LSE Data Analytics Career Accelerator

Assignment 3: Predicting Future Outcomes

Presentation Structure

1. Business Context
2. Key Insights
3. Recommendations
4. Conclusion

1. Background and Business Context

Turtle Games

- Manufactures and sells products under its brand name.
- Sources and distributes products manufactured by other companies.
- Operates worldwide and has an international customer base.
- Produces a range of products, including video game, board games, books and toys.

1.1. Analytical Techniques

Python:

- The Report employed statistical regression analysis to determine the relationship between the loyalty score (i.e., based on the point value of the purchase) and spending, remuneration and age of customers, respectively.
- In analysing specific groups within the customer base, the Elbow and Silhouette methods enabled to determine the optimal number of clusters and fit the K-means Model.
- Polarity and frequency analysis were used to determine the customer sentiment towards the products manufactured and sold by the Turtle Games.

R / R Studio:

- The Report used Q–Q plots, Skewness and Kurtosis, and a Shapiro-Wilk test to explore, prepare and explain the normality of the dataset.
- The Report built and tested the strength of the simple linear regression and multiple regression models to determine the relationships between global sales, sales in North America and the European Union.
- Based on the given values for sales in North America and Europe, future values were predicted with the confidence interval.

1.2. Business Questions

REVIEWS

- How do customers accumulate loyalty points?
- How specific groups within the customer base can be used to target specific market segments?
- How can social data be used to inform marketing campaigns?

SALES

- What impact does each product have on sales?
- How reliable is the data used to make assumptions about sales relationships?
- What is/are the relationship(s) between North American, European and global sales?

2. Key Insights: Structure

REVIEWS DATASET [Python]

- 2.1. Linear Regression and OLS Model
- 2.2. Clustering and K-Means Model
- 2.3. Natural Language Processing (NLP)

SALES DATASET [R Studio]

- 2.4. Normality Tests
- 2.5. Simple Linear Regression Model
- 2.6. Multiple Linear Regression Model
- 2.7. Prediction with Confidence Interval

2.1. Key Insights: Linear Regression

- **Approach:** Linear Regression, Ordinary Least Squares (OLS) Model
- **Tool:** Python
- **Functions:**

```
# Independent variable.
x = reviews_new['spending score']

# Dependent variable.
y = reviews_new['loyalty points']
```

```
# OLS model and summary.
f = 'y ~ x'
reviews_model = ols(f, data = reviews_new).fit()

reviews_model.summary()
```

Linear Regression and OLS Model

Spending Score (x) vs Loyalty Points (y)

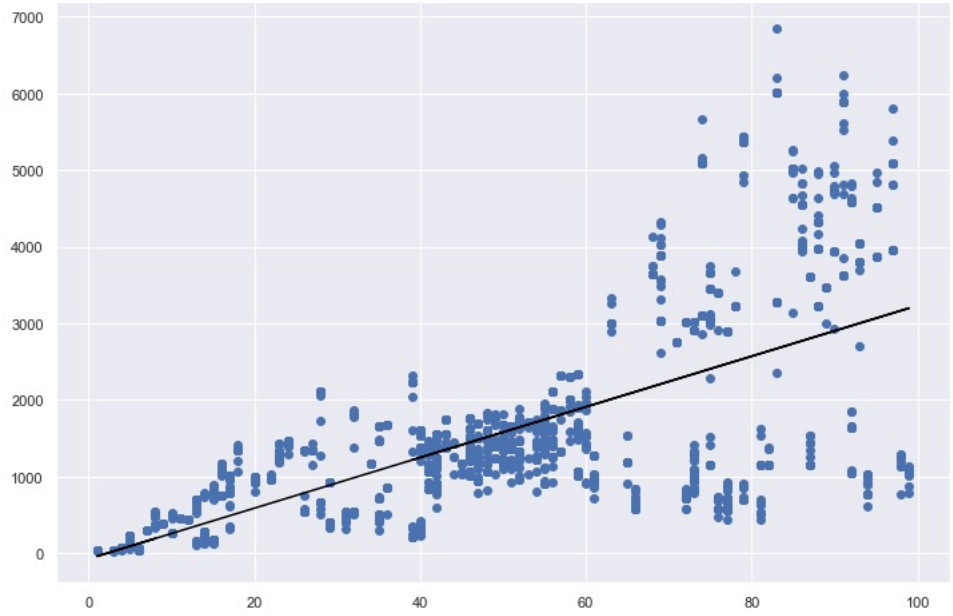
OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.
Date:	Sun, 11 Sep 2022	**Prob (F-statistic):**	2.92e-263
Time:	22:52:49	**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.

The regression model has low explanatory power (R-squared = 0.452) and low significance (P-value = 0.102).



The regression line shows a positive linear relationship, with the explanatory variable *x* (i.e., loyalty points) directly affecting the dynamics of the response variable *y* (i.e., spending score).

Linear Regression and OLS Model

Remuneration (x) vs Loyalty Points (y)

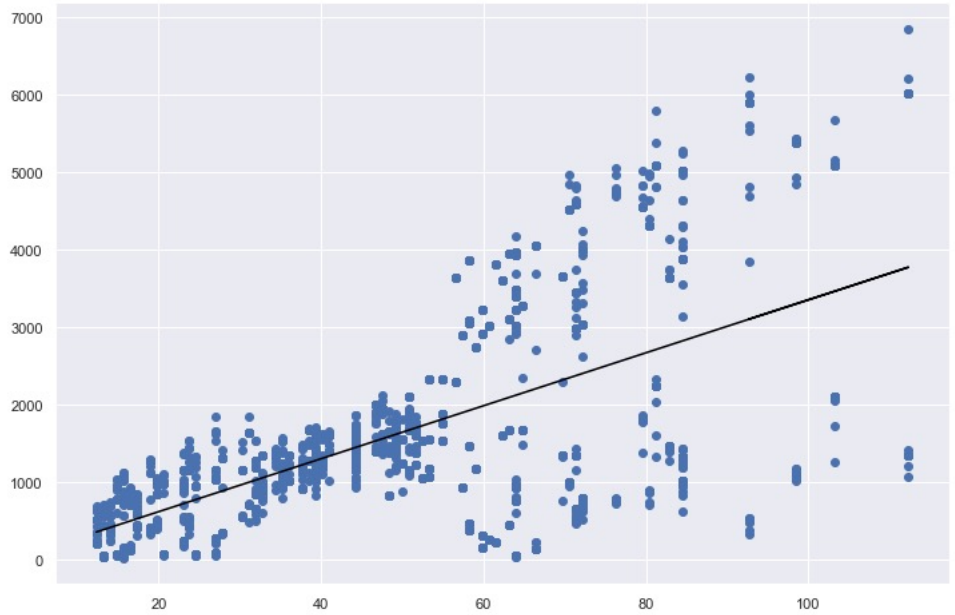
OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.380
Model:	OLS	**Adj. R-squared:**	0.379
Method:	Least Squares	**F-statistic:**	1222.
Date:	Sun, 11 Sep 2022	**Prob (F-statistic):**	2.43e-209
Time:	22:54:22	**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.

The regression model has low explanatory power (R-squared = 0.379) and low significance (P-value = 0.208).



The regression line shows a positive linear relationship, with the explanatory variable *x* (i.e., remuneration) directly affecting the dynamics of the response variable *y* (i.e., spending score).

Linear Regression and OLS Model

Age (x) vs Loyalty Points (y)

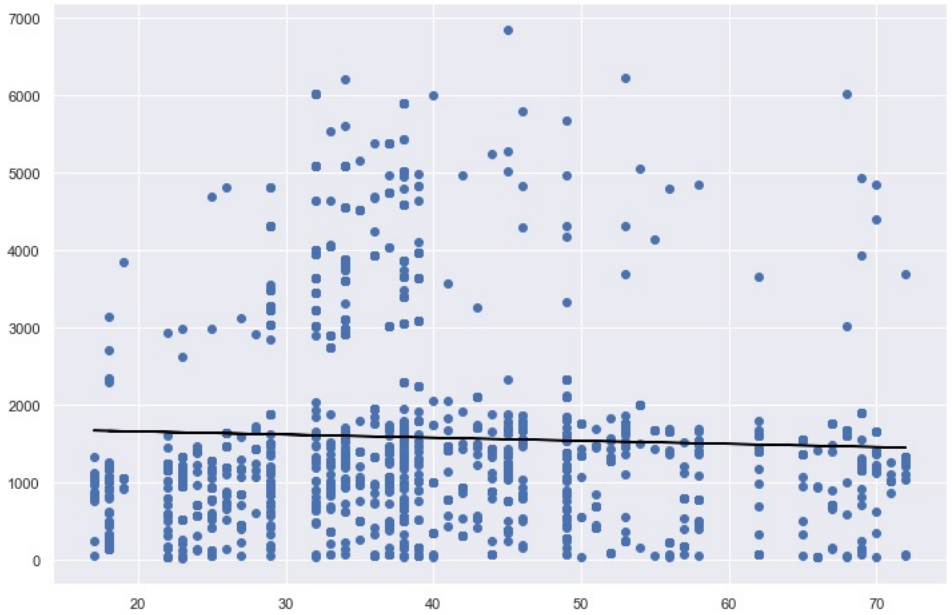
OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.606
Date:	Sun, 11 Sep 2022	**Prob (F-statistic):**	0.0577
Time:	22:54:45	**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.

The regression model has low explanatory power (R-squared = 0.001) and high significance (P-value = 0.000).



The regression line shows a slightly negative linear relationship, with the explanatory variable *x* (i.e., age) having a minor direct affect on the dynamics of the response variable *y* (i.e., spending score).

