PREDICTIVE ANALYSIS OF THE TURTLE GAMES DATA SET

3rd January 2023
Prepared by: Oksana Fedorova
e-mail address: OksanaF@email.com

1. Context of the business issue and aim of the data analytics project.

Turtle Games, a game manufacturer and retailer, has a business objective of improving overall sales performance by utilising customer trends.

This data analytics project aims to help Turtle Games improve sales performance by identifying:

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

2. Analytical approach and discovered insights.

2.1. Analytical approach.

2.1.1. GitHub repository.

The LSE_DA301_Assignment-Predicting-future-outcomes repository was created on https://github.com/ website to store, update, manage project files and allow easy collaboration for the team members working on the project.

Repository URL: https://github.com/SanaFed/LSE_DA301_Assignment-Predicting-future-outcomes.git

2.1.2. Initial exploration of the turtle_reviews.csv data file.

- The workstation was prepared by importing the necessary libraries and turtle_reviews.csv data file in a new Python3 file.
- The file was converted to a DataFrame and sense-checked.
- No missing values were identified.
- The data set contains information about 2000 product reviews including some customer demographics.

2.1.3. Investigation of the possible relationships between the loyalty points, age, remuneration, and spending scores.

- The initial DataFrame was modified by removing redundant columns ('language' and 'platform') for future analysis.
- Simple linear regression model was applied to learn relationships between:

a) spending score vs loyalty points.

Output of the model:

- t-value tests (P>|t|) for x coefficient = 0 < 0.05, thus the spending score (x) is significant in predicting loyalty points (y);
- R-squared = 0.452 => 45.2% of the total variability of y (loyalty points), is explained by the variability of x (spending score);
- Coefficient of x = 33.0617 => if the spending score (x) will increase by 1 unit, the number of loyalty points (y) will increase by 33.06 units.

Output of Breusch-Pagan test: LM Test p-value = 5.04e-139 < 0.05 => we fail to accept the H_0 and assume that **heteroscedasticity is present**.

Log10 transformation on the dependent variable (y = loyalty_points) as well as Weighted Least Squares (WLS) models with weight = 1/(y_pred_sslp squared) and weight = 1/(x_sslp squared) were applied to reduce heteroscedasticity, but unsuccessfully.

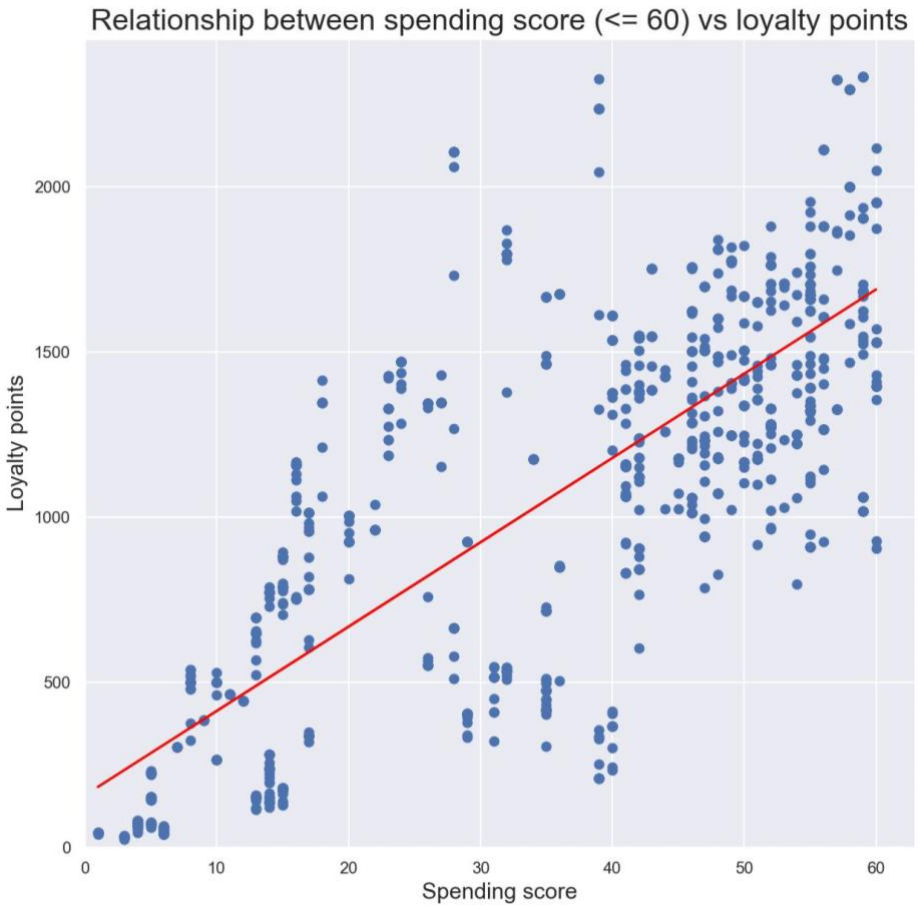
The decision was made to split the data in spending_score column by 0-60 and 61+ scores and apply the simple linear regression model on each sub-set again.

b) spending score (<= 60) vs loyalty points.

Output of the model:

- t-value tests (P>|t|) for x coefficient = 0 < 0.05, thus the spending score (x) is significant in predicting loyalty points (y);
- R-squared = 0.601 => 60.1% of the total variability of y (loyalty points), is explained by the variability of x (spending score);
- Coefficient of x = 25.5150 => if the spending score (x) will increase by 1 unit, the number of loyalty points (y) will increase by 25.52 units.

Output of Breusch-Pagan test: LM Test p-value = 0.39 > 0.05 => we fail to reject the H_0 and assume homoscedasticity.



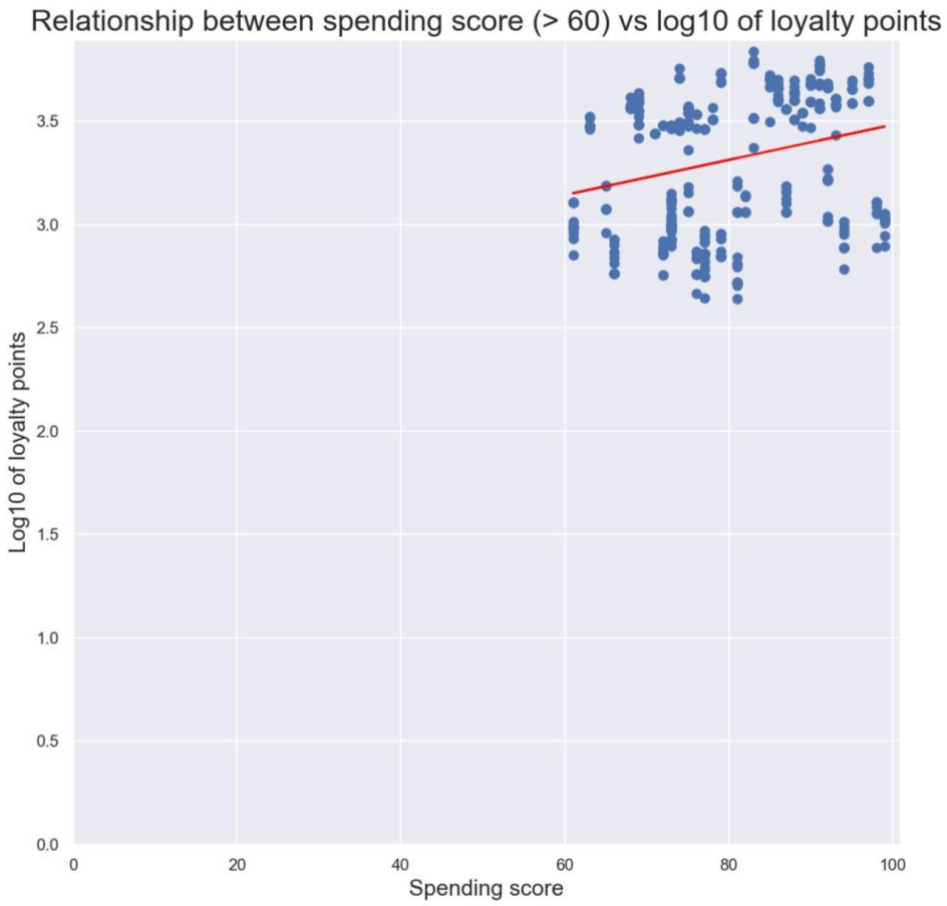
c) spending score (> 60) vs log10 of loyalty points.

Heteroscedasticity was detected when run linear regression model on spending score (> 60) vs loyalty points, therefore log10 transformation on the dependent variable (y = loyalty_points) was applied.

Output of the model:

- t-value tests (P>|t|) for x coefficient = 0 < 0.05, thus the spending score (x) is significant in predicting loyalty points (y);
- R-squared = 0.068 => 6.8% of the total variability of y (loyalty points), is explained by the variability of x (spending score);
- Coefficient of x = 0.0085 => if the spending score (x) will increase by 1 unit, the log10 of loyalty points (log10 of y) will increase by 0.0085 unit, therefore number of loyalty points (y) will increase by 1.02 units.

Output of Breusch-Pagan test: LM Test p-value = 0.23 > 0.05 => we fail to reject the H_0 and assume homoscedasticity.



d) remuneration vs loyalty points.

Output of the model:

- t-value tests (P>|t|) for x coefficient = 0 < 0.05, thus the remuneration (x) is significant in predicting loyalty points (y);
- R-squared = 0.380 => 38.0% of the total variability of y (loyalty points), is explained by the variability of x (remuneration);
- Coefficient of x = 34.1878 => if the remuneration (x) will increase by 1 unit (£1000), the number of loyalty points (y) will increase by 34.19 units.

Output of Breusch-Pagan test: LM Test p-value = 7.15e-228 < 0.05 => we fail to accept the H_0 and assume that **heteroscedasticity is present**.

