Data Analysis report – Predicting future outcomes for Turtle Games

Business context

Turtle Games, a game manufacturer and retailer, manufactures and sells their own products and products manufactured by other companies. Their products range includes books, board games, video games and toys. They have a global customer base. The stakeholders are their marketing team and sales team.

Business objective

The marketing team wants to improve overall sales performance by analysing the customer trends.

The sales team wants to find out the impact of sales per product id.

Primary sources

Data sets from turtle_reviews.csv and turtle_sales.csv are used for the analysis.

The data analysis are carried out in Python and R because both teams stated their preferred language.

Data profile of turtle_sales can be found in “Data Profiling Report – Turtle sales.pdf”.

Making predictions with regression using Python

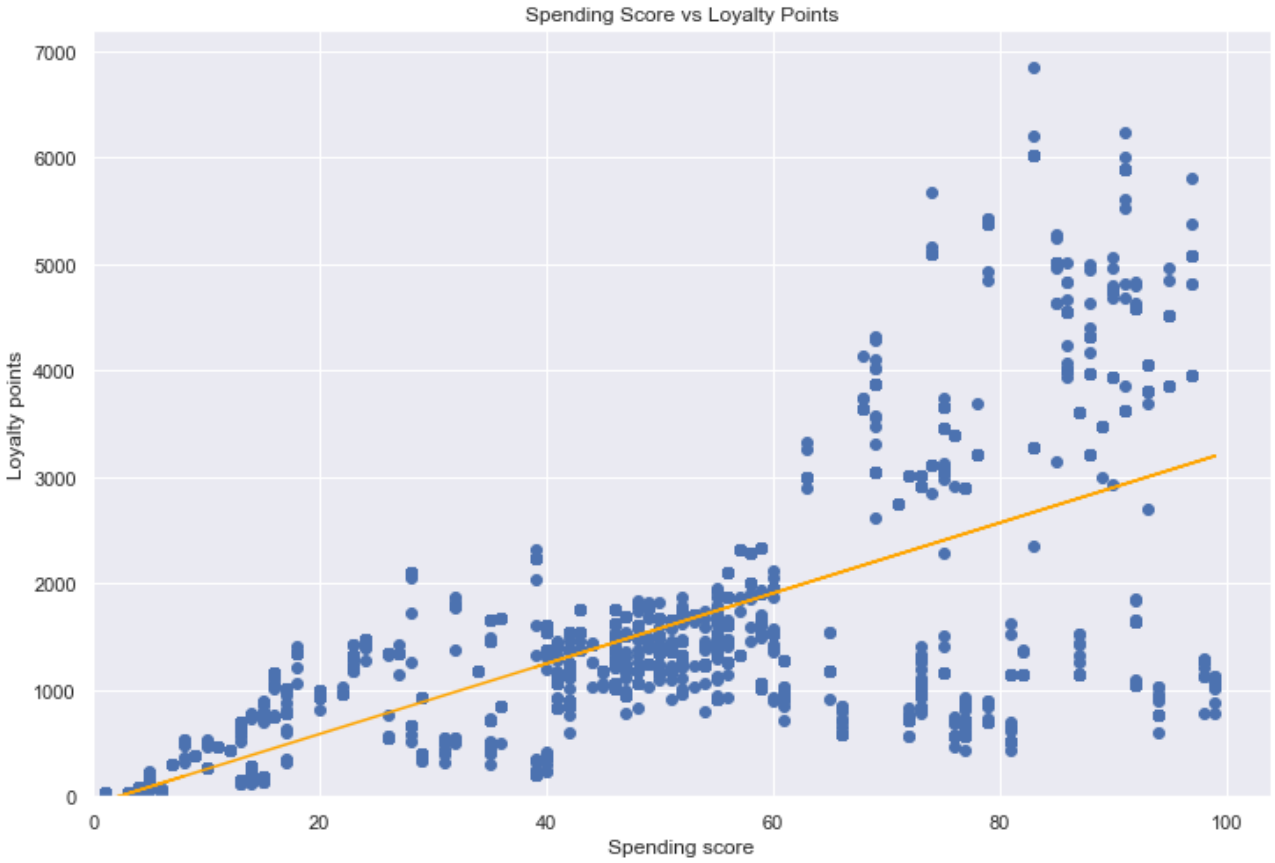
Objectives:

- Investigate the possible relationships between the loyalty points, age, remuneration and spending scores using linear regression.

There are 2000 reviews in the turtle reviews data set. There are no missing values and duplicates.

The linear regressions use the population size.

Spending scores and loyalty points



There is a positive relationship between spending scores and loyalty points. The higher the spending scores, the higher the loyalty points.

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.
Date:	Mon, 29 Aug 2022	**Prob (F-statistic):**	2.92e-263
Time:	17:17:27	**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

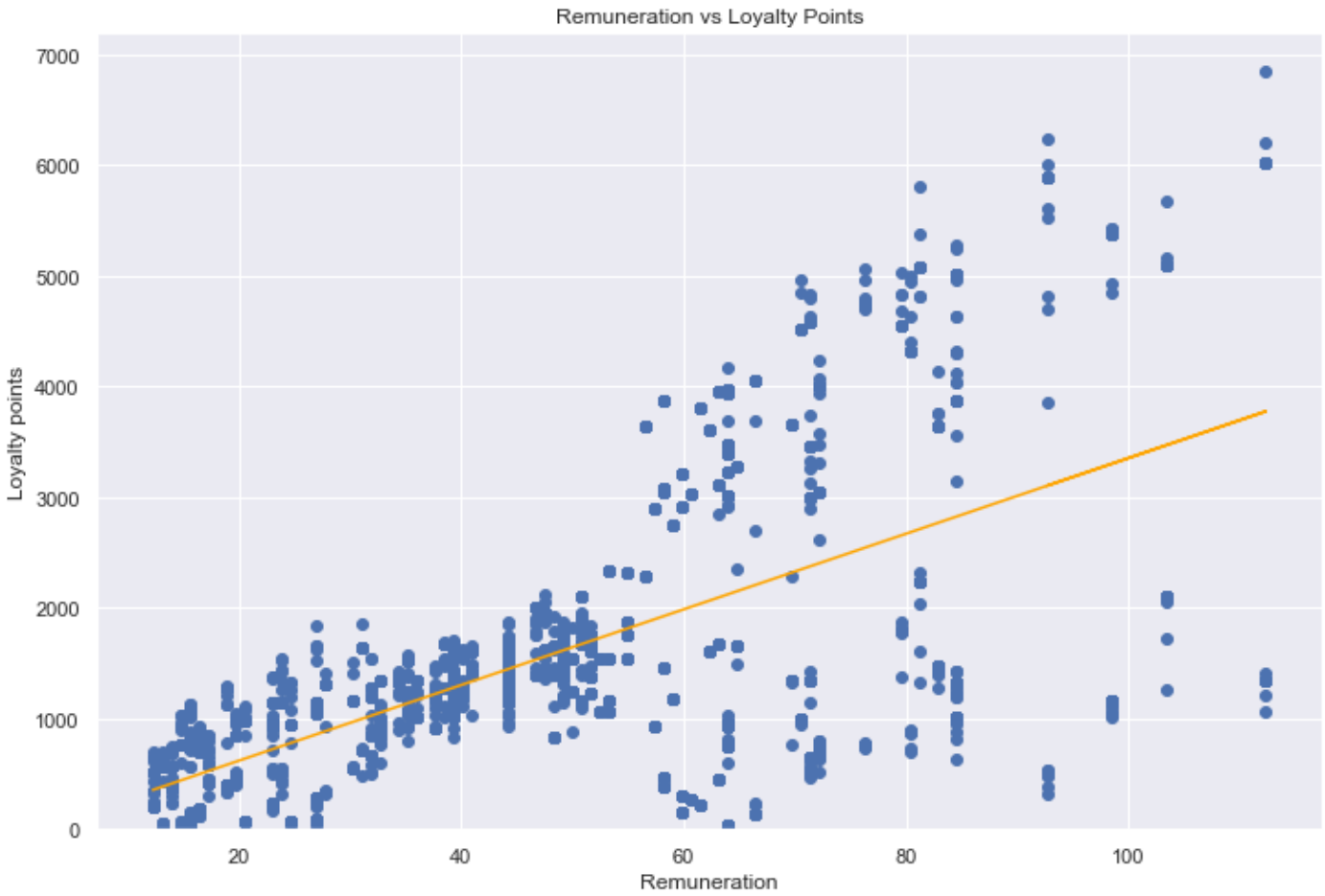
Omnibus:	126.554	**Durbin-Watson:**	1.191
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	260.528
Skew:	0.422	**Prob(JB):**	2.67e-57
Kurtosis:	4.554	**Cond. No.**	122.

R-squared is 0.452. Around 45% of the variations of the loyalty points can be explained by the customers' spending scores.

x- coefficient is 33.0617. The loyalty points will increase by 33 if the spending scores change increases by 1.

Probability of the t-test is 0 so the estimated slope is significant. Using the population size of 2000 reviews, the line does not fit perfectly among all data points. This highlights that the linear regression model could be underfitting.

Remuneration and loyalty points



There is a positive relationship between remuneration and loyalty points.

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.380
Model:	OLS	**Adj. R-squared:**	0.379
Method:	Least Squares	**F-statistic:**	1222.
Date:	Mon, 29 Aug 2022	**Prob (F-statistic):**	2.43e-209
Time:	18:39:20	**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.

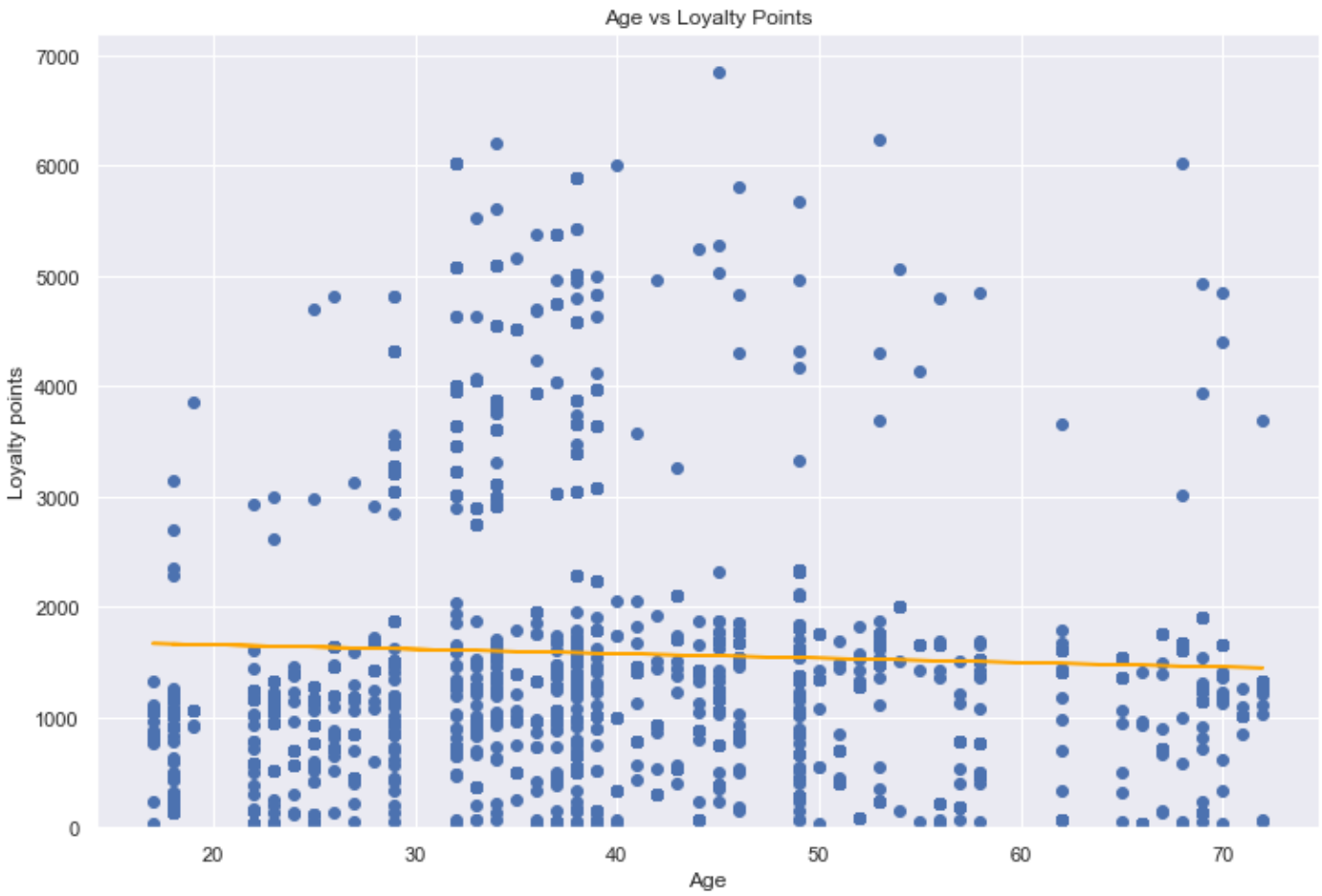
R_squared of 0.380 means around 38% of the observed variation loyalty points can be explained by the remuneration.

x- coeficient is 34.1878. The remuneration will increase by 34 if the loyalty point increases by 1.

Probability of the t-test is 0 so the estimated slope is significant.

Using the population size of 2000 reviews, the line does not fit perfectly among all data points. This highlights that the linear regression model could be underfitting.

Age and loyalty points



There is no relationship between age and loyalty points.

OLS Regression Results

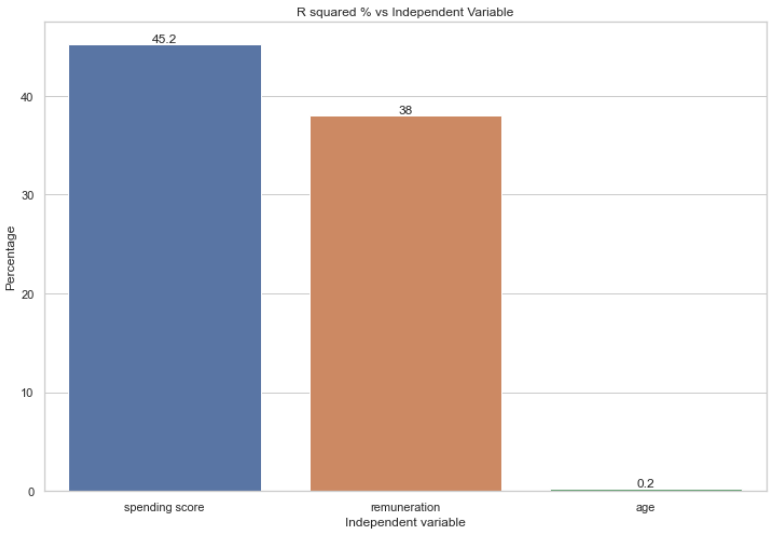
Dep. Variable:	y	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.606
Date:	Mon, 29 Aug 2022	**Prob (F-statistic):**	0.0577
Time:	19:14:56	**Log-Likelihood:**	-17150.
No. Observations:	2000	**AIC:**	3.430e+04
Df Residuals:	1998	**BIC:**	3.431e+04
Df Model:	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.

R_squared of 0.002 means only 0.2% of the loyalty points can be explained by customers' age.

Probability of the t-test is 0.58 tells us that the estimated slope is insignificant. There is very little relationship between age and loyalty points.



Insights and observations

There are positive relationships between spending scores and loyalty points, and remuneration and loyalty points. Both estimated slopes are significant. The lines fit loosely in both models. This highlights underfitting of the models because of using population of 2000 reviews. Even so, spending score and remuneration are good indicators to help marketing team to steer their marketing campaign because these groups of customers purchase more at Turtle Games.

To achieve more accurate predictions, further analysis with these variables are recommended using subsets of test and training data. Multiple regression model could then be applied using these variables to gain more insights among loyalty points, remuneration and spending scores.

There is no relationship found between age and loyalty points. Customers of any age groups spend as much

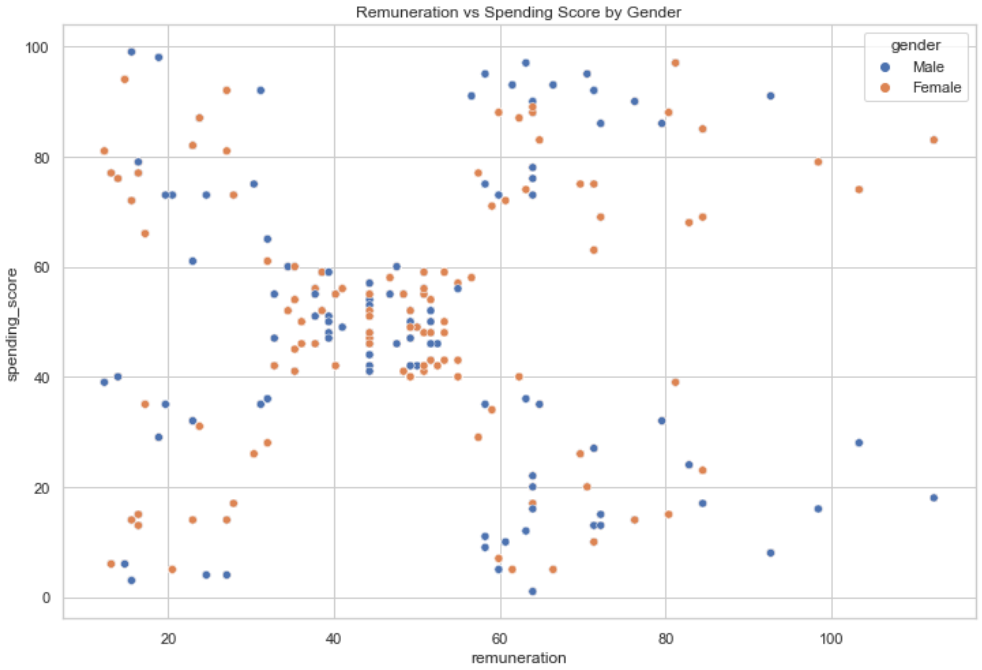
at Turtle Games.

Make predictions with clustering

The marketing department wants to better understand the usefulness of renumeration and spending scores. To identify groups within the customer base that can be used to target specific market segments, I use *k*-means clustering to identify the optimal number of clusters then apply and plot the data using the created segments.

Objectives:

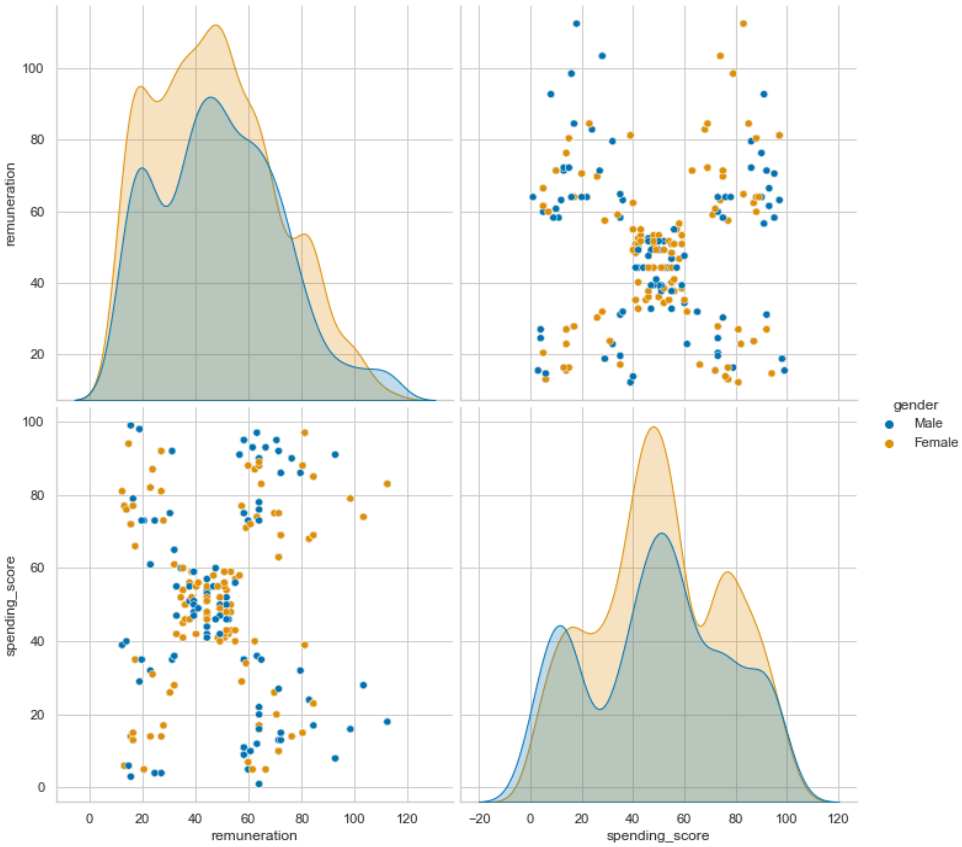
- Use k-means clustering to identify the optimal number of clusters
- Apply and plot the data using the created segments



The scatterplot shows the shape of the data derived from remuneration and spending scores.

There are 5 distinct clusters on the scatterplot grouped by gender.

