PREDICTING FUTURE OUTCOMES- TURTLE GAMES

Author: Neetu Thomas

23 December 2022

This page intentionally left blank

Contents

Abbreviations

EU	Europe
NA	North America
NLP	Natural Language Processing
OLS	Ordinary Least Squares

1 Introduction

Turtle Games wants to improve overall sales performance by utilising customer trends. The sales data as well as customer reviews are analysed to derive insights and make meaningful business decisions.

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?
- impact that each product has on sales
- data reliability
- relationship(s) (if any) between North American, European, and global sales?

2 Analytical Approach

Data provided by Turtle Games is analysed using Python to make predictions with regression.

Necessary libraries are imported into the notebook.

The dataset 'turtle_reviews.csv' is loaded into a data frame and sense checked.

- Dataset contains 2000 rows with 11 columns.
- No missing values.

```
# Determine if there are any missing values in the data set
reviews.isnull().sum()
```

```
gender                  0
age                     0
remuneration (k£)       0
spending_score (1-100)  0
loyalty_points          0
education               0
language                0
platform                0
product                 0
review                  0
summary                 0
dtype: int64
```

- Dropped unnecessary columns: 'language' and 'platform'

```
# Drop unnecessary columns.
reviews_df = reviews.drop(['language','platform'],axis=1)
```

- Columns renamed

```
# Rename the column headers.
reviews_df.rename(columns= {'remuneration (k£)' : 'remuneration',\
                            'spending_score (1-100)' : 'spending_score'},
                  inplace=True)

# View column names.
reviews_df.columns.values
```

```
array(['gender', 'age', 'remuneration', 'spending_score',
       'loyalty_points', 'education', 'product', 'review', 'summary'],
      dtype=object)
```

- Descriptive statistics

```
# Descriptive statistics.
reviews_df.describe()
```

	age	remuneration	spending_score	loyalty_points	product
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	39.495000	48.079060	50.000000	1578.032000	4320.521500
std	13.573212	23.123984	26.094702	1283.239705	3148.938839
min	17.000000	12.300000	1.000000	25.000000	107.000000
25%	29.000000	30.340000	32.000000	772.000000	1589.250000
50%	38.000000	47.150000	50.000000	1276.000000	3624.000000
75%	49.000000	63.960000	73.000000	1751.250000	6654.000000
max	72.000000	112.340000	99.000000	6847.000000	11086.000000

  - Mean age of customers is 39.5 years and the maximum is 72 years.
  - Remuneration has highest value of £112.34k.
  - Maximum spending score received - 99
  - Highest loyalty points received: 6847, lowest: 25.

- df.corr() is used to determine the correlation between the different variables in the dataset and determine their relationship.

```
reviews_df.corr()
```

	age	remuneration	spending_score	loyalty_points	product
age	1.000000	-0.005708	-0.224334	-0.042445	0.003081
remuneration	-0.005708	1.000000	0.005612	0.616065	0.305309
spending_score	-0.224334	0.005612	1.000000	0.672310	-0.001649
loyalty_points	-0.042445	0.616065	0.672310	1.000000	0.183600
product	0.003081	0.305309	-0.001649	0.183600	1.000000

Renumeration and spending score have a positive correlation with loyalty points.

Age variable shows no significant correlation with loyalty points.

3 Linear Regression

3.1 Spending_Score vs Loyalty points

The independent variable (x) or predictor is 'spending_score' and the dependant variable (y) is 'loyalty_points'.

The OLS formula y~x is applied and passed to the ols() function, and fit the model.

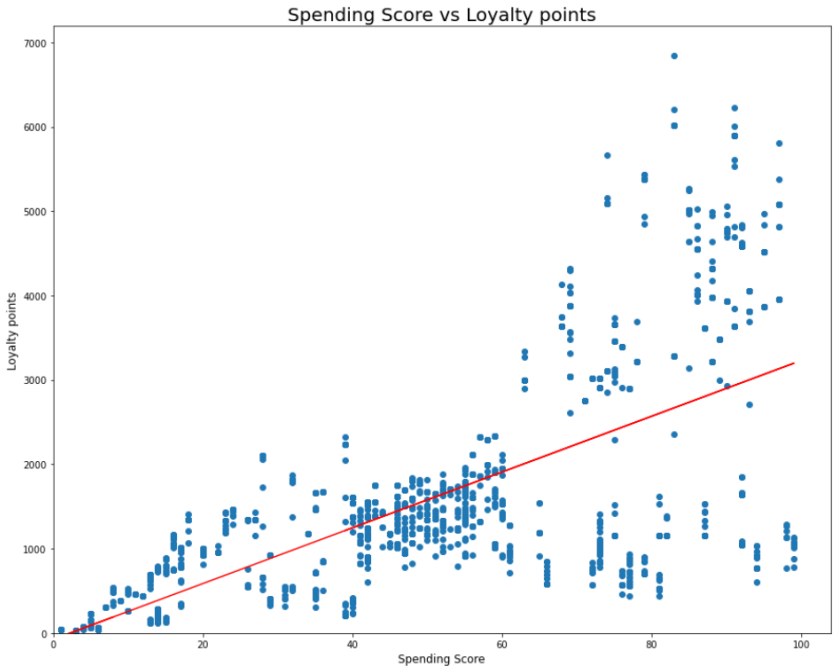
Regression summary is shown below:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.
Date:	Thu, 22 Dec 2022	**Prob (F-statistic):**	2.92e-263
Time:	11:46:51	**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.



- R^2 is 0.452 or 45.2 % of variation of loyalty points can be explained by the variation of spending score.
- Coefficient of x: if the spending score increase by 1 unit, the loyalty points(y) will increase by 33.06 units.
- The probability of the t-value is zero, thus the estimated slope is significant
- The confidence interval (95%) tells that for different set of random samples of 2000 observations, the slope will remain within the interval 31.46 and 34.66
- Moderate positive correlation seen.
- Linear regression model shows heteroscedasticity, the variation of residuals is not constant.
- It is recommended that the outliers (above 60 spending scores) are removed to get a better fit.

3.2 Renumeration Vs Loyalty points

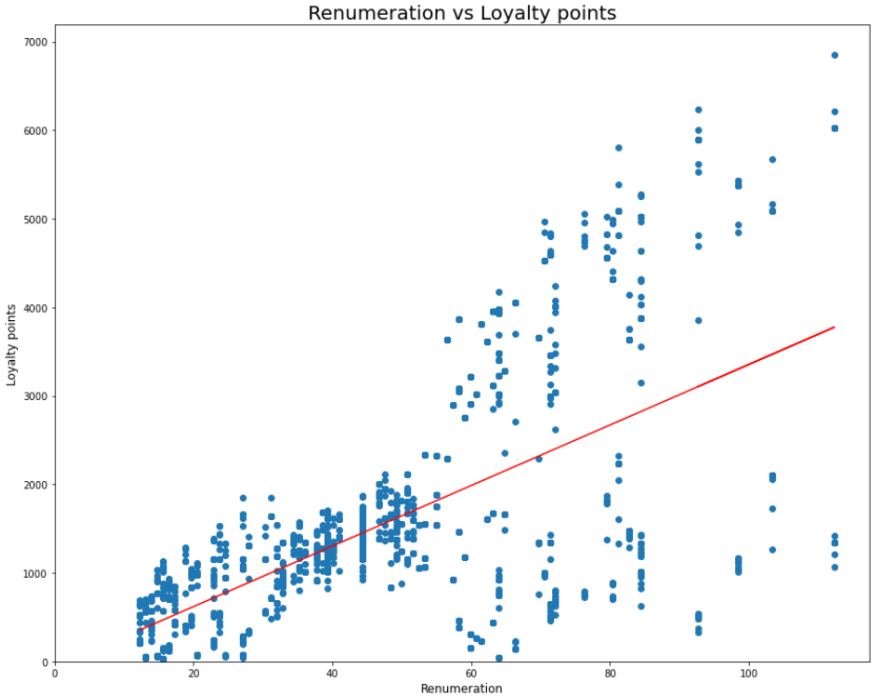
OLS Regression summary shows R^2=38%, variation of loyalty points can be explained by variation in renumeration. The slope is significant.

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.380
Model:	OLS	**Adj. R-squared:**	0.379
Method:	Least Squares	**F-statistic:**	1222.
Date:	Wed, 21 Dec 2022	**Prob (F-statistic):**	2.43e-209
Time:	23:46:10	**Log-Likelihood:**	-16674.
No. Observations:	2000	**AIC:**	3.335e+04
Df Residuals:	1998	**BIC:**	3.336e+04
Df Model:	1		
Covariance Type:	nonrobust		

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

Omnibus:	21.285	**Durbin-Watson:**	3.622
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	31.715
Skew:	0.089	**Prob(JB):**	1.30e-07
Kurtosis:	3.590	**Cond. No.**	123.



- A positive correlation exists between the renumeration and loyalty points.
- Higher the renumeration, more loyalty points accumulated.
- There exists heteroscedasticity and outliers > £55k renumeration can be removed for a more accurate fit.
- Most data is concentrated below 3000 loyalty points.

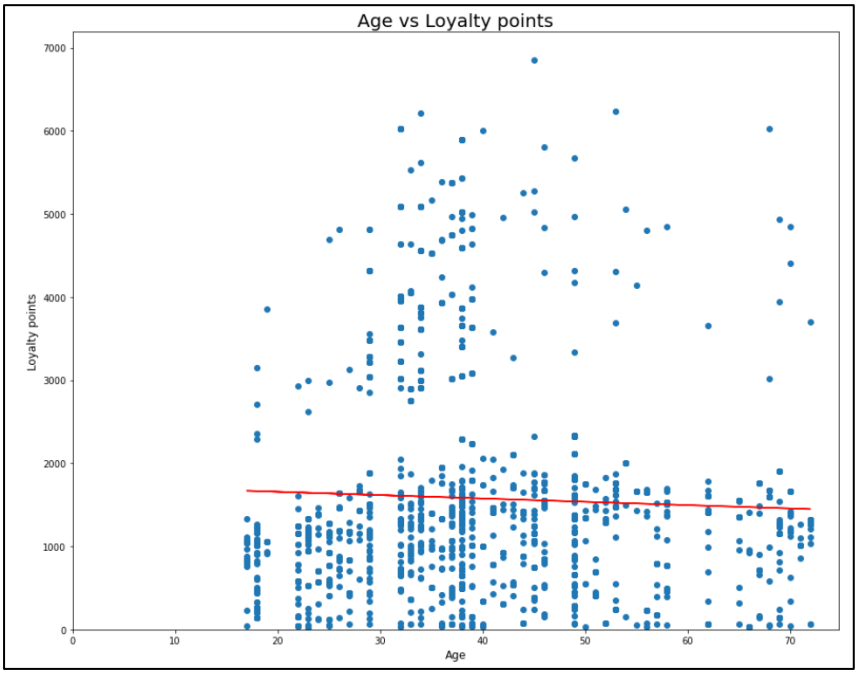
3.3 Age Vs Loyalty points

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.606
Date:	Wed, 21 Dec 2022	**Prob (F-statistic):**	0.0577
Time:	23:41:08	**Log-Likelihood:**	-17150.
No. Observations:	2000	**AIC:**	3.430e+04
Df Residuals:	1998	**BIC:**	3.431e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1736.5177	88.249	19.678	0.000	1563.449	1909.587
x	-4.0128	2.113	-1.899	0.058	-8.157	0.131

Omnibus:	481.477	**Durbin-Watson:**	2.277
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	937.734
Skew:	1.449	**Prob(JB):**	2.36e-204
Kurtosis:	4.688	**Cond. No.**	129.

