Assignment LSE_DA301_Advanced Analytics for Organisational Impact_C2_2023
Student: Elisa Basolu

Business Problem

Turtle Games aims to improve overall sales performance by leveraging customer trends. To achieve this goal, the company has identified key areas of inquiry, including understanding customer loyalty points, targeting specific market segments, utilising social data for marketing campaigns, assessing product impact on sales, ensuring data reliability, and evaluating the relationships between North American, European, and global sales.

Making predictions with regression

Linear Regression Modeling

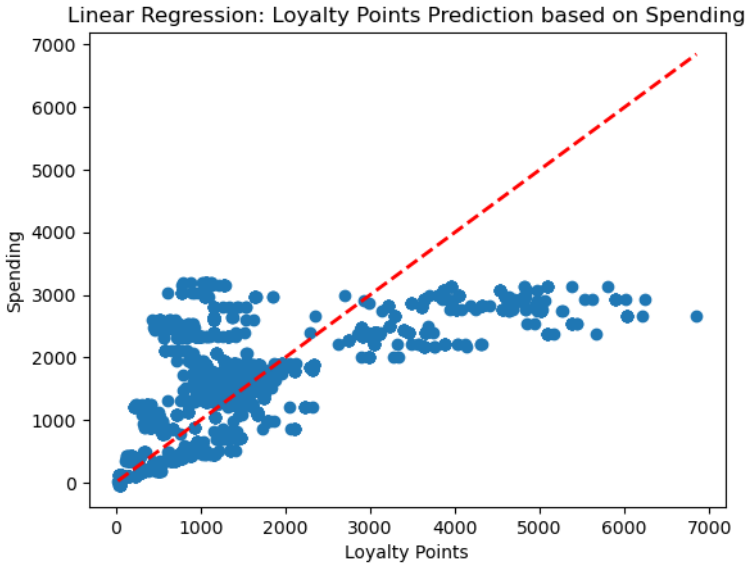
a. Spending vs. Loyalty Analysis

- Fit an Ordinary Least Squares (OLS) regression model and printed the regression summary.

```
print(summary)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         loyalty_points   R-squared:                       0.452
Model:                            OLS   Adj. R-squared:                  0.452
Method:                 Least Squares   F-statistic:                     1648.
Date:                Sun, 08 Oct 2023   Prob (F-statistic):          2.92e-263
Time:                        17:32:43   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]
------------------------------------------------------------------------------
const        -75.0527     45.931     -1.634      0.102    -165.129      15.024
spending_score    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.
==============================================================================
```

- Visualised the regression line against the actual data points.



Summary

R-squared: 0.452

This indicates that the spending score explains approximately 45.2% of the variance in loyalty points.

F-statistic: 1648

This high value suggests that the regression model fits the data well.
Prob (F-statistic): 2.92e-263
The model is statistically significant given this extremely low p-value.

Spending Score Insights:
Coefficient: 33.0617
Every unit increase in spending score corresponds to an average increase of 33.06 loyalty points.
t-value: 40.595
The high t-value confirms the statistical significance of the spending score as a predictor.
P>|t|: 0.000
The spending score is significant at standard levels.

Tests and Diagnostics for Spending Score:
Durbin-Watson: 1.191
Possible presence of autocorrelation.
Prob(JB): 2.67e-57
Residuals may not be normally distributed.
From the above we can gather that spending score is a strong predictor of loyalty points, with a notecable relationship suggesting that every unit increase in spending score leads to about 33.06 more loyalty points.

The same procedure was applied to determine the relationship between remuneration, age vs loyalty points.

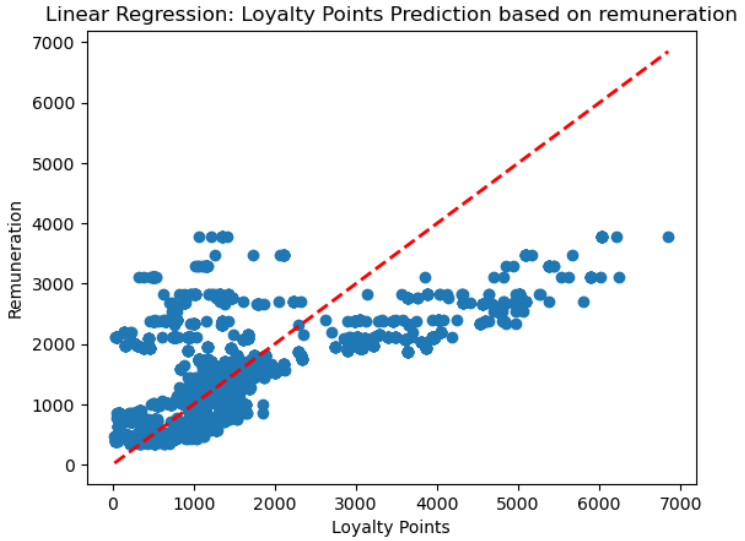
b. remuneration vs. loyalty_points

OLS Regression Results

Dep. Variable:	loyalty_points	R-squared:	0.380
Model:	OLS	Adj. R-squared:	0.379
Method:	Least Squares	F-statistic:	1222.
Date:	Sun, 08 Oct 2023	Prob (F-statistic):	2.43e-209
Time:	17:32:43	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]
const	-65.6865	52.171	-1.259	0.208	-168.001	36.628
remuneration	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.



Summary

R-squared: 0.380

Approximately 38% of the variance in loyalty points is explained by remuneration.

Remuneration Insights:

Coefficient: 34.1878

Every unit increase in remuneration corresponds to an average increase of 34.19 loyalty points.

Tests and Diagnostics for Remuneration:

Durbin-Watson: 3.622

Possible presence of autocorrelation.

