DONATA POLLOCK's GITHUB REPOSITORY:
https://github.com/donata-p/LSE_DA301_assignment

'Predicting future outcomes' stakeholder report

Context and problem statement

Turtle Games is 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 utilising customer trends.

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.

Data and general statistics

	gender	age	remuneration	spending_score	loyalty_points	education	product	review	summary
0	Male	18	12.30	39	210	graduate	453	When it comes to a DM's screen, the space on t...	The fact that 50% of this space is wasted on a...
1	Male	23	12.30	81	524	graduate	466	An Open Letter to GaleForce9*:\n\nYour unpaint...	Another worthless Dungeon Master's screen from...
2	Female	22	13.12	6	40	graduate	254	Nice art, nice printing. Why two panels are f...	pretty, but also pretty useless
3	Female	25	13.12	77	562	graduate	263	Amazing buy! Bought it as a gift for our new d...	Five Stars
4	Female	33	13.94	40	366	graduate	291	As my review of GF9's previous screens these w...	Money trap

Picture 1.

Picture 1. displays a snapshot of social media reviews data.

Filter

	Product	Platform	NA_Sales	EU_Sales	Global_Sales
1	107	Wii	34.02	23.80	67.85
2	123	NES	23.85	2.94	33.00
3	195	Wii	13.00	10.56	29.37
4	231	Wii	12.92	9.03	27.06
5	249	GB	9.24	7.29	25.72
6	254	GB	19.02	1.85	24.81
7	263	DS	9.33	7.57	24.61
8	283	Wii	11.50	7.54	23.80
9	291	Wii	11.96	5.79	23.47
10	326	NES	22.08	0.52	23.21

Picture 2.

Picture 2. displays a snapshot of the NA, EU and Global sales data.

DONATA POLLOCK's GITHUB REPOSITORY:
https://github.com/donata-p/LSE_DA301_assignment

```
    Product       Platform           NA_Sales          EU_Sales        Global_Sales
3645    :  9   Length:350         Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.010
2518    :  8   Class :character   1st Qu.: 0.4725   1st Qu.: 0.3925   1st Qu.: 1.045
3967    :  8   Mode  :character   Median : 1.8200   Median : 1.1900   Median : 4.330
3887    :  7                      Mean   : 2.5190   Mean   : 1.6507   Mean   : 5.349
9080    :  7                      3rd Qu.: 3.1200   3rd Qu.: 2.1600   3rd Qu.: 6.445
1945    :  6                      Max.   :34.0200   Max.   :23.8000   Max.   :67.850
(Other):305
```

Picture 3.

Sales data contains 352 games, 10 platforms and 12 genres.
Global sales mean 5.349 million, while the EU and NA mean sales are 1.6507 million and 2.519 million respectively.

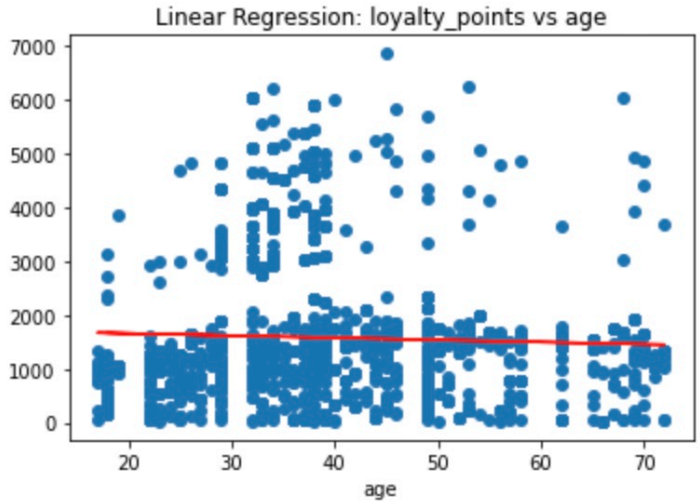
Loyalty points accumulation

Picture 4.

Picture 4. displays how loyalty points are accumulated amongst males and females.

