Course-3

Assignment

Assignment Cover Sheet

Submitted by: Atif Habib Syed

Date Submitted: 12-September-2022

Module Title: Course-3

Date/Year of Module: 2022

Word Count:

Number of Pages:

Question: *You've been hired by Turtle Games to help analyse the available data and to share insights with stakeholders.*

Declaration: *"I declare that this work is entirely my own in accordance with the academia's guidelines on plagiarism and collusion. All external references and sources are clearly acknowledged and identified within the contents.*

No substantial part(s) of the work submitted here has also been submitted by me in other assessments for accredited courses of study, and I acknowledge that if this has been done it may result in me being reported for self-plagiarism and an appropriate reduction in marks may be made when marking this piece of work."

Table of Contents

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. 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.

This report/ assignment is focused on analyzing Turtle games' available data and explain how that analyses was conducted.

The Data

1. **turtle_reviews.csv** includes details on customer gender, age, remuneration, spending_score, loyalty_points, education, language, platform, review and summary across products.
2. **turtle_sales.csv** includes details of video games sold globally, such as the rank, product, platform, genre, publisher, and their sales across North America, Europe, and worldwide.

Data Analysis Tools

We will be using PYTHON and R as data analysis tools.

Report Structure

This report is structured around each week's assignment as we will be discussing the analysis conducted during that week and the insights that were discovered during the analysis.

Report Structure will be as following:

1. Making predictions with Regression.
2. Making predictions with Clustering.
3. Analyzing customer sentiments with reviews.
4. Visualizing data to gather insights.
5. Cleaning, manipulating and visualizing data.
6. Making recommendations to the business.

1.0 Making predictions with regression

The requirement in this part of the assignment, required to investigate possible relationships between loyalty points and age, renumeration and spending scores. In order to estimate the strength and character of the relationship between Loyalty Points (dependent variable) and Spending scores, Renumeration and Age (Independent variables).

Regression Model: 'Spending Score' vs 'Loyalty Points'

First relationship that was analyzed was between 'Spending Score' and 'Loyalty Points'. That was to determine if a significant relationship exists between both variables. Also to observe, if 'Loyalty Points' accumulation can be predicted through a customer's 'Spending Score'.

The results are as following:

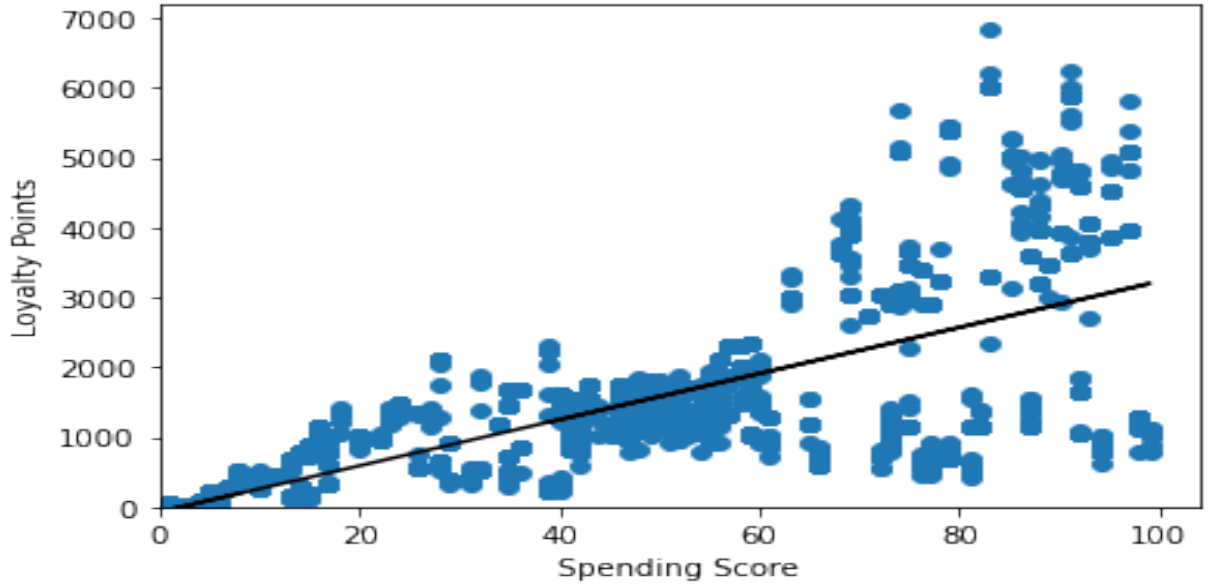
Dep. Variable:	y	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.

	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

S

Here, important parameter to note is R-Squared, which is a measure of the strength of the relationship between or regression model and dependent variable. In other words, this means how well a relationship exists between independent and dependent variable and whether dependent variable can be predicted using independent variable.

In case of 'Spending score' vs 'Loyalty Points', we can see that R-Squared is around 0.452, which shows moderate correlation (relationship b/w variables) according to Henseler (2009). This means that predictions made through 'Spending Scores' for 'Loyalty Points' will be accurate around 45% of the time. This can also be explained through a following visualization:



According to the visualization, there is significant correlation between 'Loyalty Points' and 'Spending Score' when 'Spending Score' is from 0 – 20 and somewhat between 40 – 60. For rest of the values/ Data points, regression line does not fit the model.

In conclusion, customer spending can be a moderate predictor of customer's loyalty point accumulation.

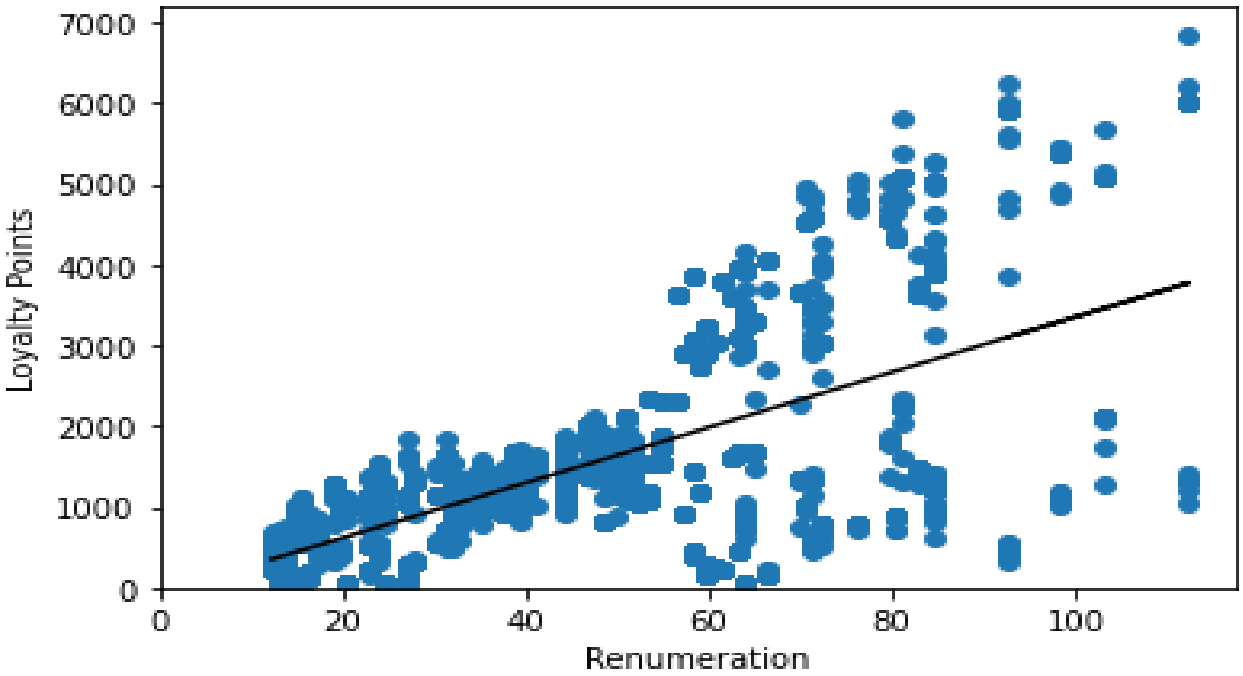
Regression Model: 'Renumeration' vs 'Loyalty Points'

In 'Renumeration' vs 'Loyalty points', following are the results of the test:

Dep. Variable:	y_renum_loyal	**R-squared:**	0.380
Model:	OLS	**Adj. R-squared:**	0.379
Method:	Least Squares	**F-statistic:**	1222.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-65.6865	52.171	-1.259	0.208	-168.001	36.628
X_renum_loyal	34.1878	0.978	34.960	0.000	32.270	36.106

Here we can observe that R-Squared value of 0.380 depicts a weak correlation (Henseler (2009)) between variables. Visualization/ Regression line of this model is as following:



Here, it is also observable that the correlation between total annual income between £15k and £58k and loyalty points is strong, however, beyond this range correlation is not significant at all.