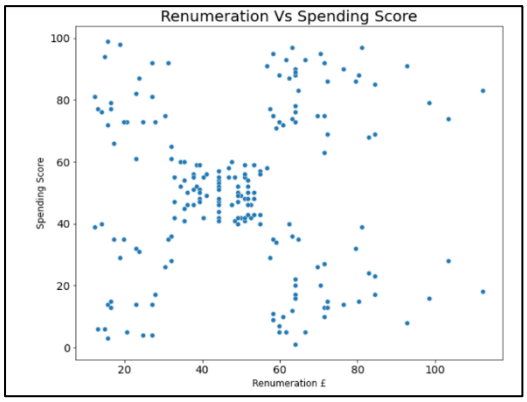


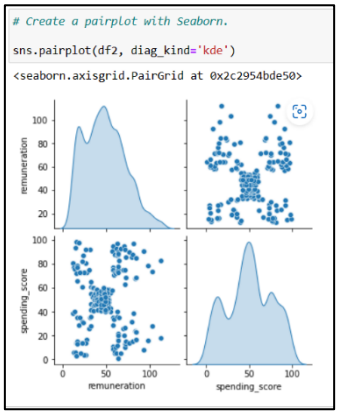
- R^2 =0.2% - loyalty points variations cannot be accurately explained by the age variations.
- The slope is not significant (p>0.05) and not a good fit.
- The scatterplot shows there is no correlation between age and loyalty points.

4 Predictions with clustering

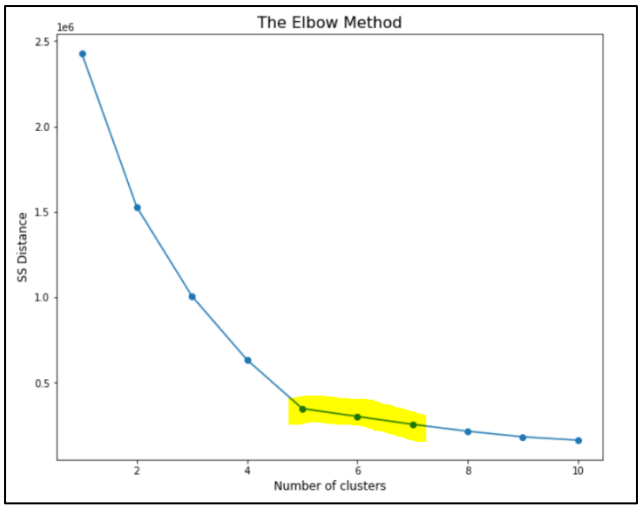
Spending_scores and renumeration relationship are explored to understand customer groups.

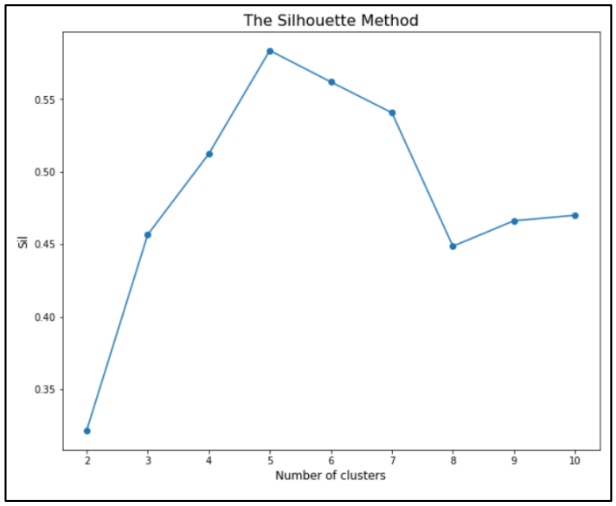
Scatterplot and pairplot shown





The Elbow and Silhouette methods used to determine the optimal number of clusters for *k*-means clustering.

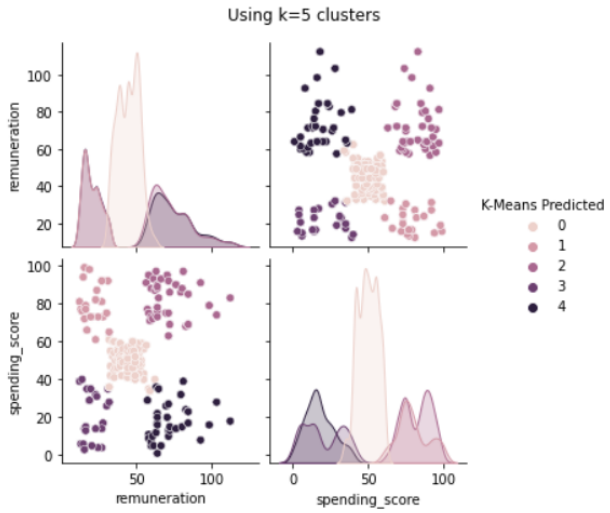




- The point where the graph becomes linear is 5 – assumed as the optimal number of clusters.
- In the silhouette method, highest score is for clusters=5
- k-means models evaluated using different k-clusters, 5,6,7 and fit and the predicted k-means are plotted.
- Clusters = 5 is the best fit.

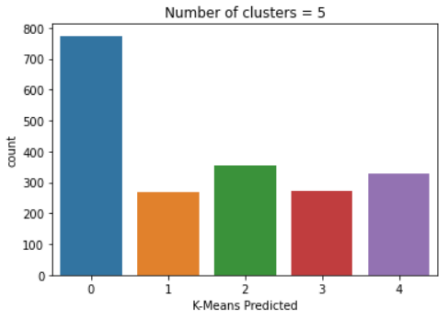
```
# Using 5 clusters
df_k5 = df2.copy()
kmeans = KMeans(n_clusters=5, max_iter=15000, init='k-means++',random_state=42).fit(x)
clusters = kmeans.labels_
df_k5['K-Means Predicted'] = clusters



# Create a pairplot
ax = sns.pairplot(df_k5, hue='K-Means Predicted', diag_kind='kde')
ax.fig.suptitle('Using k=5 clusters', y = 1.05)
```

```
Text(0.5, 1.05, 'Using k=5 clusters')
```




```
# Check the number of observations per predicted class.
df_k5['K-Means Predicted'].value_counts()
```

```
0    774
2    356
4    330
3    271
1    269
Name: K-Means Predicted, dtype: int64
```



- Although cluster 0 is the largest group, k=5 is the optimal model. The number of predicted values per cluster indicates a better distribution.

4.1 Visualising kmeans clusters =5 model:

- Spending score 40-60 and renumeration (£30k-£50k) have the largest group of customers.
- Cluster 2 - high spending score and high renumeration.
- Depending on the renumeration and spending, the different marketing strategies can be utilised for various groups of customer base.

5 Sentiment Analysis

The review and summary columns are analysed using NLP to understand customer sentiments

Dataframe containing only these columns created.

Data cleaned and wrangled for NLP.

- Missing values: None
- Duplicates: 39 removed

User defined functions utilised

- Columns converted to lower case and punctuations removed.
- Tokenised using word_tokenize to calculate frequency distribution

```
# Apply tokenisation to both columns.
# Tokenise 'review' column
df_final['review_token'] = df_final['review'].apply(word_tokenize)

# View DataFrame.
df_final['review_token'].head()
```

```
0    [when, it, comes, to, a, dms, screen, the, spa...
1    [an, open, letter, to, galeforce9, your, unpai...
2    [nice, art, nice, printing, why, two, panels, ...
3    [amazing, buy, bought, it, as, a, gift, for, o...
4    [as, my, review, of, gf9s, previous, screens, ...
Name: review_token, dtype: object
```

5.1 Word Cloud

Word Cloud for review and summary generated and stopwords removed.

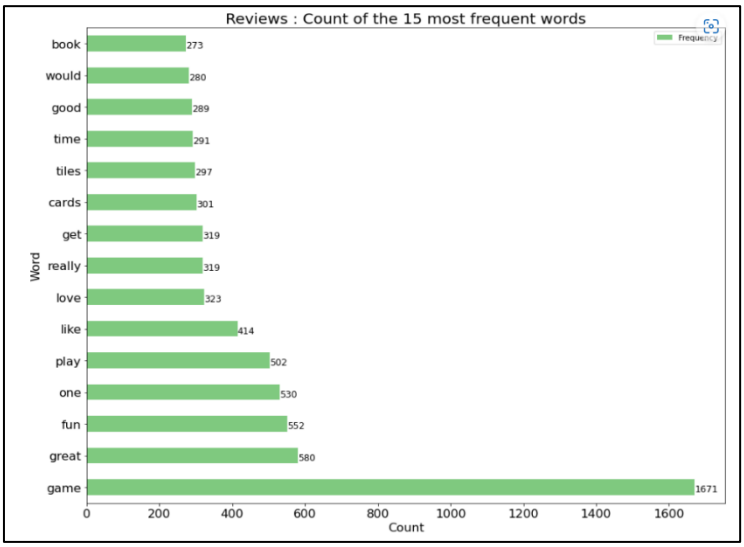
Fig: Reviews Word cloud without stopwords

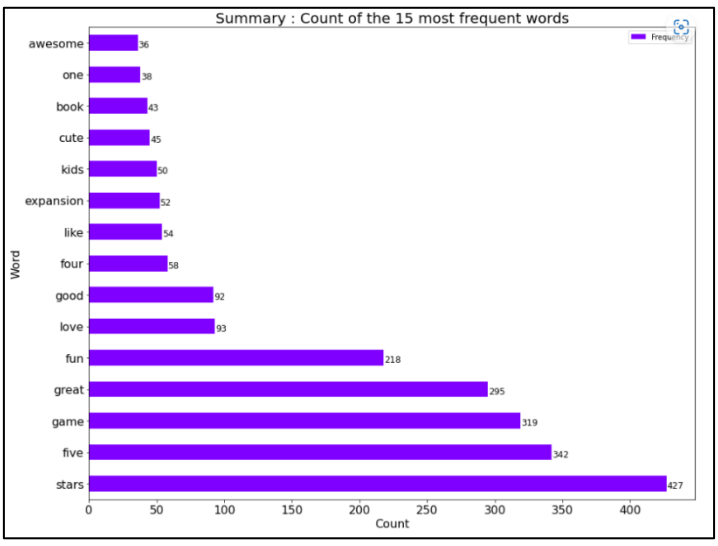
Fig: Summary – Word cloud without stopwords

Most common words determined; results visualised in a countplot.

```
# 15 Most common words in reviews
review_common_count = pd.DataFrame(Counter(review_tokens_clean).most_common(15),
                      columns = ['Word', 'Frequency']).set_index('Word')
print(review_common_count)
```

```
        Frequency
Word
game         1671
great         580
fun           552
one           530
play          502
like          414
love          323
really        319
get           319
cards         301
tiles         297
time          291
good          289
would         280
book          273
```