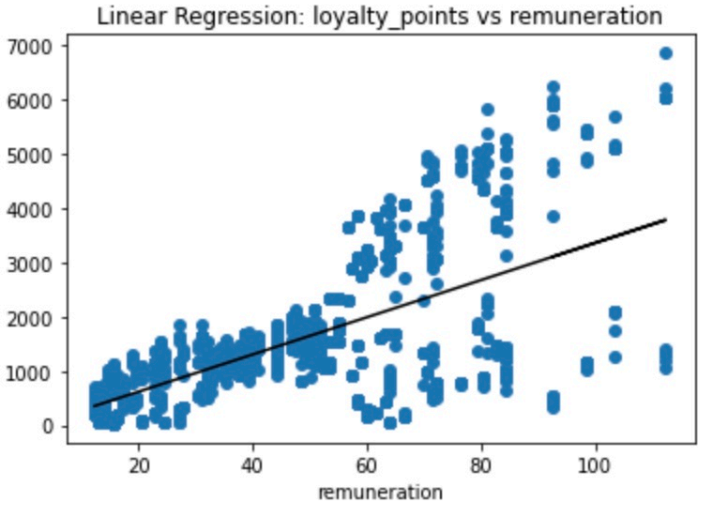
DONATA POLLOCK's GITHUB REPOSITORY:
https://github.com/donata-p/LSE_DA301_assignment

I have built linear regression models using quantitative variables loyalty points, spending score, remuneration and age.



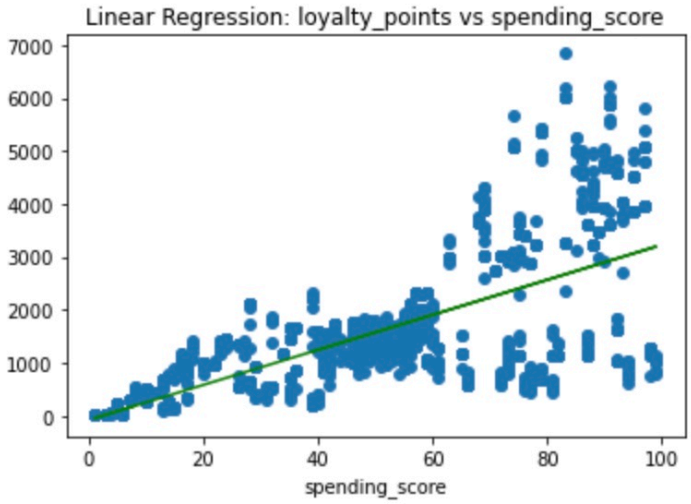
Picture 5.

Picture 5. displays there is not relationship between loyalty points and age.



Picture 6.

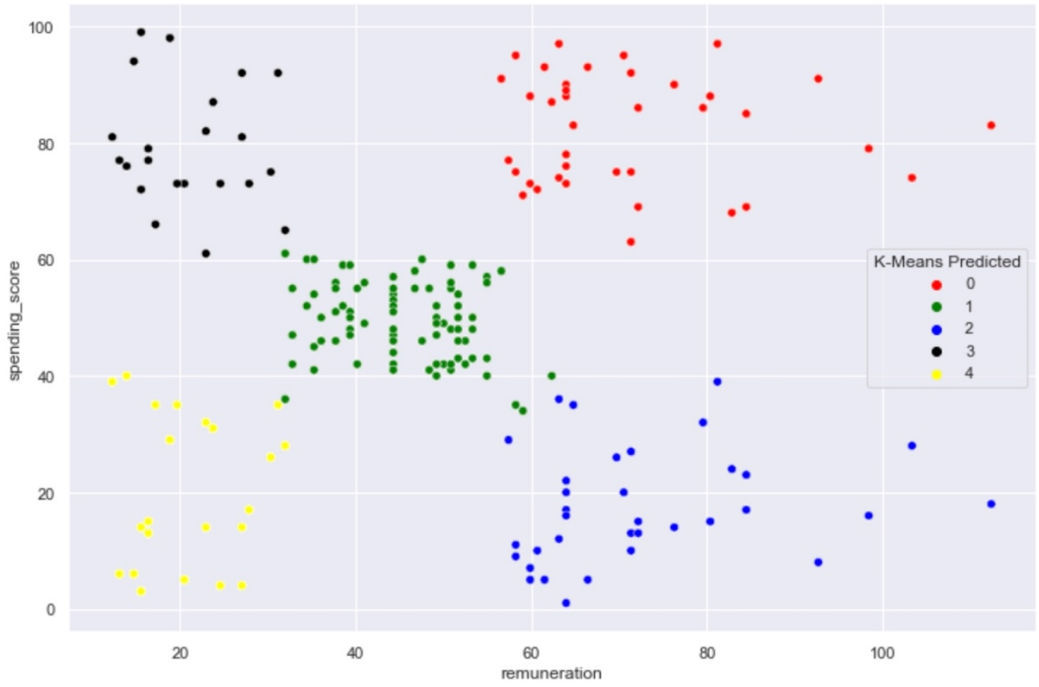
Picture 6. displays slightly stronger positive relationship between loyalty points and remuneration.