Prob(JB): 1.30e-07

Residuals might not be normally distributed.

Overall, remuneration is a notable predictor of loyalty points. An increase in remuneration relates to an approximate 34.19 unit rise in loyalty points.

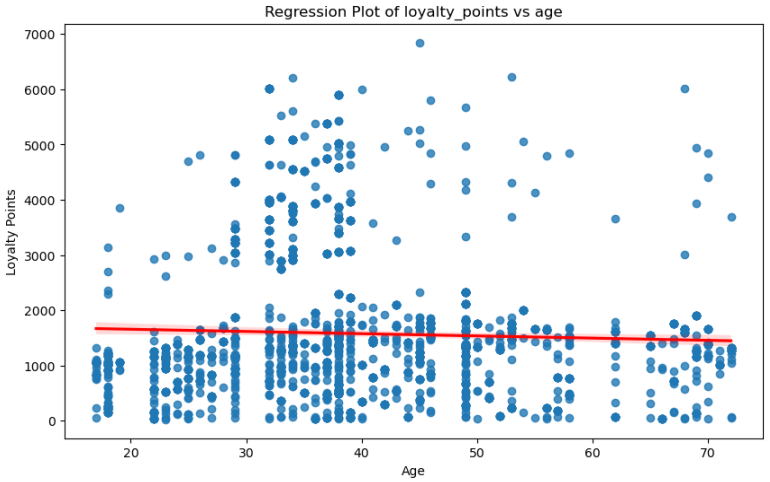
c. age vs. loyalty_points

OLS Regression Results

Dep. Variable:	loyalty_points	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	3.606
Date:	Sun, 08 Oct 2023	Prob (F-statistic):	0.0577
Time:	17:32:43	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]
const	1736.5177	88.249	19.678	0.000	1563.449	1909.587
age	-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.



Summary

R-squared: 0.002

Age explains a mere 0.2% of the variance in loyalty points, suggesting a weak relationship.

Age Insights:

Coefficient: -4.0128

Each additional year in age relates to an average decrease of 4.01 loyalty points.

Tests and Diagnostics for Age:

Prob(JB): 2.36e-204

Strong evidence that residuals aren't normally distributed.

The relationship between age and loyalty points is weak. An increase in age associates with a minor decrease in loyalty points, but the connection is only marginally significant.

Insights:

- Spending Score and Remuneration: Both are significant predictors of loyalty points. Spending score explains 45.2% of loyalty points variance, and remuneration accounts for 38%.
- Age: Weak link to loyalty points, explaining just 0.2% of its variance.
- Standard Errors: Lower values for spending score and remuneration models suggest reliable coefficient estimates, while age presents greater uncertainty.
- The data suggests that while spending and remuneration have strong relationships with loyalty points, age doesn't serve as a robust predictor in this context.

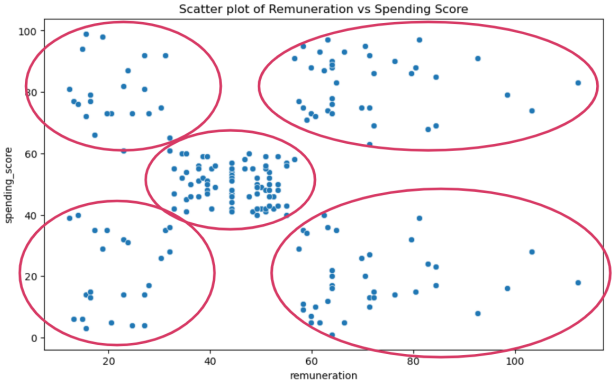
Make predictions with clustering

Objective

To identify clusters within customer reviews based on two features: remuneration and spending_score.

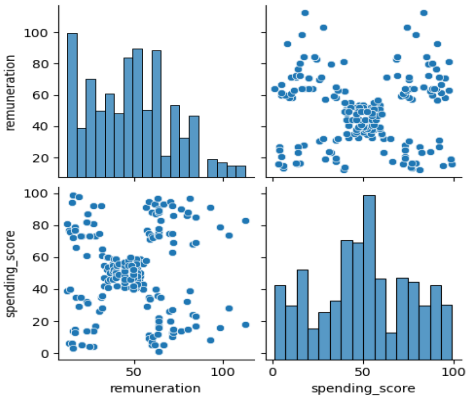
Data Visualization

- Plotted a scatter plot using Seaborn to visualise remuneration against spending_score.



This provides an initial visual inspection of potential clusters. The clusters (highlighted in the red circles) are visually identifiable, suggesting a natural segmentation present in the dataset.
This supposition was then confirmed by using the elbow and the silhouette method (see below).

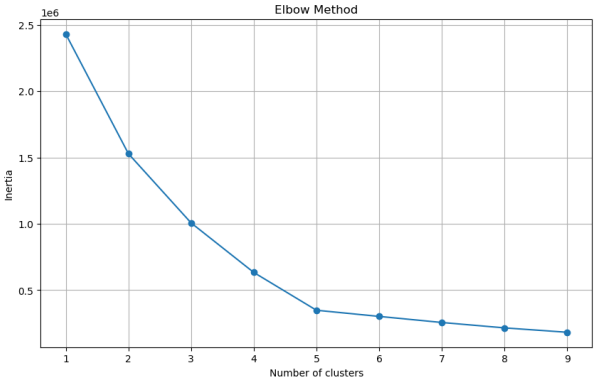
- Visualised the data distribution using a pairplot.



Remuneration histogram: Most individuals seem to have remuneration in the mid/low-range, with fewer people at the extreme high end.
Spending score histogram: Pronounced peak around the mid-range, indicating that a majority of individuals have a spending score in that range. There's also a smaller concentration of individuals with high spending scores.