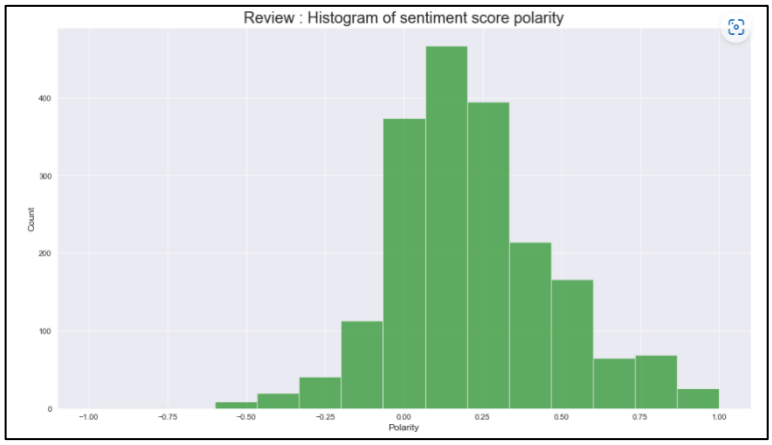


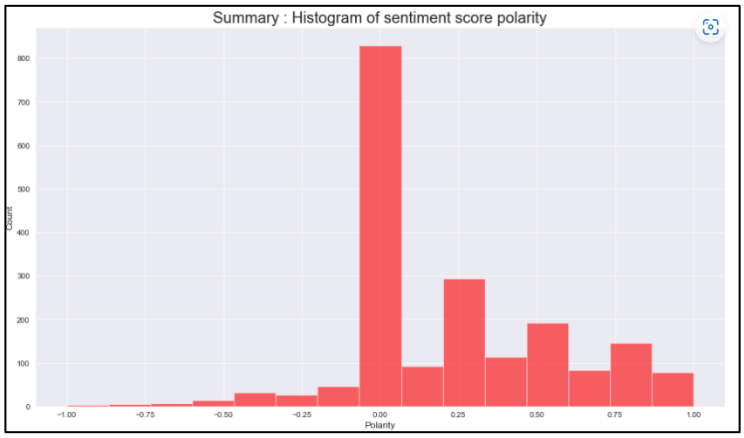
5.2 Polarity

Polarity scores calculated using TextBlob(), results plotted in histograms.

```
# Function :- To plot histogram for polarity
def plot_histogram(col, color, title):
    # Set the plot area
    plt.figure(figsize=(16,9))
    sns.set_style('darkgrid')
    # Define the bars
    plt.hist(col, bins=15, facecolor=color, alpha=0.6)

    # Set the labels
    plt.xlabel('Polarity', fontsize=12)
    plt.ylabel('Count', fontsize=12)
    plt.title(f'{title} : Histogram of sentiment score polarity', fontsize=20)
    plt.show()
```

```
# Histogram of sentiment score polarity for 'review'
plot_histogram(df_final['review_polarity'], 'green', 'Review')
```





- Positive skewed distribution seen. There is a majority of neutral sentiment score for the reviews and summary by customers but less negative sentiment.

5.3 Reviews

The top 20 positive and negative review and summaries are identified

```
# Top 20 positive summaries
# Create a DataFrame
positive_summary = df_final.nlargest(20, 'summary_polarity')

# Eliminate unnecessary columns
positive_summary = positive_summary[['summary',
                                     'summary_polarity']]

# Adjust the column width
positive_summary.style.set_properties(subset=['summary'], **{'width' : '900px'})
```

	summary	summary_polarity
6	best gm screen ever	1.000000
28	wonderful designs	1.000000
32	perfect	1.000000
76	theyre the perfect size to keep in the car or a diaper	1.000000
129	perfect for preschooler	1.000000
135	awesome sticker activity for the price	1.000000
156	awesome book	1.000000
158	he was very happy with his gift	1.000000
182	awesome	1.000000
205	awesome and welldesigned for 9 year olds	1.000000
407	perfect	1.000000
463	excellent	1.000000
531	excellent	1.000000
536	excellent therapy tool	1.000000
567	the pigeon is the perfect addition to a school library	1.000000
586	best easter teaching tool	1.000000
634	wonderful	1.000000
638	all f the mudpuppy toys are wonderful	1.000000
644	awesome puzzle	1.000000
649	not the best quality	1.000000

	review	review_polarity
203	booo unles you are patient know how to measure i didnt have the patience neither did my daughter boring unless you are a craft person which i am not	-1.000000
177	incomplete kit very disappointing	-0.780000
1769	im sorry i just find this product to be boring and to be frank juvenile	-0.583333
357	one of my staff will be using this game soon so i dont know how well it works as yet but after looking at the cards i believe it will be helpful in getting a conversation started regarding anger and what to do to control it	-0.550000
112	i bought this as a christmas gift for my grandson its a sticker book so how can i go wrong with this gift	-0.500000
222	this was a gift for my daughter i found it difficult to use	-0.500000
225	i found the directions difficult	-0.500000
285	instructions are complicated to follow	-0.500000
295	difficult	-0.500000
1492	expensive for what you get	-0.500000
169	i sent this product to my granddaughter the pompom maker comes in two parts and is supposed to snap together to create the pompoms however both parts were the same making it unusable if you cant make the pompoms the kit is useless since this was sent as a gift i do not have it to return very disappointed	-0.491667
340	my 8 yearold granddaughter and i were very frustrated and discouraged attempting this craft it is definitely not for a young child i too had difficulty understanding the directions we were very disappointed	-0.446250
526	i purchased this on the recommendation of two therapists working with my adopted children the children found it boring and put it down half way through	-0.440741
300	very hard complicated to make these	-0.439583
416	kids i work with like this game	-0.400000
425	this game although it appears to be like uno and have an easier play method it was still too time consuming and wordy for my children with learning disabilities	-0.400000
485	my son loves playing this game it was recommended by a counselor at school that works with him	-0.400000
788	this game is a blast	-0.400000
791	i bought this for my son he loves this game	-0.400000
807	was a gift for my son he loves the game	-0.400000

The sentiment analysis for the summary column cannot be considered very accurate-needs further investigation. Eg: 'not the best quality' is considered positive due to presence of word 'best and 'quality' although it implies negative sentiment.

The higher neutral and positive sentiments show customers relate to turtle games products and therefore marketing campaigns can be utilised to improve sales.

6 Data Visualisation and Insights using R

The necessary libraries and packages loaded into RStudio.

Dataset turtle_sales.csv loaded

dim() and str() gives the dimensions and structure. : contains 352 rows and 9 columns.

- Unnecessary columns are removed using select()

```
# Remove unnecessary columns.
sales_new <- select(sales, -c(Ranking, Year, Genre, Publisher))
# View the data frame.
head(sales_new)
Product Platform NA_Sales EU_Sales Global_Sales
    107      Wii    34.02    23.80        67.85
    123      NES    23.85     2.94        33.00
    195      Wii    13.00    10.56        29.37
    231      Wii    12.92     9.03        27.06
    249       GB     9.24     7.29        25.72
    254       GB    19.02     1.85        24.81
```

- 'product' column converted to factor as it is a categorical variable
- Dataset explored using glimpse()

```
> # Add new column and convert the 'product' column to factor
> sales_new <- mutate(sales_new, Product_Id = as.factor(Product))
> glimpse(sales_new)
Rows: 352
Columns: 6
$ Product      <int> 107, 123, 195, 231, 249, 254, 263, 283, 291, 3…
$ Platform     <chr> "Wii", "NES", "Wii", "Wii", "GB", "GB", "DS", …
$ NA_Sales     <dbl> 34.02, 23.85, 13.00, 12.92, 9.24, 19.02, 9.33,…
$ EU_Sales     <dbl> 23.80, 2.94, 10.56, 9.03, 7.29, 1.85, 7.57, 7.…
$ Global_Sales <dbl> 67.85, 33.00, 29.37, 27.06, 25.72, 24.81, 24.6…
$ Product_Id   <fct> 107, 123, 195, 231, 249, 254, 263, 283, 291, 3…
>
```

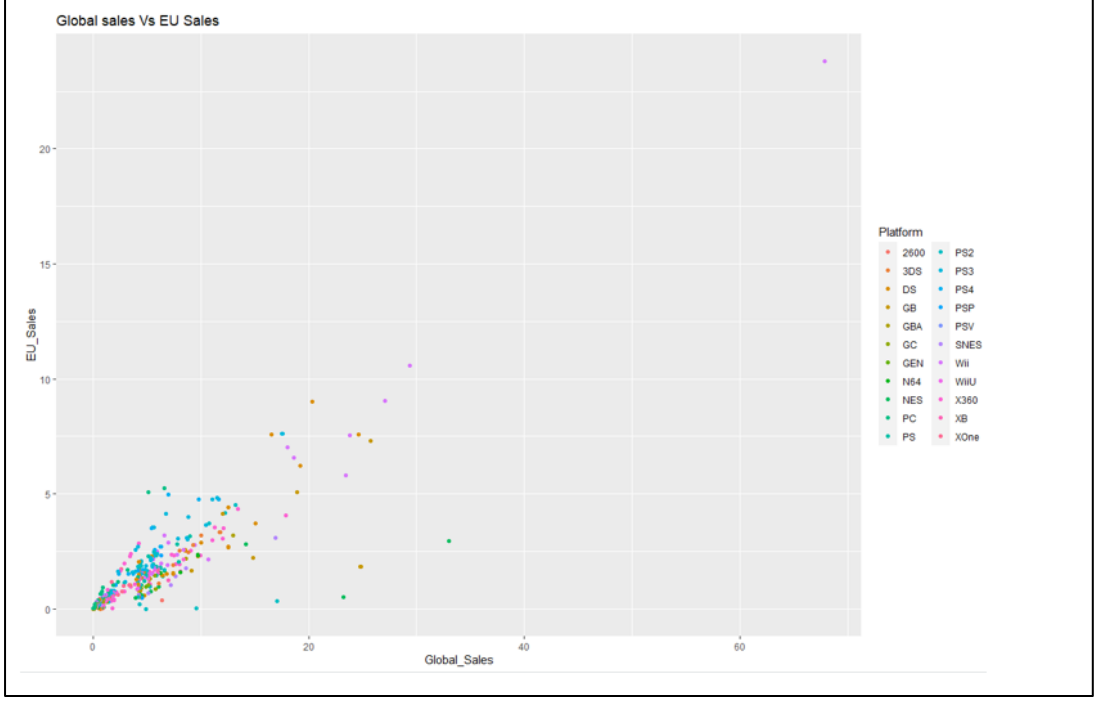
- Summary() gives the descriptive statistics and the maximum and minimum sales across North America, Europe and global regions.

