LSE Data Analytics Online Career Accelerator

DA301 Assignment: Predicting future outcomes

Background:

Turtle Games is a global game manufacturer and retailer who manufactures and sells its own products, along with sourcing and selling products manufactured by other companies. The company collects data from sales as well as customer reviews.

To improve overall sales performance, Turtle Games would like to understand:

I. how customers accumulate loyalty points

II. how groups within the customer base can be used to target specific market segments

III. how social data (e.g. customer reviews) can be used to inform marketing campaigns

IV. the impact that each product has on sales

V. how reliable the data is (e.g., normal distribution, skewness, or kurtosis)

VI. what the relationship(s) is/are (if any) between North American, European, and global sales

Approach:

1. Prepare a GitHub repository:

https://github.com/chiusinchun/Chiu_SinChun_DA301_Assignment.git

As a part of a team of data analysts contracted by Turtle Games, share the working progress in GitHub is a efficient way to collaborate with teammate.

2. Import and explore the data

Turtle Games provided the team with two CSV files data sets which is turtle_reviews and turtle_sales, accompany a metadata txt file.

Python and R will be utilised for data analysis.

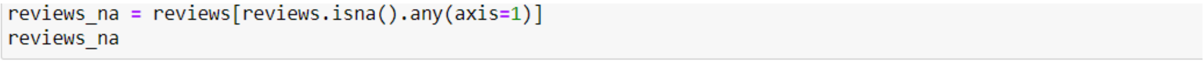
3. Analysis

I. How customers accumulate loyalty points:

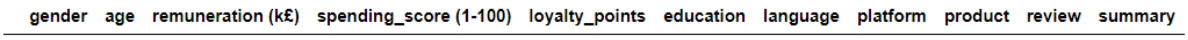
This question can be investigated by utilising Jupyter Notebook and open a new Python3 file. First step is to import the Python libraries and packages: Numpy, Pandas, Matplotlib, Seaborn, Statsmodels.api and Statsmodels.formula.api , in order to perform linear regression.

Load turtle_reviews.csv file provided by Turtle Games and create a new DataFrame named reviews.

Confirm there is no missing values:

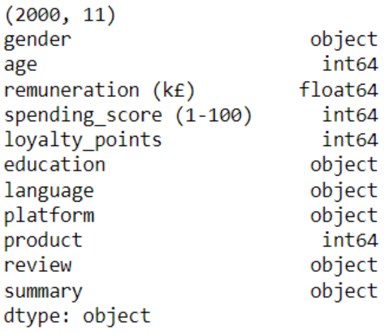


```
reviews_na = reviews[reviews.isna().any(axis=1)]
reviews_na
```

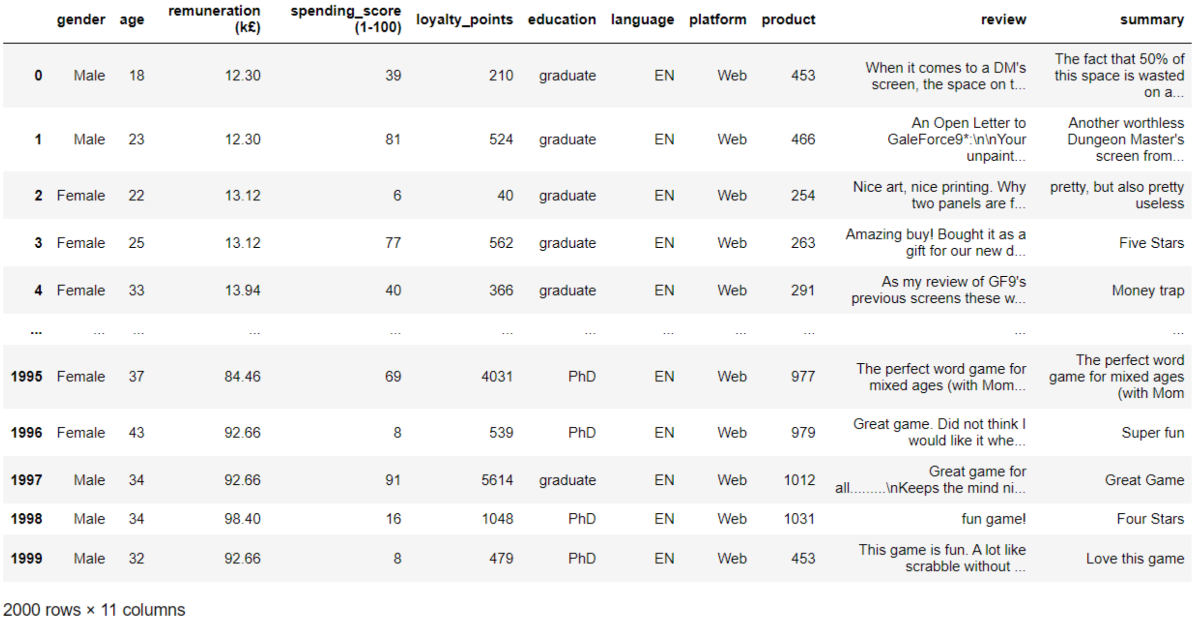


	gender	age	remuneration (k£)	spending_score (1-100)	loyalty_points	education	language	platform	product	review	summary

Explore the data by shape, dtypes, and view the DataFrame:



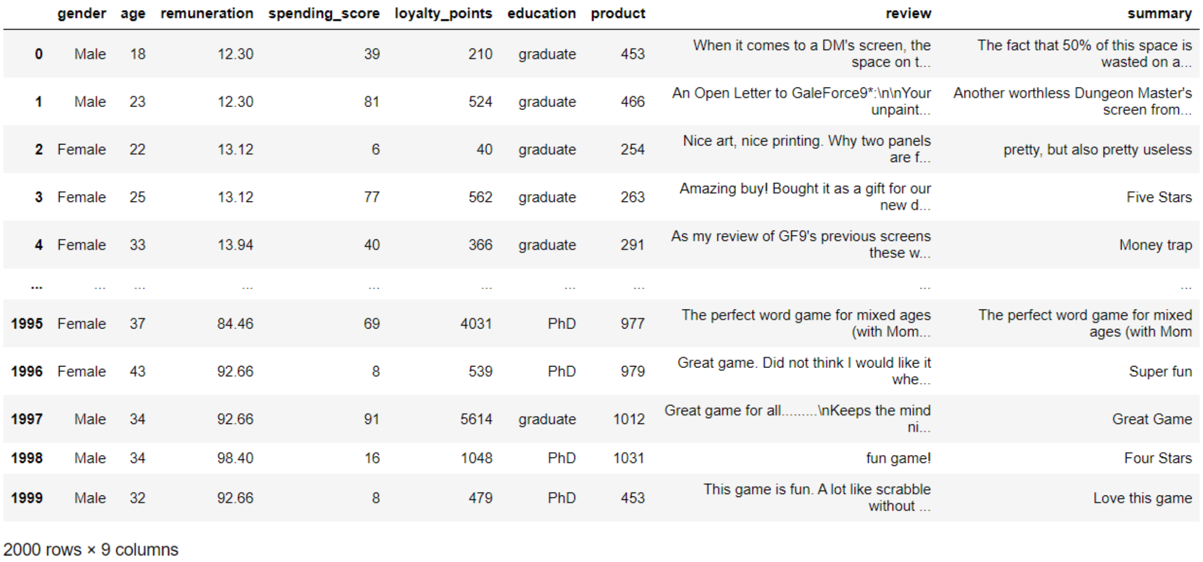
```
(2000, 11)
gender                      object
age                          int64
remuneration (k£)          float64
spending_score (1-100)       int64
loyalty_points               int64
education                   object
language                    object
platform                    object
product                      int64
review                      object
summary                     object
dtype: object
```



	gender	age	remuneration (k£)	spending_score (1-100)	loyalty_points	education	language	platform	product	review	summary
0	Male	18	12.30	39	210	graduate	EN	Web	453	When it comes to a DM's screen, the space on t...	The fact that 50% of this space is wasted on a...
1	Male	23	12.30	81	524	graduate	EN	Web	466	An Open Letter to GaleForce9*:\n\nYour unpaint...	Another worthless Dungeon Master's screen from...
2	Female	22	13.12	6	40	graduate	EN	Web	254	Nice art, nice printing. Why two panels are f...	pretty, but also pretty useless
3	Female	25	13.12	77	562	graduate	EN	Web	263	Amazing buy! Bought it as a gift for our new d...	Five Stars
4	Female	33	13.94	40	366	graduate	EN	Web	291	As my review of GF9's previous screens these w...	Money trap
...	...	...	...	...	...	...	...	...	...	...	...
1995	Female	37	84.46	69	4031	PhD	EN	Web	977	The perfect word game for mixed ages (with Mom...	The perfect word game for mixed ages (with Mom
1996	Female	43	92.66	8	539	PhD	EN	Web	979	Great game. Did not think I would like it whe...	Super fun
1997	Male	34	92.66	91	5614	graduate	EN	Web	1012	Great game for all.........\nKeeps the mind ni...	Great Game
1998	Male	34	98.40	16	1048	PhD	EN	Web	1031	fun game!	Four Stars
1999	Male	32	92.66	8	479	PhD	EN	Web	453	This game is fun. A lot like scrabble without ...	Love this game

2000 rows × 11 columns

For easier to reference, use drop and rename function to remove redundant columns and rename some headings, then save the cleaned DataFrame and import the file back to sense-check:



