This study is conducted to explore opportunities for improving overall sales performance by analyzing consumer insights and sales data. There are six key questions to focus on as follows:

- how customers accumulate loyalty points
- how useful are remuneration and spending scores data
- can social data (e.g. customer reviews) be used in marketing campaigns
- what is the impact on sales per product
- the reliability of the data
- if there is any possible relationship(s) in sales between North America, Europe, and global sales

First part of study is conducted in Python. Data file "tutle_reviews.csv" is used in this study. There are total of 10 columns with 2000 entries. No null data was found according to the info table, Table1.

Table1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 11 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   gender                 2000 non-null   object
 1   age                    2000 non-null   int64
 2   remuneration (k£)      2000 non-null   float64
 3   spending_score (1-100)  2000 non-null   int64
 4   loyalty_points         2000 non-null   int64
 5   education              2000 non-null   object
 6   language               2000 non-null   object
 7   platform               2000 non-null   object
 8   product                2000 non-null   int64
 9   review                 2000 non-null   object
 10  summary                2000 non-null   object
dtypes: float64(1), int64(4), object(6)
```

Columns "language" and "platform" are dropped as not relevant to customers' loyalty points, remuneration/ spending relationship, and customer review which are the main focus of this study.

Column 2 and 3 originally with units stated in the column headers. Rename for better presentation.
No adjustment is needed to datatype.

Table2. Descriptive statistics of cleaned data

	age	remuneration	spending_score	loyalty_points	product
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	39.495000	48.079060	50.000000	1578.032000	4320.521500
std	13.573212	23.123984	26.094702	1283.239705	3148.938839
min	17.000000	12.300000	1.000000	25.000000	107.000000
25%	29.000000	30.340000	32.000000	772.000000	1589.250000
50%	38.000000	47.150000	50.000000	1276.000000	3624.000000
75%	49.000000	63.960000	73.000000	1751.250000	6654.000000
max	72.000000	112.340000	99.000000	6847.000000	11086.000000

Factors driving Loyalty Points

Scatterplots are plotted to check for linearity between each variable and loyalty points.
Blue: Age vs Loyalty Points
Orange: Remuneration vs Loyalty Points
Green: Spending Score vs Loyalty Points

Fig,1. Different variables vs Loyalty Points

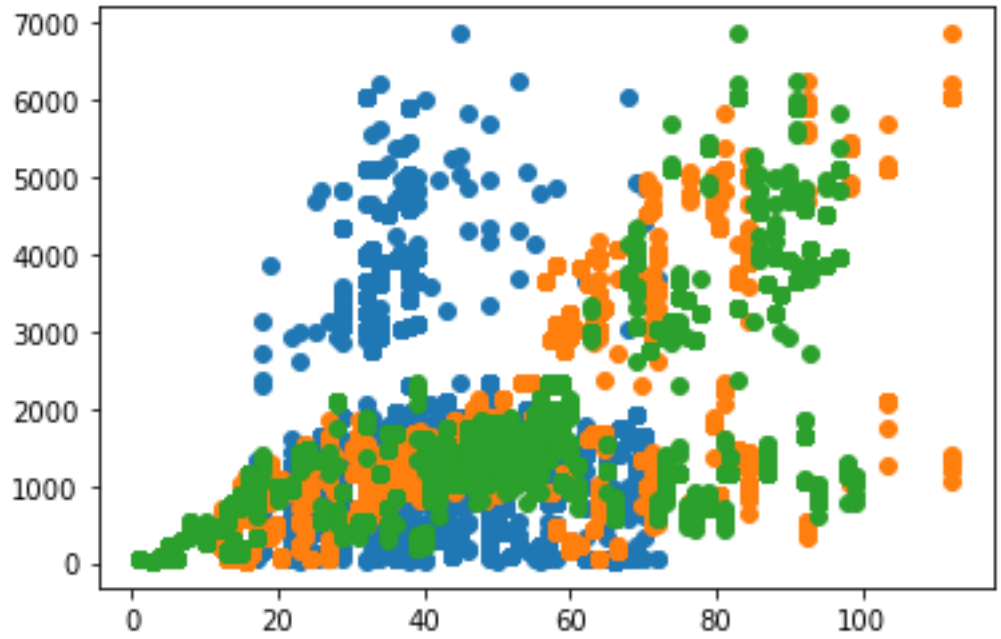


Fig1 shows no clear linear relationship between Age and Loyalty Points. Both Remuneration and Spending scores show a potential possible relationship with Loyalty Points, and both trends look similar, indicating that Remuneration and Spending Scores may also have a strong relationship.

OLS regression used to further look into the relationship between Loyalty Points and Remuneration.

Table3. Key results from OLS regression

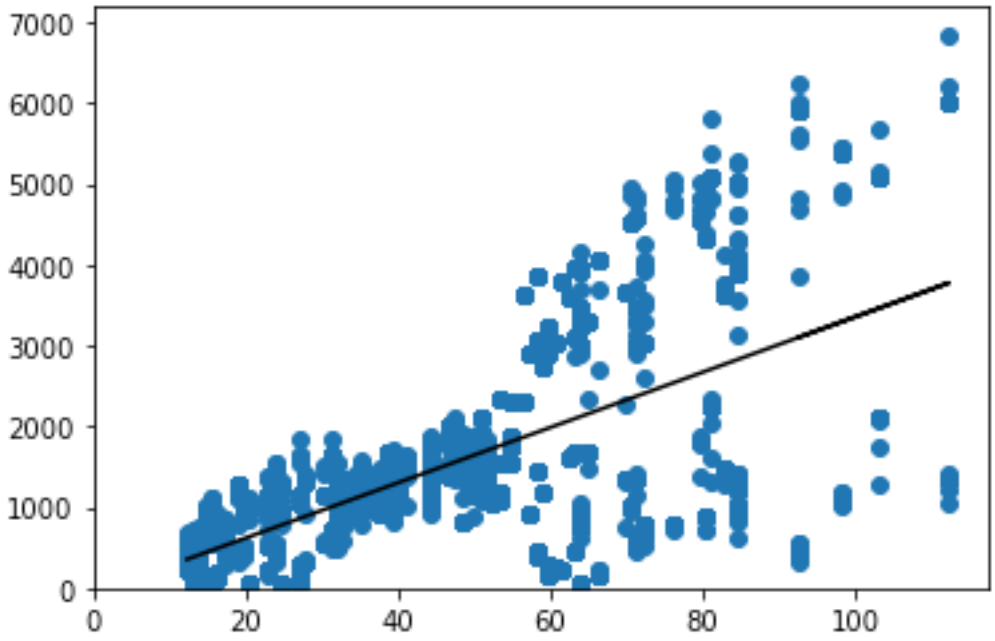
OLS Regression Results

Dep. Variable:		y_lp		**R-squared:**		0.380
Model:		OLS		**Adj. R-squared:**		0.379
Method:		Least Squares		**F-statistic:**		1222.
Date:		Fri, 23 Dec 2022		**Prob (F-statistic):**		2.43e-209
Time:		02:08:19		**Log-Likelihood:**		-16674.
No. Observations:		2000		**AIC:**		3.335e+04
Df Residuals:		1998		**BIC:**		3.336e+04
Df Model:		1				
Covariance Type:		nonrobust				
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-65.6865	52.171	-1.259	0.208	-168.001	36.628
x_r	34.1878	0.978	34.960	0.000	32.270	36.106

R-squared=0.38, only 38% of data can be explained by the regression line
P>|t|=0, i.e. the estimated slope (x_r coefficient) is significant
p-value = 2.43e-209 < 0.05, i.e. significant

The statistics support that the regression result is significant and can be used for prediction, despite the fact that the R-squared value is relatively low. In this case in future studies, we may try to improve the regression by adding more variables, and/ or clustering methods to break down the data into different groups with similar behavior.

Fig2. Remuneration vs Loyalty Regression Line



The same method applied to spending score vs loyalty points and similar results were driven.

The usefulness of Remuneration and Spending data

This section only focuses on remuneration and spending score, therefore all other columns are dropped for smoother operation.

Scatter plots and Pairplots are used to overview the relationship between the two variables.