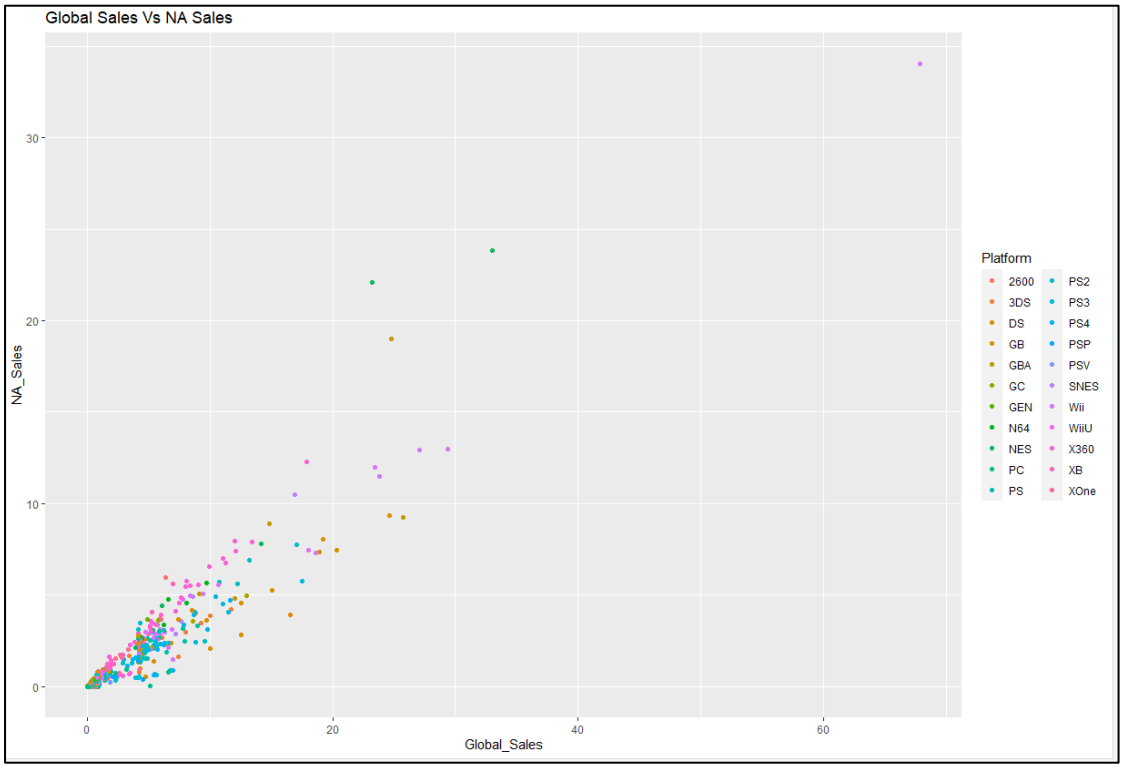
```
> # View the descriptive statistics.
> summary(sales_new)
    Product       Platform           NA_Sales
 Min.   : 107   Length:352         Min.   : 0.0000
 1st Qu.:1945   Class :character   1st Qu.: 0.4775
 Median :3340   Mode  :character   Median : 1.8200
 Mean   :3607                      Mean   : 2.5160
 3rd Qu.:5436                      3rd Qu.: 3.1250
 Max.   :9080                      Max.   :34.0200

    EU_Sales       Global_Sales      Product_Id
 Min.   : 0.000   Min.   : 0.010   3645   :  9
 1st Qu.: 0.390   1st Qu.: 1.115   2518   :  8
 Median : 1.170   Median : 4.320   3967   :  8
 Mean   : 1.644   Mean   : 5.335   3887   :  7
 3rd Qu.: 2.160   3rd Qu.: 6.435   9080   :  7
 Max.   :23.800   Max.   :67.850   1945   :  6
                                   (Other):307
```

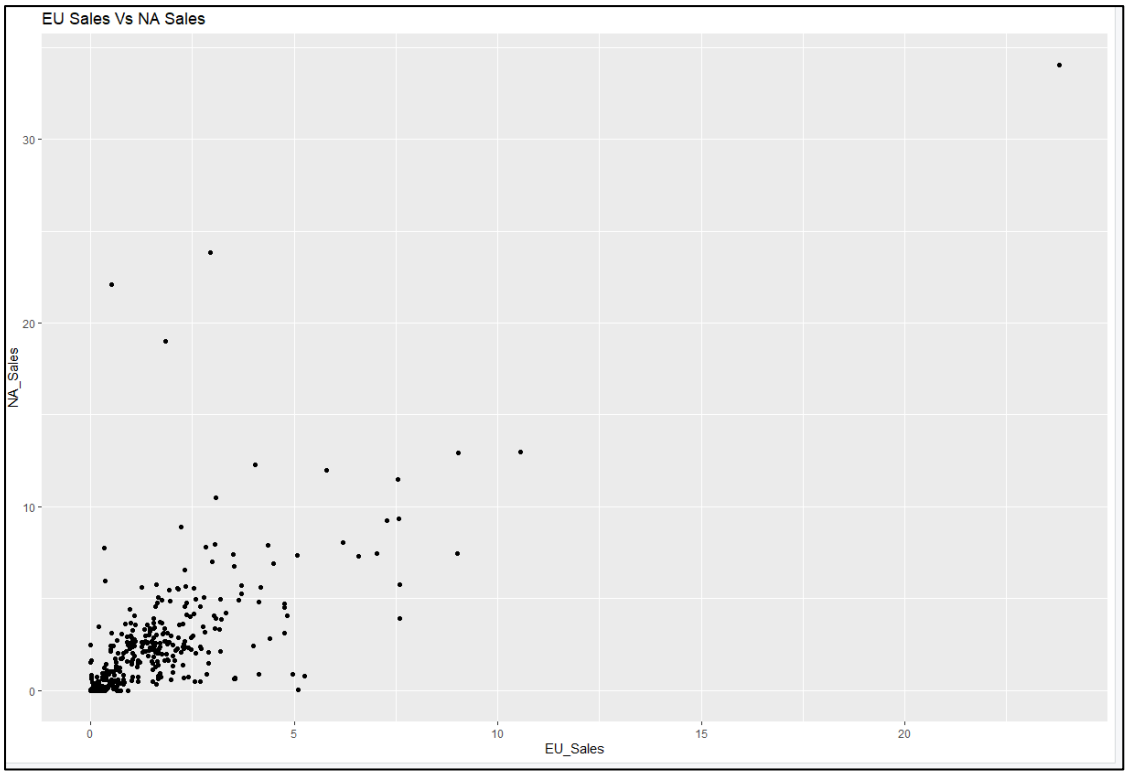
- Missing values : is.na() – none

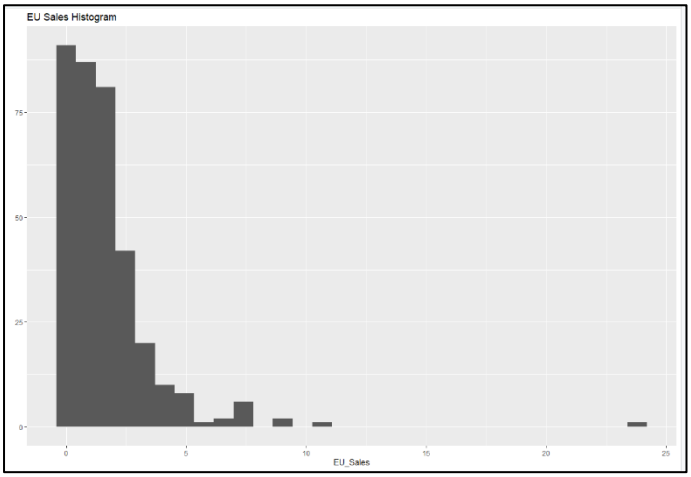
- Qplot() used to plot different sales data to gather insights

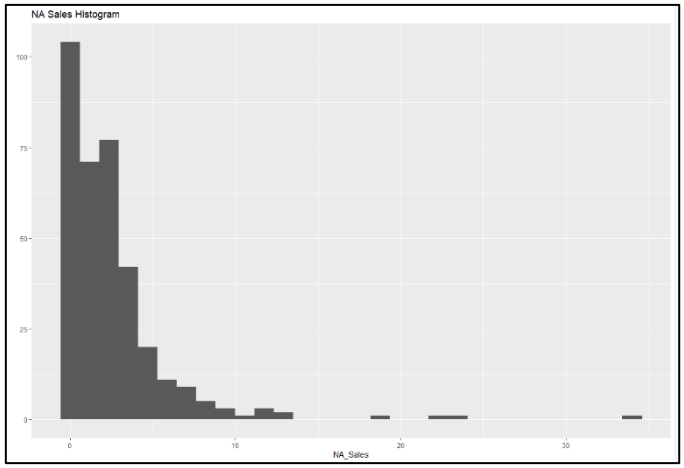
```
qplot(Global_Sales, EU_Sales, data = sales_new, col=Platform,
      geom=c('point','jitter'),main = 'Global sales Vs EU Sales')
```