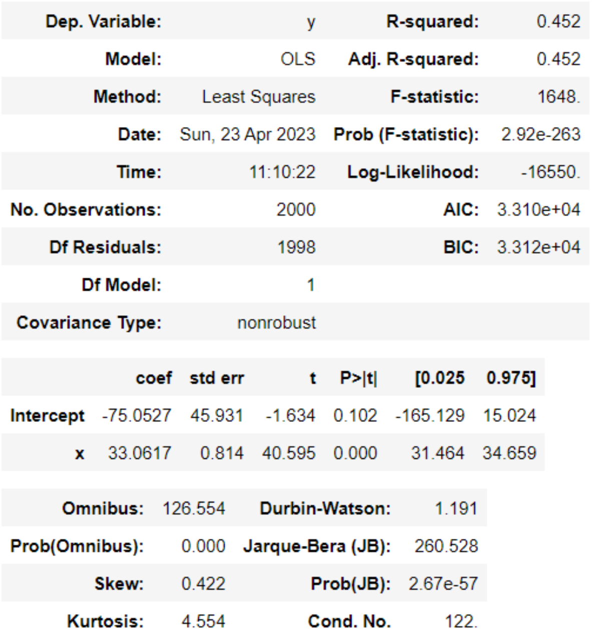
	gender	age	remuneration	spending_score	loyalty_points	education	product	review	summary
0	Male	18	12.30	39	210	graduate	453	When it comes to a DM's screen, the space on t...	The fact that 50% of this space is wasted on a...
1	Male	23	12.30	81	524	graduate	466	An Open Letter to GaleForce9*:\n\nYour unpaint...	Another worthless Dungeon Master's screen from...
2	Female	22	13.12	6	40	graduate	254	Nice art, nice printing. Why two panels are f...	pretty, but also pretty useless
3	Female	25	13.12	77	562	graduate	263	Amazing buy! Bought it as a gift for our new d...	Five Stars
4	Female	33	13.94	40	366	graduate	291	As my review of GF9's previous screens these w...	Money trap
...	...	...	...	...	...	...	...	...	...
1995	Female	37	84.46	69	4031	PhD	977	The perfect word game for mixed ages (with Mom...	The perfect word game for mixed ages (with Mom
1996	Female	43	92.66	8	539	PhD	979	Great game. Did not think I would like it whe...	Super fun
1997	Male	34	92.66	91	5614	graduate	1012	Great game for all.........\nKeeps the mind ni...	Great Game
1998	Male	34	98.40	16	1048	PhD	1031	fun game!	Four Stars
1999	Male	32	92.66	8	479	PhD	453	This game is fun. A lot like scrabble without ...	Love this game

2000 rows × 9 columns

To answer the question about loyalty points, we will apply linear regression model by setting loyalty points as dependent variable against spending, remuneration, and age as independent variable.

a. **Spending score as independent variable vs loyalty points as dependent variable**

Setting x as spending score and y as loyalty points, we formulate ordinary least squares method for linear regression model: f='y~x', then use summary function to print the OLS regression result:

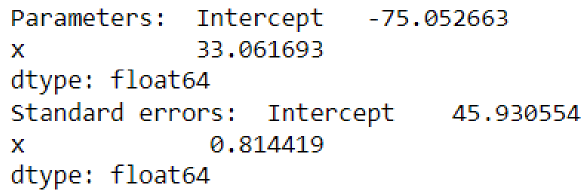


Dep. Variable:	y	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.
Date:	Sun, 23 Apr 2023	**Prob (F-statistic):**	2.92e-263
Time:	11:10:22	**Log-Likelihood:**	-16550.
No. Observations:	2000	**AIC:**	3.310e+04
Df Residuals:	1998	**BIC:**	3.312e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-75.0527	45.931	-1.634	0.102	-165.129	15.024
x	33.0617	0.814	40.595	0.000	31.464	34.659

Omnibus:	126.554	**Durbin-Watson:**	1.191
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	260.528
Skew:	0.422	**Prob(JB):**	2.67e-57
Kurtosis:	4.554	**Cond. No.**	122.

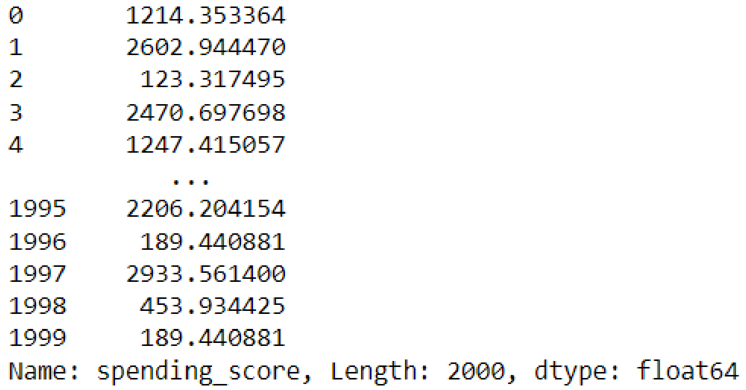
Then extract the parameters, standard errors:



```
Parameters:  Intercept   -75.052663
x            33.061693
dtype: float64
Standard errors:  Intercept    45.930554
x             0.814419
dtype: float64
```

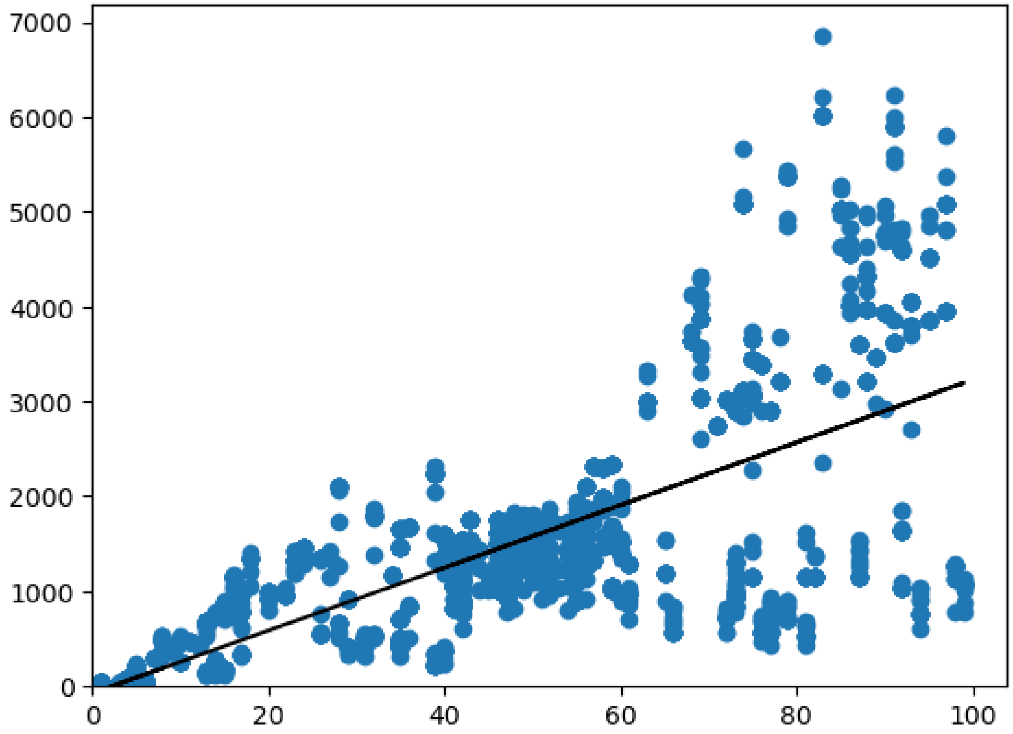
Set the x coefficient and the constant to formulate the regression table:

y_pred = (-75.052663)+33.061693*x



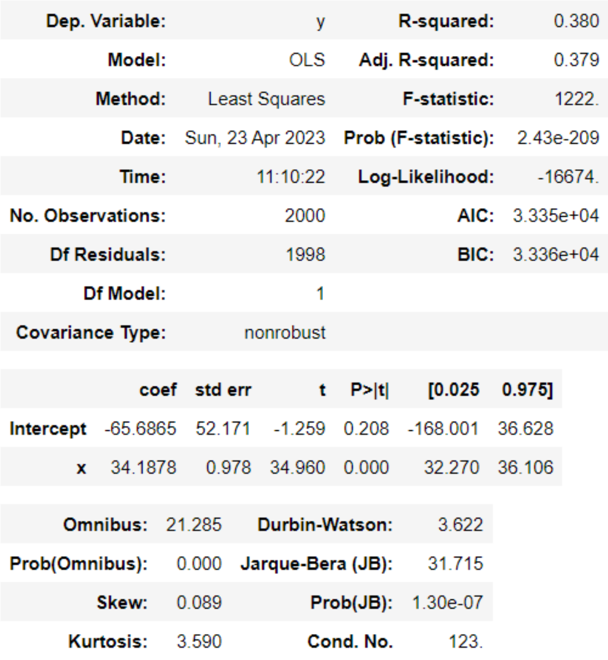
```
0       1214.353364
1       2602.944470
2        123.317495
3       2470.697698
4       1247.415057
           ...
1995    2206.204154
1996     189.440881
1997    2933.561400
1998     453.934425
1999     189.440881
Name: spending_score, Length: 2000, dtype: float64
```

And plot the graph with a regression line:



b. **Remuneration as independent variable vs loyalty points as dependent variable**

We do similar process setting remuneration as independent variable and loyalty points as dependent variable, the OLS regression results is:

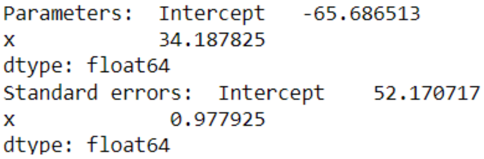


Dep. Variable:	y	**R-squared:**	0.380
Model:	OLS	**Adj. R-squared:**	0.379
Method:	Least Squares	**F-statistic:**	1222.
Date:	Sun, 23 Apr 2023	**Prob (F-statistic):**	2.43e-209
Time:	11:10:22	**Log-Likelihood:**	-16674.
No. Observations:	2000	**AIC:**	3.335e+04
Df Residuals:	1998	**BIC:**	3.336e+04
Df Model:	1		
Covariance Type:	nonrobust		

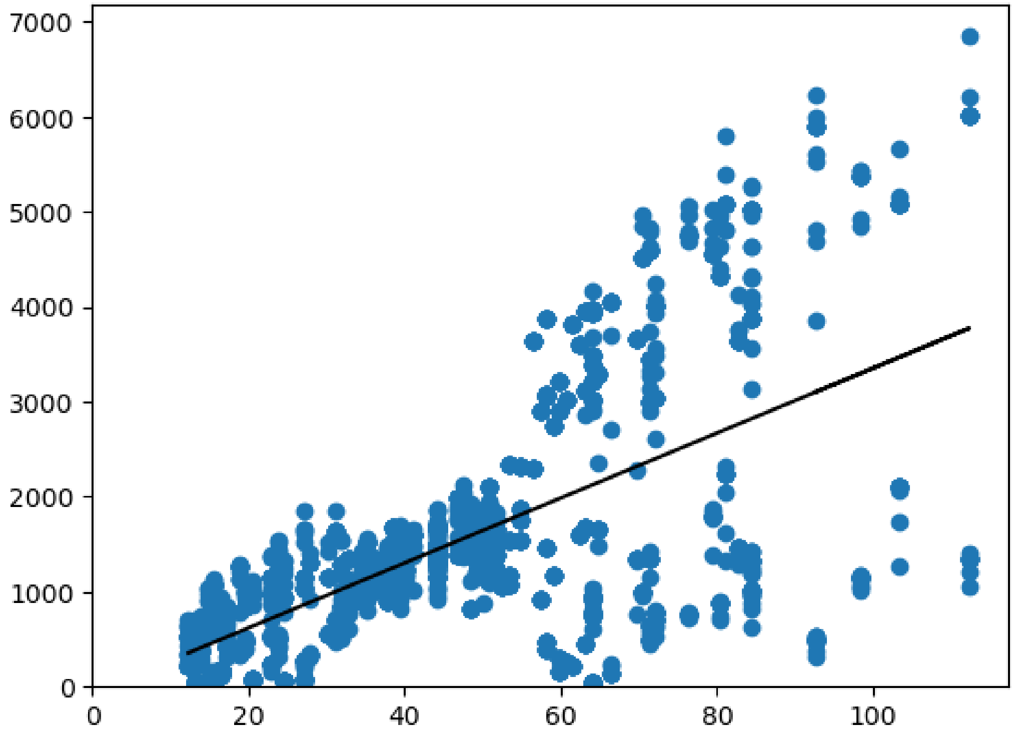
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-65.6865	52.171	-1.259	0.208	-168.001	36.628
x	34.1878	0.978	34.960	0.000	32.270	36.106

Omnibus:	21.285	**Durbin-Watson:**	3.622
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	31.715
Skew:	0.089	**Prob(JB):**	1.30e-07
Kurtosis:	3.590	**Cond. No.**	123.