In conclusion, a customer's annual income is not a good predictor of their loyalty points accumulation, however, there is a room for further analysis where a model can be built for customers with an annual salary between the range of £15k and £58k and then train it accordingly for loyalty points predictions.

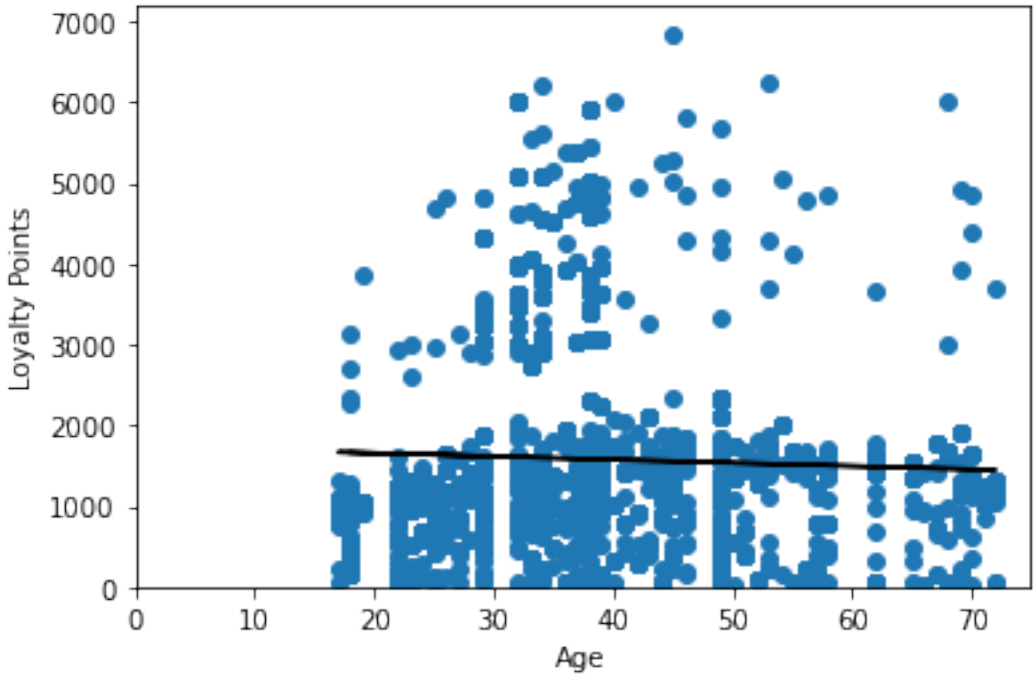
Regression Model: 'Age' vs 'Loyalty Points'

In order to understand if there is a significant statistical relationship between 'Age' and 'Loyalty Points', we conducted regression test with following results:

Dep. Variable:	y_age_loyal	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.606

R-squared value speaks for itself, there is no correlation between 'Age' and 'Loyalty Points'.

Visualization will be able to strengthen R-squared perspective:



It can be seen that there is no relationship between ‘loyalty Points’ and ‘Age’.

In conclusion, Age of a customer is not a predictor of their loyalty points accumulation.

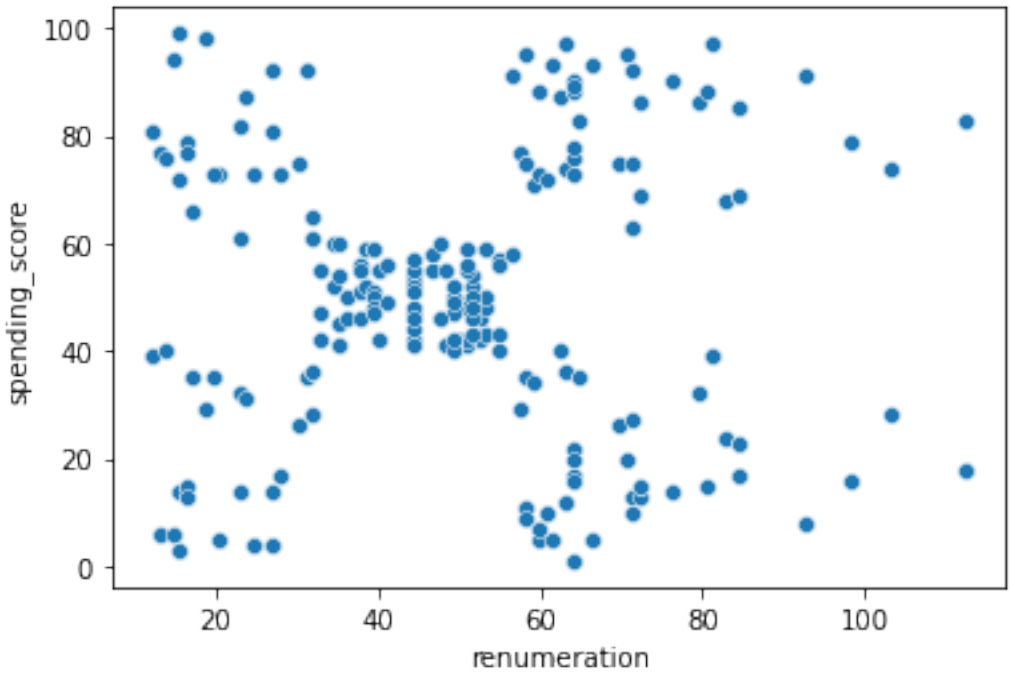
Although out of the scope of this project, I would like to further analyze relationships between loyalty points and gender, education and product/platform type.

2.0 Making predictions with Clustering

Clustering is a method of unsupervised learning in order to gain insights into data by observing what groups the data points fall into after clustering algorithm has been applied.

In this part of the assignment, we will analyse customer’s annual income i.e ‘renumeration’ and ‘spending score’ to observe correlation and identify customer groups.

