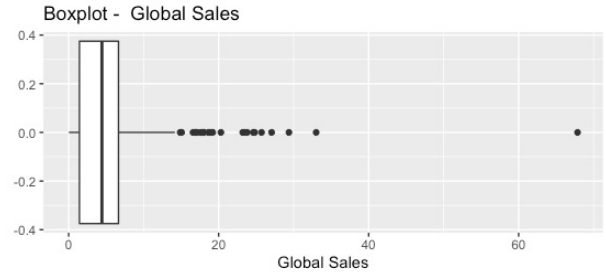










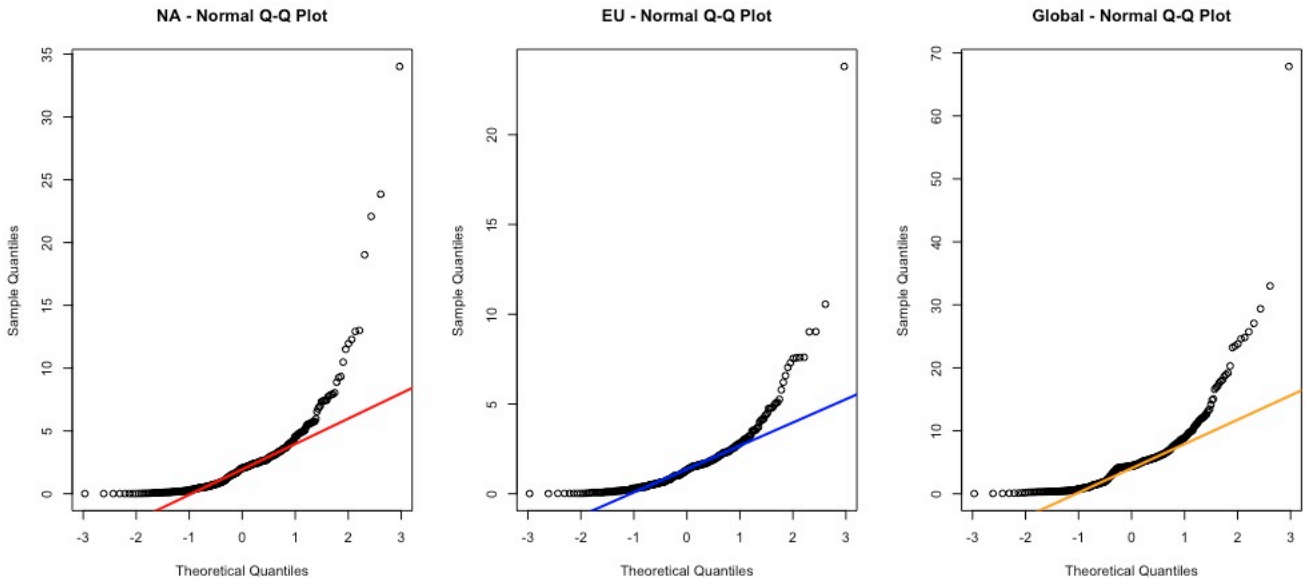


Both the scatter and box plot shows the presense of outliers.

For this analysis, the outliers are kept as it is as more insights on the outliers are needed before it can be removed or not. The outliers have certainly skewed the data.

CHECK FOR NORMALITY

Q-Q PLOT



The values in the tails of the distribution of all 3 plots are quite extreme from the normal distribution.

SHAPIRO-WILK TEST

The null hypothesis in the Shapiro-Wilk test is that the sample comes from a normal distribution. The alternative hypothesis is that the sample does not come from a normal distribution.

In this data set, the p-value is less than 0.05, and the W-value is less than 68% in all three cases. So the values are not distributed normally.

```
> shapiro.test(sales_prd$NA_Sales_sum)

        Shapiro-Wilk normality test

data:  sales_prd$NA_Sales_sum
W = 0.63115, p-value < 2.2e-16

> shapiro.test(sales_prd$EU_Sales_sum)

        Shapiro-Wilk normality test

data:  sales_prd$EU_Sales_sum
W = 0.64802, p-value < 2.2e-16

> shapiro.test(sales_prd$Global_Sales_sum)

        Shapiro-Wilk normality test

data:  sales_prd$Global_Sales_sum
W = 0.68245, p-value < 2.2e-16
```

SKWENESS AND KURTOSIS

```
> skewness(sales_prd$NA_Sales_sum)
[1] 4.286135
> skewness(sales_prd$EU_Sales_sum)
[1] 4.811984
> skewness(sales_prd$Global_Sales_sum)
[1] 4.050553
```

The values indicate that the distribution is a positive skew.

```
> kurtosis(sales_prd$NA_Sales_sum)
[1] 30.8299
> kurtosis(sales_prd$EU_Sales_sum)
[1] 44.21582
> kurtosis(sales_prd$Global_Sales_sum)
[1] 32.45573
```

