Turtle Games Project – Technical Report

Intro

Data Analysis project contracted by Turtle Games, a game manufacturer and retailer with a global customer base. The company manufactures and sells its own products, along with sourcing and selling products manufactured by other companies. Its product range includes books, board games, video games, and toys. The company collects data from sales as well as customer reviews. Turtle Games has a business objective of improving overall sales performance by utilizing customer trends.
To improve overall sales performance, Turtle Games has come up with an initial set of questions. You'll explore these questions in greater depth through the weekly assignment activities. Turtle Games wants to understand:

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

Make predictions with regression.

We commence analysis by the dataset obtained via csv ("turtle_reviews) for exploratory analysis via Python.
Imported packages for data wrangling (numpy, pandas), visualization (matplotlib, seaborn), statistics (statsmodels, sklearn), natural language analysis (nltk).
Dataframe for this part of the analysis ("turtle_reviews.csv") contains mainly demographic data along with 2 interesting columns, ie customers reviews and summaries of products purchased. The data was checked for missing values (none found), duplicates (none found) and then removed columns being repeated (language, platform) and renamed columns (renumeration, spending_score) for easier handling.
We performed some extra data wrangling and found:

- `Female 56.00% Male 44.00% Name: gender, dtype: float64`

Insight: *Almost balanced gender diversity with a slight edge to Female customers and if this edge (or sample) is representative of actual situation it can give a direction to stakeholders as it is quite an advantage to attract female clientele on a stereotypically "male endorsed" market*

- Women might earn yearly 2k less than(from mean renumeration index) but spend almost same as men

gender	renumeration	spending_score
Female	47.3	50.7
Male	49.1	49.1

Insight: *Once more, women's strong spending_score should showcase demographic opportunity to be further exploited and used to increase profit by engaging more female clientele*

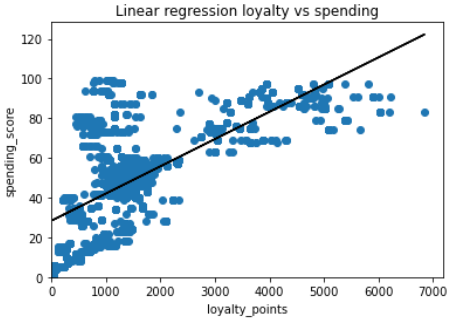
- We grouped by education field and aggregated basis renumeration to get "the educational profile of our sample" where almost 88% of our clientele hold a university degree

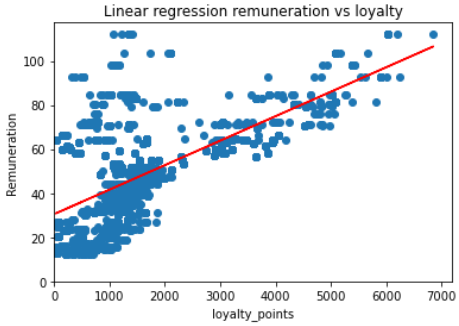
education	renumeration
graduate	44.272850
PhD	24.504972
postgraduate	19.512050

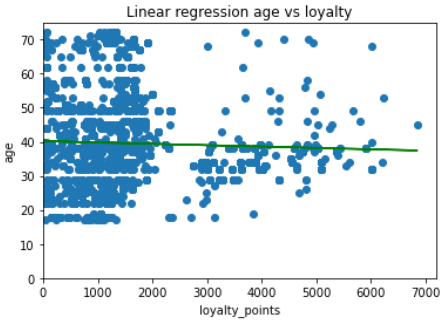
After data wrangling and cleaning our new DataFrame was saved and exported in new CSV for further analysis (turtle_reviews_clean.csv)

We continued by producing by running a correlation matrix on the new df to witness that:

a) Renumeration has a reasonable correlation to loyalty points (0.62)
b) Spending score has a slightly bigger correlation to loyalty points (0.67)
c) Renumeration has a very small correlation to spending score (**insight:** *good news for marketing, as not only big wallets increase sales)*
d) **On the other hand, we should remember, correlation does not imply causation**. Since we work with a sample of data, if we obtain a different sample, it's possible, we could have different correlation scores. As such we need to assess the significance of the correlation values we calculated, which depends on the sample size.







Case1 - Spending vs loyalty points

R^2 45% of the total variability of `y` (spending score), is explained by the variability of `X` (how many purchases done).

F-stat:is very high but p-value is very low (2.92e -263) as such greater the statistical significance.

**X`**: The coefficient of `X` tells us, if the length that the customer has been a member (`X`) changes by 1 unit (please check units used) the money spent (`y`) will change by 64.2187 units (spending score assigned).

The **`t`-value** being at 41.5 shows that slope is not significant.

Case2 - Renumeration vs loyalty points

R^2 38% of the total variability of `y` (total income of each customer per year), is explained by the variability of `X` (how many purchases done).

F-stat :is very high but p-value is very low as such greater the statistical significance.

X: The coefficient of `X` shows if the length that the customer has been a member (`X`) changes by 1 unit (ie loyalty point) the money he should be earning (`y`) will change by 30.56 units (k =1000 GBP).

The **`t`-value** being at 47 shows that slope is not significant.

Case 3 - Age vs loyalty points

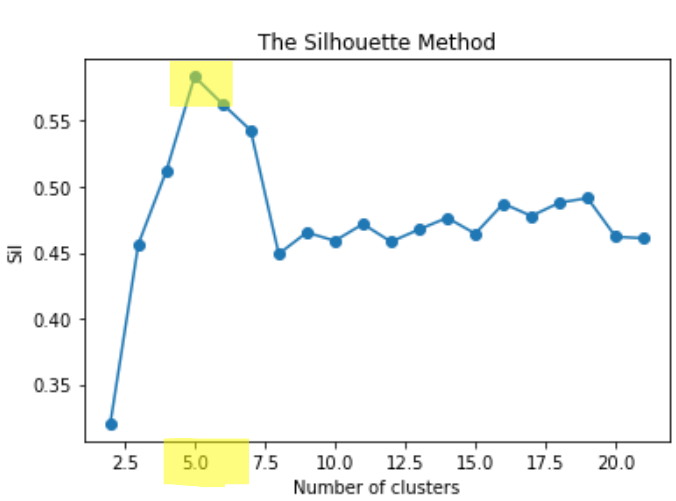
R^2: 0.2% of the total variability of `y` (age), is explained by the variability of `X` (how many purchased they have done).

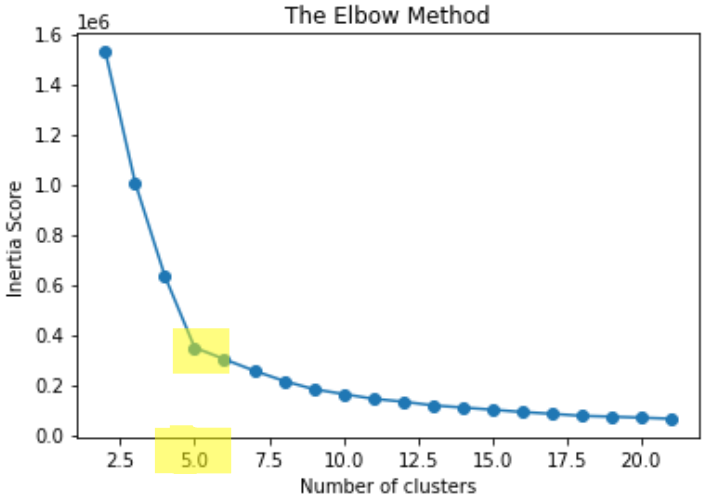
No need to investigate further as it looks that suggested model is not successful