2.3. Key Insights: K-Means Clustering

- **Approach:** Elbow Method, Silhouette Method, K-means Clustering
- **Tool:** Python
- **Functions:**

```python
# Import the KMeans class.
from sklearn.cluster import KMeans

# Determine the optimal number of clusters: Elbow method.
cs = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++',
                    max_iter = 500, n_init = 10, random_state = 0)
    kmeans.fit(x)
    cs.append(kmeans.inertia_)

ax = plt.plot(range(1, 11), cs, marker='o')
plt.title("The Elbow Method")
plt.xlabel("Number of clusters")
plt.ylabel("CS")
plt.show()
```

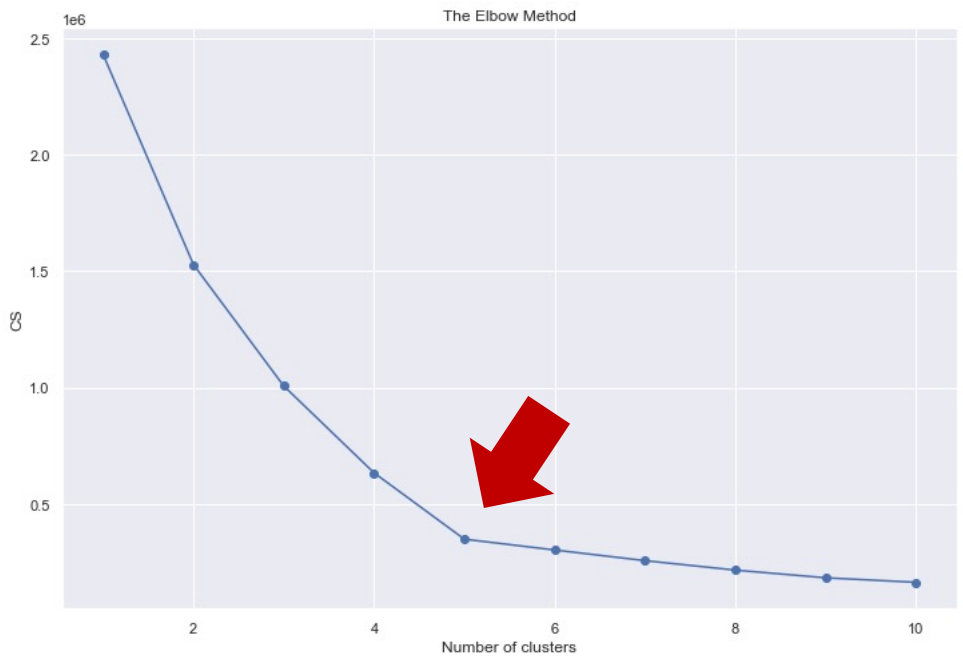
```python
# Import silhouette_score class from sklearn.
from sklearn.metrics import silhouette_score

# Determine the optimal number of clusters: Silhouette method.
sil = []
kmax = 10

for k in range(2, kmax+1):
    kmeans_s = KMeans(n_clusters = k).fit(x)
    labels = kmeans_s.labels_
    sil.append(silhouette_score(x, labels, metric = 'euclidean'))

# Plot the silhouette method.
ax = plt.plot(range(2, kmax+1), sil, marker='o')
plt.title("The Silhouette Method")
plt.xlabel("Number of clusters")
plt.ylabel("Sil")
plt.show()
```

```python
# Apply the final model:
# The optimal number of clusters is 5 as determined by the Elbow and Silhouette method.
kmeans = KMeans(n_clusters = 5, max_iter = 15000, init='k-means++', random_state=0).fit
clusters = kmeans.labels_
x['K-Means Predicted'] = clusters

# Plot the predicted.
ax = sns.pairplot(x, hue='K-Means Predicted', diag_kind= 'kde')
```

```python
# Visualising the clusters.
# Set plot size.
sns.set(rc = {'figure.figsize':(12, 8)})

ax = sns.scatterplot(x='remuneration' ,
                     y ='spending score',
                     data=x , hue='K-Means Predicted',
                     palette=['red', 'green', 'blue', 'black', 'orange'])
```

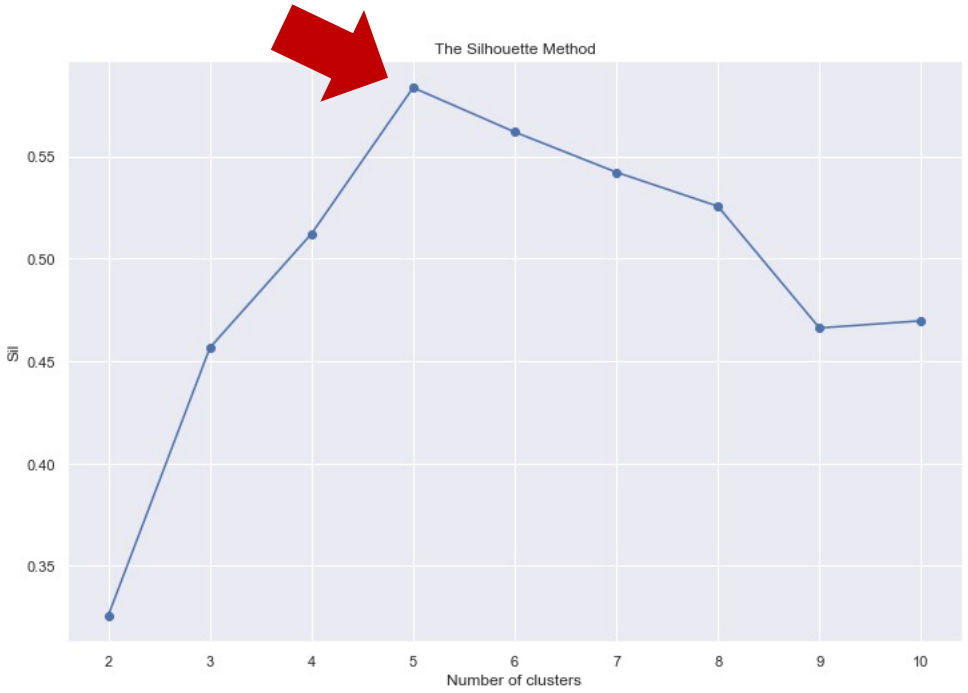
Determine the optimal number of clusters, *k*

Elbow Method



- Error measurement (SSE) is from the data point to the centroid.
- Optimal number of clusters, *k* is where the curve starts to strengthen out.

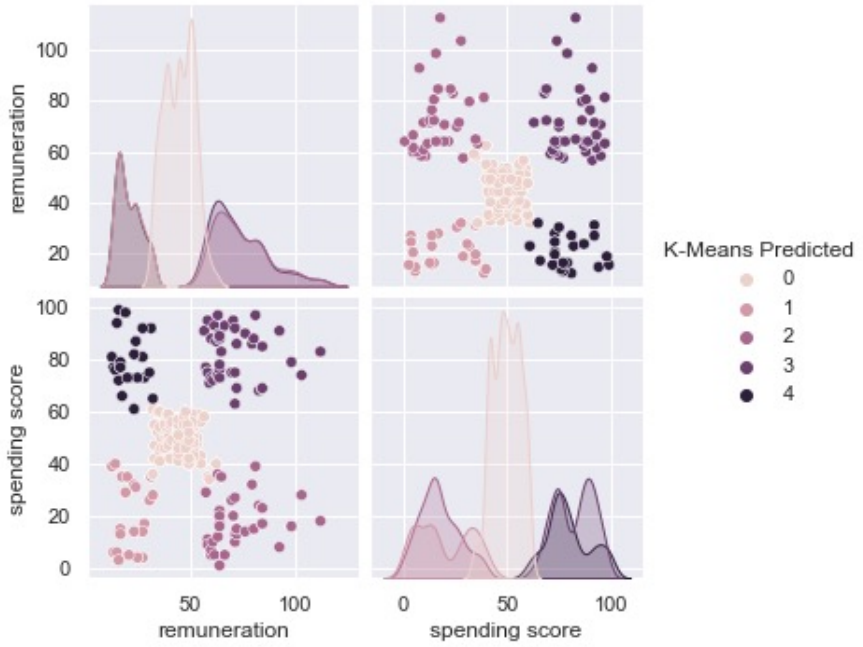
Silhouette Method



- Number of clusters, k averages intra-cluster distance.
- Optimal number of clusters, *k* is peak.

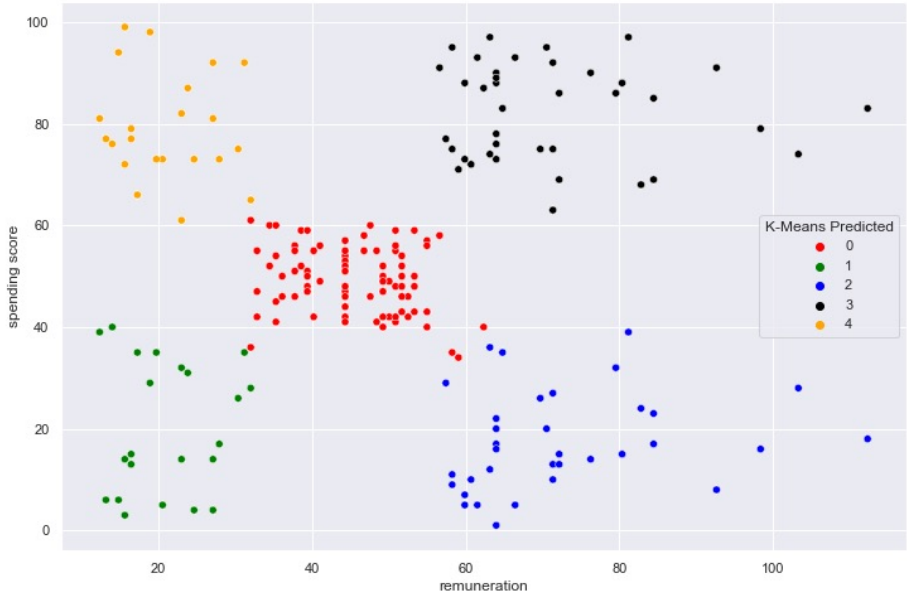
K-means model at value of clusters k = 5

K-means Predicted



After fitting the final model to the optimal number of clusters, the plotting of the K-Means Predicted shows that the reviews data does not fit normal distribution.

K-means Predicted, k = 5



	remuneration	spending score
0	12.30	39
1	12.30	81
2	13.12	6
3	13.12	77
4	13.94	40

There are 5 specific groups of clusters based on the respective remuneration and spending score values.