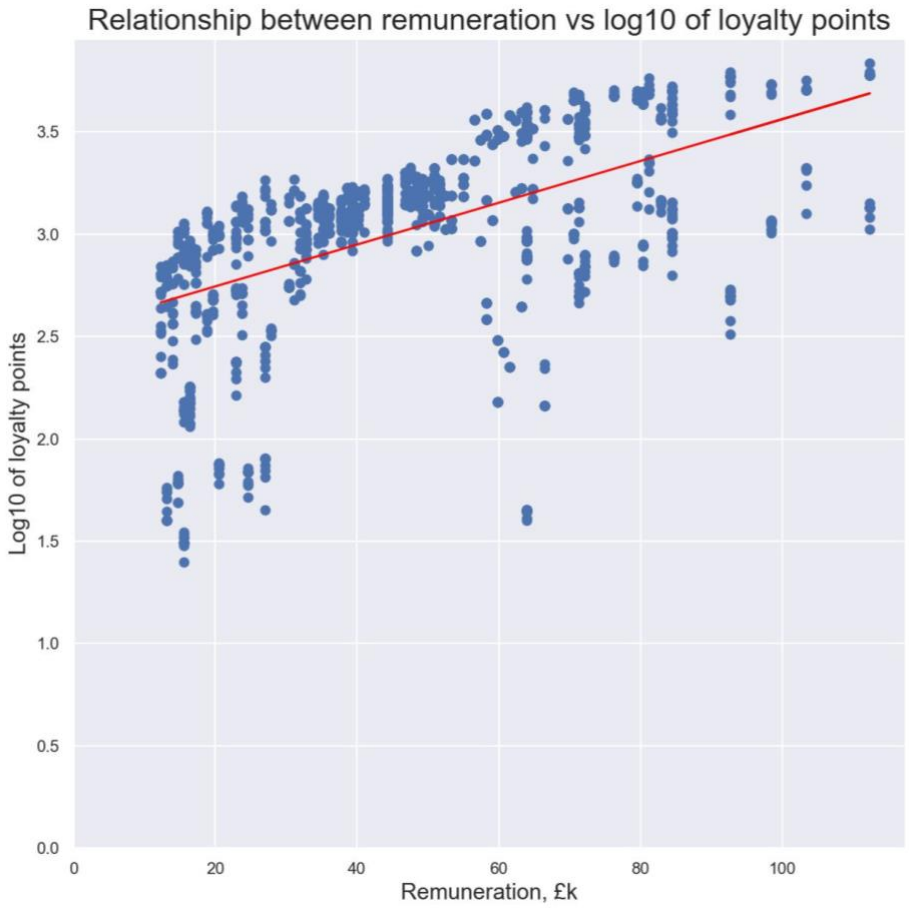
Log10 transformation on the dependent variable (y = loyalty_points) was applied to reduce heteroscedasticity.

e) remuneration vs log10 of loyalty points.

Output of the model:

- t-value tests (P>|t|) for x coefficient = 0 < 0.05, thus the remuneration (x) is significant in predicting loyalty points (y);
- R-squared = 0.284 => 28.4% of the total variability of y (loyalty points), is explained by the variability of x (remuneration);
- Coefficient of x = 0.0102 => if the remuneration (x) will increase by 1 unit (= £1,000), the log10 of loyalty points (y) will increase by 0.0102 units, therefore number of loyalty points (y) will increase by 1.024 units (= 1.02).

Output of Breusch-Pagan test: LM Test p-value = 0.83 > 0.05 => we fail to reject the H_0 and assume homoscedasticity.



f) age vs loyalty points.

Output of the model:

- t-value tests (P>|t|) for x coefficient = 0.058 > 0.05, thus the spending score (x) is not significant in predicting loyalty points (y);
- R-squared = 0.002 => 0.2% of the total variability of y (loyalty points), is explained by the variability of x (spending score);
- Coefficient of x = -4.0128 => if the spending score (x) will increase by 1 unit, the number of loyalty points (y) will decrease by 4.01 units.

Output of Breusch-Pagan test: LM Test p-value = 0.0003 < 0.05 => we fail to accept the H_0 and assume that **heteroscedasticity is present**.

Log10 transformation on the dependent variable (y = loyalty_points) was applied to reduce heteroscedasticity.

- Multiple linear regression model was applied to learn relationships between:

a) spending score and remuneration vs loyalty points.

Output of the model:
- t-value tests (P>|t|) for x coefficients = 0 < 0.05, thus the spending score and remuneration (X) are significant in predicting loyalty points (y);
- R-squared = 0.827 => 82.7% of the total variability of y (loyalty points), is explained by the variability of X (spending score and remuneration);
- Adj. R-squared = 0.827.

Test for multicollinearity: VIF factors = 1 => no correlation between independent variables.
Output of Breusch-Pagan test: LM Test p-value = 2.15e-12 < 0.05 => we fail to accept the H_0 and assume that **heteroscedasticity is present**.

Log10 transformation on the dependent variable (y = loyalty_points) as well as a data split (by spending score values <= 60 and >60) were applied to reduce heteroscedasticity but unsuccessfully.

b) spending score, remuneration and age vs loyalty points.

Output of the model:
- t-value tests (P>|t|) for x coefficients = 0 < 0.05, thus the spending score, remuneration and age (X) are significant in predicting loyalty points (y);
- R-squared = 0.840 => 84.0% of the total variability of y (loyalty points), is explained by the variability of X (spending score, remuneration and age);
- Adj. R-squared = 0.840.

Test for multicollinearity: VIF factors = between 1 and 1.05 => no correlation between independent variables.
Output of Breusch-Pagan test: LM Test p-value = 2.10e-09 < 0.05 => we fail to accept the H_0 and assume that **heteroscedasticity is present**.

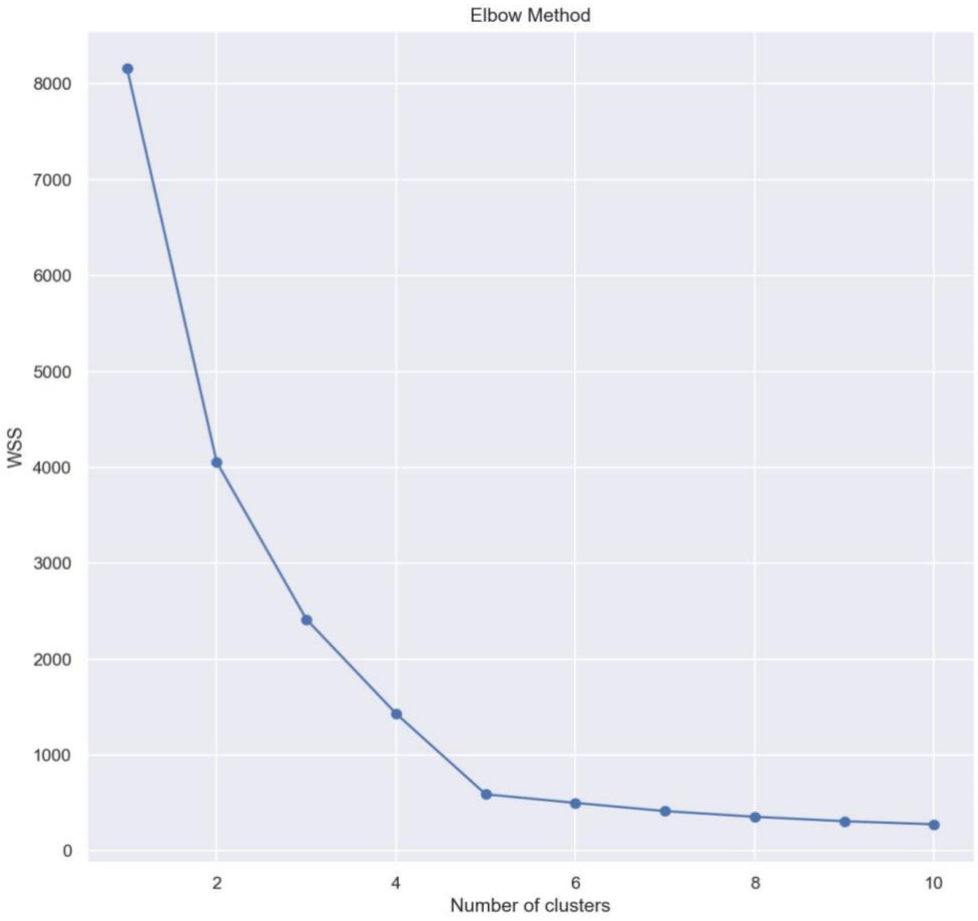
Log10 transformation on the dependent variable (y = loyalty_points) as well as a data split (by spending score values <= 60 and >60) were applied to reduce **heteroscedasticity** but unsuccessfully.

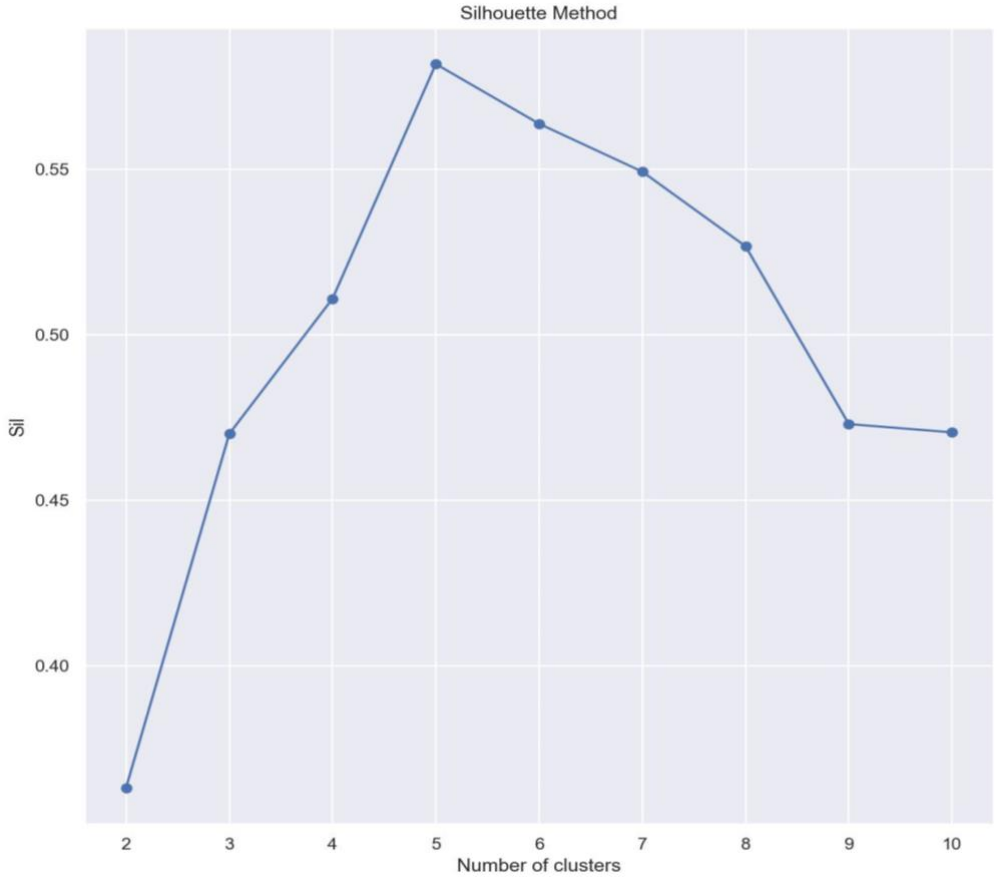
- **Summary:**
Multiple linear regression models have produced larger R-squared values than simple linear regression models but we were unable to eliminate heteroscedasticity, therefore we will be relying on the homoscedastic simple linear regression models.

