Assignment report: Predicting future outcomes.

1. Background.

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. Using collected data from sales as well as customer reviews, Turtle Games aims at improving its overall sales performance. To achieve this goal, we will provide the company with actionable insights using data analytics on:

- How customers accumulate loyalty points
- How the customer base could be segmented
- How customer reviews can be used for marketing campaigns
- What relationships are there in the sales from the different regions and the Global sales?

2. Make predictions with regression.

So, our first step is to prepare our workstation, starting with importing the necessary libraries and the Turtle reviews data set. The dataset contains 11 variables with 2000 entries. There were no missing values in any of the 11 variables.

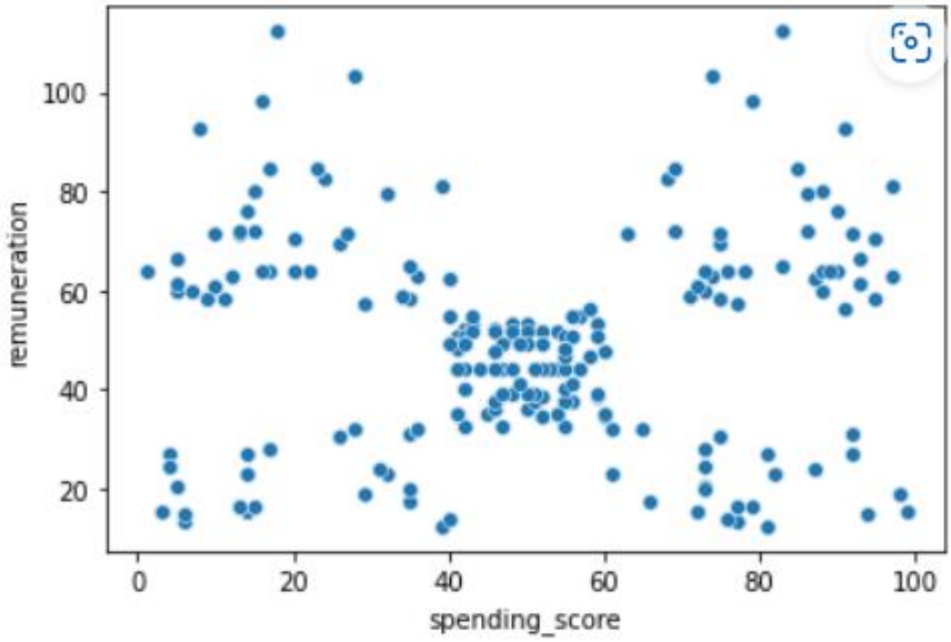
As requested, to understand how customers accumulate loyalty points, we would evaluate any potential relationship between accumulated loyalty points by the customer and the attributes such as customer age, spending, etc. Therefore, looking at the summary of the different ordinary least squares models, we could conclude that:

- Compared to the spending habit of customers alone, almost half, 45% precisely, of the total change in the loyalty points is explained by the variation in the spending.
- In the same logic, when compared to customer remuneration alone, 38% of the total variation in the loyalty points accumulated is explained by the change in remuneration.
- The age attribute on the other hand though has no connection to how customers accumulate loyalty points as the regression model showed an R^2 of 0, therefore age cannot be used to predict loyalty points.

What is interesting to note though is that when running a multiple linear regression model with both customer remuneration and the spending habit, we could see that the combination of these two variables resulted in a way significant R^2 which means that 82% of each variation in the loyalty points accumulated the customer is explained by a combined variation of customer remuneration and the spending score.