2.3. Key Insights: Natural Language Processing (NLP)

- **Approach:** Tokenisation, WorldCloud, Counter Function, Frequency Distribution, Polarity and Sentiment Analysis
- **Tool:** Python
- **Functions:**

```
# Tokenise the words.
df4['tokens_review'] = df4['review'].apply(word_tokenize)
df4['tokens_summary'] = df4['summary'].apply(word_tokenize)

df4 = df4.reset_index()

# Preview data.
df4.head()
```

```
# Review: Create a word cloud.
textt = " ".join(review for review in df4.review)
wordcloud = WordCloud(width = 1600, height = 900,
                background_color ='white',
                colormap='plasma',
                min_font_size = 10).generate(textt)
```

```
# Review: Plot the WorldCloud image.
ax = plt.figure(figsize = (16, 9), facecolor = None)
plt.imshow(wordcloud)
plt.axis('off')
plt.tight_layout(pad = 0)
plt.show()

all_tokens = []

for i in range(df4.shape[0]):
    all_tokens = all_tokens + df4['tokens_review'][i]
```

```
# Import the Counter class.
from collections import Counter

# Generate a DataFrame from Counter.
counts = pd.DataFrame(Counter(tokens2).most_common(15),
                      columns=['Word', 'Frequency']).set_index('Word')

# Preview data.
counts
```

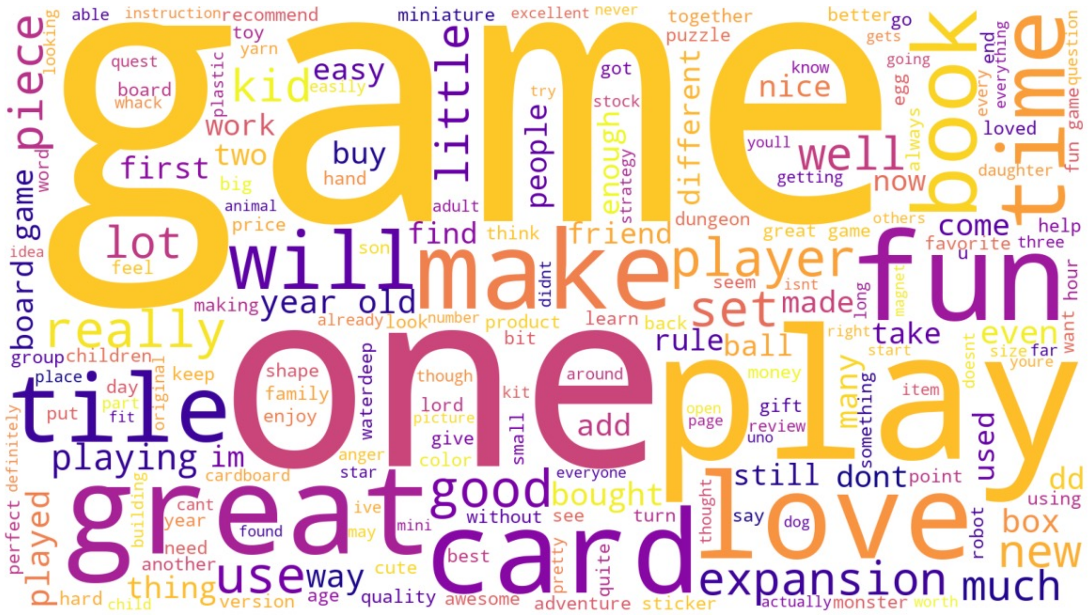
```
# Set the plot type.
ax = counts.plot(kind='barh', figsize=(16, 9), fontsize=12,
                 colormap ='plasma')
# Set the labels.
ax.set_xlabel('Count', fontsize=12)
ax.set_ylabel('Word', fontsize=12)
ax.set_title("Count of the 15 most frequent words",
             fontsize=20)
# Draw the bar labels.
for i in ax.patches:
    ax.text(i.get_width()+.41, i.get_y()+.1, str(round((i.get_width()), 2)),
            fontsize=12, color='red')
```

```
# Determine polarity of both columns.

# Populate a new column with polarity scores for each review.
df4['polarity_review'] = df4['review'].apply(generate_polarity)
df4['polarity_summary'] = df4['summary'].apply(generate_polarity)

# Preview the result.
df4.head()
```

```
# Review: Create a histogram plot with bins = 15.

# Histogram of polarity
num_bins = 15
ax = plt.hist(df4['polarity_review'], num_bins, facecolor='red', alpha=0.6)

# Histogram of sentiment score
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
mu, std = norm.fit(df4['polarity_review'])
p = norm.pdf(x, mu, std)

ax = plt.figure(figsize=(12, 18))
plt.plot(x, p, 'k', linewidth=2)
plt.xlabel('Polarity', fontsize=12)
plt.ylabel('Count', fontsize=12)
plt.title('Histogram of sentiment score polarity of reviews', fontsize=20)
plt.show()
```

Wordcloud Plot for Review and Summary

REVIEW WordCloud



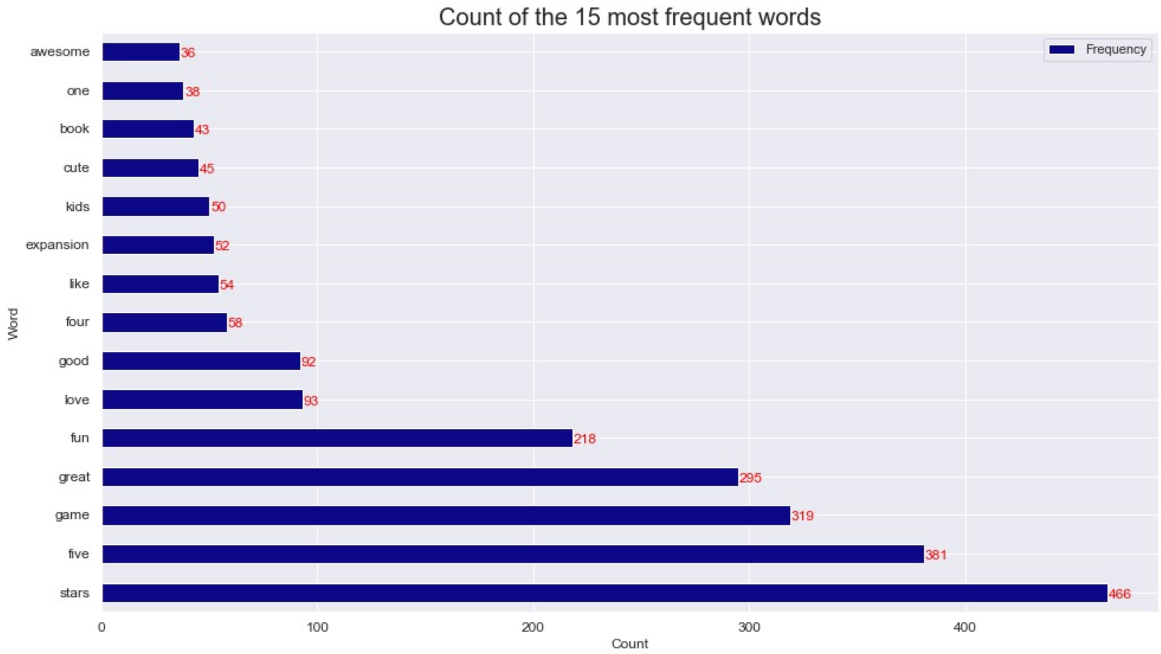
SUMMARY WordCloud

WordCloud Plot and Frequency Distribution

WorldCloud without stopwords

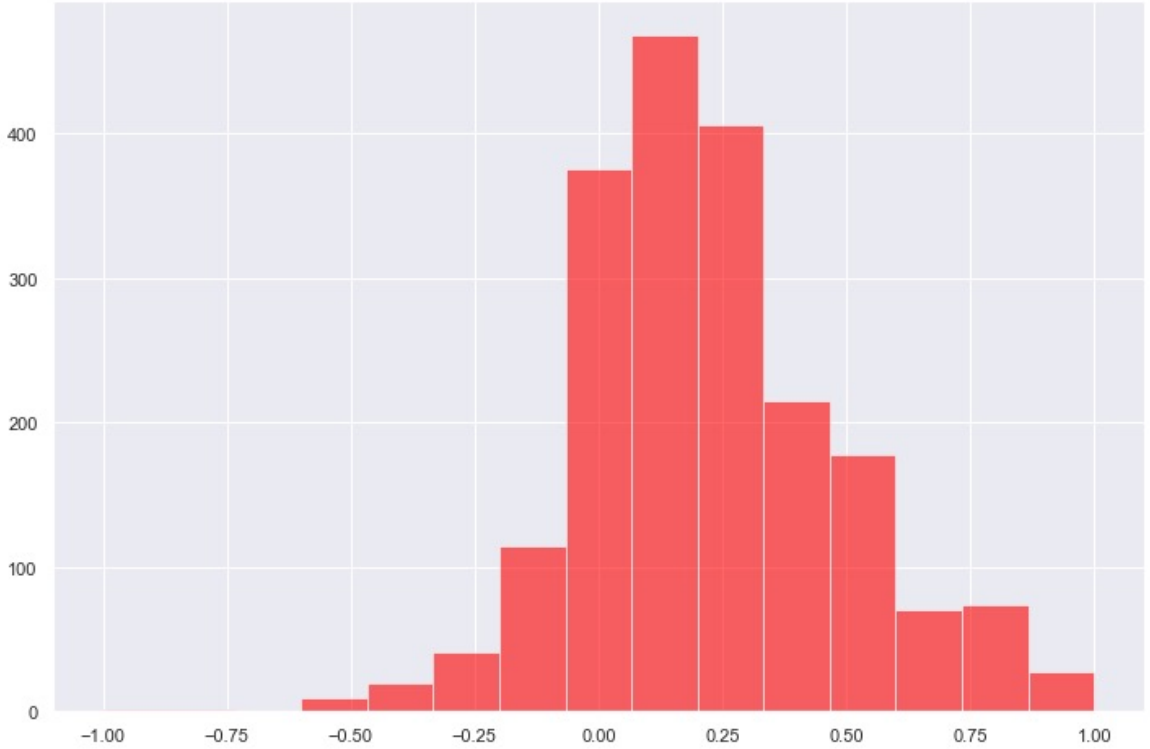
The words with the highest frequency counts include: **‘stars’**, **‘five’**, game, **‘great’**, **‘fun’**, **‘love’**, **‘good’**, four, **‘like’**, expansion, kids, **‘cute’**, book, one, **‘awesome’**.