At this point we can provide certain feedback to the question:" ***how customers accumulate loyalty points***" . There is a strong correlation between spending_score and loyalty points as well between spending score and renumeration. Which means that certain clients with higher yearly earnings produce bigger spending_score and therefore increase their loyalty_points.
While the latter might be more evident to spot or even assume, in reference to the loyalty points our metadata file is not clearly communicating whether there is any confirmed loyalty programs or the loyalty points are just an "insiders" metric.
Insight: *If there is not a confirmed loyalty scheme for clientele of Turtle Games (involving possession of members cards with discounts, exclusive advantages etc), marketing should act on this initiative as it will most certainly increase the spending score which will increase the profits (provided that either clients increase or purchases are of high value)*

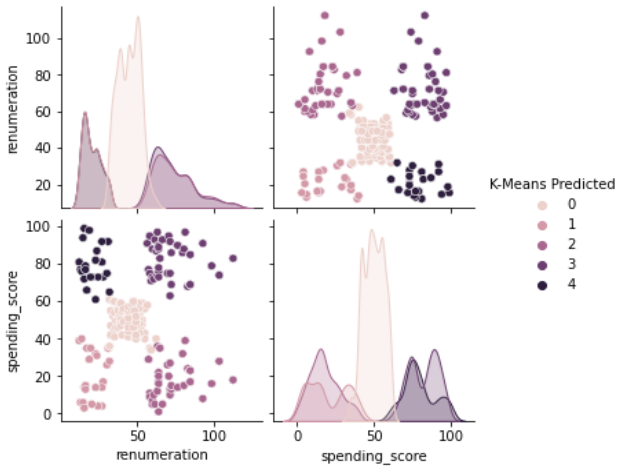
Make predictions with clustering.

We continue our exploratory analysis with the clean dataframe from previous steps. We further remove columns and focus on **remuneration** and **spending_score**. We run descriptive statistics to get a feel of the values in our reduced dataframe and conclude that no scaling is required for this case as comparable variables are aligned in terms of values and both lie on the same scale without over-showing the other.
We run the elbow method (k-means clustering on the dataset for a range of values) and the silhouette method (silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters), both statistical and we evaluate the produced values.

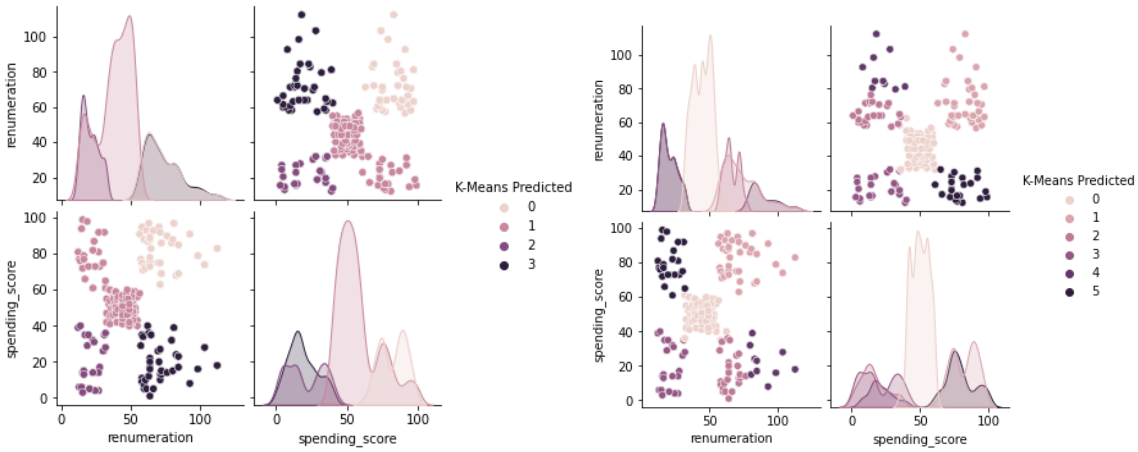




We kept running evaluations basis KMEANS function with different k, starting from 5 and reducing but also increasing as well. We concluded with final model to be using five clusters as it gives the more balanced predicted results and also distinctly useful groups in the renumeration diagrams.
The below diagram showing different correlations of our cluster groups can visually designate that 5 is the correct separation to be made for our model



Compare the above with a group of 4 and a group of 6 clusters.



We can easily identify that over 5 some clusters overlap each other and with less than 5 we don't fully differentiate all the groups.

Our final cluster model as follows:

In order to give some business context in the analysis above, and reply to the question: *"**how groups within the customer base can be used to target specific market segments**"* we can translate our 5 clusters in following profiles:

1. **The low key - logicals** - Starting from the bottom left, cluster group 1 (GREEN) represents our customers with low yearly renumeration that spend according to their budget and can stress until limits of following category.
2. **Our average Joe - middle class** - Middle class, cluster group 0 (RED) not necessarily justifies her name due to remuneration classification in our plot but view that a big chunk of our datapoints are centred between an average of 40-60 spending score as well as renumeration
3. **The high earners - savers** - Representing cluster group 2 (BLUE), where we can see individuals with mid and high earnings but keeping same spending score as individuals earning 1/3 of their yearly income
4. **The high earners - Big Spenders** - Cluster group 3 (BLACK) could represent our high profile VIP customers base which seems to earn a considerable high year income and spend it accordingly or (relationally) to our store
5. **The high spenders - Low Earners** - Cluster group4 (ORANGE) representing the trusted clientele our clientele that even with low to moderated yearly renumeration their spending score equals and in certain cases surpasses individuals with 2 or 3 more times their year income

Analyse customer sentiments with reviews.

We commence with our original clean dataframe which we explore again to drop emtpty rows (none), duplicates (none) and confirm that we have a dataframe with 2000 entries (original size).
We will be focusing only on the columns **review** and **summary** which we keep and we drop the rest of the columns.
We run lambda functions to remove punctuation, special characters and to make all characters in lower case for each column separately. (above steps all needed for dataframe not to show false positives in the sentiment analysis).
IMPORTANT: Checking to reveal duplicates shows that different users having bought the same product might submit reviews or summaries that don't represent originality in terms of content but they appear to be 100% original in terms of data validity. Meaning that different users might have inserted the exact same words as reviews or summaries. In terms of data validity, no duplicate should be removed **THOUGH** for this part of the analysis as we are focused on the general sentiment, we will drop duplicates from **both combined columns**.
Our dataframe is being reduced to 1961 rows and we proceed with tokenization of both columns.
IMPORTANT: we can see from the tokenized data that hat we will have different spellingfor the same product (eg. "galeforce9" or "gf9s"), though in this part of the analysis we are interest in the sentiment of users and not product classification or review per se

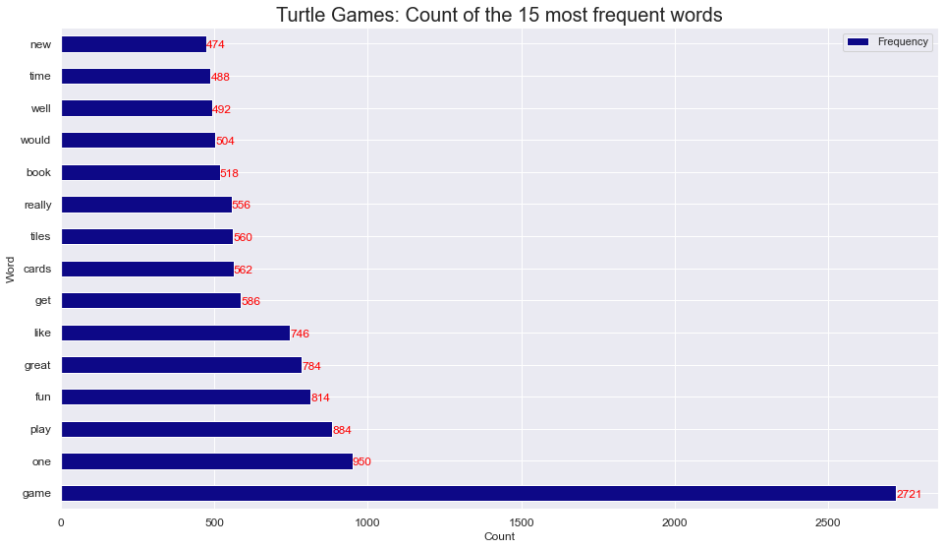
Running the function of **wordcloud** on our tokenised set of data doesn't reveal too much as we still have to isolate and remove the common stopwords. Once we pass from the stopwords function the produced wordcloud reveals very positive sentiment for our data