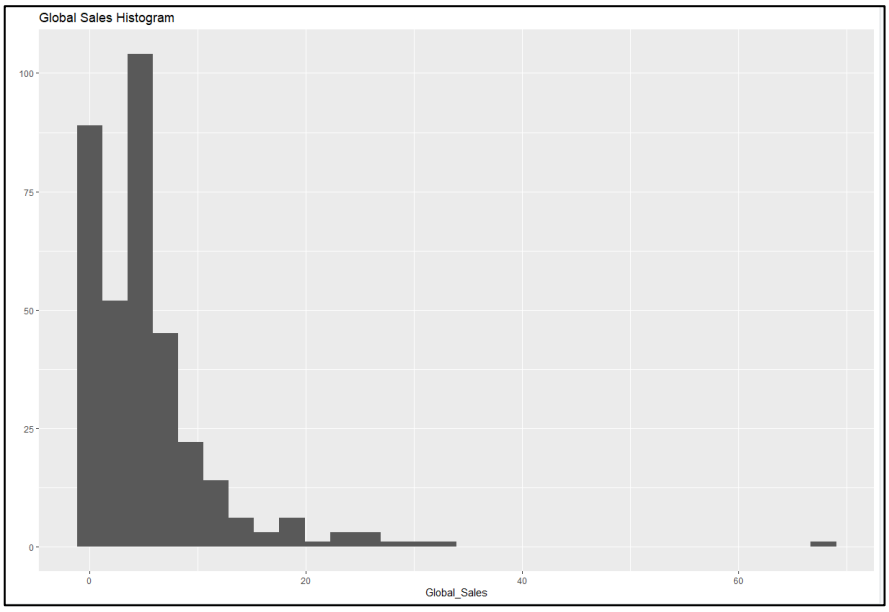


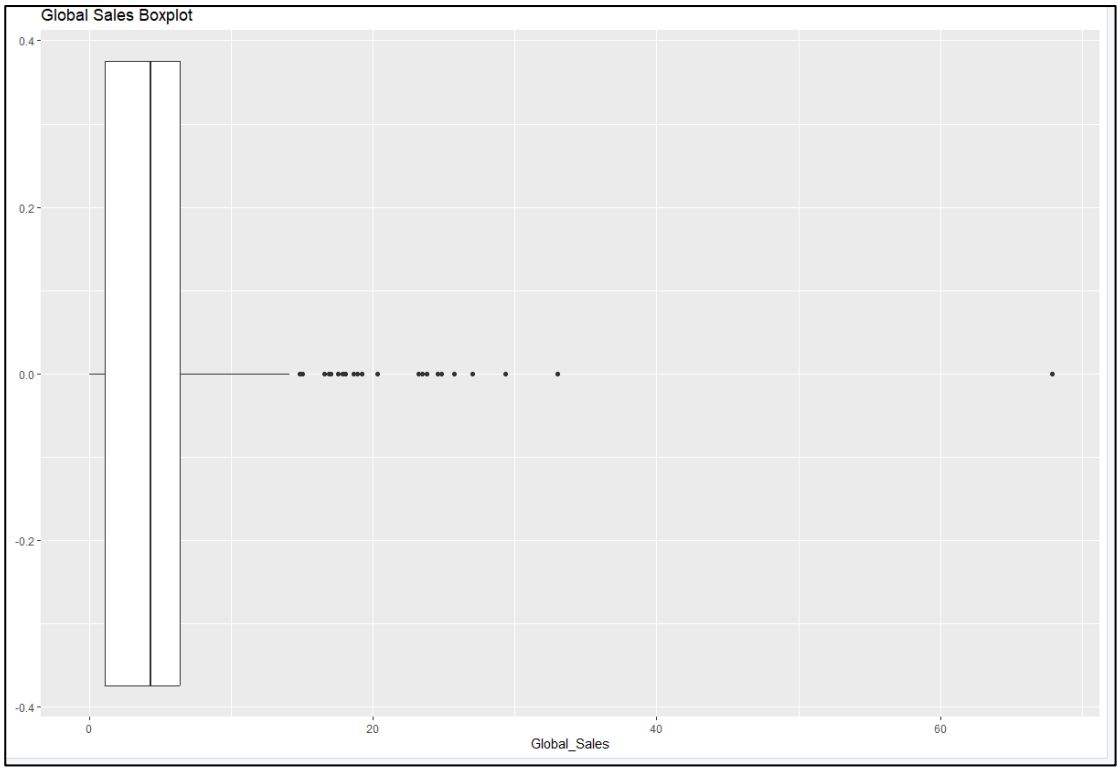



EU Sales Vs NA Sales
NA_Sales
EU_Sales
0
10
20
30
0
5
10
15
20


EU Sales Histogram
EU_Sales
0
25
50
75
0
5
10
15
20
25

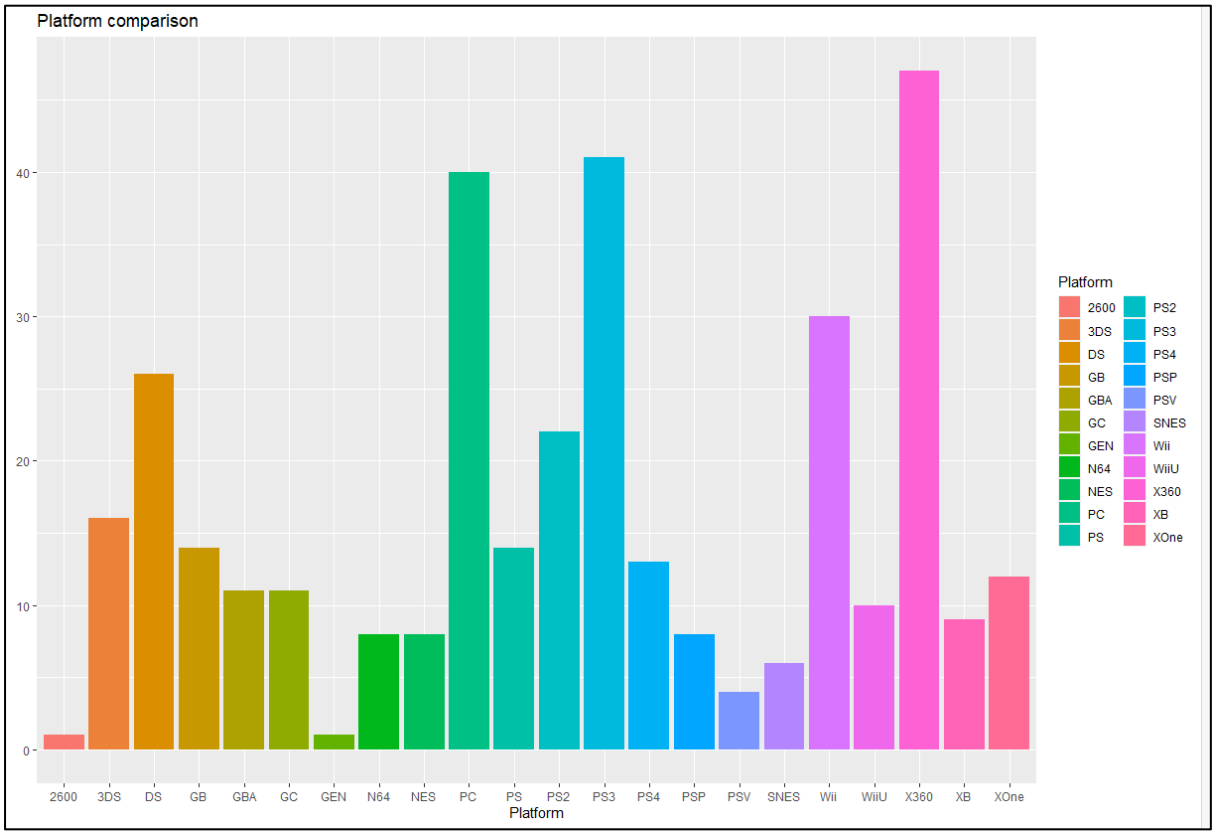

NA Sales Histogram
NA_Sales
0
25
50
75
100
0
10
20
30


Global Sales Histogram
Global_Sales
0
25
50
75
100
0
20
40
60



- Histograms of EU ,NA and Global sales show that NA_sales are higher with a light tail .
- For overall global sales, the NA sales > EU sales.
- Boxplot shows outliers in the sales data

Explore platforms: X360 , PS3 and PC - most popular



- Sales data grouped by product and their impact analysed
- Group_by , summarise() is used with pipe operator %>%.

```
# Group data based on Product and determine the sum per Product.
sales_product <- sales_new %>% group_by(Product_Id) %>%
  summarise(Total_EU_Sales = sum(EU_Sales),
            Total_NA_Sales = sum(NA_Sales),
            Total_Global_Sales = sum(Global_Sales),
            Total_Other_Sales = sum(Global_Sales) - sum(EU_Sales) - sum(NA_Sales),
            .groups = 'drop')
# View the data frame.
head(sales_product)
A tibble: 6 x 5
Product_Id Total_EU_Sales Total_NA_Sales Total_Global_Sales Total_Other_Sales
<fct>               <dbl>          <dbl>              <dbl>             <dbl>
107                23.8            34.0               67.8              10.0
123                 4.01           26.6               37.2               6.51
195                10.6            13                 29.4               5.81
231                 9.03           12.9               27.1               5.11
249                 7.29            9.24              25.7               9.19
254                 2.42           21.5               29.4               5.51
```

- arrange() is used to see the Global sales in descending order.

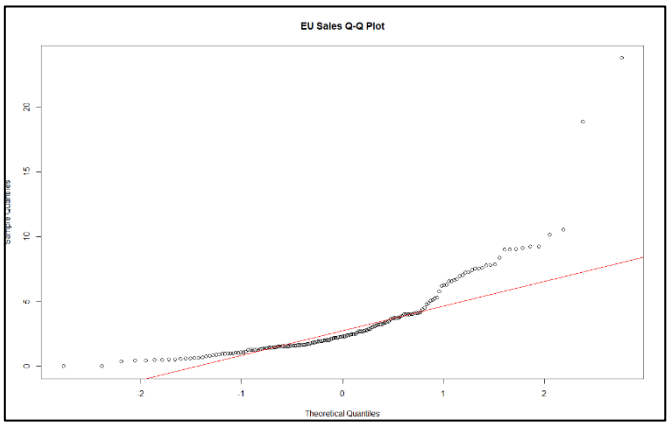
```
> arrange(sales_product, desc(Total_Global_Sales))
# A tibble: 175 x 5
   Product_Id Total_EU_Sales Total_NA_Sales Total_Global_Sales Total_Other_Sales
   <fct>               <dbl>          <dbl>              <dbl>             <dbl>
 1 107                23.8            34.0               67.8              10.0
 2 515                18.9            19.2               45.9               7.73
 3 123                 4.01           26.6               37.2               6.51
 4 254                 2.42           21.5               29.4               5.51
 5 195                10.6            13                 29.4               5.81
 6 231                 9.03           12.9               27.1               5.11
 7 249                 7.29            9.24              25.7               9.19
 8 948                 7.79           14.4               25.4               3.24
 9 876                 9.25           12.8               25.3               3.26
10 263                 7.57            9.33              24.6               7.71
# … with 165 more rows
```

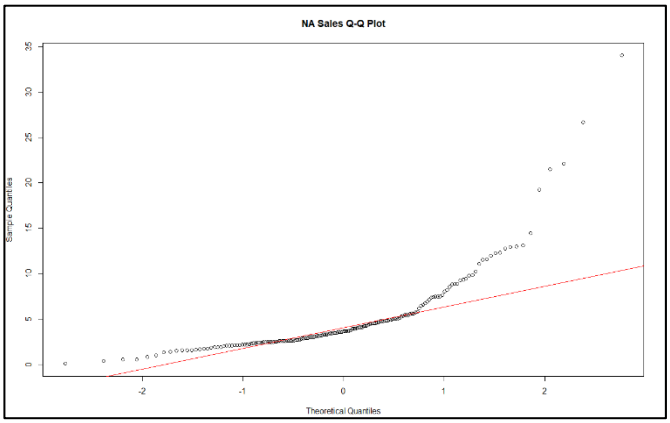
- Product_Id 107 has the maximum sales.
- The sales data is also grouped based on Platform and genre and the top platforms and genres with the most sales can be seen.

```
# Group sales based on Platform:
sales_platform<- sales_new %>% group_by(Platform) %>%
  summarise(Total_EU_Sales = sum(EU_Sales),
            Total_NA_Sales = sum(NA_Sales),
            Total_Global_Sales = sum(Global_Sales),
            .groups = 'drop')
top_platforms <- head(arrange(sales_platform,desc(Total_Global_Sales)),n=5)
top_platforms
A tibble: 5 x 4
Platform Total_EU_Sales Total_NA_Sales Total_Global_Sales
<chr>              <dbl>          <dbl>              <dbl>
Wii                105.            150.               313.
X360                76.0           153.               254.
PS3                 88.5            77.8              212.
DS                  65.6            72.6              205.
GB                  28.2            68.7              134.
```

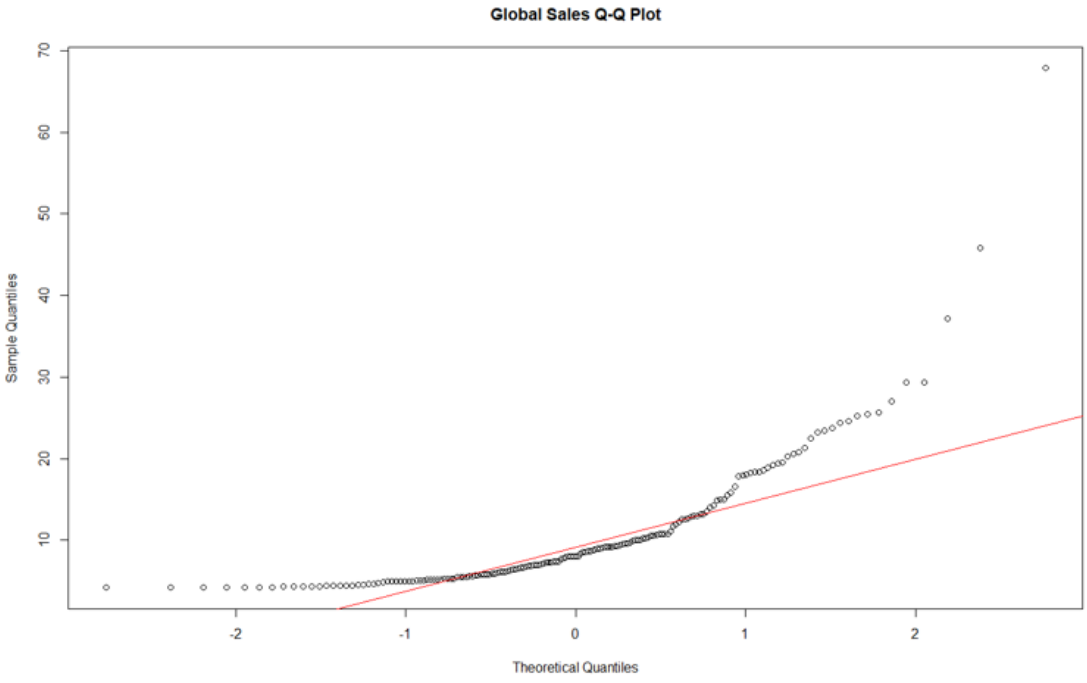
6.1 Data Reliability

Q-Q plots created to determine normality of sales data.






```
> qqnorm(sales_df$Total_Global_Sales,
+        main = 'Global Sales Q-Q Plot')
> qqline(sales_df$Total_Global_Sales, col = 'red')
>
```



Q-Q plots show the sales data is not normally distributed. There is a tail and outliers in the distribution which can skew the results.