The variability of loyalty points can be the best explained by the spending score (when it is <= 60) followed by the remuneration, and the spending score (when it is > 60). Age is not significant in predicting loyalty points unless it is part of the multiple linear regression model.

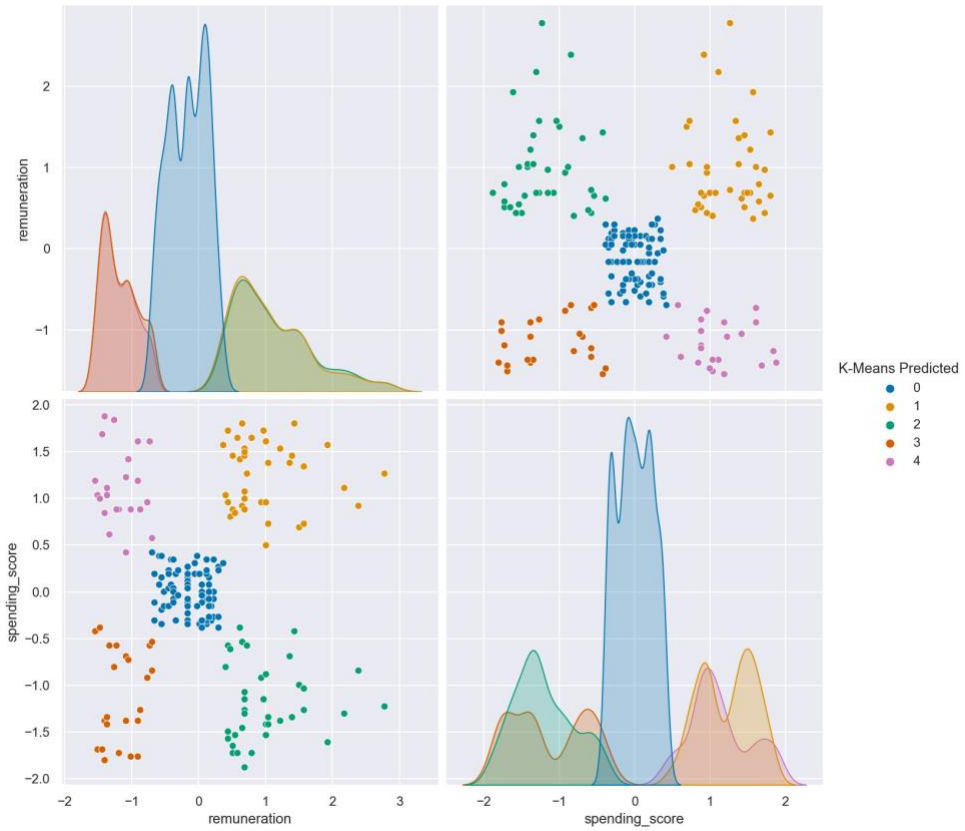
2.1.4. Segmenting customer base by groups.

- The initial DataFrame was modified by selecting relevant columns (remuneration and spending score).
- Colour-blind pallet/colours were used to create visualisations included in the report.
- Elbow and silhouette methods were used to identify the best number of clusters for applying k-means clustering method for segmenting the data into groups.


Elbow Method
WSS
8000
7000
6000
5000
4000
3000
2000
1000
0
2
4
6
8
10
Number of clusters


Silhouette Method
Sil
0.55
0.50
0.45
0.40
2
3
4
5
6
7
8
9
10
Number of clusters

Clustering models with k = 4, 5, 6 and 7 were investigated. The model with 5 clusters have been chosen as the best one with five clear groups of customers.



- **Clusters characteristics:**

a) Cluster sizes:

- Cluster 0 - 744 customers;
- Cluster 1 - 356 customers;
- Cluster 2 - 351 customers;
- Cluster 3 - 280 customers;
- Cluster 4 - 269 customers.

b) Age group:

Similar for all clusters.

c) Remuneration:

- Cluster 0 - between £31.98k and £56.58k;
- Cluster 1 and 2 are similar - between £56.58k(cluster 1)/ £57.40k(cluster 2) and £112.34k;
- Cluster 3 and 4 - between £12.30k and £31.98k.

d) Spending score:

- Cluster 0 - between 40 and 61;
- Cluster 1 and 4 are similar - 63(cluster 1)/ between 61(cluster 4) and 97(cluster 1)/ 99(cluster 4);
- Cluster 2 and 3 are similar - between 1(cluster 2)/ 3(cluster 3) and 40.

e) Loyalty points:

- Cluster 0 - between 603 and 2332;
- Cluster 1 - between 2289 and 6847;

- Cluster 2 - between 40 and 2325;
- Cluster 3 - between 25 and 854;
- Cluster 4 - between 436 and 1851.

Customers with the lowest number of loyalty points have been grouped in cluster 3; customers with the highest number of loyalty points have been grouped in cluster 1.

f) Gender ratio:
Clusters 2 is the only cluster with higher number of male than female customers.

g) Education level ratio:
- The lowest %% of customers with graduate level of education is in cluster 2.
- The highest %% of customers with PhD level of education is in cluster 2.
- The lowest %% and absolute number of customers with the basic level of education is in cluster 4.

- **Summary:**
Cluster 0 is the largest cluster identified. It contains 37.20% of all customers included in the data set. Customers included in this cluster have spendings from the middle remuneration bracket identified (£31.98k-£56.58k) and the spending scores from the middle spending score bracket identified (40-61).

Cluster 1 is the second largest cluster identified. It contains 17.80% of all customers. Customers included in this cluster have spendings from the high remuneration bracket identified (£56.58k-£112.34k) and the spending scores from the high spending score bracket identified (63-97). Customers in this cluster also have the highest number of loyalty points (2289-6847) in the data set.

Cluster 2 is similar in size to cluster 1 and contains 17.55% of all customers. Customers included in this cluster, similarly to customers from cluster 1, have spendings from the high remuneration bracket identified (£57.40k-£112.34k) but, unlike customers from cluster 1, have the spending scores from the low spending score bracket identified (1-40). This is the only identified cluster with the higher number of male than female customers.

Cluster 3 contains 14.00% of all customers. Customers included in this cluster have spendings from the low remuneration bracket identified (£12.30k-£31.98k) and, similarly to customers from cluster 2, have the spending scores from the low spending score bracket identified (3-40). Customers in this cluster also have the lowest number of loyalty points (25-854) in the data set.

Cluster 4 is similar in size to cluster 3 and overall the smallest cluster identified. It contains 13.45% of all customers. Customers included in this cluster, similarly to customers from cluster 3, have spendings from the low remuneration bracket identified (£12.30k-£31.98k) but, unlike customers from cluster 3, have the spending scores from the high spending score bracket identified (61-99). This cluster contains the lowest percentage and absolute number of customers with the basic level of education.

2.1.5. Exploration of social media data (customer reviews and summary of customer reviews).

- Social data was provided in two columns: online reviews submitted by customers who purchased and used the products and summary for each customer's review.
- The data was pre-processed by changing it to the lower case, removing punctuation and stop words.
- 39 duplicates (1.95% of all observations) were identified in review and summary columns. Not enough evidence were observed to conclude that identified observations are indeed duplicates when considering data in other columns of the data set.

- NLP sentiment analysis were conducted on data with and without duplicates; no significant changes in the outputs were observed. Below will be present insights and observations from the original data set (including duplicates in review and summary columns).
- Tokenisation was applied to create word clouds.
  Word cloud for reviews column:

- 15 most frequently used words in reviews with polarity scores:

	Word	Frequency	Polarity
1	game	1685	-0.4
2	great	596	0.8
3	fun	553	0.3
4	one	530	0.0
5	play	502	0.0
6	like	414	0.0
7	love	331	0.5
8	really	319	0.2
9	get	319	0.0
10	cards	301	0.0
11	tiles	297	0.0
12	good	294	0.7
13	time	291	0.0
14	would	280	0.0
15	book	273	0.0

15 most frequently used words in summaries of reviews with polarity scores:

	Word	Frequency	Polarity
1	stars	466	0.0
2	five	381	0.0
3	game	319	-0.4
4	great	295	0.8
5	fun	218	0.3
6	love	93	0.5
7	good	92	0.7
8	four	58	0.0
9	like	54	0.0
10	expansion	52	0.0
11	kids	50	0.0
12	cute	45	0.5
13	book	43	0.0
14	one	38	0.0
15	awesome	36	1.0