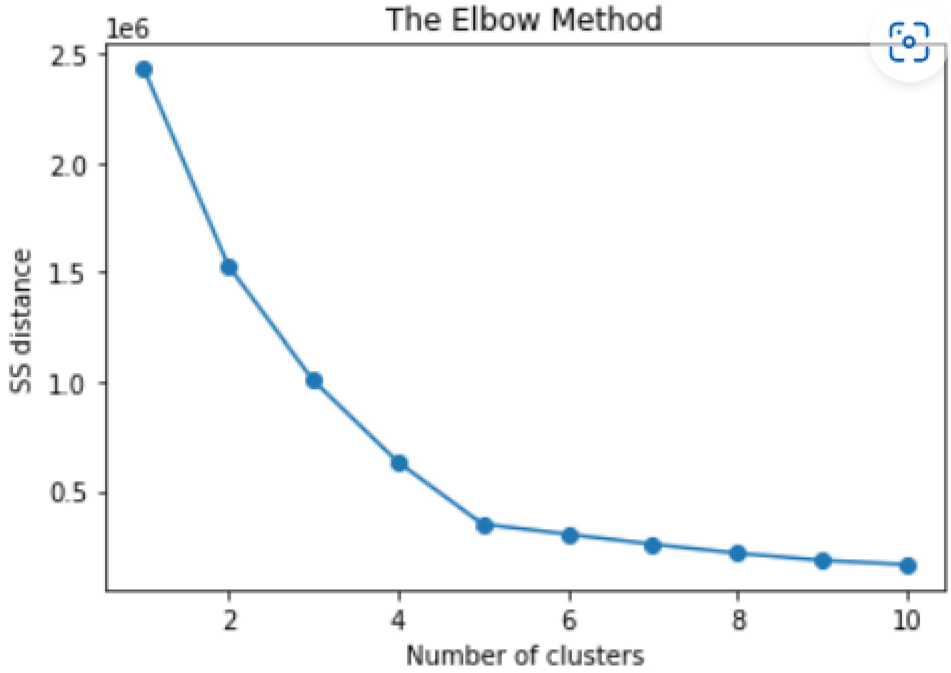
3. Make predictions with clustering.

Next, to determine the different groups within the customer base that can be used to target specific market segments, we used the k-means clustering method. At a first glance of a scatterplot to determine the relationship between remuneration and spending score we could see some sort of groups formed by the data points that we would need to confirm using the elbow and silhouette methods.

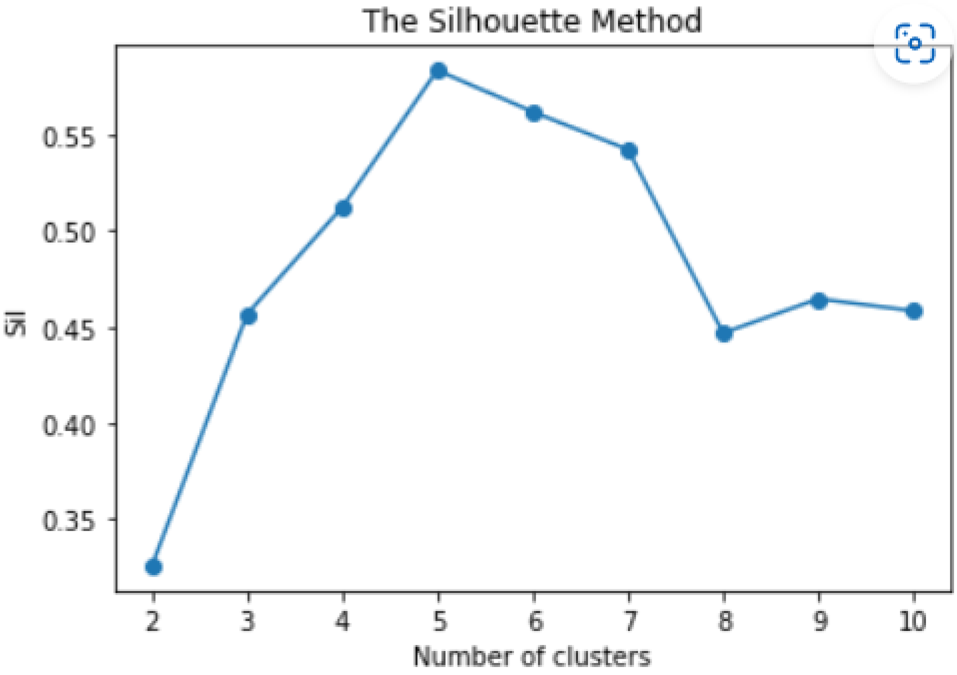


From the above-mentioned clustering techniques to determine the optimal number of clusters we could conclude the following:

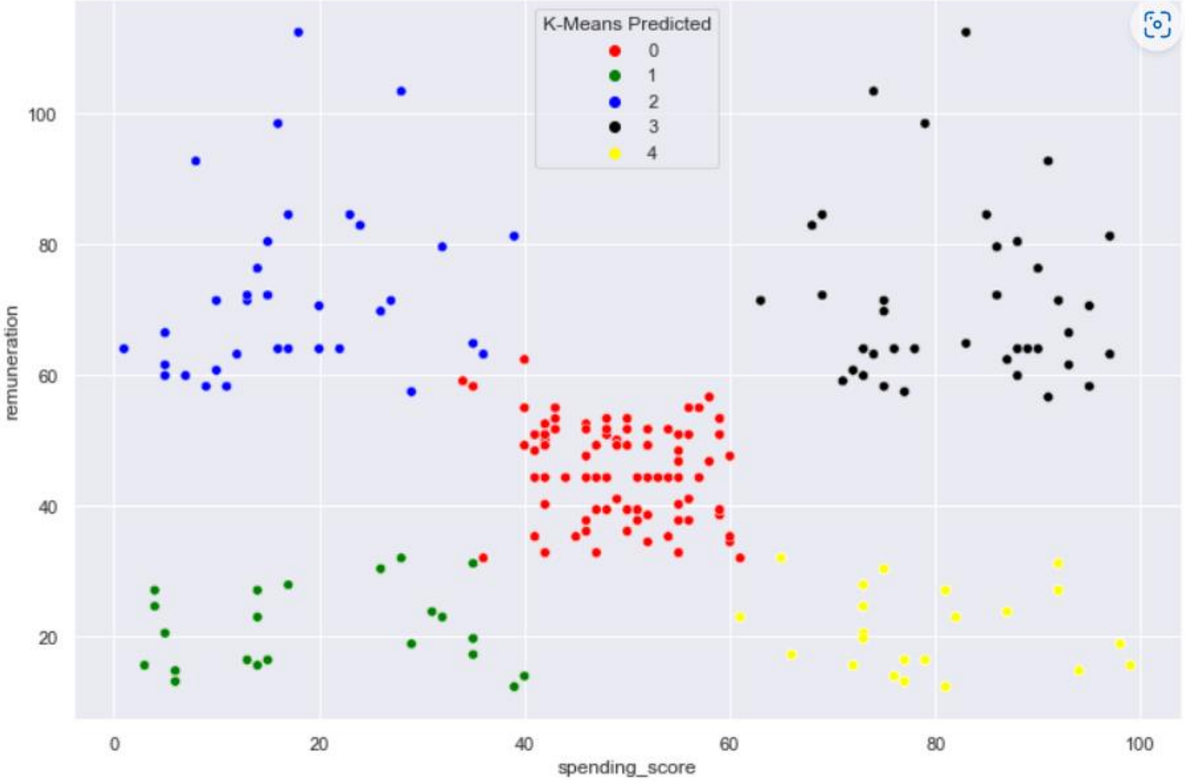
- Using the elbow method, we could see that the plot suggests that the elbow is formed with values of k between 3 and 5 and the curve started to flatten up after 5.



- Looking at the silhouette method, the highest point on the curve appears to be 5, therefore the optimum number of clusters is 5.



When plotted using the 5 clusters, 'k=5', we could see a very distinguishable grouping of the 5 clusters that are sufficiently homogenous with almost insignificant overlap.



