Assignment Report

Background: This report is produced for Turtle Games. They manufacture and sell their own products, along with sourcing and selling products manufactured by other companies. Their product range includes books, board games, video games and toys. They have a global customer base and have a business objective of improving overall sales performance by utilising customer trends. This data analysis is carried out to understand and gather insight into the following business problems.

- how customers accumulate loyalty points?
- how useful are remuneration and spending scores data.
- can social data (e.g. customer reviews) be used in marketing campaigns
- what is the impact on sales per product?
- the reliability of the data (e.g. normal distribution, Skewness, Kurtosis)
- if there is any possible relationship(s) in sales between North America, Europe, and global sales.

Analytical approach: The software used in this data analytical process are Python and R.

The basic approach and goal has been to solve the business problems by applying various statistical methodologies in gathering insights, understanding patterns and understanding relations between different entities. I worked on two datasets(.csv files) for the purposes of data analysis. Firstly, important Python libraries were identified to carry out the procedure. This includes Pandas(for data analysis), Numpy, matplotlib, seaborn(for Visualisations) , statsmodels(for statistical libraries),scikitlearn, nltk, wordcloud, nltk.corpus etc.

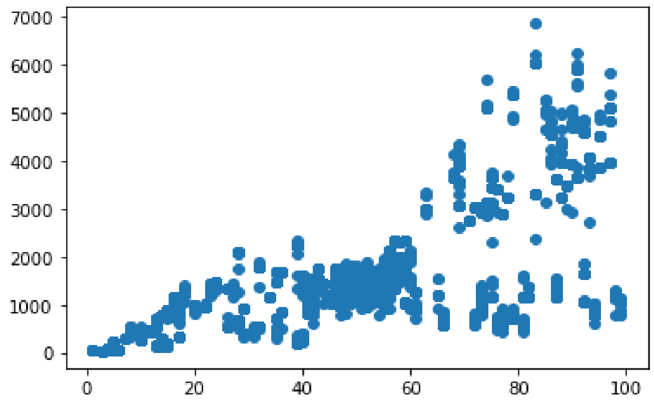
Data Ingestion: The datasets are loaded into Pandas and converted into dataframes. Dataframes have a large functions to support data wrangling, manipulations. The imported data is aimed to be cleaned by checking the following criteria: missing values, incorrect data formats, extremely skewed data, outliers. Any irrelevant columns are dropped from the dataframe. The column names are corrected if they aren't according to the best practices. Once the data is cleaned it is written to a new file and then re imported. This ensures that the work is carried out on the cleaned version. If required data wrangling process are carried out like merging the dataframes, creating the subsets from existing dataset, dropping duplicate values, extracting rows of data based on criteria etc.

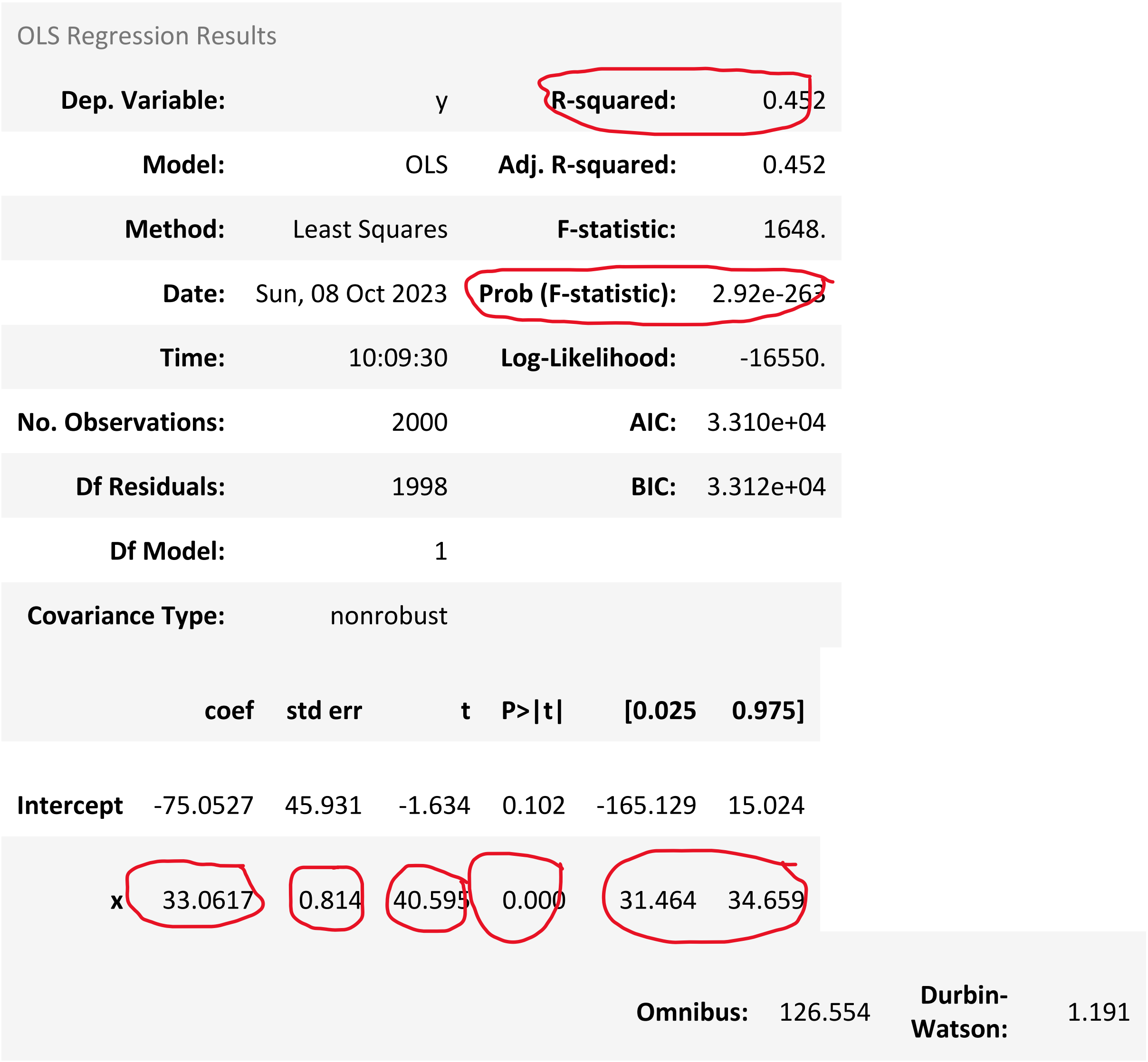
To address the first business problem of, finding out how customers accumulate loyalty points, the following process was undertaken.