Fig3. Scatter plot remuneration vs spending score

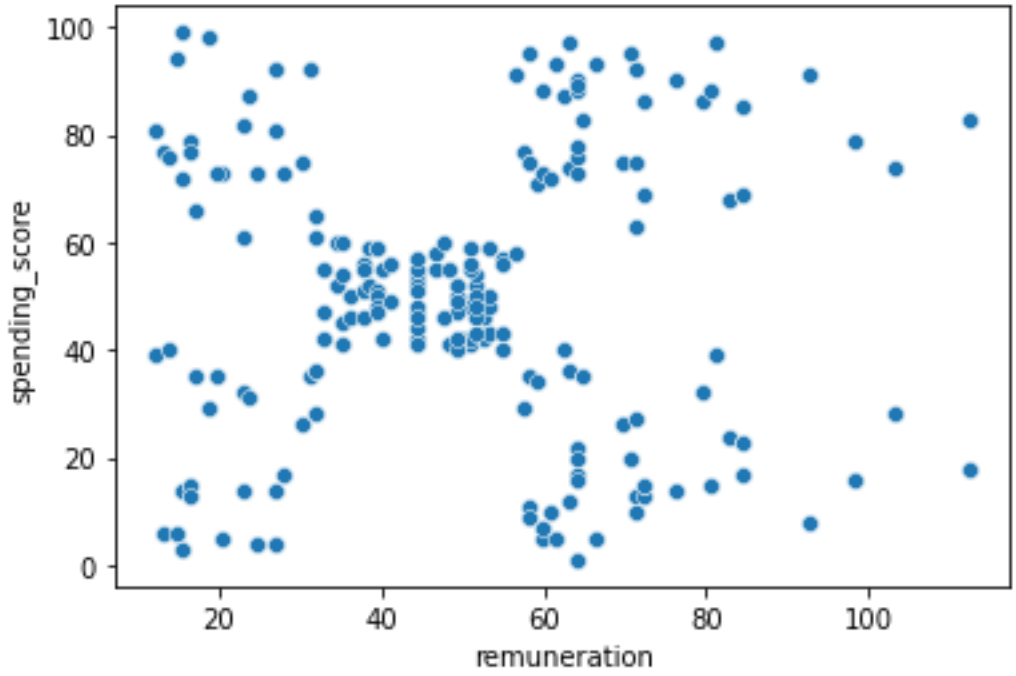
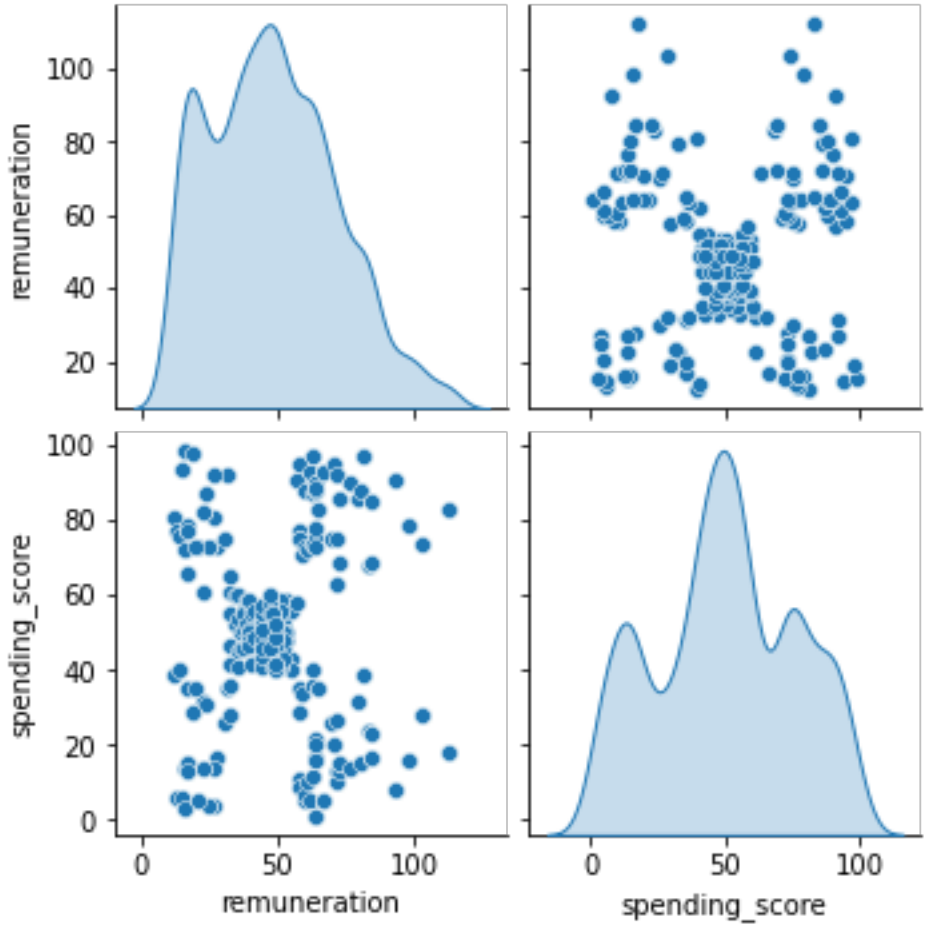


Fig4. Pair plot



From Fig3 and Fig4 we can see clear chunks of data points distributed in different areas on the graphs. It looks like there are five groups of data points, but we will use the Elbow and Silhouette methods to determine the best number of clusters to use.

Fig5. The Elbow Method result

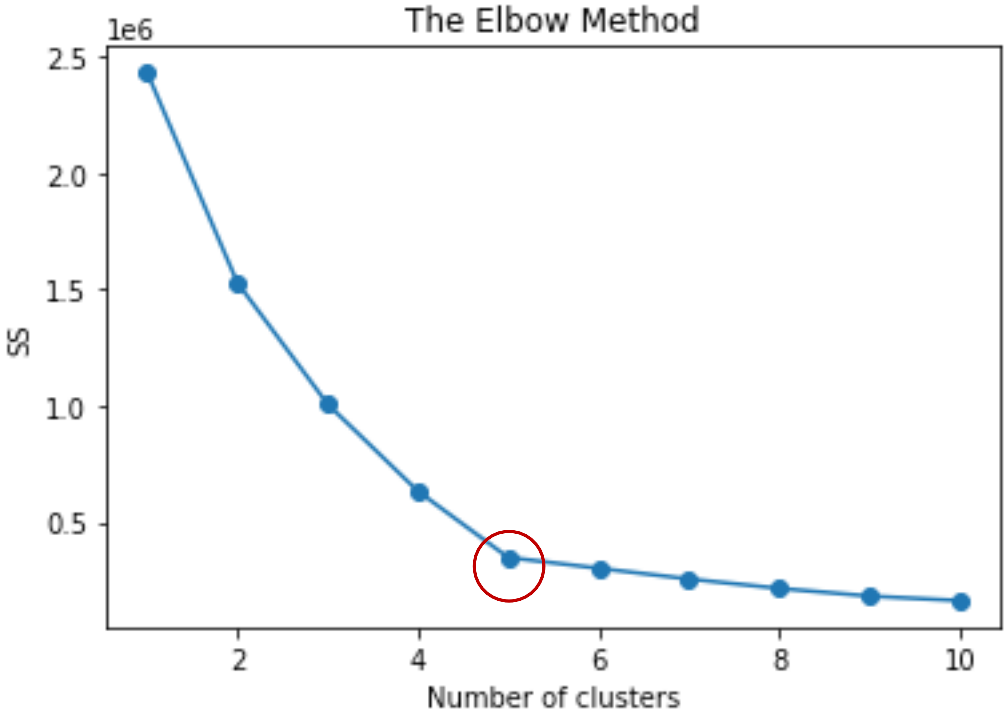
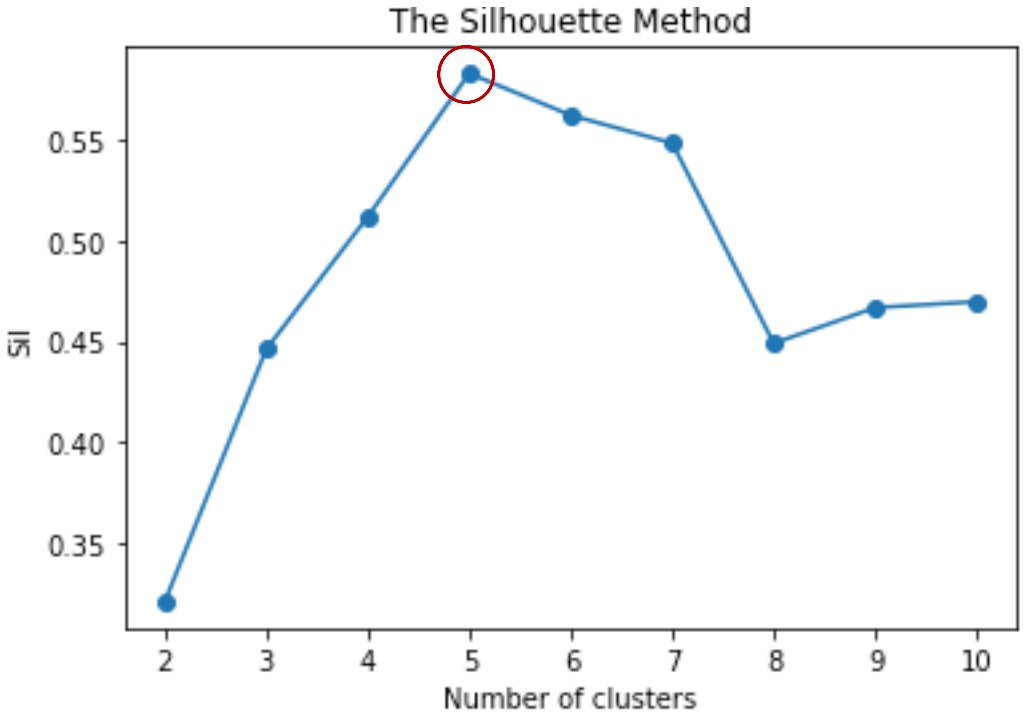


Fig6. The Sihouette Method result

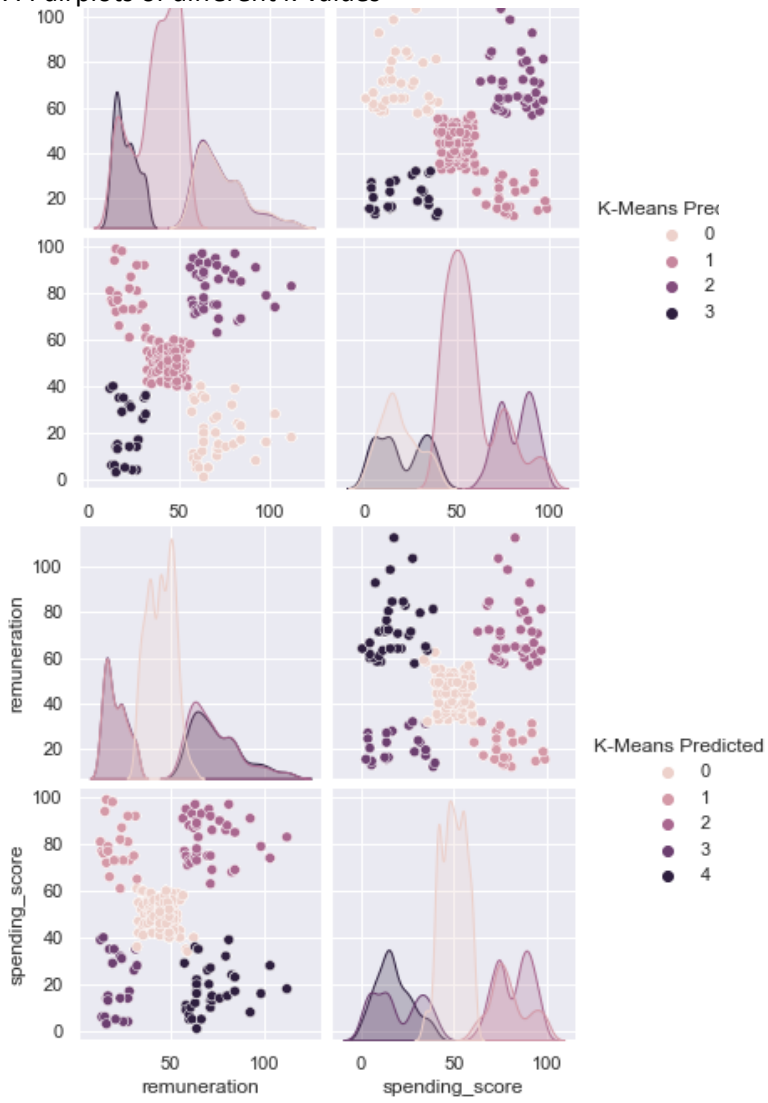


Both methods show 5 seems to be reasonable number of clusters to use. K-means model is then applied to three different k-values 4,5,6 to confirm.

Table4. Data points distribution under different k-values

	K=4	***K=5***	***K=6***
No. of data points in each cluster (in descending order)	1013	774	767
	356	356	356
	351	330	271
	280	271	269
		269	214
			123

Fig7. Pairplots of different k-values



Results supporting the 5 would be a good number to use for k-value. From both the table and pariplots we can see that when k=4 the distribution of cluster 1 has two peaks which are significantly different from each other. If we use k=4 the mean of cluster 1 is likely to offset the high and low peaks which gives a less reliable result. On the other hand, if k=6 we can see that cluster 0 and cluster 5 have a very similar distribution. Therefore there seems to be no reason to keep them as two groups.

The result indicates that there are 5 groups of customers with different spending habits that we should keep in mind in future analysis. The 5 groups can be considered as

1. High income and high spending