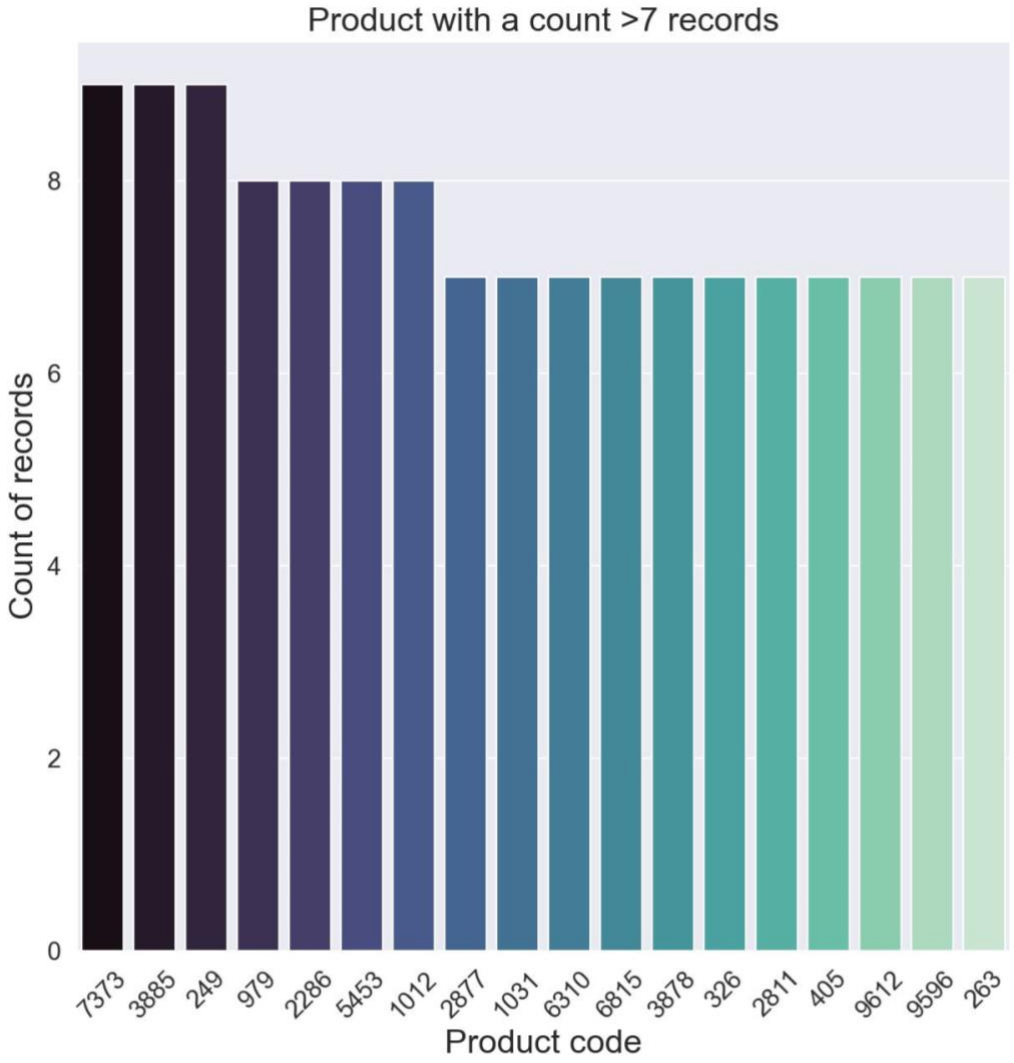
- Textblob pre-trained inbuilt Python classifier and VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon and rule-based sentiment analysis tool were used for the sentiment analysis of review and summary columns. Outputs of both analysis were overall positive for both columns (Textblob (mean): 0.217688 for reviews; 0.226667 for summaries. VADER (mean): 0.614441 for reviews; 0.378276 for summaries).
- Textblob output:
  - 225 observations (11.25%) have inconsistency between polarity score for review and summary columns (positive polarity score for review column and negative polarity score for summary column and vice versa).
  - 72 observations (3.6%) have negative polarity score for review and summary columns and were marked those as 'true negative'.
  - 930 observations (46.5%) have positive polarity score for review and summary columns and were marked those as 'true positive'.

- Clusters identified in p.2.1.4 were added to 'true negative' and 'true positive' data sets:
  - 44.44% of all customers in 'true negative' data set belong to cluster 0, followed by cluster 4 and cluster 1 (16.67% each) => customers have spending score from middle-high bracket (40-99). Equal amount of male and female customers are present in this data set.
  - 37.53% of all customers in 'true positive' data set belong to cluster 0, followed by cluster 1 (17.74%) and cluster 2 (17.31% each) => customers have spendings from middle-high remuneration bracket (£31.98k-£112.34k). 56.99% of female customers is present in this data set.

- Product:
  - 61 products (30.5% of all products in the original data set) were mentioned in 'true negative' data set. 11 products were mentioned the most (twice):

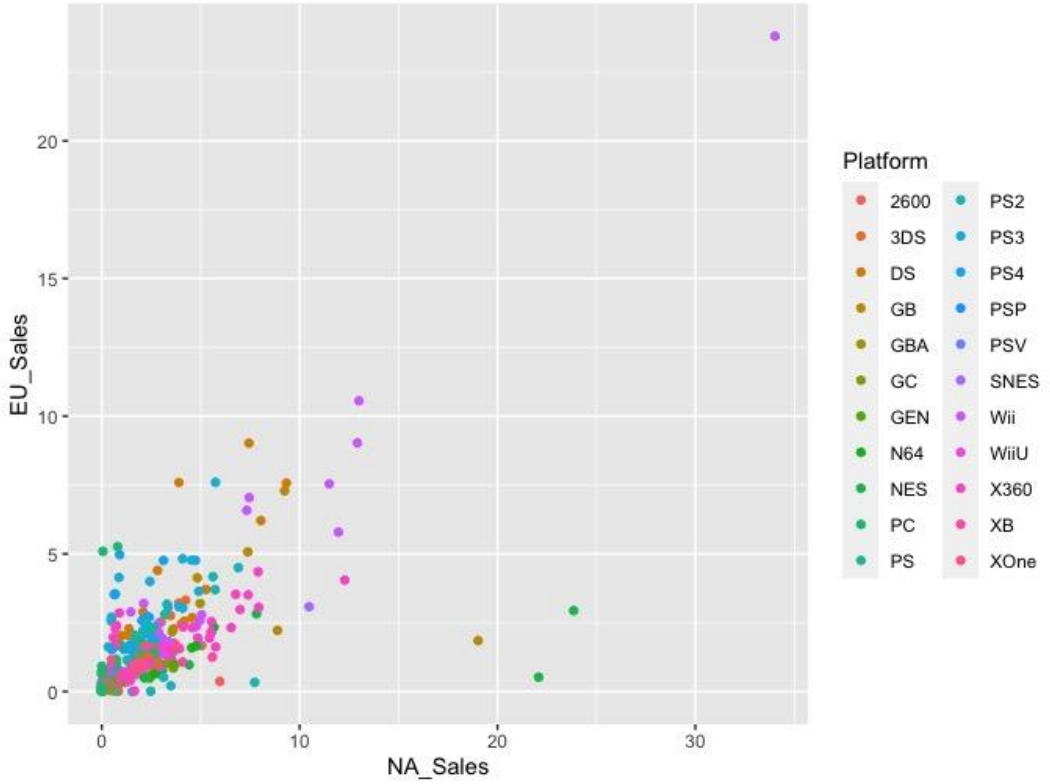
	Product code	Count
1	486	2
2	518	2
3	760	2
4	876	2
5	2285	2
6	2870	2
7	3436	2
8	3524	2
9	3967	2
10	6233	2
11	9597	2

- 200 products (100% of all products in the original data set) 'true positive' data set. 18 products were mentioned the most (7-9 times):