The dataset was imported to Pandas dataframe. The data was sense checked and descriptive statistics were applied to get insight of the data. Two columns were dropped as they were not required for the analysis purposes. Two of the columns were also renamed appropriately. To get predictive insights into accessing how the customer would accumulate loyalty points, I thought to apply linear regession model. This is the most established statistical model in predictive analysis.

In order for me to apply linear regression, I had to correctly identify the independent variable and the dependent variable(s). As per the scenario, Loyalty points was decided to be the independent variable and Age, Remuneration and Spending Score were the dependent variables. A correlation between the dependent variables was done. There was slight positive correlation between spending score and remuneration but none with Age. The variables were then plotted to understand linearity using a scatter plot. Further, OLS model was applied to the data and a summary table was printed.

On this instance Spending score vs loyalty points were plotted.





OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.
Date:	Sun, 08 Oct 2023	**Prob (F-statistic):**	2.92e-263
Time:	10:09:30	**Log-Likelihood:**	-16550.
No. Observations:	2000	**AIC:**	3.310e+04
Df Residuals:	1998	**BIC:**	3.312e+04
Df Model:	1		
Covariance Type:	nonrobust		

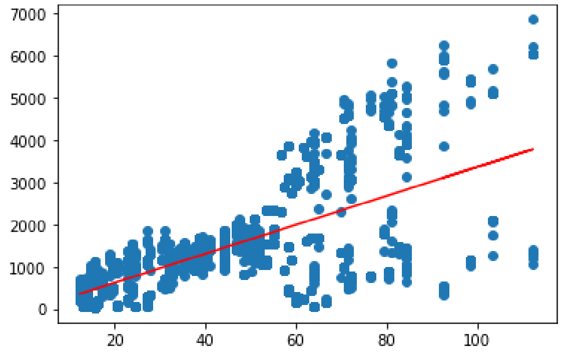
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-75.0527	45.931	-1.634	0.102	-165.129	15.024
x	33.0617	0.814	40.595	0.000	31.464	34.659

Omnibus:	126.554	**Durbin-Watson:**	1.191

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

Observations and interpretations: The data points are not strongly linear, as there are outliers. R^2 expalins that the variabiltiy of the model is 45%. The intercept being negative is insignificant, x coefficient is 33.0617, means that every 1 unit increases the predicted value would increase by 33.0617. F-stat values is below the value of 0.05, which suggests that the regression model is significant. t- value is calculated by coeff / std error. So when the std error is smaller the better. Here std error is smaller.Std error is also inversely proportional to t-value. The last 2 values are the confident intervals which suggests that 95% of chances are the values predicted wpould be between 31.4 & 34.65. The observation sample is also not too small(n= 2000).

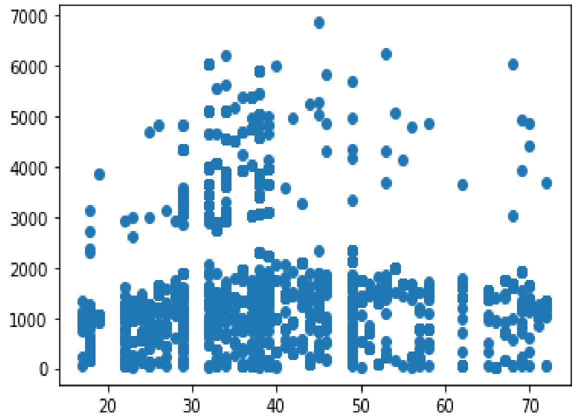
This model suggests that as the spending score increases it increases chances of accumulating loyalty points. Similarly, remuneration vs loyalty was plotted to check linearity.



The line of best fit or the line of regression is plotted in red. This model suggests that slightly more loyalty points could be accumulated by increases in remuneration. The third comparison model was Age vs Loyalty.

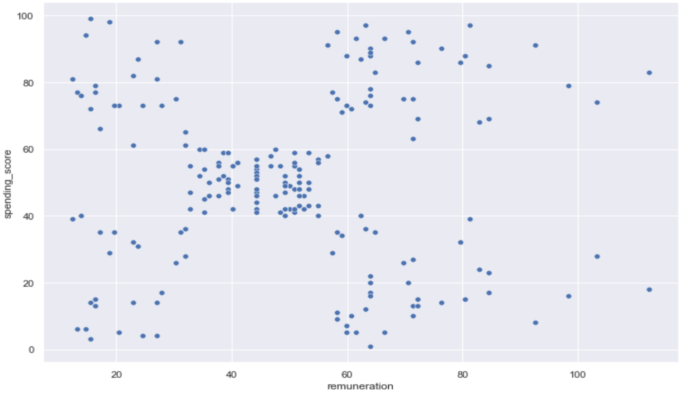
This model suggests no linearity and the regression tables suggests no statistical significance. So Age variable has been ruled out.

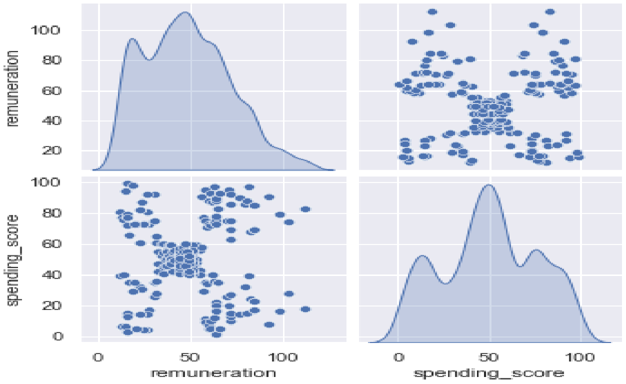
So to conclude in this observation Age has no impact in gaining loyalty points but spending scores has positive chances of gaining more loyalty points and remuneration has slight chance of gaining loyalty points This solves Turtle games first business problem.