2. High income but low spending
3. Low income and low spending
4. Low income but high spending
5. Mutual Group

Customers Reviews

This section continues to use the cleaned data, but keeps only columns “review” and “summary”.

In order to apply NLP, data needed to be further adjusted for smooth processing. Few steps are applied to data (both columns) in Python.

1. Change all to lowercase and join with space
2. Punctuation removed
3. Drop duplicates entries
4. Tokenise and create wordcloud
5. Remove alphanumeric characters and stopwords
6. Create wordcloud without stopwords
7. Identify 15 most common words and polarity
8. Identify top 20 positive and negative reviews and summaries

Step 1 to 4 can basic adjument made to text to remove any non-words element which will not be processed by NLP. After tokenising the data, frequency distributions are calculated and show the words appear most frenquently are mainly stopwords such as “the”, “and”, “of”, etc. Therefore alphanumeric characters and stopwords have to be removed and wordcloud is created again.

Fig.8 WordCloud from customers reviews

Fig 9. WordCloud from customers reviews summary

WordClouds give an overview on which word is most used in customers' comments. We can see that most words on the wordclouds look positive. However, by looking at one word at a time without any context could be misleading. Furthermore, some of the words on the wordclouds are kind of neutral and give no information on customers' preferences. For example, the word "game" is the biggest on both of the wordclouds. However, this gives no insights about whether customers like the product of not. This is simply a descriptive noun as these are all reviews for gaming products.