Picture 7.

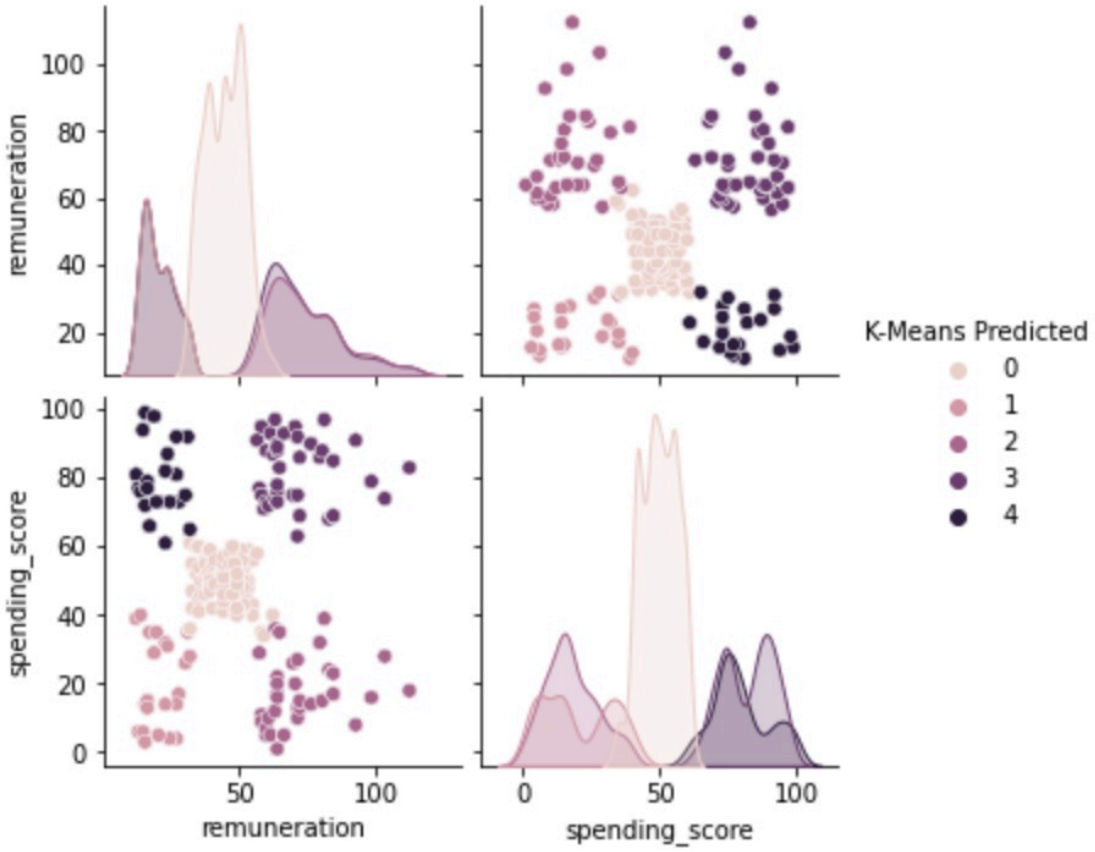
Picture 7. displays there is a slightly stronger positive relationship between loyalty points and spending.

Customer market segments



Picture 8.

Picture 8. displays 5 customer segments with some outliers based on remuneration and spending.



Picture 9.

Picture 9. Displays 5 customer segments based on remuneration and spending using k-means model.

Social media impact on marketing campaigns

Picture 10.