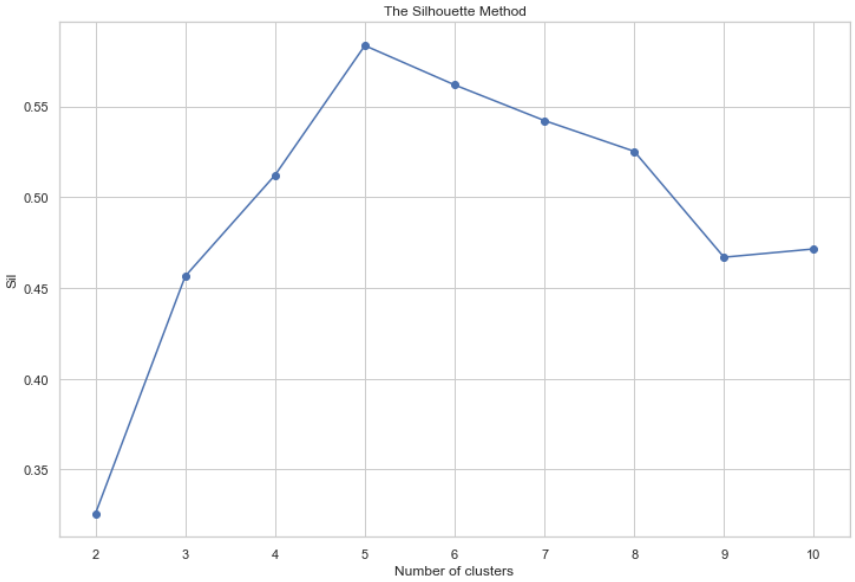
.



Pairplot shows there are many overlapped data points.

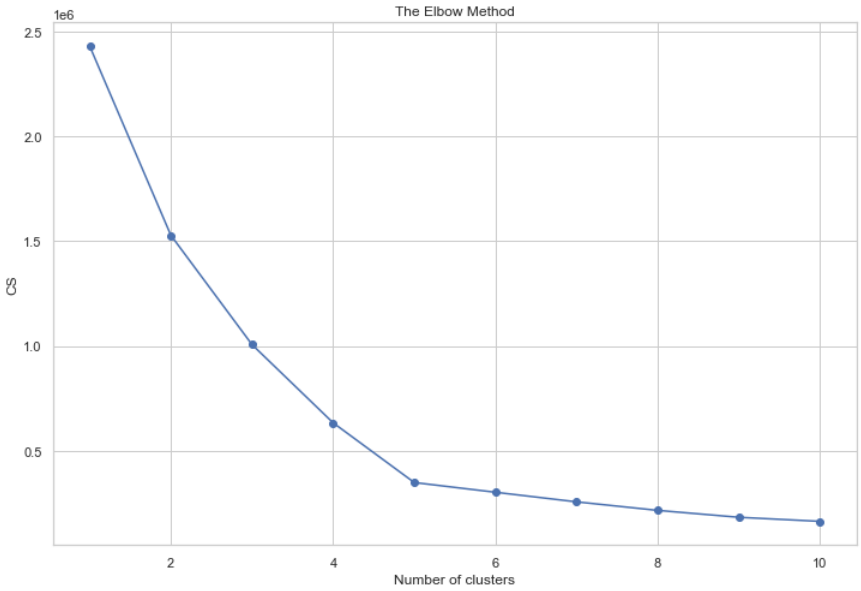
Determine the optimal number of clusters for k-means clustering

Silhouette method



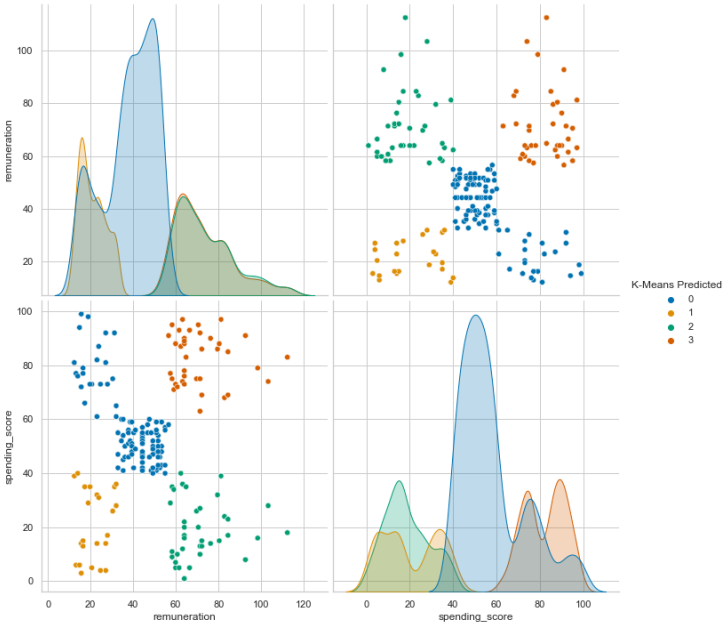
From Silhouette Method, the optimal number for k is 5.

Elbow method



From Elbow Method, the optimal number for k could be either 4 or 5.

Evaluate the usefulness of three possible values for k-means 4, 5 and 6.



k–means = 4

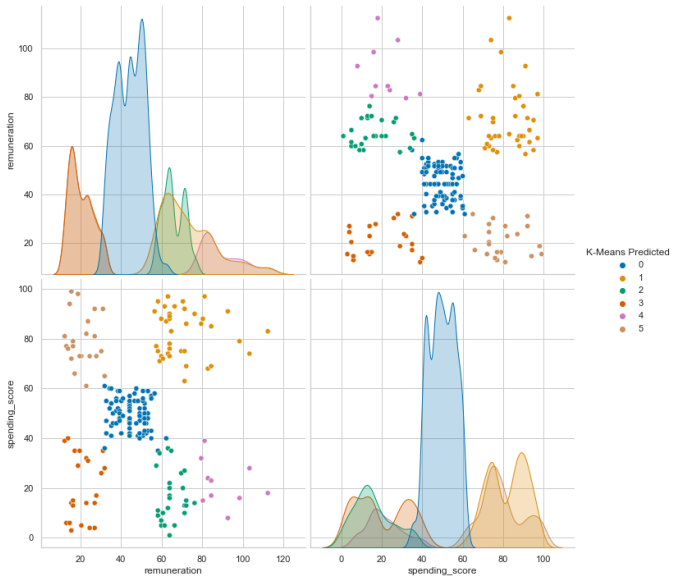
When k-means is 4. There are many overlapped data points. Cluster 0 has the most data points.

```
0    1013
3     356
2     351
1     280
Name: K-Means Predicted, dtype: int64
```

k-means = 5

When k-means is 5. There are 5 distinct clusters. Cluster 0 has the most data points.

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

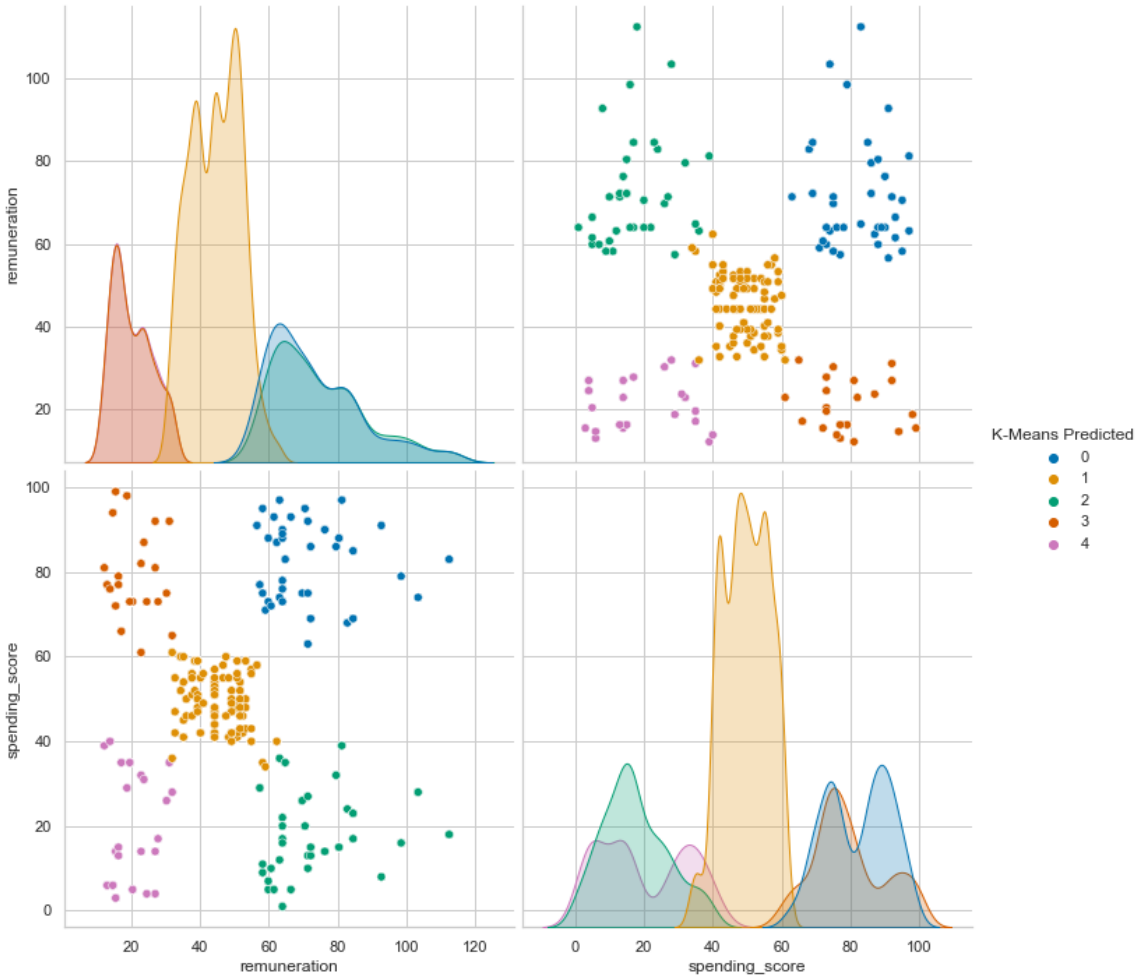


k-means = 6

When k-means is 6, some data points are overlapped. Cluster 0 has the most data points.

```
0    767
1    356
3    271
5    269
2    214
4    123
Name: K-Means Predicted, dtype: int64
```

Fit the model using k-means = 5



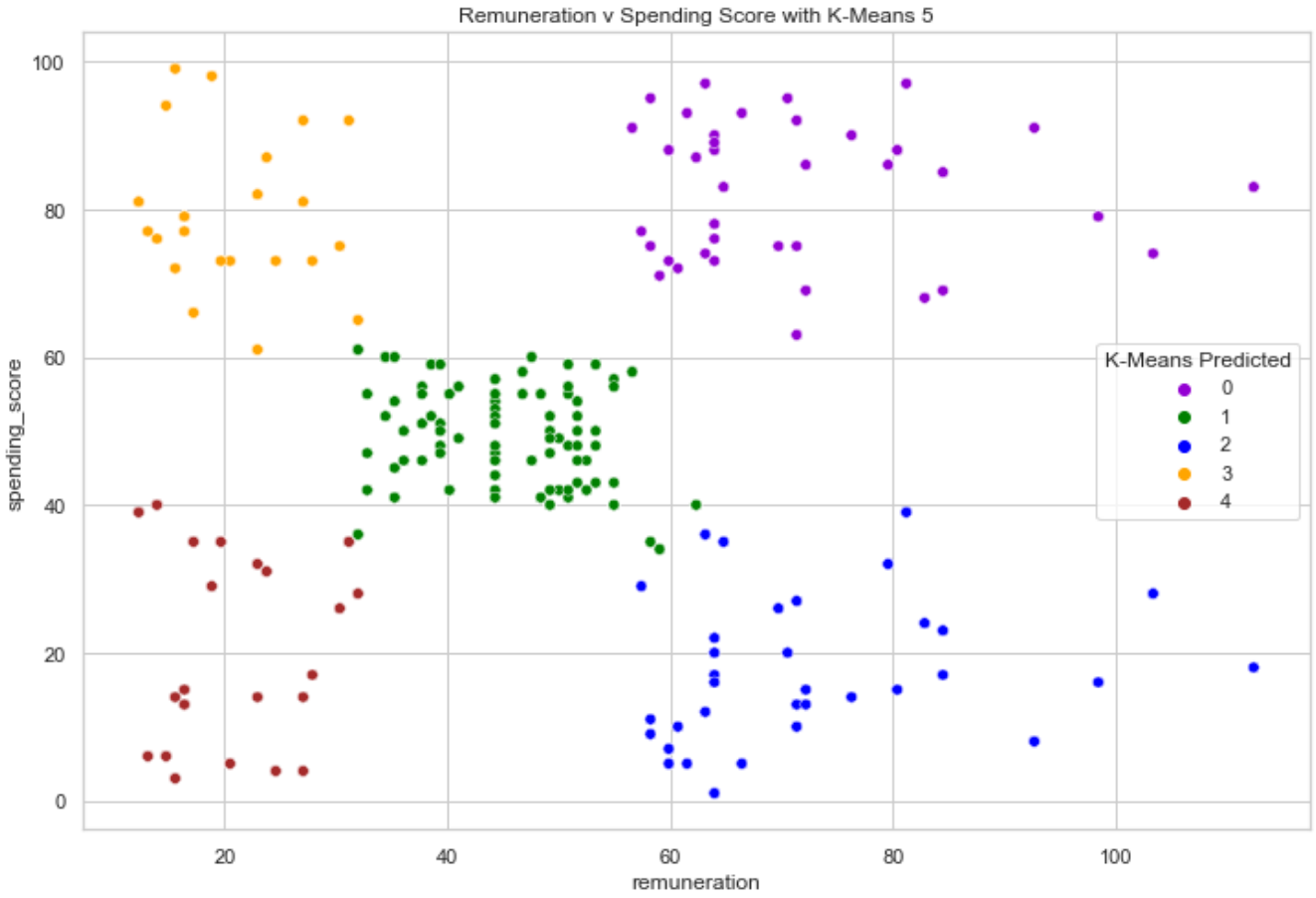
With k-means = 5, there are 5 distinct clusters. Cluster 1 has the most data points followed by clusters 0 and 2. Cluster 1 has the least overlaps with other clusters.

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

	remuneration	spending_score	K-Means Predicted
0	12.30	39	4
1	12.30	81	3
2	13.12	6	4
3	13.12	77	3
4	13.94	40	4

The k-Means predicted indicated that the first 5 data points allocated to clusters 3 and 4.

.



Insights and Observations:

Using the k-means, the optimal number of clusters is 5. The clusters are distinct with fewer overlaps.

Data points are more evenly allocated in 5 clusters.

Customers base could be grouped into 5 groups using the relationship between remuneration and spending score:

- Cluster 0 has high remuneration with high spending score with some visible outliers.
- Cluster 1 has both average remuneration and spending score. It has the hardest cluster among all clusters.
- Cluster 2 has high remuneration with low spending score and some visible outliers
- Cluster 3 has low remuneration with higher spending score.
- Cluster 4 has low remuneration with low spending score.

Clusters 0 and 3 have higher spending score, this tells us that these segments of customer base are more willing to spend, and are useful for target marketing. The rest of the groups could be differentiated for different marketing strategies.

Analyse customer sentiments with reviews

To help marketing department of Turtle Games to inform future campaigns, NLP is used to analyse products reviews downloaded from Turtle Games website about customer sentiments.

Objectives:

- Identify the 15 most common words online products reviews
- Identify top 20 positive reviews
- Identify top 20 negative reviews

There are 2000 reviews. Duplicates are defined as same reviews and summary. 39 duplicates removed.

No missing values found. In total 1961 reviews are used for the analysis.

Reviews word cloud

In reviews word cloud, many alphanumeric characters and stop words dominate. They have little value for analysis so will have to be removed.

Summary word cloud

In summary word cloud, "five stars", "fun" , “good” and "game" stand out. The initial summary word cloud gives a positive sentiment. “Game” appears many times in the word clouds, this could tell us that among all the products range, sales in games could be more popular than books.

Reviews word cloud after removing stop words

In reviews, the common words include “game”, “play”, “card”, “love”, “one” and “great”. “Game” stands out in reviews like in the summaries, this could tell us that games are popular products among all products sales in Turtle Games.

Summary word cloud after removing stop words

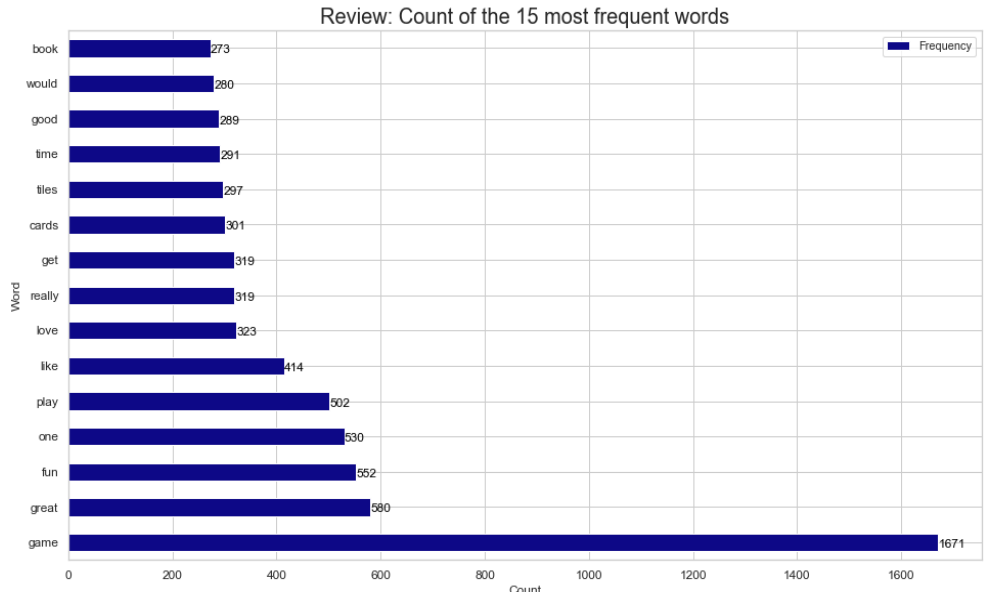
Similar patterns in summary, popular words are “five stars”, “four stars”, “love”, “great”, “fun” and “game”.

The frequency of “game” is high visible in the word cloud.

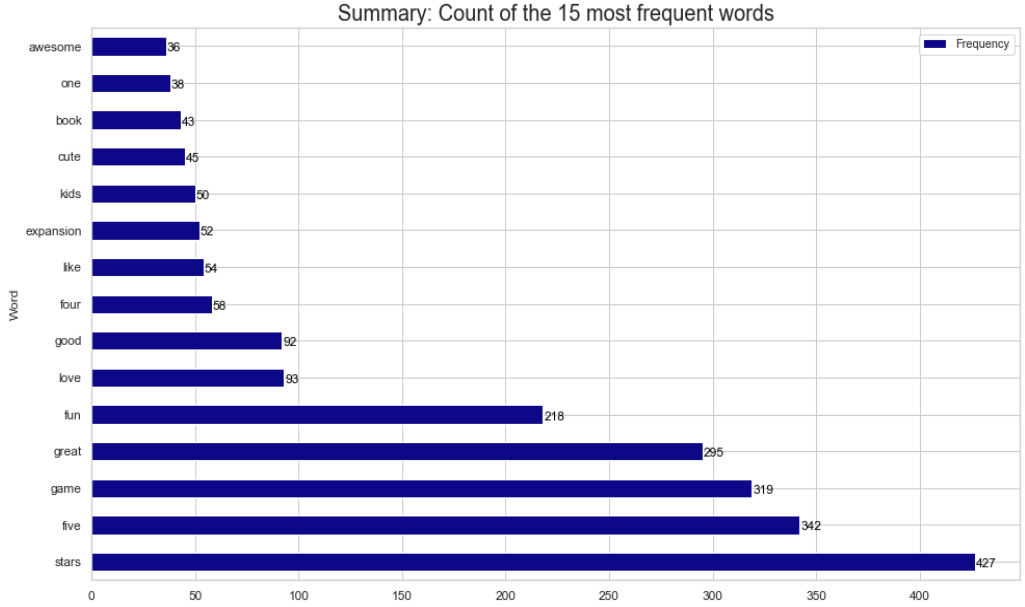
Word cloud from both reviews and summary

In both reviews and summary, “game” is the most visible. This tells us that most reviews are about game products sold at Turtle Games. “five star” gives a positive sentiment about the customers’ feedbacks. Interestingly, “one” stands out in the word cloud. It would be interesting to find out in the both contexts, what does “one” imply.

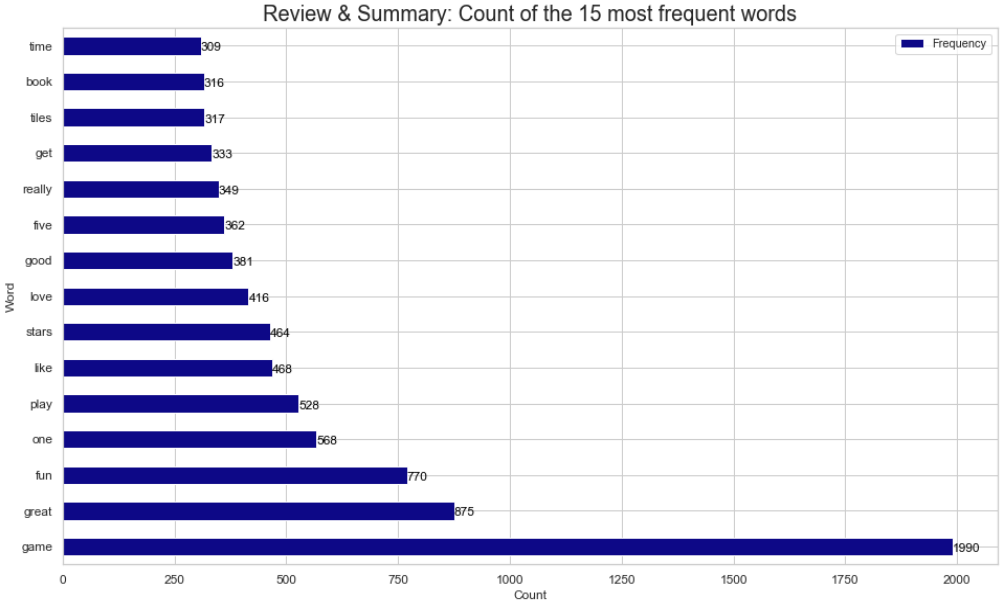
Top 15 most common words in Review, Summary, Review and Summary



Top 15 common words in reviews are "game", "great", "fun", "one" and a few more. "book" is in the top 15 words in review. It is interesting to find out what “one” means in reviews.