BEFORE

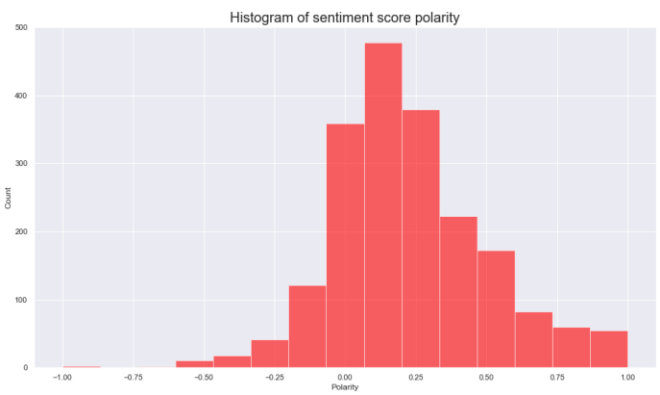
AFTER

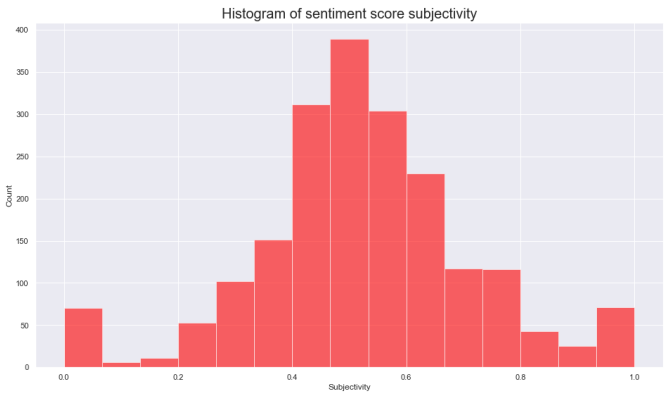
<u>Insight</u>: *What we are looking here is is a collection, of words depicted in different sizes. The bigger and bolder the word appears, the more often it's mentioned within a given text and the more important it is. Marketing should take note that words **game, fun, great, play** and **card** are quite bold and big. Reviews show that clients demonstrate a rather positive sentiment towards purchases from Turtle Games.*

Next step is to pass the wordcloud via a counter an produce again visually the top 15 most used words. Take note that play, fun great and like are among the top 10:



After having displayed the most frequently use words we continue with NLP and run the TextBlob library in research for polarity and subjectivity in our list of words. When a sentence is passed into Textblob it gives Polarity which is the output that lies between [-1,1], where -1 refers to negative sentiment and +1 refers to positive sentiment. Subjectivity is the output that lies within [0,1] and refers to personal opinions and judgments. We passed both reviews and summary columns and results can be seen in following histograms. Review show a good portion skewed towards positive polarity and subjectivity shows a normal distribution which makes our sample credible enough.





Our next step continues with the TextBlob library and this time we will focus not on the tokens pool but on the complete reviews and summaries (top 20 positive and negative) and relate it with other useful columns for further insights. Technically, we used the same variables which created in previous step with polarity score and applied sorting list to gather useful insights.
While we already partially answered to the question:" ***how social data (e.g. customer reviews) can be used to inform marketing campaigns***" by showcasing the info we can extract from any online review (or social media page like tweeter) we complete our answer by providing following examples/
Starting from top 20 negative which was sorted by remuneration and spending score, we have :

	review	gender	age	renumeration	spending_score	loyalty_points	education
989	If you, like me, used to play D&D, but now you...	Female	38	84.46	85	5019	graduate
182	Incomplete kit! Very disappointing!	Male	44	80.36	15	881	PhD
1524	Expensive for what you get.	Female	25	72.16	13	522	graduate
174	I sent this product to my granddaughter. The p...	Female	51	72.16	13	696	postgraduate
364	One of my staff will be using this game soon, ...	Male	49	69.70	26	1344	Basic
347	My 8 year-old granddaughter and I were very fr...	Female	34	63.14	74	3111	PhD
538	I purchased this on the recommendation of two ...	Male	18	60.68	10	266	graduate
117	I bought this as a Christmas gift for my grand...	Female	49	53.30	59	2332	postgraduate
306	Very hard complicated to make these.	Female	68	51.66	50	1668	postgraduate
301	Difficult	Female	49	50.84	48	1809	graduate
497	My son loves playing this game. It was recomme...	Female	29	49.20	50	1503	postgraduate
290	Instructions are complicated to follow	Female	70	48.38	55	1658	graduate
437	This game although it appears to be like Uno a...	Female	32	27.88	73	1314	postgraduate

Insight: *The underlined cases should be considered as an example of our VIP customers. We have underlined customers with many loyalty points (if the mean is about 1'500, cases with 2'300 and more points worth the business' attention) and decent remuneration and spending_score for more focus/attention/customer "extra support".*

Applying same for the summary column and we get following:

summary	gender	age	renumeration	spending_score	loyalty_points	education
The worst value I've ever seen	Male	27	19.680000	73	840	postgraduate
BORING UNLESS YOU ARE A CRAFT PERSON WHICH I AM ...	Male	66	15.580000	3	31	PhD
Boring	Female	25	23.780000	87	1151	graduate
before this I hated running any RPG campaign dealing with towns because it ...	Male	40	70.520000	20	1004	PhD
Another worthless Dungeon Master's screen from GaleForce9	Male	23	12.300000	81	524	graduate
Disappointed	Male	27	63.140000	12	443	postgraduate
Disappointed.	Female	23	24.600000	73	944	graduate
Disappointed	Female	38	92.660000	91	5895	graduate
Disappointed	Male	37	39.360000	59	1606	postgraduate
Promotes anger instead of teaching calming methods	Female	33	66.420000	93	4052	graduate
Too bad, this is not what I was expecting.	Male	46	44.280000	46	1501	diploma
Bad Quality-All made of paper	Female	70	48.380000	55	1658	graduate
At age 31 I found these very difficult to make ...	Male	58	76.260000	14	772	diploma
Small and boring	Female	49	50.840000	48	1809	graduate
It's UNO for the angry!	Male	49	50.840000	56	2111	PhD
Mad dragon	Female	50	54.940000	43	1752	postgraduate

Insight: *It is worth mentioning that libraries like TextBlob don't always give us 100% credible results. Underlined examples of high profile clients (always basis loyalty points and remuneration fields) show*

that their summary doesn't necessarily mean disgruntle. (ie example "It's UNO for the angry!" , might be a genuine summary without hints of negativity) . As such in future analysis, these exemptions should be well noted for better understanding of our results.