15 most frequent words in customer reviews

Fig.10 15 most frequent words in Customer Reviews

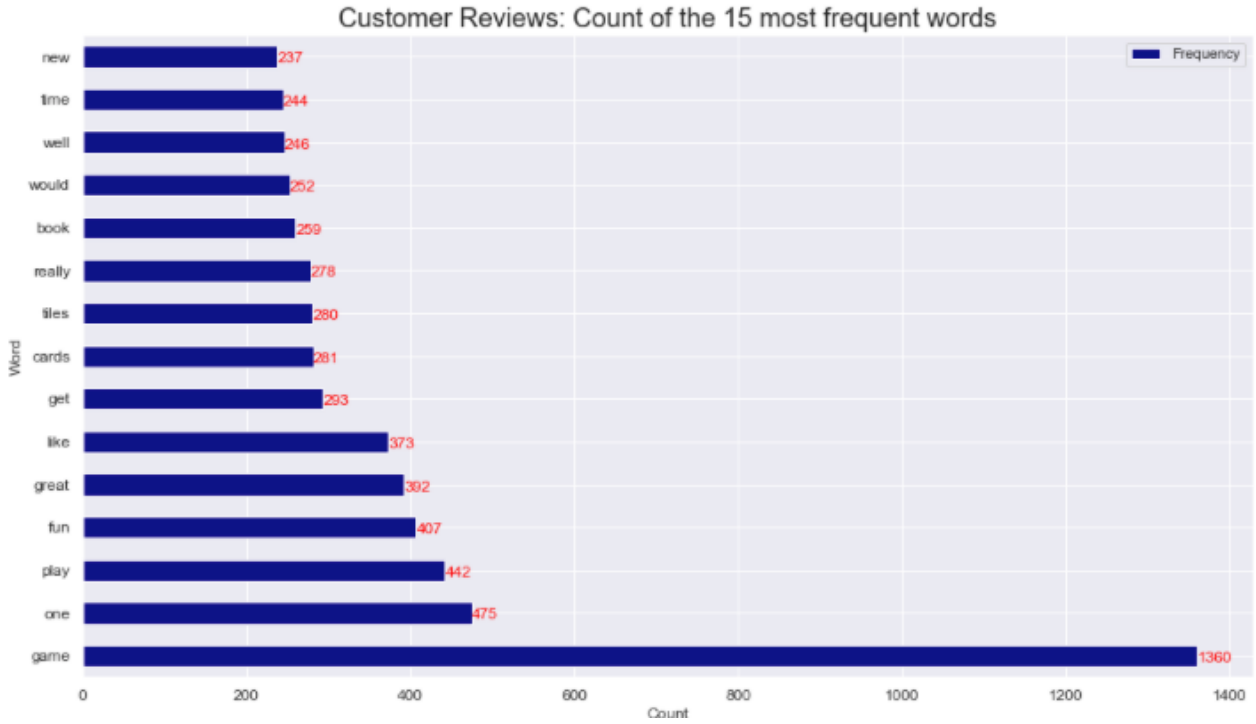
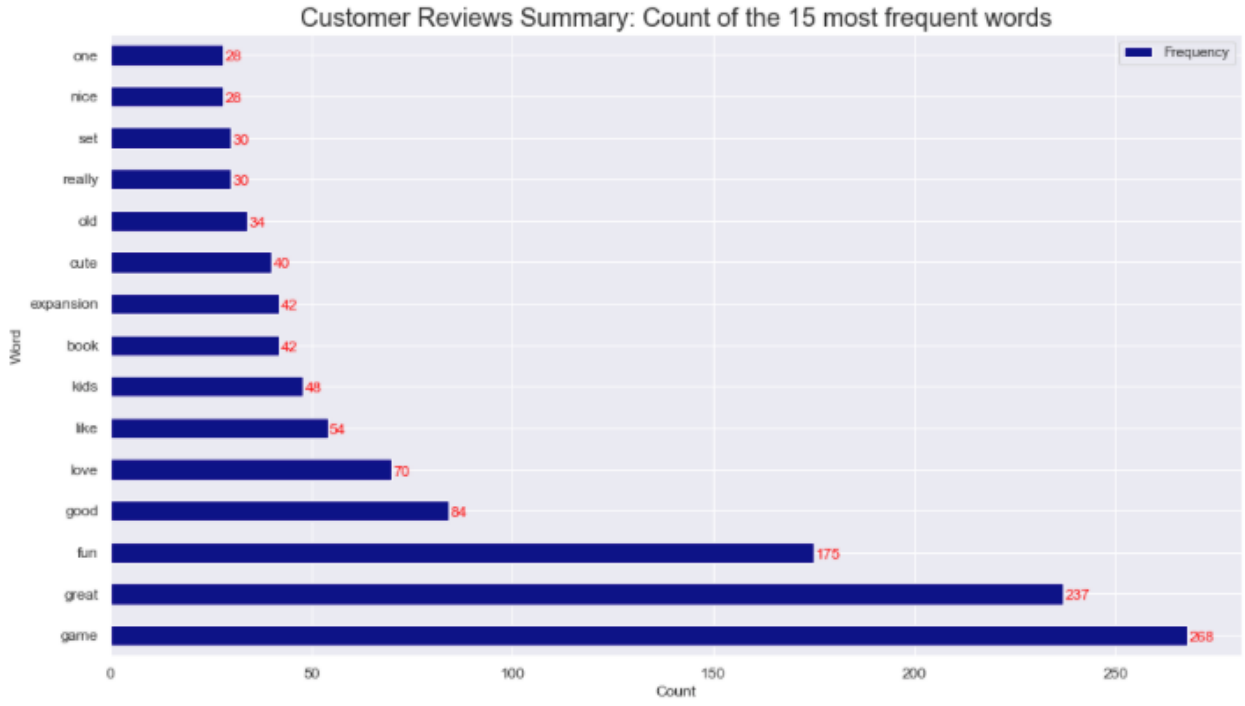
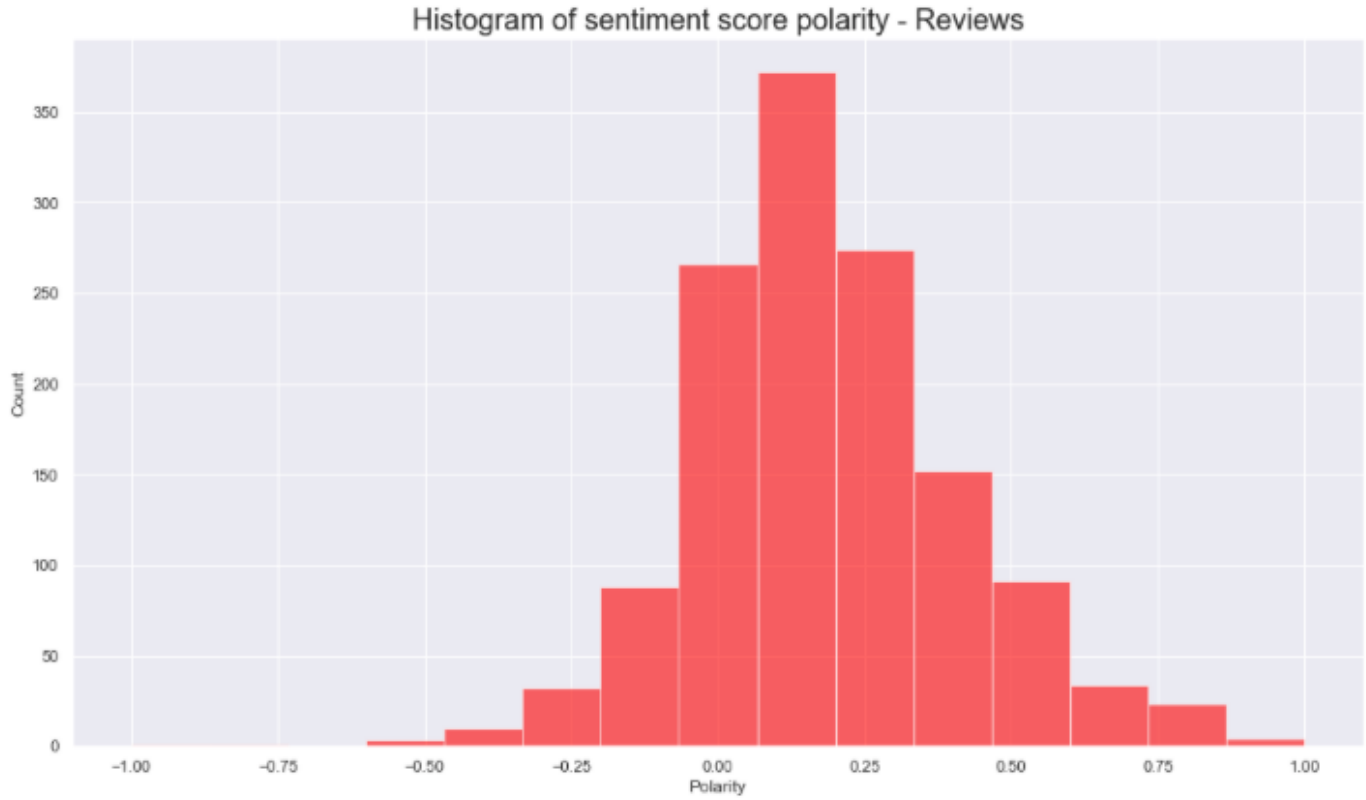


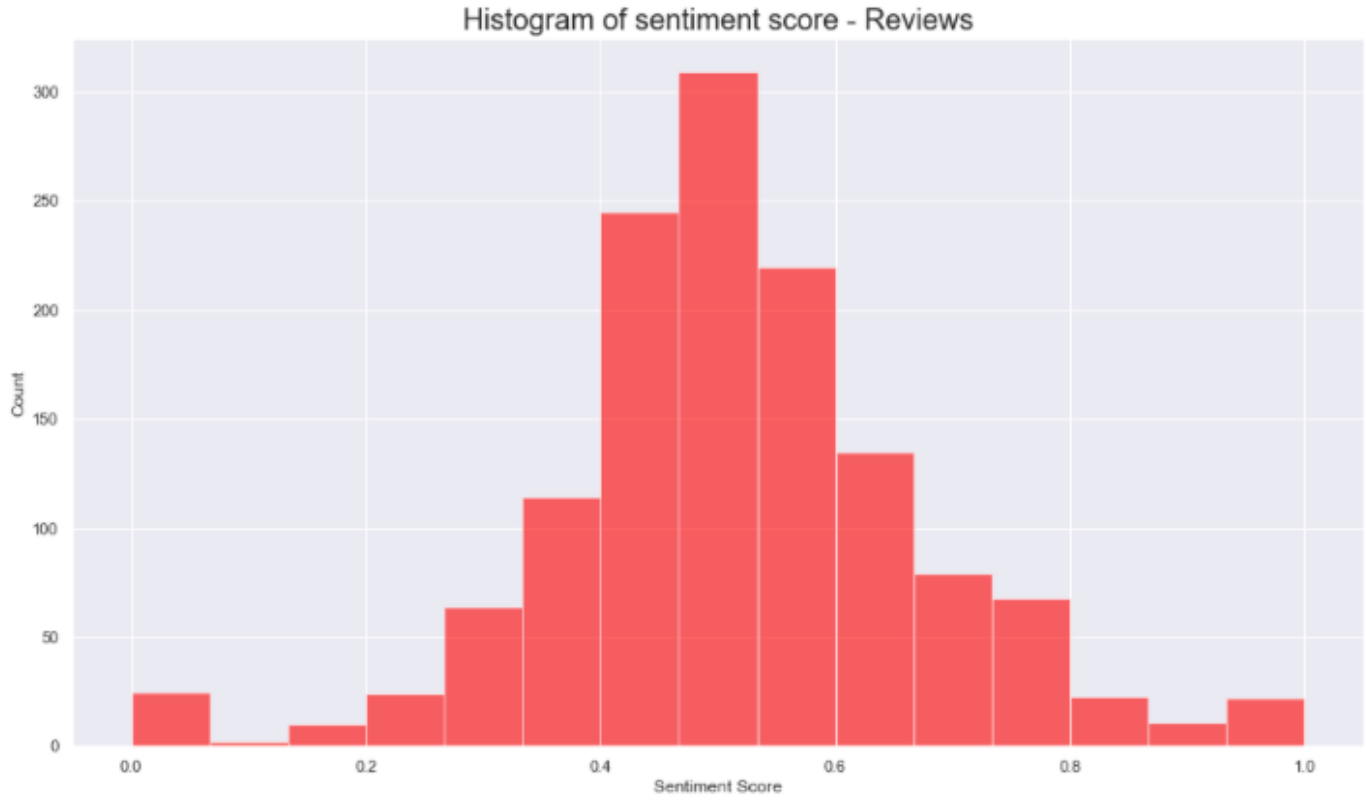
Fig.11 15 most frequent words in Customer Reviews Summary



To get more insights from the reviews comments, further sentiment analysis is conducted.