Top 15 common words in summary are "stars", "five", "game", "great" and a few more. Again, "book" is in the top 15 words. "five" "stars" send a positive sentiment in the summaries.



Top 15 words in both review and summary are “game”, “great”, “fun”, “play” and others. This tells us that games could be the best products sales among the all product ranges followed by “book”. Interesting to find out what “one” means.

Vader Sentiment Analysis

Top 15 positive common words and polarity in reviews

	neg	neu	pos	compound
entertaining	0.0	0.0	1.0	0.4404
fun gift	0.0	0.0	1.0	0.7351
ok	0.0	0.0	1.0	0.2960
cool	0.0	0.0	1.0	0.3182
great	0.0	0.0	1.0	0.6249
fantastic	0.0	0.0	1.0	0.5574
satisfied thanks	0.0	0.0	1.0	0.6908
awesome	0.0	0.0	1.0	0.6249
satisfied	0.0	0.0	1.0	0.4215
nice	0.0	0.0	1.0	0.4215
cute	0.0	0.0	1.0	0.4588
outstanding	0.0	0.0	1.0	0.6124
loved loved loved	0.0	0.0	1.0	0.9136
fine	0.0	0.0	1.0	0.2023
fun	0.0	0.0	1.0	0.5106

Overall, there is a fairly strong positive opinions in the reviews. They reflect more about the enjoyment of the products.

Top 15 common negative words and polarity in reviews

	neg	neu	pos	compound
difficult	1.000	0.000	0.000	-0.3612
incomplete kit very disappointing	0.538	0.462	0.000	-0.5413
no more comments	0.524	0.476	0.000	-0.2960
a crappy cardboard ghost of the original hard to believe they did this but they did shame on hasbro disgusting	0.487	0.455	0.058	-0.9052
not a hard game to learn but not easy to win	0.470	0.456	0.075	-0.7946
i found the directions difficult	0.455	0.545	0.000	-0.3612
who doesnt love puppies great instructions pictures fun	0.445	0.334	0.221	-0.5207
different kids had red faces not sure they like	0.368	0.632	0.000	-0.4717
got the product in damaged condition	0.367	0.633	0.000	-0.4404
i bought this thinking it would be really fun but i was disappointed its really messy and it isnt nearly as easy as it seems also the glue is useless for a 9 year old the instructions are very difficult	0.362	0.592	0.045	-0.9520
great game poor quality	0.337	0.217	0.446	0.2500
we really did not enjoy this game	0.325	0.675	0.000	-0.4389
not as easy as it looks	0.325	0.675	0.000	-0.3412
hard to put together	0.318	0.682	0.000	-0.1027
my 8 yearold 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	0.318	0.613	0.069	-0.8674

In the negative reviews, they tend to be about the poor quality of the products and the ease of use of the products.

Top 15 positive common words and polarity in summary

	neg	neu	pos	compound
love	0.0	0.0	1.0	0.6369
great helper	0.0	0.0	1.0	0.7579
great	0.0	0.0	1.0	0.6249
cute	0.0	0.0	1.0	0.4588
super fun	0.0	0.0	1.0	0.8020
ok ok	0.0	0.0	1.0	0.5267
wonderful	0.0	0.0	1.0	0.5719
perfect	0.0	0.0	1.0	0.5719
good fun	0.0	0.0	1.0	0.7351
thanks	0.0	0.0	1.0	0.4404
perfect gift	0.0	0.0	1.0	0.7650
pretty cool	0.0	0.0	1.0	0.6705
wow	0.0	0.0	1.0	0.5859
fun fun fun	0.0	0.0	1.0	0.8720
okay	0.0	0.0	1.0	0.2263

In summary, the positive sentiments reflect more about the enjoyment of the products.

Top 15 negative common words and polarity in summary

	neg	neu	pos	compound
disappointing	1.000	0.000	0.0	-0.4939
meh	1.000	0.000	0.0	-0.0772
frustrating	1.000	0.000	0.0	-0.4404
boring	1.000	0.000	0.0	-0.3182
disappointed	1.000	0.000	0.0	-0.4767
defective poor qc	0.857	0.143	0.0	-0.7184
not great	0.767	0.233	0.0	-0.5096
mad dragon	0.762	0.238	0.0	-0.4939
no 20 sided die	0.753	0.247	0.0	-0.7269
damaged product	0.744	0.256	0.0	-0.4404
faulty product	0.697	0.303	0.0	-0.3182
money trap	0.697	0.303	0.0	-0.3182
nothing special	0.693	0.307	0.0	-0.3089
wimpy magnets	0.655	0.345	0.0	-0.2263
anger control game	0.649	0.351	0.0	-0.5719

In summary, negative opinions are about the quality of the products and ease of use.

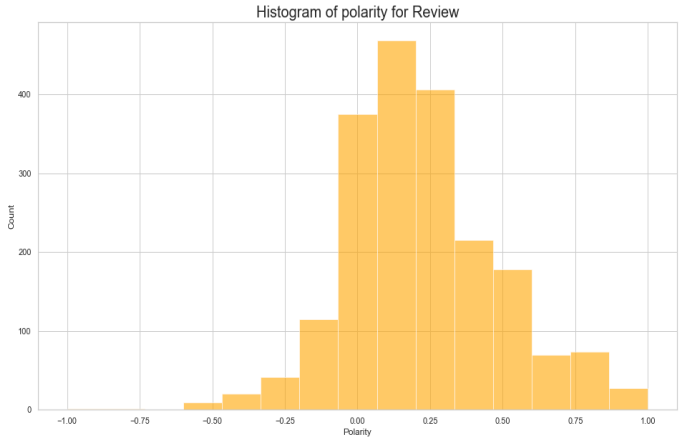
Top 20 positive reviews

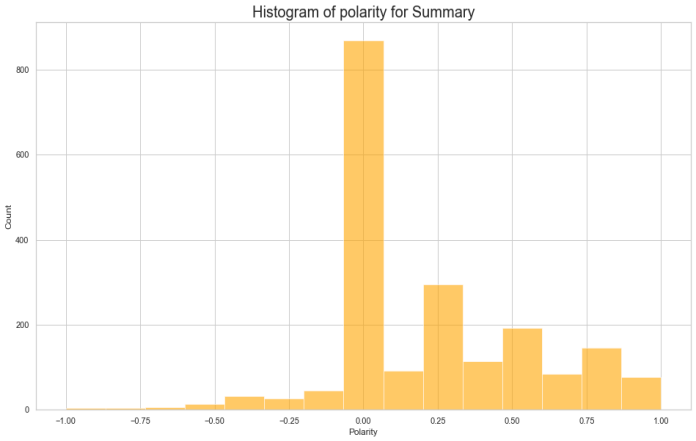
	review	polarity_review
7	came in perfect condition	1.000000
165	awesome book	1.000000
194	awesome gift	1.000000
496	excellent activity for teaching selfmanagement skills	1.000000
524	perfect just what i ordered	1.000000
591	wonderful product	1.000000
609	delightful product	1.000000
621	wonderful for my grandson to learn the resurrection story	1.000000
790	perfect	1.000000
933	awesome	1.000000
1037	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	its 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

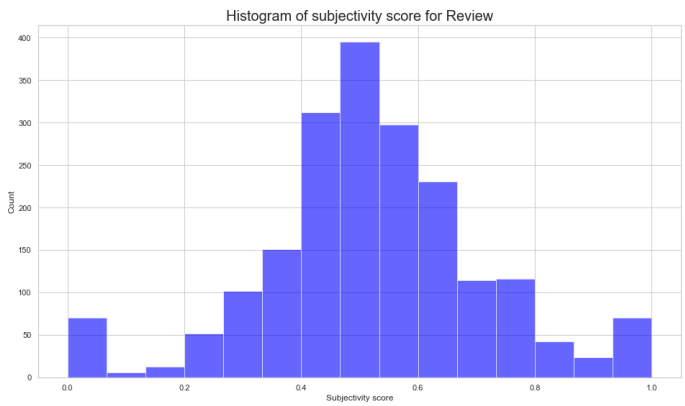
“awesome” appear many times in top reviews. We could try to find out what are awesome, are the products, service or the delivery or any unknown. A very useful insight for marketing team to focus on their strength.

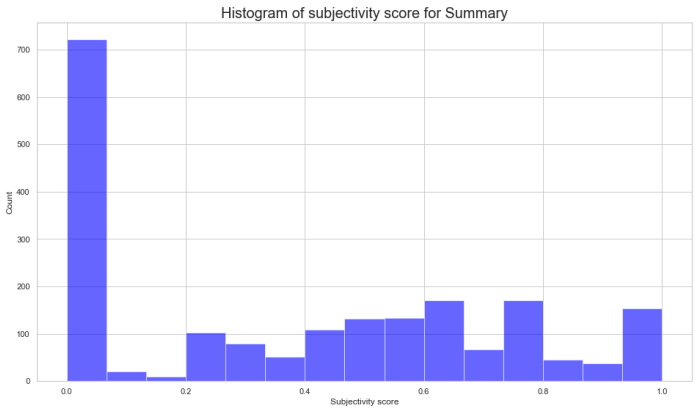
Top 20 negative reviews

	review	polarity_review
208	booo unles you are patient know how to measure i didnt 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	im 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 dont 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 pompom maker comes in two parts and is supposed to snap together to create the pompoms however both parts were the same making it unusable if you cant make the pompoms 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 yearold 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









Overall the polarity in reviews and summary tend to be neutral to positive.

The subjectivity score of review is fairly evenly distributed on both side of 0.5. This suggests detects more customers' opnions.

In summaries, the subjectivity score is rather low. A high 0 counts suggest most summaries are factual. Summaries do not detect much customers' opinions.

Top 20 positive summaries

	summary	polarity_summary
6	best gm screen ever	1.000000
28	wonderful designs	1.000000
32	perfect	1.000000
80	theyre 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 welldesigned 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

Results from summary are positive opinions about suitability of the products for the age groups or as a gift.

Top 20 negative summaries

	summary	polarity_summary
21	the worst value ive 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 masters 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 qualityall 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

“disappointing” appear many times in the negative summary. Further analysis could help to find out what customers are disappointed of, the products, service and delivery.

Observations and insights:

To understand the customers' sentiments both frequency distribution and VADER sentiment analysis are used for the analysis. Results from frequency distributions evaluate the top 15 common words in both reviews and summary. These words include "game", "great", "five", "stars", "book" and a few more. These imply that games could be the best selling products among all products range, followed by "book". Overall, there is a positive sentiment.

To have more insights, VADER sentiments analysis is used to measure the sentiments both positive and negative. With recurring words like "awesome" in top positive reviews, a very useful insight for marketing team to

find out what customers are very satisfied of, the products, quality, service or any other unknowns. They could focus on their strengths in their next marketing campaigns.

Negative feedbacks from summary like "disappointing" need further analysis to find out what customers are disappointed of. This will help marketing team to improve on them.

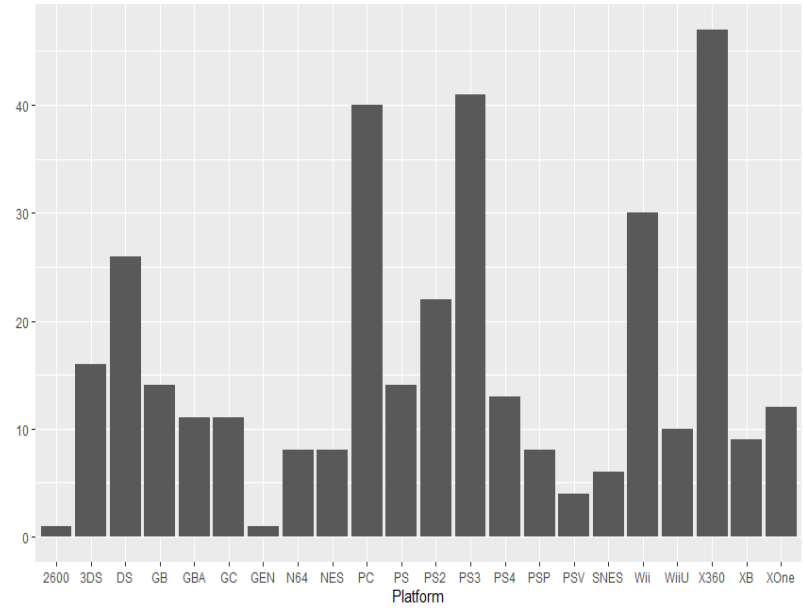
Overall, reviews detect a wider variety of sentiments than summaries.

Visualise data to gather insights

Objectives:

- What insights gathered from scatterplots, histograms, and boxplots
- The impact on sales per product id

There are 175 unique products with products ids ranging from 107 to 9080. There are 352 products across 22 different platforms. X360 platform runs across most products. On average the platforms run across 16 different products.

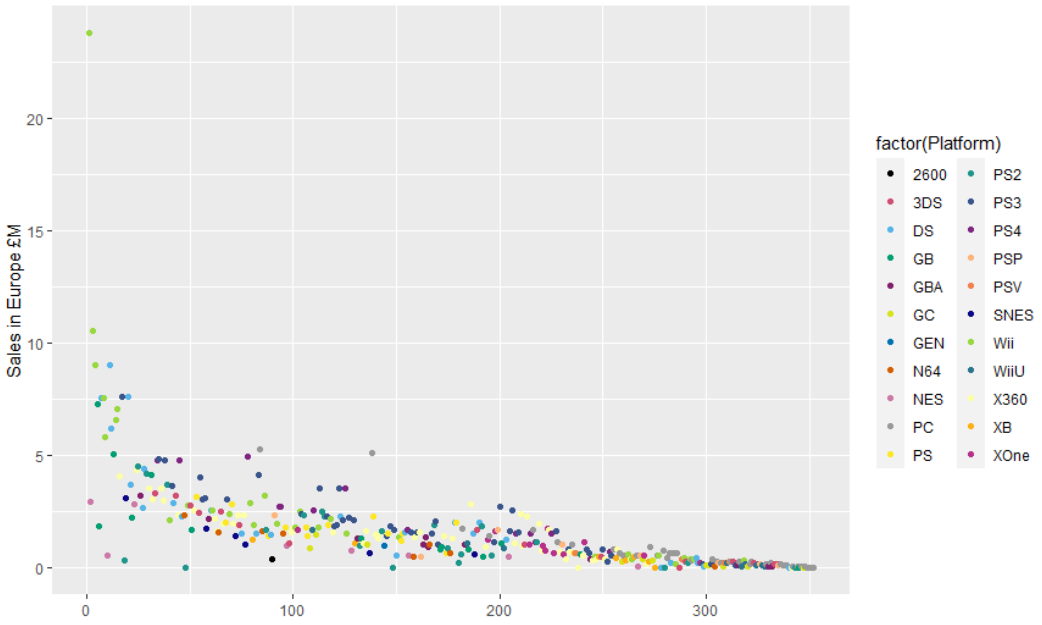


Platform	n
2600	1
GEN	1
PSV	4
SNES	6
N64	8
NES	8
PSP	8
XB	9
WiiU	10
GBA	11
GC	11
XOne	12
PS4	13
GB	14
PS	14
3DS	16
PS2	22
DS	26
Wii	30
PC	40
PS3	41
X360	47

Top 5 popular platforms are X360, PS3, PC, Wii and DS.

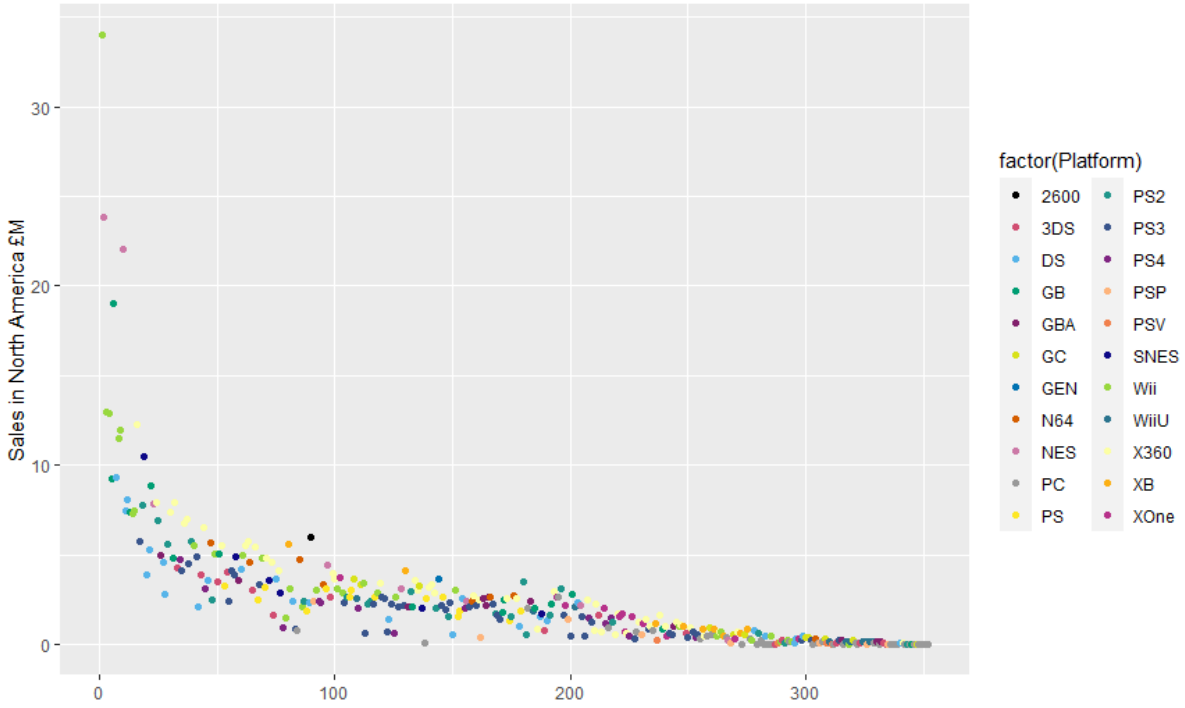
Top 5 least popular platforms are 2600, GEN, PSV, SNES, N64 and NES.

Europe sales (£M) by Platforms



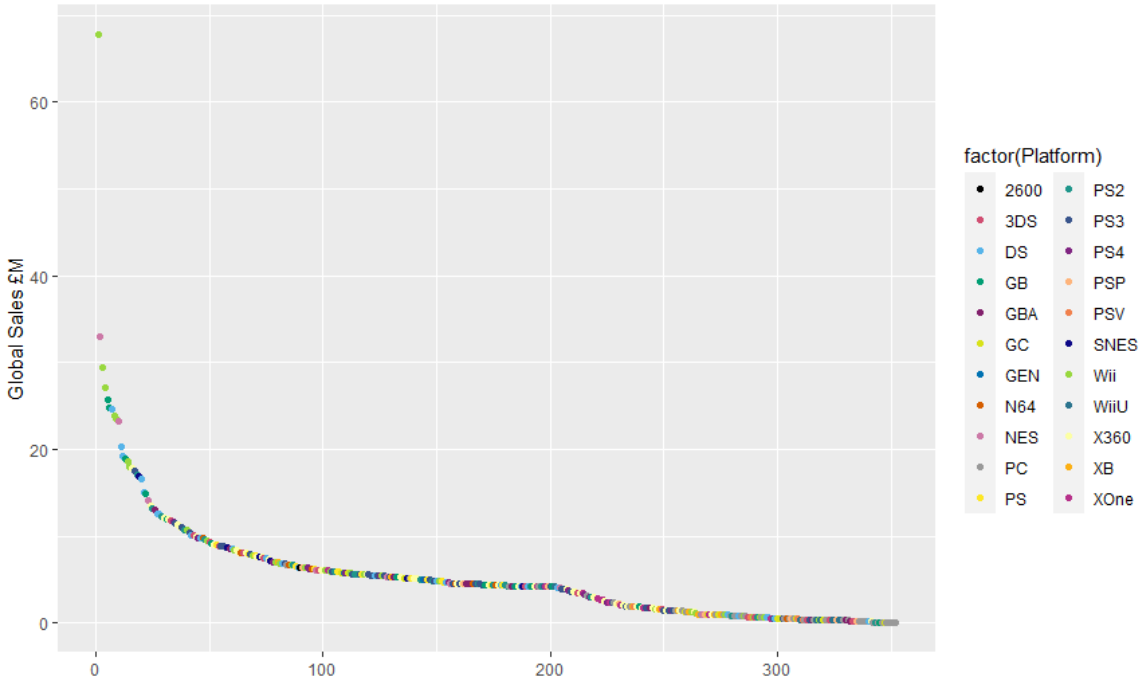
Most sales in the Europe are around £2.5M. There is an outlier with over £22.5M with Wii platform.