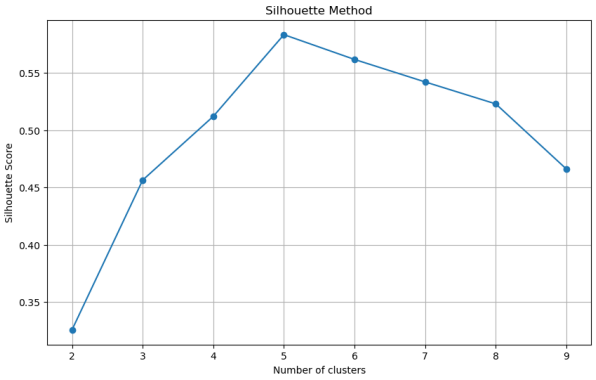
Cluster Determination:

- Used the Elbow method to identify a suitable number of clusters.



In the above plot, the elbow is at k=5 (i.e., the Sum of squared distances falls suddenly), indicating the optimal k for this dataset is 5.

- Utilised the Silhouette method to confirm the number of clusters.



The silhouette score is maximised at k = 5.

The optimal k from the Silhouette method aligns with the Elbow method and visual inspection which indicates a strong cluster structure.

In the jupyter notebook k equal to 4 and 6 was also tested but k=5 remained the optimal choice.

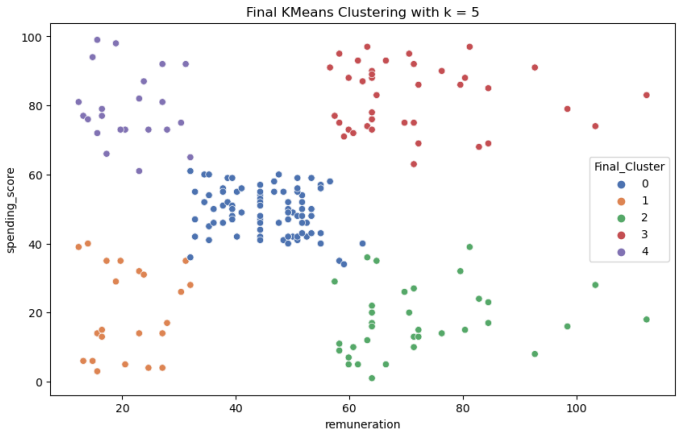
Model Application:

- Fit a KMeans clustering model with the optimal number of clusters (k=5).
- Predicted clusters for the data.
- Checked the number of observations per cluster and calculated their respective percentages (below).

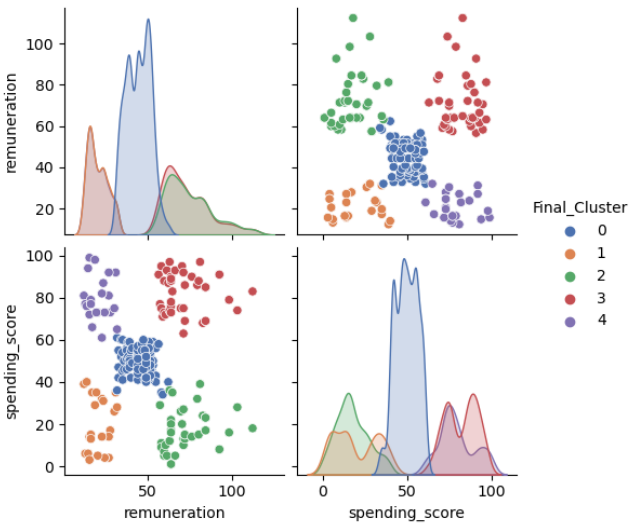
```
0    774
3    356
2    330
1    271
4    269
Name: Final_Cluster, dtype: int64

0    38.70
3    17.80
2    16.50
1    13.55
4    13.45
Name: Final_Cluster, dtype: float64
```

Cluster Visualisation:

- Plotted a scatter plot highlighting the clusters.



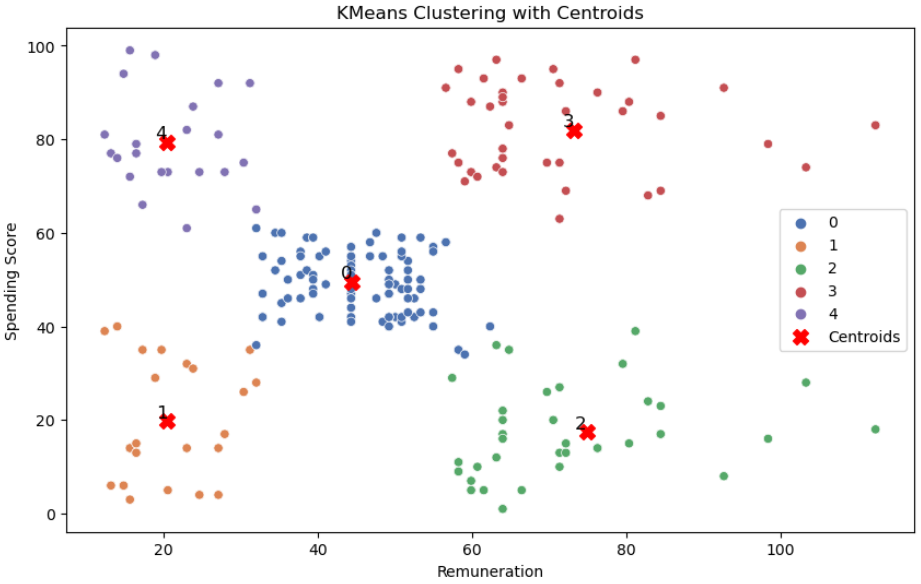
- Saved the dataframe with cluster assignments to a CSV file.
- Visualised the clusters using a pairplot.