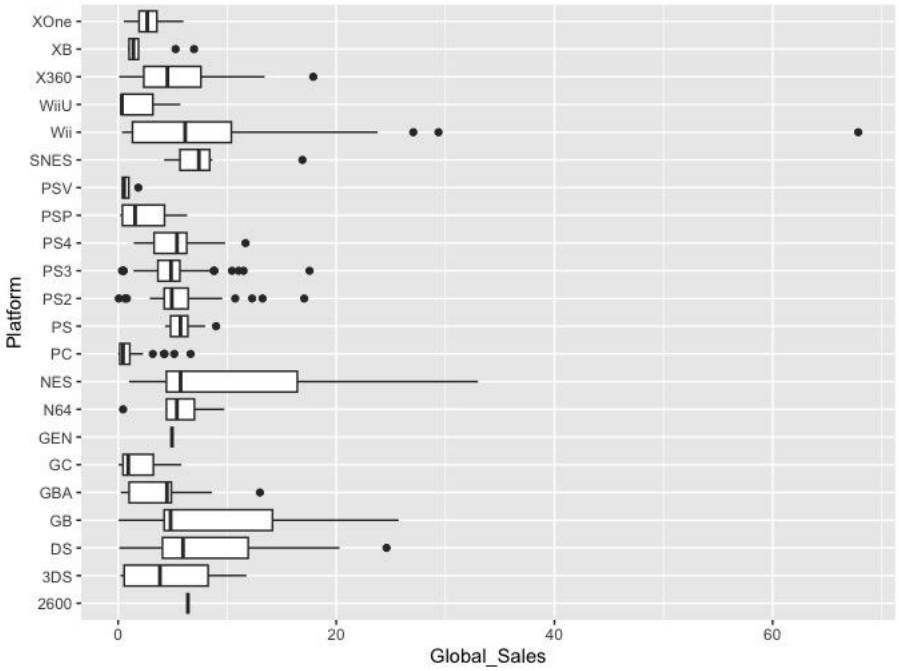


2.1.6. Exploration of an impact of each product on sales.

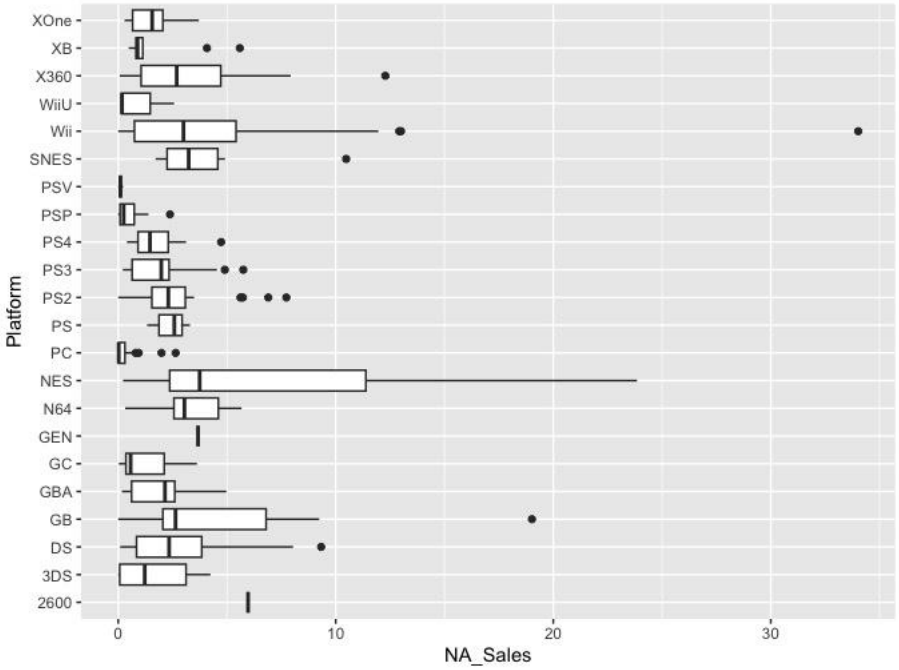
- The workstation was prepared by importing the necessary libraries and turtle_sales.csv data file in a new R file.
- The file was converted to a data frame and sense-checked.
- The data set contains 352 observations of Turtle Games sales by product/platform/year/etc. in North America (NA), Europe (EU) and globally (sum of NA, EU and other).
- No missing values were identified.
- 17 observations have no sales recorded for NA and 3 observations have no sales recorded for EU.
- The below scatterplot shows some positive moderate relationships between sales in NA and in EU with four obvious outliers: one is at the upper right corner and three are at the bottom centre of the plot. The scatterplot has a cone shape meaning that the data might be heteroscedastic. We can see some sales for NA when EU sales=0 and vice versa; as well as the data point when both NA and EU = 0 (possibly, it relates to 'other' sales).



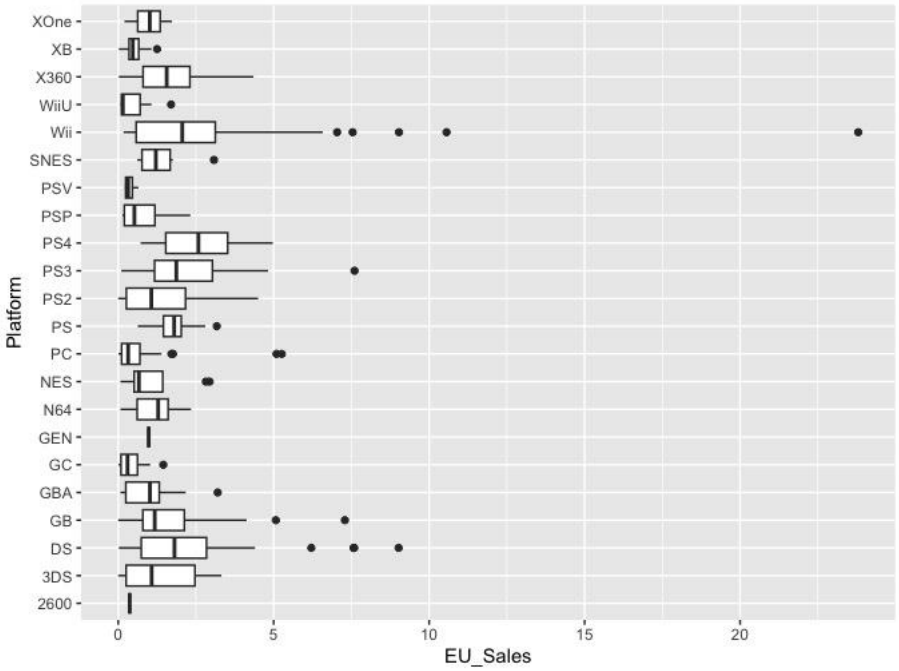
- Average EU sales are lower than average NA sales.
- Each set of boxplots (below) by platform (Global, NA and EU) includes a variation of symmetric, right-skewed, and left-skewed unimodal data sets.
  For Global sales per platform: SNES has the highest average sales; NES has the largest range and IQR (overall spread between data and the middle 50% of the data); 59% of platforms have outliers; GEN and 2600 platforms need to be investigated further as might have one recorded sale and possibly are outliers.



For NA sales per platform: NES has the highest average sales, the largest range and IQR; 50% of platforms have outliers.



For EU sales per platform: PS4 has the highest average sales; Wii has the largest range and IQR; 55% of platforms have outliers.

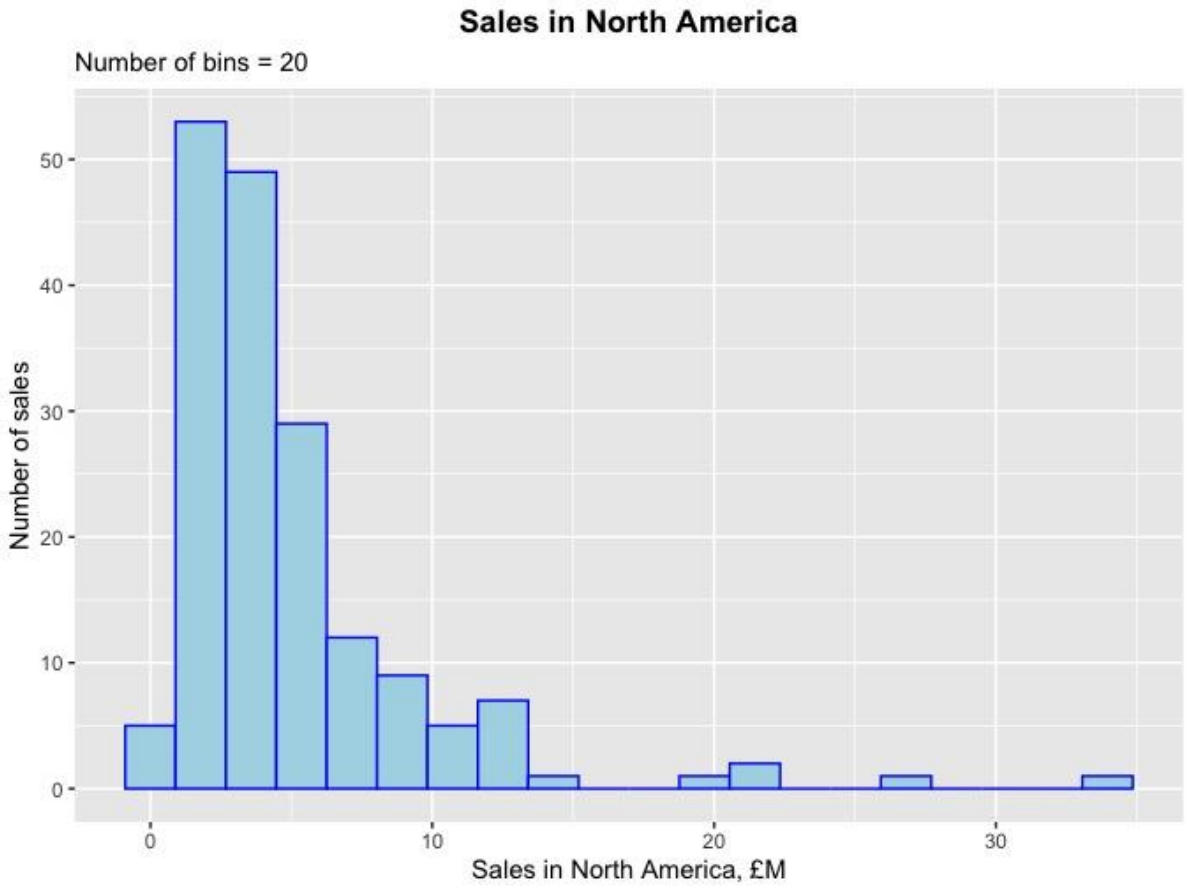


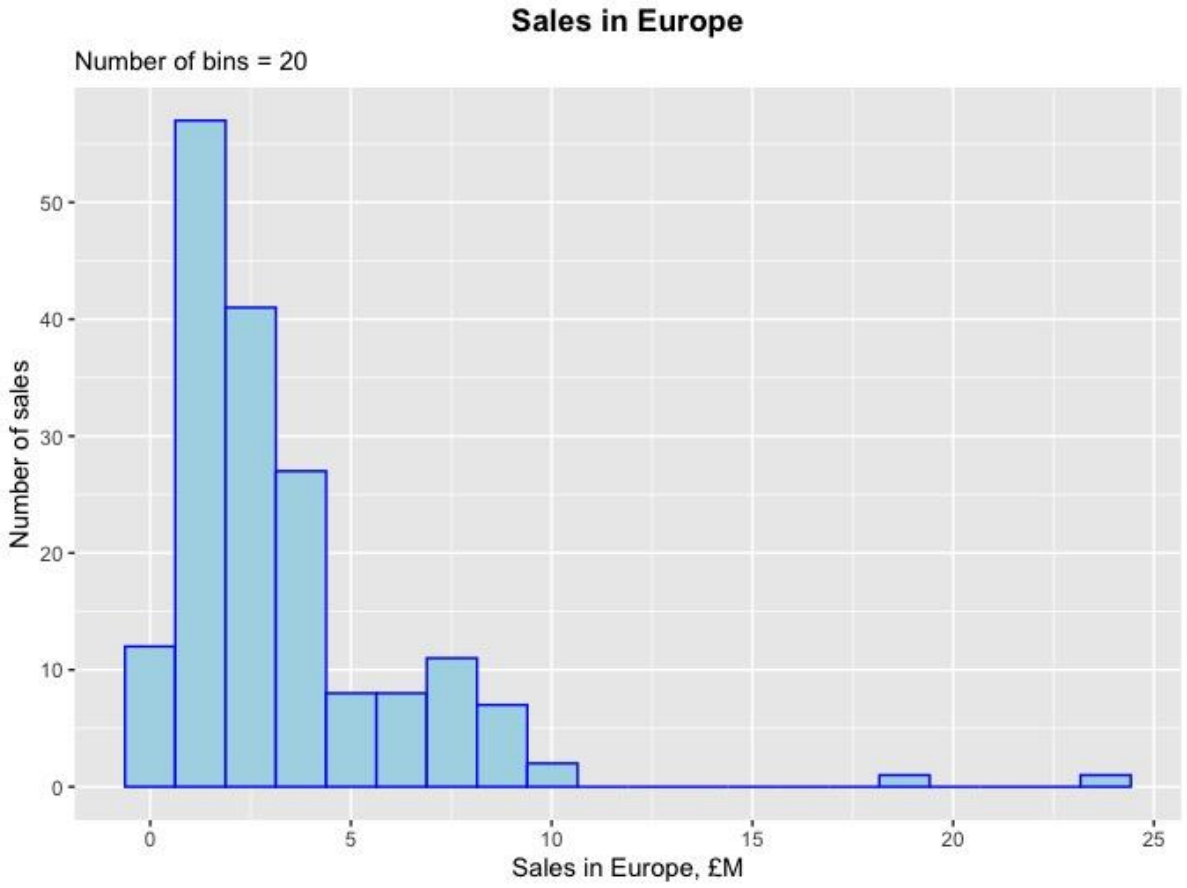
2.1.7. Exploration of how reliable the data is (e.g. normal distribution, skewness, or kurtosis).