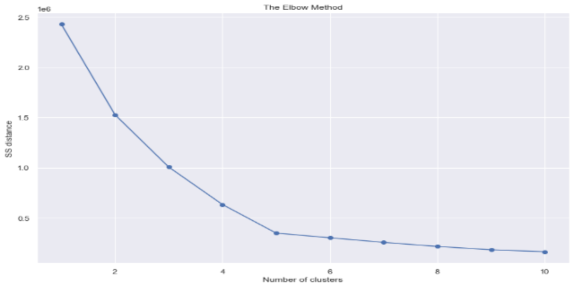


In order to address the second business problem to identify groups within the customer base that can be used to target specific market segments. I chose to apply the k means clustering model to identify the optimal number of clusters and then apply and plot the data using the created segments.

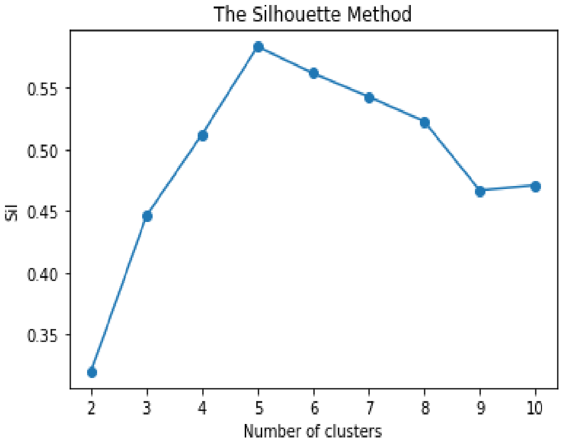
First up, the data is loaded to Pandas dataframe, the data sense check is carried out. Unnecessary columns are dropped from the dataset and descriptive statistics is done to assess the dataset. A scatter plot is plotted with spending score and remuneration as variables.





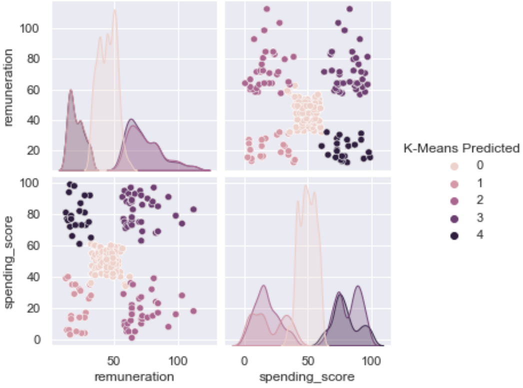


A pair plot is also plotted. Then two standard methods are applied to identify the number of k means clusters. They are Silhouette method and Elbow method. After implementing the Elbow method, we get the following graph:

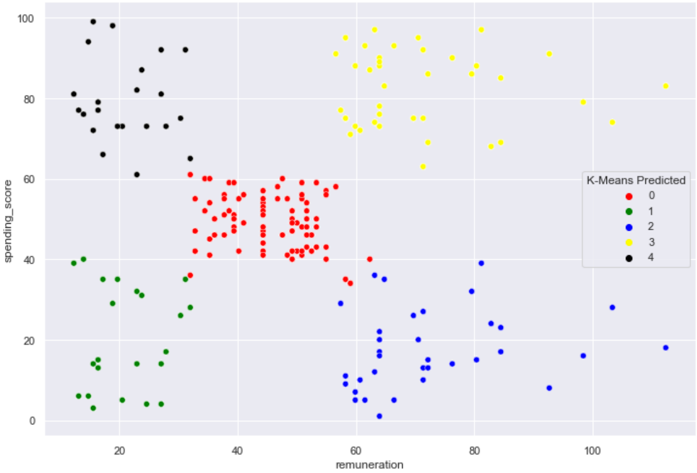


In Elbow method, the idea is to use logical interpretation as to what number to pick. The point where there is the kink, could be a suggested option. In Silhouette method, the point closer to 1 is better but the data shows the most available point as close to .60 which corresponds to 5. I am applying the method therefore as K means for 4,5,6.

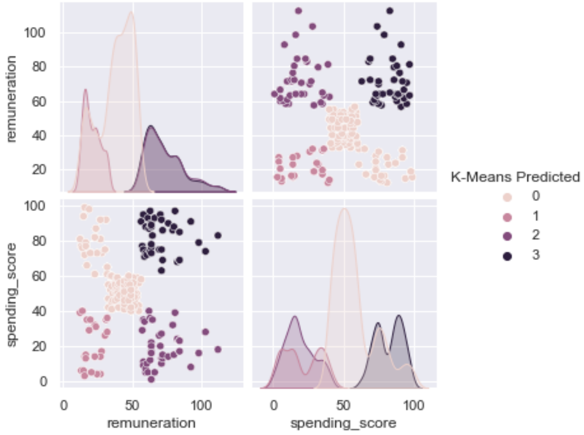
Evaluating for k=5, this pairplot is plotted, which distinctly have three different clusters with 0 having more observations. Then the predicted values in the clusters are generated. If we visualise this cluster as this graph.

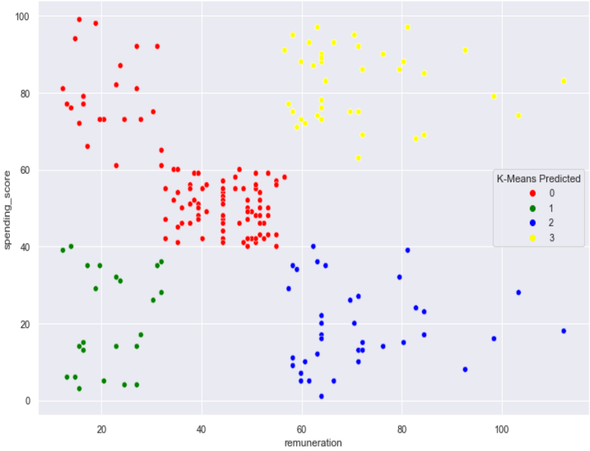


To address the third business problem of identifying,



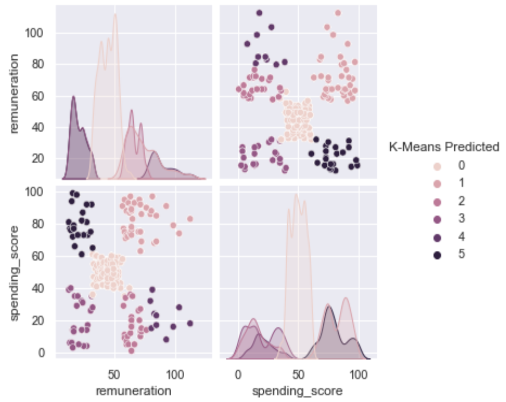
This image clearly shows the red clusters is more cantered from the sum of squares. Again, this process is repeated with k=4 and the following as the graphs.