The density plots for remuneration (top left) provides an understanding of the distribution of 'remuneration' for each cluster. Cluster 0 (blue) has a higher concentration in the mid to high remuneration range, while Cluster 1 (orange) seems to be focused on the lower to mid remuneration range.
The density plots for spending score (bottom right) gives insights into the spending score distribution across clusters. Cluster 3 (red) dominates the high spending score range, whereas Cluster 2 (green) has a peak in the mid spending score range.

Centroid Visualization:

- Plotted the cluster centroids on the scatter plot to represent the central tendency of each cluster.



Outlier Detection:

- Utilised the Interquartile Range (IQR) method to identify and count outliers in both the remuneration and spending_score columns: no outliers detected.

```python
#Determining outliers with Interquartile Range (IQR) method.
# Calculate Q1, Q3 and IQR for 'remuneration'
Q1_remu = df2['remuneration'].quantile(0.25)
Q3_remu = df2['remuneration'].quantile(0.75)
IQR_remu = Q3_remu - Q1_remu

# Define bounds for outliers
lower_bound_remu = Q1_remu - 1.5 * IQR_remu
upper_bound_remu = Q3_remu + 1.5 * IQR_remu

# Calculate outliers for 'remuneration'
outliers_remu = df2[(df2['remuneration'] < lower_bound_remu) | (df2['remuneration'] > upper_bound_remu)]

# Calculate Q1, Q3 and IQR for 'spending_score'
Q1_score = df2['spending_score'].quantile(0.25)
Q3_score = df2['spending_score'].quantile(0.75)
IQR_score = Q3_score - Q1_score

# Define bounds for outliers
lower_bound_score = Q1_score - 1.5 * IQR_score
upper_bound_score = Q3_score + 1.5 * IQR_score

# Calculate outliers for 'spending_score'
outliers_score = df2[(df2['spending_score'] < lower_bound_score) | (df2['spending_score'] > upper_bound_score)]

# Print count of outliers
print(f"Number of outliers in 'remuneration': {len(outliers_remu)}")
print(f"Number of outliers in 'spending_score': {len(outliers_score)}")
```

```
Number of outliers in 'remuneration': 0
Number of outliers in 'spending_score': 0
```

Insights

The KMeans clustering with k=5 resulted in five distinct clusters.

Each cluster represents a group of customers with similar financial behaviour, which can be pivotal for targeted marketing campaigns. For instance, customers with high remuneration but low spending might be categorised as "High Potential, Low Activity" and could be targeted with exclusive promotions to boost spending (more on the presentation).

The cluster sizes varied, with one group being notably larger (774 members). Such a significant cluster indicates a dominant customer segment, which might be considered the "core" or "typical" customer base. Considering tailoring general marketing strategies to this segment would in my opinion be beneficial.

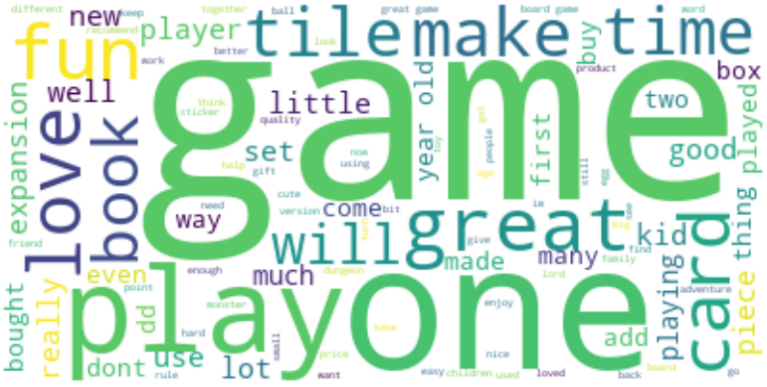
On the other hand, the smaller clusters (like the ones with 269 or 271 members) represent niche segments. These might require specialised marketing approaches or could represent opportunities for growth or expansion.

Analysing customer sentiments with reviews (NPL)

1. **Visual Representation of Words:**

- Created word clouds to display the frequency of words for both reviews and summaries.

Review word cloud:



Summary word cloud:

- Defined a function plot_wordcloud to streamline the visualisation process.

2. **Word Frequency Distribution:**

- Calculated the frequency of words in the 'review' column.
- Filtered out common stopwords and non-alphabetical tokens for clarity.
- Generated another word cloud without stopwords for a focused visualisation.

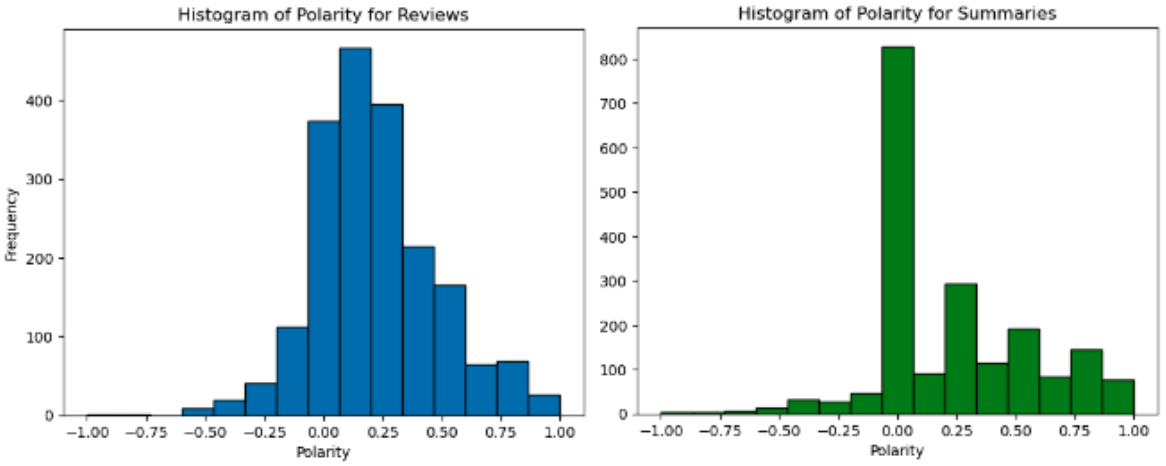
3. **Identification of the 15 Most Common Words:**

```
# Identify 15 most common words and polarity
top_15 = filtered_freq_dist.most_common(15)
print(top_15)

for word, freq in top_15:
    print(f"{word}: {TextBlob(word).sentiment.polarity}")
```

```
[('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)]
game: -0.4
great: 0.8
fun: 0.3
one: 0.0
play: 0.0
like: 0.0
love: 0.5
really: 0.2
get: 0.0
cards: 0.0
tiles: 0.0
time: 0.0
good: 0.7
would: 0.0
book: 0.0
```