Passing dataframes basis the largest polarity scores we also identify interesting insights:

review	gender	age	renumeration	spending_score	loyalty_points	education
Came in perfect condition.	Female	25	14.760000	94	772	graduate
Absolutely great pictures even before coloring!	Female	49	31.980000	28	664	diploma
Great!	Male	45	35.260000	41	1062	graduate
Awesome book	Female	38	69.700000	75	3654	PhD
Awesome gift	Female	45	98.400000	16	1156	graduate
Great product! Arrived on time.	Female	37	17.220000	35	417	graduate
Great buy!! My granddaughter loves it!!	Female	50	54.940000	43	1752	postgraduate
Great!	Male	32	71.340000	75	3455	diploma
Great resource for BHIS care coordinators!! Works well with kids and teens on what it says it does!!	Male	51	18.860000	29	406	diploma
Great Seller!!! Happy with my purchase!!! 5 starrrr	Male	58	44.280000	47	1504	graduate
Excellent activity for teaching self-management skills!	Female	45	49.200000	47	1698	postgraduate
Great game...I use it a lot!	Male	28	50.840000	55	1673	PhD
Great therapy tool!	Female	49	53.300000	59	2332	postgraduate
Perfect, just what I ordered!!	Female	25	57.400000	29	926	graduate
Wonderful product	Female	34	84.460000	69	3880	diploma
Delightful product!	Female	32	15.580000	72	724	PhD
Great Easter gift for kids!	Male	37	19.680000	35	476	postgraduate
Wonderful for my grandson to learn the resurrection story.	Male	27	19.680000	73	840	postgraduate
These are great!	Male	46	44.280000	46	1501	diploma

Insight: *Maybe by limiting the review and summary fields to a minimum of 150 characters will creatively push more innovative reviews entries which will a) help Turtle games recover better results in future analysis b) populate the comments of a product with something more useful for others to consider than just a single word.*

	summary	gender	age	renumeration	spending_score	loyalty_points	education
199	Great product! Darling puppies!	Male	32	112.34	83	6020	PhD
187	Awesome	Male	32	82.82	68	3636	PhD
163	He was very happy with his gift	Female	33	66.42	93	4052	graduate
161	Awesome Book...	Female	29	64.78	83	3285	graduate
140	Awesome sticker activity for the price	Female	56	61.50	5	225	diploma
335	Another great book by Klutz!	Female	29	59.86	88	3218	graduate
134	Perfect for Preschooler	Male	22	59.86	5	152	graduate
122	Great for a gift!	Female	38	56.58	58	2294	graduate
80	They're the perfect size to keep in the car or...	Male	56	44.28	51	1651	graduate
57	great!	Male	71	36.08	46	1014	PhD
457	This is a great product! I use it as a therape...	Male	71	36.08	46	1014	PhD
449	Great resource!	Female	33	32.80	42	904	graduate
40	So beautiful!	Female	67	31.16	35	715	graduate
37	Great buy! Can't wait to work on this book	Female	32	27.88	73	1314	postgraduate

Insight: *On the same note with the negative reviews Turtle Games could probably give a bit more attention to customer profiles with above average score in loyalty_points, above average spending_score and the means (remuneration figures) to support and keep supporting purchases.*

Clean and manipulate data.

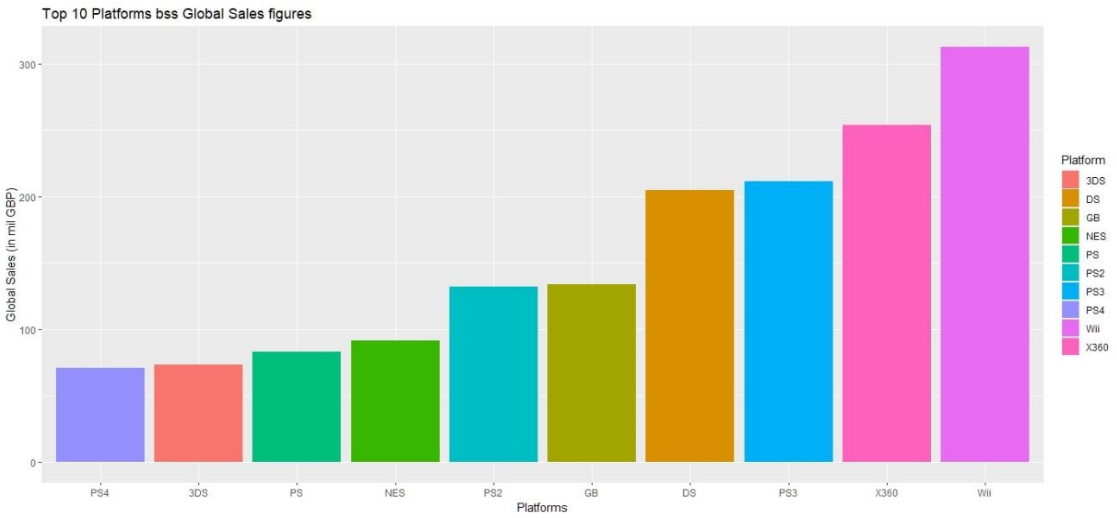
Using R we analyse the data received from csv turtles_sales. Our aim is to gather insight on the actual sales report data apart from the demographics of the clientele. At this point we should note, that we could join both datasets (ie turtle_reviews and turtle_sales) on common field Product but they don't share the same amount of records which mean that the combined dataframe wouldn't be coherent.

Dataframe is checked for duplicates, and NA (where we found 2 in the column Year)

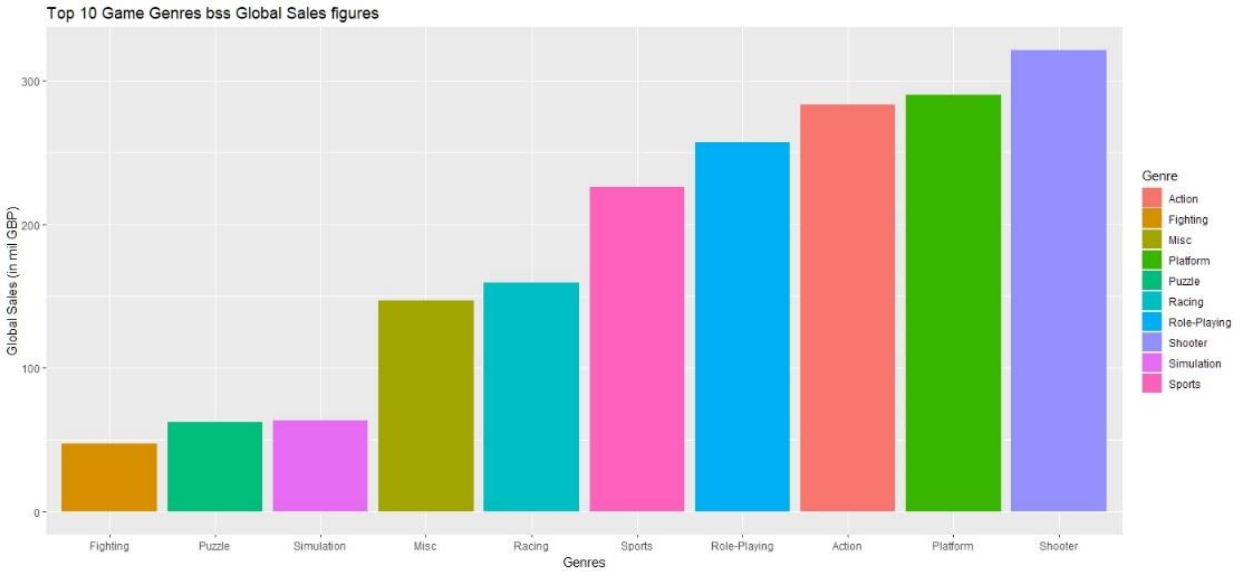
```
$Year
[1] 180 258
```

Before starting to drop columns for further analysis into sales, we have performed some queries for :