North America sales (£M) by Platforms



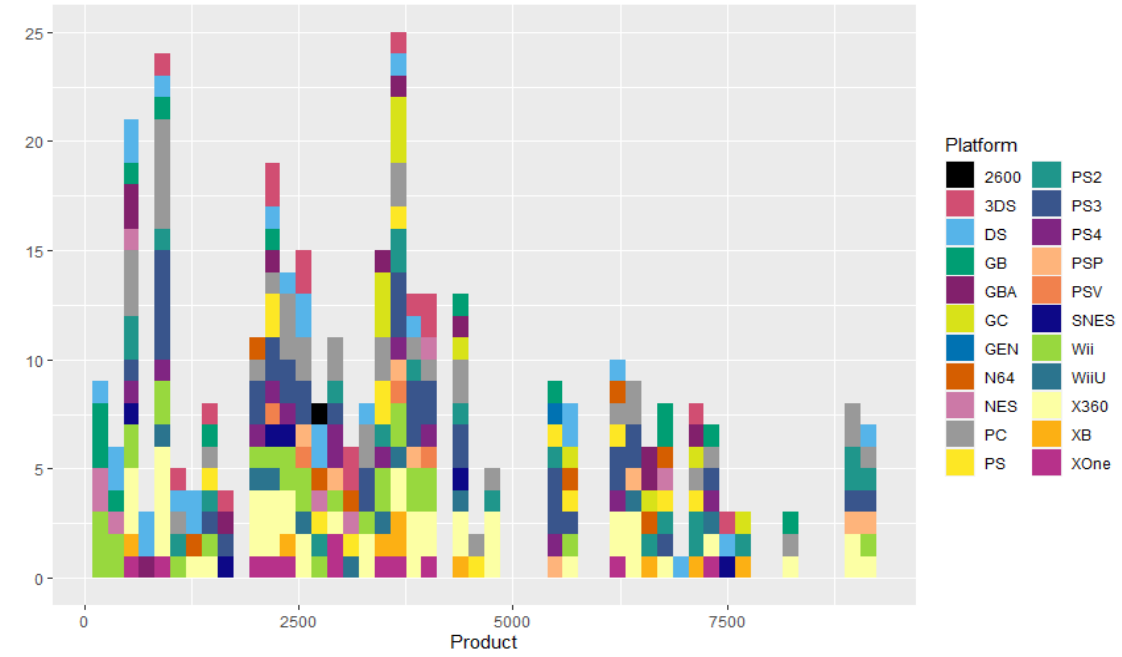
Most sales in the North America are around £5M. There are a few outliers with sales around £20M and one over £30M with Wii platform. North America has a stronger sales than Europe overall.

Global sales (£M) by Platforms

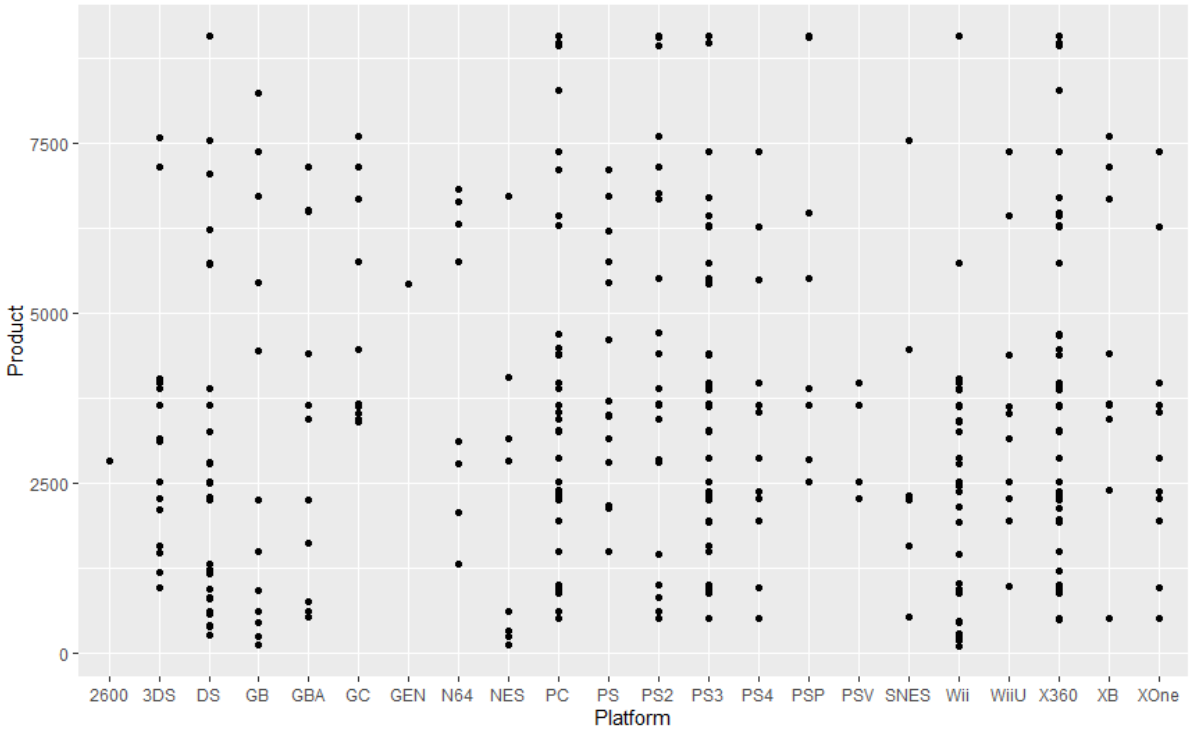


Global sales show an outlier of over £60M with Wii platform. It is interesting to find out which products with Wii platform produce such strong sales.

Products trends by Platforms

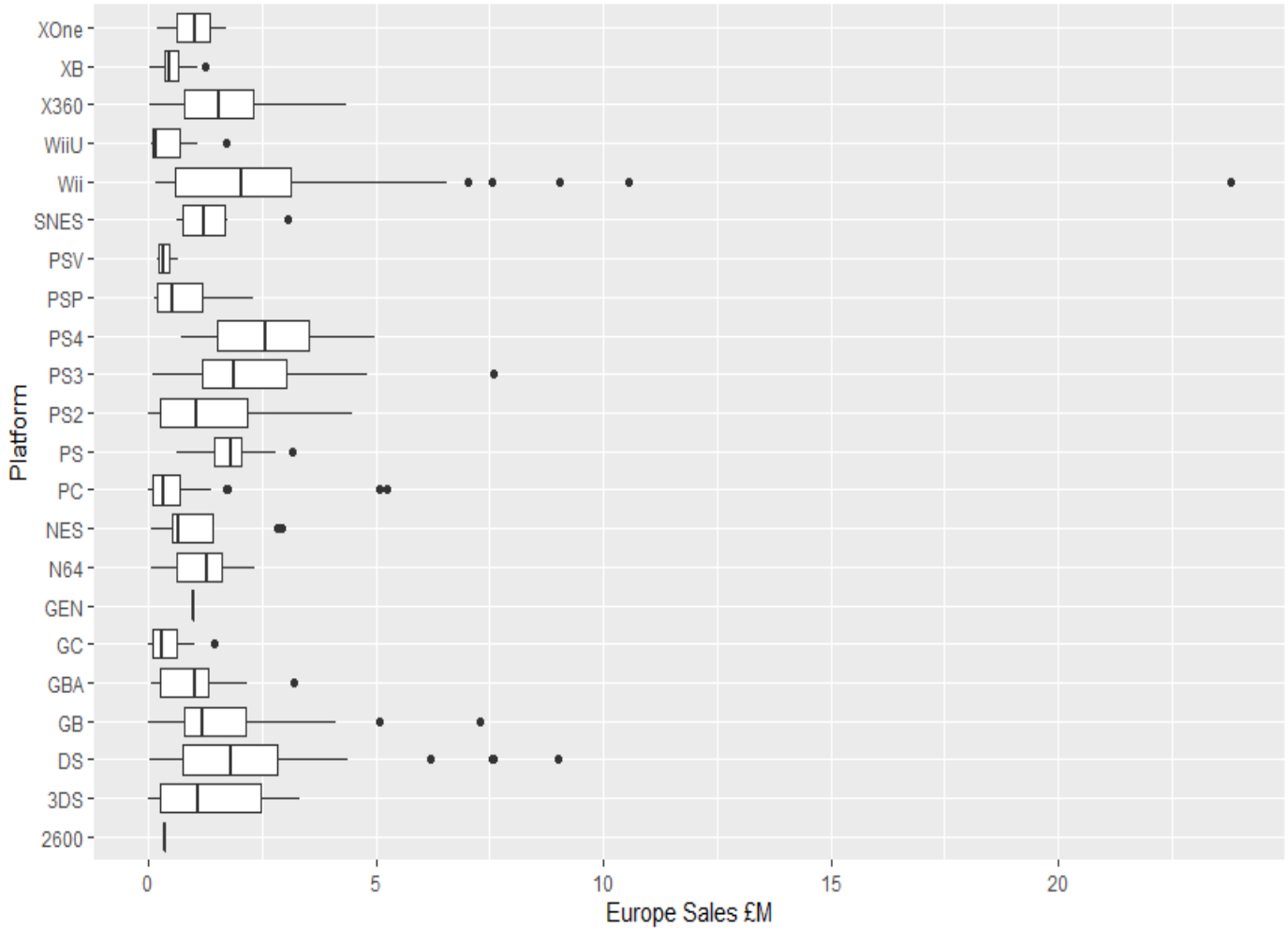


X360 platform is the most popular patform across all product ranges. Product ids around 3750 have the highest sales.



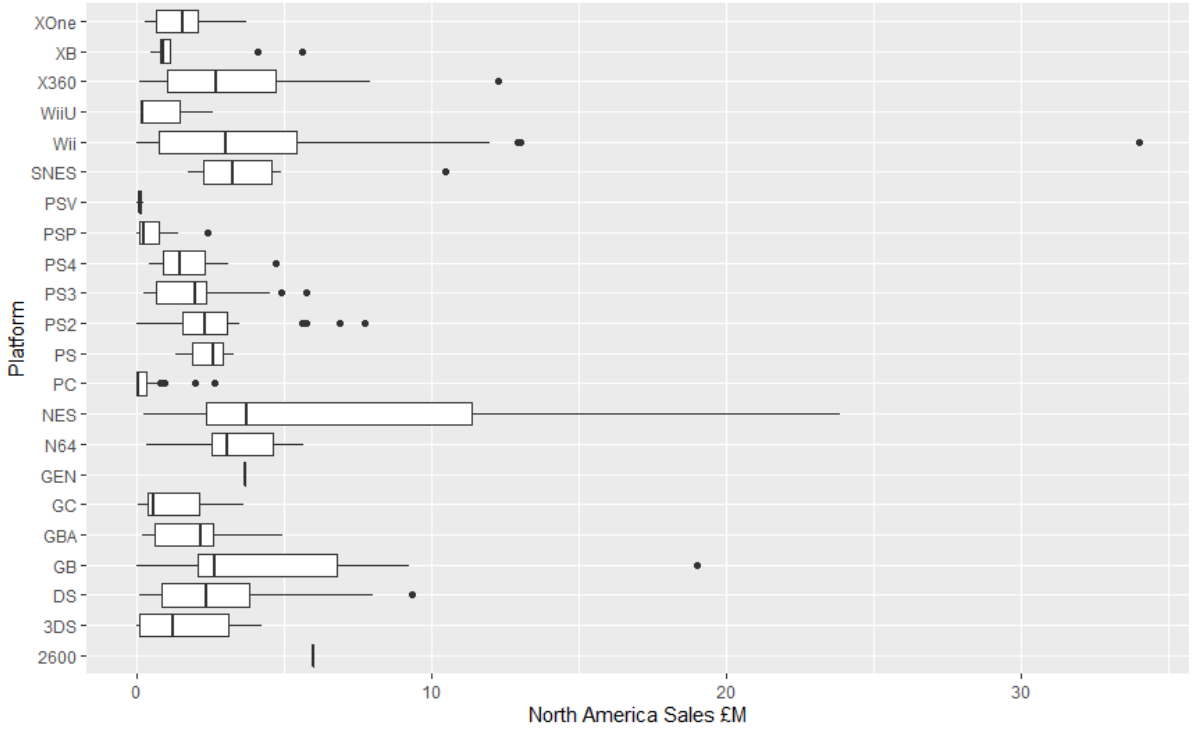
PC, PS2, PS3 and X360 platforms have most products range. DS and Wii platforms have product ranges mostly spread below 4000. PC, PS2, PS3 and X360 platforms have product ids higher than 7500.

Europe sales (£M) by Platforms



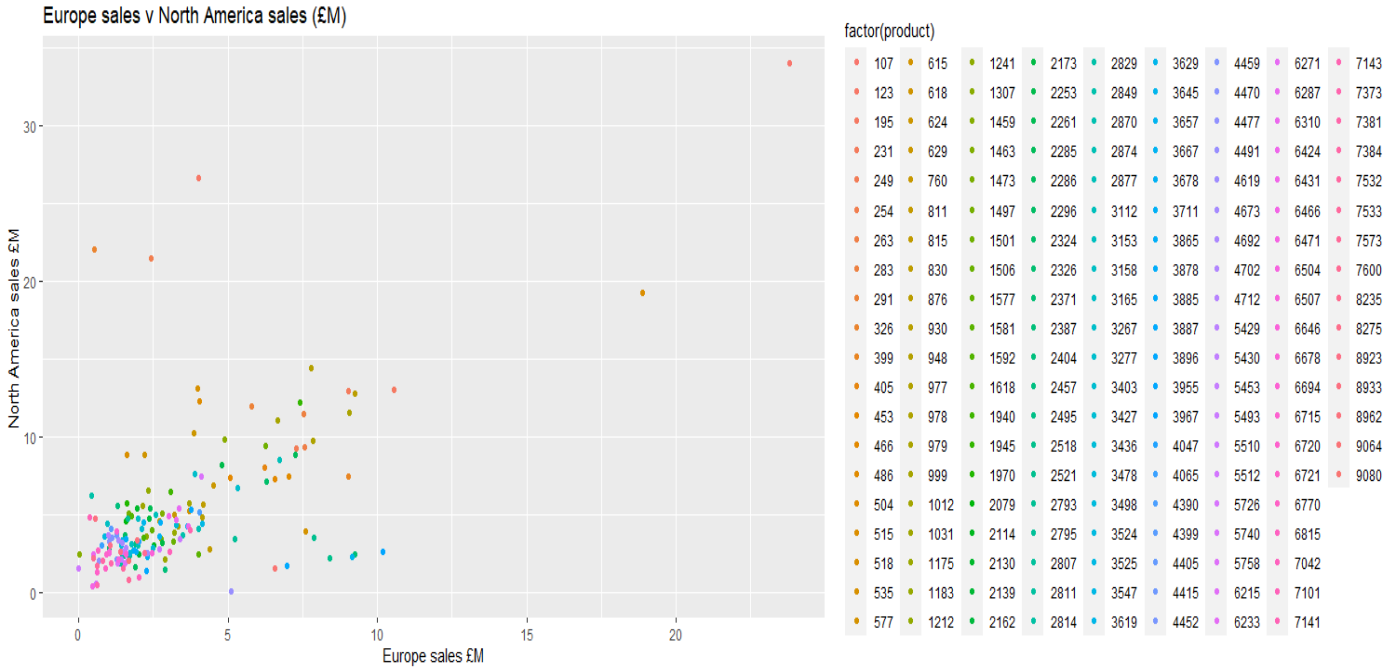
Wii platform in Europe has the many outliers, followed by PC.

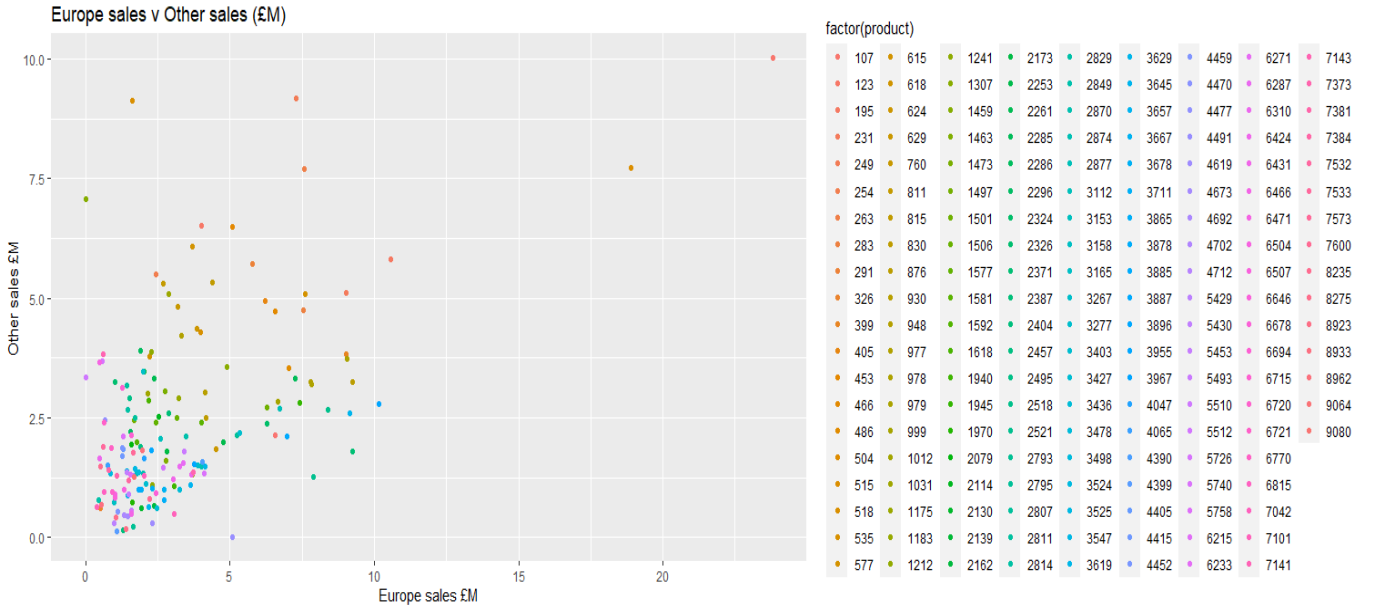
North America sales (£M) by Platforms

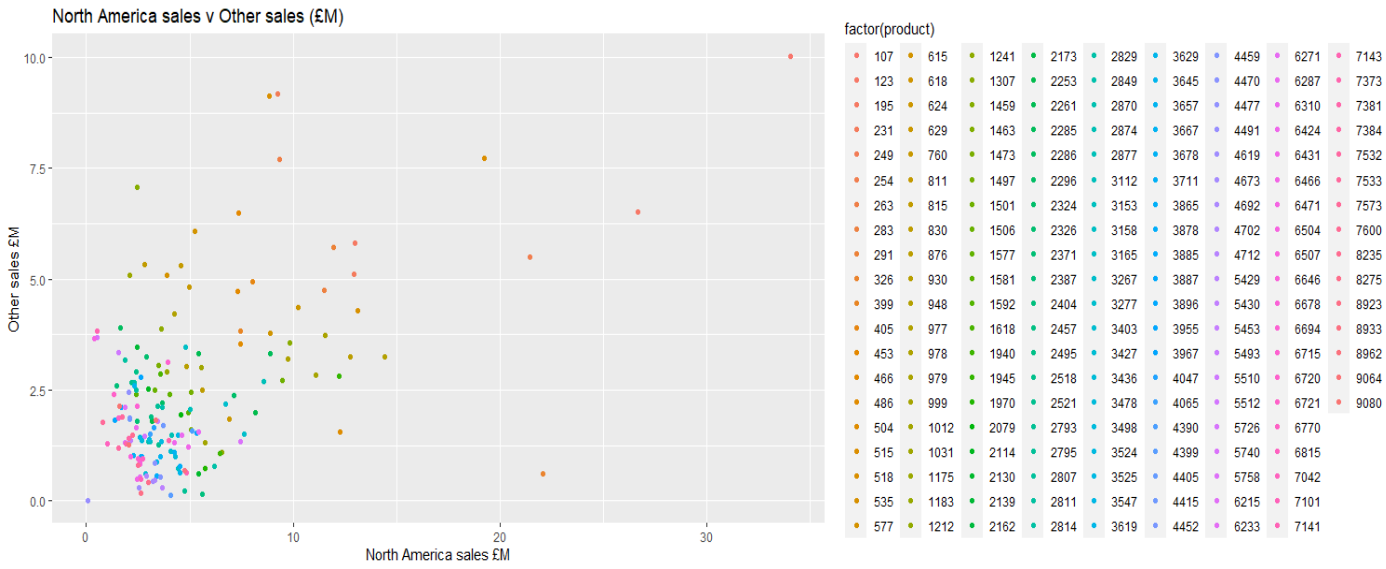


Wii platform has also outliers in North America. PC, PS2 and PS3 platforms have outliers in North America.

There some positive correlations in sales across Europe, North America and Other with some visible outliers.