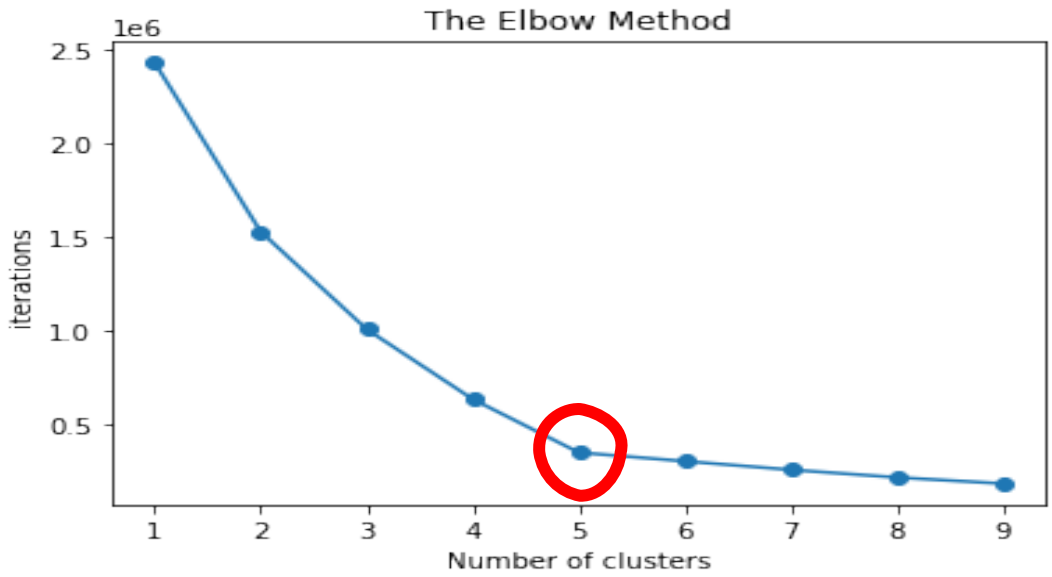
First we need to visualize realtionship between ‘Spending_Score’ and ‘Renumeration’ by plotting a scatter plot:

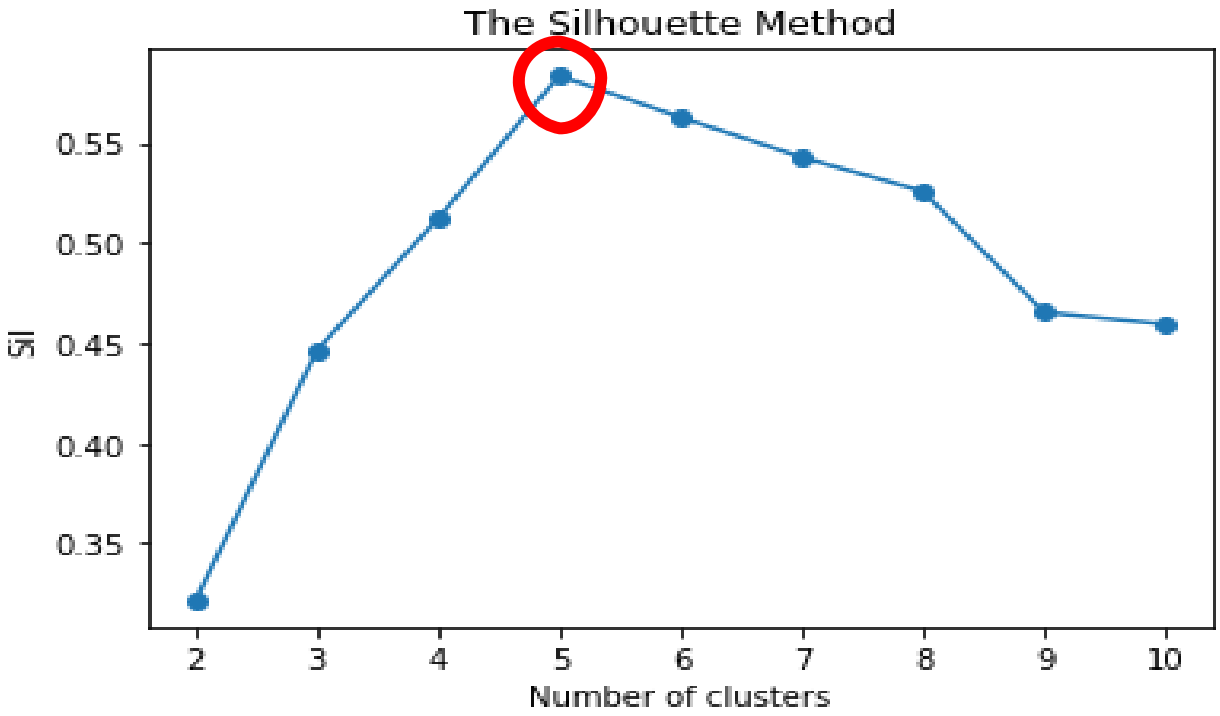


Scatter plot shows no apparent correlation between two variables. However, in this case we will apply Clustering algorithm to identify groups within these variables and observe any correlation.

To identify groups, we first need to determine the number of clusters. To determine that, we will use ELBOW method and to cross-check whether ELBOW method's recommended clusters are appropriate, we will also deploy SILHOUTTE method.

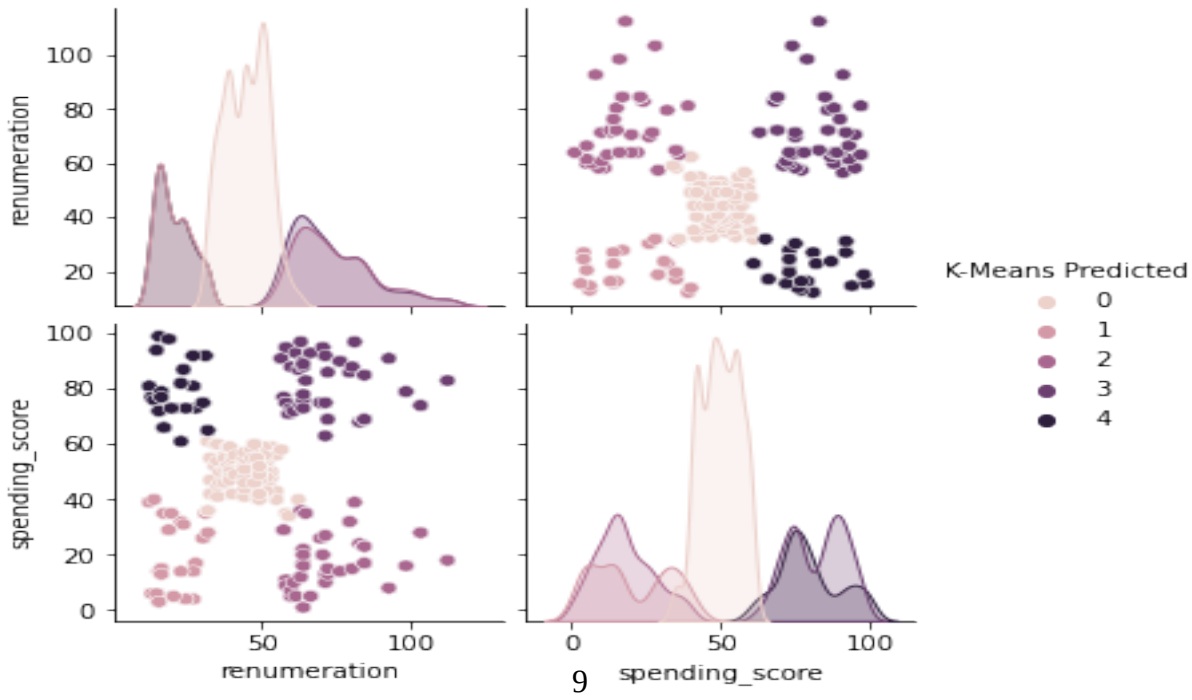
Visualization of ELBOW method and Silhouette methods are below:

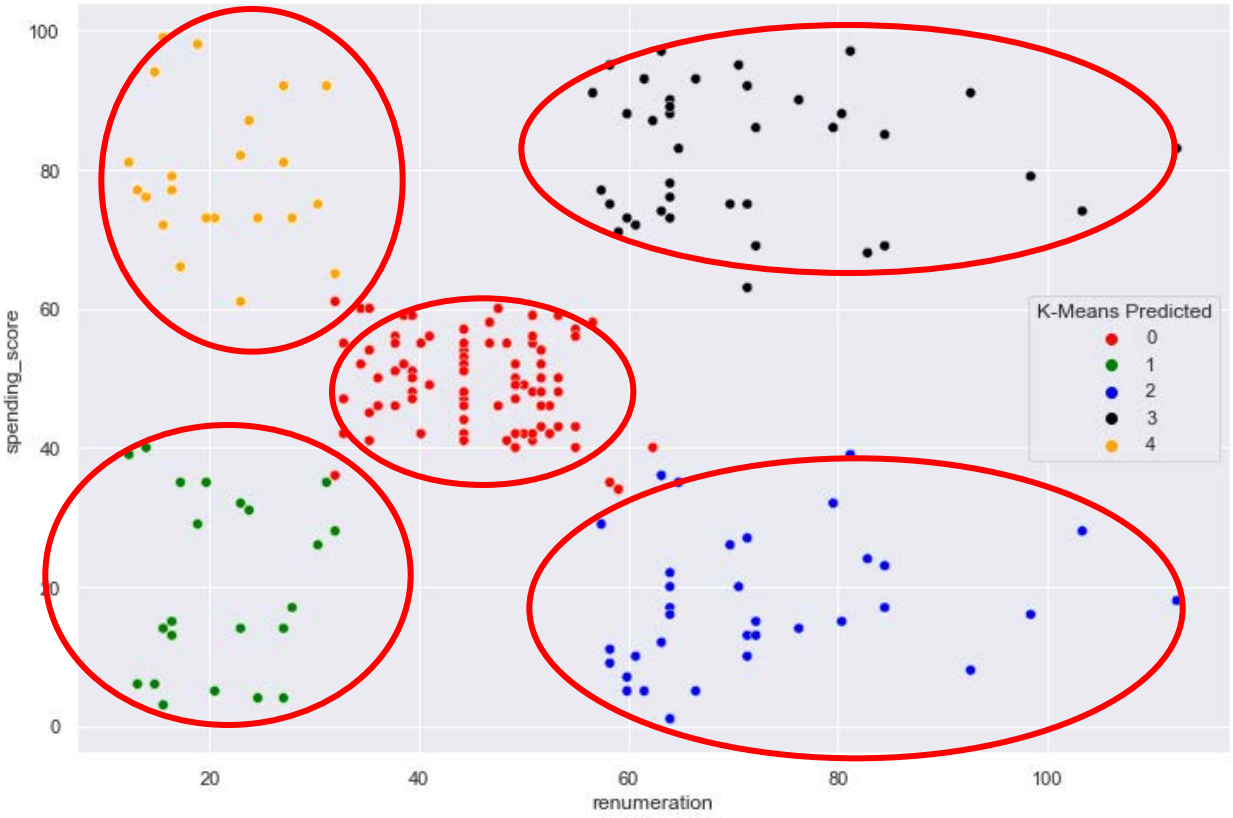




Both methods indicate that number of clusters that can be generate from the provided data are 5 (Five).

Now, that we identified the number of clusters, we applied KMEANS clustering algorithm on the data to generate groups. Visualization of the model is as following:





This shows an interesting insight into the data. Customers groups can be identified as following:

Group-0
Annual Income: 30k to 60k
Spend Score: 40 - 60

Group-1
Annual Income: 10k to 30k
Spend Score: 1 - 40

Group-2
Annual Income: 60k to >100k
Spend Score: 0 - 40

Group-3
Annual Income: 60k to 100k
Spend Score: 60 – 100

Group-4
Annual Income: 10k to 30k
Spend Score: 60 – 100

The interesting part is that Group-4 and Group-1 has similar annual income i.e. 10k – 30k, however, their spending scores are in different quadrants. Same is the case with Group-3 and Group-2 with annual income in the range of 60k - >100k, but different spending scores. This leaves room for further analysis such as knowing customer's age, education, profession etc to understand which type of customer has high spending scores and how to address their needs and how to covert customers with low spending score to higher spending ones.

3.0 Analyzing customer sentiments with reviews

Sentiment analysis is today's world is becoming essential for a business to understand customer opinion and strategize accordingly. The time consumers take to ask questions, resolve issues, and share both positive and negative experiences can be used to help an organization evolve.

We will be using Python's NLTK and textblob libraries.

In our case, we used sentiment analysis techniques to gauge customer sentiment and their views.

We started off by using a method called TOKENSATION. Tokenization basically breaks sentences into words. These words can then be analyzed using multitude of techniques. We removed all stop words such as 'I, its, myself, what, which etc' to analyze the sentiment considering words that convey customer's sentiments. Important point to note here is that sentiment analysis was conducted on two columns:

1. Review: Online reviews submitted by customers who purchased and used the products
2. Summary: Summary of the customer's review.

A Word cloud is a visual representation of text data which makes it easier to visualize words according to their size and color. Word Cloud that was generated from both columns are as following:

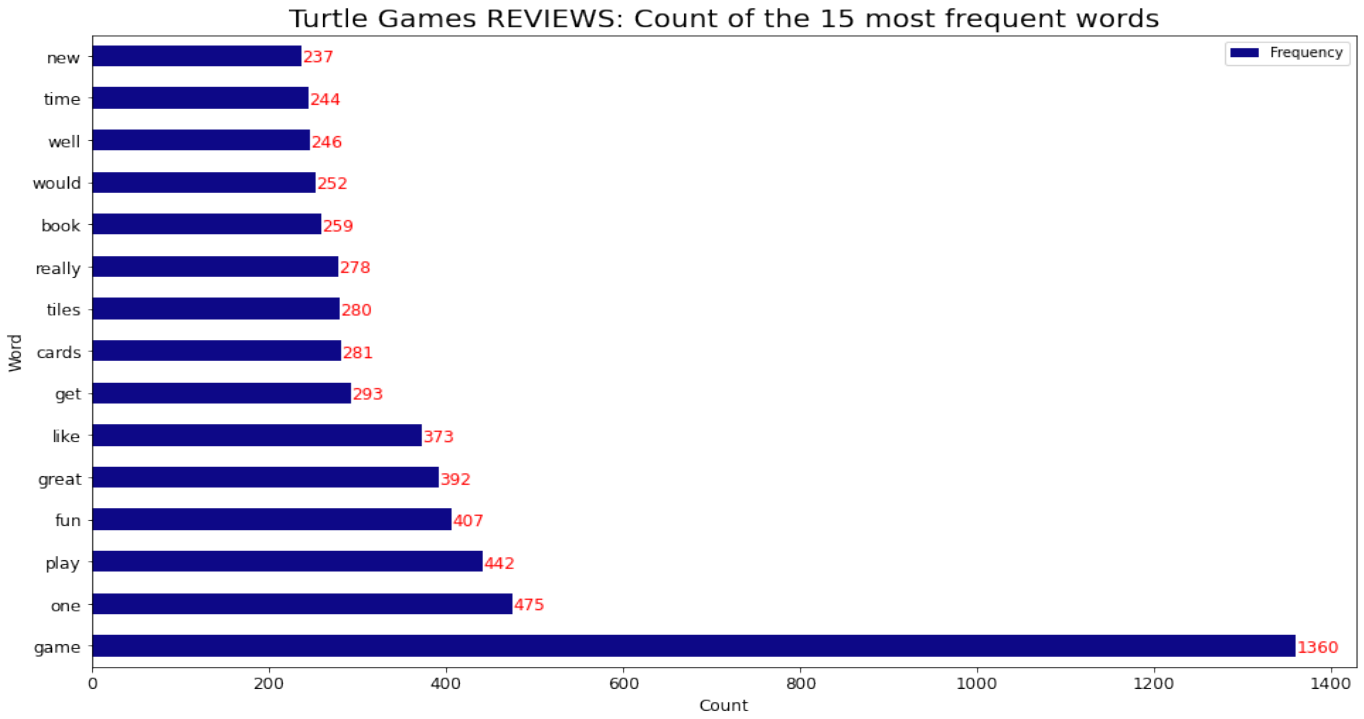
Word cloud from REVIEW column

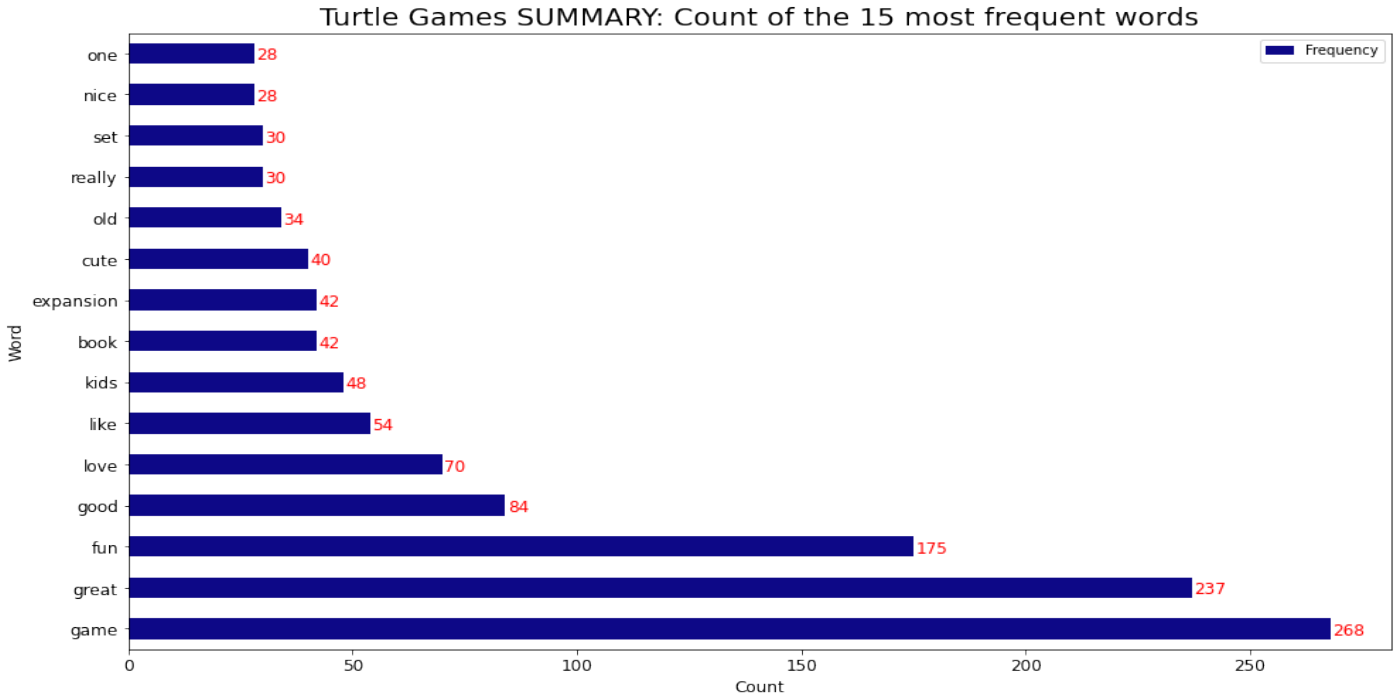
Word cloud from SUMMARY column

Another way of looking at word data is to visualize most frequently occurring first 15 common words:

Course-3

Assignment

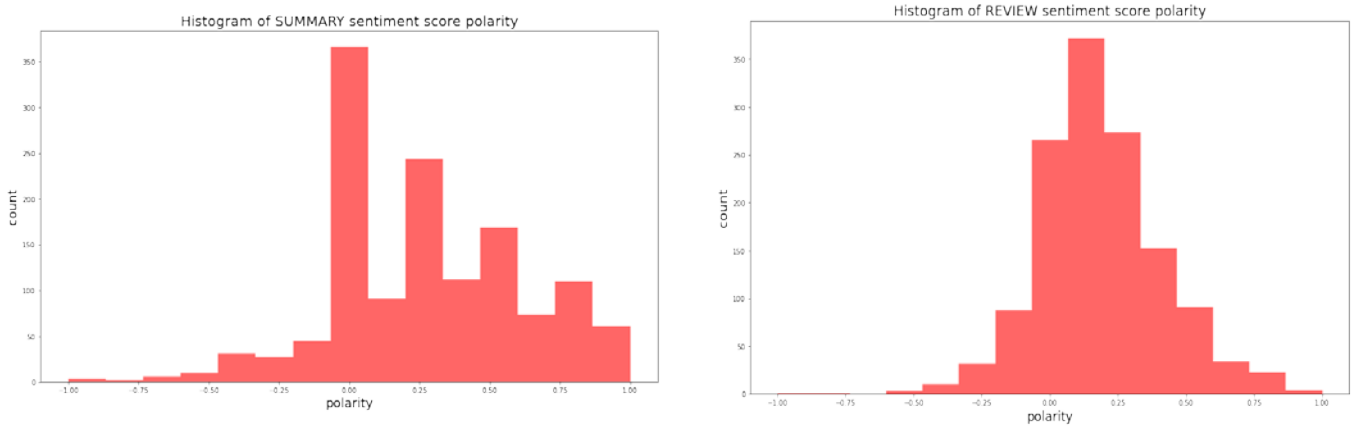




It appears that most frequently used word is GAME which is not a surprise as we are analyzing a game company, however, looking at rest of the words it seems that a positive tone is most commonly used in the both columns.

In order to know the exact sentiment of the sentences in reviews and summary columns we are going to deploy Polarity analysis model which show the exact sentiment of the reviews.

Visualizing the polarity of both columns provides us with sentiments of customers:



Histograms shows that sentiments in review and summary columns are neutral or positively skewed.

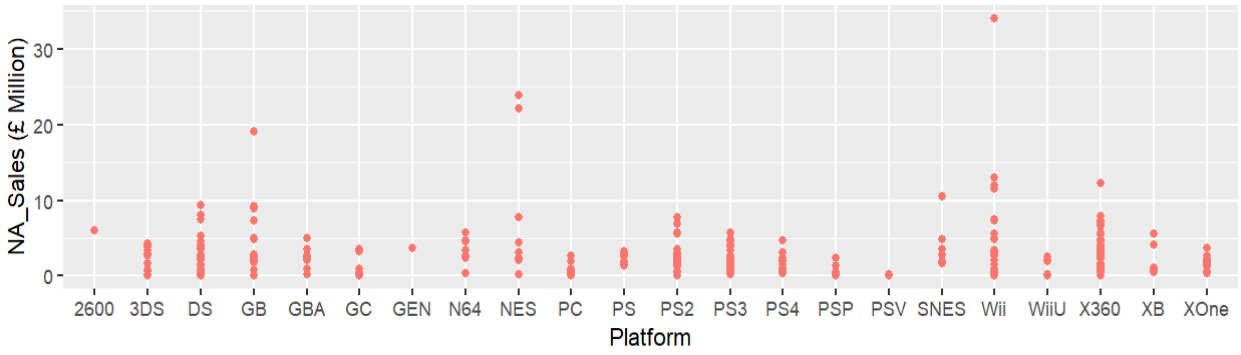
Conclusion: The sentiment analysis was conducted on Turtle games' website's reviews. In order to gain further insight, Turtle games should also collect reviews and statements from Social Media and Review websites like Trustpilot etc.