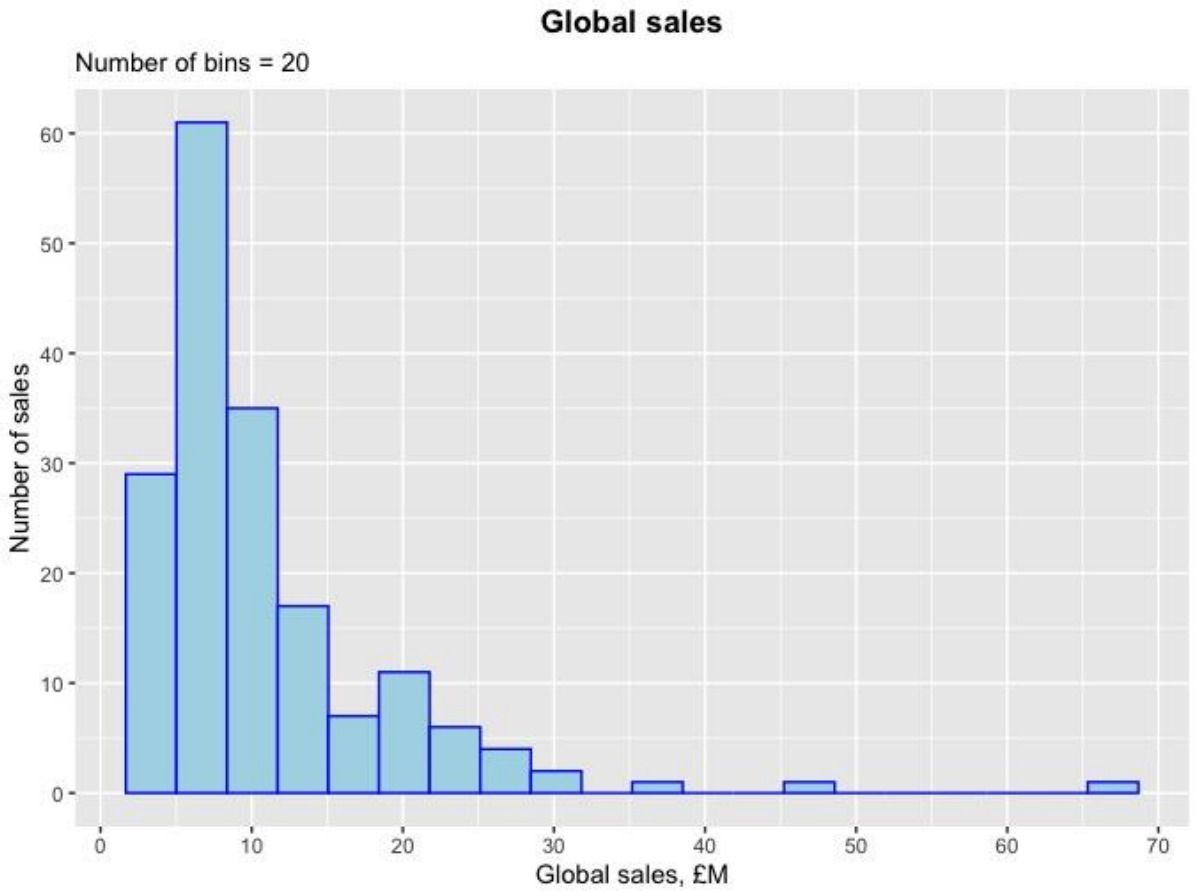
- For further exploratory data analysis (EDA) the new data frame was created by keeping product and sales columns from the original data frame.
- 41.14% of all unique products were mentioned more than twice in the data set. The decision was made to aggregate data by unique product; as a result, the number of observations was reduced by 49.72% (from 352 to 175 observations). It also affected values of descriptive statistics for sales columns; in particular mean and median of all sales columns have increased.

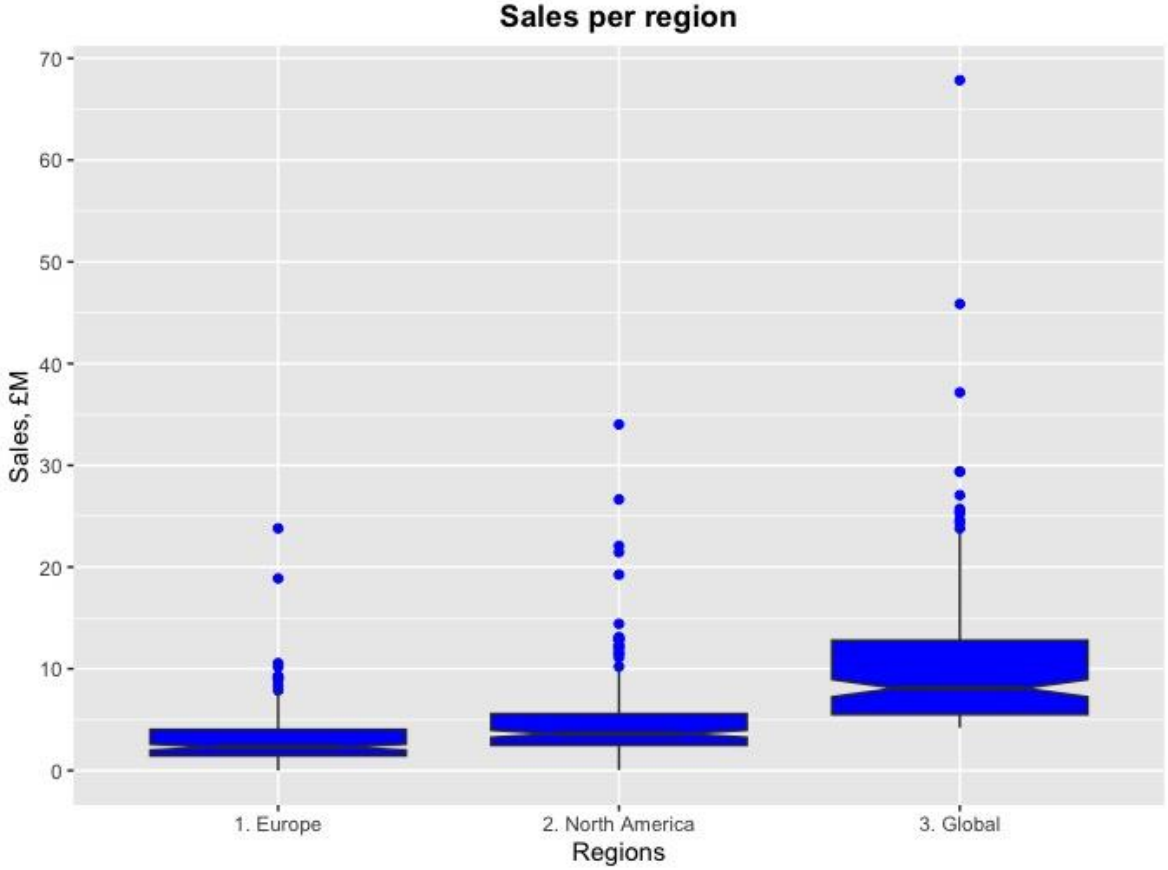
NA_Sales, £M	EU_Sales, £M	Global_Sales, £M
Min. : 0.060	Min. : 0.000	Min. : 4.200
1st Qu.: 2.495	1st Qu.: 1.460	1st Qu.: 5.515
Median : 3.610	Median : 2.300	Median : 8.090
Mean : 5.061	Mean : 3.306	Mean :10.730
3rd Qu.: 5.570	3rd Qu.: 4.025	3rd Qu.:12.785
Max. :34.020	Max. :23.800	Max. :67.850

- The sales columns in the new aggregated data frame were investigated for the normalisation. All three variables (sum_NA_Sales, sum_EU_Sales and sum_Global_Sales) have highly skewed (positive) leptokurtic distribution.