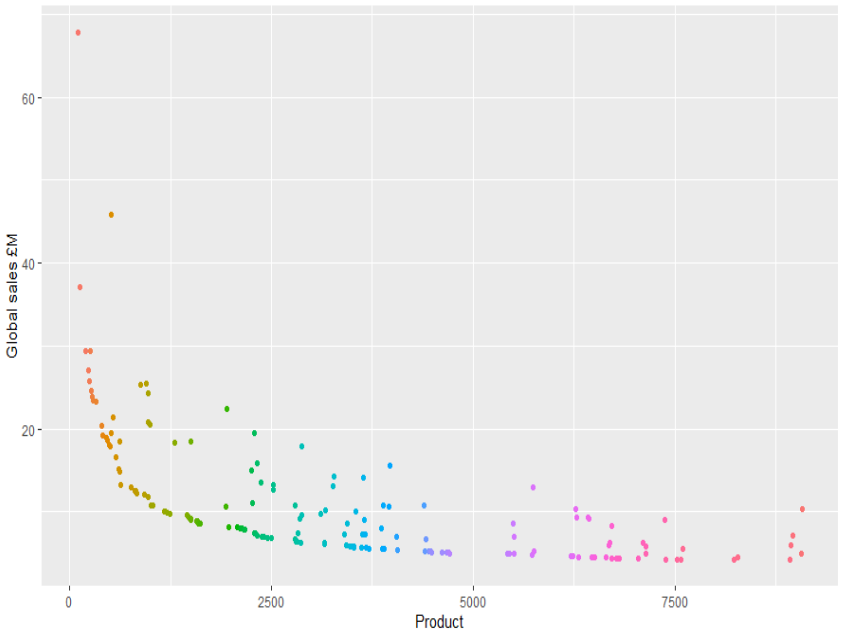


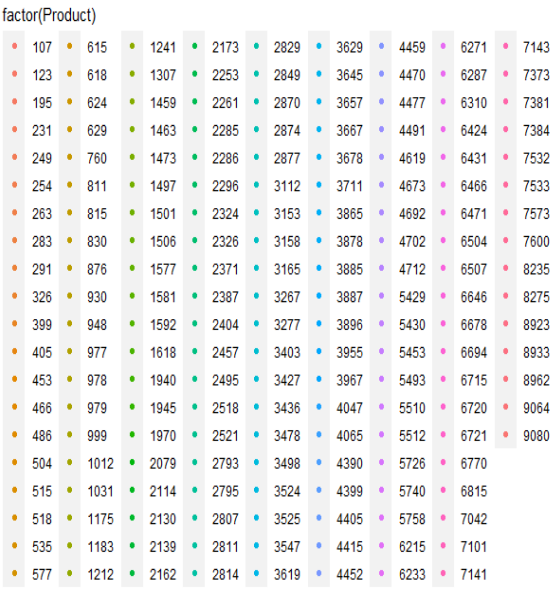
Product ids by Global sales (£M)

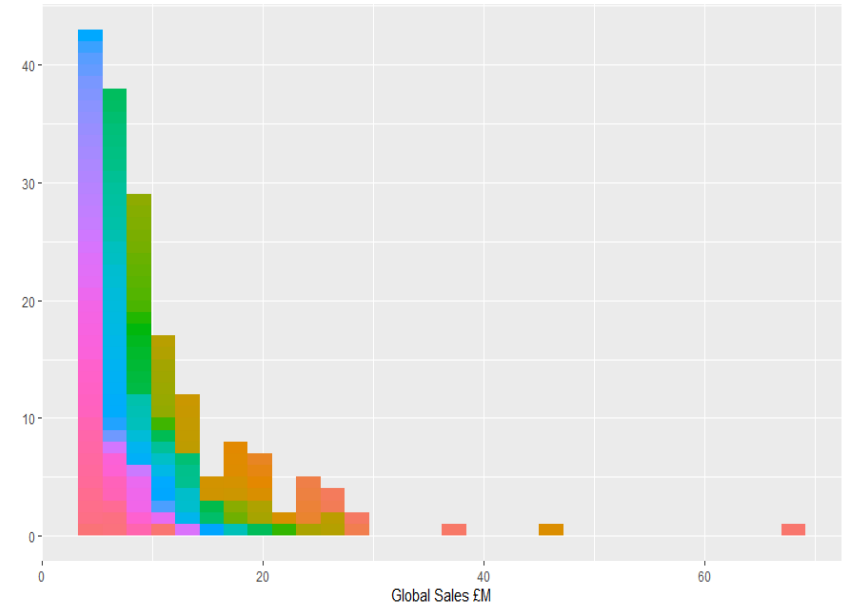
	Product	global_sum	global_no_platforms
1	107	67.85	1
2	515	45.86	5
3	123	37.16	2
4	254	29.39	2
5	195	29.37	1

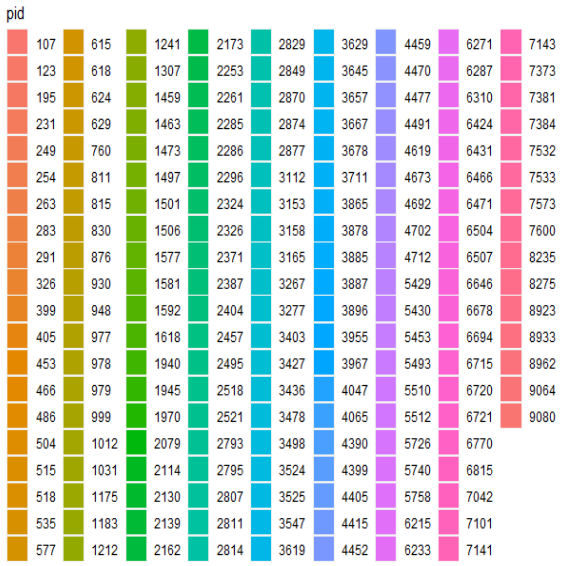
	Product	global_sum	global_no_platforms
1	3645	14.06	9
2	2518	13.26	8
3	3967	15.59	8
4	3887	10.79	7
5	9080	10.30	7

Top 5 products ids with the highest global sales are **107, 515, 123, 254 and 195**. They run across from 1 to maximum 5 platforms. The product ids which run on the most platforms do not produce the highest global sales. The product ids has a negative relationship with global sales, the higher the product ids the lower the global sales.

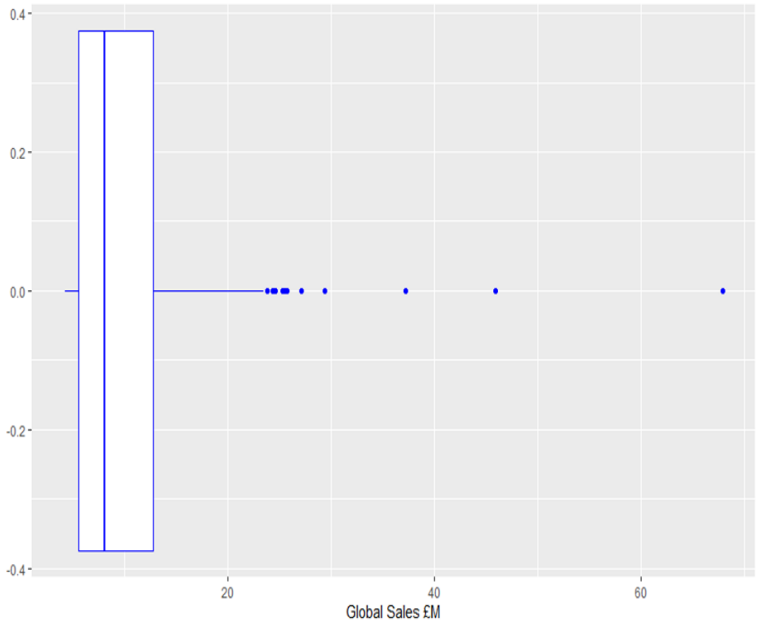








The highest global sales shown as an outlier is £67.8 M. On average the global sales is around £8.09M. There are many outliers with sales over £20M.



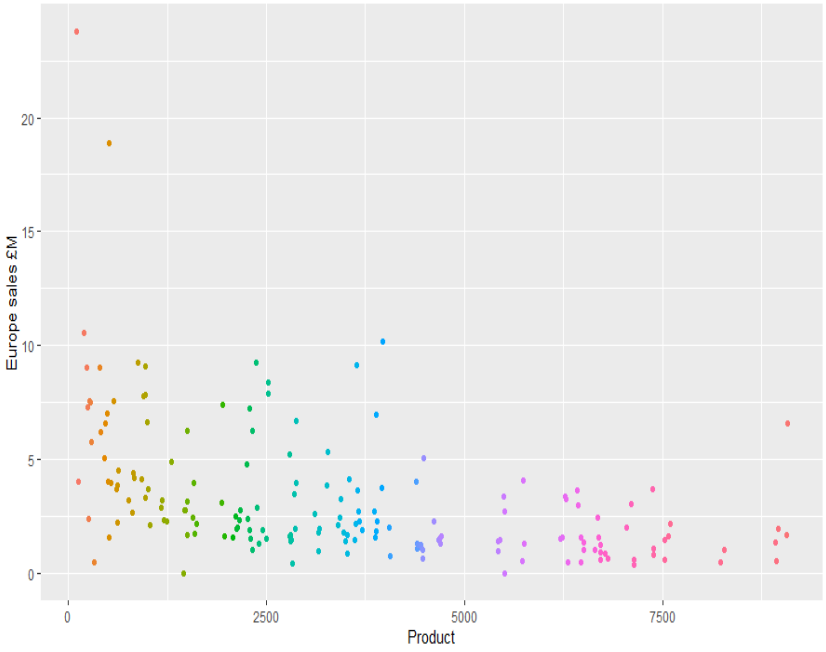

```
> # summary global sales by product
> summary(global_product_sales)
    Product       global_sum     global_no_platforms
 Min.   : 107   Min.   : 4.200   Min.   :1.000
 1st Qu.:1468   1st Qu.: 5.515   1st Qu.:1.000
 Median :3158   Median : 8.090   Median :1.000
 Mean   :3490   Mean   :10.730   Mean   :2.011
 3rd Qu.:5442   3rd Qu.:12.785   3rd Qu.:2.000
 Max.   :9080   Max.   :67.850   Max.   :9.000
>
```

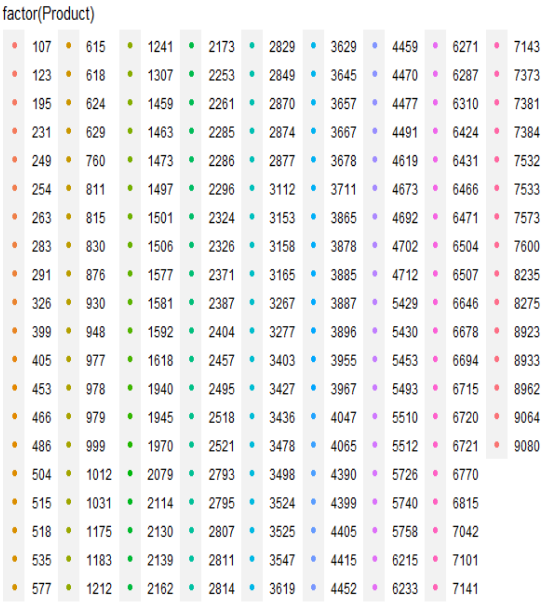
Product ids by Europe sales (£M)

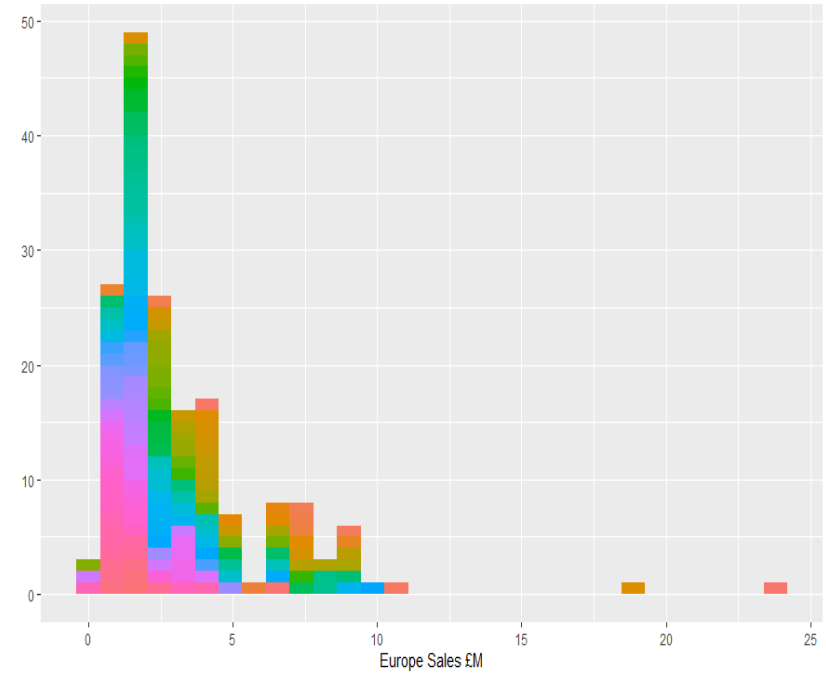
	Product	eu_sum	eu_no_platforms
1	107	23.80	1
2	515	18.88	5
3	195	10.56	1
4	3967	10.17	8
5	2371	9.26	5

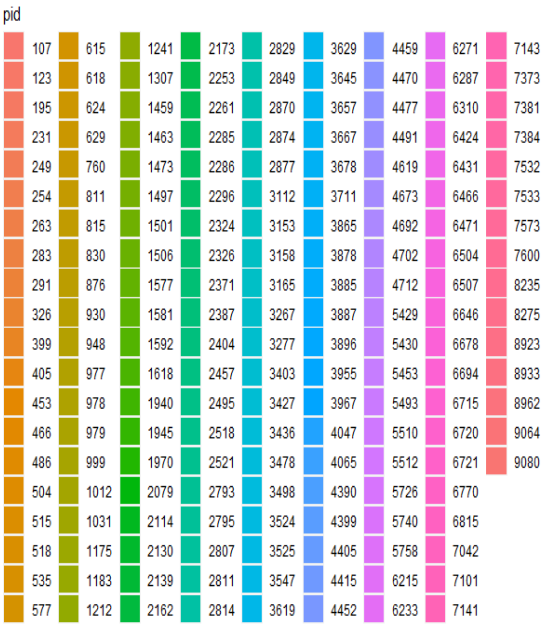
	Product	eu_sum	eu_no_platforms
1	3645	9.14	9
2	2518	8.40	8
3	3967	10.17	8
4	3887	6.97	7
5	9080	6.57	7

Top 5 products ids with the highest Europe sales are **107, 515, 195, 3967 and 2371**. They run across from 1 to maximum 8 platforms. The product id **3967** which runs on the 8 platforms generates one of top sales in Europe.

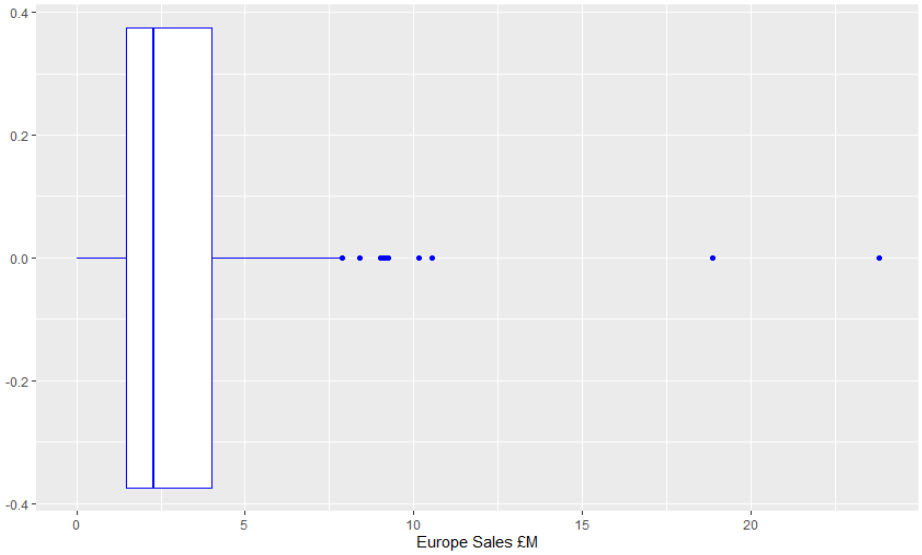








The highest Europe sales shown as outliers of around £20M. On average the Europe sales is lower and around £2.3M.



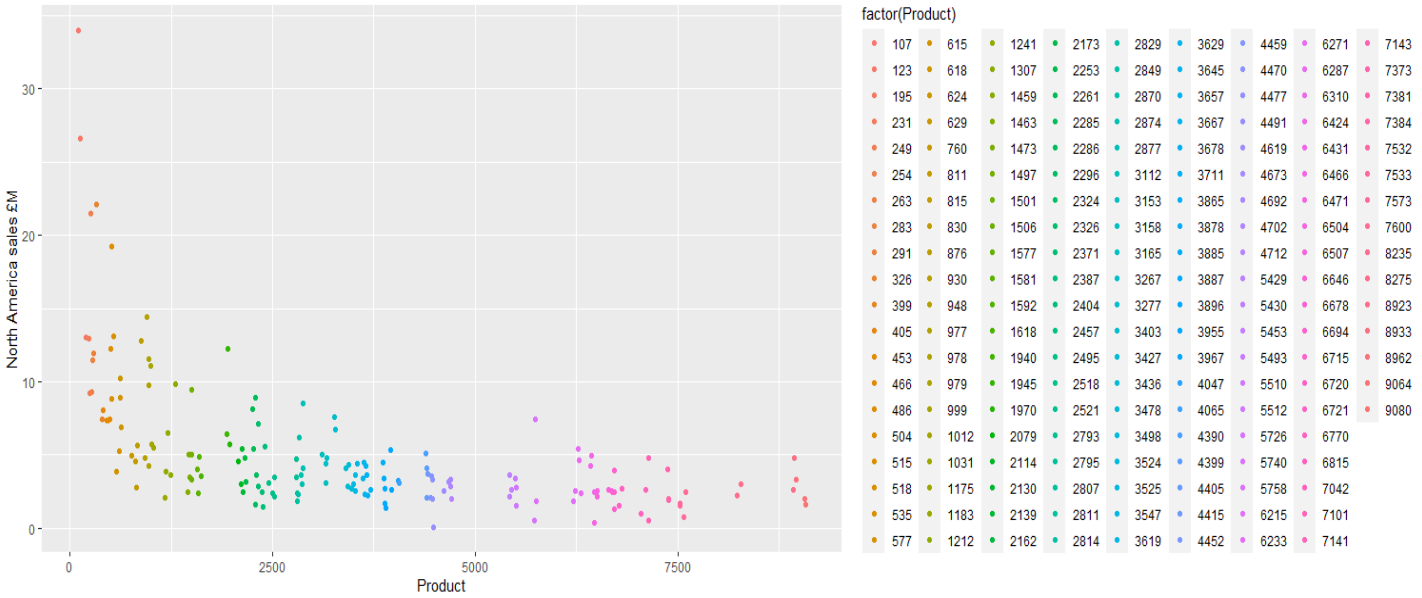

```
> summary(eu_product_sales)
    Product        eu_sum        eu_no_platforms
 Min.   : 107   Min.   : 0.000   Min.   :1.000
 1st Qu.:1468   1st Qu.: 1.460   1st Qu.:1.000
 Median :3158   Median : 2.300   Median :1.000
 Mean   :3490   Mean   : 3.306   Mean   :2.011
 3rd Qu.:5442   3rd Qu.: 4.025   3rd Qu.:2.000
 Max.   :9080   Max.   :23.800   Max.   :9.000
```

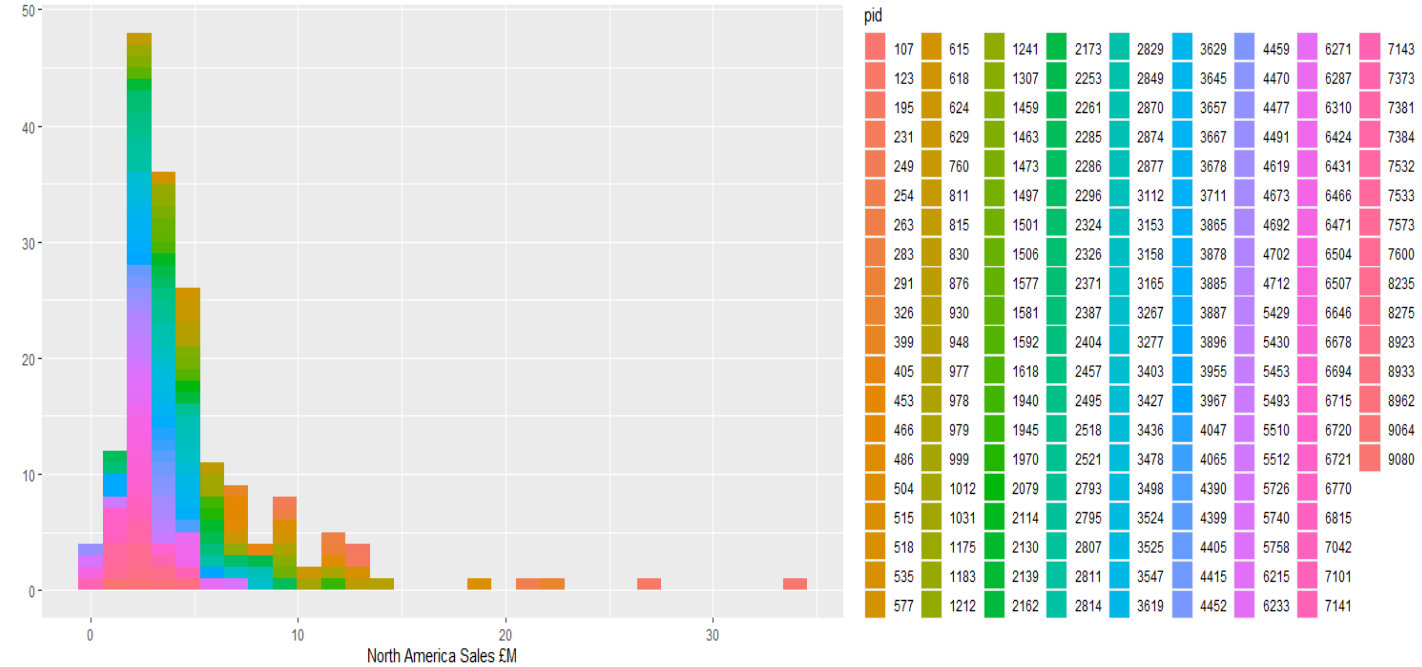
Product ids by North America sales (£M)

	Product	na_sum	na_no_platforms
1	107	34.02	1
2	123	26.64	2
3	326	22.08	1
4	254	21.46	2
5	515	19.25	5

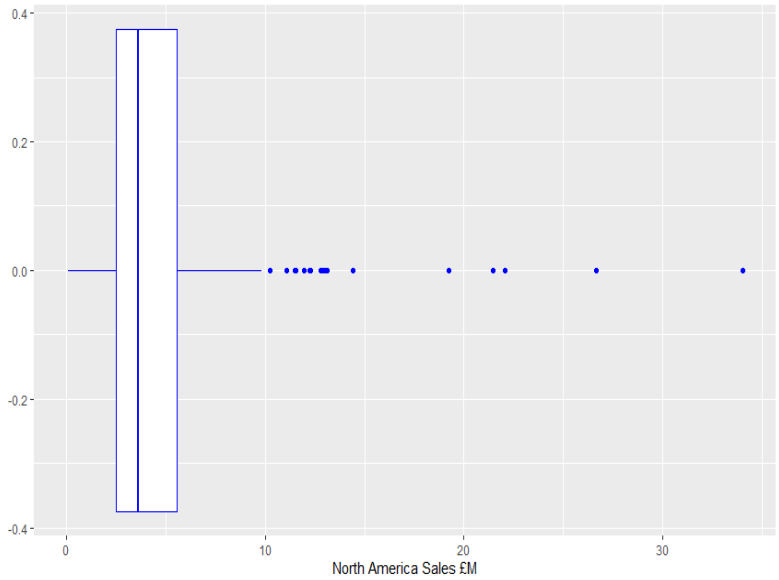
	Product	na_sum	na_no_platforms
1	3645	2.33	9
2	2518	2.18	8
3	3967	2.63	8
4	3887	1.71	7
5	9080	1.59	7

Top 5 products ids with the highest North America sales are **107, 123, 326, 254** and **515**.They run across from 1 to maximum 5 platforms. The product ids which runs on many different platforms do not generate the highest sales in North America and are in the higher range products ids from 3000 and above.