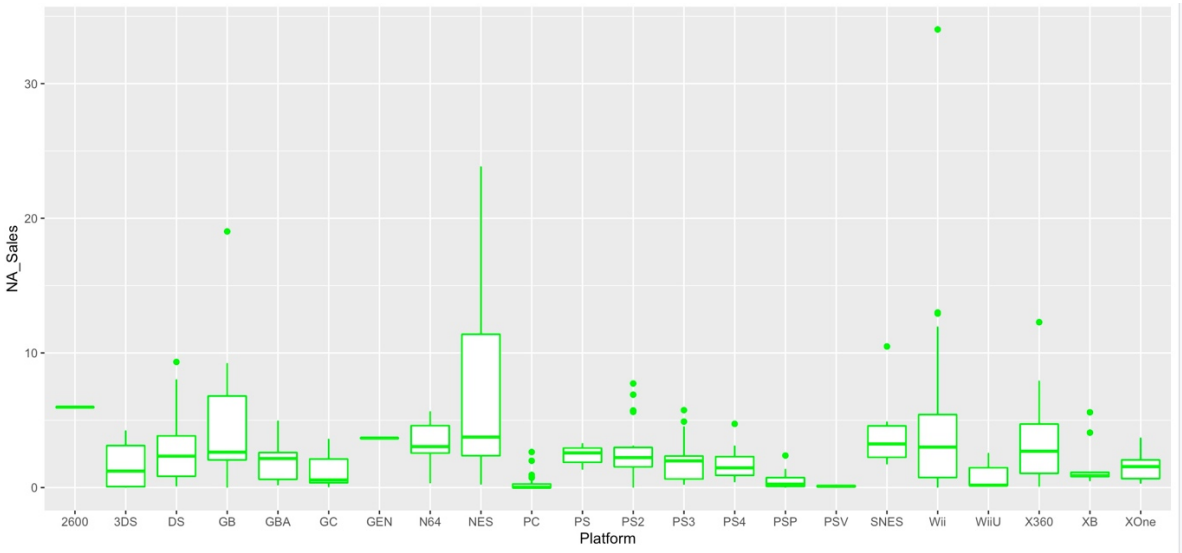
Picture 10. displays most popular words in social media reviews.

Picture 11.

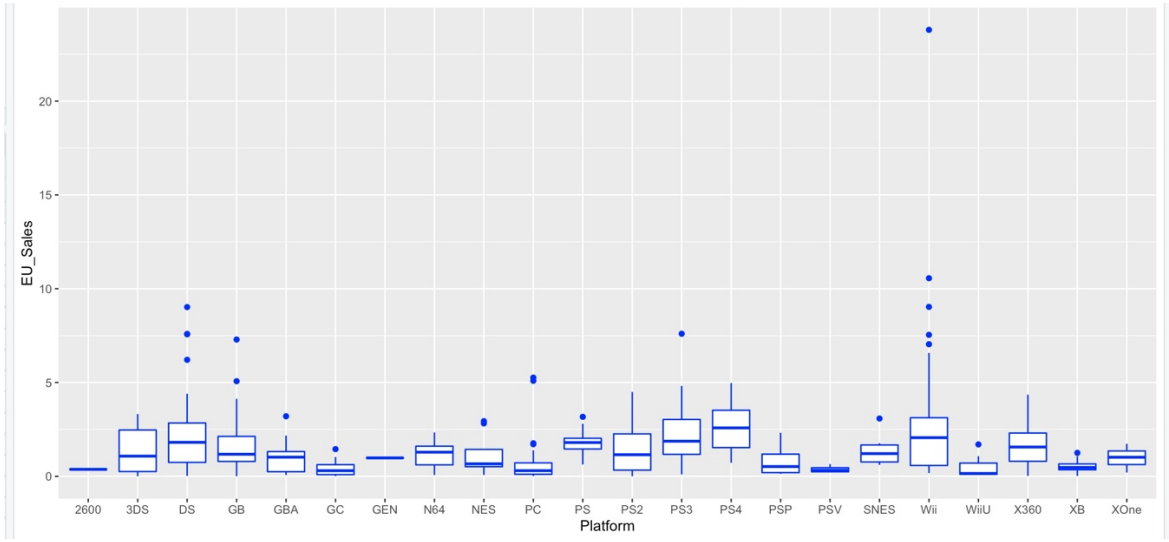
Picture 11. Displays most popular words in ratings summary.

The above visualisations are produced using sentiment analysis.