For both columns “review” and “summary”, polarity and subjectivity of each comments is determined. Histogram is then plotted to represent the result.

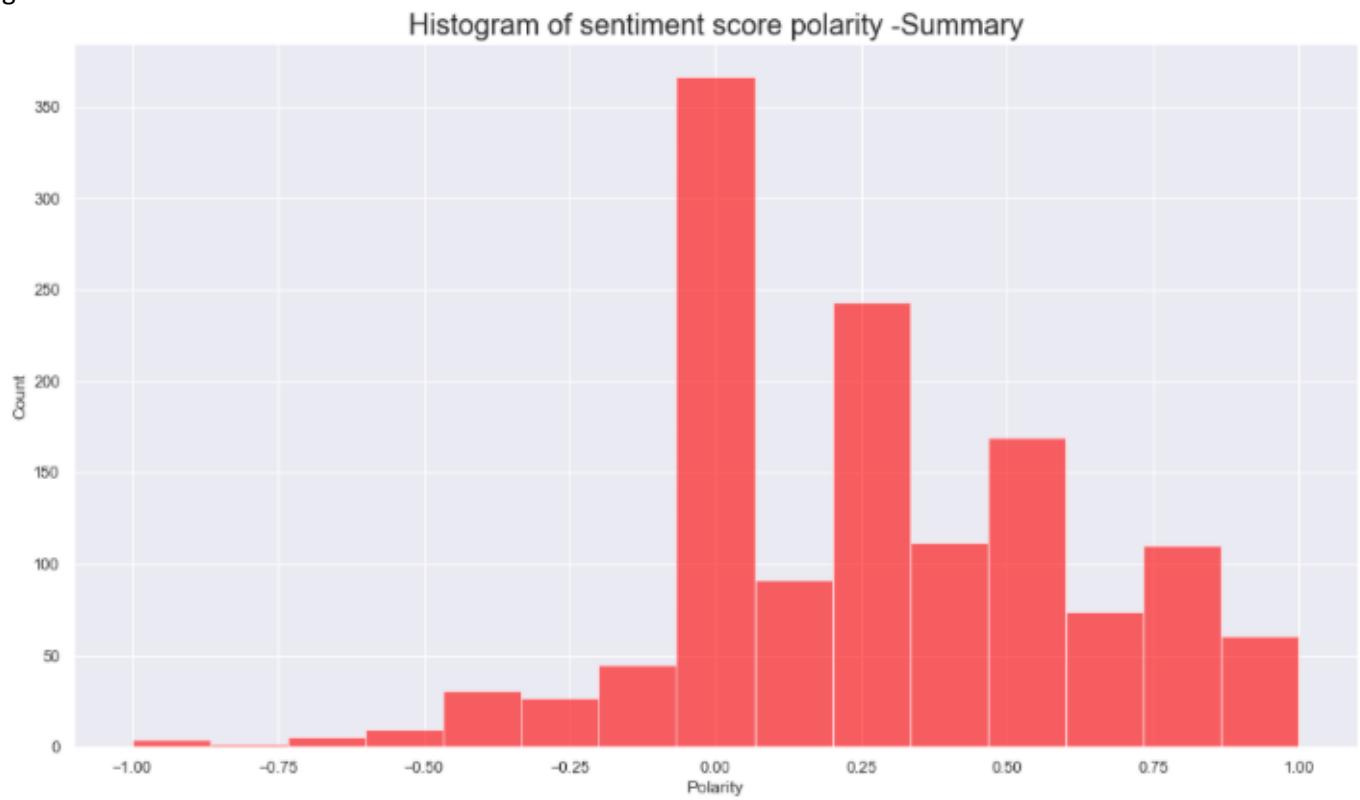
Fig11.

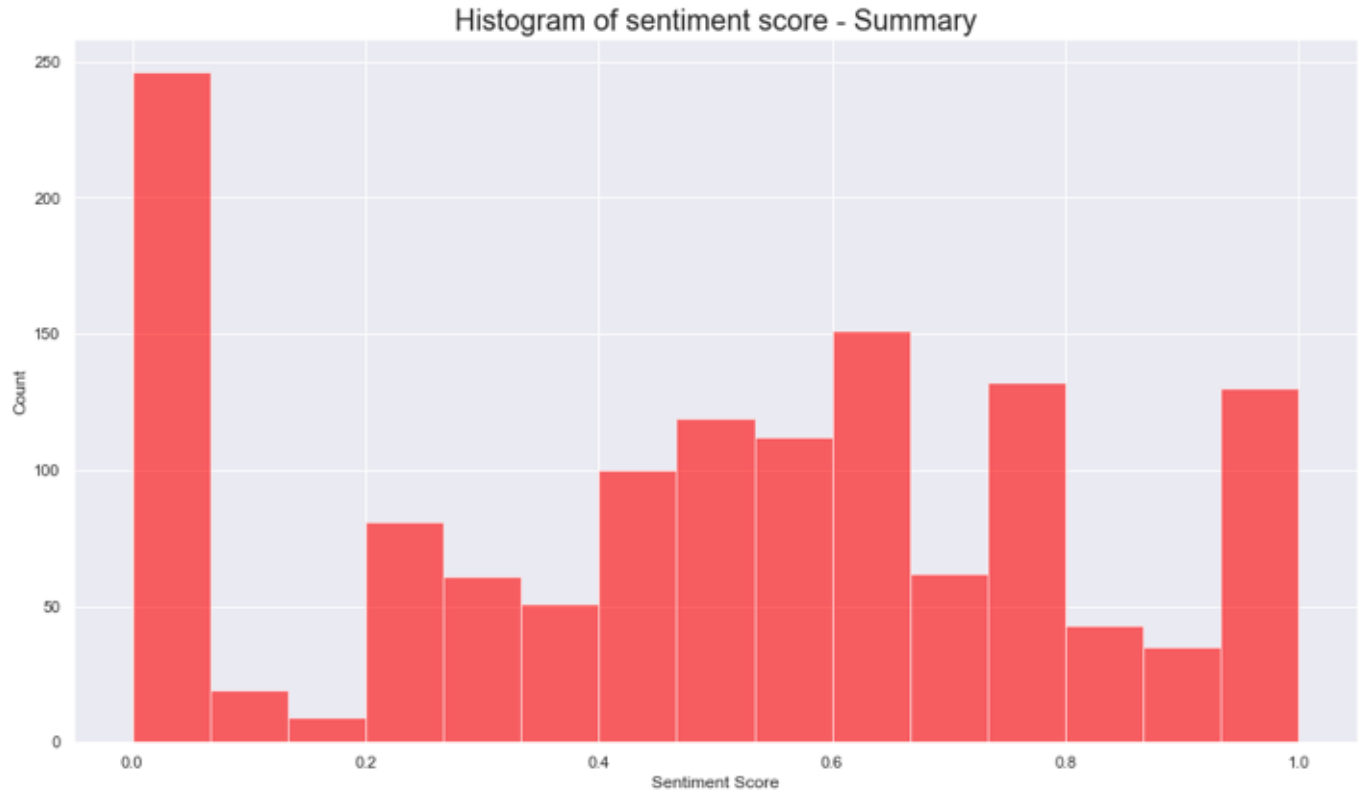




Polarity of Customer Reviews is slightly skewed to the positive side, which indicates that most comments received are fairly positive, lying between 0-0.3. Sentiment score is concentrated between 0.4-0.6.