In North America, on average the sales is around £5 M. There are many outliers with which generate sales over £10M. One significant outlier which product a sales of over £34 M.




```
> summary(na_product_sales)
    Product          na_sum        na_no_platforms
 Min.   : 107   Min.   : 0.060   Min.   :1.000
 1st Qu.:1468   1st Qu.: 2.495   1st Qu.:1.000
 Median :3158   Median : 3.610   Median :1.000
 Mean   :3490   Mean   : 5.061   Mean   :2.011
 3rd Qu.:5442   3rd Qu.: 5.570   3rd Qu.:2.000
 Max.   :9080   Max.   :34.020   Max.   :9.000
> 
```

Insights and Observations:

There are over 100 products with over 20 different platforms. A mix of summary tables, scatterplots, histograms and boxplots are used to explore the trends and insights of the game sales at Turtle Games. Overall, scatterplots are best to compare the game sales because of the wide range of products. The visualisations are used to identify the sales trends, at this stage, it is not possible to visualise the individual product by colours.

X360, **PS3, PC, Wii** and **DS** are the most popular platforms. **Wii** is the most popular platform which generate the highest sales in both Europe and North America. Product ids **107** and **515** generate the top 5 highest sales in both Europe and North America. In Europe, one of the highest product id sales **3967** run on the maximum of **9 platforms**. The product ids with the highest sales in the North America tends to be lower numbers in comparison to Europe. On average, the sales in both Europe and North America are between £2M to £3M. There are many interesting outliers in both markets which generate significant higher sales. It is important to further analyse these product ids and platforms of these outliers to better understand the market trends. Product id **107** runs only on 1 platform which is **Wii** has a very significant impact across all sales in both Europe and America. Followed by product id **515** which runs on most of the top popular platforms **X360**, **PC** and **Wii**.

Clean, manipulate and visualise the data with R

Objectives:

- Explore, prepare and explain the normality of the data set based on plots, Skewness, Kurtosis and a Sharpiro-Wilk test
- Determine the impact on sales per product id

Explore the data sets grouped by products id

Min, Max, Mean and Median of all markets

```
> # Determine descriptive statistics of df from week 4 group by products
> summary(df)
    product         eu_sales         na_sales       other_sales      global_sales
 Min.   : 107   Min.   : 0.000   Min.   : 0.060   Min.   : 0.000   Min.   : 4.200
 1st Qu.:1468   1st Qu.: 1.460   1st Qu.: 2.495   1st Qu.: 1.095   1st Qu.: 5.515
 Median :3158   Median : 2.300   Median : 3.610   Median : 1.850   Median : 8.090
 Mean   :3490   Mean   : 3.306   Mean   : 5.061   Mean   : 2.363   Mean   :10.730
 3rd Qu.:5442   3rd Qu.: 4.025   3rd Qu.: 5.570   3rd Qu.: 3.155   3rd Qu.:12.785
 Max.   :9080   Max.   :23.800   Max.   :34.020   Max.   :10.030   Max.   :67.850
```

Europe sales have mean £3.31M and median £2.30M.

North America sales have mean £5.06M and median £3.61M.

Global sales have mean £10.73M and median £8.09M.

The means and medians of all markets sales are not very close and similar with a difference of £1M to £2M in all markets.

Mininum sales in Europe is £0M. North America has minimum sales of £0.06M. Global sales have a maximum with £67.85M.

Min, Max, Mean and Median of all markets after removing outliers

Subset the data sets with sales under £30M, removed the outliers.

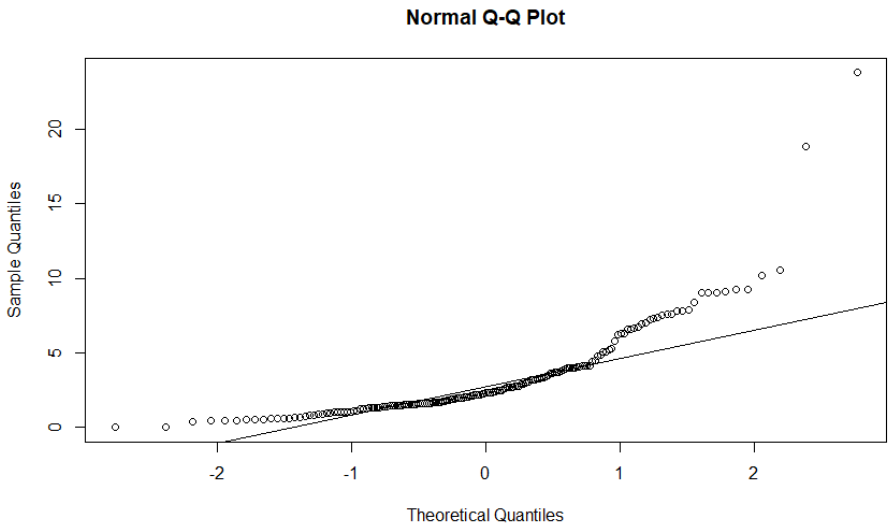
```
> # View descriptive statistics of df1 after removing outliers
> summary(df1)
    product         eu_sales         na_sales       other_sales      global_sales
 Min.   : 195   Min.   : 0.000   Min.   : 0.060   Min.   :0.000   Min.   : 4.200
 1st Qu.:1500   1st Qu.: 1.458   1st Qu.: 2.478   1st Qu.:1.085   1st Qu.: 5.500
 Median :3216   Median : 2.275   Median : 3.580   Median :1.815   Median : 8.035
 Mean   :3546   Mean   : 3.093   Mean   : 4.684   Mean   :2.263   Mean   :10.040
 3rd Qu.:5463   3rd Qu.: 4.000   3rd Qu.: 5.418   3rd Qu.:3.045   3rd Qu.:12.553
 Max.   :9080   Max.   :10.560   Max.   :22.080   Max.   :9.190   Max.   :29.390
> |
```

After removing the outliers, Europe sales have mean £3.09M. North America sales have mean £4.68M.

Global sales have mean £10.04M. The means and medians across all markets are not very close. Maximum sales in Europe and North America decrease by around £10M after removing the outliers. Other sales maintain similar maximum sales.

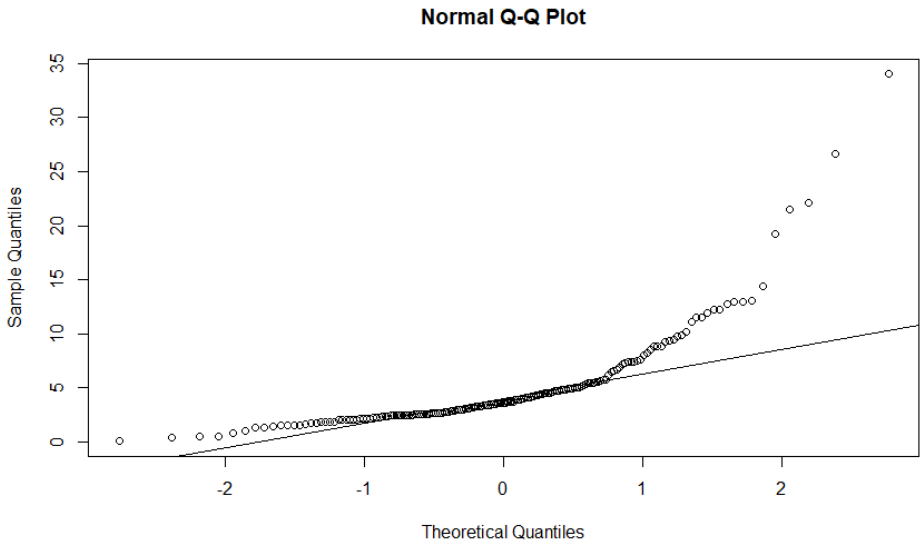
Determine the normality of data sets using Q-Q plots

Q-Q Plot Europe sales



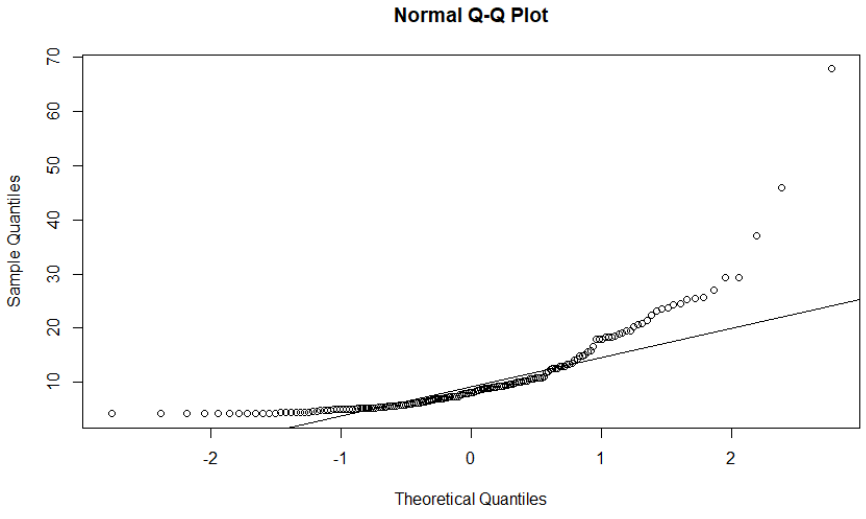
In Europe sales, the data sets in the middle range follow a straight line. The data sets in the bottom and top are further away from the line. The normality is unclear.

Q-Q plot North America sales



In North America, the data sets follow similar pattern as in Europe. Some data points are further away from the reference line.

Q-Q plot Global sales



Determine the normality of data sets using Shapiro-Wilk test

The p-values of Europe 1.42e-11, North America 4.618e-14 and Global 8.212e-13 are very small, less than 5%. The p-values suggest that the assumption of normality is a poor fit for the data sets.

Determine the normality of data sets using skewness and kurtosis

Europe sales:

1.247008 positive skew, right-skewed and biased towards higher values

3.719443 means Europe sales has a heavier tail, data is leptokurtik

North America sales:

2.097169 positive skew, right-skewed and biased towards higher values

8.877347 means North America sales has a heavier tail, data is leptokurtik

Global sales:

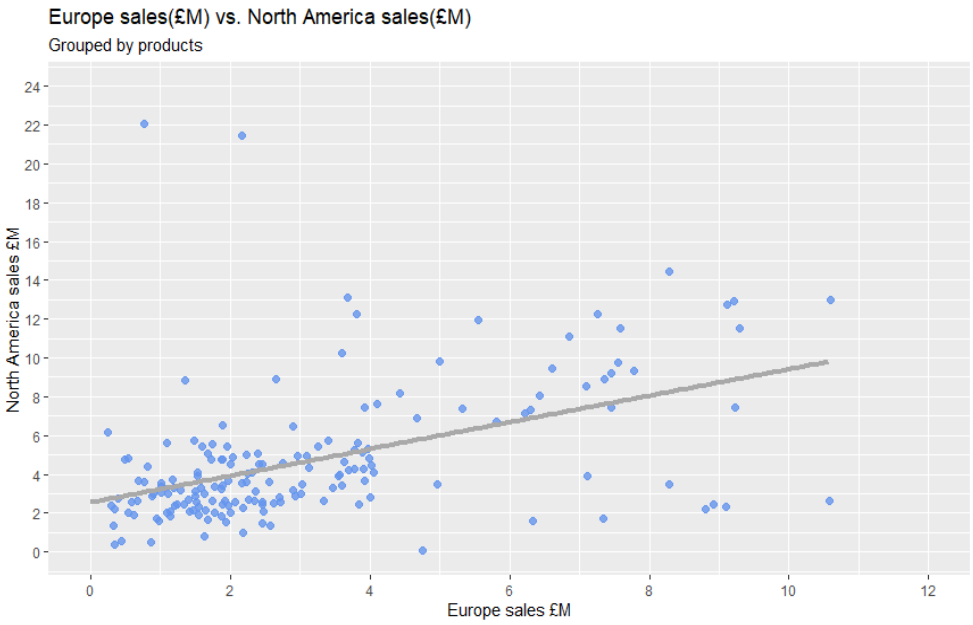
1.352425 positive skew, right-skewed and biased towards higher values

4.033009 means Global sales has a heavier tail, data is leptokurtic

Determine if there is any correlation between the sales data in Europe, North America and Other markets.

```
> # Determine the correlation for the whole data frame df1.
> round (cor(df1),
+        digits=2)
              product eu_sales na_sales other_sales global_sales
product          1.00    -0.47    -0.58       -0.56        -0.68
eu_sales        -0.47     1.00     0.47        0.41         0.78
na_sales        -0.58     0.47     1.00        0.38         0.87
other_sales     -0.56     0.41     0.38        1.00         0.66
global_sales    -0.68     0.78     0.87        0.66         1.00
```

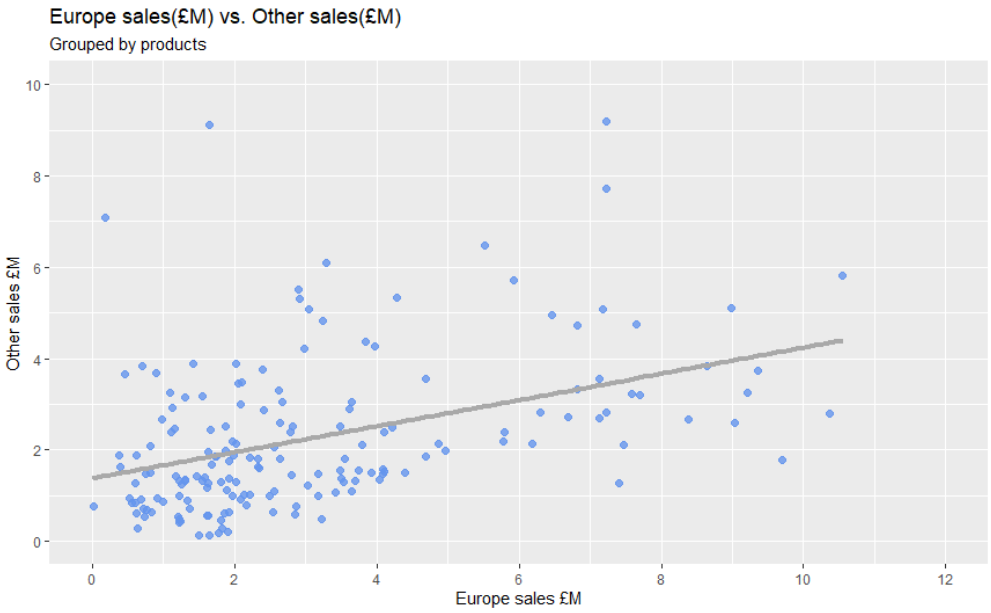
The correlation coefficients highlighted suggest positive correlations between Europe with North America and Others.



In Europe and North America, there is a positive trend in sales. Sales in both markets are more clustered when they are less than £4M.

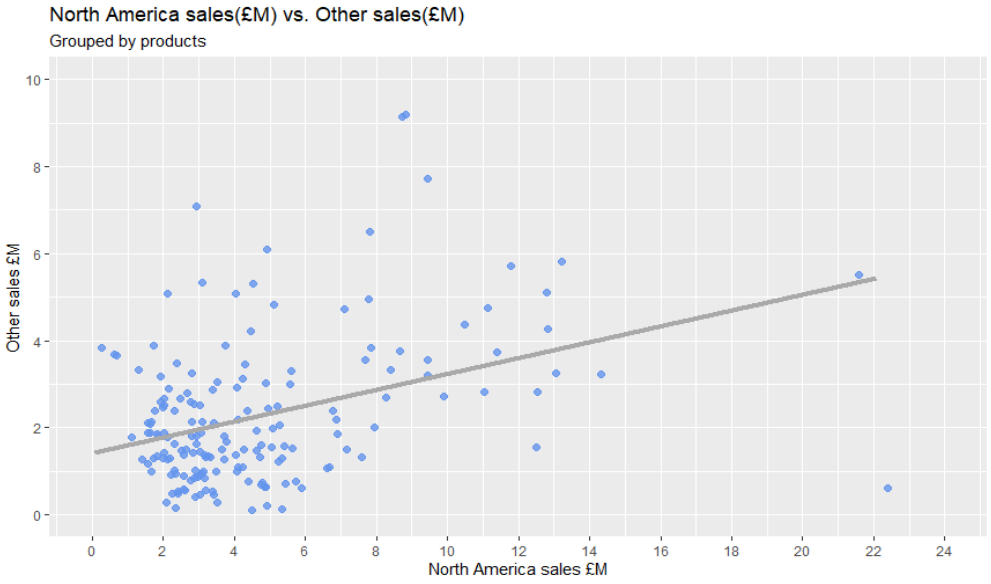
As the sales increase in Europe, similar pattern follow in North America. North America sales show pattern of higher figures compared to Europe for the same products. This implies North America is a bigger market.

Determine if there is any correlation between the sales data in Europe, North America and Other markets.



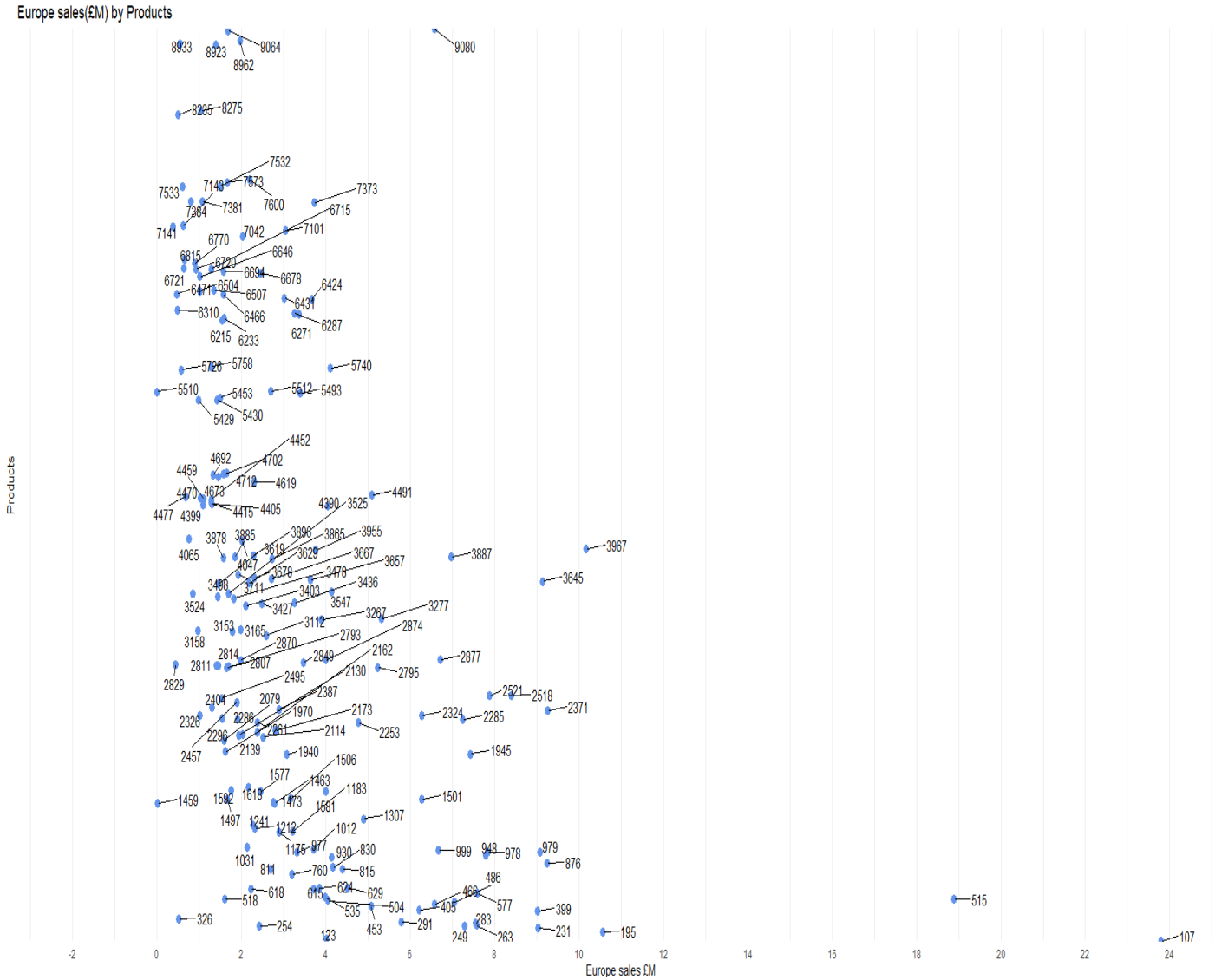
In Europe and Other sales, similar positive trends around £5M and lower.