Top-15 most frequent words

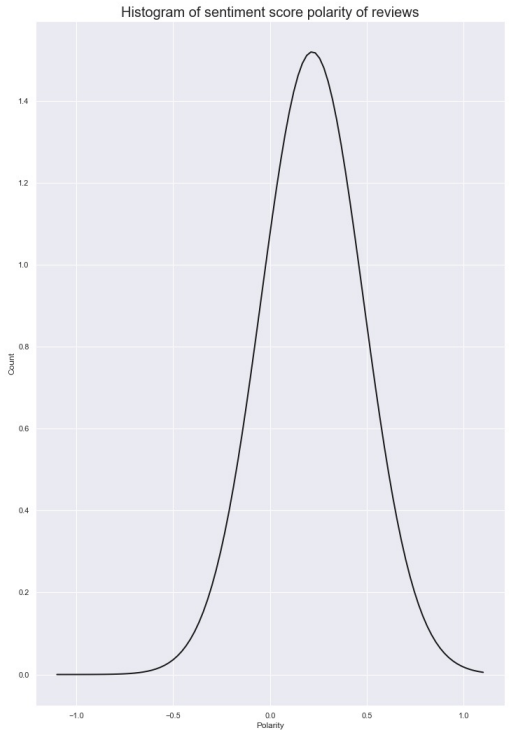


REVIEW: Polarity and Sentiment

Histogram of polarity



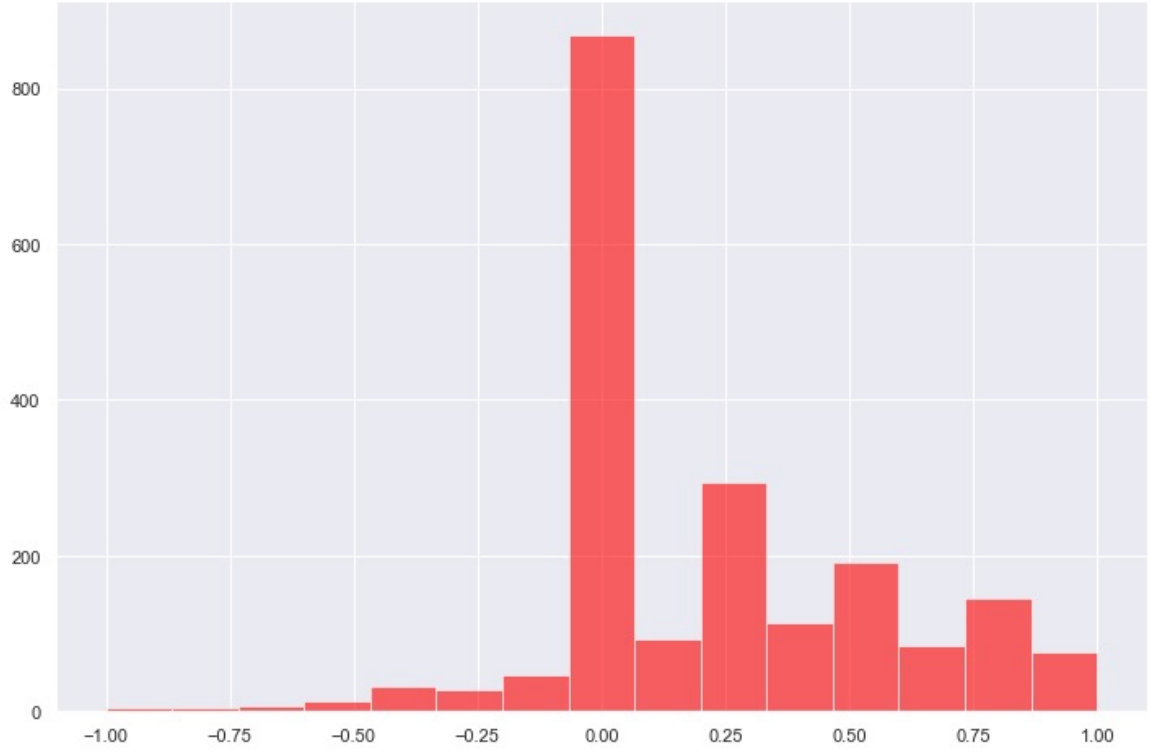
Histogram of sentiment scores



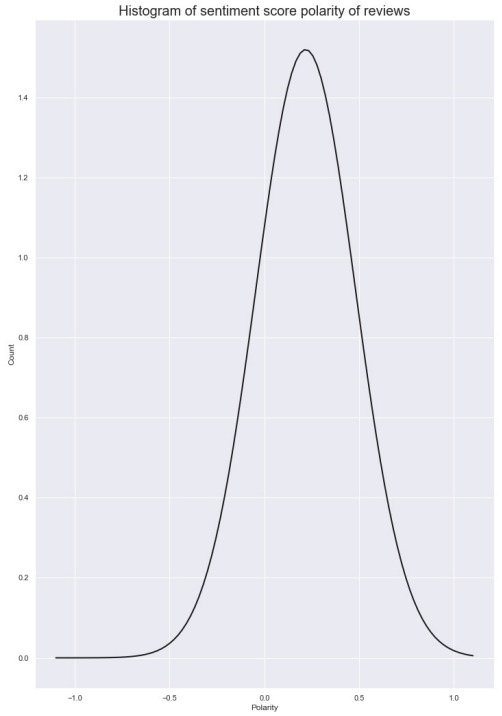
Left-skewed distribution
Mean < Mode ~ 0.25

SUMMARY: Polarity and Sentiment

Histogram of polarity



Histogram of sentiment scores



Left-skewed distribution
Mean < Mode ~ 0.25

2.4. Key Insights: Normality Tests

- **Approach:** Normal Q–Q Plots, Shapiro-Wilk Test, Skewness and Kurtosis, Correlation
- **Tool:** R Studiuo
- **Functions:**

```
# 3. Determine the normality of the data set.

## 3a) Create Q-Q Plots
# Create Q-Q Plots.
qqnorm(combined_sales_per_product$Total_Sales)
qqline(combined_sales_per_product$Total_Sales)
```

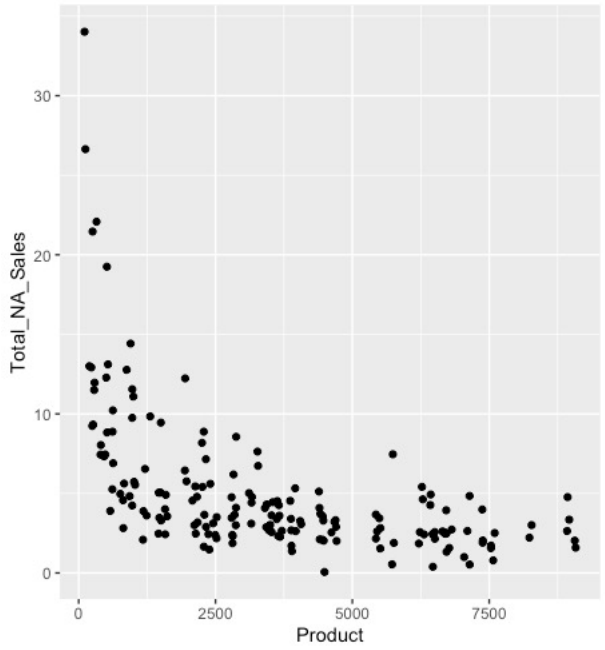
```
# Perform Shapiro-Wilk test.
shapiro.test(combined_sales_per_product$Total_Sales)
```

```
## 3c) Determine Skewness and Kurtosis
# Skewness and Kurtosis.
skewness(combined_sales_per_product$Total_Sales)
kurtosis(combined_sales_per_product$Total_Sales)
```

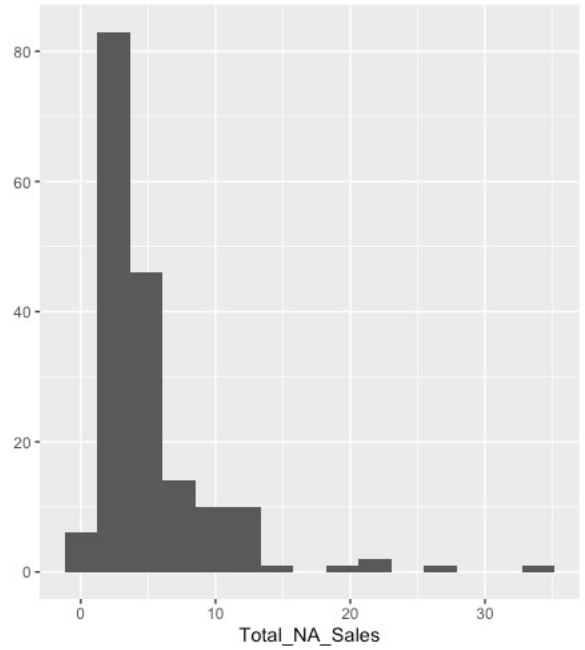
```
## 3d) Determine correlation
# Determine correlation.
cor(combined_sales_per_product$Product, combined_sales_per_product$Total_Sales)
```

Aggregate Impact of Each Product on Sales

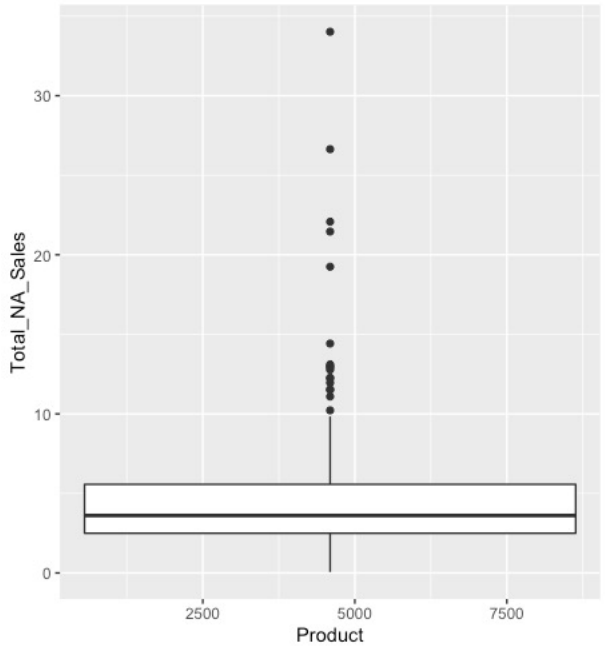
Sales in North America (x) vs Product (y)



The scatterplot shows several outliers in the dataset.