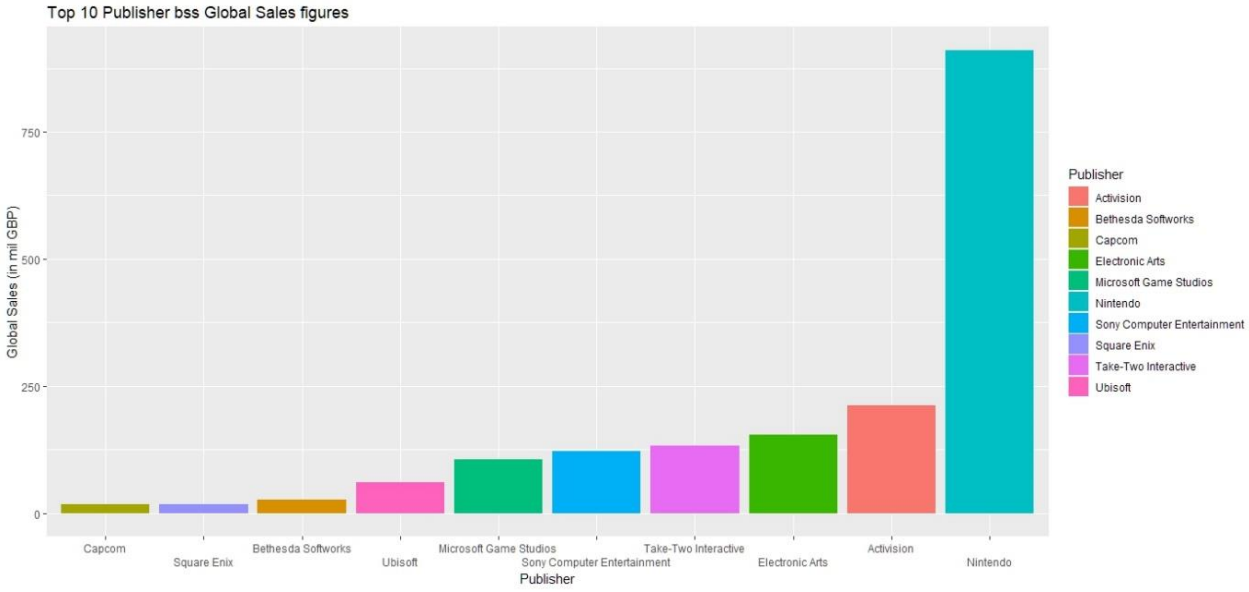
1) Top 10 Platforms of games sold at Turtle Games



2) Top 10 genres of games sold at Turtle Games



3) Top 10 Publishers of Games sold at Turtle Games



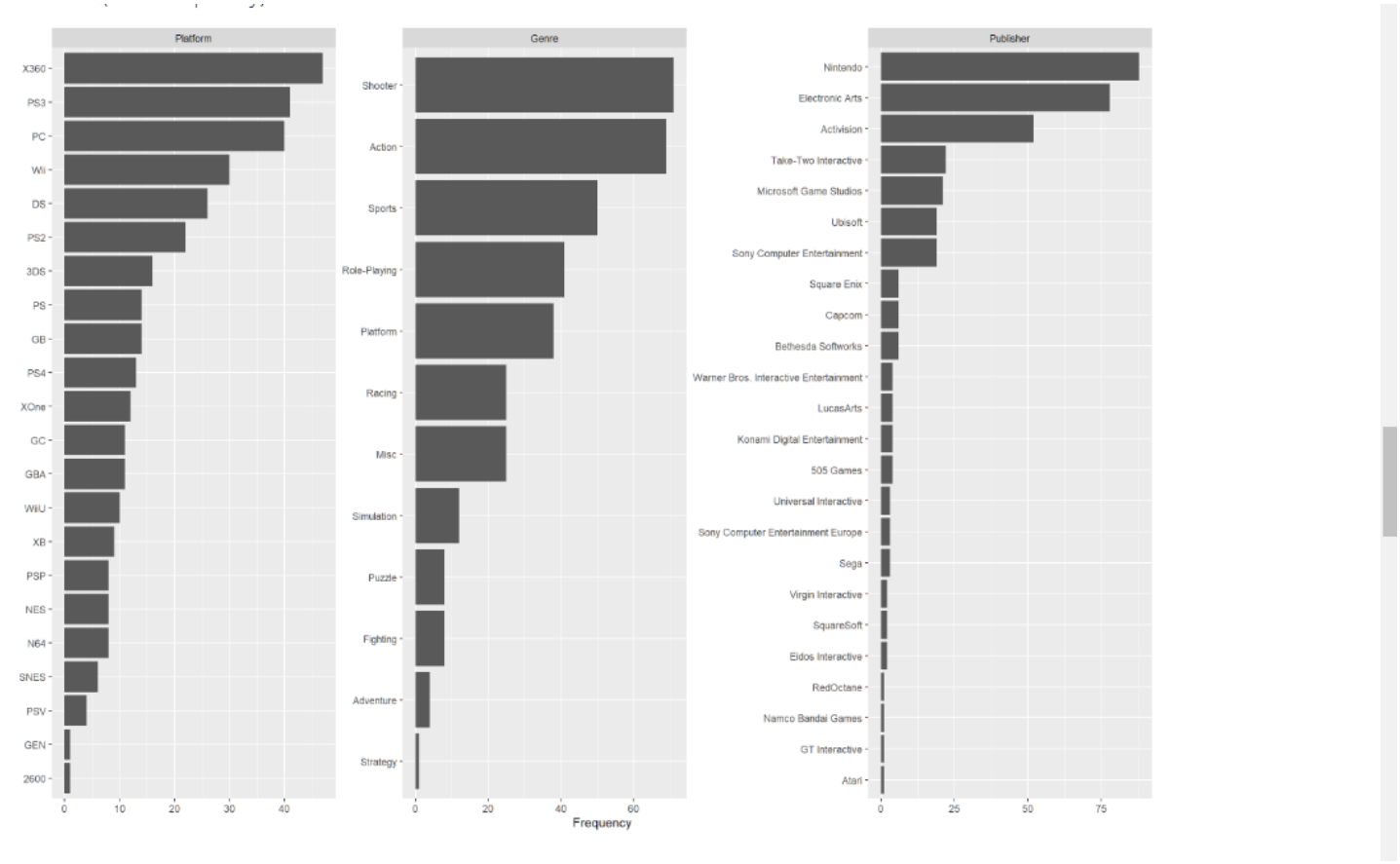
And last but not lease of the 10 Top Products

	Ranking	Product	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	Global_Sales
1	1	107	Wii	2006	Sports	Nintendo	34.02	23.80	67.85
2	2	123	NES	1985	Platform	Nintendo	23.85	2.94	33.00
3	3	195	Wii	2008	Racing	Nintendo	13.00	10.56	29.37
4	4	231	Wii	2009	Sports	Nintendo	12.92	9.03	27.06
5	5	249	GB	1996	Role-Playing	Nintendo	9.24	7.29	25.72
6	6	254	GB	1989	Puzzle	Nintendo	19.02	1.85	24.81
7	7	263	DS	2006	Platform	Nintendo	9.33	7.57	24.61
8	8	283	Wii	2006	Misc	Nintendo	11.50	7.54	23.80
9	9	291	Wii	2009	Platform	Nintendo	11.96	5.79	23.47
10	10	326	NES	1984	Shooter	Nintendo	22.08	0.52	23.21

The above screenshot gives us very quickly a big part of the answer to the question: "***the impact that each product has on sales*** "