Other sales are less than Europe sales by product, it tells that market of Other sales have smaller market compare to North America and Europe.



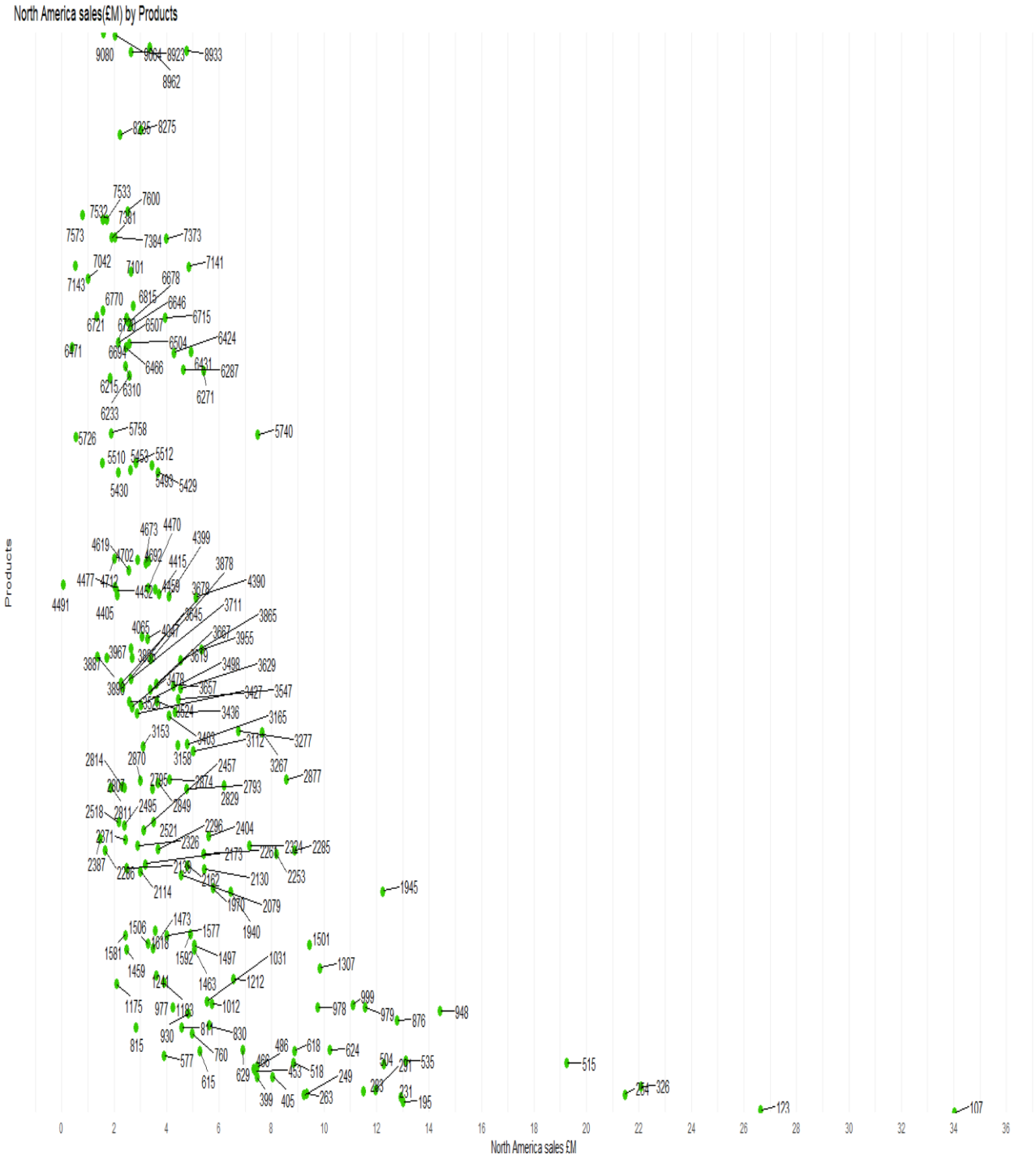
In North America and Other sales, a strong positive trend around £8M and lower. The market size of North America is bigger than Other sales.

Impact on Sales per product id



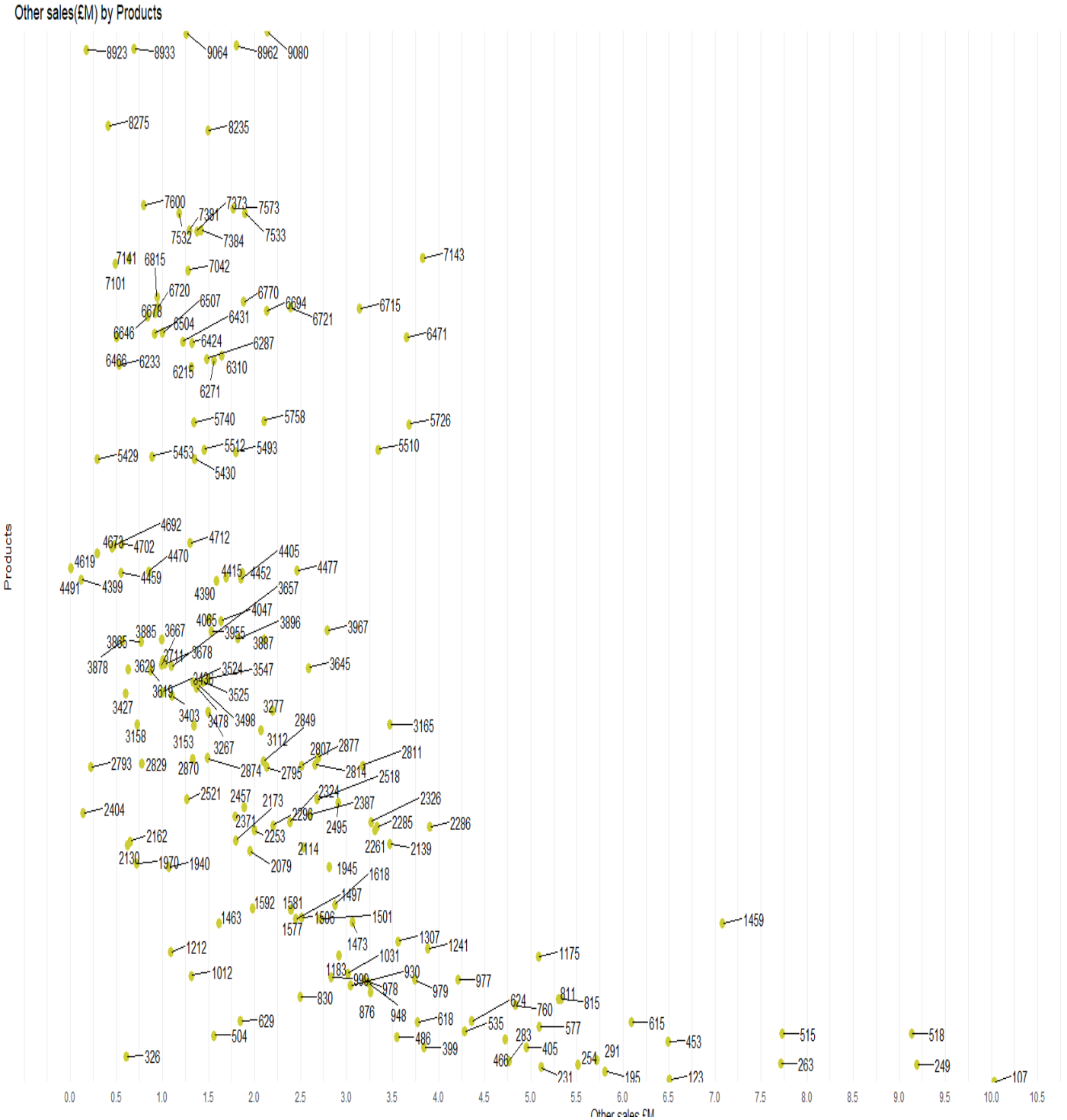
In Europe, product ids is **515** and **107** generate a very high significant sales compared to other products ids.

Most products generate sales around £6M or less. There is a group of products (eg. **195**, **231**, **2371**, **3645**, **3967**) which generate a higher middle range of sales.

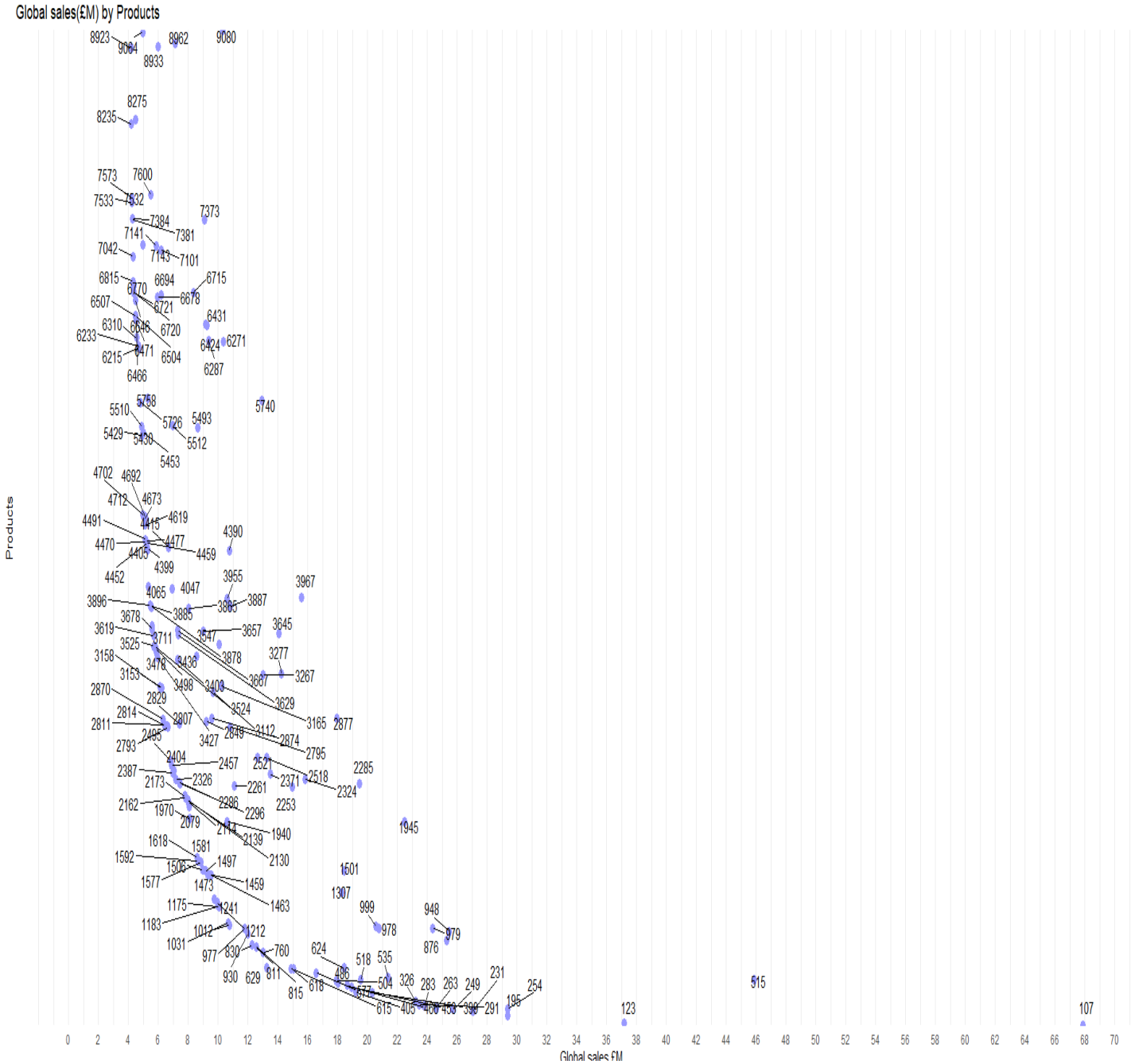


In North America, product ids is **515** and **107** generate a very high significant sales compared to other products ids, similar pattern as in Europe. North America has more products which generate higher sales such as products **123**, **254** and **326**.

Most products generate sales around £14M or less.



In other market, product ids is **515, 263, 518, 249** and **107** generate a very high significant sales compared to other products ids. These products follow similar trends in Europe and America. Markets sales per product is lower in other.



Global trends shows more product ids with higher number generate more sales but the sales tend to be less than £20M. A few product ids with lower numbers ranging from 100 to around 1000, in contrast, generate higher significant sales.

Observations and insights:

The means and medians across all markets are not close and similar. Even after removing outliers of over £30M, the patterns are similar. The maximum sales decrease in Europe and North America while other sales maintain. This implies the outliers have a significant effect on the maximum sales in Europe and North America.

The normality of the data sets are tested using 3 different methods:

Q-Q plots:

The data sets of all sales tend to follow the reference line in the middle ranges, and with data points move further away at the bottom and the top. The Q-Q plots cannot suggest normality of the data sets.

Shapiro-Wilk test:

The p-values of all market sales are very small less than 5%. These suggest normality is a poor fit for the sales data.

Skewness and kurtosis:

The results from all markets suggest data sets are leptokurtic, positive skew, right-skewed and biased towards higher values with heavier tail.

The results and trends show positive correlations between Europe, North America and Others.

The correlations suggest stronger positive correlations with sales lower than £6M. For the same product ids, North America tends to generate higher sales compared to Europe with the same products. One reason could be that North America is a bigger country with higher population.

The product ids have an impact on the sales in all markets. Products with lower number ranging from 100 to 1000 generate higher sales in all markets. Product ids **515, 263, 518, 249** and **107** are top products which generate the highest sales. Global trends show product ids with higher number generate more sales but lower. They tend to be less than £20M.

Making recommendations to the business

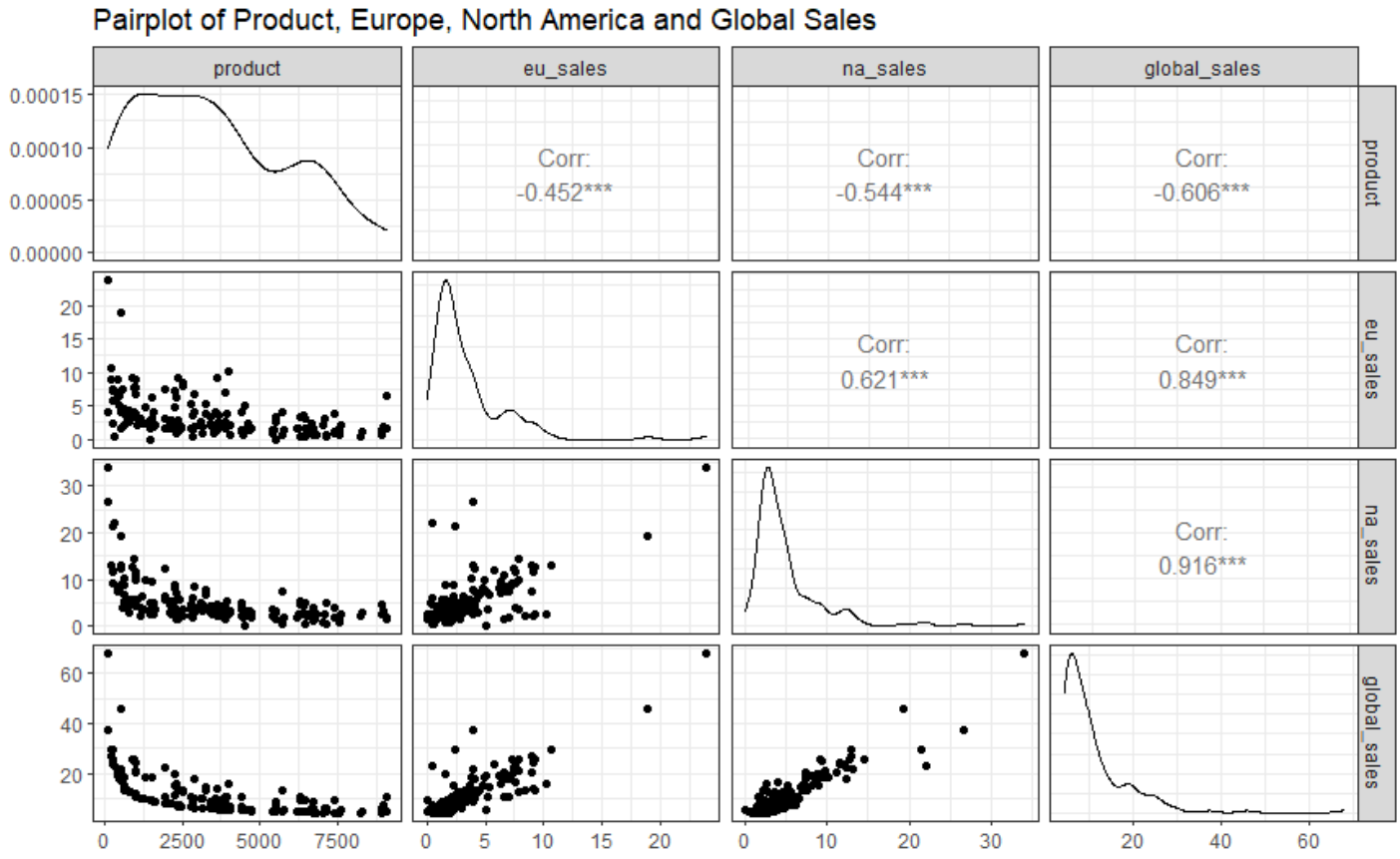
The sales department wants to understand the whether there are any relationships between sales in Europe, North America and Global.

Objectives:

- **Determine if there is any relationships in sales by continents**

Using the sales data sets with 175 rows including the outliers, identify variables with correlation.

```
> # Determine a summary of the sales data frame.
> summary(sales_df)
    product        eu_sales         na_sales       global_sales
 Min.   : 107   Min.   : 0.000   Min.   : 0.060   Min.   : 4.200
 1st Qu.:1468   1st Qu.: 1.460   1st Qu.: 2.495   1st Qu.: 5.515
 Median :3158   Median : 2.300   Median : 3.610   Median : 8.090
 Mean   :3490   Mean   : 3.306   Mean   : 5.061   Mean   :10.730
 3rd Qu.:5442   3rd Qu.: 4.025   3rd Qu.: 5.570   3rd Qu.:12.785
 Max.   :9080   Max.   :23.800   Max.   :34.020   Max.   :67.850
`
> ## Determine the correlation between columns
> cor(sales_df)
                product   eu_sales   na_sales global_sales
product       1.0000000 -0.4524737 -0.5435505   -0.6061376
eu_sales     -0.4524737  1.0000000  0.6209317    0.8486148
na_sales     -0.5435505  0.6209317  1.0000000    0.9162292
global_sales -0.6061376  0.8486148  0.9162292    1.0000000
```



Global sales shows very strong positive correlations with North America and Europe sales, and fairly strong negative correlation with products.