Parameters and standard errors result:

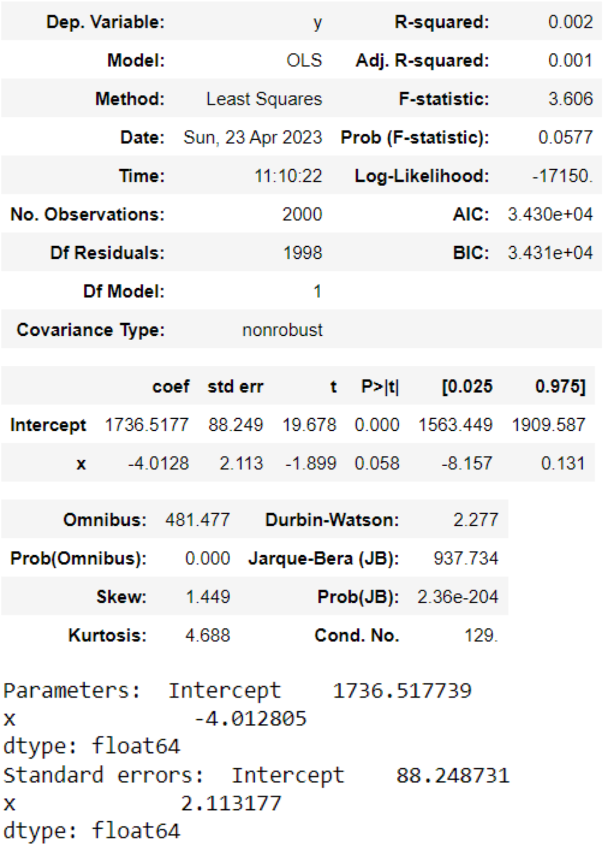


```
Parameters:  Intercept   -65.686513
x            34.187825
dtype: float64
Standard errors:  Intercept    52.170717
x             0.977925
dtype: float64
```



c. Age as independent variable vs loyalty points as dependent variable

Utilise OLS model again:

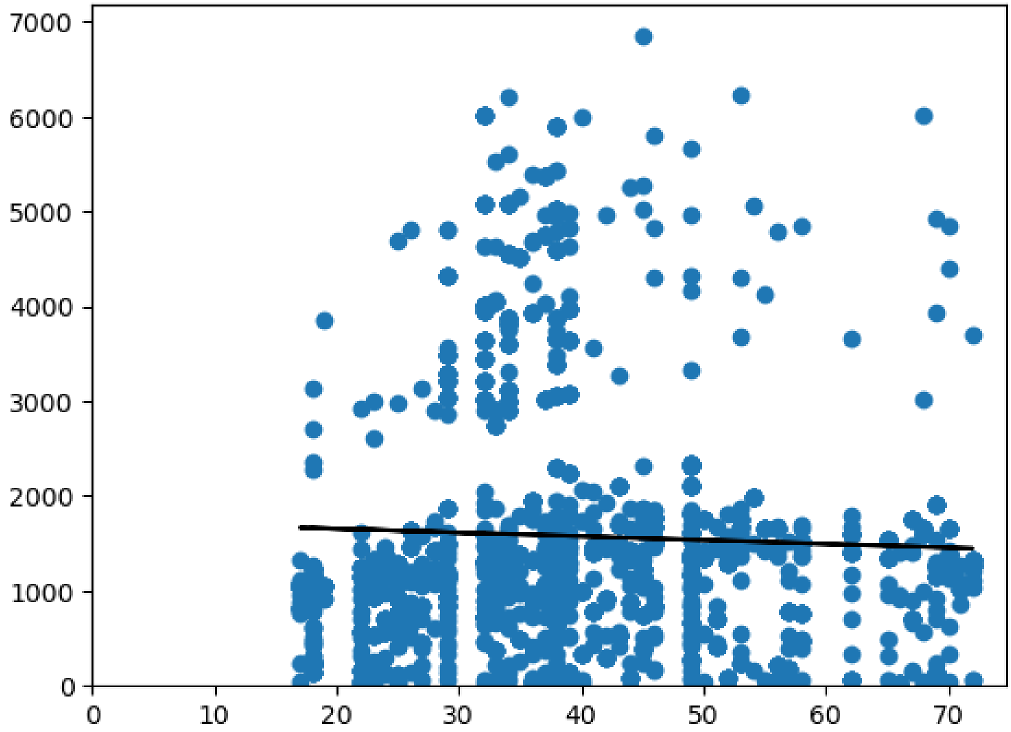


Dep. Variable:	y	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.606
Date:	Sun, 23 Apr 2023	**Prob (F-statistic):**	0.0577
Time:	11:10:22	**Log-Likelihood:**	-17150.
No. Observations:	2000	**AIC:**	3.430e+04
Df Residuals:	1998	**BIC:**	3.431e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1736.5177	88.249	19.678	0.000	1563.449	1909.587
x	-4.0128	2.113	-1.899	0.058	-8.157	0.131

Omnibus:	481.477	**Durbin-Watson:**	2.277
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	937.734
Skew:	1.449	**Prob(JB):**	2.36e-204
Kurtosis:	4.688	**Cond. No.**	129.

```
Parameters:  Intercept    1736.517739
x              -4.012805
dtype: float64
Standard errors:  Intercept    88.248731
x             2.113177
dtype: float64
```

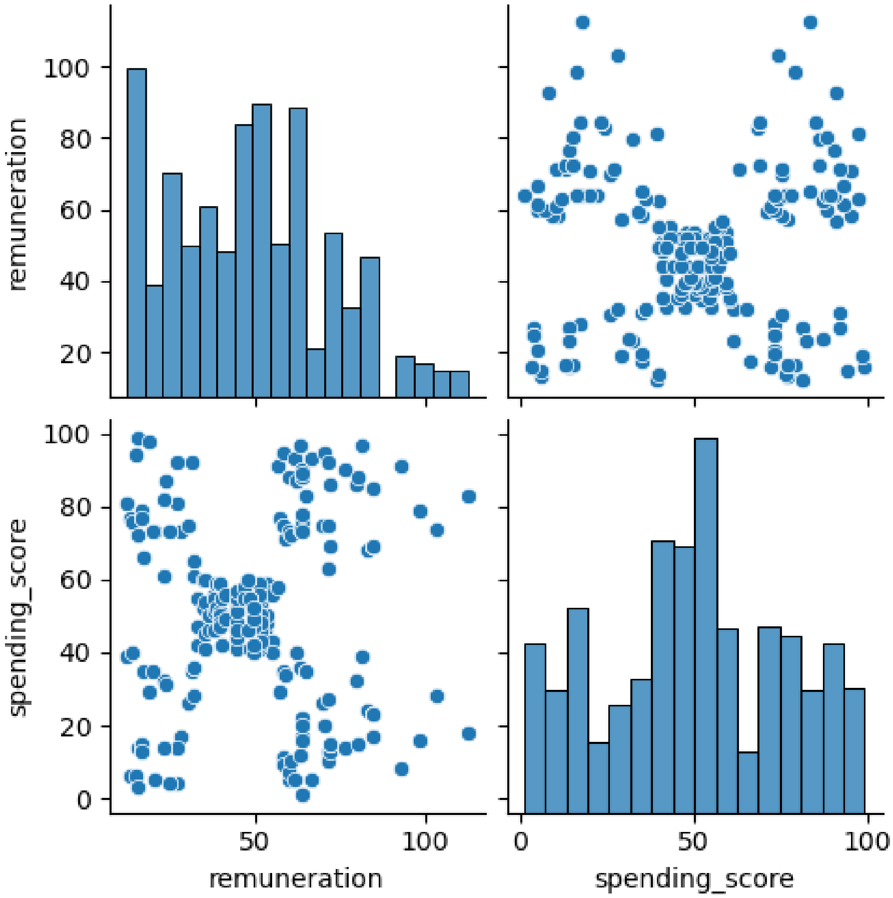


By reviewing the R-squared, which represents the proportion of the variance of the dependent variable can be explained by the independent variable, **spending score** is the highest amongst the 3 independent variables, showing it is the main attribute of loyalty points. Besides, Prob(F-statistic), i.e., p-value, is extremely small, indicates that spending score is statistically significant. In contrast, age R-squared is only 0.002 and p-value is 0.0577, reveal that there is no significant relationship between age and the loyalty points. The conclusion is also supported by visualization of the linear regression plots.