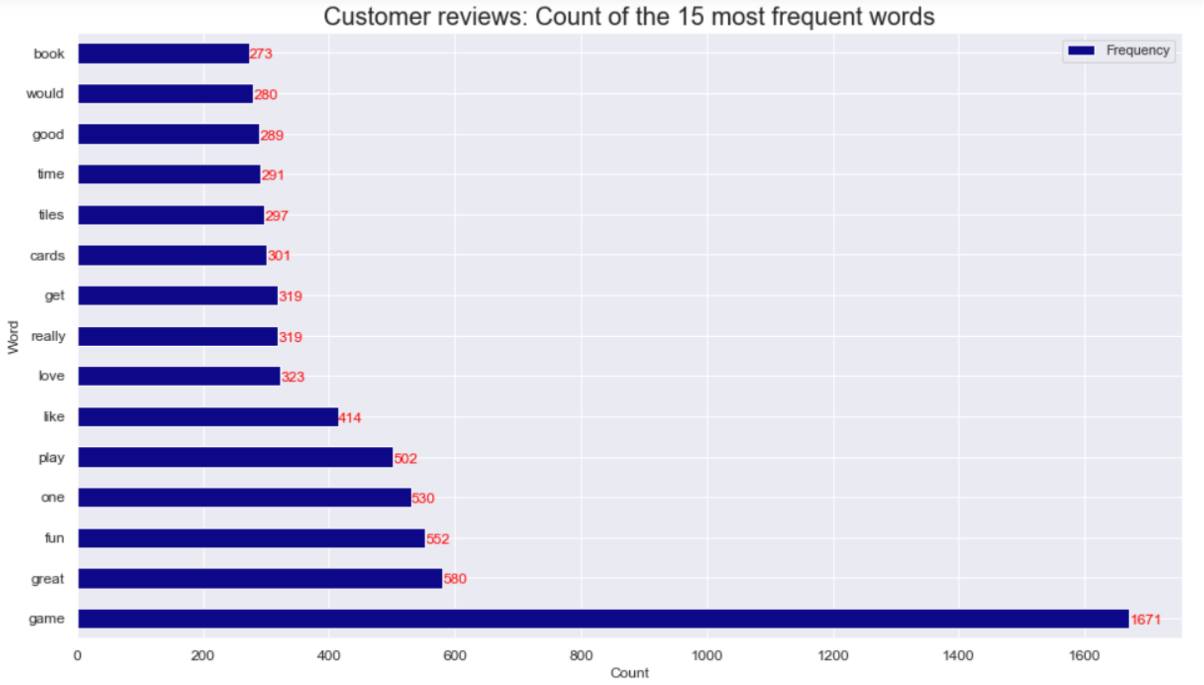
Further exploration of the dataset, using descriptive statistics such as each group’s mean or standard deviation will help provide more characteristics profiles to show how meaningful the segments are.

4. <u>Analyse customer sentiments with reviews.</u>

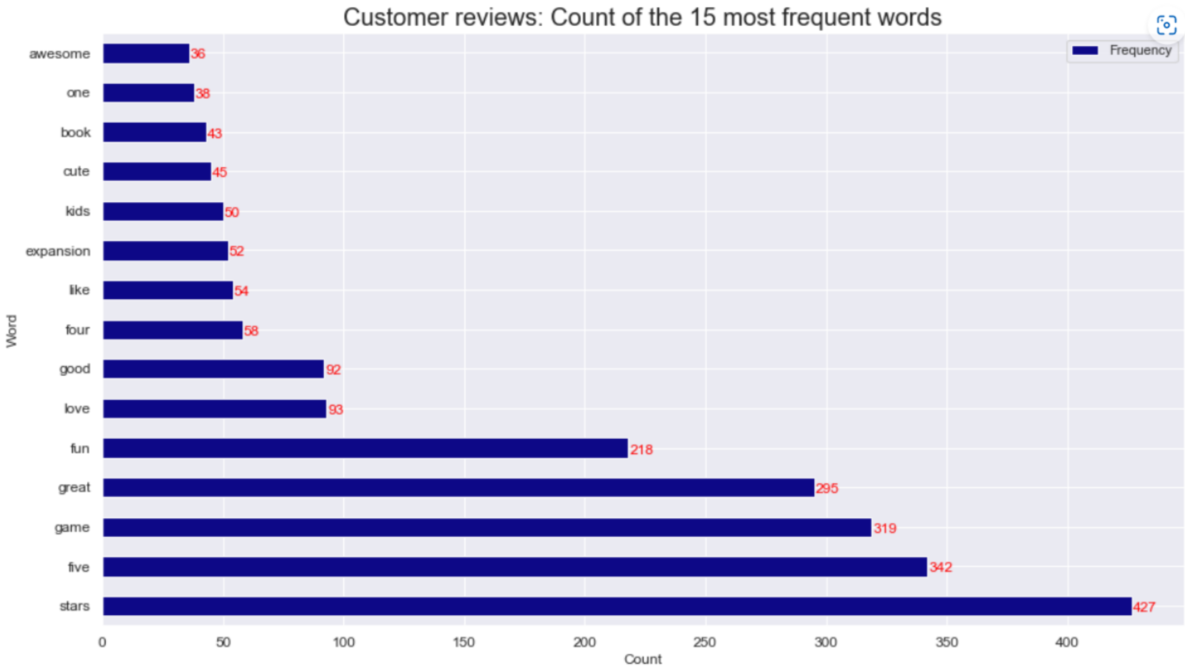
To derive insights from the customer reviews data, our focus would be this time on the 'review' and 'summary' columns. As these columns are formed with text format data, we would first need to pre-process it to get the appropriate structured data for the sake of our analysis. This pre-processing will include removing the punctuations, standardizing the words to lowercase, removing alphanumeric characters and stop words as well as tokenization.