Sales in North America
Number of bins = 20
Number of sales
0
10
20
30
40
50
0
10
20
30
Sales in North America, £M


Sales in Europe
Number of bins = 20
Number of sales
0
10
20
30
40
50
0
5
10
15
20
25
Sales in Europe, £M

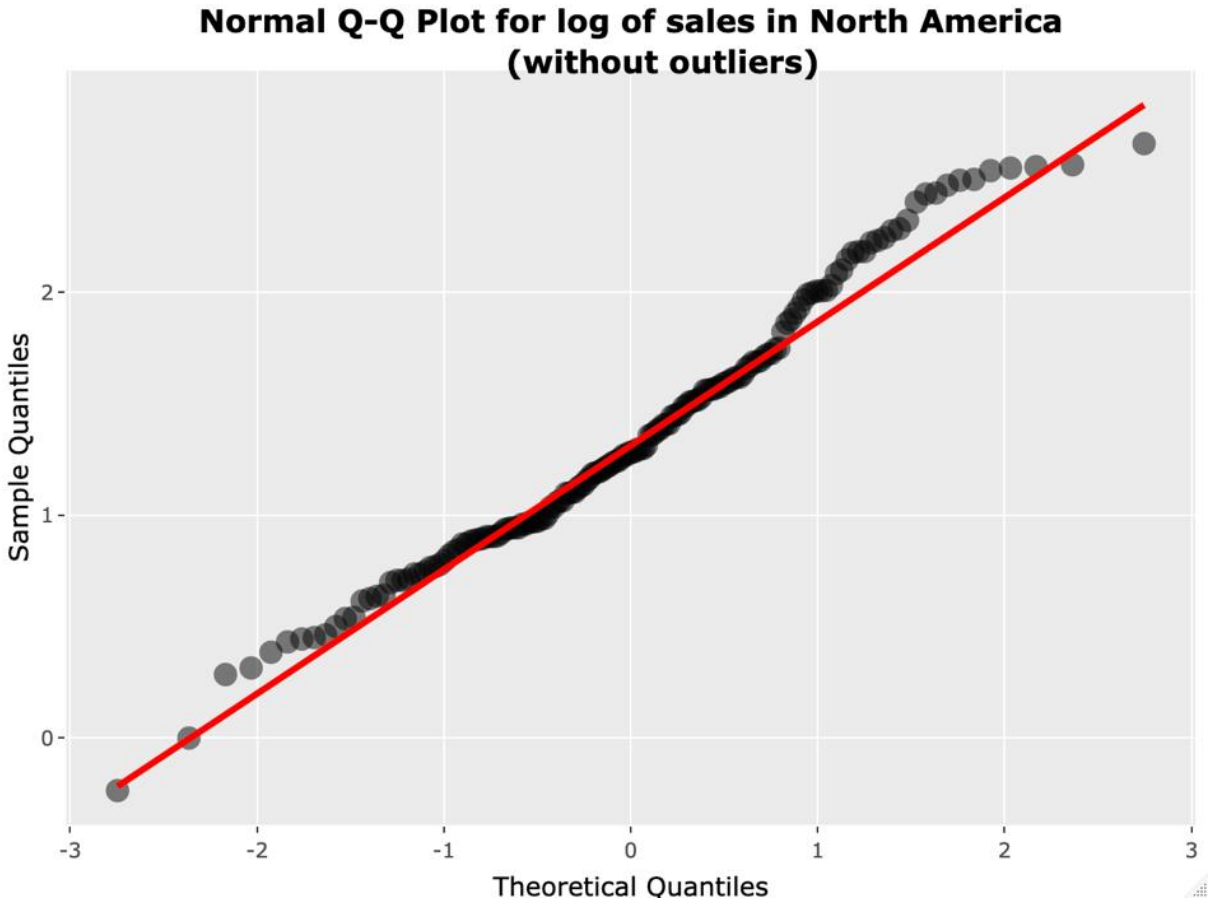

Global sales
Number of bins = 20
Number of sales
0
10
20
30
40
50
60
0
10
20
30
40
50
60
70
Global sales, £M


Sales per region
Sales, £M
0
10
20
30
40
50
60
70
1. Europe
2. North America
3. Global
Regions

- Removing outliers did not help to normalise data. Three additional methods - square root, log and inverse of data, were employed in order to normalise data.
- Log method in combination with removing outliers gave the best result for sales in NA and EU:

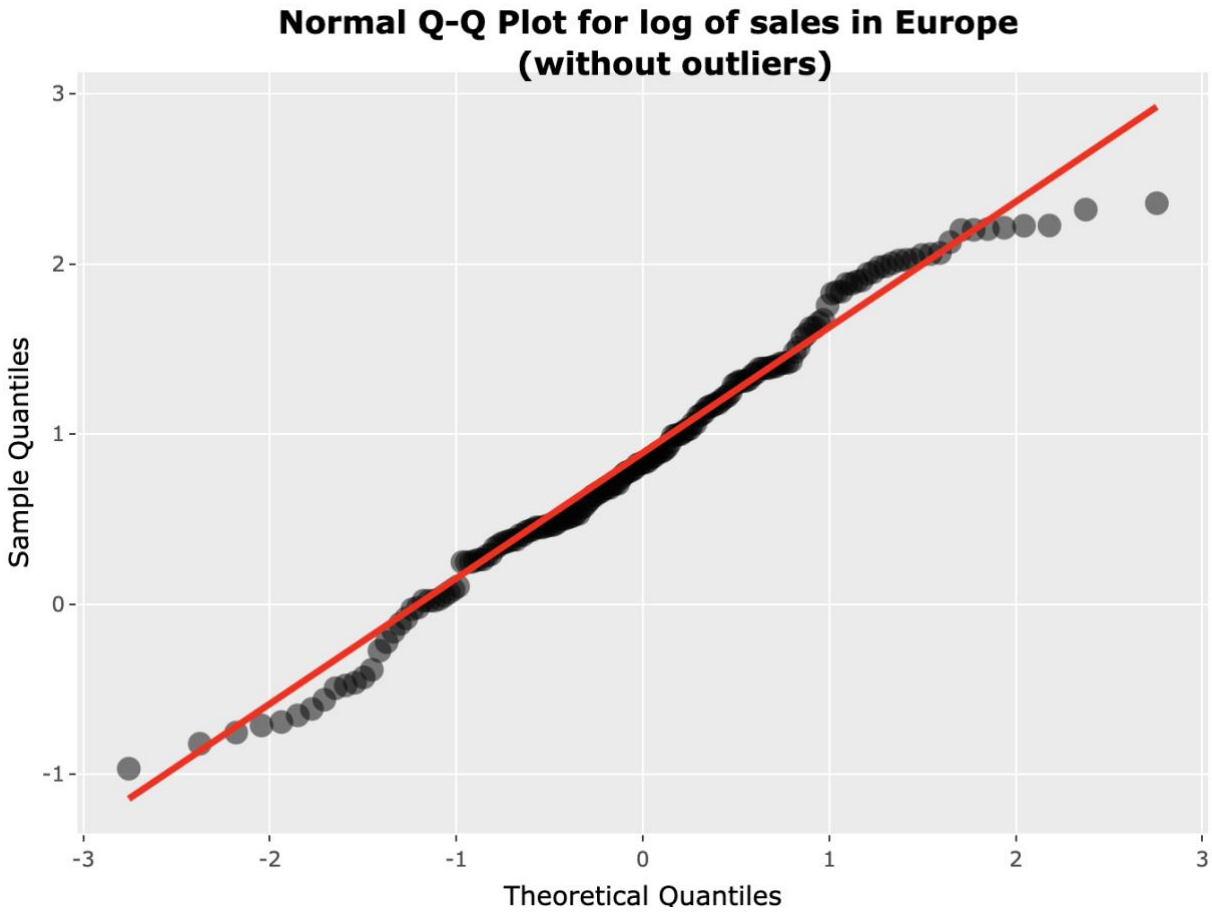
a) log of sales in North America (without outliers):

- Shapiro-Wilk normality test: p-value = 0.02838;
- skewness = 0.2562564 is between -0.5 and 0.5 => the data is fairly symmetrical;
- kurtosis = 2.678189 is close to 3 => mesokurtic distribution (medium-tailed, so outliers are neither highly frequent, nor highly infrequent).



b) log of sales in Europe (without outliers):

- Shapiro-Wilk normality test: p-value = 0.03709;
- skewness = -0.08772597 is between -0.5 and 0.5 => the data is fairly symmetrical;
- kurtosis = 2.484654 is close to 3 => mesokurtic distribution (medium-tailed, so outliers are neither highly frequent, nor highly infrequent).



c) Global sales data was not normalised.

- Correlation matrix shows the relationship between variables:

  - Global and NA sales: 0.92 => strong positive correlation;
  - Global and EU sales: 0.85 => strong positive correlation;
  - NA and EU sales: 0.62 => moderate positive correlation;
  - log NA and log EU sales: = 0.44 => low positive correlation.

	sum_NA_Sales	sum_EU_Sales	sum_ Global_Sales	log_NA_Sales	log_EU_Sales	log_Global_Sales
sum_NA_Sales	1.0000000	0.6209317	0.9162292	0.8214719	0.4433215	0.8204364
sum_EU_Sales	0.6209317	1.0000000	0.8486148	0.4814316	0.8029432	0.7880406
sum_Global_Sales	**0.9162292**	**0.8486148**	1.0000000	0.7329525	0.6263640	0.9177421
log_NA_Sales	0.8214719	0.4814316	**0.7329525**	1.0000000	0.4420464	0.7858555
log_EU_Sales	0.4433215	0.8029432	0.6263640	0.4420464	1.0000000	0.7040250
log_Global_Sales	**0.8204364**	**0.7880406**	0.9177421	0.7858555	0.7040250	1.0000000

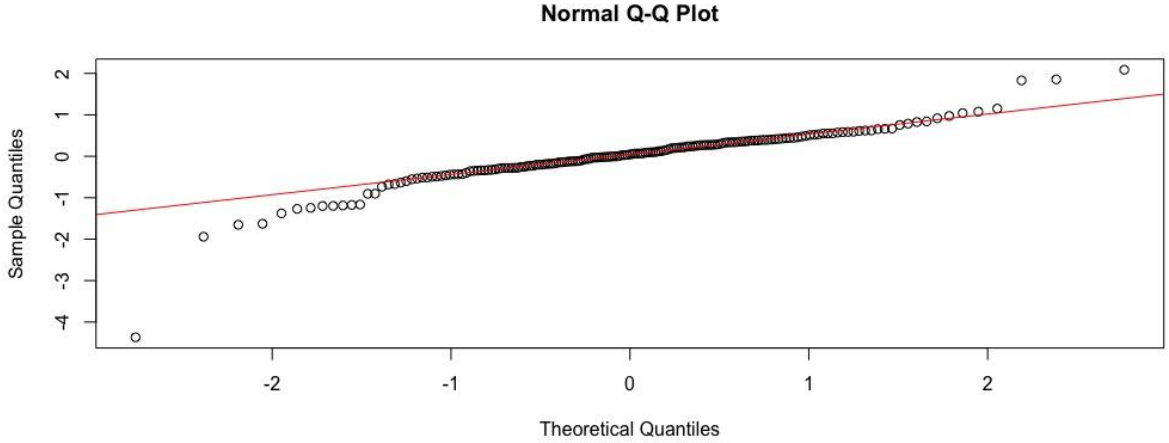
2.1.8. Exploration of relationship(s) between North American, European, and global sales.