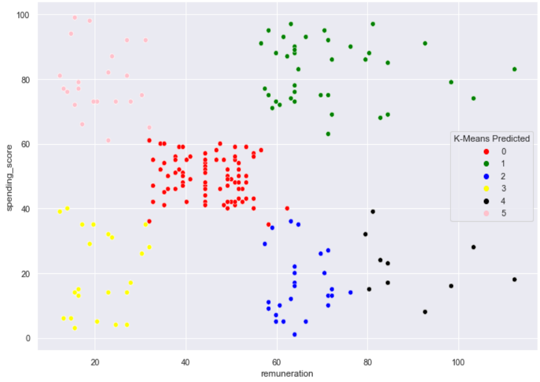




Here there is clear distinction of cluster 0 having more data points, red cluster in the predicted value clusters.

Finally, the process is iterated for k=6, with 6 clusters, to find the suitable clusters.





I chose to

I choose the ideal number of k means to be 5 cluster and 0 group has the largest objects gathered with a very close distance from the sum of squares.

To address the third business problem, can social data (e.g. customer reviews) be used in marketing campaigns, I aimed to produce 15 most commonly used words for online reviews and 20 positive and negative reviews from the website. To proceed with this requirement, I thought to approach the Natural Language Processing (NLP) and sentiment analysis methods. To proceed, I have installed the required libraries like NLTK, wordcloud, FreqDist,TextBlob, nltk.tokenize etc.

The dataset is loaded and required sense check and cleaning is carried out as before. The data is then removed of punctuations, converted to lower cases, duplicates are dropped. The words are tokenised to create a group of words from sentences and cfreate word clouds (graphical representation of words). A list of English stop words are also created. Then tokenised listed of words are checked in the stop words and removed. Then the frequency distribution is generated to produce the most common frequent words.

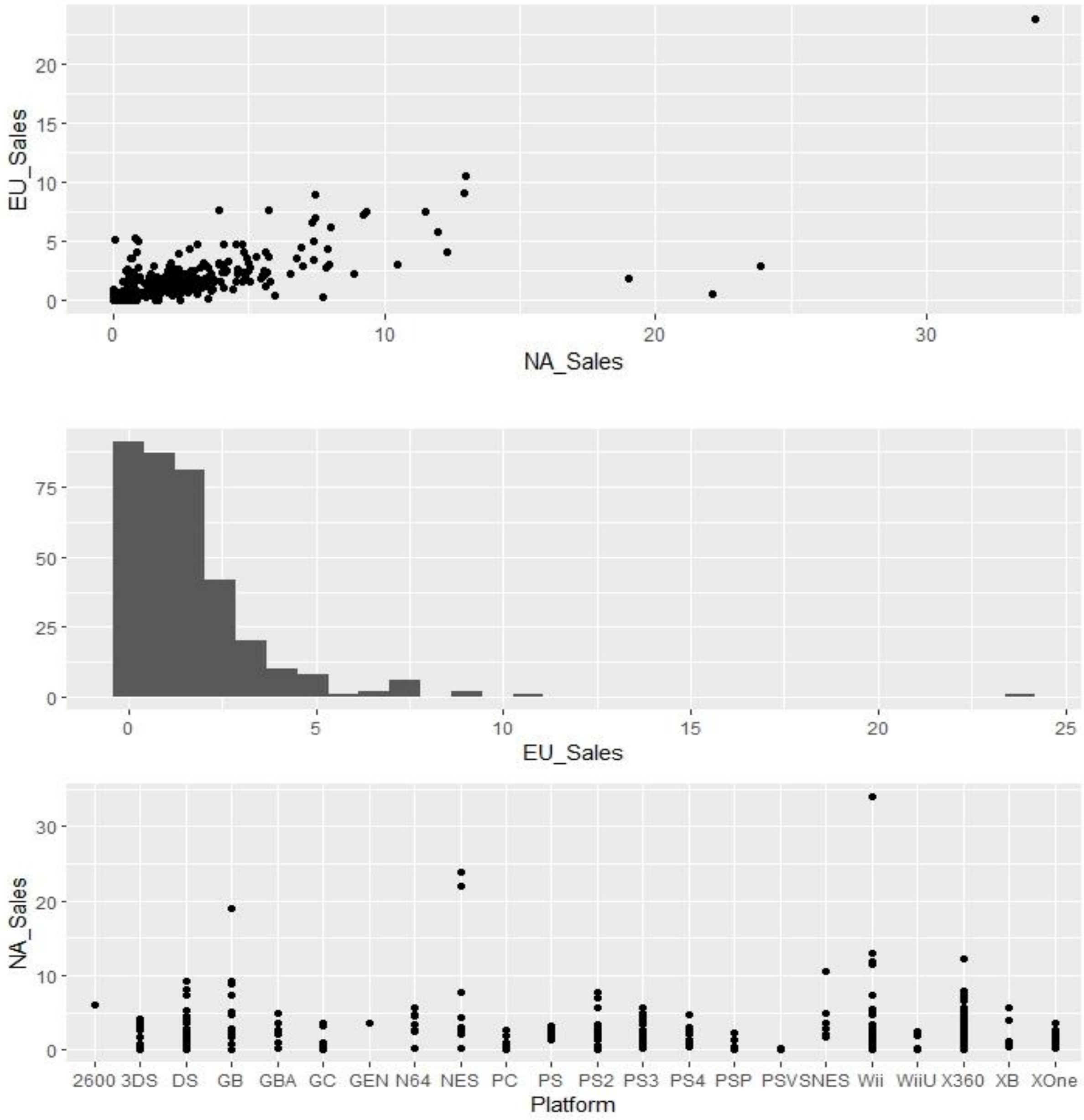
The next set olf business problems as per below:

what is the impact on sales per product?