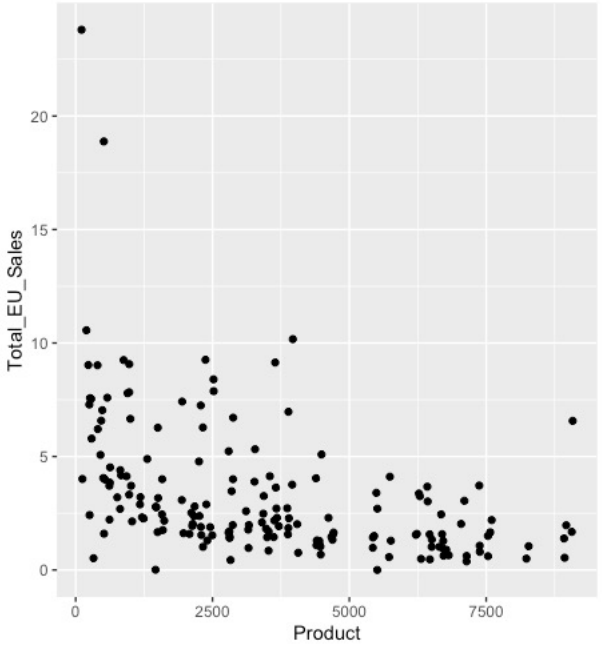
The histogram shows that aggregate sales in North America are distributed across 10 bins. The distribution is skewed to the right.



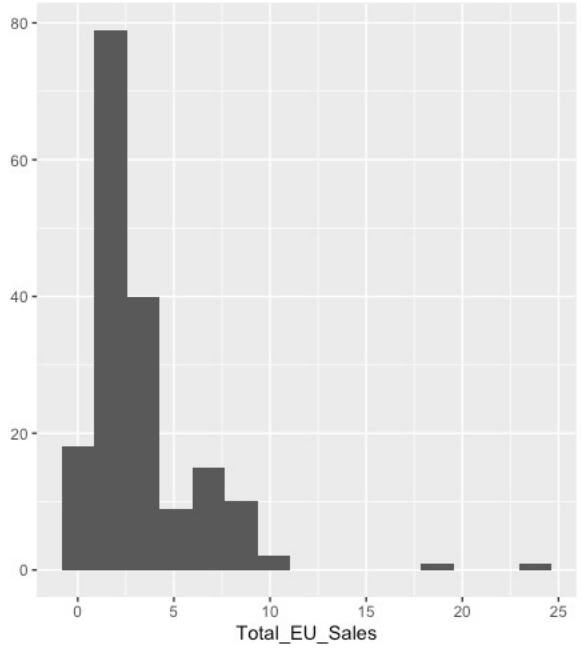
The boxplot shows several outliers in the dataset.

Aggregate Impact of Each Product on Sales

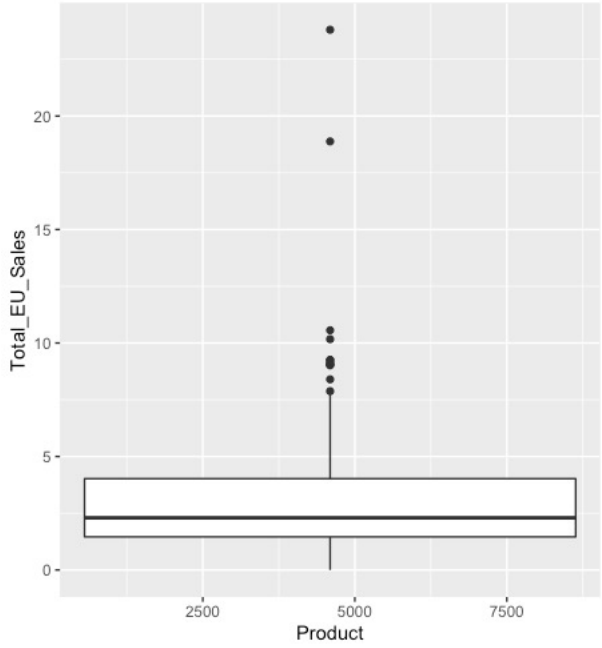
Sales in Europe (x) vs Product (y)



The scatterplot measuring the relationship between Product shows several outliers in the dataset.



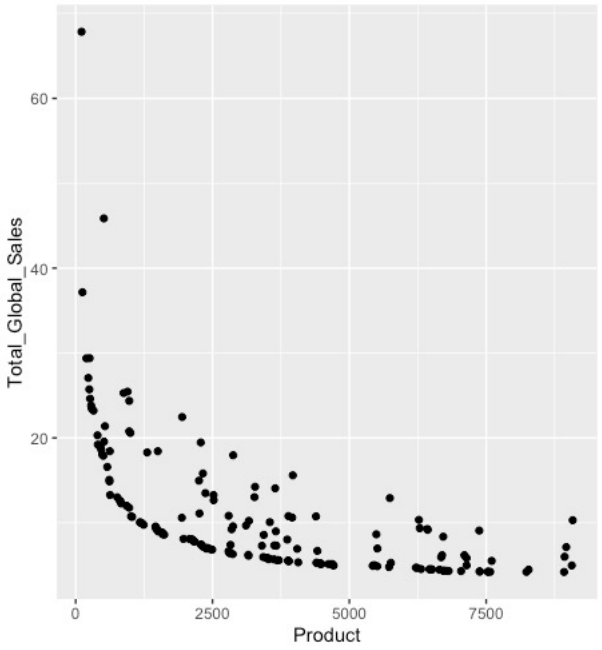
The histogram shows that aggregate sales in the European Union are distributed across 10 bins. The distribution is skewed to the right.



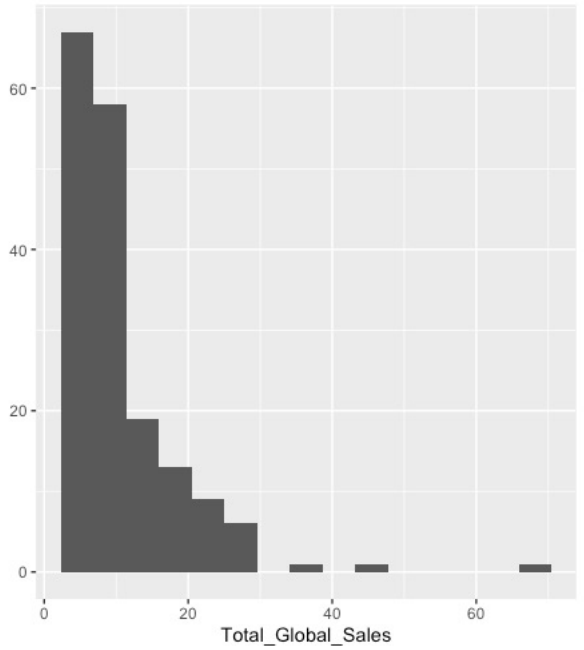
The boxplot shows several outliers in the dataset.

Aggregate Impact of Each Product on Sales

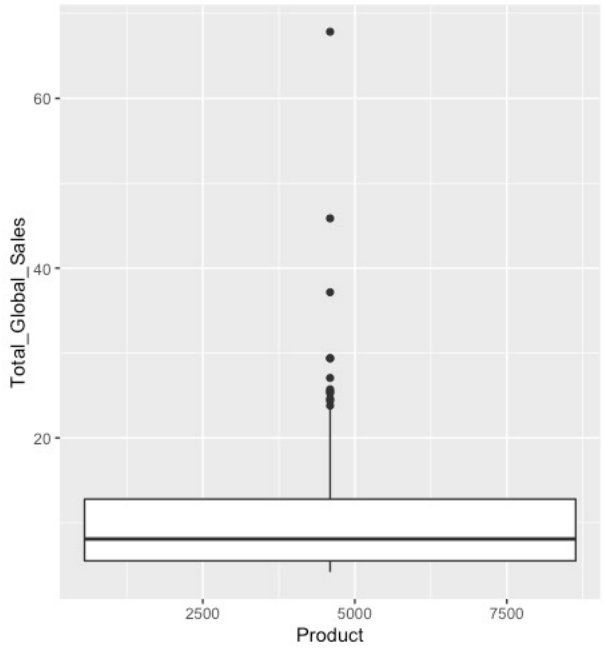
Global Sales (x) vs Product (y)



The scatterplot measuring the relationship between Product shows several outliers in the dataset.



The histogram shows that aggregate sales in the world are distributed across 10 bins. The distribution is skewed to the right.



The boxplot shows several outliers in the dataset.

Relationship between Product ID and Total Sales

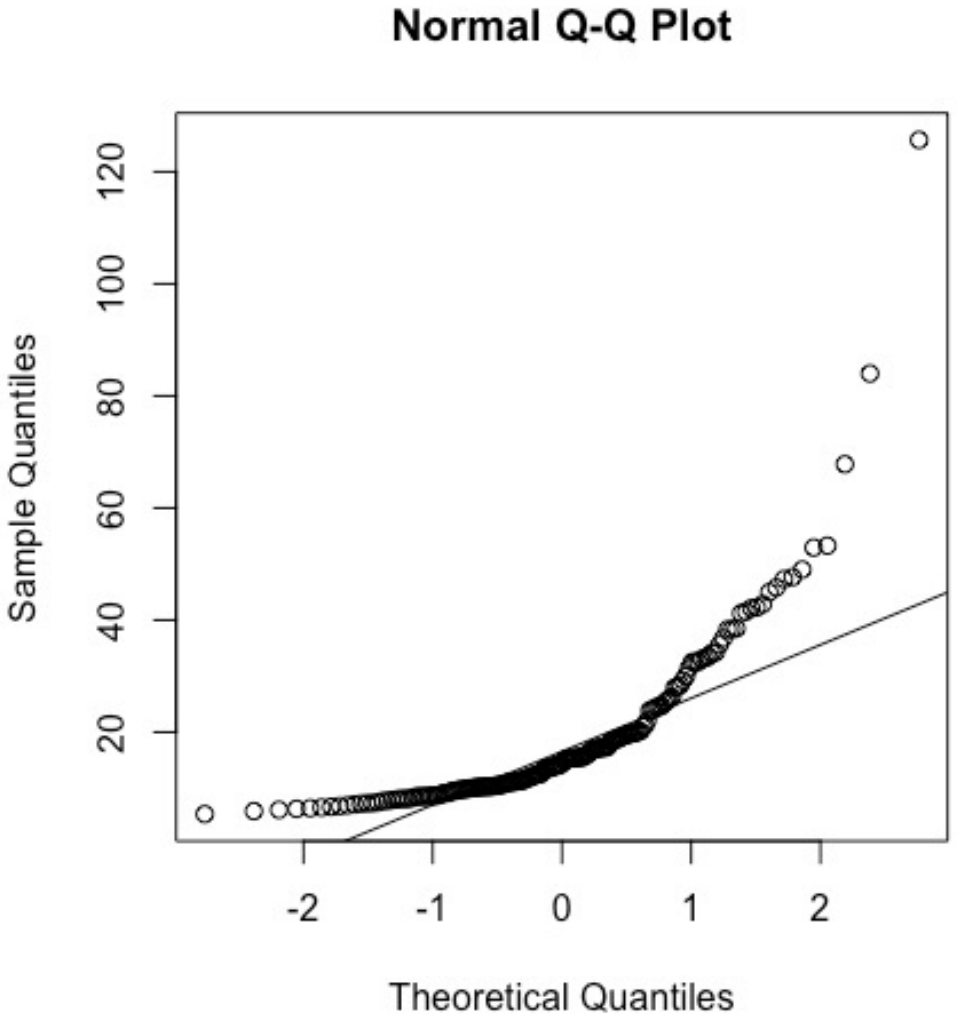
Normal Q–Q plots

- The quantile-quantile plot (Q–Q plot) compares the two probability distributions by plotting their standard normal quantiles against the quantiles of the points from the standard normal distribution.
- The Q-Q plot measures how close to normal the standardised residuals are.