Fig12.





The distribution of the summary polarity and sentiment score is not as regular but we can still see most of the comments have positive polarity. There is no obvious trend in sentiment score, as the summary is usually simple and short, with less subjectivity tracked.

Top 20 positive and negative reviews and summaries

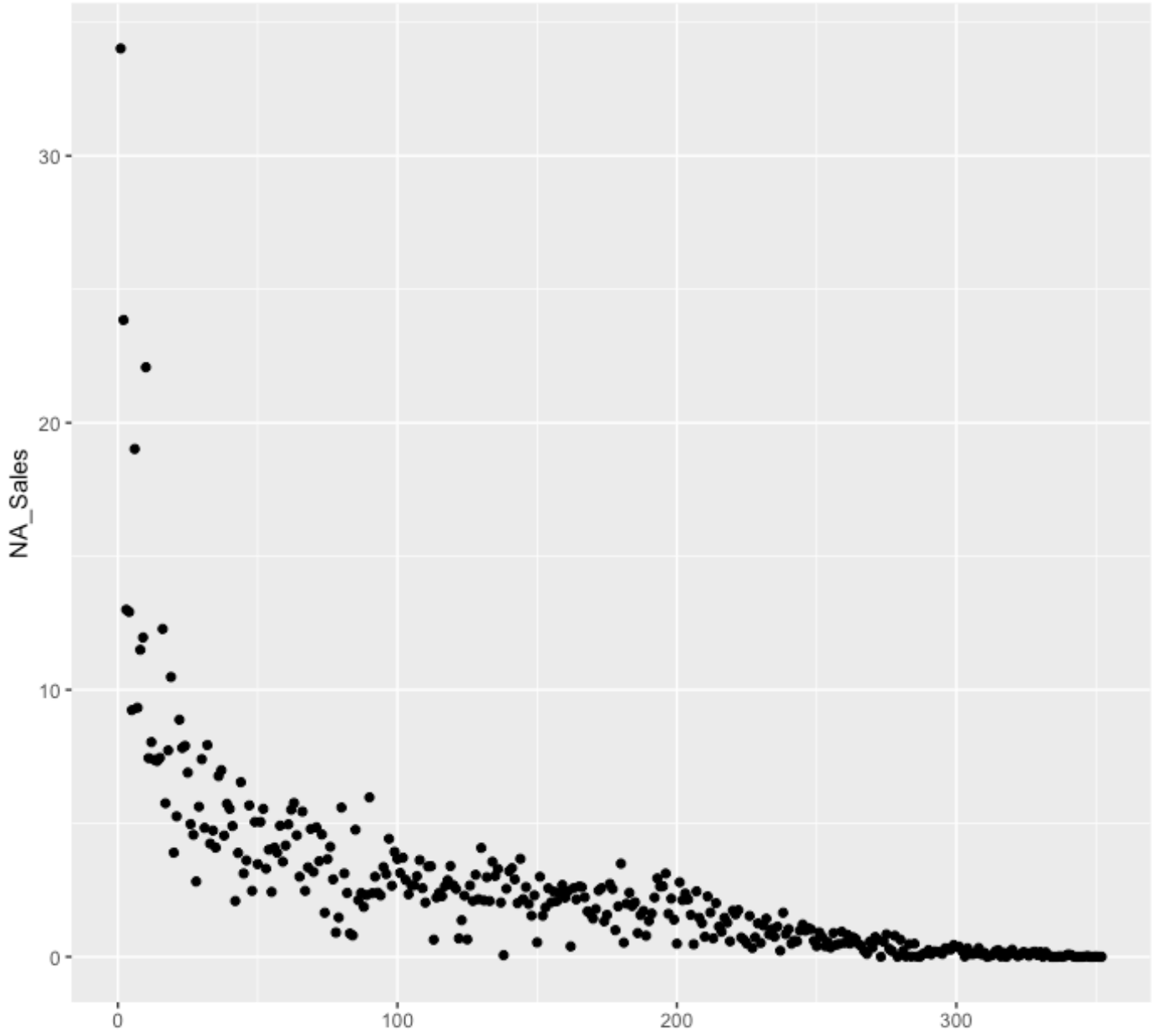
Top 20 positive and negative reviews and summaries are found (separately), according to polarity score.
Columns “polarity” and “subjectivity” are added in the result table together with the comments as a reference.
Result please refer to attached Jupyter Notebook.

Impact on sales per product

From this section of study we will use R to continue the analysis.
Data file “turtle_sales.csv” is imported to R and unnecessary columns “Ranking”, “Year”, “Genre” and “Publisher” are removed.

Scatterplots are used to get an overview of the sales data. For each sales column, “NA_Sales”, “EU_Sales”, “Global_Sales”, two graphs are plotted, one against “Platform” and one shows the sequence of the sales data.

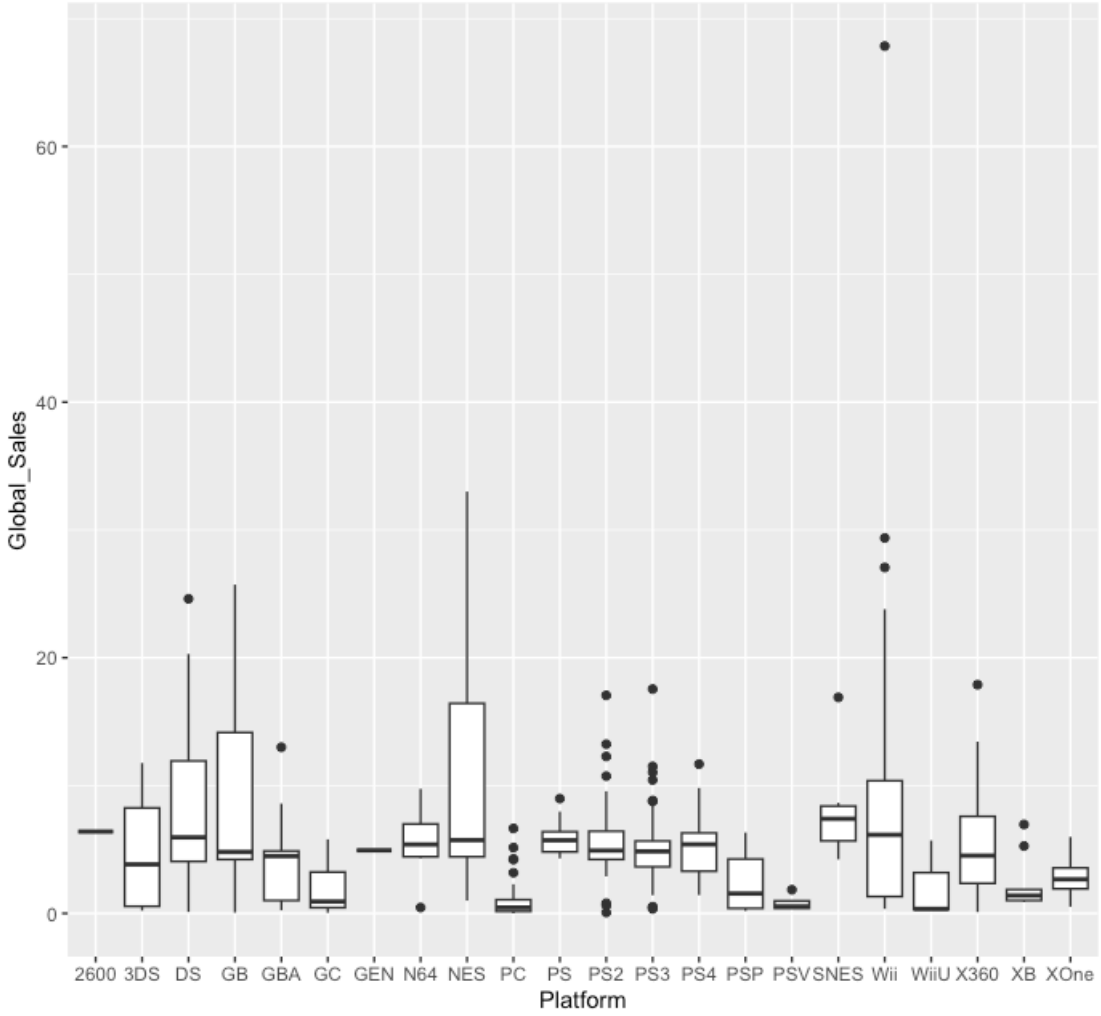
Fig13. Scatterplot showing sequence of sales data of NA area.