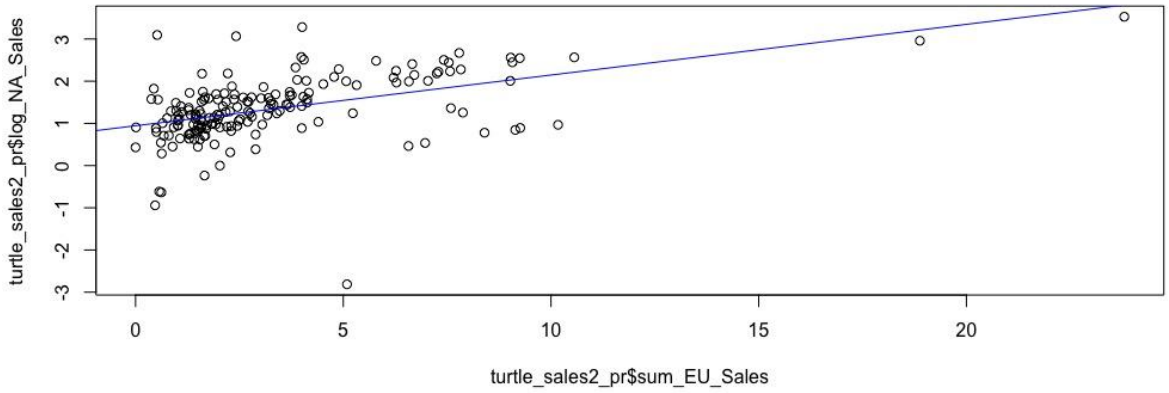
- Eight simple linear regression models were created in order to determine the relationships between sales in North America and Europe. 5 of which had signs of heteroscedasticity. Various techniques were applied to fix it, including removing outliers, applying log and square root transformations to dependent variable.

Two best models are below:

a) Model2 (y = log_NA_Sales and x = sum_EU_Sales) was identified as the best out of four simple linear regression models for predicting sales in North America:

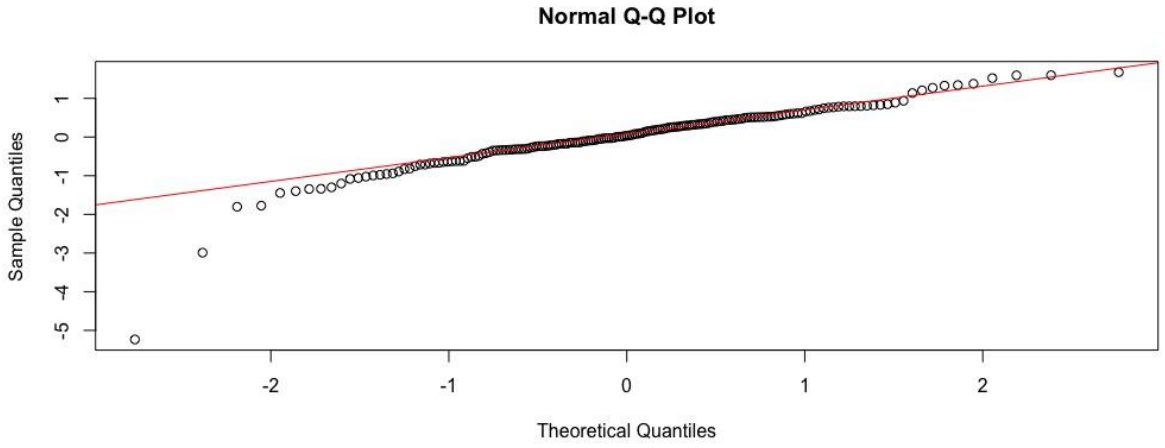
- p value = 1.54e-11;
- R^2 value = 0.2318;
- Residual standard error: 0.6771 on 173 degrees of freedom;
- BP test p-value = 0.5233 > 0.05 => heteroscedasticity is not present.

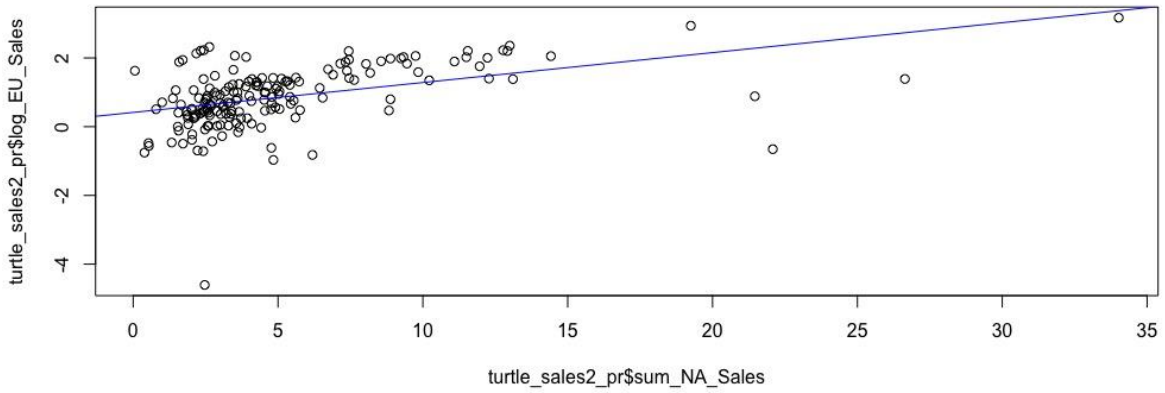




b) Model2_1 (y = log_EU_Sales and x = sum_NA_Sales) is the best out of four created simple linear regression models for predicting sales in Europe:

- p value = 8.05e-10;
- R^2 value = 0.1965;
- Residual standard error: 0.8044 on 173 degrees of freedom;
- BP test p-value = 0.4609 > 0.05 => heteroscedasticity is not present.





Comparing model2 and model2_1, we can conclude that model2 is stronger and we can better predict NA sales by using EU sales than vice versa.

- Eight simple linear regression and six multiple linear regression models were created to predict Global sales. In four simple linear regression and three multiple linear regression models heteroscedasticity was detected. Various techniques were applied to fix it, including removing outliers, applying log and square root transformations to dependent variable.
  After the first round of evaluations two best simple linear regression and three multiple linear regression models have been chosen to be tested by the test data set. One simple linear regression and two multiple linear regression models produced the best output and were compared by RMSE and R2 values.
  Model_ml (y = sum_Global_Sales and X = sum_NA_Sales, sum_EU_Sales) was identified as the strongest with error rate 9.11% and R^2 = 0.998 (observed and the predicted outcome values are highly correlated):

  - both p values <2e-16;
  - R^2 value = 0.9668;
  - Adj R^2 value = 0.9664;
  - Residual standard error: 1.49 on 172 degrees of freedom;
  - VIF = 1.63 => between 1 and 5 => variables are moderately correlated (but < 3 => is not a cause for concern);
  - BP test p-value = 0.1718 > 0.05 => heteroscedasticity is not present.
  - The actual data is compared below to the output of the model:

	Global_Sales FACT	Global_Sales MODEL	lw range MODEL	up range MODEL
1	**67.85**	**68.056548**	66.429787	69.683310
2	**6.04**	**7.356754**	7.099418	7.614090
3	**4.32**	**4.908353**	4.614521	5.202185
4	**3.53**	**4.761039**	4.478855	5.043223
5	**23.21**	**26.625558**	25.367353	27.883763

3. Recommendations and questions for further investigation.

- Positive linear trends were observed between loyalty points and spending score and between loyalty points and remuneration. For customers with spending scores <=60, increase in spending score by 1 unit will increase loyalty points by 25.52 units. For customers with spending scores >60, increase in spending score by 1 unit will increase loyalty points by 1.02 units. For customers in total increase in remuneration by £1,000 will increase loyalty points by 1.024 units (= 1.02).
  Therefore, it is recommended to focus on customers with spending scores <=60, as each increase in spending score will result in a proportionately larger increase in loyalty points for those customers.
- Using remuneration and spending scores as main factors, five groups/clusters of customers were identified in Turtle Games data set.
  The largest group contains 37.20% of all customers included in the data set (cluster 0). Customers included in this cluster have spendings from the middle remuneration bracket and the spending scores from the middle spending score bracket.
  The next two clusters (cluster 1 and cluster 2) are similar in size (contain 17.80% and 17.55% of all customers respectively). Both have customers with spendings from the high remuneration bracket but first cluster has customers with the highest number of spending scores and the second with the lowest.
  The last two clusters (cluster 3 and cluster 4) are also similar in size (contain 14.00% and 13.45% of all customers respectively). Both have customers with spendings from the low remuneration bracket but first cluster has customers with the lowest number of spending scores and the second with the highest.
  Customers in cluster 1 (high remuneration – high spending score) have the highest number of loyalty points in the data set.
  Customers in cluster 3 (low remuneration – low spending score) have the lowest number of loyalty points in the data set.
  Taking into the consideration insights highlighted in the first point, encouraging customers from cluster 3 particularly to increase spending scores, will improve their loyalty points also. We also can predict that the number of loyalty points will increase faster for customers in cluster 2 (high remuneration – low spending score).
- Overall customers sentiment towards Turtle Games product is positive. Applied NLP models in 11.25% observations highlighted difference in polarity signs between the reviews and summary of the reviews when first ones had a positive sentiment and the second ones – negative, and vice versa. For the future sentiment analysis it might be beneficial to investigate this discrepancies further.
  30.5% of all products from the original data set were mentioned in the 'true negative' data set (review and summary of the review have negative polarity score). 11 products were mentioned the most (twice). Saying that the 'true negative' data set contains only 3.6% of all observations, which can be seen as a positive trend. Majority customers in this data set have spending score from middle-high bracket. Equal number of male and female customers is present in this data set.
  All products from the original data set were mentioned in the 'true positive' data set (review and summary of the review have positive polarity score). 18 products were mentioned the most (7-9 times). The 'true positive' data set contains 46.5% of all observations. Majority customers in this

data set have spendings from middle-high remuneration bracket. 56.99% of female customers is present in this data set.

- Building linear model regressions were challenging due to the highly skewed (positive) leptokurtic distribution of sales data. Nevertheless, two best simple linear regression models were selected to predict North American and European sales and one multiple linear regression model was selected to predict Global sales. Overall positive linear trend was observed between NA, EU and Global sales, meaning that if one variable will be increasing, it is a high possibility that other variables will be increasing also.