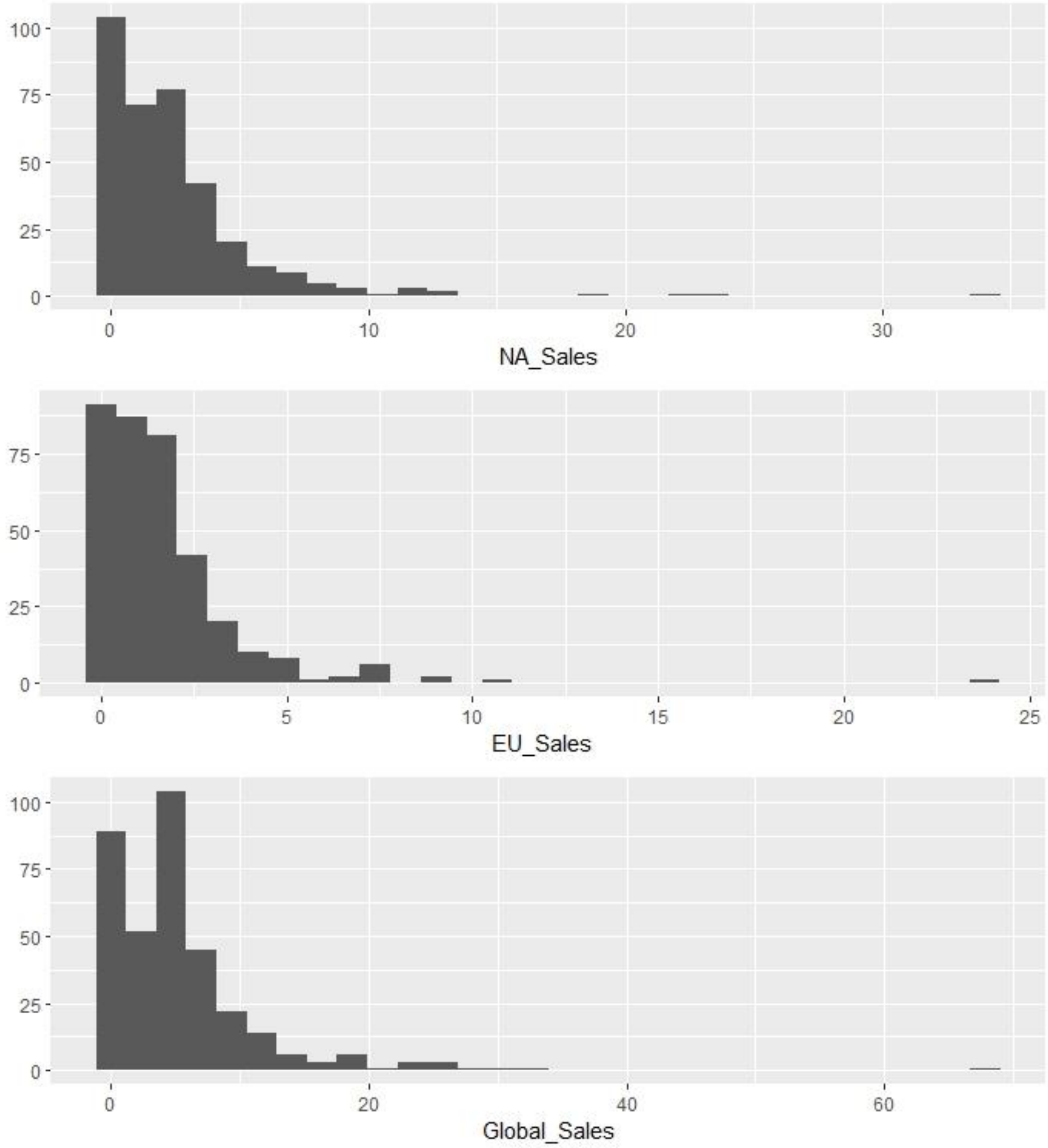
- the reliability of the data (e.g. normal distribution, Skewness, Kurtosis) ,

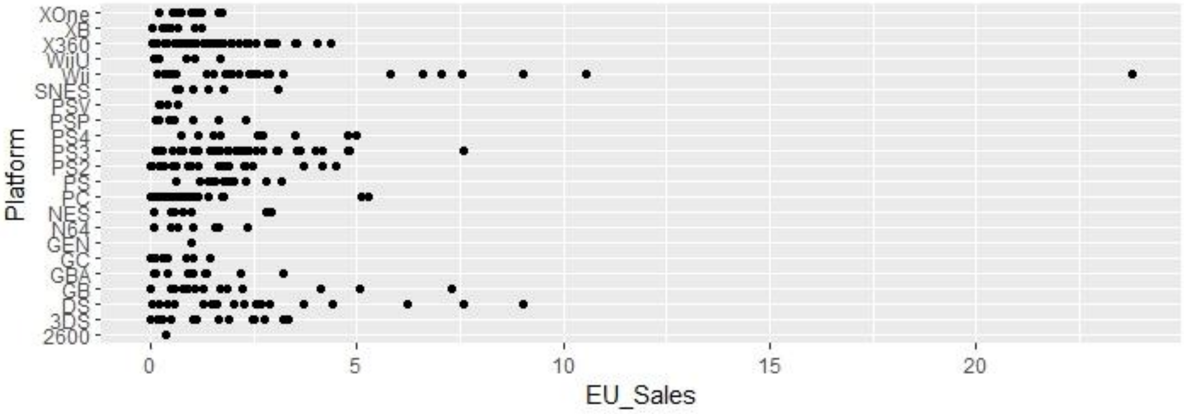
  I had carried out the following procedures in R.