→ The Normal Q–Q Plot indicates that the standardised residuals of the dataset deviate from the normal.

- Normal Q-Q plot could be used to evaluate how well the distribution of a dataset matches a standard normal (Gaussian) distribution. The linearity of the points suggests that the data are normally distributed.

→ The Normal Q–Q Plot shows that the Total Sales data is not normally distributed given no strict linearity of the points.



Relationship between Product ID and Total Sales

- **The Shapiro-Wilk Test:**

W = 0.71023

→ Since the W Statistic has a relatively large value, the relationship between Product and Total Sales is close to normal distribution.

```
        Shapiro-Wilk normality test

data:  combined_sales_per_product$Total_Sales
W = 0.71023, p-value < 2.2e-16
```

- **Skewness and Kurtosis:**

Skewness() = 3.14

→ Based on the relative size of the two tails, the distribution of data is not symmetric.

Kurtosis () = 18.54

→ Based on the combined size of the two tails, the dataset is not symmetric.

→ The relationship between Product and Total Sales is not normally distributed.

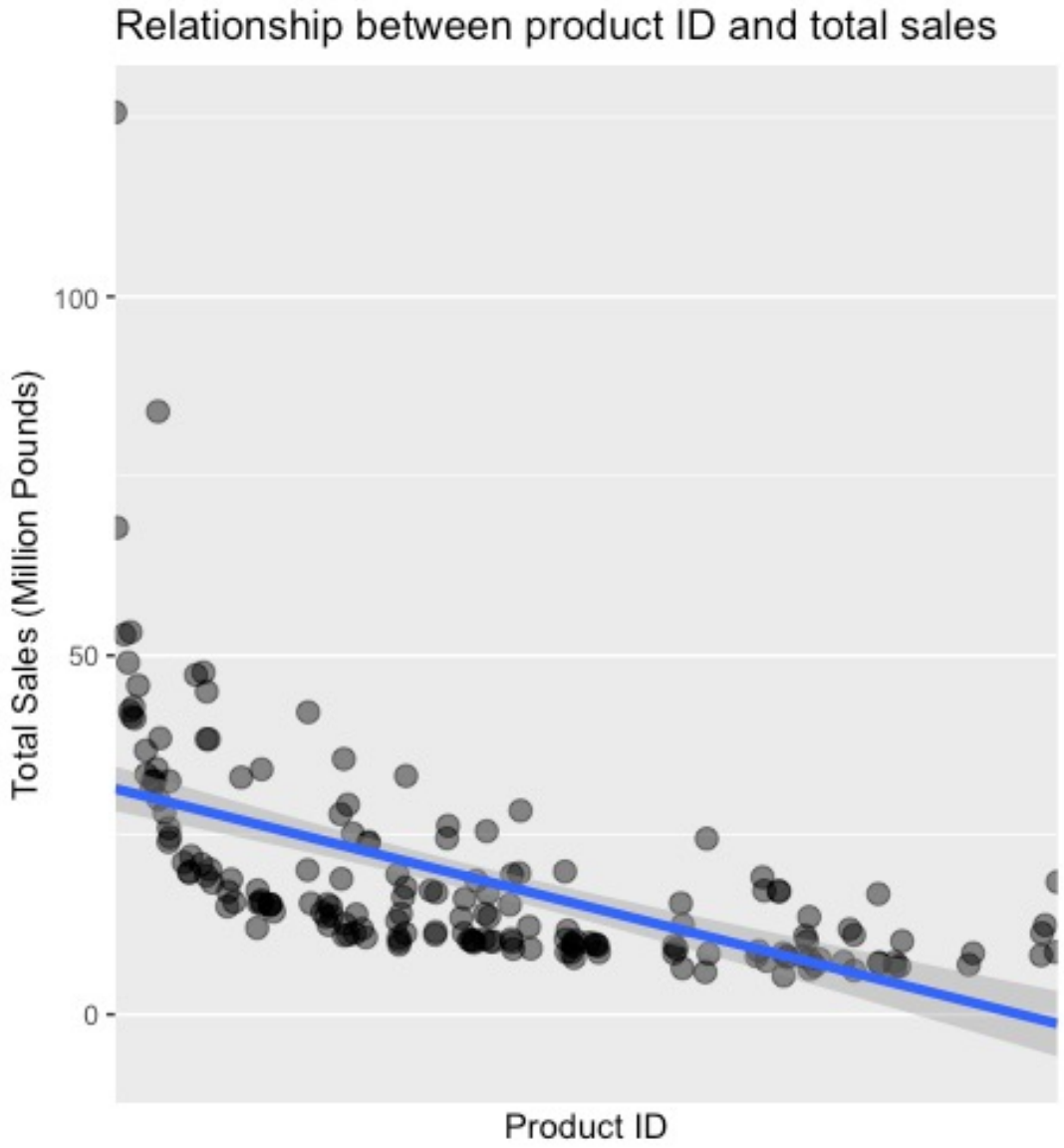
```
> ## 3c) Determine Skewness and Kurtosis
> # Skewness and Kurtosis.
> skewness(combined_sales_per_product$Total_Sales)
[1] 3.141885
> kurtosis(combined_sales_per_product$Total_Sales)
[1] 18.5417
```

Relationship between Product ID and Total Sales

Correlation:

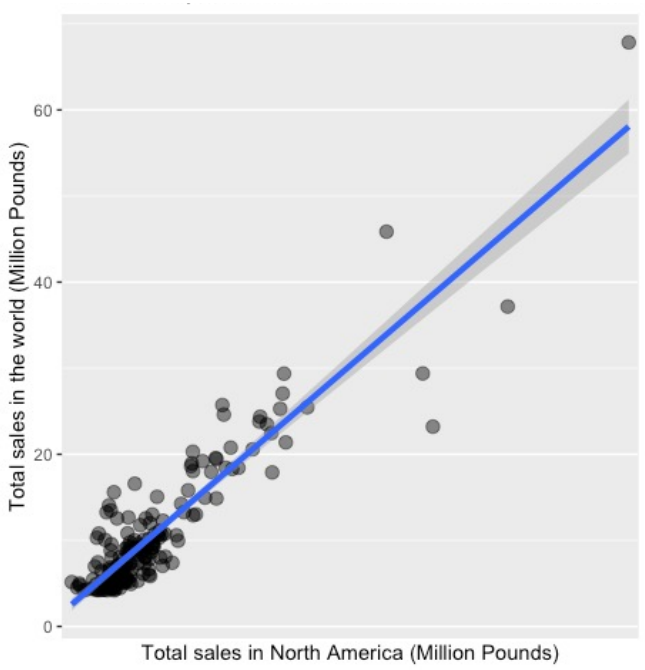
→ Product and Total Sales variables are negatively correlated.

```
> ## 3d) Determine correlation
> # Determine correlation.
> cor(combined_sales_per_product$Product, combined_sales_per_product$Total_Sales)
[1] -0.5876374
```



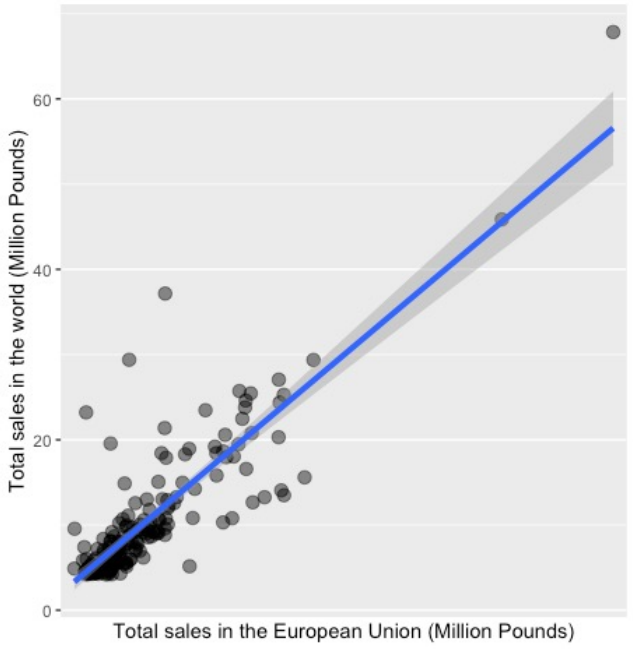
Relationship between Global Sales, Sales in North America and Europe

Global Sales vs Sales in North America



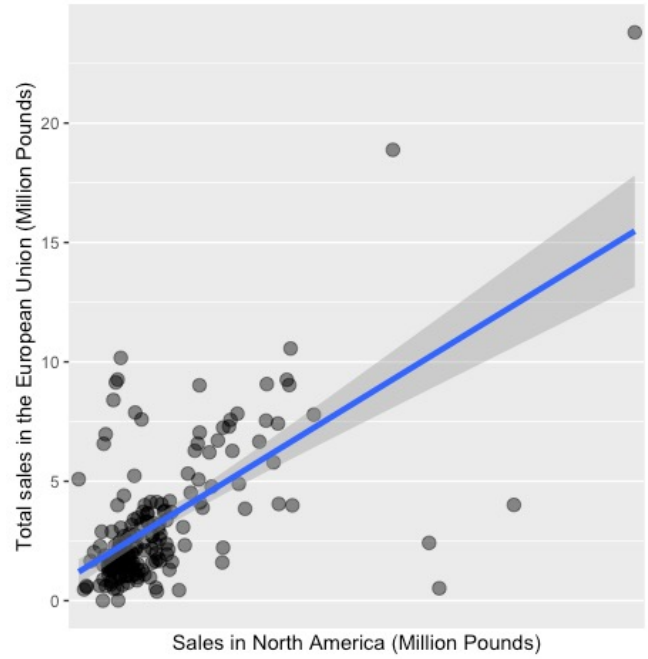
The Q-Q plot shows that Global Sales and Sales in North America data does not match standard normal distribution.