4. **Sentiment Analysis:**

- Defined a function, generate_polarity, to derive sentiment scores for reviews and summaries.
- Calculated sentiment scores for both columns.
- Created histograms (below) to visually assess the distribution of sentiment scores across reviews and summaries.



5. **Top Positive and Negative Reviews & Summaries:**

- Extracted the top 20 reviews and summaries with the most negative sentiment scores using the nsmallest function.
- Similarly, identified the top 20 reviews and summaries with the most positive sentiment scores using the nlargest function.

Insights:

Sentiment Values Interpretation:

The word "game" showed a surprising negative sentiment score of -0.4, even though it's typically a neutral term. This could be due to limitations in the pre-trained model of TextBlob.

Words like "great", "fun", "love", "really", and "good" had expected positive sentiment scores ranging from 0.2 to 0.8.

Several words (e.g., "one", "play", "like") were rightly assessed as neutral, having sentiment scores of 0.0.

Given the skewed sentiment score of "game", it would be worth considering its removal from the dataset might provide a clearer sentiment overview. This word, being a domain-specific term (pertaining to Turtle games), could be skewing sentiment outcomes due to its prevalent use and inaccurate sentiment score.

Histogram Insights:

	review_polarity	summary_polarity
count	1961.000000	1961.000000
mean	0.213170	0.223678
std	0.260360	0.337507
min	-1.000000	-1.000000
25%	0.045833	0.000000
50%	0.177222	0.100000
75%	0.351562	0.475000
max	1.000000	1.000000

Both review and summary polarities span from -1 (highly negative) to +1 (highly positive).

- **Central Tendency**

Both reviews and summaries lean towards a positive sentiment, with mean scores of 0.213 and 0.224, respectively. The similarities suggest that summaries are representative of the reviews.

- **Variability**

Summaries exhibit slightly more sentiment variability than reviews, with standard deviations of 0.338 and 0.260, respectively.

- **Outliers and Distribution Spread**

Both reviews and summaries are positively skewed, as their medians (0.177 and 0.1) are below their respective means.

Summaries possess a broader interquartile range (IQR) than reviews, suggesting more sentiment variability within their middle 50% data range.

Insights from Top 20 Negative Reviews:

The negative sentiment score associated with the word "game" may distort the overall sentiment of otherwise neutral or positive reviews.

To refine sentiment evaluation:

- The word "game" could be removed or treated neutrally for sentiment re-assessment.
- The sentiment dictionary's weights can be reviewed and adjusted based on domain relevance and expertise.

Visualise data to gather insights

Calculate Other Sales:

- Derived 'Other_Sales' by subtracting 'NA_Sales' and 'EU_Sales' from 'Global_Sales'.
- Added this column to the subset.

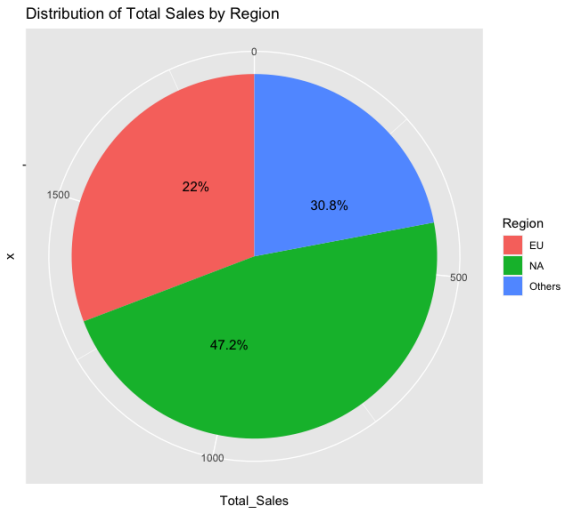
Grouping data:

- Grouped by 'Product' to determine sales impact.
- Summarised total and average sales for different regions and globally.

Visualisation:

Pie Chart:

- Calculated total sales for each region: EU, NA, Others.
- Derived sales percentages and visualised them as a pie chart showing distribution of sales by region.



North America represents 47.2% of the total sales, followed by the other and European markets.