NA, EU, and Global sales all have similar characteristics, according to the graphs. We can see the data points have higher density at low number sales, and only a few reaching above 25, in this case for NA sales, which may

considered as outliers. This result is also supported by the histograms and box plots which suggest Wii and PS2,3,4 games sold tend to have more obvious outliers. This could be due to the popularity of the game consoles and the game itself.

Fig13. Box Plot of global sales data



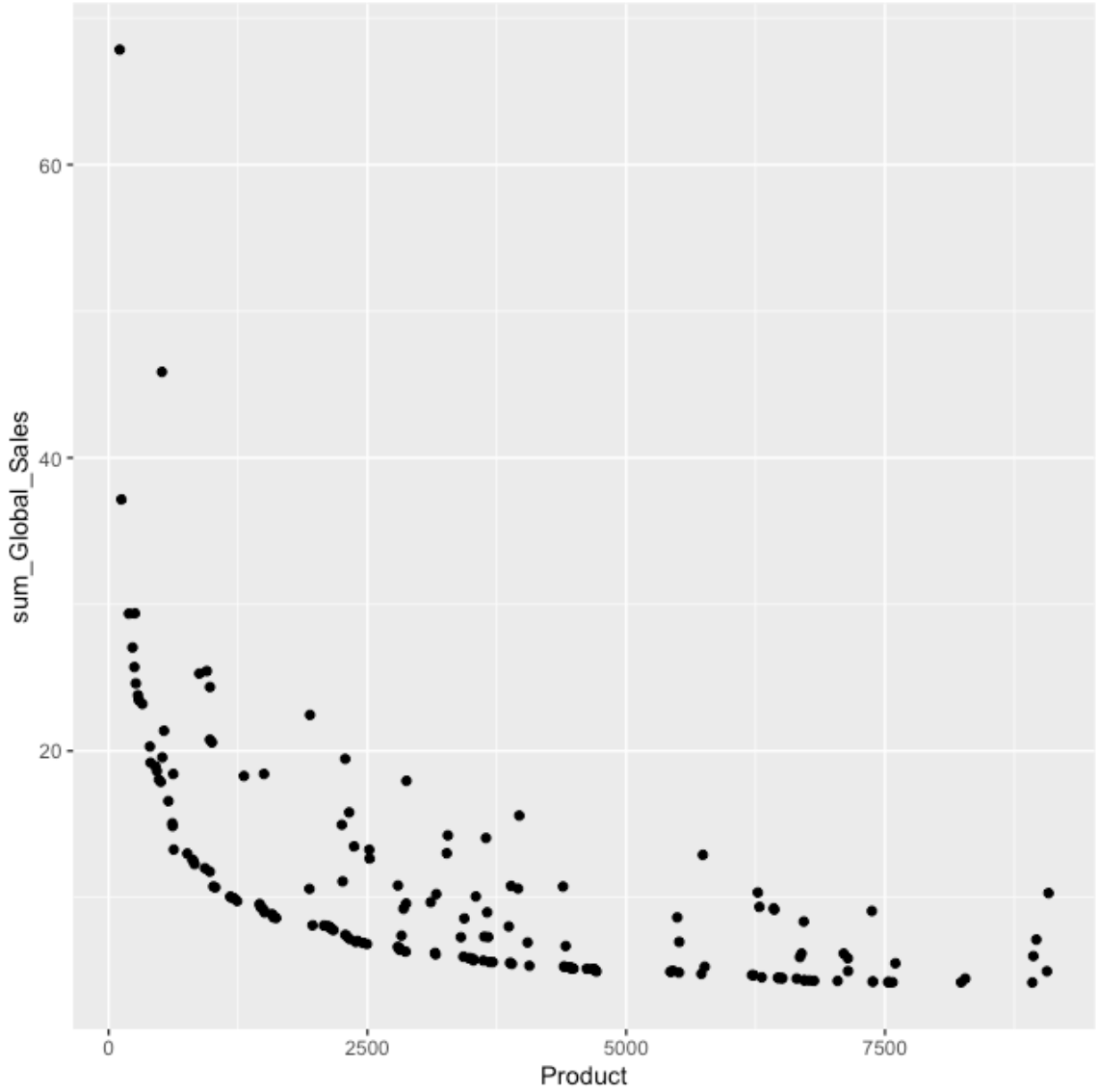
The sales data is then further grouped by product id, so we can get the number of sales per product.
175 rows of data after grouping. Basic statistics as follows.

Table5. Descriptive statistics of Sales by Product

```
> summary(df_product)
    Product       sum_NA_Sales     sum_EU_Sales    sum_Global_Sales
 Min.   : 107   Min.   : 0.060   Min.   : 0.000   Min.   : 4.200
 1st Qu.:1468   1st Qu.: 2.495   1st Qu.: 1.460   1st Qu.: 5.515
 Median :3158   Median : 3.610   Median : 2.300   Median : 8.090
 Mean   :3490   Mean   : 5.061   Mean   : 3.306   Mean   :10.730
 3rd Qu.:5442   3rd Qu.: 5.570   3rd Qu.: 4.025   3rd Qu.:12.785
 Max.   :9080   Max.   :34.020   Max.   :23.800   Max.   :67.850
```

The statistics show that NA has more sales opportunities in NA then in EU. Since Global sales is the sum of NA, EU and other area, it shows here that “other area” is also contributing a significant part in total sales and we may consider to separate it out as well in future analysis.