Global Sales vs Sales in the European Union



The Q-Q plot shows that Global Sales and Sales in Europe data does not match standard normal distribution.

Sales in the European Union vs Sales in North America



The Q-Q plot shows that Sales in Europe and North America data does not match standard normal distribution.

2.5. Key Insights: Simple Linear Regression Model

- **Approach:** Simple Regression Model
- **Tool:** R Studiuo
- **Functions:**

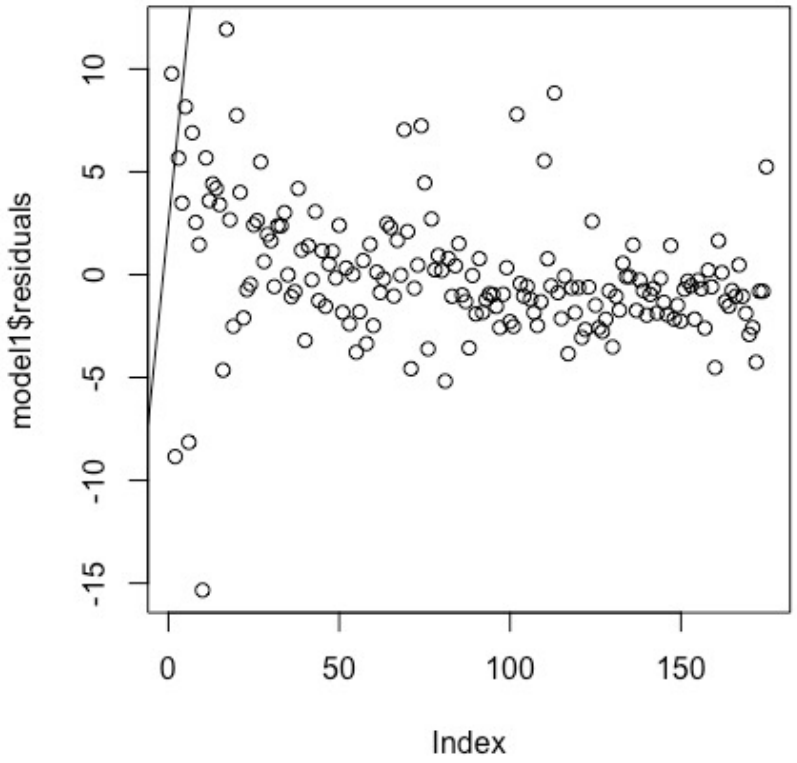
```
## 2a) Determine the correlation between columns
# Create a linear regression model on the original data.
cor(sales_per_product)
View(cor(sales_per_product))

### (1) Test the relationship between Global Sales and NA Sales
plot(sales_per_product$Total_NA_Sales, sales_per_product$Total_Global_Sales)

model1 <- lm(Total_Global_Sales ~ Total_NA_Sales, data = sales_per_product)
summary(model1)

### (2) Test the relationship between Global Sales and EU Sales
plot(sales_per_product$Total_EU_Sales, sales_per_product$Total_Global_Sales)

model2 <- lm(Total_Global_Sales ~ Total_EU_Sales, data = sales_per_product)
summary(model2)

### (3) Test the relationship between NA Sales and EU Sales
plot(sales_per_product$Total_NA_Sales, sales_per_product$Total_EU_Sales)

model3 <- lm(Total_EU_Sales ~ Total_NA_Sales, data = sales_per_product)
summary(model3)
```

MODEL 1: Total Global Sales ~ Total Sales in North America

The sum of squares error (SSE) = 1845.812
→ The forecasting model is not accurate.

The adjusted R-squared = 0.8385
→ The linear regression model is statistically significant.

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

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
```




```
> SSE1 = sum(model1$residuals^2)
> SSE1
[1] 1845.812
```

MODEL 2: Total Global Sales ~ Total Sales in Europe

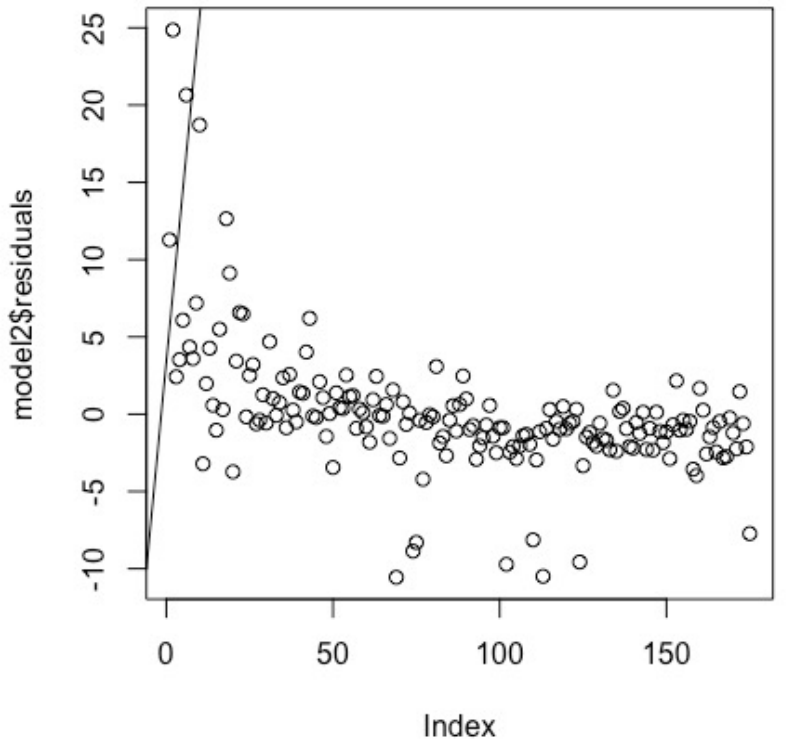
The sum of squares error (SSE) = 3217.935

→ The forecasting model is not accurate.

The adjusted R-squared = 0.7185

→ The linear regression model is statistically significant.

```
Call:
lm(formula = Total_Global_Sales ~ Total_EU_Sales, data = sales_per_product)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5583  -1.7530  -0.5371   0.9586  24.8556

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.3343     0.4787   6.965 6.57e-11 ***
Total_EU_Sales   2.2369     0.1060  21.099  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.313 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
```




```
> SSE2 = sum(model2$residuals^2)
> SSE2
[1] 3217.935
```

MODEL 3: Total Sales in Europe ~ Total Sales in North America

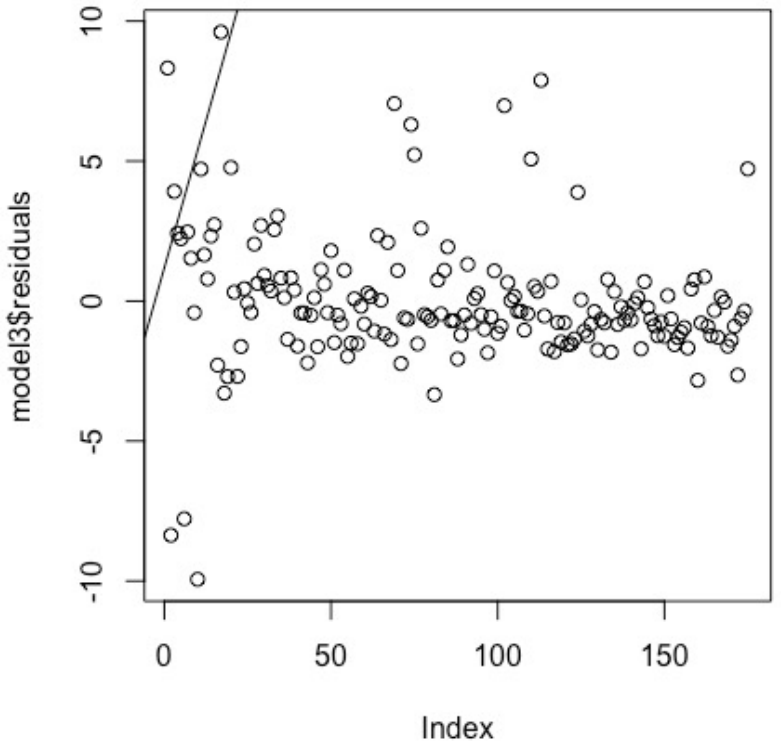
The sum of squares error (SSE) = 1016.824

→ The forecasting model is not accurate.

The adjusted R-squared = 0.382

→ The linear regression model is not statistically significant.