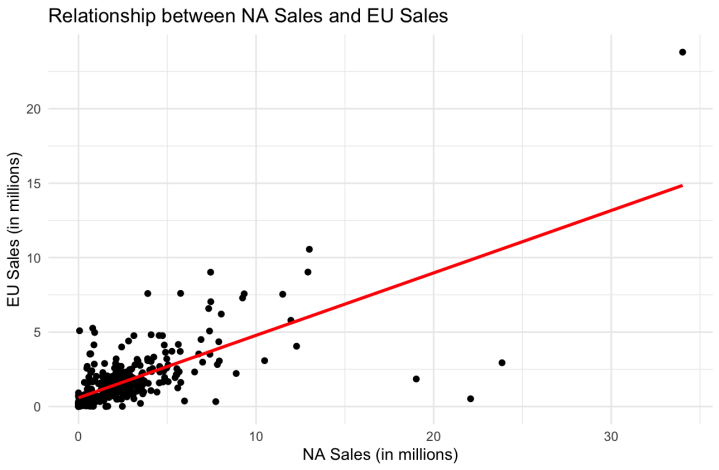
Descriptive Statistics:

- Generated a summary of the 'turtle_sales_subset' dataset.

```
> summary(turtle_sales_subset)
   Product       Platform           NA_Sales          EU_Sales       Global_Sales     Other_Sales
 Min.   : 107   Length:352         Min.   : 0.0000   Min.   : 0.000   Min.   : 0.010   Min.   : 0.000
 1st Qu.:1945   Class :character   1st Qu.: 0.4775   1st Qu.: 0.390   1st Qu.: 1.115   1st Qu.: 0.110
 Median :3340   Mode  :character   Median : 1.8200   Median : 1.170   Median : 4.320   Median : 0.610
 Mean   :3607                      Mean   : 2.5160   Mean   : 1.644   Mean   : 5.335   Mean   : 1.175
 3rd Qu.:5436                      3rd Qu.: 3.1250   3rd Qu.: 2.160   3rd Qu.: 6.435   3rd Qu.: 1.458
 Max.   :9080                      Max.   :34.0200   Max.   :23.800   Max.   :67.850   Max.   :10.030
```

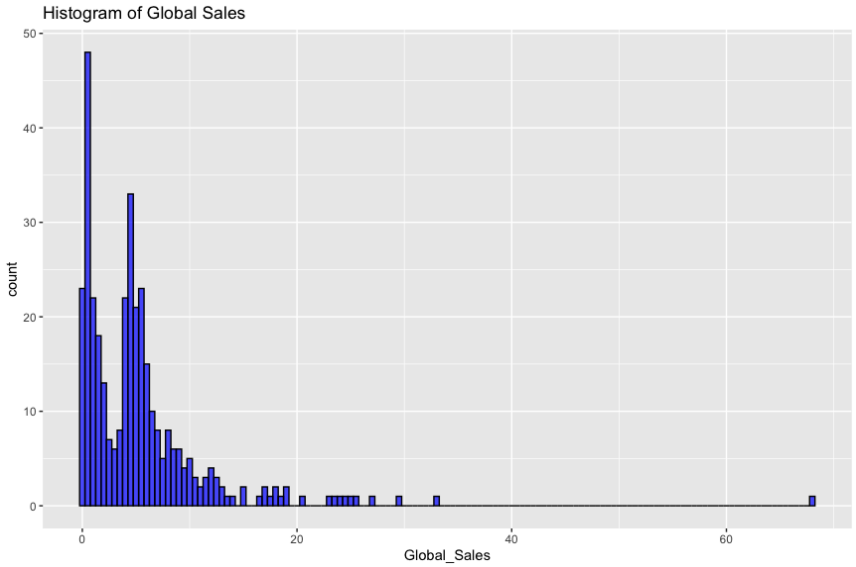
Scatterplot:

- Plotted a scatterplot for 'NA_Sales' vs. 'EU_Sales'.
- Incorporated a linear regression line to better visualise the relationship.



Histogram:

- Showcased the distribution of 'Global_Sales' using a histogram.



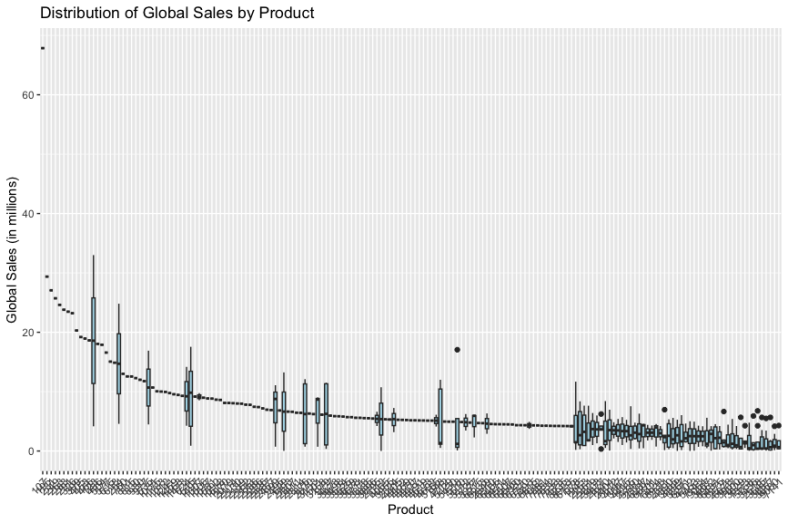
Most products have sales under 10 million, forming a peak on the left. Few products reach the 60-million mark, highlighting the rarity of top-sellers.

Top Sellers Analysis:

- Set a threshold for high sales (top 10% - result: 11.2m)
- Filtered products that surpassed this threshold as top sellers.

This can be useful for the business in order to optimise marketing budgets, refining inventory management, and setting strategic pricing. This insight informs product development based on proven demand and directs targeted marketing to key demographics.
Exploring bundling and cross-promotion strategies could also further boost sales of other related items.

Boxplots:
Designed boxplots to represent the distribution of 'Global_Sales' for each product.



A few products achieved notably high sales. There's a noticeable decline in sales as we move right, indicating few top-performing products and many with low sales. The X axis is very crowded, excluding some products could enhance readability of the graph.

Clean, manipulate and visualise the data

- Determine the min, max, and mean values using the "summary" function.