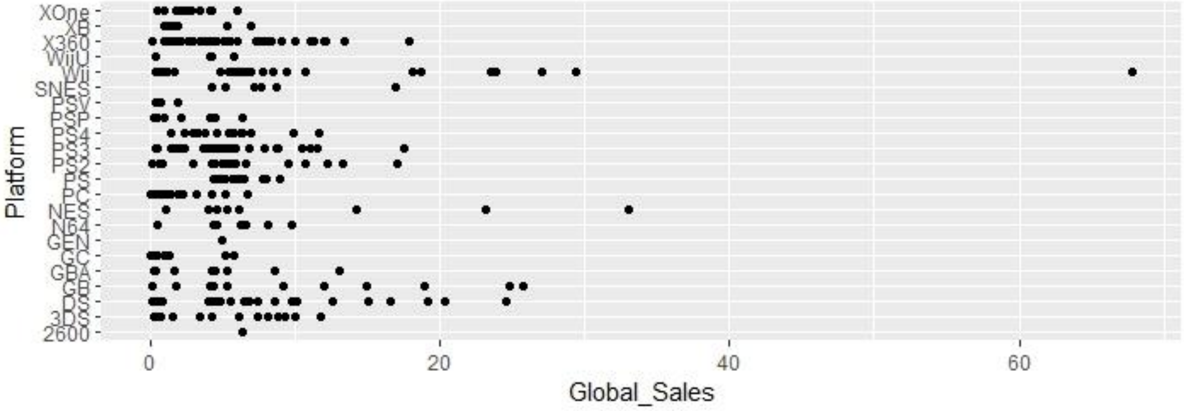
  The dataset has been loaded to R. The data sense check is completed. The necessary libraries are imported. Descriptive statistics are applied to understand the data. The data is also plotted to understand the data better. Mainly scatter plots and boxplots are used for this.

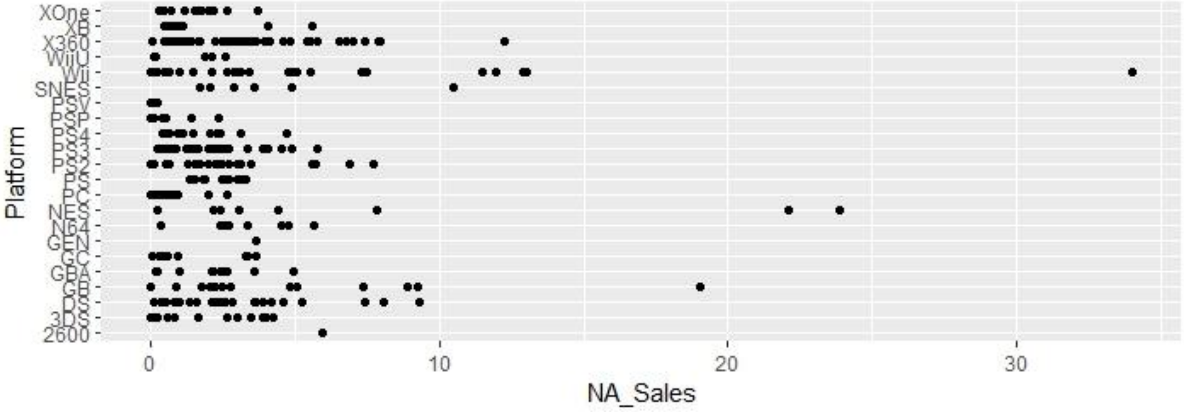


EU_Sales
20
15
10
5
0
0
10
20
30
NA_Sales
75
50
25
0
0
5
10
15
20
25
EU_Sales
NA_Sales
30
20
10
0
2600
3DS
DS
GB
GBA
GC
GEN
N64
NES
PC
PS
PS2
PS3
PS4
PSP
PSV
SNES
Wii
WiiU
X360
XB
XOne
Platform


100
75
50
25
0
0
10
20
30
NA_Sales
75
50
25
0
0
5
10
15
20
25
EU_Sales
100
75
50
25
0
0
20
40
60
Global_Sales





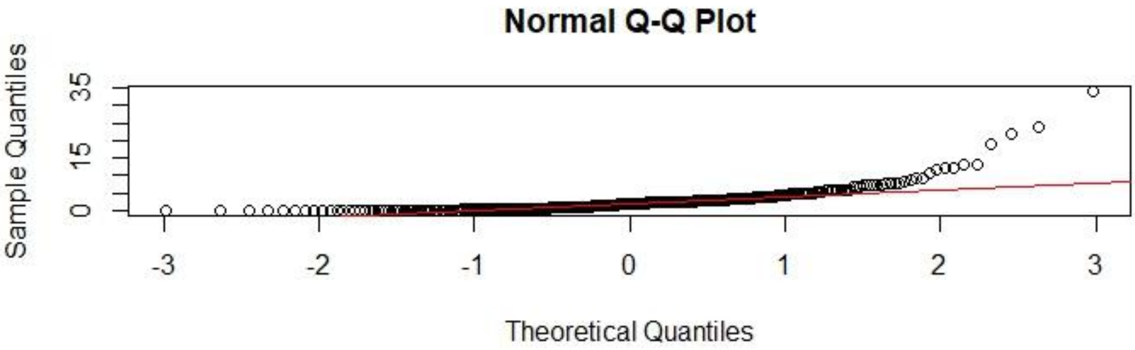


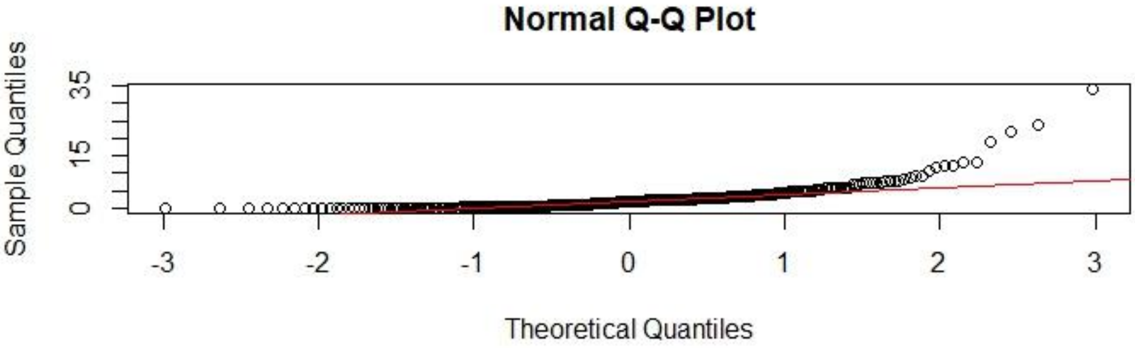
These plots shows the distribution of sales data across NA area, EU and global. A new dataframe is created by grouping the product and calculating the sales by NA, EU and globally.