Shapiro Wilk test performed.

```
> ## Perform Shapiro-Wilk test
> shapiro.test(sales_df$Total_EU_Sales)

        Shapiro-Wilk normality test

data:  sales_df$Total_EU_Sales
W = 0.74058, p-value = 2.987e-16

>
> shapiro.test(sales_df$Total_NA_Sales)

        Shapiro-Wilk normality test

data:  sales_df$Total_NA_Sales
W = 0.69813, p-value < 2.2e-16

>
> shapiro.test(sales_df$Total_Global_Sales)

        Shapiro-Wilk normality test

data:  sales_df$Total_Global_Sales
W = 0.70955, p-value < 2.2e-16

>
```

p-value < 0.05 - data is **not normally distributed.** Outliers in the distribution.

Skewness and kurtosis of the sales data show positive skewness and a high kurtosis value(>3). The data is not platykurtic but with heavy tails and outliers.

```
> # Skewness and Kurtosis.
>
> # EU Sales:
> skewness(sales_df$Total_EU_Sales)
[1] 2.886029
> kurtosis(sales_df$Total_EU_Sales)
[1] 16.22554
>
> # NA Sales
> skewness(sales_df$Total_NA_Sales)
[1] 3.048198
> kurtosis(sales_df$Total_NA_Sales)
[1] 15.6026
>
> # Global sales
> skewness(sales_df$Total_Global_Sales)
[1] 3.066769
> kurtosis(sales_df$Total_Global_Sales)
[1] 17.79072
>
```

- Cor() – determine correlation between EU, NA and global sales.

```
cor(sales_df$Total_EU_Sales, sales_df$Total_NA_Sales)
[1] 0.6209317
cor(sales_df$Total_Global_Sales, sales_df$Total_EU_Sales)
[1] 0.8486148
cor(sales_df$Total_Global_Sales, sales_df$Total_NA_Sales)
[1] 0.9162292
```