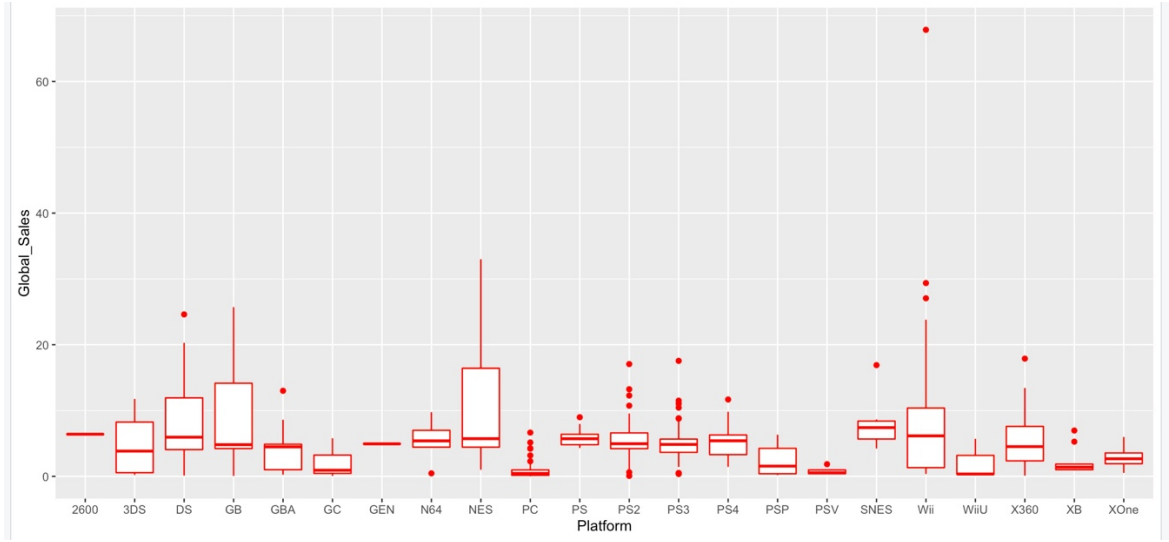
Product impact on sales



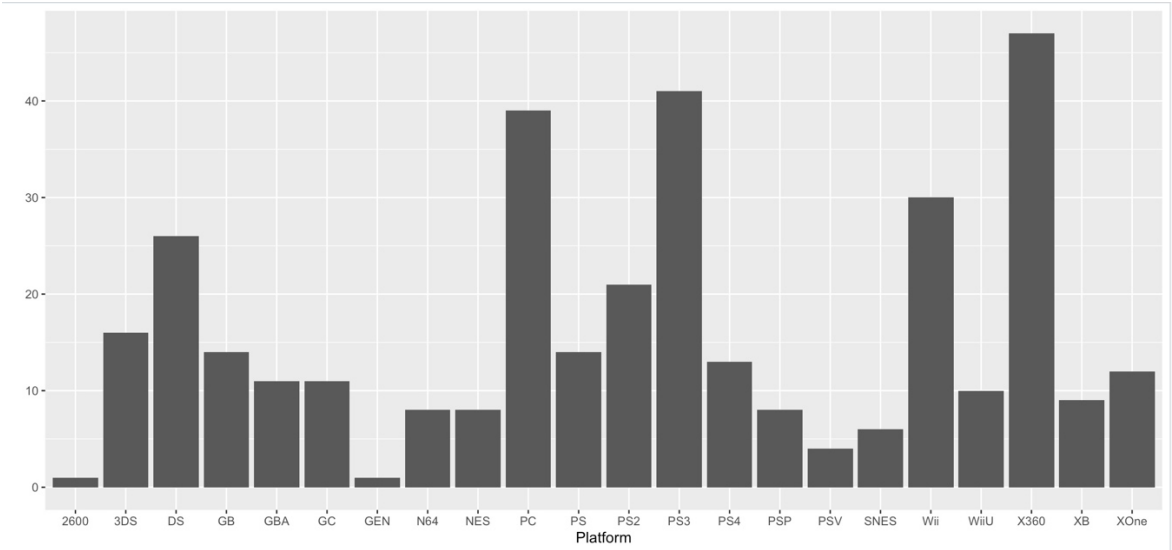
Picture 12. displays product sales per platform in North America (NA).



Picture 13. displays product sales per platform in European Union (EU).



Picture 14. displays product sales per platform Globally.



Picture 15. displays product sales on all planforms across all regions (NA, EU and Globally).