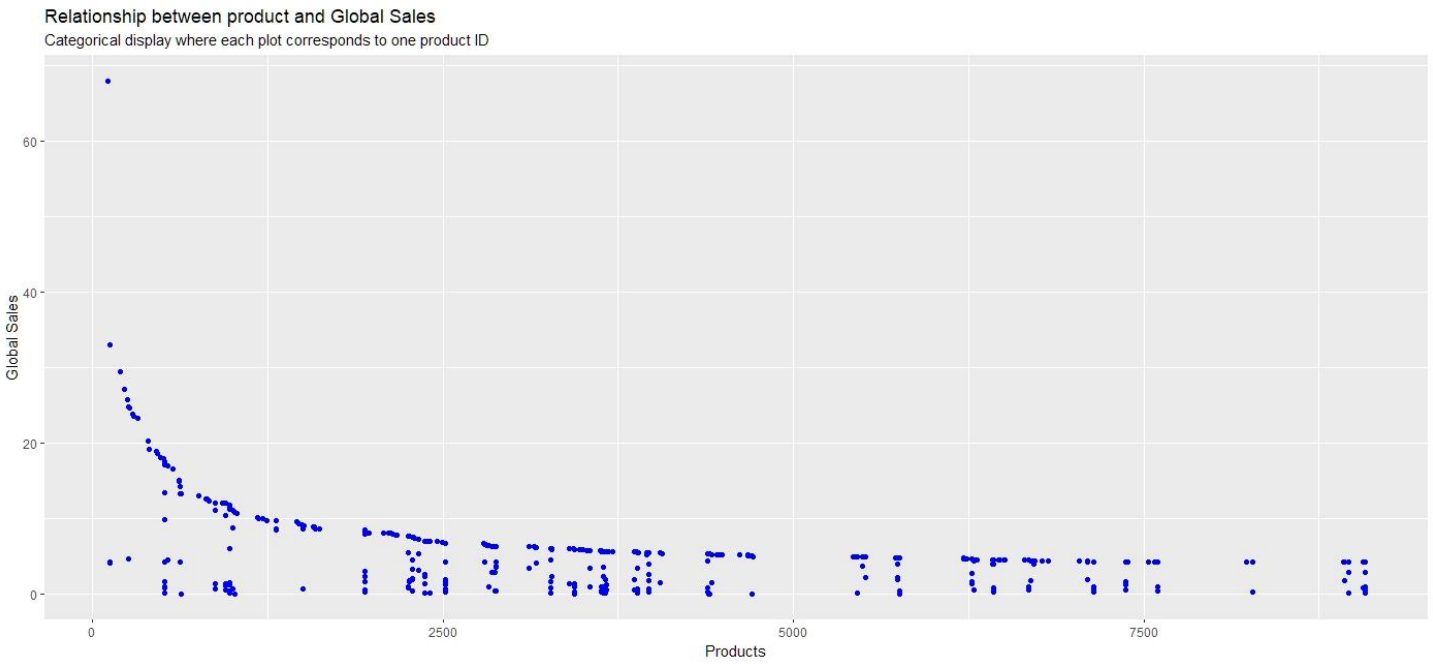
Technical insight: *The simplicity of R in data wrangling can be only demonstrated by the View() function which single handily answers the top 3 questions (sorting of data, ranking, and summarizing) all in one handy table format*

Same also results and easier with the plots provided before we can get by running the DataExplorer



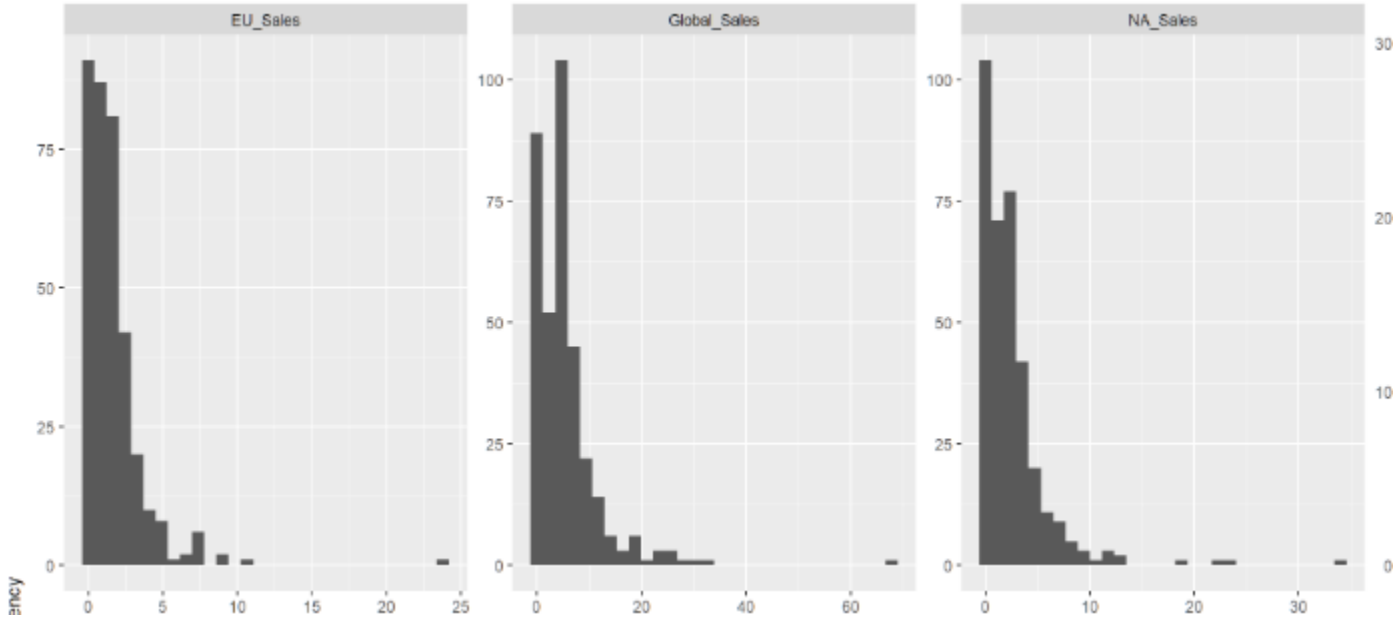
We continue our analysis by dropping columns (Ranking, Platform, Year, Genre and Publisher) in order to focus our exploratory analysis on the sales in different regions.

A quick scatterplot between Product (a categorical variable) and Global Sales (a continuous variable) verifies table with top 10 products above

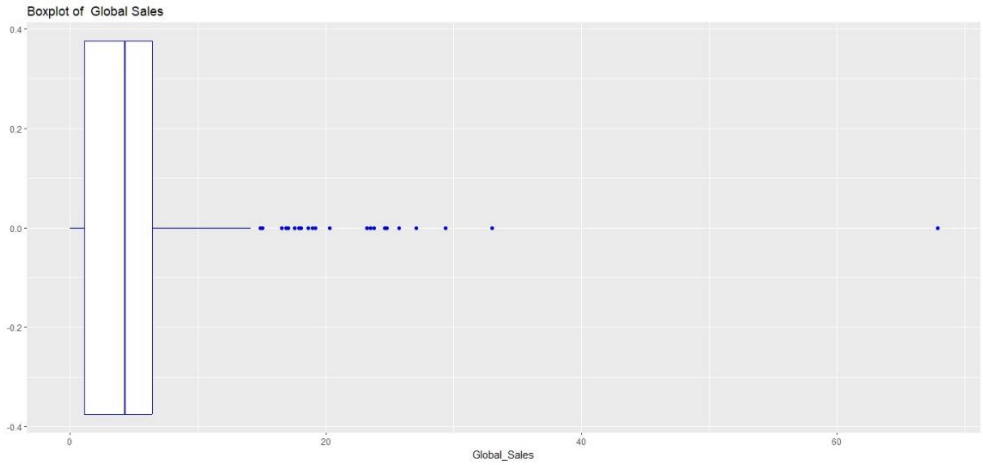


We can see the different product IDs and their sales figures in Global, NA and EU region. (all 3 regions show about same number of sales for same products)