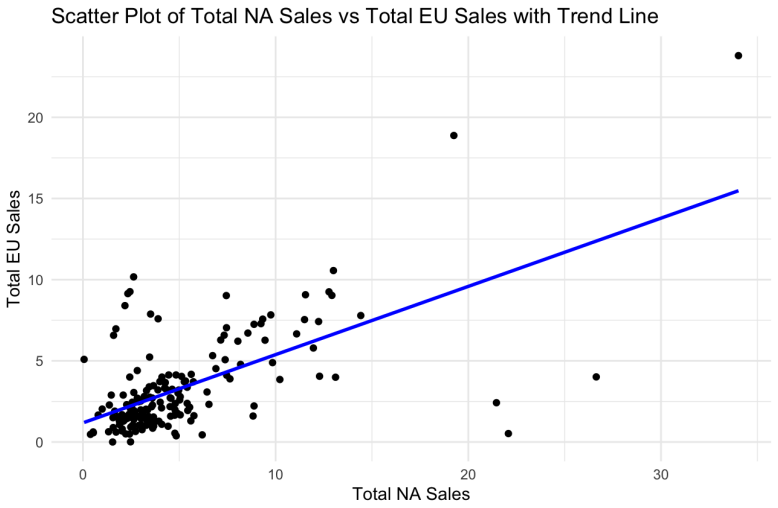
```
> summary(turtle_sales_subset)
    Product       Platform           NA_Sales         EU_Sales       Global_Sales     Other_Sales
 Min.   : 107   Length:352         Min.   : 0.0000   Min.   : 0.000   Min.   : 0.010   Min.   : 0.000
 1st Qu.:1945   Class :character   1st Qu.: 0.4775   1st Qu.: 0.390   1st Qu.: 1.115   1st Qu.: 0.110
 Median :3340   Mode  :character   Median : 1.8200   Median : 1.170   Median : 4.320   Median : 0.610
 Mean   :3607                      Mean   : 2.5160   Mean   : 1.644   Mean   : 5.335   Mean   : 1.175
 3rd Qu.:5436                      3rd Qu.: 3.1250   3rd Qu.: 2.160   3rd Qu.: 6.435   3rd Qu.: 1.458
 Max.   :9080                      Max.   :34.0200   Max.   :23.800   Max.   :67.850   Max.   :10.030
```

Determine the impact on sales per product_id

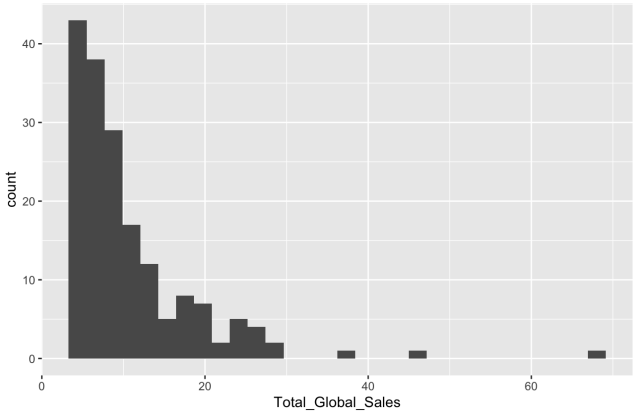
- Group data by Product and calculate total sales using "group_by" and "summarise".
- View and explore the grouped data.

Product	Total_NA_Sales	Total_EU_Sales	Total_Global_Sales	Total_Other_Sales
107	34.02	23.80	67.85	1.003000e+01
123	26.64	4.01	37.16	6.510000e+00
195	13.00	10.56	29.37	5.810000e+00
231	12.92	9.03	27.06	5.110000e+00
249	9.24	7.29	25.72	9.190000e+00

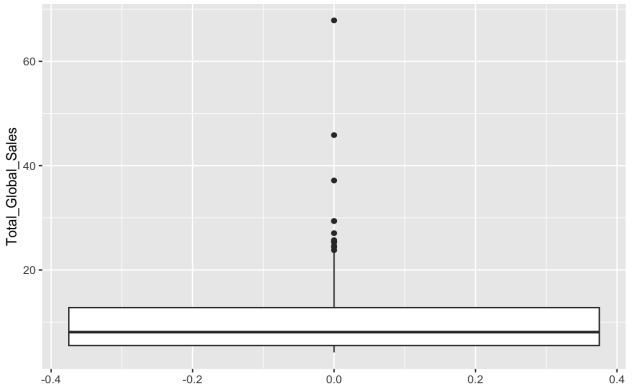
- Scatterplots for Total NA Sales vs Total EU Sales with a trend line.



- Total Global Sales Histogram

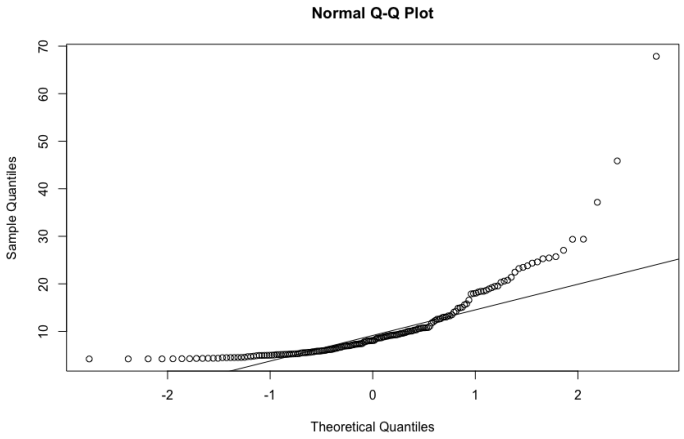


- Total Global Sales boxplot for outliers detection.



Determine the normality of the data set

- Create Q-Q Plots for Total Global Sales.



- Perform the Shapiro-Wilk test for normality.

```
        Shapiro-Wilk normality test

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

- Determine Skewness and Kurtosis values.