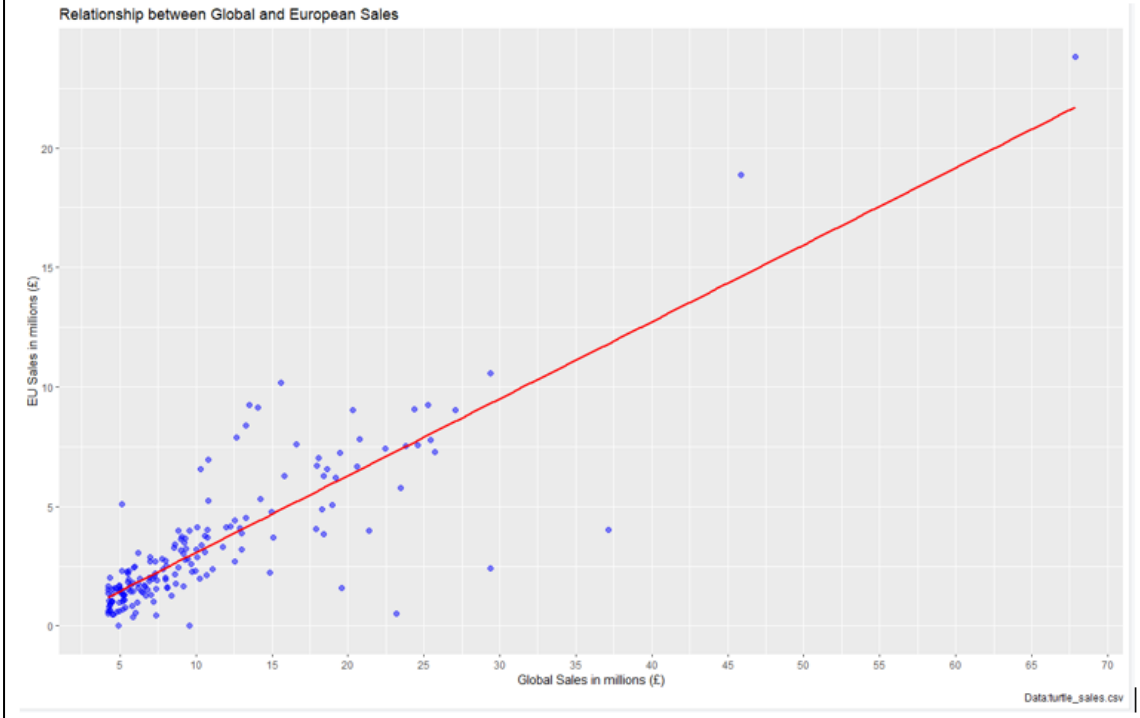
EU-Global – good correlation 84.9%

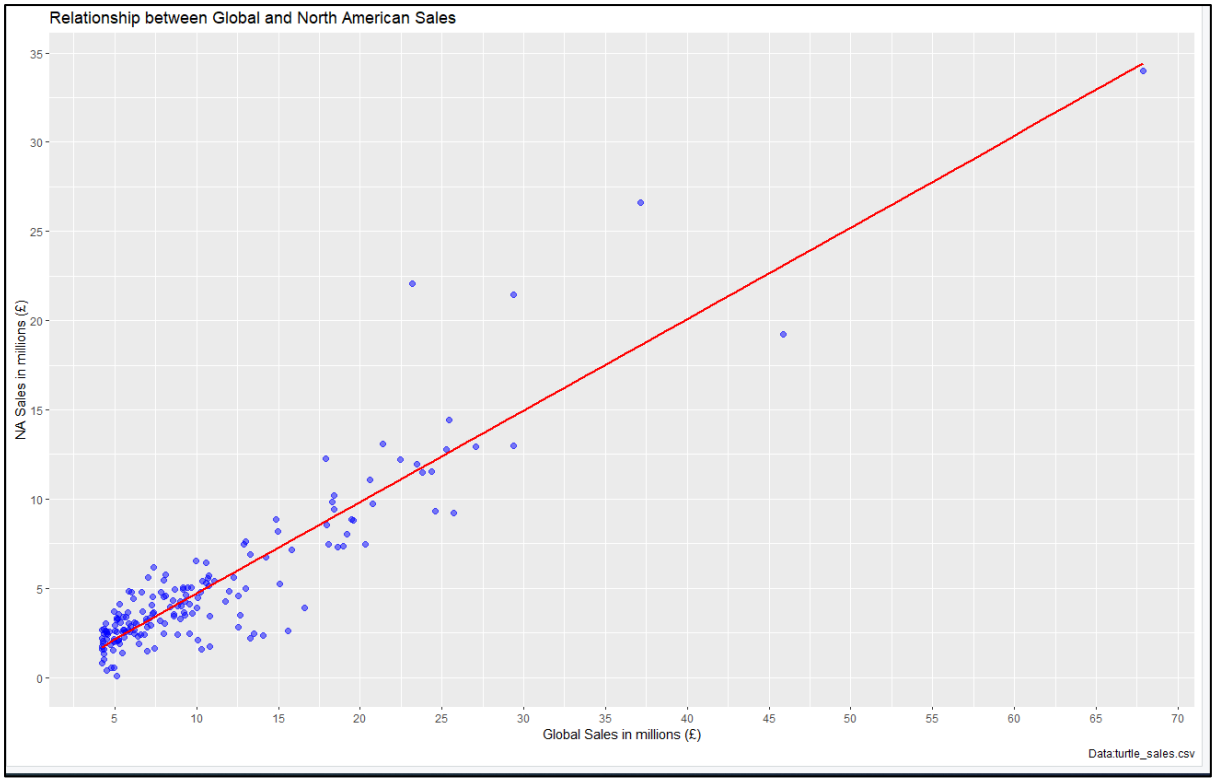
NA-Global – strong correlation 91.6%

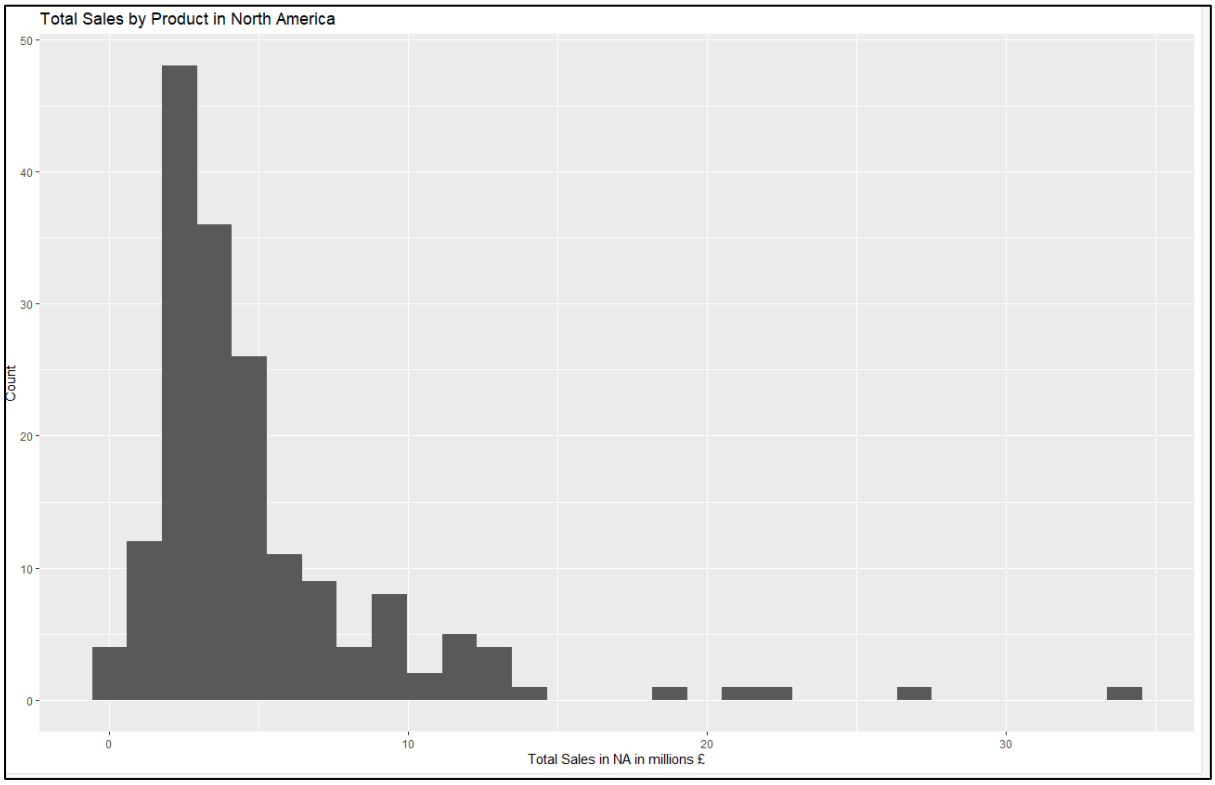
6.2 Visualisations

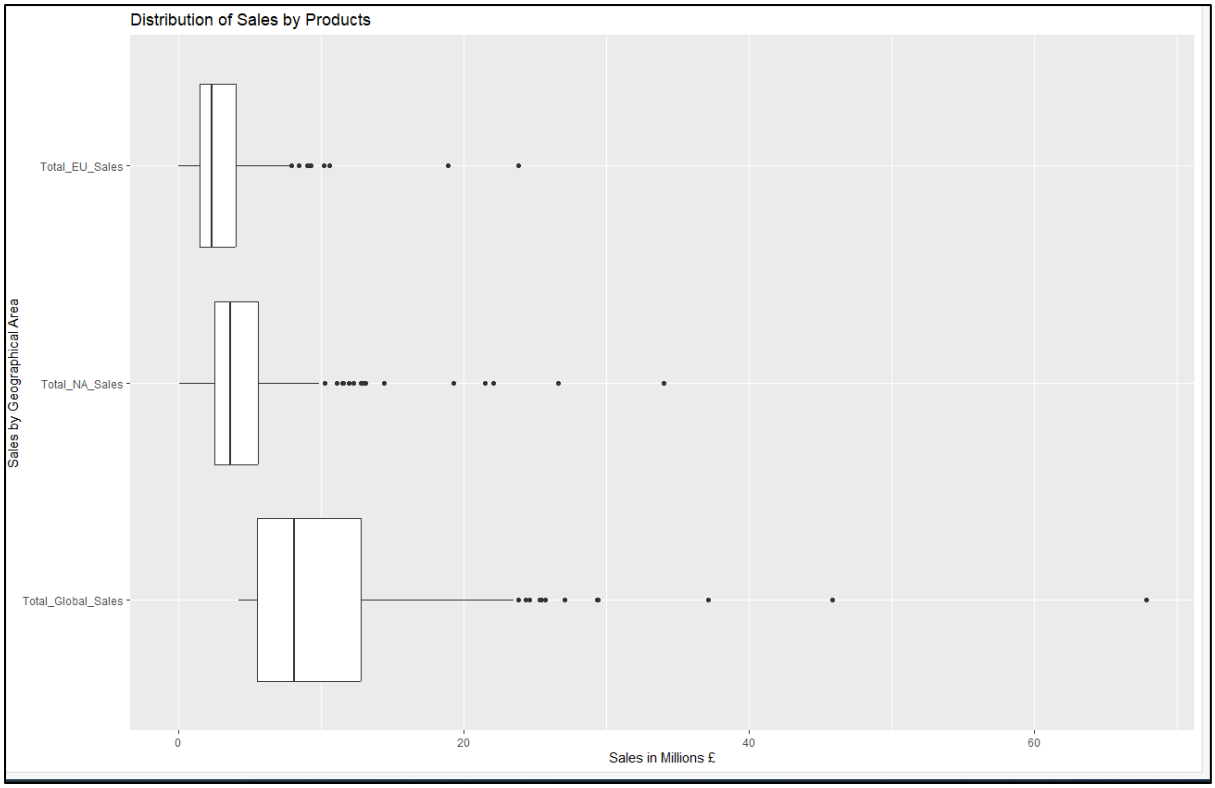
- ggplot() and geom_ utilised to created plots to gain insights in sales data

```
# Global Sales Vs European Sales
ggplot(data = sales_df,
       aes(x = Total_Global_Sales, y = Total_EU_Sale
  geom_point(color ='blue', alpha = 0.5, size = 2) +
  geom_smooth(method = 'lm', se=FALSE, col='red') +
  scale_x_continuous(breaks = seq(0, 80, 5)) +
  scale_y_continuous(breaks = seq(0, 50, 5)) +
  labs(title = 'Relationship between Global and Euro
     x = 'Global Sales in millions (£)',
     y = 'EU Sales in millions (£)',
     col = 'Sales rank of Products',
     caption = 'Data:turtle_sales.csv')
geom_smooth()` using formula = 'y ~ x'
```



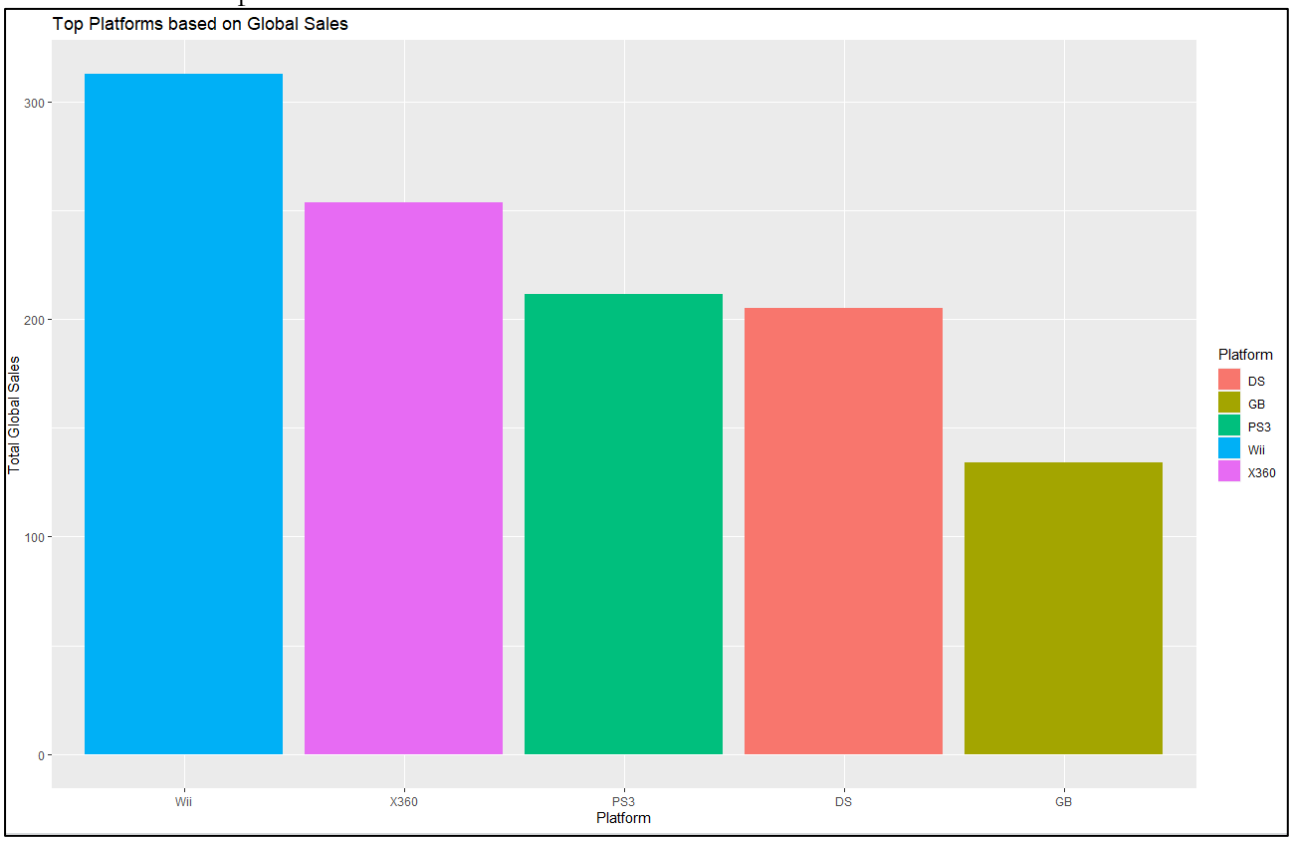

Relationship between Global and North American Sales
NA Sales in millions (£)
Global Sales in millions (£)
Data:turtle_sales.csv


Total Sales by Product in North America
Count
Total Sales in NA in millions £

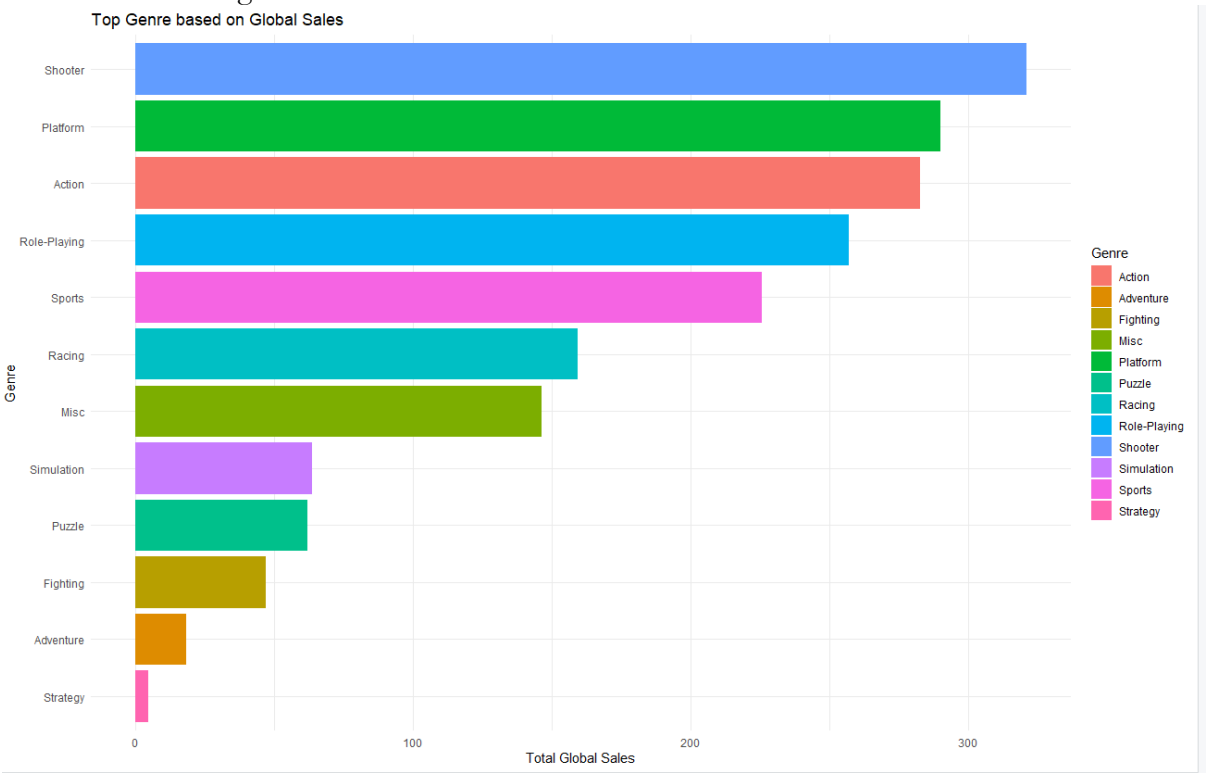


NA sales higher than EU Sales , NA seems to have a higher customer market for turtle games.