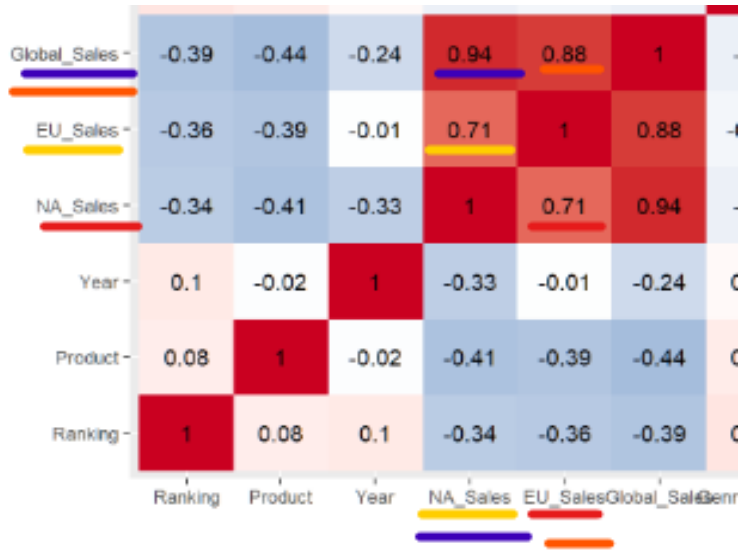
We perform a histogram on Sales column (EU, NA & Global) and all three show right (or positive) skewed distribution. We practically expected to have right skewed distribution as sales data can never be less than zero but they can have unusually large values (ref outliers).This should indicate that we should consider a further normality study of our variables and advanced plotting



We proceed on boxplots where we view following:



As expected from the right skewed histogram, in the boxerplot we find extreme values far from the peak on the high end more frequently than on the low. View of course that we analyse sales data we do expect to have outliers positive and high. We will return to check same with normality test.
As we want to examine the correlation between different regions of sales, easier and more practical than the plots is the correlation heatmap offered from the DataExplorer



- *Global Sales* show strong correlation **0.94** with *NA Sales*
- *Global Sales* show almost same strong correlation **0.88** with *EU Sales*
- *EU and NA Sales* share a smaller but still significant correlation of value **0.71**

Above give us a general idea on question: "***what the relationship(s) is/are (if any) between North American, European, and global sales?***" ,even though this question refers to possible relationships to be discovered vie simple and multiple linear regression models, there are other "relationships" in terms of different products or platform Top 10 in EU and NA regions accordingly. The correlation, being one of them, will give us a ratio basis the comparable numerical values column.

NB By executing a simple view of our studied dataframe we can see that top products are different in each region (which consequently means different focus on Top Genres, Platforms, Publishers).

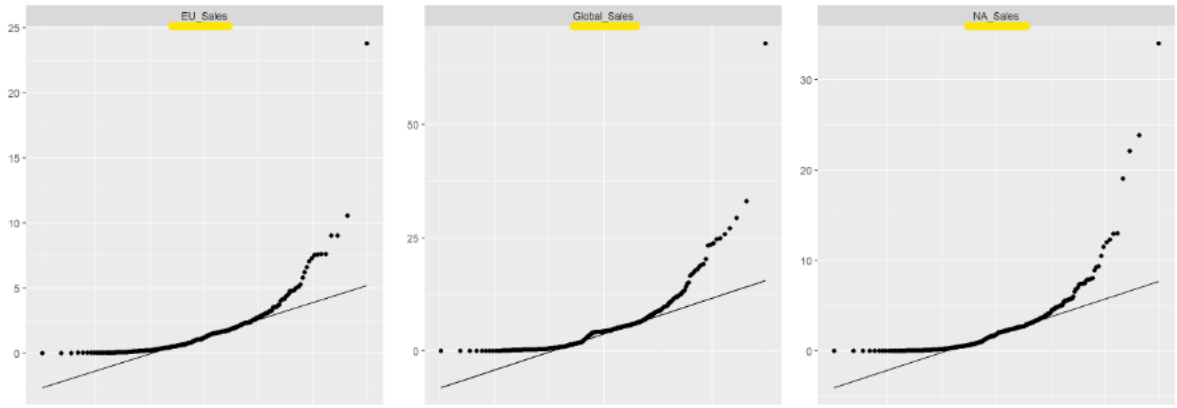
Product	Platform	Year	Genre	Publisher	↓ NA_Sa...
107	Wii	2006	Sports	Nintendo	34.02
123	NES	1985	Platform	Nintendo	23.85
326	NES	1984	Shooter	Nintendo	22.08
254	GB	1989	Puzzle	Nintendo	19.02
195	Wii	2008	Racing	Nintendo	13
231	Wii	2009	Sports	Nintendo	12.92
504	X360	2010	Misc	Microsoft Game Studios	12.28
291	Wii	2009	Platform	Nintendo	11.96
283	Wii	2006	Misc	Nintendo	11.5
535	SNES	1990	Platform	Nintendo	10.48
263	DS	2006	Platform	Nintendo	9.33
249	GB	1996	Role-Playing	Nintendo	9.24
618	GB	1989	Platform	Nintendo	8.88
405	DS	2005	Racing	Nintendo	8.04
948	X360	2010	Shooter	Activision	7.93
515	X360	2013	Action	Take-Two Interactive	7.9
624	NES	1988	Platform	Nintendo	7.82
518	PS2	2004	Action	Take-Two Interactive	7.73
486	Wii	2009	Sports	Nintendo	7.45
399	DS	2005	Simulation	Nintendo	7.44
876	X360	2011	Shooter	Activision	7.4

Product	Platform	Year	Genre	Publisher	NA_Sales	↓ EU_Sa...
107	Wii	2006	Sports	Nintendo	34.02	23.8
195	Wii	2008	Racing	Nintendo	13	10.56
231	Wii	2009	Sports	Nintendo	12.92	9.03
399	DS	2005	Simulation	Nintendo	7.44	9.02
515	PS3	2013	Action	Take-Two In...	5.75	7.6
577	DS	2005	Misc	Nintendo	3.9	7.59
263	DS	2006	Platform	Nintendo	9.33	7.57
283	Wii	2006	Misc	Nintendo	11.5	7.54
249	GB	1996	Role-Playing	Nintendo	9.24	7.29
486	Wii	2009	Sports	Nintendo	7.45	7.04
466	Wii	2007	Sports	Nintendo	7.33	6.58
405	DS	2005	Racing	Nintendo	8.04	6.21
291	Wii	2009	Platform	Nintendo	11.96	5.79
2521	PC	2009	Simulation	Electronic A...	0.8	5.26
4491	PC	2004	Role-Playing	Activision	0.06	5.09
453	GB	1999	Role-Playing	Nintendo	7.38	5.07
2371	PS4	2015	Sports	Electronic A...	0.91	4.97
979	PS3	2012	Shooter	Activision	4.09	4.82

<u>Insight:</u> *We have identified quite different trends in both regions that might be absorbed from Global figures but show that each region requires its own attention and adapted marketing strategy.*

Predict sales with regression.
report you will clearly articulate a response to the questions posed by Turtle Games and summarise your analysis by providing recommendations to Turtle Games. You can also share any obstacles you faced in the process and how you overcame them.