```
> skewness(grouped_data$Total_Global_Sales)
[1] 3.066769
> kurtosis(grouped_data$Total_Global_Sales)
[1] 17.79072
```

- Calculate correlation between Total NA Sales and Total EU Sales.

```
> cor(grouped_data$Total_NA_Sales, grouped_data$Total_EU_Sales)
[1] 0.6209317
```

Observations and insights:

- North America shows higher average sales (5.061 million units) compared to Europe (3.306 million units).

Data indicates a broad sales range, hinting at varying product success across regions.
Few products show extremely high sales, indicating certain blockbuster hits in the market.

- The Q-Q plot showed deviations from a normal distribution.
- The Shapiro-Wilk test for global sales has a p-value of < 2.2e-16 (less than a typical alpha level of 0.05), confirming non-normal distribution.
- Skewness (3.066769) suggests positive skew with some products having exceptionally high sales.

- Kurtosis (17.79072) indicates the presence of extreme values in global sales data.
- Lastly, a correlation of 0.6209317 between NA and EU sales signifies a moderate positive relationship (confirming what we saw in the original scatterplot), indicating parallel sales trends between the two regions.

Making recommendations to the business

Creating a Simple Linear Regression Model:

- Determine the correlation between NA_Sales and EU_Sales.
- Calculate correlation coefficient and store it in correlation_coefficient_NA_EU.

```
> correlation_coefficient_NA_EU <- cor(turtle_sales$NA_Sales, turtle_sales$EU_Sales)
> print(paste("Correlation between NA_Sales and EU_Sales:", correlation_coefficient_NA_EU))
[1] "Correlation between NA_Sales and EU_Sales: 0.705523648917198"
```

- Construct a linear regression model using EU_Sales as the dependent variable and NA_Sales as the independent variable.

```
Call:
lm(formula = EU_Sales ~ NA_Sales, data = turtle_sales)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3248 -0.5791 -0.2776  0.3439  8.9501

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.58911    0.09528   6.183 1.75e-09 ***
NA_Sales     0.41919    0.02251  18.625  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.438 on 350 degrees of freedom
Multiple R-squared:  0.4978,    Adjusted R-squared:  0.4963
F-statistic: 346.9 on 1 and 350 DF,  p-value: < 2.2e-16
```

- The summary showed significant predictors and explained variance.

Creating a Multiple Linear Regression Model:

- Extract only numeric columns (NA_Sales, EU_Sales, and Global_Sales) from turtle_sales and store them in numeric_data.
- Build a multiple linear regression model with Global_Sales as the dependent variable and NA_Sales and EU_Sales as predictors.

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

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
```

Making Predictions Based on Given Values:

- Use predefined model coefficients (coeff_intercept, coeff_NA_Sales, and coeff_EU_Sales) for prediction.
- Given specific NA_Sales and EU_Sales values, predict Global_Sales values.
- Extract actual observed Global_Sales values corresponding to the given NA_Sales and EU_Sales. This was done by using a loop as in the screenshot below.

```r
# Extracting the observed Global_Sales for those specific NA_Sales and EU_Sales values
observed_Global_Sales <- numeric(length(NA_Sales_values))

for (i in 1:length(NA_Sales_values)) {
  observed_Global_Sales[i] <- turtle_sales$Global_Sales[which(turtle_sales$NA_Sales == NA_Sales_values[i] &
                                                              turtle_sales$EU_Sales == EU_Sales_values[i])]
}
```

- Create and display a comparison dataframe showcasing NA_Sales, EU_Sales, predicted Global_Sales, and observed Global_Sales.

	NA_Sales	EU_Sales	Predicted_Global_Sales	Observed_Global_Sales
1	34.02	23.80	71.468365	67.85
2	3.93	1.56	6.856063	6.04
3	2.73	0.65	4.248354	4.32
4	2.26	0.97	4.134733	3.53
5	22.08	0.52	26.431469	23.21

Insights:

NA_Sales and EU_Sales Correlation:

A correlation of 0.7055 indicates a strong positive relationship between sales in North America and Europe (noticed before).

Simple Linear Regression (EU_Sales vs. NA_Sales):

- NA_Sales significantly predicts EU_Sales with p-value less than 2e-16.
- For every unit increase in NA_Sales, EU_Sales rises by 0.41919 units.
- This model accounts for about 49.78% of EU_Sales variance (R-squared).

Multiple Linear Regression (Predicting Global_Sales):

- Both NA_Sales and EU_Sales are crucial predictors of Global_Sales.
- For every unit rise in NA_Sales and EU_Sales, Global_Sales climbs by 1.15543 units and 1.34197 units, respectively.
- This model covers 96.85% of the variance in Global_Sales (Adjusted R-squared: 0.9685).

Redundancy & Prediction Accuracy:

- Using NA_Sales and EU_Sales to predict Global_Sales can be seen as redundant since Global_Sales is essentially a sum of sales from different regions, including NA and EU. However, this could still help to to understand the weighted influence of these two regions on global sales.
- The model's predictions are in most of the time overestimating the observed global sales. This could be due to the extraordinarily high sales data points (outliers) we saw during our previous analysis. Adjusting them (by using log transformation for example), removing them, or using models that are less sensitive to them (robust modelling), can lead to more accurate sales predictions.

Recommendations:

Data & Sales:

- Incorporate sales data from other regions and more features like game genres and publishers for a comprehensive model.
- Based on the observed relationship between NA_Sales and EU_Sales, if Turtle Games is launching a new game and expects good sales in North America, they might want to prepare for increased demand in Europe as well.

Business Strategy & External Factors:

- External factors, such as marketing campaigns, geopolitical events, or economic conditions, can influence sales. Understanding these factors can add another layer of depth to the sales prediction models.