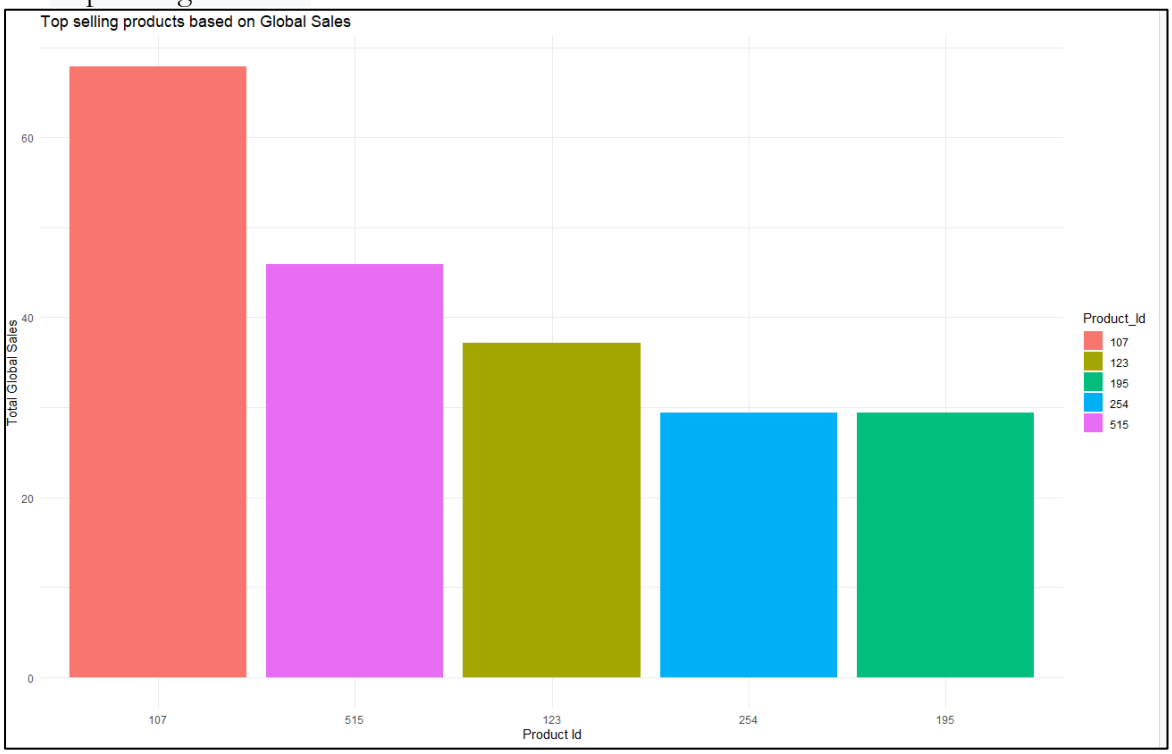
- Sales based on platform



- Sales based on genre



- Top Selling Products

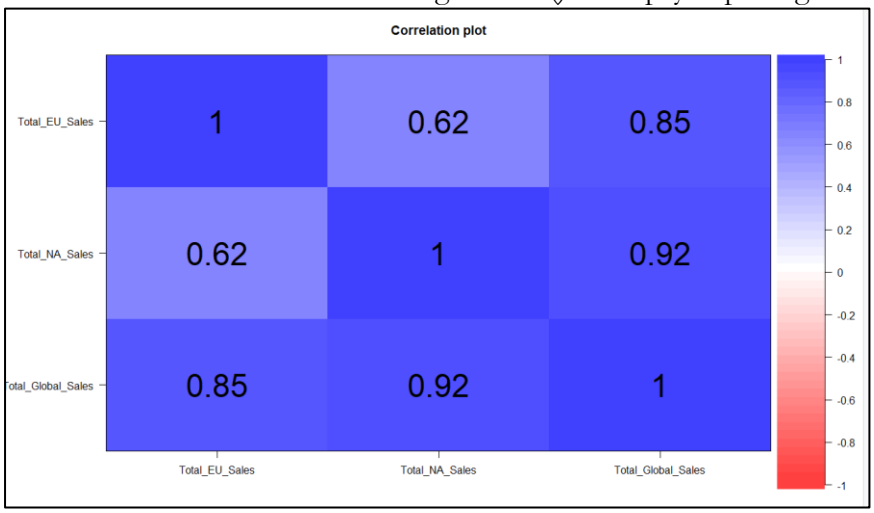


- Top selling product ids are 107, 515, 123, 254 and 195
- Top platforms are Wii, X360 PS3, Ds and GB
- Top genres are Shooter, Platform, Action, Role playing and Sports.

6.3 Insights

Simple linear models are created to determine if Global_sales can be predicted with the NA and EU sales.

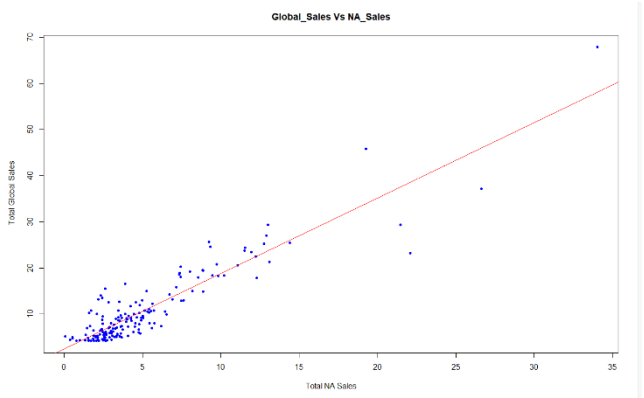
- Correlation matrix created using corPlot() from psych package.

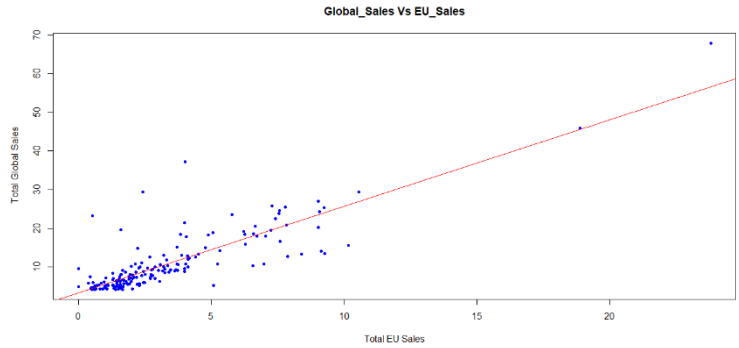


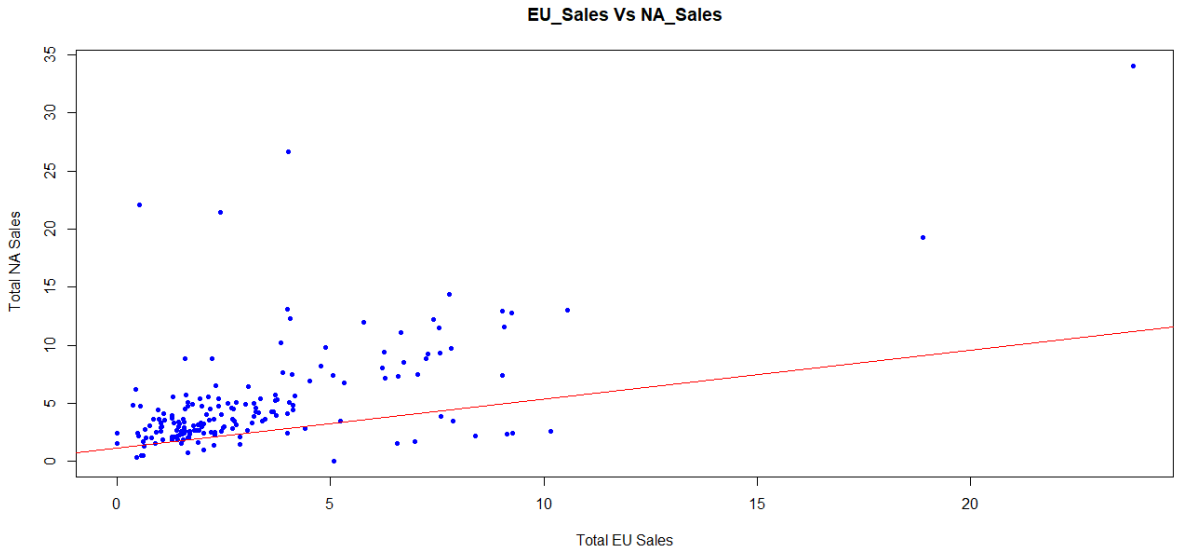

```
> # Global sales - NA Sales :
> model_NA <- lm(Total_Global_Sales ~ Total_NA_Sales, sales_df)
> # Summary of the model:
> summary(model_NA)

Call:
lm(formula = Total_Global_Sales ~ Total_NA_Sales, data = sales_df)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3417  -1.8198  -0.5933   1.4322  11.9345

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.45768    0.36961   6.649 3.71e-10 ***
Total_NA_Sales  1.63469    0.05435  30.079  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.266 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
```

- abline() used to determine coefficients of the model and add a line of best fit to the linear regression.

```
plot(sales_df$Total_NA_Sales, sales_df$Total_Global_Sales,
     main = 'Global_Sales Vs NA_Sales',
     xlab = 'Total NA Sales',
     ylab = ' Total Global Sales',
     col = 'blue', pch = 20)
# Add line of best fit
abline(coefficients(model_NA), col ='red')
```







R^2 of EUsales 72% and NASales 83.9% explains the variability in global sales. They are highly significant. The plot shows a positive correlation.

EU – NA have no correlation. R^2 =38%, not accurate fit.

- Multiple linear regression model is created between NA_sales + EU_sales to determine the Global_sales.

```
Call:
lm(formula = Total_Global_Sales ~ Total_EU_Sales + Total_NA_Sales,
    data = sales_df)

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 ***
Total_EU_Sales  1.19992    0.04672  25.682  < 2e-16 ***
Total_NA_Sales  1.13040    0.03162  35.745  < 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 coefficients of EU and NA sales have stars(***) - highly significant in determining the dependent variable(Global_Sales).The adjusted R^2 of 96.6% variability in global sales is explained by NA-EU_sales and the model can predict global sales.

7 Conclusion and Recommendations

- Spending score and Renumeration have a positive correlation with loyalty points.
- There are 5 different customer groups for targeted marketing campaigns
- The social data from customer reviews show positive sentiment
- There exists a strong correlation with between North America, Europe and global sales. NA has a higher overall sale.
- Popular genres Shooting and Action games can be can be developed on the popular platform Wii and X360 to further improve sales
- Address the negative reviews issues to avoid similar in future.

Further work:

- Outliers in data to be explored and understand the reason for the same
- Data to be analysed with outliers removed for higher accuracy in predictions.
- The gender and education of customers can be analysed to understand loyalty points relationships.
- In depth analysis of negative reviews to be done.