Looking at word count by identifying the frequency of distribution and polarity, as a start of analysis:

- In the review column, the word ***'game'*** appears most frequently followed by ***'great'***, **'fun'**, etc. when we display the top 15 frequent words using the counter class function. Details can be found below.



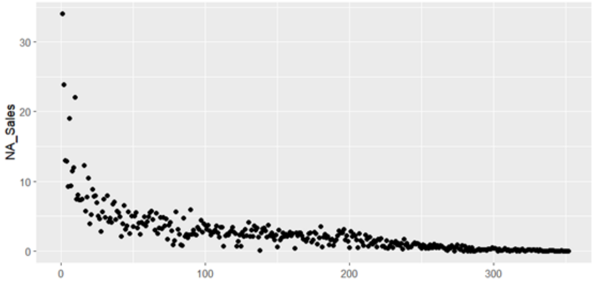
Likewise, when plotting the frequency distribution of the column summary, we could see a different order in the top 15 words where 'stars' comes first and 'game' at the third position, see the below visual.

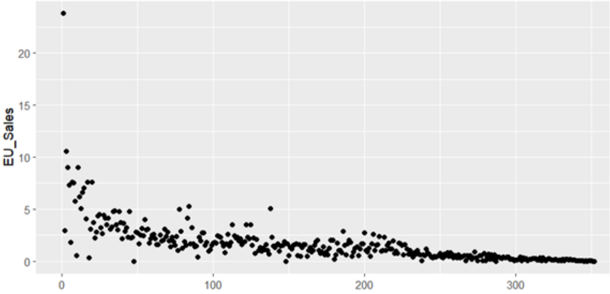


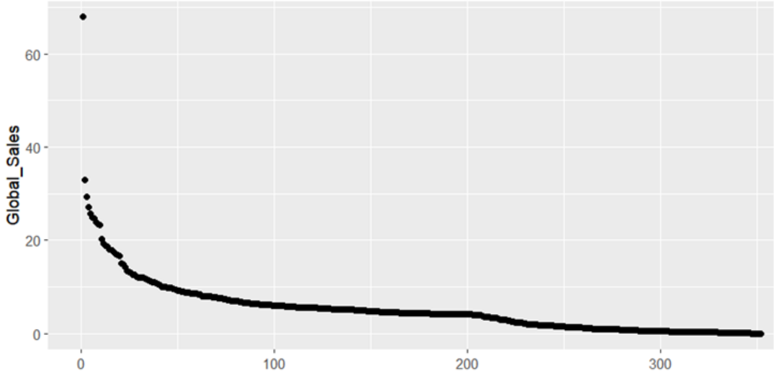
But finally using the polarity review, we could obtain more insights on the positive (top 20) and negative (top 20) comments to inform future marketing campaigns.

5. Visualise data to gather insights.

Using the scatter plot to view the distribution of the different sales data, we could clearly notice that the data points are clustered along the X-axis. There are outliers that we could spot visually though, a group of high sales outliers across the different regions and in the Global sales variables. See details in the below visual.