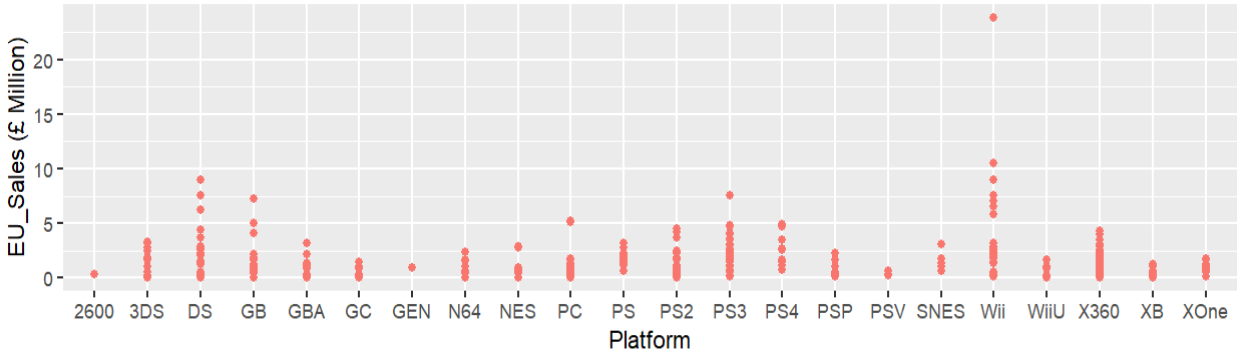
4.0 Visualise data to gather insights

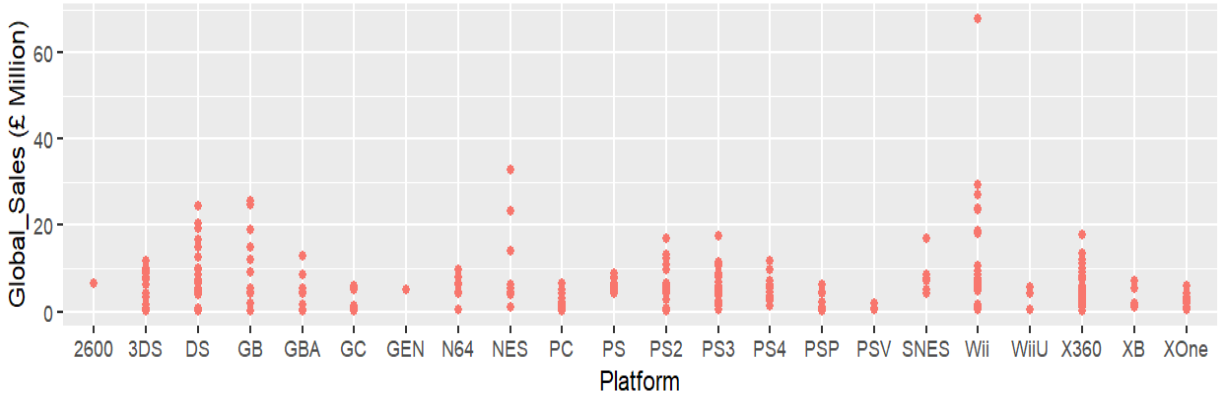
In this section of our assignment, we will turn our focus to Sales data of Turtle Games. This data is focused on video games that are sold for gaming consoles such as Xbox 360, Nintendo Wii, DS lite etc.

We will be exploring and visualizing sales data using 'R' programming language.

Visualizing sales for North America, European Union and Globally according to the gaming platforms:





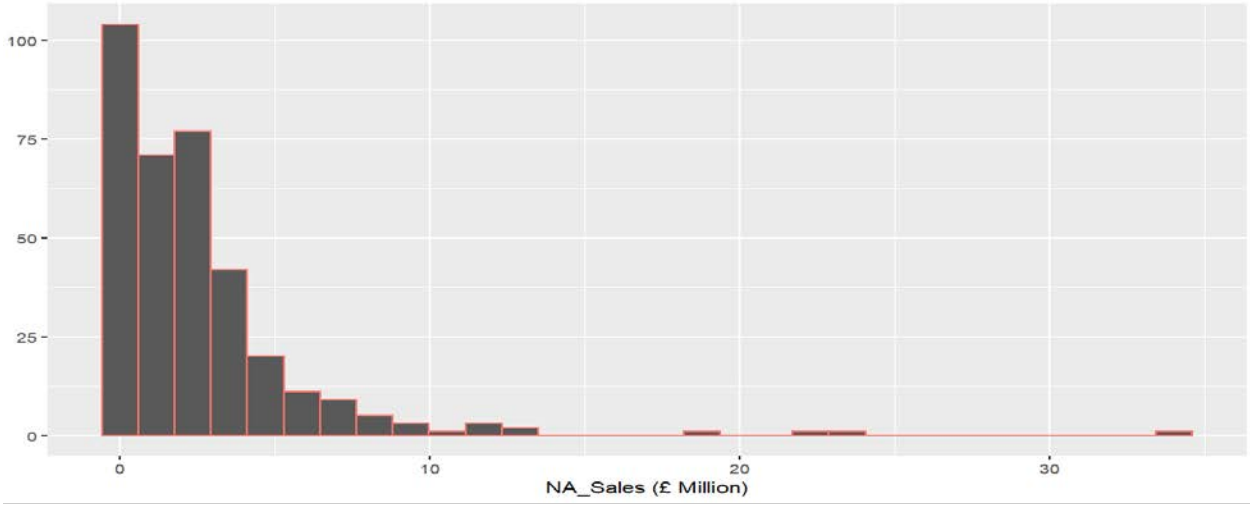


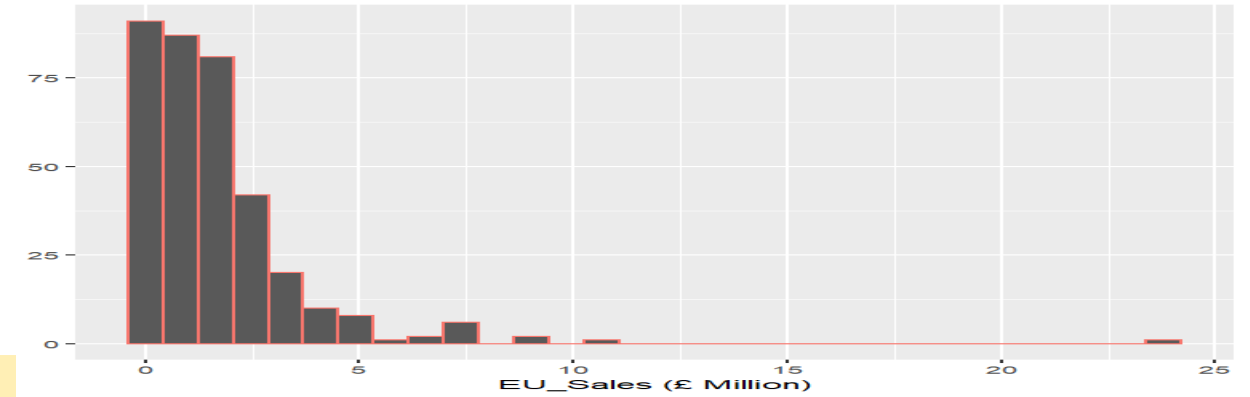
It can be observed that for some products on several platforms (Wii, Xbox 360, NES & GB) sales in North America is above £10 million as compared to European Union where £10 million mark is crossed by couple of games on Wii platform. Global sales on the other hand shows that majority of platforms cross £10 million mark.

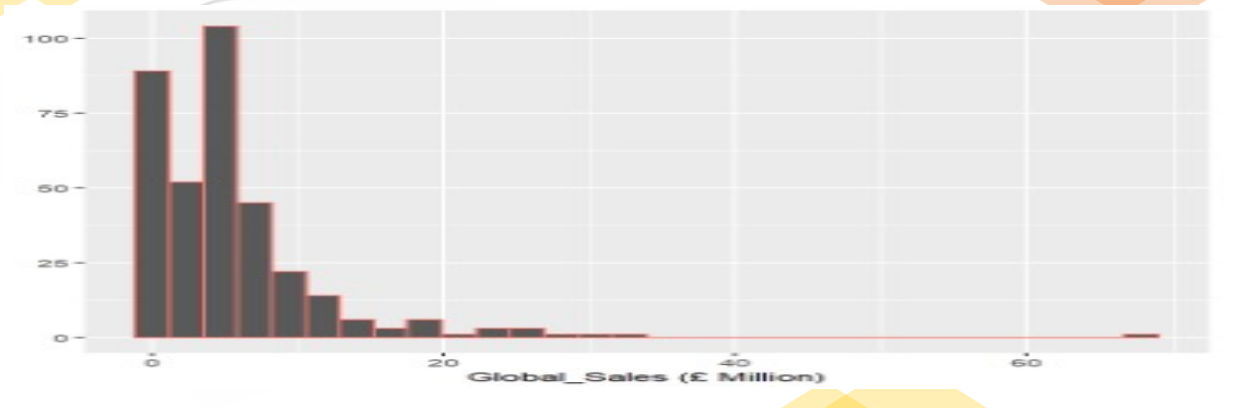
Another way of looking at the sales data is to see the count of products and revenue they generated. This was achieved through visualizing sales data through Histograms of each region:

Course-3

Assignment







We can observe that most of the games in all regions have generated a revenue between £1 million and £20 million mark. However, there is small number of products which have also crossed £60 million mark in North America and Globally.