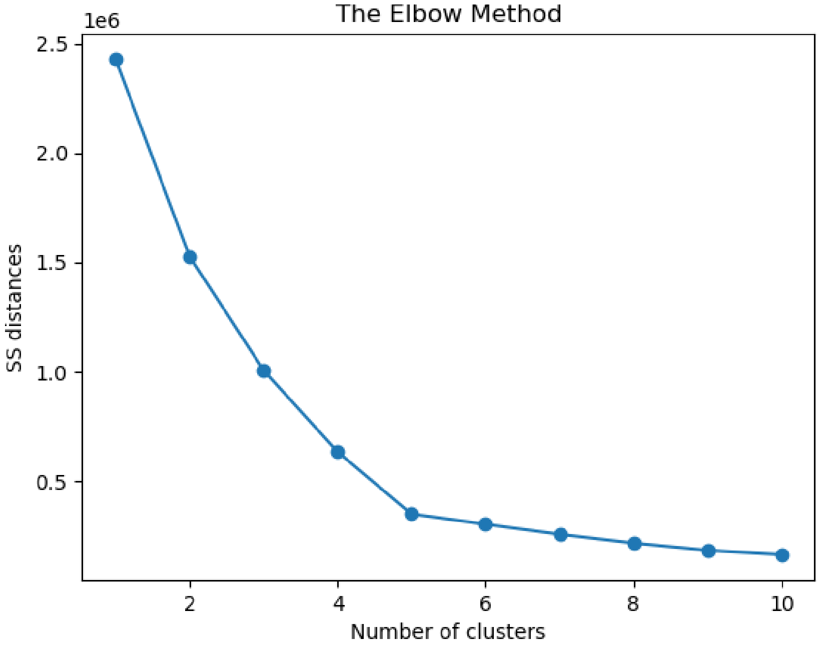
II. How groups within the customer base can be used to target specific market segments:

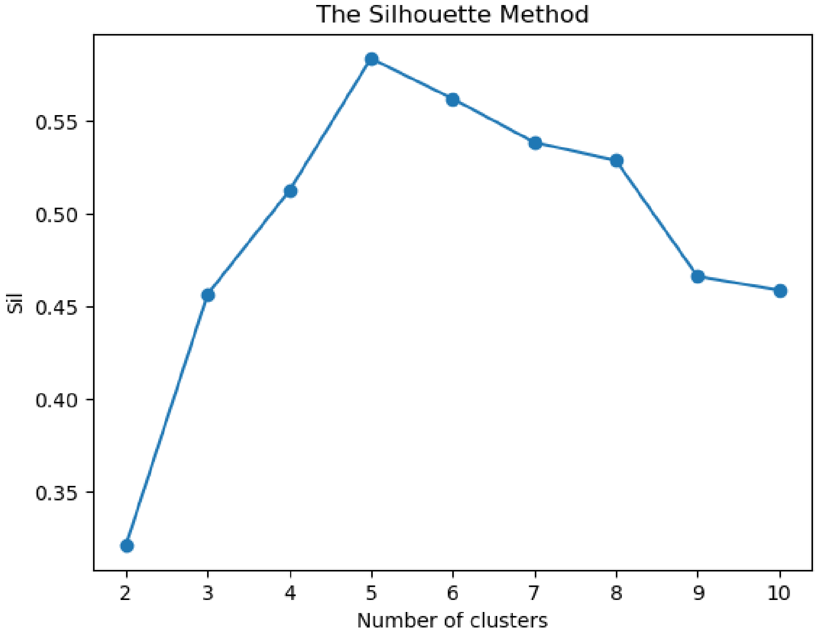
The best way to investigate this question is to use k-means clustering, we will import the library from Scikit-learn, including StandardScaler, KMeans, silhouette_score, accurary_sscore, and cdist. In this case, we group customer by remuneration against spending score to target specific market segments.

Firstly, we can have a general intuition by plotting a pairplot:



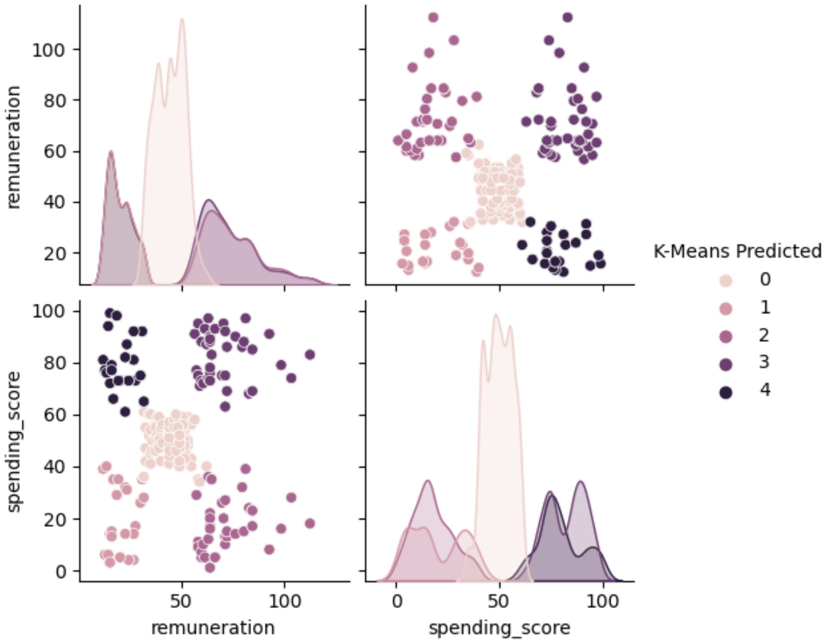
Then determine the number of clusters by utilising elbow and silhouette method:





In this case, the plot of elbow method is ambiguous to determine how many groups to be divided. However, silhouette method guidance is relatively clear, suggest to divide customers into 4 or 5 groups.

By evaluating k-means model at different values of k by visualization, 5 groups would be reasonable:

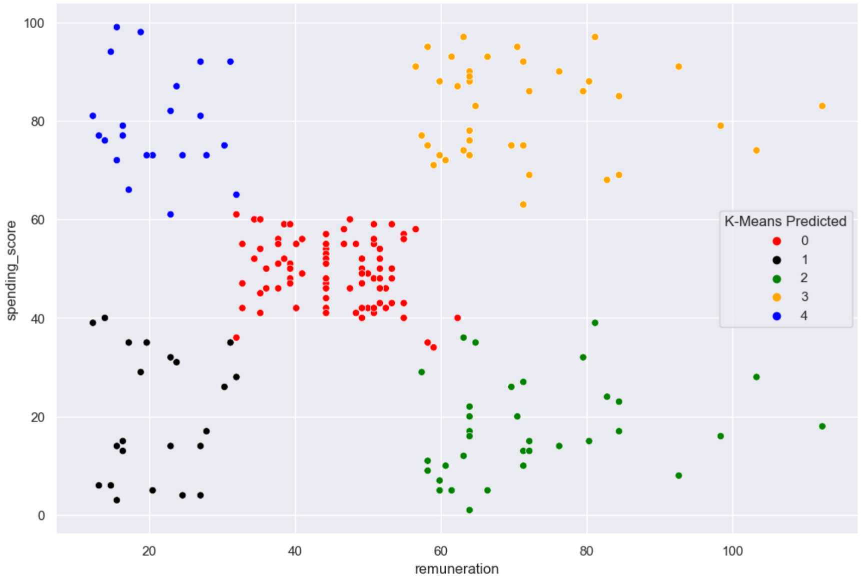


And the number of observations per predicted class would be:

```
0    774
3    356
2    330
1    271
4    269
Name: K-Means Predicted, dtype: int64
```

The 5 groups of customers should be:

1. low remuneration and spending
2. low remuneration but high spending
3. medium remuneration and spending
4. high remuneration and spending
5. high remuneration but low spending



III. How social data (e.g. customer reviews) can be used to inform marketing campaigns:

Natural Language Processing will be the perfect tool to utilise customer reviews to inform marketing campaigns.

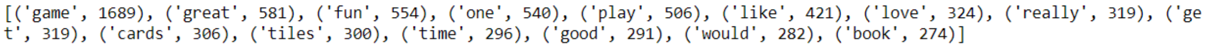
To implement NLP, more packages and library must be imported: e.g., nltk(Natural Languuage Toolkit), os, word_tokenize, sent_tokenize, FreqDist, stopwords, textblob and scipy.stats.

Create a new dataframe by only selecting the necessary columns: ‘review’ and ‘summary’. Sense check the dataframe and confirm there are no missing values.

The first step is the wrangling the data, change both the columns’ word to lower case, drop the duplicates, replace all the punctuation and join the elements with a space.

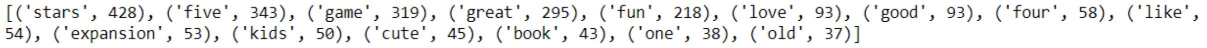
Then tokenise the word and identify 15 most common words:

Review:



```
[('game', 1689), ('great', 581), ('fun', 554), ('one', 540), ('play', 506), ('like', 421), ('love', 324), ('really', 319), ('ge
t', 319), ('cards', 306), ('tiles', 300), ('time', 296), ('good', 291), ('would', 282), ('book', 274)]
```

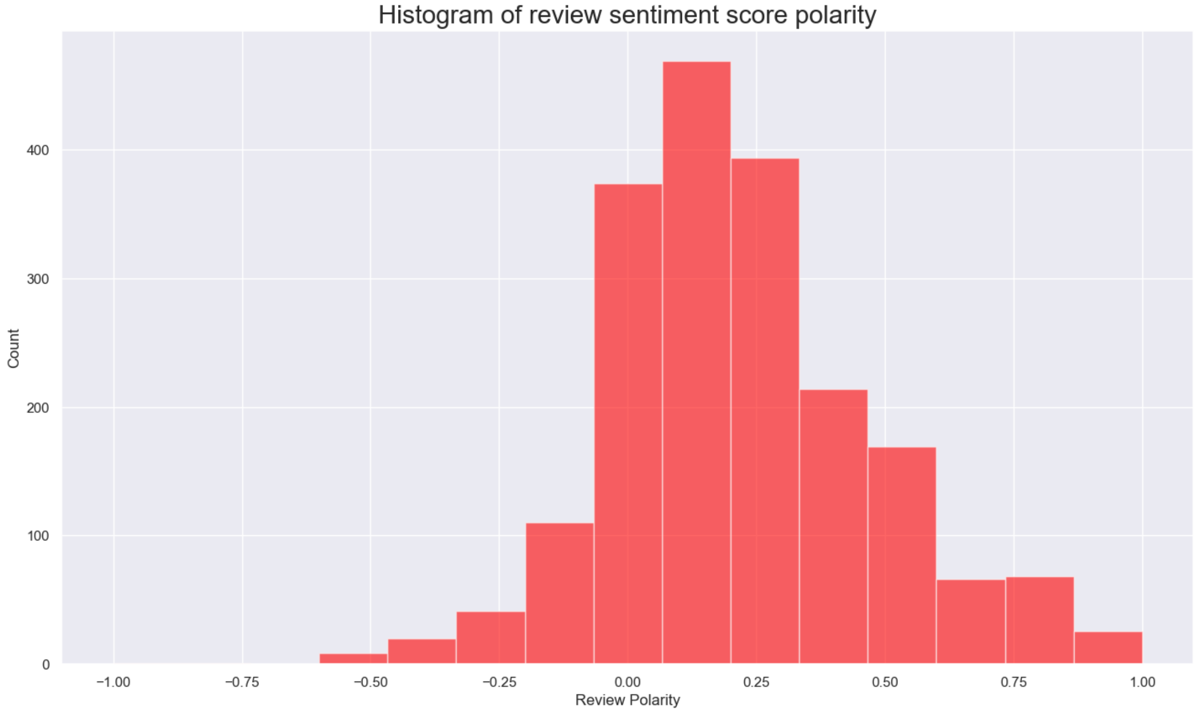
Summary:



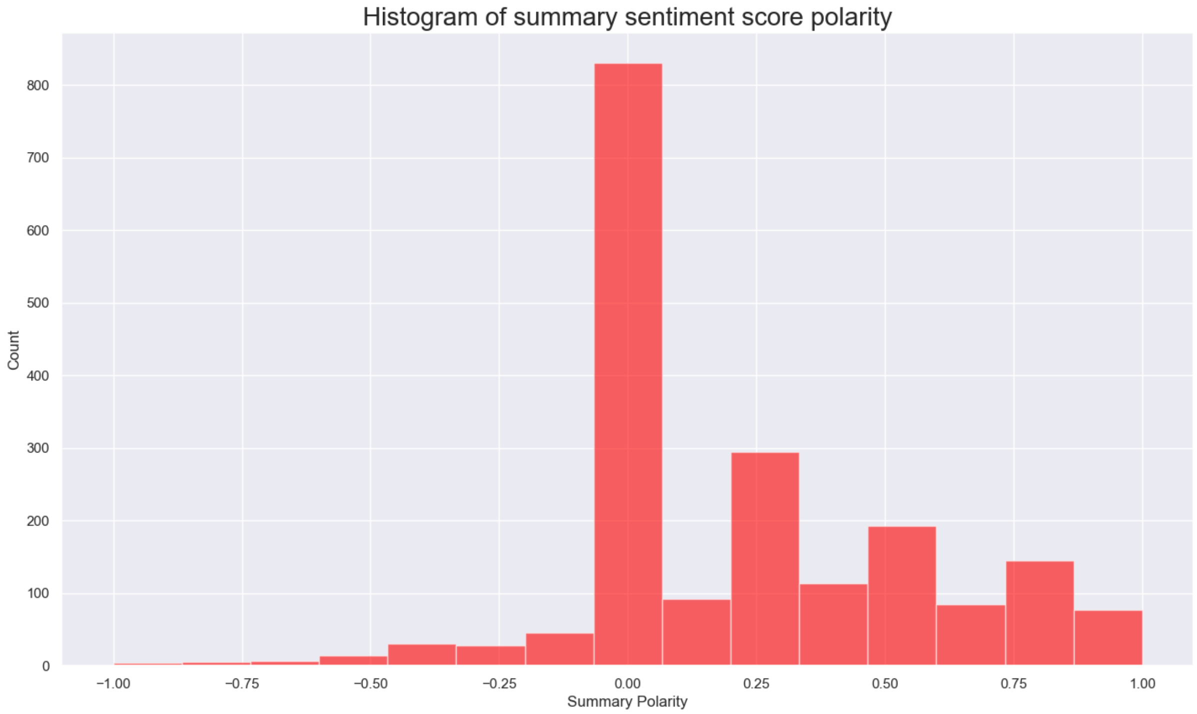
```
[('stars', 428), ('five', 343), ('game', 319), ('great', 295), ('fun', 218), ('love', 93), ('good', 93), ('four', 58), ('like',
54), ('expansion', 53), ('kids', 50), ('cute', 45), ('book', 43), ('one', 38), ('old', 37)]
```

By leveraging vaderSentiment library, we can analyse the sentiment of the feedback across products received from customers.

Review:



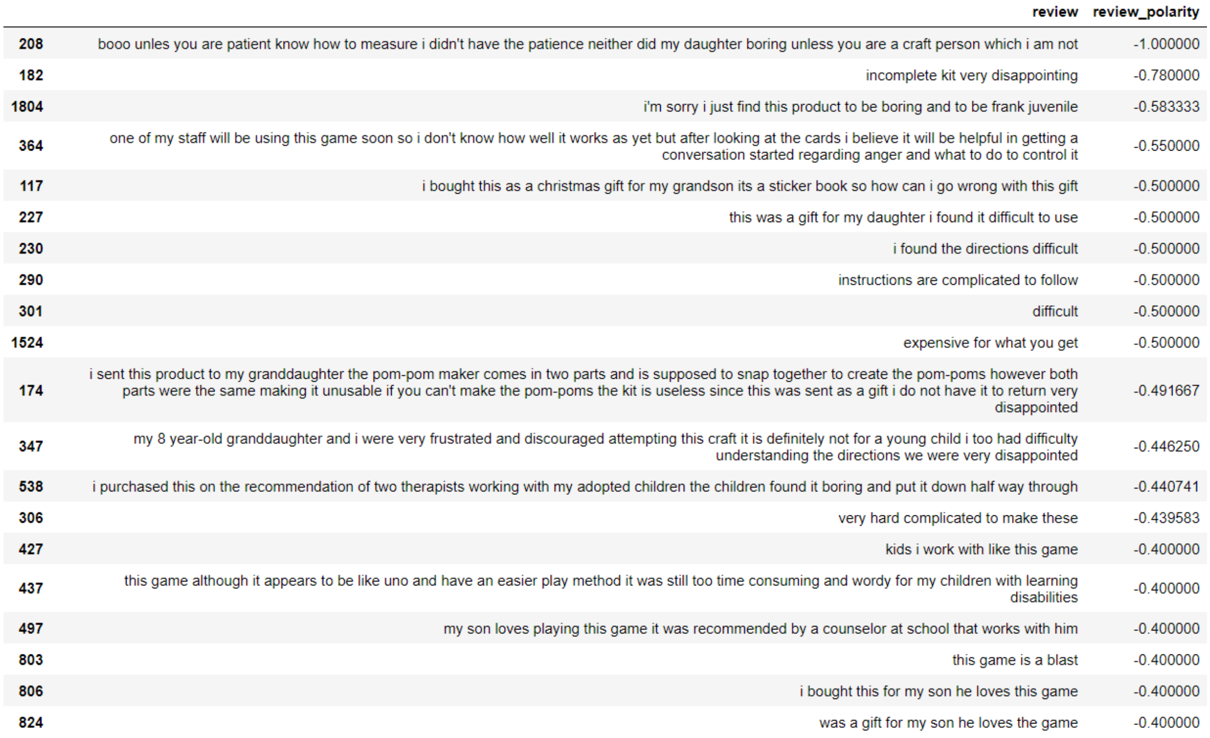
Summary:



Both review and summary skewed to the right, which means higher proportion of customers have positive impression on Turtle Games products.

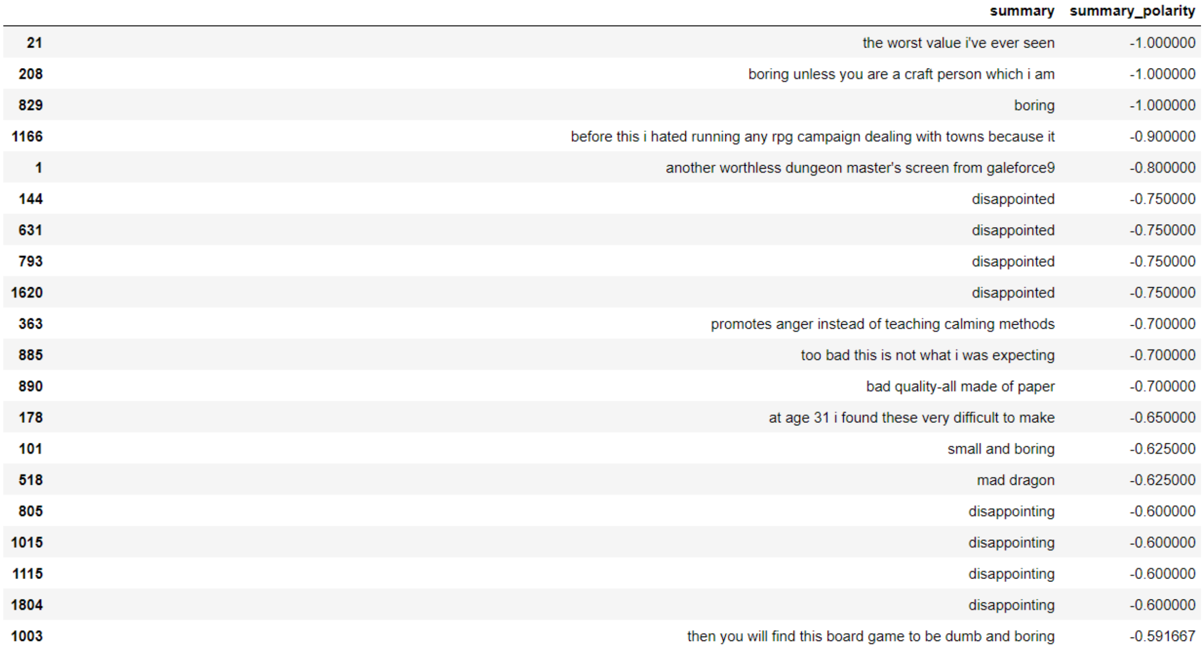
We can also identify top 20 negative and positive reviews and summaries respectively for maintaining customer loyalty and retention:

Top 20 review negative reviews:



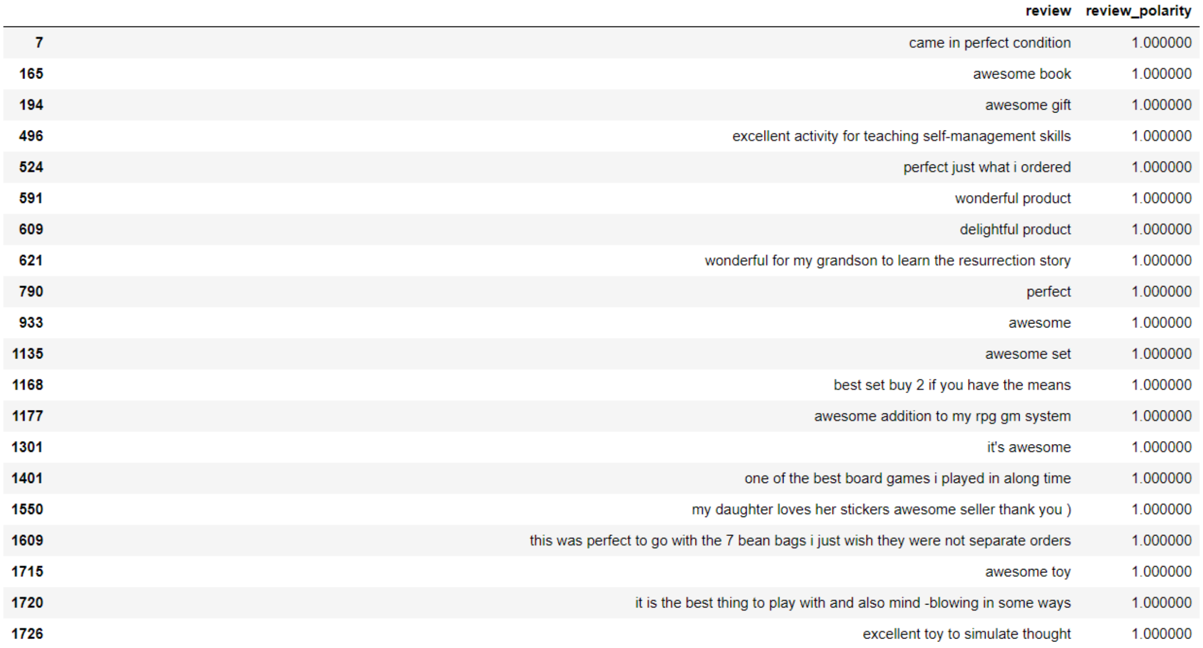
	review	review_polarity
208	booo unles you are patient know how to measure i didn't have the patience neither did my daughter boring unless you are a craft person which i am not	-1.000000
182	incomplete kit very disappointing	-0.780000
1804	i'm sorry i just find this product to be boring and to be frank juvenile	-0.583333
364	one of my staff will be using this game soon so i don't know how well it works as yet but after looking at the cards i believe it will be helpful in getting a conversation started regarding anger and what to do to control it	-0.550000
117	i bought this as a christmas gift for my grandson its a sticker book so how can i go wrong with this gift	-0.500000
227	this was a gift for my daughter i found it difficult to use	-0.500000
230	i found the directions difficult	-0.500000
290	instructions are complicated to follow	-0.500000
301	difficult	-0.500000
1524	expensive for what you get	-0.500000
174	i sent this product to my granddaughter the pom-pom maker comes in two parts and is supposed to snap together to create the pom-poms however both parts were the same making it unusable if you can't make the pom-poms the kit is useless since this was sent as a gift i do not have it to return very disappointed	-0.491667
347	my 8 year-old granddaughter and i were very frustrated and discouraged attempting this craft it is definitely not for a young child i too had difficulty understanding the directions we were very disappointed	-0.446250
538	i purchased this on the recommendation of two therapists working with my adopted children the children found it boring and put it down half way through	-0.440741
306	very hard complicated to make these	-0.439583
427	kids i work with like this game	-0.400000
437	this game although it appears to be like uno and have an easier play method it was still too time consuming and wordy for my children with learning disabilities	-0.400000
497	my son loves playing this game it was recommended by a counselor at school that works with him	-0.400000
803	this game is a blast	-0.400000
806	i bought this for my son he loves this game	-0.400000
824	was a gift for my son he loves the game	-0.400000

Top 20 summary negative summaries:



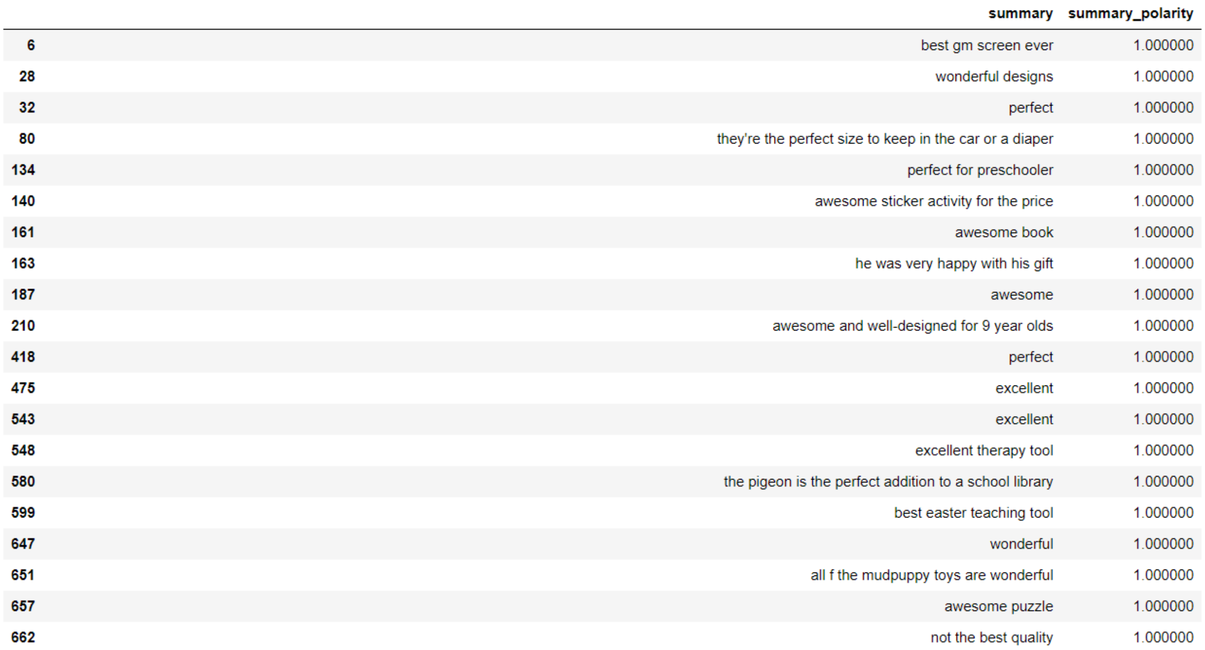
	summary	summary_polarity
21	the worst value i've ever seen	-1.000000
208	boring unless you are a craft person which i am	-1.000000
829	boring	-1.000000
1166	before this i hated running any rpg campaign dealing with towns because it	-0.900000
1	another worthless dungeon master's screen from galeforce9	-0.800000
144	disappointed	-0.750000
631	disappointed	-0.750000
793	disappointed	-0.750000
1620	disappointed	-0.750000
363	promotes anger instead of teaching calming methods	-0.700000
885	too bad this is not what i was expecting	-0.700000
890	bad quality-all made of paper	-0.700000
178	at age 31 i found these very difficult to make	-0.650000
101	small and boring	-0.625000
518	mad dragon	-0.625000
805	disappointing	-0.600000
1015	disappointing	-0.600000
1115	disappointing	-0.600000
1804	disappointing	-0.600000
1003	then you will find this board game to be dumb and boring	-0.591667

Top 20 review positive reviews:



	review	review_polarity
7	came in perfect condition	1.000000
165	awesome book	1.000000
194	awesome gift	1.000000
496	excellent activity for teaching self-management skills	1.000000
524	perfect just what i ordered	1.000000
591	wonderful product	1.000000
609	delightful product	1.000000
621	wonderful for my grandson to learn the resurrection story	1.000000
790	perfect	1.000000
933	awesome	1.000000
1135	awesome set	1.000000
1168	best set buy 2 if you have the means	1.000000
1177	awesome addition to my rpg gm system	1.000000
1301	it's awesome	1.000000
1401	one of the best board games i played in along time	1.000000
1550	my daughter loves her stickers awesome seller thank you)	1.000000
1609	this was perfect to go with the 7 bean bags i just wish they were not separate orders	1.000000
1715	awesome toy	1.000000
1720	it is the best thing to play with and also mind -blowing in some ways	1.000000
1726	excellent toy to simulate thought	1.000000

Top 20 summary positive reviews:



	summary	summary_polarity
6	best gm screen ever	1.000000
28	wonderful designs	1.000000
32	perfect	1.000000
80	they're the perfect size to keep in the car or a diaper	1.000000
134	perfect for preschooler	1.000000
140	awesome sticker activity for the price	1.000000
161	awesome book	1.000000
163	he was very happy with his gift	1.000000
187	awesome	1.000000
210	awesome and well-designed for 9 year olds	1.000000
418	perfect	1.000000
475	excellent	1.000000
543	excellent	1.000000
548	excellent therapy tool	1.000000
580	the pigeon is the perfect addition to a school library	1.000000
599	best easter teaching tool	1.000000
647	wonderful	1.000000
651	all f the mudpuppy toys are wonderful	1.000000
657	awesome puzzle	1.000000
662	not the best quality	1.000000

IV. The impact that each product has on sales

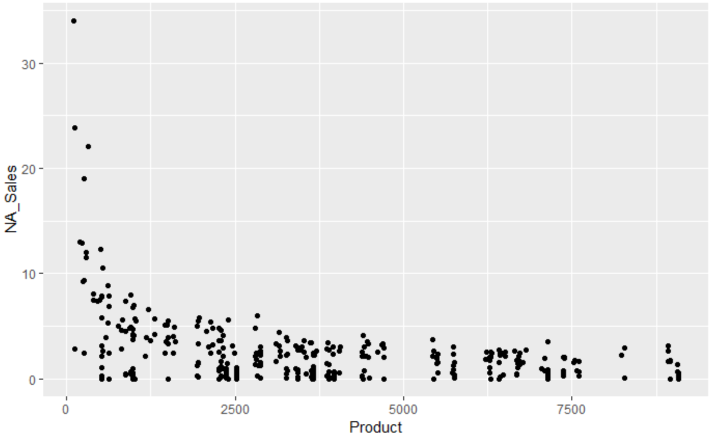
The best way to convey the impact of sales per product to the stack holder, the sales department of Turtle Games, by R is through visualisation.

To kick start R script, it is better to set up a working directory, make sure the data source is in the same directory with the working R script file, then import turtle_sales csv file and remove redundant columns, sense check the new data frame and view the descriptive statistics.