Linear regression Global sales ~ Europe sales

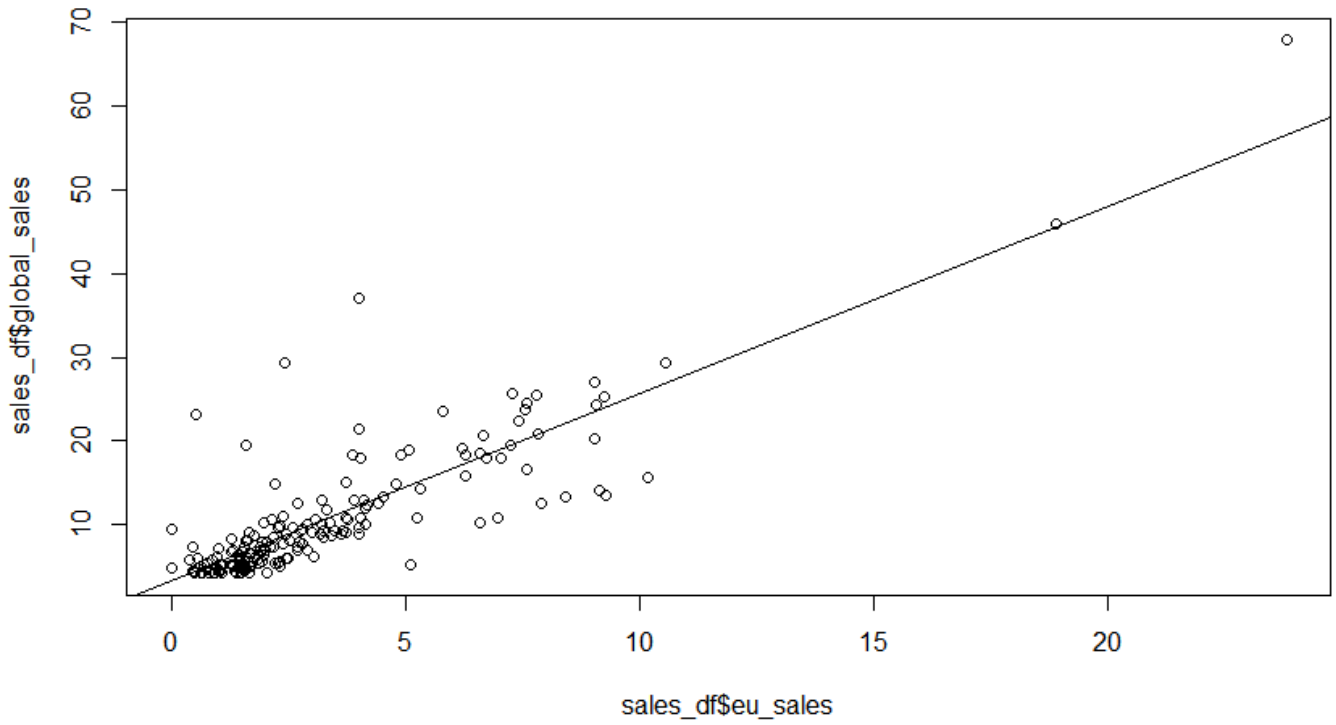
```
> # View the model1 global sale ~ Europe sales
> model1

Call:
lm(formula = global_sales ~ eu_sales, data = sales_df)

Coefficients:
(Intercept)     eu_sales
      3.334        2.237
```

Coefficient shows global sales will go up by £2.237M with every increment of Europe sales

```
> # View full regression table of model1 - global_sales~eu_sales
> summary(model1)

Call:
lm(formula = global_sales ~ eu_sales, data = sales_df)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5583  -1.7530  -0.5371   0.9586  24.8556

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3343     0.4787   6.965 6.57e-11 ***
eu_sales      2.2369     0.1060  21.099  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.313 on 173 degrees of freedom
Multiple R-squared:  0.7201,    Adjusted R-squared:  0.7185
F-statistic: 445.2 on 1 and 173 DF,  p-value: < 2.2e-16
```

t-value of Europe sales is 21.099, means estimates of the slope coefficient is 21.099 standard errors away from 0, a lot of standard errors. The p-value of Europe sales is 2e-16, very small. This suggests that Europe sales is a highly significant variable. The multiple R-squared of 72.01% explains that Europe sales explains 72.01% of the variability in the global sales variable.

Europe sales is a highly significant value, explaining over 72.01% of the variability.



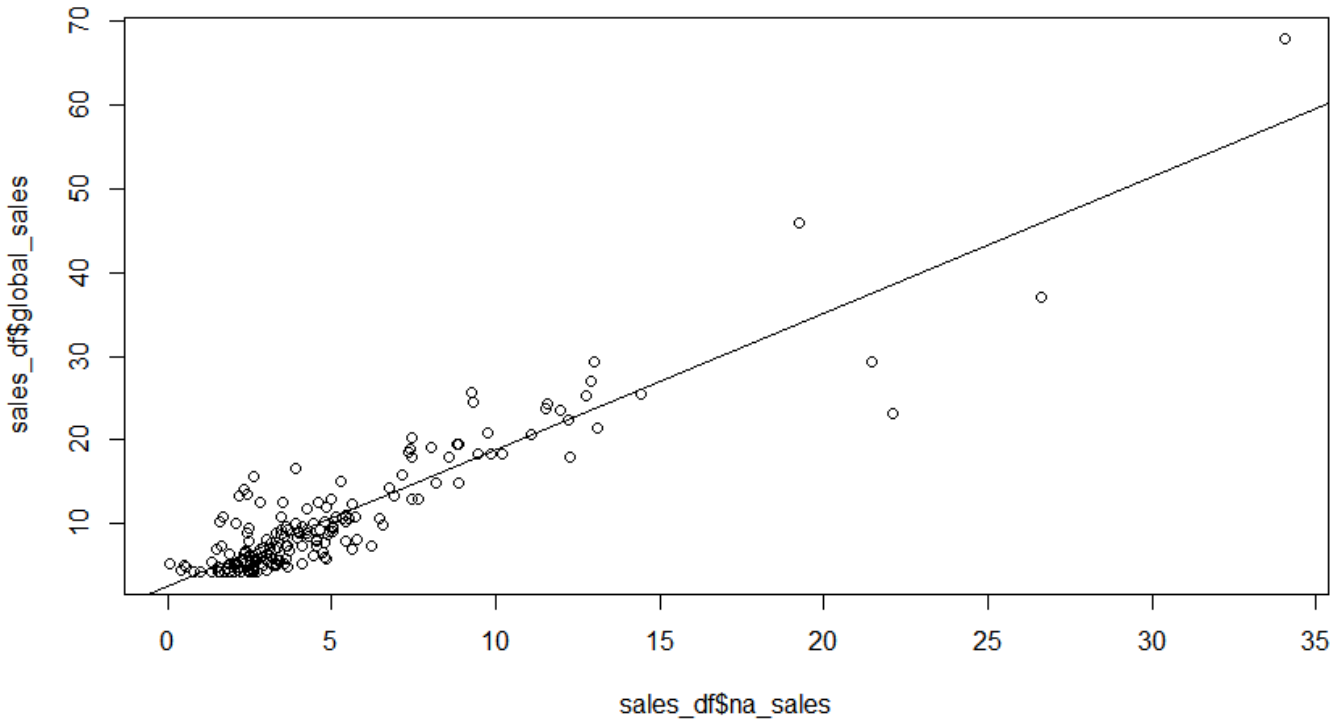
Linear regression Global sales ~ North America sales

```
> # View the model2 global sale ~ North America sales
> model2

Call:
lm(formula = global_sales ~ na_sales, data = sales_df)

Coefficients:
(Intercept)     na_sales
      2.458        1.635
```

Coefficient shows global sales will go up by £1.635M with every increment of North America sales

```
> # View full regression table of model1 - global_sales~na_sales
> summary(model2)

Call:
lm(formula = global_sales ~ na_sales, data = sales_df)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3417  -1.8198  -0.5933   1.4322  11.9345

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.45768    0.36961   6.649 3.71e-10 ***
na_sales     1.63469    0.05435  30.079  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.266 on 173 degrees of freedom
Multiple R-squared:  0.8395,    Adjusted R-squared:  0.8385
F-statistic: 904.7 on 1 and 173 DF,  p-value: < 2.2e-16
```

t-value of North America sales is 30.079, means estimates of the slope coefficient is 30.079 standard errors away from 0, a lot of standard errors. The p-value of North America sales is 2e-16, very small. This suggests that North America sales is a highly significant variable. The multiple R-squared of 83.95% explains that North America sales explains 83.95% of the variability in the global sales variable.



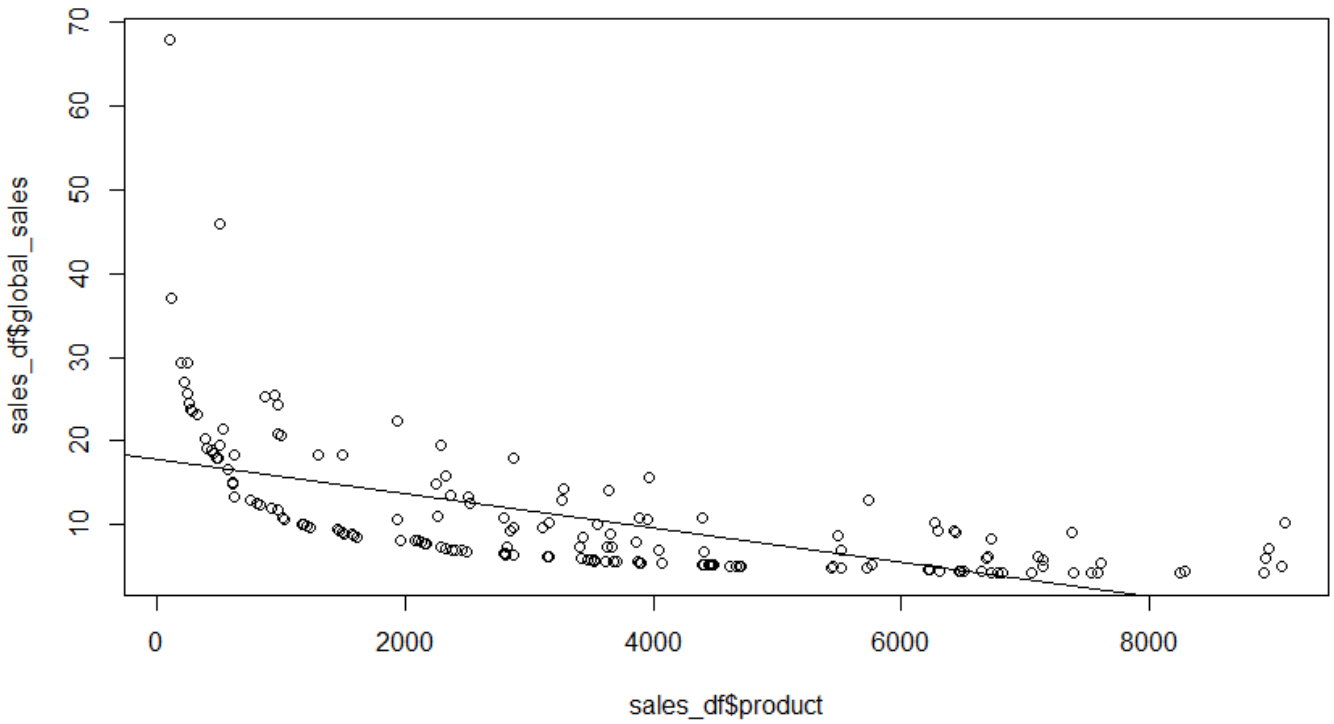
Linear regression Global sales ~ product

```
> # View the model3 global sales ~ product
> model3

Call:
lm(formula = global_sales ~ product, data = sales_df)

Coefficients:
(Intercept)      product
  17.859540    -0.002043
```

Coefficient shows global sales will decrease by £0.002043M with every increment of product id.

```
> # View full regression table of model3 - global_sales~product
> summary(model3)

Call:
lm(formula = global_sales ~ product, data = sales_df)

Residuals:
   Min     1Q Median     3Q    Max
-6.023 -4.685 -1.372  2.638 50.209

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.8595399  0.8637784   20.68   <2e-16 ***
product     -0.0020428  0.0002038  -10.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.484 on 173 degrees of freedom
Multiple R-squared:  0.3674,    Adjusted R-squared:  0.3637
F-statistic: 100.5 on 1 and 173 DF,  p-value: < 2.2e-16
```

t-value of product is 10.02, means estimates of the slope coefficient is 10.02 standard errors away from 0, a fair number of standard errors. The p-value of product is 2e-16, very small. This suggests that product is a highly significant variable. The multiple R-squared of 36.74% explains that product explains 36.74% of the variability in the global sales variable.

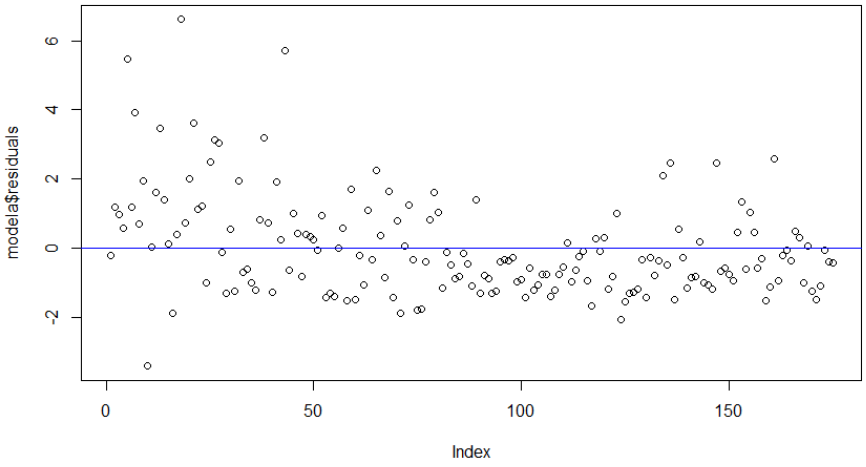


Multiple Linear Regression Model – Global sales ~ Europe Sales and North America Sales

```
> # Summary statistics of modela
> summary(modela)

Call:
lm(formula = global_sales ~ eu_sales + na_sales, data = sales_df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4156 -1.0112 -0.3344  0.6516  6.6163

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.04242    0.17736   5.877 2.11e-08 ***
eu_sales     1.19992    0.04672  25.682  < 2e-16 ***
na_sales     1.13040    0.03162  35.745  < 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
```

Adjusted R-squared is 0.9664, very high, it means that the model is a good fit with very high correlation.

Multiple R-squared of 0.9668 means that 96.68% of the variability observed of the Global sales is explained by Europe Sales and North America sales.



There are many positive and negative residual errors, meaning some predicitons are either too high or too low.

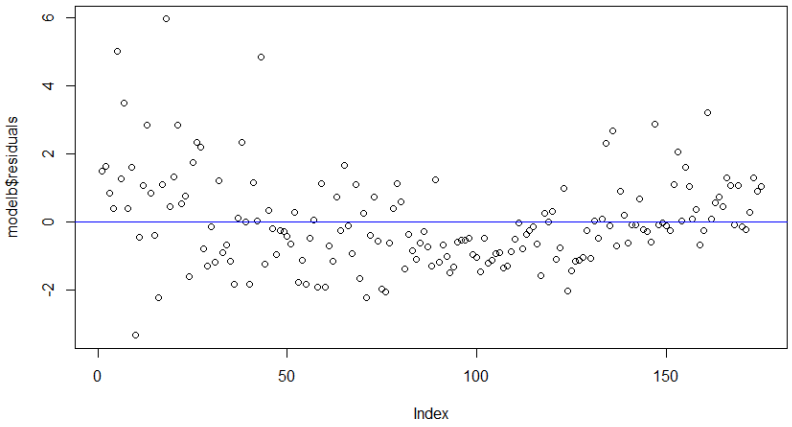
Multiple Linear Regression Model – Global sales ~ Product, Europe Sales and North America Sales

```
> # Summary statistics of modelb
> summary(modelb)

Call:
lm(formula = global_sales ~ product + eu_sales + na_sales, data = sales_df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3388 -0.9149 -0.2399  0.7364  5.9643

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.451e+00  3.167e-01   7.741 8.24e-13 ***
product     -2.753e-04  5.278e-05  -5.215 5.26e-07 ***
eu_sales     1.160e+00  4.421e-02  26.233  < 2e-16 ***
na_sales     1.068e+00  3.179e-02  33.601  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.388 on 171 degrees of freedom
Multiple R-squared:  0.9714,    Adjusted R-squared:  0.9709
F-statistic:  1933 on 3 and 171 DF,  p-value: < 2.2e-16
```

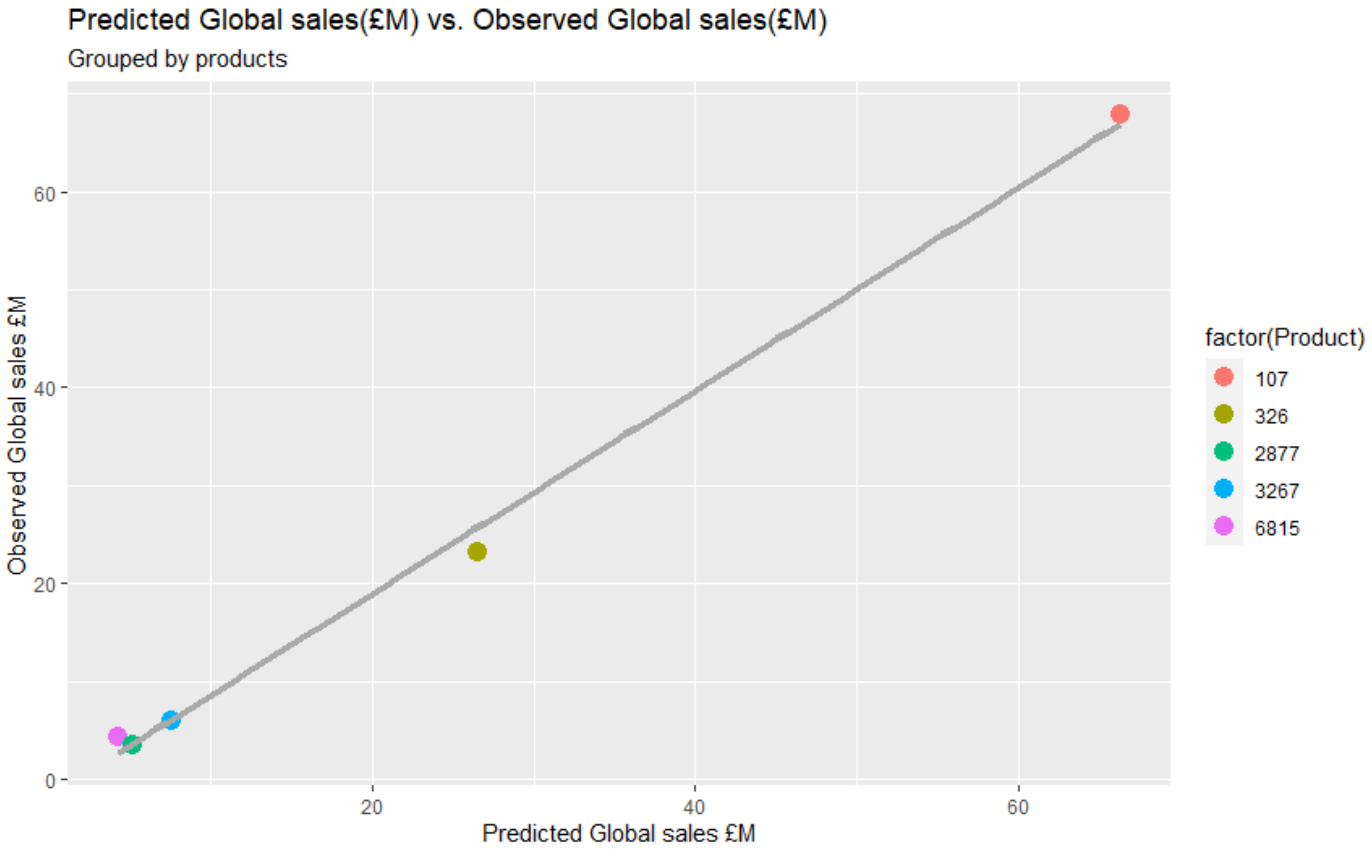
Adjusted R-squared is 0.9709, very high, it means that the model is a good fit with very high correlation.

Multiple R-squared of 0.9714 means that 97.14% of the variability observed of the Global sales is explained by product, Europe Sales and North America sales.

Predictions using Model – Global Sales ~ Product, Europe Sales and North America Sales with the highest adjusted R-squared



Residual errors of modelb tells that many predictions are negative, meaning predictions could be too high. A few along the 0 which are very accurate. Some positive ones which tell that predictions could be too low.



	Product	Platform	NA_Sales	EU_Sales	Global_Sales	fit	lwr	upr
1	107	Wii	34.02	23.80	67.85	66.358702	64.712584	68.004821
99	3267	X360	3.93	1.56	6.04	7.558873	7.307227	7.810518
176	6815	N64	2.73	0.65	4.32	4.245235	3.873852	4.616618
211	2877	X360	2.26	0.97	3.53	5.198279	4.887641	5.508917
10	326	NES	22.08	0.52	23.21	26.548792	25.376282	27.721303

The predicted global sales are fairly closed to the observed ones so modelb is a good fit.

Observations and insights:

There are some very strong positive correlations among global sales, Europe and North America sales. With products, the global sales have a negative correlation. Since they are all significant, they are tested to model the prediction of global sales.

Both modela (global sales~Europe Sales+ North America sales) and modelb (global sales ~ Product + Europe sales + North America sales) have very high adjusted R-squared of over 0.9 and closer to 1.

Modelb is slightly higher than modela, and thus a better fit. From the results of the predicted global sales, modelb produces fairly accurate predictions comparing to observed values. Sales team could use modelb to help predicting future sales.

To further improve the models, it is recommended to perform log transformation on the variables because the data sets do not follow a bell shape and have some visible outliers. Transforming the data using log may help to reduce or remove the skewness of the data sets and give even better predictions.

Conclusion and recommendations:

Spending scores and remuneration have a positive correlation with loyalty point. These groups of customers have higher spending power, and spend more at Turtle Games. They could be grouped into 5 segments based on their spending scores and remunerations. Marketing team could use these segments to target their marketing campaigns.

Overall, the reviews and summary from customers are positive about the products.. Many insights are identified by frequency of words. Games are the most popular products. Sales team could, therefore, continue to push the sales in games products range.

Further analysis could be carried out to have more insights about causes of negative reviews, from defective products, service, delivery or any unknows. This will help sales team to improve on the areas identified.

North America tend to have higher sales than Europe for the same products, the reason could be that it is a bigger country with a higher population and more spending power. Sales team could target the sales more in North America to generate higher revenues.

Products sales have follow products ids trend. Sales are higher with products ids lower than 1000, and decreases as the product ids get higher. Product id 107 running on a single platform produces the highest sales in all. Sales team could segment the products by ids to target the sales.