```
Call:
lm(formula = Total_EU_Sales ~ Total_NA_Sales, data = sales_per_product)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9391 -1.1930 -0.4267  0.7023  9.6102

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.17946    0.27433   4.299 2.85e-05 ***
Total_NA_Sales 0.42028    0.04034  10.419  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.424 on 173 degrees of freedom
Multiple R-squared:  0.3856,    Adjusted R-squared:  0.382
F-statistic: 108.6 on 1 and 173 DF,  p-value: < 2.2e-16
```




```
> SSE3 = sum(model3$residuals^2)
> SSE3
[1] 1016.824
```

2.6. Key Insights: Multiple Regression Model

- **Approach:** Multiple Linear Regression
- **Tool:** R Studiuo
- **Functions:**

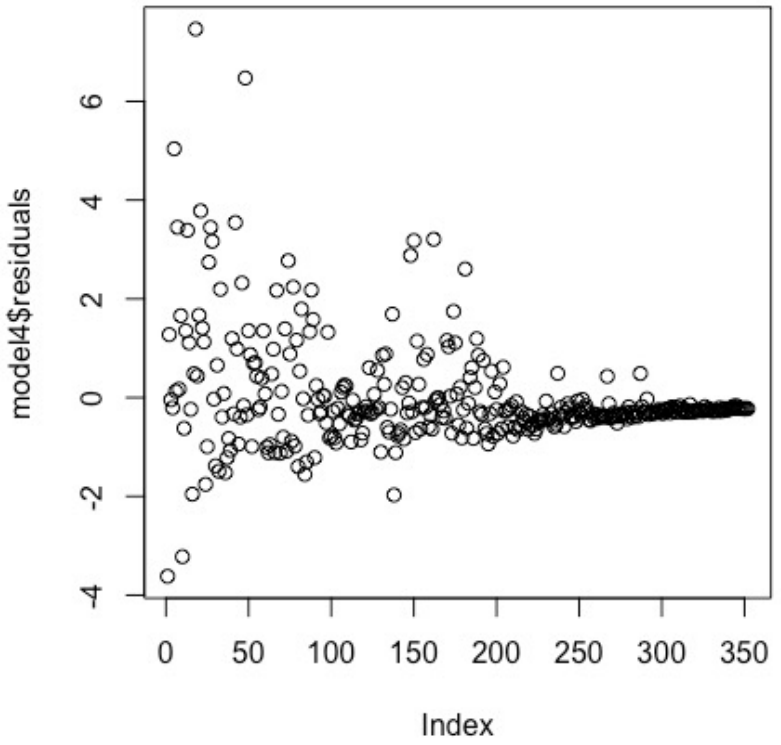
```
# 3. Create a multiple linear regression model

# Select only numeric columns from the original data frame.
sales_data_numeric <- subset(sales_data, select = c(NA_Sales, EU_Sales, Global_Sales))
View(sales_data_numeric)

# Multiple linear regression model.
cor(sales_data_numeric)
View(cor(sales_data_numeric))
model4 <- lm(Global_Sales ~ NA_Sales+EU_Sales, data = sales_data_numeric)

# Descriptve statistics
summary(model4)
```

MODEL 4: Global Sales ~ Sales in North America + Sales in Europe

P-value = 0.00453
→ The linear regression model is accurate.

The adjusted R-squared = 0.9685
→ The linear regression model is statistically significant.

```
Call:
lm(formula = Global_Sales ~ NA_Sales + EU_Sales, data = sales_data_numeric)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6186 -0.4234 -0.2692  0.0796  7.4639

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.22175    0.07760   2.858  0.00453 **
NA_Sales     1.15543    0.02456  47.047  < 2e-16 ***
EU_Sales     1.34197    0.04134  32.466  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.112 on 349 degrees of freedom
Multiple R-squared:  0.9687,    Adjusted R-squared:  0.9685
F-statistic:  5398 on 2 and 349 DF,  p-value: < 2.2e-16
```



2.7. Key Insights: Prediction with Confidence Interval

- **Approach:** Predict Function
- **Tool:** R Studiuo
- **Functions:**

```
## (1) NA_Sales_sum of 34.02 and EU_Sales_sum of 23.80.
given_values1 <- data.frame(NA_Sales=c(34.02), EU_Sales=c(23.80))
predict(model4, newdata=given_values1)

# Predict the values with confidence interval.
predict(model4, newdata=given_values1, interval = 'confidence')
```

2.7. Predict the values with confidence interval

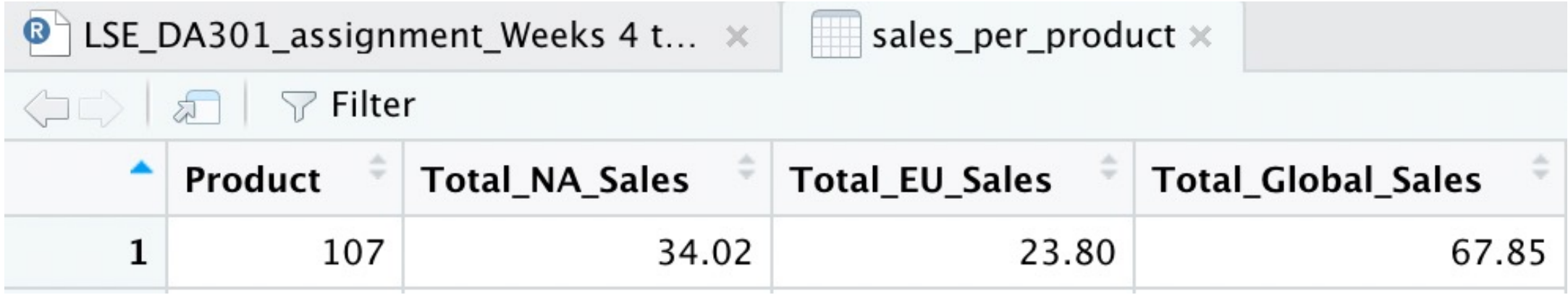
	Global_Sales = 67.85		
	FIT	LOWER LIMIT	UPPER LIMIT
NA_Sales = 34.02 EU_Sales = 23.80	71.46857	70.16242	72.77472
NA_Sales = 3.93 EU_Sales = 1.56	6. 856083	6.71842	6.993745
NA_Sales = 2.73 EU_Sales = 0.65	4.248367	4.102094	4.394639
NA_Sales = 2.26 EU_Sales = 0.97	4.124744	4.009122	4.260365
NA_Sales = 2.73 EU_Sales = 0.65	26.43157	25.41334	27.44979

2.7. Check the Accuracy of Prediction

PREDICTED VALUE = 71.47 Million Pounds

	Global_Sales = 67.85		
	FIT	LOWER LIMIT	UPPER LIMIT
NA_Sales = 34.02 EU_Sales = 23.80	71.46857	70.16242	72.77472

ACTUAL VALUE = 67.85 Million Pounds



	Product	Total_NA_Sales	Total_EU_Sales	Total_Global_Sales
1	107	34.02	23.80	67.85

2. Key Insights: Sum Up

REVIEWS DATASET [Python]

- 2.1. The regression models measuring the relationship between Loyalty Points (dependent variable) and the three independent variables, including Spending Score, Remuneration and Age, have low explanatory power and statistical significance. In relative terms, Spending Score has the greatest impact on the accumulation of Loyalty Points compared to other variables.
- 2.2. The K-means clustering model is not normally distriubuted. The optimal number of clusters was determined to be 5, which corresponds to the specific groups in the customer base arranged by their respective annual salary and spending scores. Given the non-normally distributed data, the remuneration vs spending clusters, alongside other specific groups within the Turtle Games customer base, should be used with caution when making extrapolations from sample customer base to targer specific market segments, such as elite or niche gaming products.
- 2.3. The WordCloud plot and frequency distribution computed based on the social data revealed that the customers' online reviews and summaries of these reviews tend to be slightly skewed toward positive sentiments. This might signal high customer satisfaction rate as well as the success of the previously implemented marketing strategies. To further boost sales performance, negative reviews should be considered in-detail since the current NLP model contains several inaccuracies in the sentiment and polarity analysis (namely, failure to distinguish some of the negative sentiments).

2. Key Insights: Sum Up

SALES DATASET [R Studio]

- 2.4. There is a negative linear relationship between Product and Sales, whereas Global Sales and Sales in North Americae and Europe have a positive linear relationship. The probability distributions for sales dataset do not match a standard normal (Gaussian) distribution, which might suggest that the sample data used to make assumptions about sales is either insufficient or unreliable.
- 2.5. The simple linear regression models measuring the relationship between Global Sales (dependent variable) and Sales in North America (independent variable) and Sales in Europe (independent variable) are statistcially significant, but not accurate in terms of forecasting. Although establishing the existence of a statistical relationship between Global Sales and Sales in North America and Europe, Models 1–3 might not be adequately accurate to make assumptions about sales relationships and thus a better-fit model is needed.

2. Key Insights: Sum Up

SALES DATASET [R Studio]

- 2.6. The multiple regression model measuring the relationship between Global Sales (dependent variable) and Sales in North America and Europe (independent variables) are statisticially significant and shows greater degree of accuracy. This indicates that the volume of product sales worldwide is directly affected by the changes in the volume of sales in North America and the European Union.
- 2.7. The predicted values for Global Sales using the multiple regression model (Model 4) roughly match the actual values recorded in the ‘sales_per_product’ dataframe. The predicted values lie within the confidence interval that changes in line with the changes in the given values for Sales in North America and Europe.

3.1. Recommendations

Based on the articulated insights, the following recommendations for the further analysis could be conducted using the models built in Python and R:

(1) Based on the identified factors that may affect optimal sales, suggest other variables for further analysis.

Other variables that affect sales performance may include customer satisfaction and retention rate, shipping and delivery costs and related factors that could directly impact the volume of global and regional sales. In the clustering analysis, more profound engagement with the gender and level of education demographics might yield more significant correlation patterns which, in turn, could be employed to target specific market segments.

(2) Suggest an alternative way to segment the products for further analysis. Can Turtle Games possibly introduce new variables for product segmentation?

The Turtle Games can use such new variables as geography/location of customers, customer product type preferences, customer engagement with the gaming community, etc. to create more nuanced product segmentation and thereby boost sales.

3.1. Recommendations

Based on the articulated insights, the following recommendations for the further analysis could be conducted using the models built in Python and R:

(3) Based on the outputs of the customer sentiment analysis, suggest any other analysis approaches or considerations that may provide more useful insights into areas of opportunity for improving sales performance.

One possible alternative to the customer sentiment analysis could be the collection of direct feedbacks or 'willingness to recommend' questionaries that reflect the level of satisfaction with the products quality per se as well as shipping services, client support, etc. Other viable alternative would be to access open-sourced social media data (for example, via Twitter or Facebook) to count and analyse shares and mentions of the gaming products manufactured and sold by the Turtle Games.

4. Conclusion

- The Report has built and tested data analytical models built to answer the business questions outlined by the Turtle Games.
- The proposed insights communicated to the Turtle Games stakeholders could inform the company's marketing and sales strategies to improve the sales performance.
- The examined patterns and predictions could be further enhanced understand how the specific demographics in the customer base relates to the borader market segments, as well as to study in detail the impact of product on sales and to discern the underlying relationships between global and regional sales.
- The limitations of the adopted data analytics techniques articualted in the Report should be carefully considered in the future statistical models used by the Turtle Games to improve the quality and efficiency of decision-making.