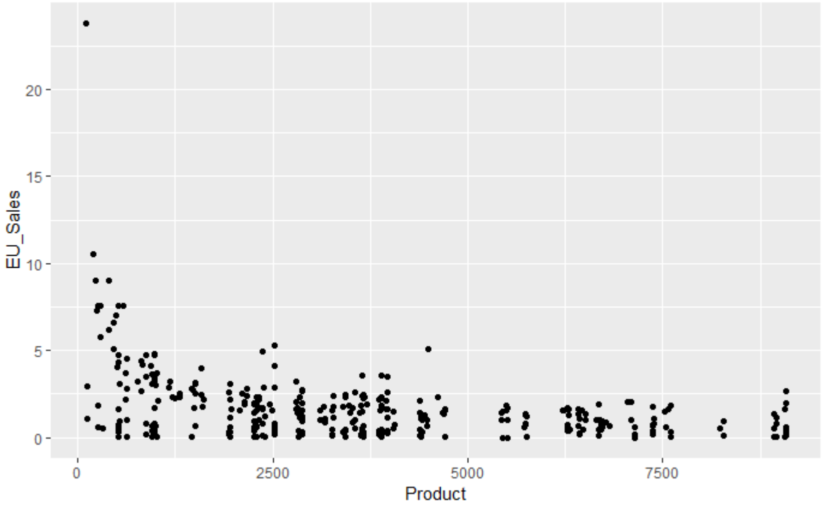
We will plot 3 types of common insightful graph to examine the patterns of sales per product:

a. Scatterplots

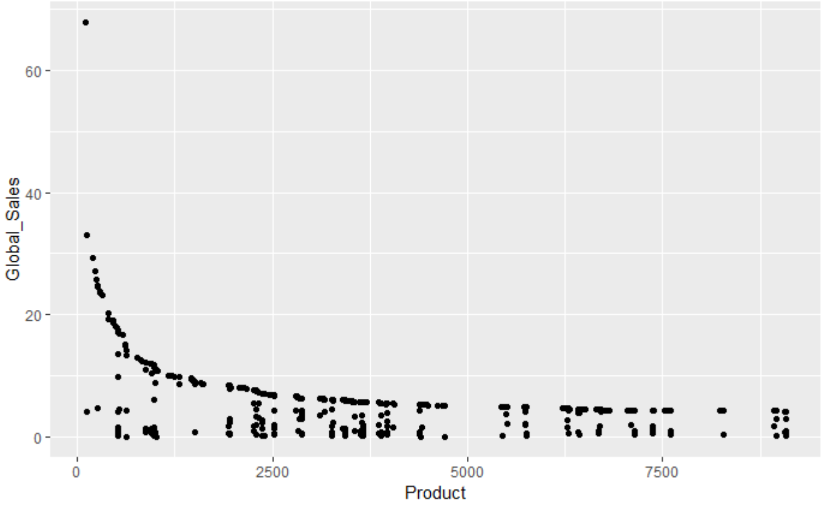
North American sales per product



European sales per product

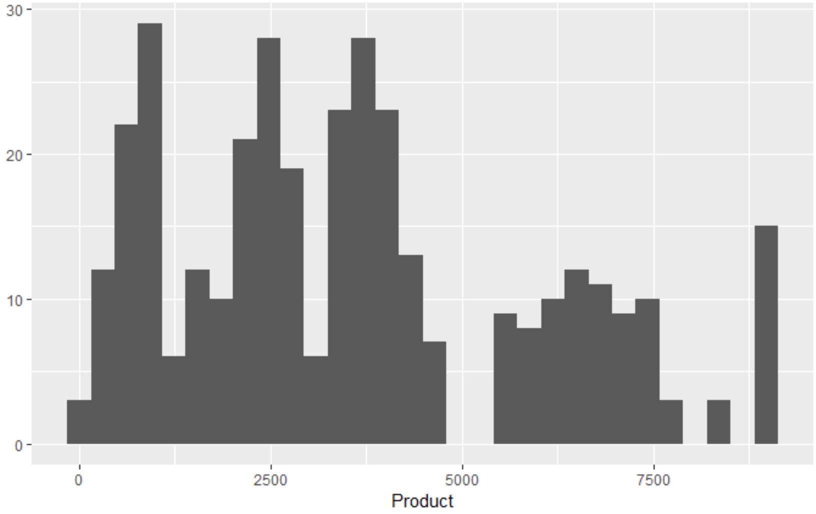


Global sales per product



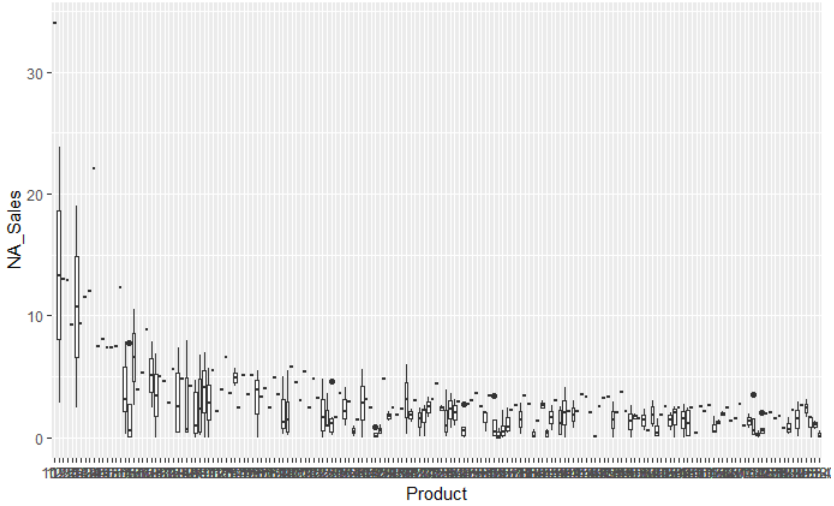
b. Histogram

Counts of each product

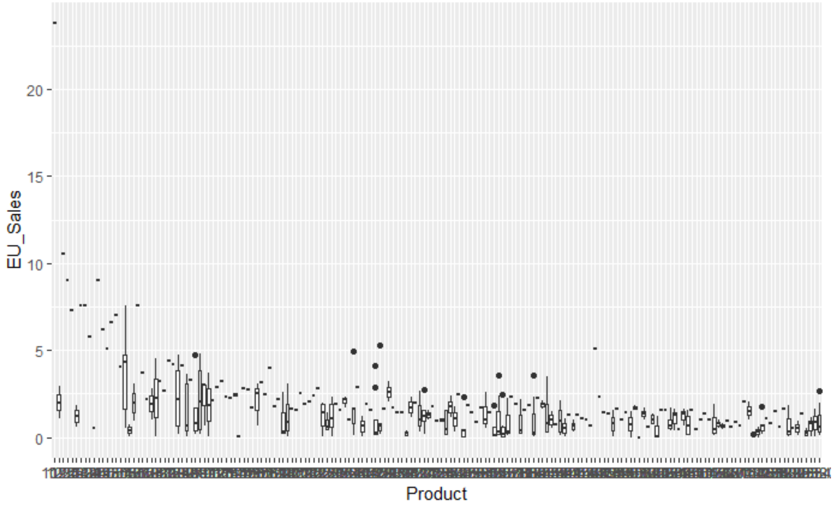


c. Boxplots

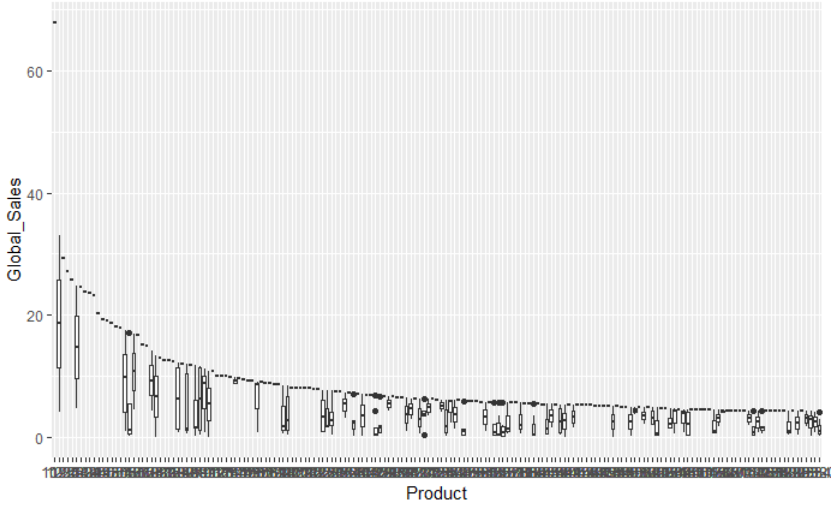
North American sales per product



European sales per product



Global sales per product



From these Plots, the first impression is that the highest sales value are from those product with smaller number.

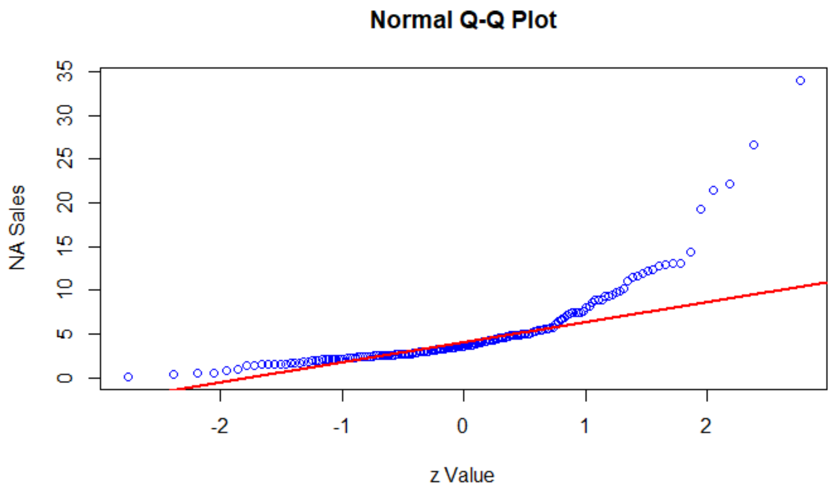
V. How reliable the data is (e.g. normal distribution, skewness, or kurtosis)

To examine how reliable are the data, we will utilise R library tidyverse, dplyr, skimr, moments to perform statistical test and visualization.

At the beginning, we need to load the data frame created in the previous question, then use the group_by and aggregate function, sum the values grouped by product, deduce the sum of North American sales, European sales and global sales. Besides, explore the descriptive statistics by skim and summary method.