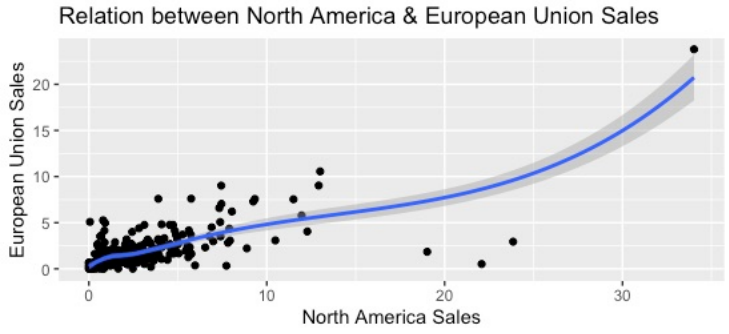
The kurtosis values of all three are more than 3, which means that all three are heavy-tailed, meaning all are further away from the mean.

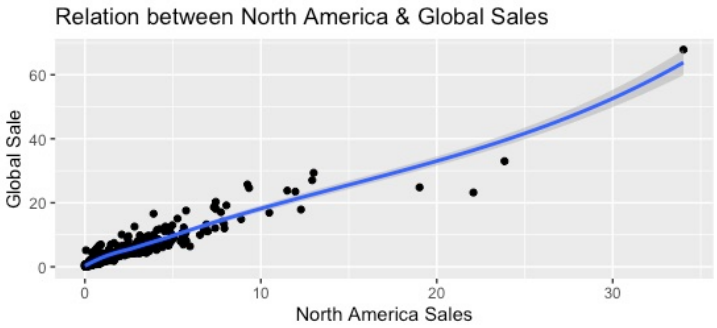
CORRELATION

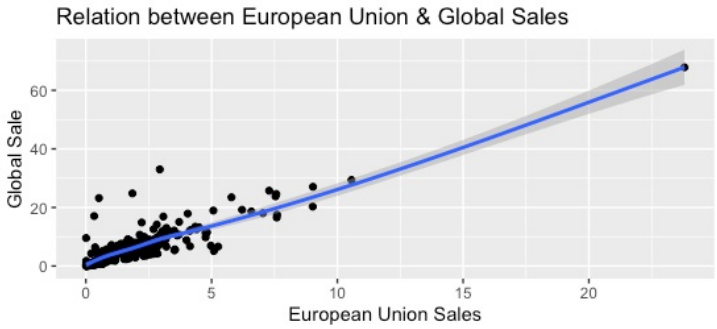
```
> cor(sales_prd$NA_Sales_sum, sales_prd$EU_Sales_sum)
[1] 0.6976798
> cor(sales_prd$NA_Sales_sum, sales_prd$Global_Sales_sum)
[1] 0.9331351
> cor(sales_prd$EU_Sales_sum, sales_prd$Global_Sales_sum)
[1] 0.8748341
```

All three sales data have a positive correlation and are closer to 1. Hence they have a strong positive correlation, meaning as one increases, the other also increases. Among the 3 correlation is stronger between NA Sales & Global sales and EU Sales & Global Sales.

The correlation between the regions can be plotted for better understanding.







A simple linear regression supports the correlation shown in the plots above with R^2 values of 48.68%, 87.07%, and 76.53%, respectively.

A much better model can be produced by MLR, with Global_sales as the dependent variable and NA_sales & EU_sales as the independent variable. The summary of the model is as below:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.22583    0.08222   2.747  0.00635 **
NA_Sales_sum  1.15237    0.02502  46.064  < 2e-16 ***
EU_Sales_sum  1.34554    0.04213  31.939  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.129 on 331 degrees of freedom
Multiple R-squared:  0.9683,    Adjusted R-squared:  0.9681
F-statistic:  5061 on 2 and 331 DF,  p-value: < 2.2e-16
```

The R2 value shows that the model is a good fit, and it can explain 96.83% of the variance in the response variable. And the residual plots show no particular pattern.

However, when this model was tested, the predicted values and actual values were not matching, as in the actual values were not present in the upper and lower limit of the predicted values.

CONCLUSION

- People with basic education seem to be accumulating more loyalty points and their spending scores, along with the graduates, are the highest. So they are our loyal customers, whom we need to retain.
- Female customers are found to be more.
- There is a segment of customers who are high earners, but the spending score is very small.
- There is another segment that can be potential high spenders as they are in the middle of the spending score.
- More analysis is required to find out these customer segments.
- The brand image or brand value of the Turtle Games is non-existent.
- The overall sentiment of people is positive, but there is not enough evidence to suggest that the positive sentiment is due to the brand Turtle Games.
- The marketing team should come up with offers to cater to the low-spending group and also change the existing strategy to market the brand better.
- The data is not normally distributed and is positively skewed.
- Global sales have a linear relation with North American and European Sales.