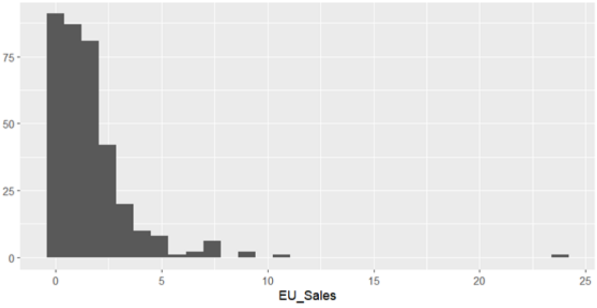


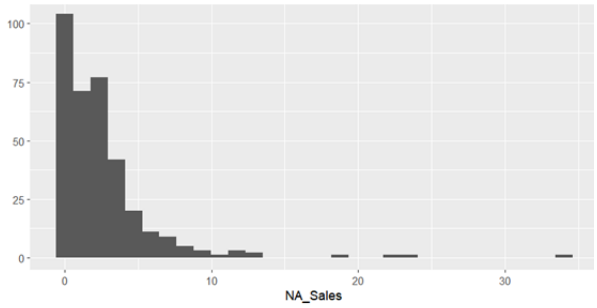


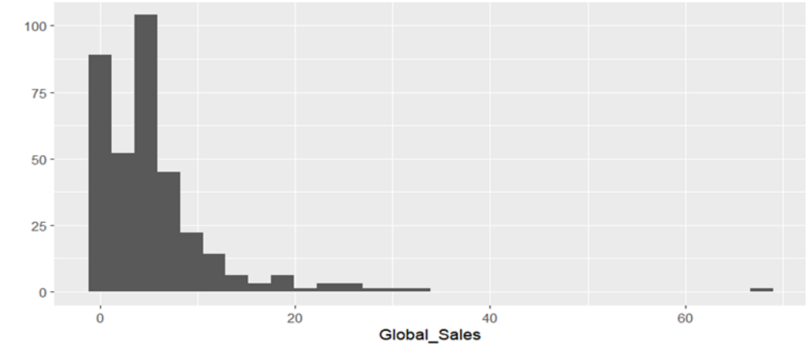


When visualizing the sales through histogram the following were spotted: the distribution of the data is very much skewed to the right, and the frequency of the sales data is lower than the left side due to fewer extreme values, and this is the case for all the sales across the different regions (EU, NA or Global).

The different histograms provide a better insight into where most of the data points are where in this case many have sales data close to the zero value.







6. Clean, manipulate and visualize the data.

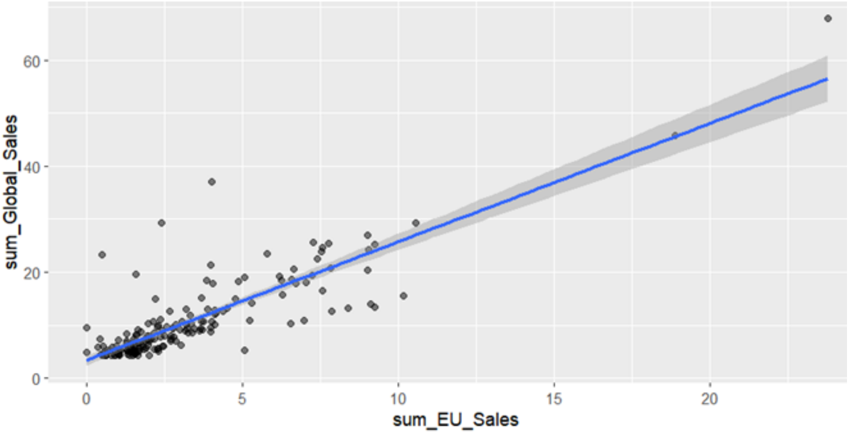
Exploring the shape of the distribution of the sales, using histogram and boxplot we could see, more clearly with the histogram though, that the sales data are extremely skewed to the right as the right tail is longer than the left one and this quite aligns with what the descriptive statistics showed: for Global sales for instance, 75% of the data have a sales value below ~ 13 units when the maximum value is up to ~ 68 units hence the presence of significant outliers in the dataset.