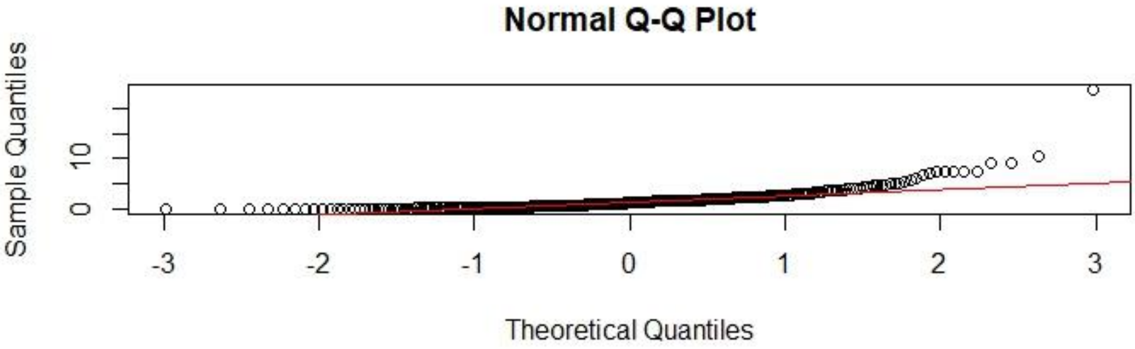
A range of statistical methods are employed to check the normal distribution of the data

As follows:

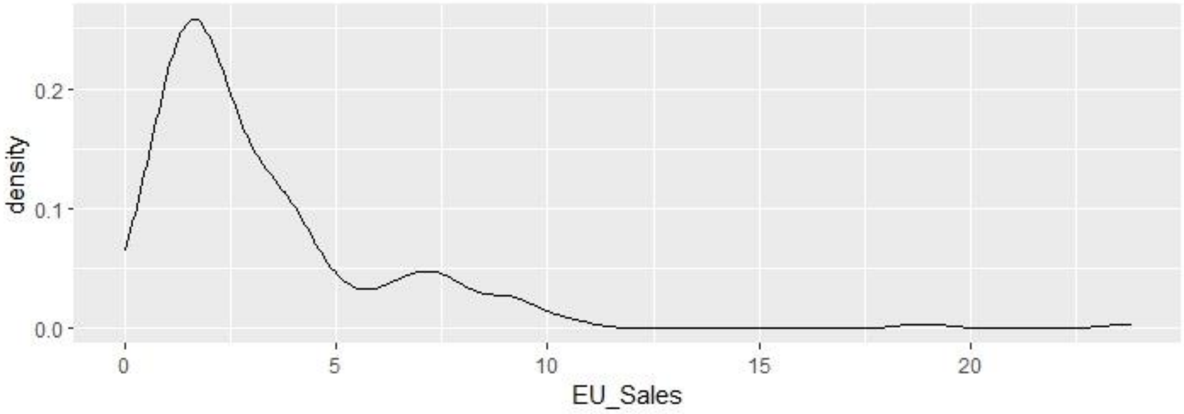
- Q-Q plot
- Shapiro-wilk test
- Skewness
- Kurtosis
- Correlation.

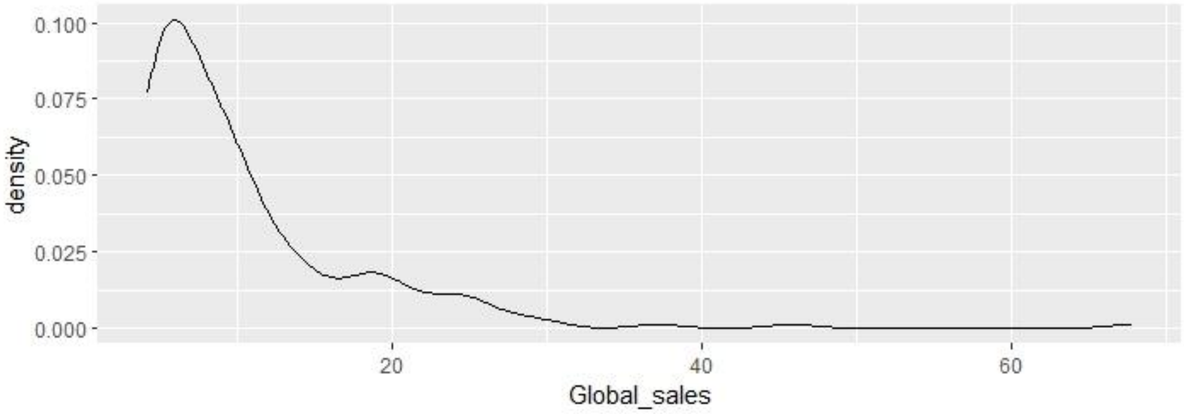


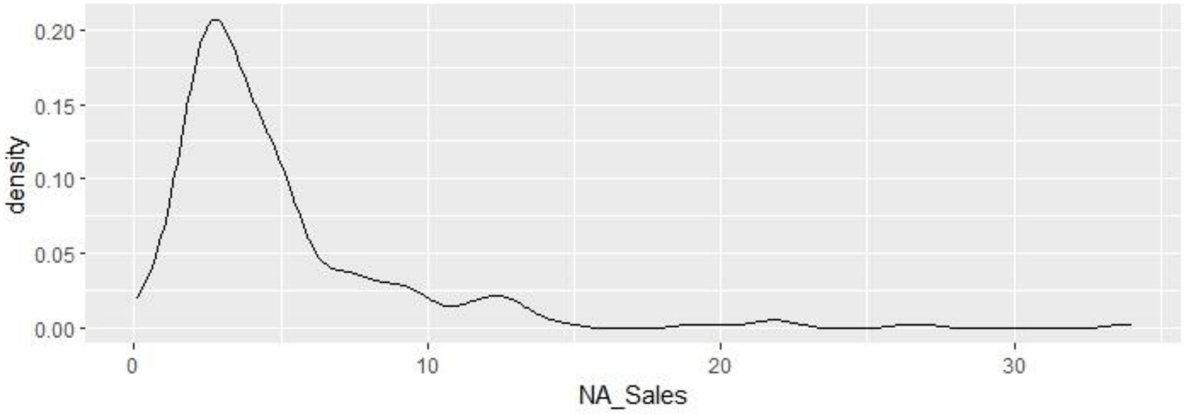




The statistical analysis data shows the dataset is skewed, with a lot of outliers. The correlation was positive among all the sales data. Then a set of GGplots- like density plots and scatter plots were plotted.