Conclusion: Turtle games' sales in North America and Globally have been generating more revenue as compared to European Union. Further analysis may require some additional analysis such as Sales in each country etc.

5.0 Clean, Manipulate and Visualise the data

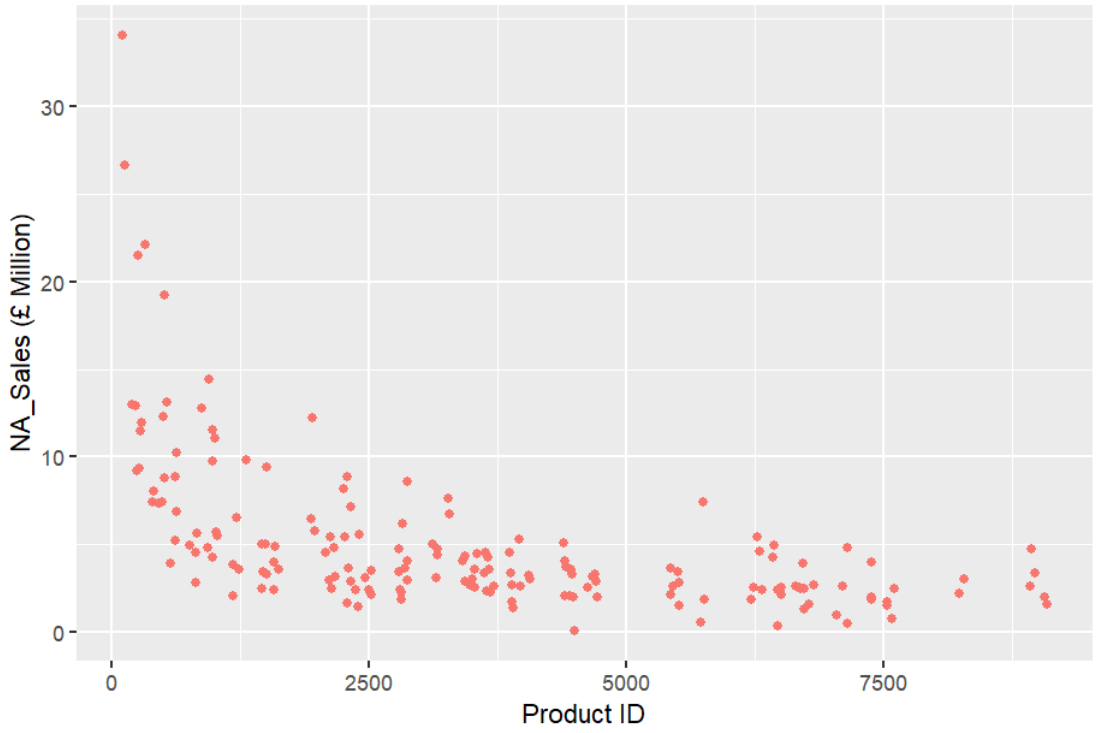
In this part of the assignment, we explore and prepares the sales data set and compare regional sales for analysis. We combined sales according to the product idea and analyzed which product did well in respective regions. In addition, central tendencies of each region were calculated.

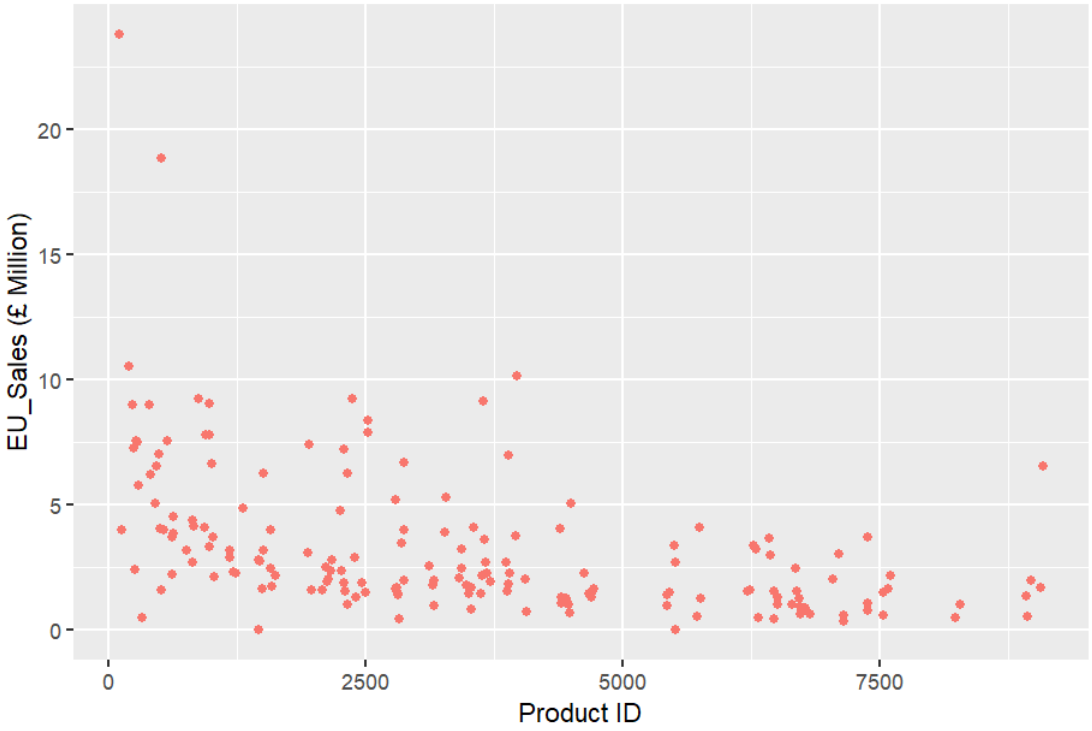
We started off by calculating Minimum, Maximum and mean values of all three regions, which are as following:

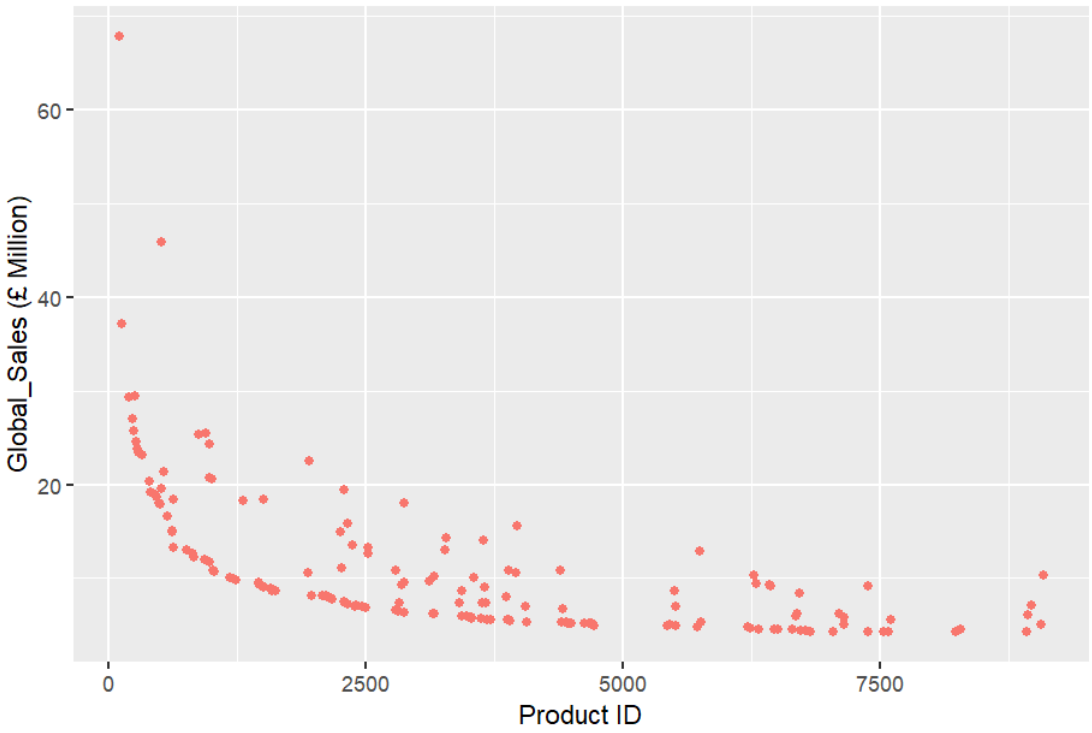
Region	Minimum Sales (£ Million)	Maximum Sales (£ Million)	Mean Sales (£ Million)
North America	**0**	**34.02**	**2.51**
European Union	**0**	**23.8**	**1.64**
Global	**0.01**	**67.85**	**5.33**

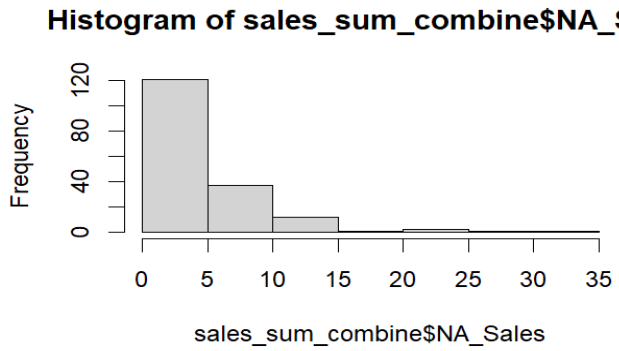
Aggregation of sales according to Product ID

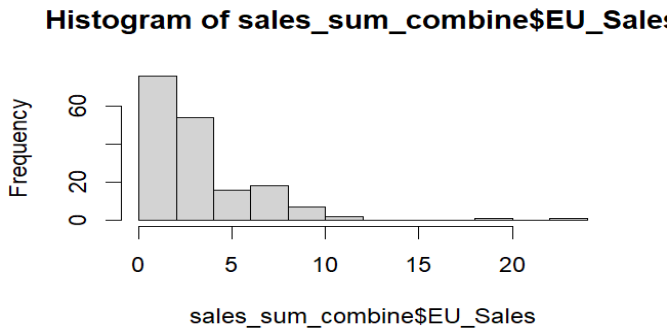
In order to observe which products (product IDs) did well sales wise in different regions, we aggregated regional sales according to Product ID and summed up the sales.

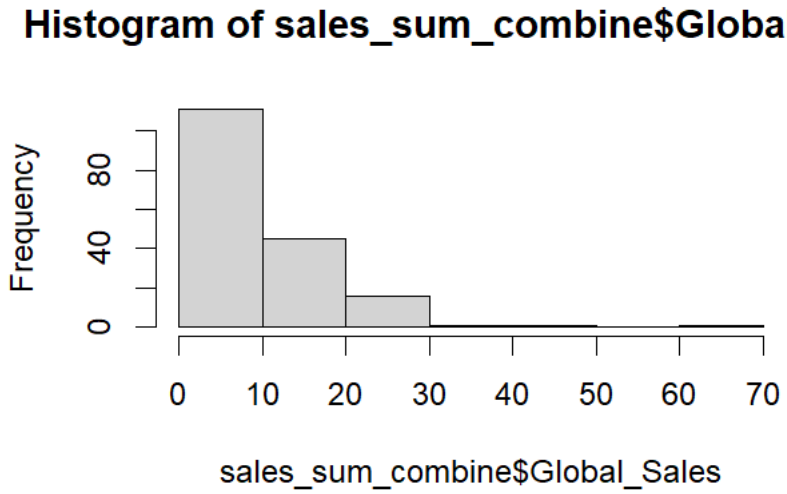



EU_Sales (£ Million)
20
15
10
5
0
0
2500
5000
7500
Product ID


Global_Sales (£ Million)
60
40
20
0
2500
5000
7500
Product ID







Skewness

Creating a histogram of the sales data, it is observed that data in all three regions is skewed to the right. Skewness can also be verified through SKEWNESS function from MOMENTS library. The values for each region are as following:

NA Sales Skewness: 3.048
EU Sales Skewness: 2.88
Global Sales Skewness: 3.066

All of the above values show that all regional sales data is positively skewed and tail is on the right side of the distribution.

Kurtosis

We can also analyze if distribution is heavy-tailed or light-tailed relative to normal distribution through kurtosis. Kurtosis values for each region are as following:

NA Sales Kurtosis: 15.6
EU Sales Kurtosis: 16.2
Global Sales Kurtosis: 17.7
Kurtosis values are greater than 3, which implies that it is leptokurtic which essentially means that it tends to produce more outliers than a normal distribution.

Correlation

Calculating the correlation between regional sales:

Correlation between European Union and North America: 0.62
Correlation between European Union and Global Sales: 0.84
Correlation between European Union and Global Sales: 0.91

Correlation values between EU - Global sales and between NA – Global sales are close to 1, which depicts that correlation is very strong. Where as correlation between EU and NA is 0.62 which can be considered as moderately strong.

This can also be depicted through a visualization:

6.0 Predicting Sales (Making Recommendations to the Business)

Final part of the assignment required us to apply Simple and multiple linear regression on the sales data set.

We start off by plotting all three regional sales data against each other to observe relationship between them. Then we run simple Linear Regression between each regional sales column. Results are as following:

Simple Linear Regression

European Union vs North America

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.17946    0.27433   4.299 2.85e-05 ***
NA_Sales     0.42028    0.04034  10.419  < 2e-16 ***

Multiple R-squared:  0.3856,    Adjusted R-squared:  0.382
```

P-value (2e-16) of this regression model indicates that NA_Sales is statistically significant variable. R-Squared value of 0.3856 indicates that 38% of NA_Sales can explain variability in EU_Sales column.

North America vs Global Sales

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.44975    0.22960  -1.959   0.0517 .
Global_Sales  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
```

P-value (2e-16) of this regression model indicates that Global_Sales is statistically significant variable. R-Squared value of 0.8395 indicates that 83% of Global_Sales can explain variability in NA_Sales column.

European Union vs Global Sales

```
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  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
```

P-value (2e-16) of this regression model indicates that Global_Sales is statistically significant variable. R-Squared value of 0.7201 indicates that 72% of Global_Sales can explain variability in European_Sales column.

Multiple Linear Regression

Next we created a Multiple Regression Model in order predict Global_Sales through EU and NA sales.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.04242    0.17736   5.877 2.11e-08 ***
NA_Sales     1.13040    0.03162  35.745  < 2e-16 ***
EU_Sales     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 regression table speaks for itself stating that 96% of Global_Sales can be predicted through both EU and NA sales.

We also predicted Global sales by explicitly supplying EU and NA values. The result was as following:

	EU_Sales	NA_Sales	Global_Sales
1	23.80	34.02	68.056548
2	1.56	3.93	7.356754
3	0.65	2.73	4.908353
4	0.97	2.26	4.761039
5	0.52	22.08	26.625558

These values are close to actual values indicating that our MLR model is able to predict significant results.