To gain the first sight of normal distribution from visualisation, we can employ q-q plot:

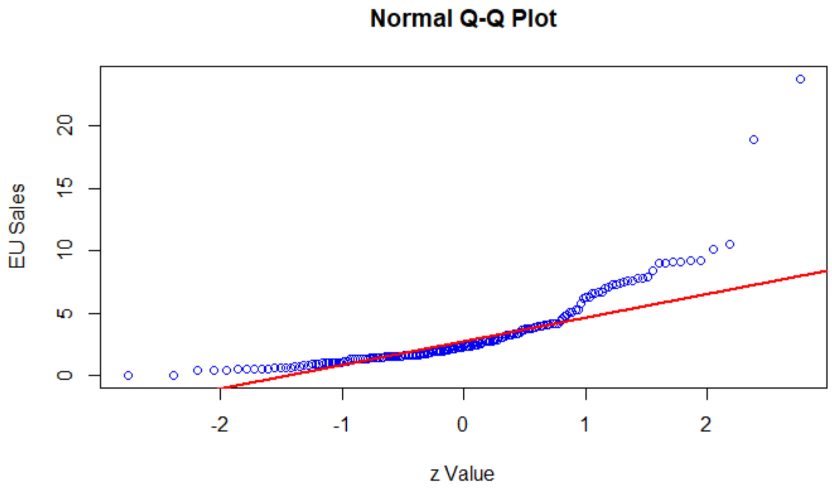
Sum of North American Sales:



```
> shapiro.test(turtle_sales3$sum_NA_Sales)

        Shapiro-Wilk normality test

data:  turtle_sales3$sum_NA_Sales
W = 0.69813, p-value < 2.2e-16


> skewness(turtle_sales3$sum_NA_Sales)
[1] 3.048198
> kurtosis(turtle_sales3$sum_NA_Sales)
[1] 15.6026
```

For the North American Sales Shapiro-Wilk test, p<0.05 indicates that the data is not normally distributed, the North American Sales is not likely normal distribution, W always falls between 0 and 1, here 0.69813, is not high, not probably come from a normal distribution, positive skewness skewed to the right, kurtosis is a lot larger than 3 indicate it has a sharp peak and fat tails

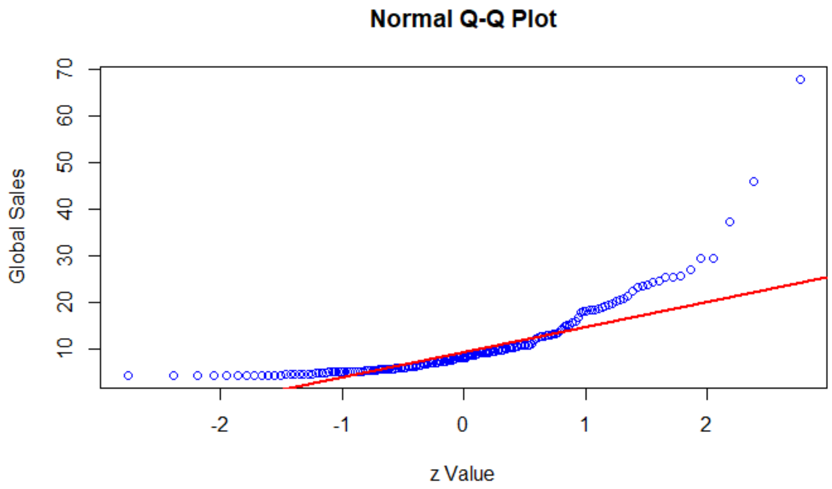
Sum of European Sales:



```
> shapiro.test(turtle_sales3$sum_EU_Sales)

	Shapiro-Wilk normality test

data:  turtle_sales3$sum_EU_Sales
W = 0.74058, p-value = 2.987e-16


> skewness(turtle_sales3$sum_EU_Sales)
[1] 2.886029
> kurtosis(turtle_sales3$sum_EU_Sales)
[1] 16.22554
```

For the European Sales Shapiro-Wilk test, p<0.05 indicates that the data is not normally distributed, the European Sales is not likely normal distribution, W always falls between 0 and 1, here 0.74058, is fair, but can not confirm from a normal distribution, positive skewness skewed to the right, kurtosis is a lot larger than 3 indicate it has a sharp peak and fat tails.

Sum of Global Sales:

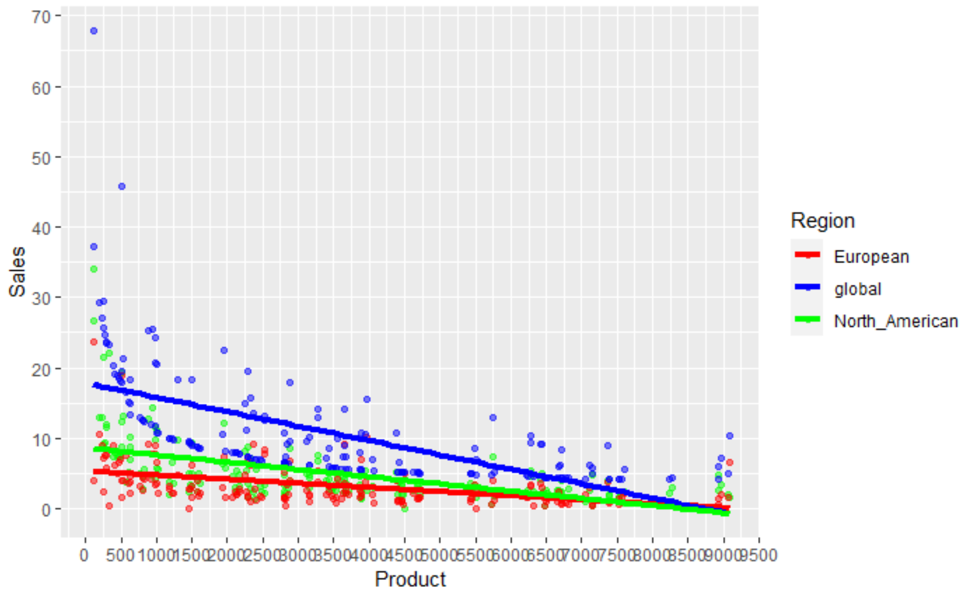


```
> shapiro.test(turtle_sales3$sum_Global_Sales)

	Shapiro-Wilk normality test

data:  turtle_sales3$sum_Global_Sales
W = 0.70955, p-value < 2.2e-16


> skewness(turtle_sales3$sum_Global_Sales)
[1] 3.066769
> kurtosis(turtle_sales3$sum_Global_Sales)
[1] 17.79072
```

For the Global Sales Shapiro-Wilk test, p<0.05 indicates that the data is not normally distributed, the Global Sales is not likely normal distribution, W always falls between 0 and 1, here 0.70955, is fair, but can not confirm from a normal distribution, positive skewness skewed to the right, kurtosis is a lot larger than 3 indicate it has a sharp peak and fat tails.

```
> cor(turtle_sales3$sum_NA_Sales,turtle_sales3$sum_EU_Sales)
[1] 0.6209317
> 
> cor(turtle_sales3$sum_EU_Sales,turtle_sales3$sum_Global_Sales)
[1] 0.8486148
> 
> cor(turtle_sales3$sum_NA_Sales,turtle_sales3$sum_Global_Sales)
[1] 0.9162292
```

The correlation function cor( ) indicate that both North American and European sales are correlated to Global sales. However, the correlation between North American and European is not strong.

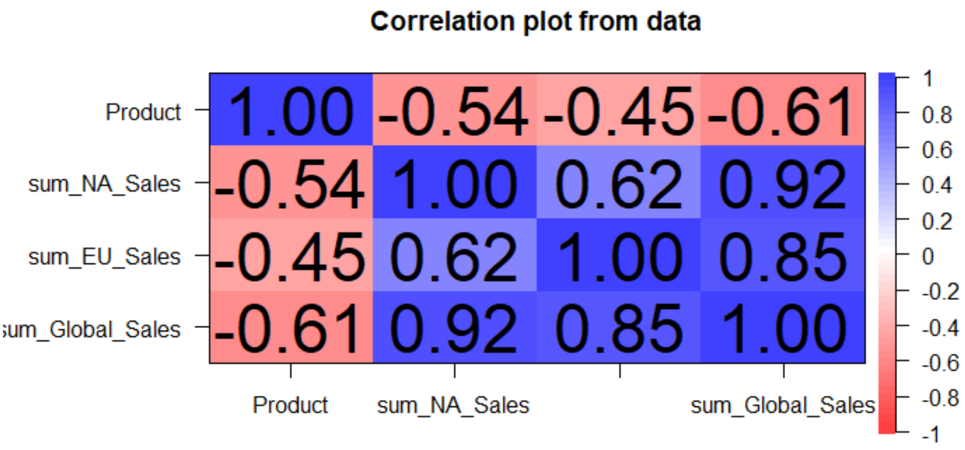


The scatterplot with regression line indicate that North American, European and Global sales trend emerged as product number glow bigger, they have obvious correlation.

VI. What the relationship(s) is/are (if any) between North American, European, and global sales?

Here we import psych library for analysing correlation and predict model.

Using cor( ) function and corPlot ( ) visualisation function, we understand the correlation between the 3 variables more:



And as there is no obvious correlation between North American and European Sales, we can formulated a linear regression of Global sales and dependent variable and North American sales and European sales as independent variable, utilising predict( )function, to do some prediction:

Create a data frame named turtle_sales_pred, input the forecast value of North American sales and European sales provided by Turtle Game, we can predict Global sales value and double check with some exact NA_sales and EU_sales figure:

```
> sum_Global_pred
        fit       lwr       upr
1 68.056548 66.429787 69.683310
2  7.356754  7.099418  7.614090
3  4.908353  4.614521  5.202185
4  4.761039  4.478855  5.043223
5 26.625558 25.367353 27.883763
```

```
> turtle_sales3_check
# A tibble: 3 × 4
  Product sum_NA_Sales sum_EU_Sales sum_Global_Sales
    <int>        <dbl>        <dbl>            <dbl>
1     107        34.0         23.8              67.8
2     326        22.1          0.52             23.2
3    6815         2.73         0.65              4.32
```

The predict value is reasonably accurate.

4. Conclusion:

All the question have been answered, Turtle Games relied on the product of smaller product code, the customer group should be classified by remuneration into 5 group, but relatively spread; the reputation of the company’s product and service is positive; the sales depend on North American more.