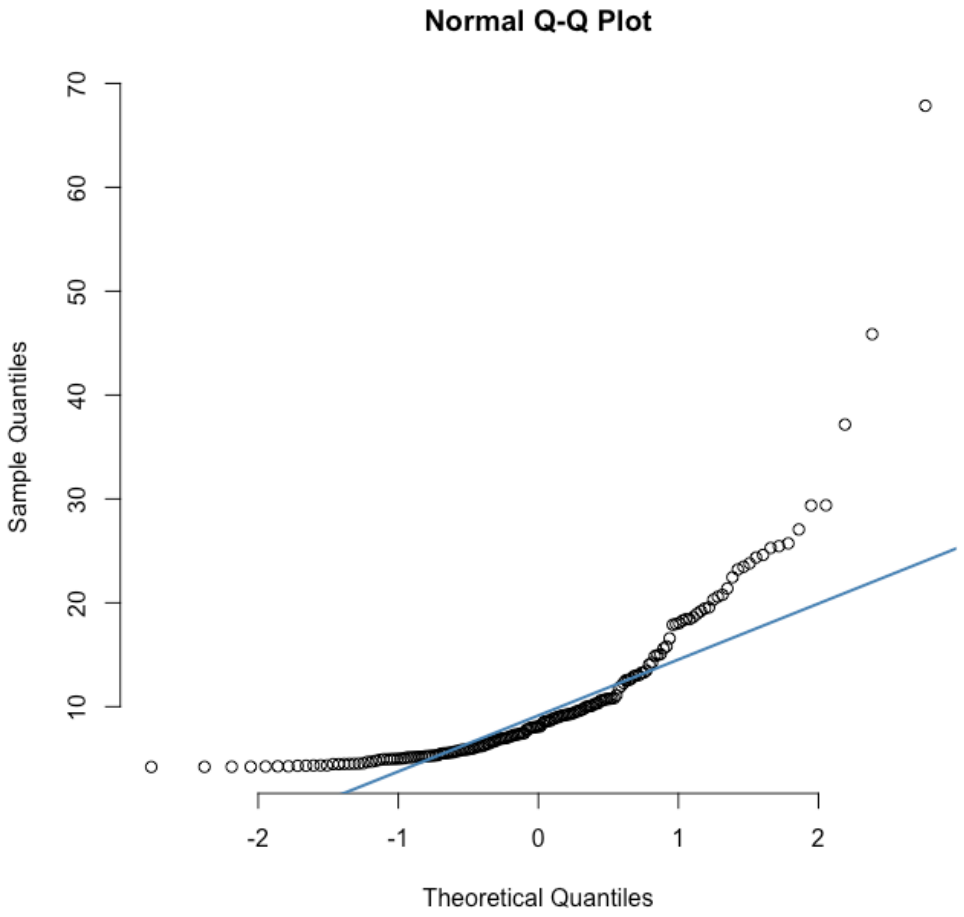
Fig14. Global Sales vs Product



It's interesting to see that when the global sales plotted against product there is a clear trend showing the bigger the product id number the less number of sales in most cases. This could be due to the bigger the product id number means a later release of the game hence a lower sale due to the timing issue. However, this could also hint that the new games are less popular or customers nowadays are turning to mobile/ online games and spending less money on the traditional game products.

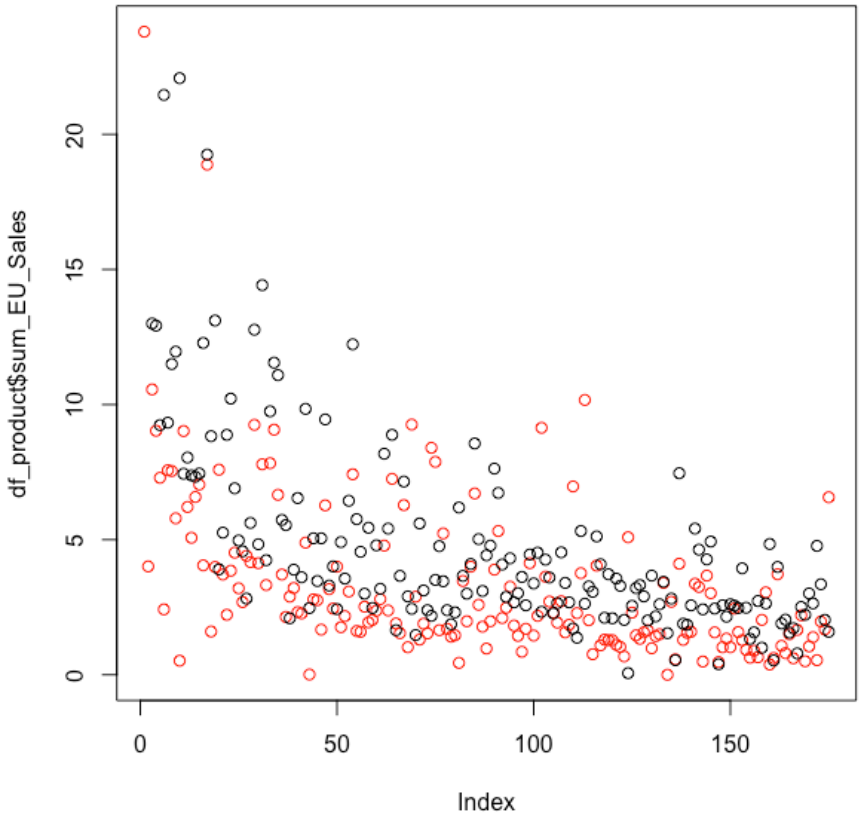
Checking normality, skewness and kurtosis of sales data

Fig15. Q-Q Plot of Global Sales



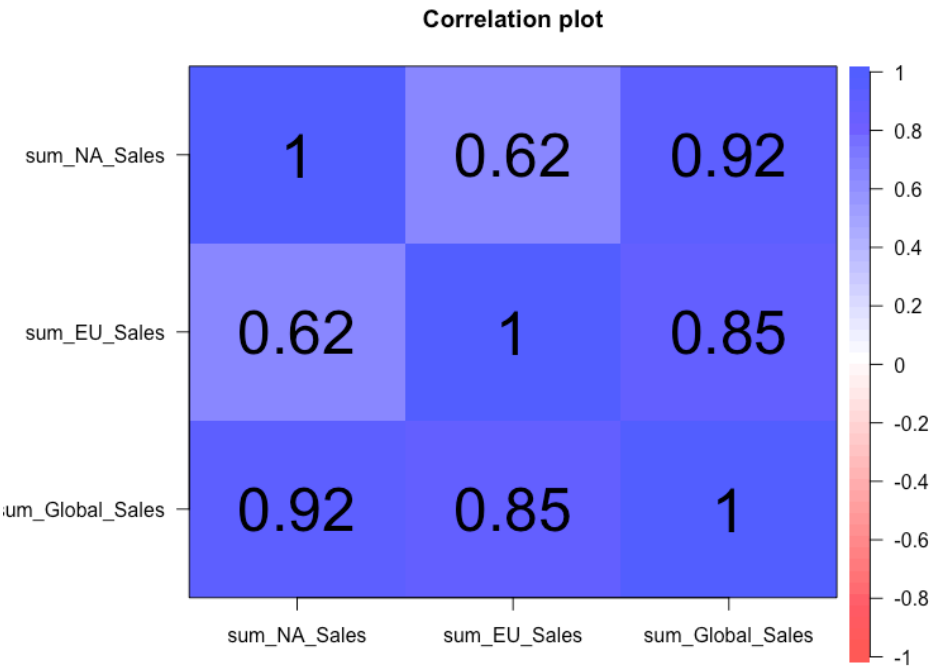
Q-Q plots are used to determine the normality of the sales data. Results, as in Fig15, suggest that the sales data is not normally distributed. Shapiro Wilk test also gives the same result with $p<0.05$. Skewness is found and Kurtosis is also significantly high, around 15-17 across all sets of sales data, indicating fat tails in the distribution.

Fig16. EU Sales (red) and NA Sales (black) per product



We can see from Fig16 that NA tends to have higher sales than EU, but the range of max and min is also larger while most products are sold with a smaller number of sales.

At last, we will investigate the correlation between each sales data, EU vs NA vs Global.
A correlation plot is created in Fig17.



It is not surprising to see both NA and EU highly correlated with Global Sales, as they are part of the total of Global Sales. NA and EU also strongly positively correlated. As mentioned previously, in future analysis we can further study the relationship between NA, EU and other area.

Since they are all correlated, we can use multiple regression model to predict furture sales. Model result gives following coefficients:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.04242    0.17736   5.877 2.11e-08 ***
sum_NA_Sales  1.13040    0.03162  35.745  < 2e-16 ***
sum_EU_Sales  1.19992    0.04672  25.682  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.49 on 172 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.9664
F-statistic:  2504 on 2 and 172 DF,  p-value: < 2.2e-16
```

Both t-test and f-test result suggests the regression model is reliable to use.

Example to predict Global Sales.

```
> NA_Sales <- c(34.02,3.93,2.73,2.26,22.08)
> EU_Sales <- c(23.80,1.56,0.65,0.97,0.52)
> Global_Sales <- 1.0424 + 1.13040*NA_Sales + 1.9992*EU_Sales
> Global_Sales
[1] 87.079568  8.603624  5.427872  5.536328 27.041216
```

The correlation between EU, NA and Global sales suggests that they are also positively correlated to other area sales. However, using two areas of sales data to predict one other area of sales does not seems to be an efficient method.

It would be more effective and accurate if further study could identify other factors impacting each area's sales and predict each area of sales independently.

-End-