We continue with the examined dataframe containing the columns (Product, NA_Sales, EU_Sales and Global_Sales) to view qqplots for sales values:



The reference line along with shape of qqplot for Global Sales show that our values
do not follow a normal distribution (which is also is what positive skewed results confirmed).

Next we proceed with Shapiro-Wilk test to examine how close the sample data (of our sales column) fit to a normal distribution .

```
        Shapiro-Wilk normality test

data:  (tsales2_pro$Global_Sales_sum)
W = 0.70955, p-value < 2.2e-16
```

All Shapiro-Wilk test show p-values below 0.05 which is another confirmation that we can reject such a null hypothesis and assume that the **sample** has **not** been generated from a **normal distribution**

Similar results produce also the Skewness and Kurtosis sets.
For the skewness in all 3 examines sales columns we witness similar results of the kind:

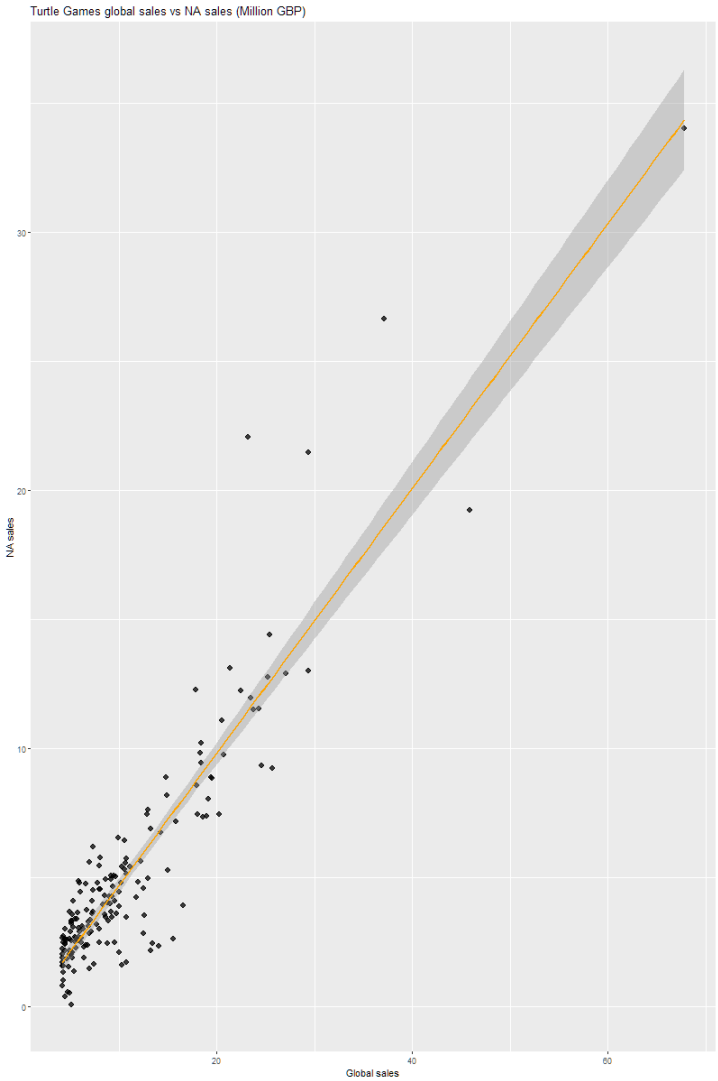
```
> skewness(tsales2_pro$Global_Sales_sum)
[1] 3.066769
```

As already confirmed, all 3 sales variables are heavily skewed to the right.
Following with Kurtosis we notice similar results

```
> kurtosis(tsales2_pro$Global_Sales_sum)
[1] 17.79072
>
```

Some interpretations for high results of kurtosis is the outliers in a sample which have been verified to be significant. Also higher kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations but all that applies only to the sample of our dataset and not the whole population.
To the question:" ***how reliable the data is (e.g. normal distribution, skewness, or kurtosis)***" wehave replied with many examples showing that this data (as most of real world examples) is not product of normal distribution but is heavily positively skewed with high figures of kurtosis confirming the plethora and high values of outliers.

Plotting Global Sales against NA and EU Sales shows the correlation mentioned before but adding also the trend line we are being guided that we can apply linear regression model (or possibly logistical) to address the issue of predictive functions.



Predict Sales With Regression

During this part of our analysis we built simple linear regression models to confirm whether there is strong correlation between the examined variables:

```
> summary(model1)

Call:
lm(formula = NA_Sales_sum ~ Global_Sales_sum, data = tsales2_pro)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9263 -0.6760  0.0729  0.7721 10.6105

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.44975    0.22960  -1.959   0.0517 .
Global_Sales_sum  0.51354    0.01707  30.079   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.831 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
```

```
> summary(model2)

Call:
lm(formula = EU_Sales_sum ~ Global_Sales_sum, data = tsales2_pro)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -0.6114 -0.0654  0.5079  5.2992

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.14813    0.20519  -0.722    0.471
Global_Sales_sum  0.32194    0.01526  21.099   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.636 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
```

For the compared Global Sales to NA Sales and Global Sales to EU Sales we have an adjusted R-Squeared Value of 0.8385 and 0.7185 which denotes that both models have very good chances of predicting values in the target field. Compared to model between EU and NA sales where besides the extremely small p-value the adjusted R-Squared is 0.382.

```
Call:
lm(formula = EU_Sales_sum ~ NA_Sales_sum, data = tsales2_pro)

Coefficients:
 (Intercept)  NA_Sales_sum

Residuals:
    Min      1Q  Median      3Q     Max
-9.9391 -1.1930 -0.4267  0.7023  9.6102

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.17946    0.27433   4.299 2.85e-05 ***
NA_Sales_sum  0.42028    0.04034  10.419  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.424 on 173 degrees of freedom
Multiple R-squared:  0.3856,    Adjusted R-squared:  0.382
F-statistic: 108.6 on 1 and 173 DF,  p-value: < 2.2e-16
```

We then proceed on implementing a multiple regression model

```
> summary(modelA)

Call:
lm(formula = Global_Sales_sum ~ NA_Sales_sum + EU_Sales_sum,
    data = tsales3)

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 ***
NA_Sales_sum  1.13040    0.03162  35.745  < 2e-16 ***
EU_Sales_sum  1.19992    0.04672  25.682  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

The initial multiple variables regression model shows promising adjusted r-squared very close to 1.

We therefore proceed for testing its reliability by comparing predicted values vs observation values (figures from tsales dataframe). We perform 5 comparisons where we get 3 average and 2 good results.

Example of good prediction

```r
# A. NA_Sales_sum of 34.02 and EU_Sales_sum of 23.80.
NA_Sales_sum <- c(34.02)
EU_Sales_sum <- c(23.80)

tsales4 <- data.frame(NA_Sales_sum, EU_Sales_sum)

# Predicted Global_Sales value.
predict(modelA, newdata = tsales4)
# Predicted value 68.056 vs observation value 67.85: good .
```

Example of average prediction

```r
# B. NA_Sales_sum of 3.93 and EU_Sales_sum of 1.56.
# Proceed with 3.94/1.28 with observed Global_sales value 8.36 as found in DF.
NA_Sales_sum <- c(3.94)
EU_Sales_sum <- c(1.28)

tsales5 <- data.frame(NA_Sales_sum, EU_Sales_sum)

# Predicted Global_Sales value.
predict(modelA, newdata = tsales5)
# Predicted value 7.03 vs observation value 8.36: average.
```

Insight: *The logistical regression model seems to provide quite correct results and it will get improved if we add more variables and also feed it with more data to improve its accuracy.*