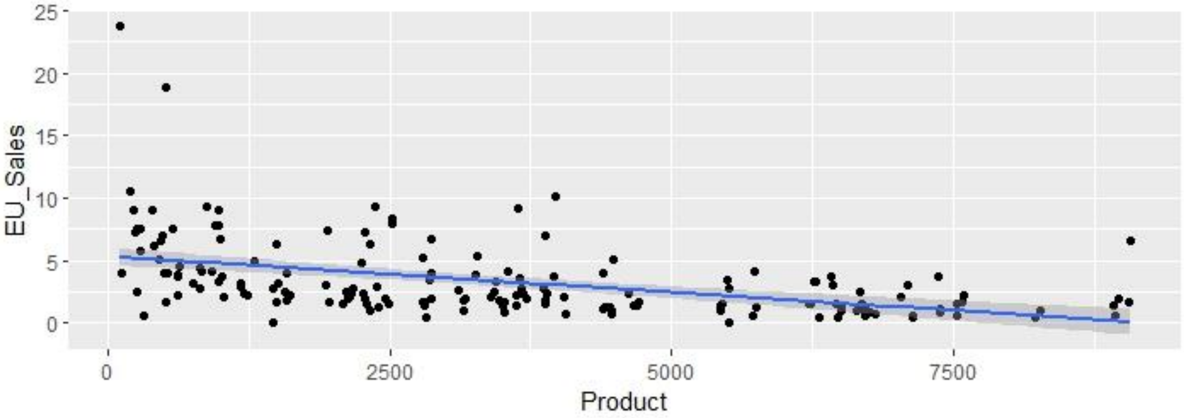


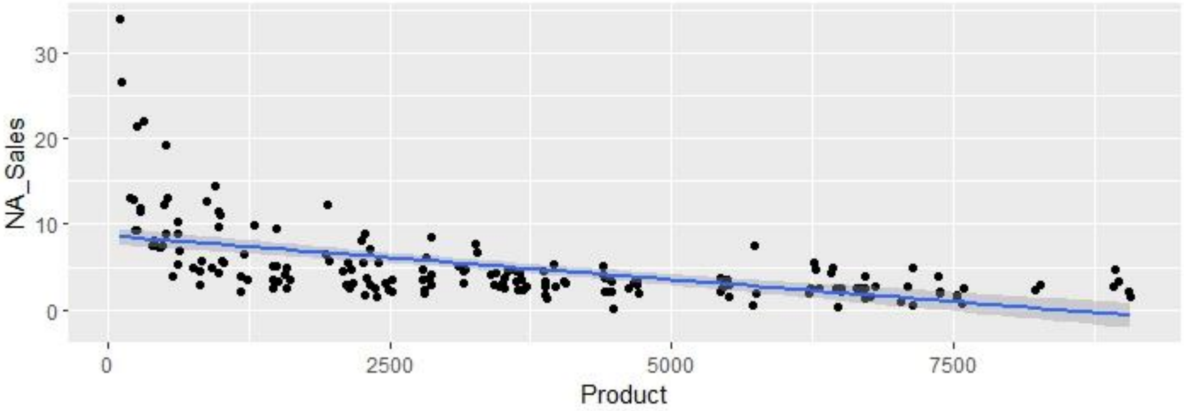


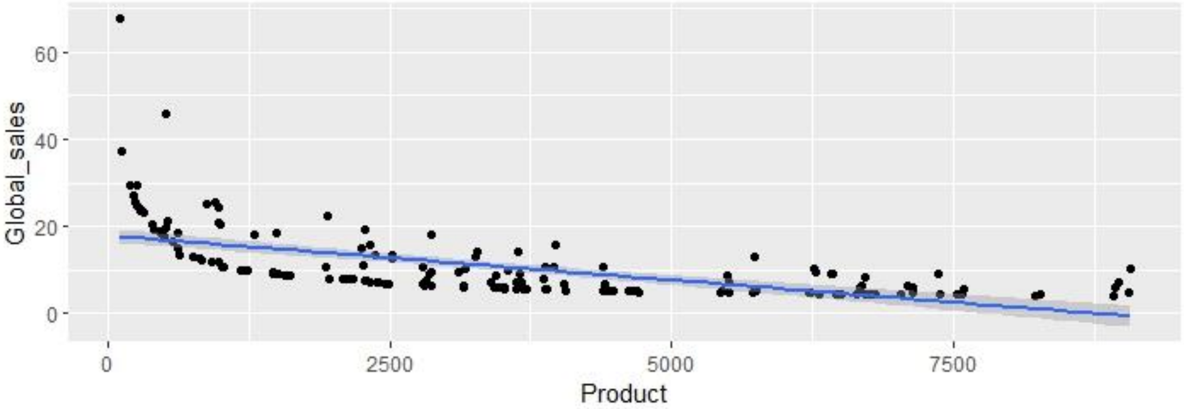


This shows the sales data distribution on GGPLOT-density plot.

Finally, I chose to plot the sales data to see the impact on each product on GGPLOT scatter plot and have inserted a trend line or the line of best fit. I have chosen the the scatter plot for this was because it plots all the data points clearly giving us the indication of trend as these graphs:







This procedure determines the sales data across three regions of NA, EU and Global. The procedure was correct identifying about the reliability of as mentioned was skewed with varying degree, which also means the data was not normally distributed. The analysis also found out about the positive correlations among the sales data.

To conclude , I have been able to address most of the business problems identified by Turtle Games.