Data reliability

Normal distribution is performed with Shapiro-Wilk normality test:
data: data5$sum_NA_Sales
W = 0.69832, p-value < 2.2e-16

data: data5$sum_EU_Sales
W = 0.74114, p-value = 3.11e-16

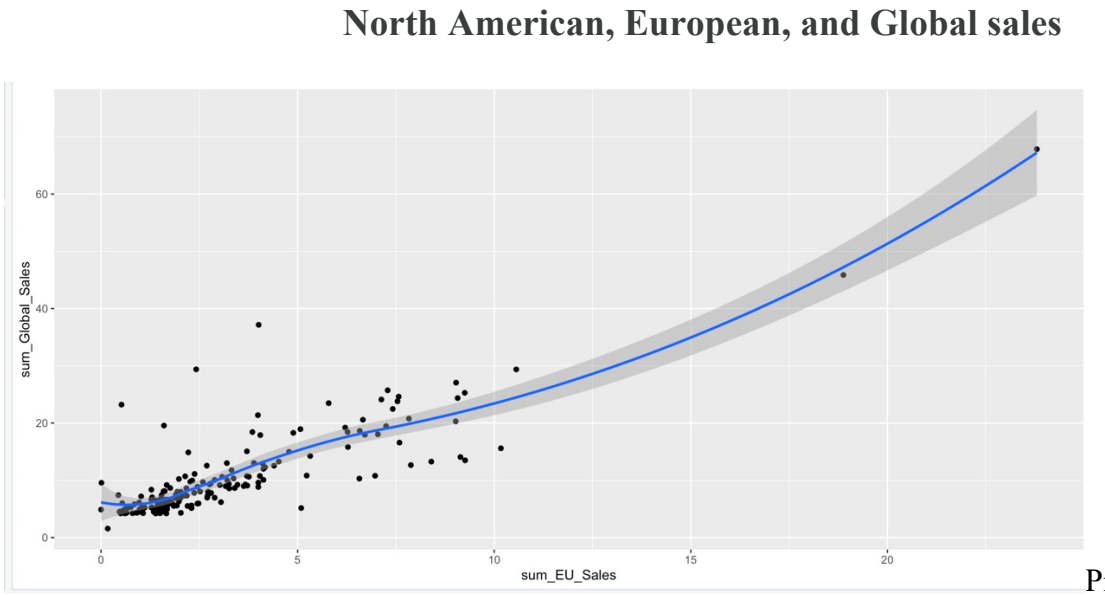
data: data5$sum_Global_Sales
W = 0.71671, p-value < 2.2e-16
The p-value is small – less than 5%, say – we would conclude that the assumption of normality is a poor fit for the data.
There is little evidence that NA, EU and Global sales are normally distributed.

Skewness:
> skewness(data5$sum_NA_Sales)
[1] 3.046616
> skewness(data5$sum_EU_Sales)
[1] 2.894049
> skewness(data5$sum_Global_Sales)
[1] 3.054206
Provided data skewness is greater than 0. This is positive skewness, which suggests that the distribution is very right-skewed and based towards higher values. Positive skewness means the mean (avg) is larger than the median and data leans to the right.

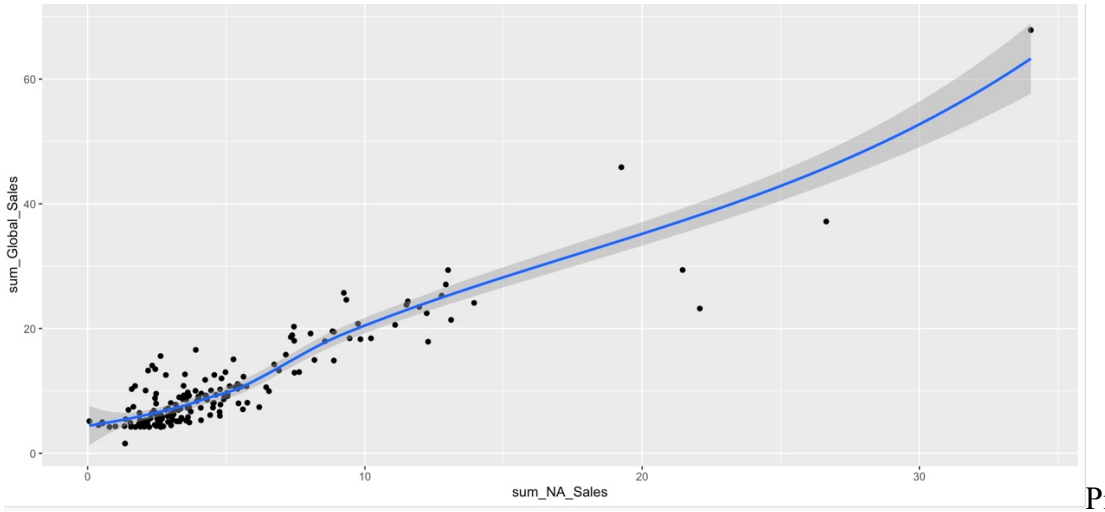
Kurtosis:
> kurtosis(data5$sum_NA_Sales)
[1] 15.60427
> kurtosis(data5$sum_EU_Sales)
[1] 16.31649