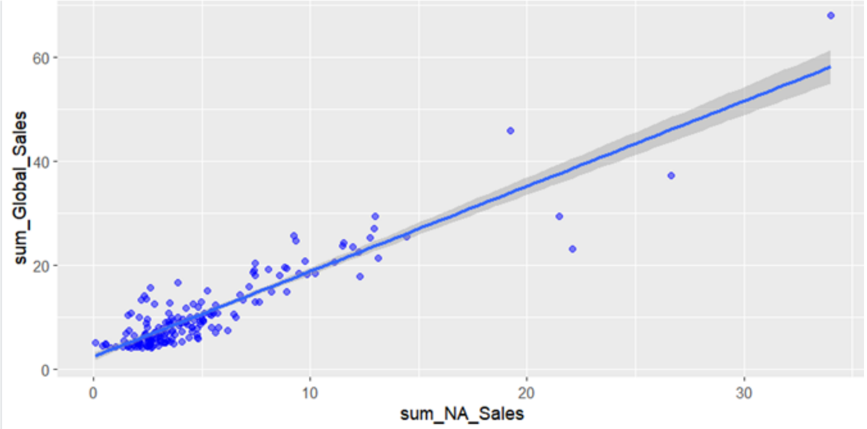
Plotting the quantile of the distribution of the different sales to compare with a normal distribution of the same, the following could be noticed visually: for all the sales data, the points were straight to the line until almost one standard deviation above the mean of the normal where from there the values tend to progressively further from the line until the extreme data points from +2 above the mean of the normal which suggests a heavier tail.

Also, using the hypothesis test such as the Shapiro. Wilk test, the very small p-value, suggests rejecting the null hypothesis (the distribution is normal).

As kurtosis is way above the normal 3 (excess kurtosis of 13 for EU sales for instance), which suggests that the distribution of the different sales has a heavy tail than the normal and signals the presence of extreme data points, that have a significant impact on the distribution of the dataset.

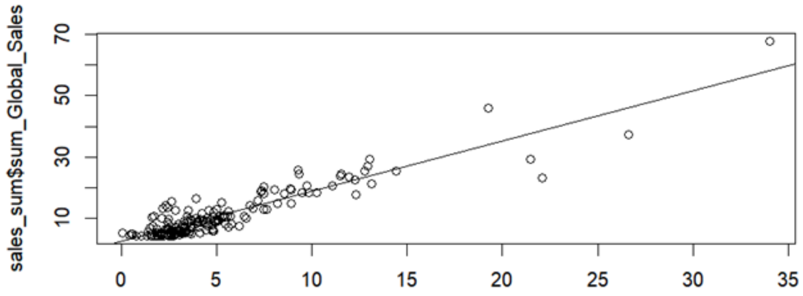
As for evaluating any potential relationship between the different sales data, the sales in the different regions (EU & NA) correlated with Global sales, at least visually at this stage.

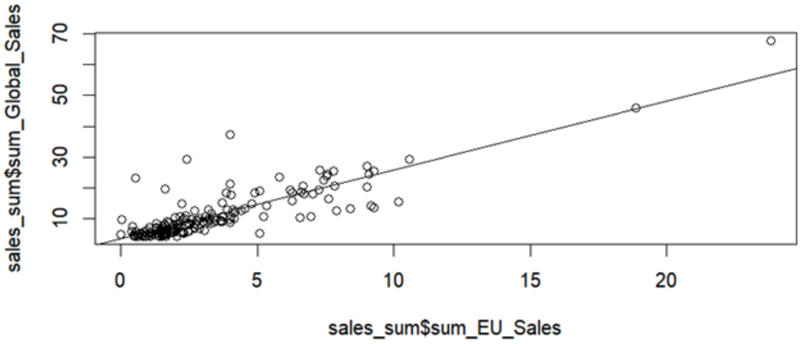




7. Making recommendations to the business.

Finally, speaking about the correlation between the sales columns, we could conclude that they are highly correlated: NA sales for instance, alone could explain the variability of the Global sales by ~ 84% and that for every unit of increase in North America sales, there will be an increase of 1.6 units in the Global sales. Also, the relationship between sales in the EU region with Global sales is somewhat linear which explains the best fit of the linear model of the two variables.





Furthermore, when looking at the multi-linear regression model between the different sales columns, there is no doubt that the model is strong with an R² of ~ 97% and the two explanatory variables for the Global sales are very significant to the model when combined: a change in sales of the different region very well explain the variation in Global sales.