> kurtosis(data5$sum_Global_Sales)
[1] 17.75875
Provided data kurtosis is higher than 3 indicates a leptokurtic (or heavy-tailed) distribution, that is one with more extreme outliers than the normal distribution. This data is leptokurtic and will produce extreme outliers rather than the normal distribution.

Correlation:
> cor(data5$sum_NA_Sales, data5$sum_Global_Sales)
[1] 0.9168662
> cor(data5$sum_EU_Sales, data5$sum_Global_Sales)
[1] 0.8487806
> cor(data5$sum_EU_Sales, data5$sum_NA_Sales)
[1] 0.622516
Provided data has positive correlation coefficient suggests that the two variables vary in the same direction. This means that as one sales figure increases the other sales figures will also increase. The highest correlation is between NA and Global sales of 0.91.

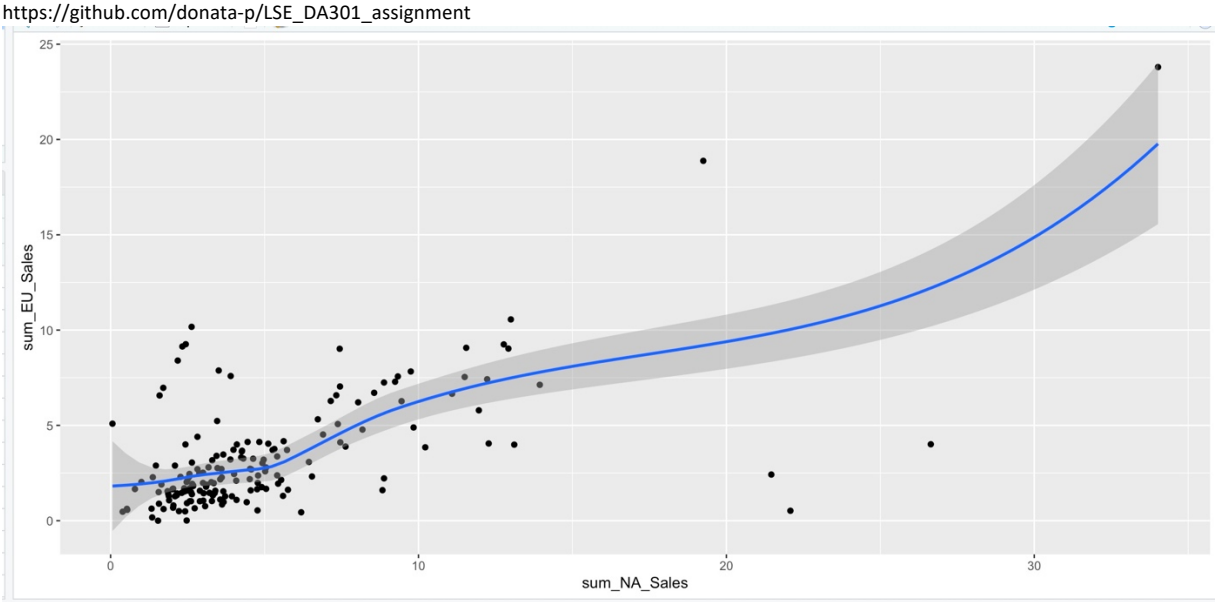


Picture 16.
Picture 16 displays relationship between EU sales and Global sales. Dependant on one another.



Picture 17.
Picture 17 displays relationship between NA sales and Global sales. Dependant on one another.

DONATA POLLOCK's GITHUB REPOSITORY:
https://github.com/donata-p/LSE_DA301_assignment



Picture 18. displays relationship between NA sales and EU sales. Dependant on one another.