Assignment – Course 3 – Report – Sean Meade

Background/context of the business

Turtle Games is a prominent game manufacturer and retailer, serving a diverse global customer base. They offer an extensive range of products including books, board games, video games, and toys, both self-manufactured and sourced from other suppliers. In their pursuit of enhancing overall sales performance, Turtle Games has enlisted the expertise of a team of data analysts. Their primary goal is to harness customer trends and insights to drive growth. The company seeks answers to critical questions, such as understanding customer loyalty point accumulation, identifying market segments within their customer base, leveraging social data for marketing, assessing the impact of individual products on sales, evaluating data reliability, and uncovering sales relationships between North America, Europe, and the global market. Turtle Games aims to leverage data-driven strategies for informed decision-making and ultimately boost their sales performance. As part of a data analyst team, I employed Python and R for data analysis, addressing specific tasks outlined by Turtle Games.

Analytical approach:

To address Turtle Games' objectives, a comprehensive analytical approach was adopted using both Python and R. In Python, the pandas library facilitated data manipulation, exploration, and cleaning. (See Figure 1)

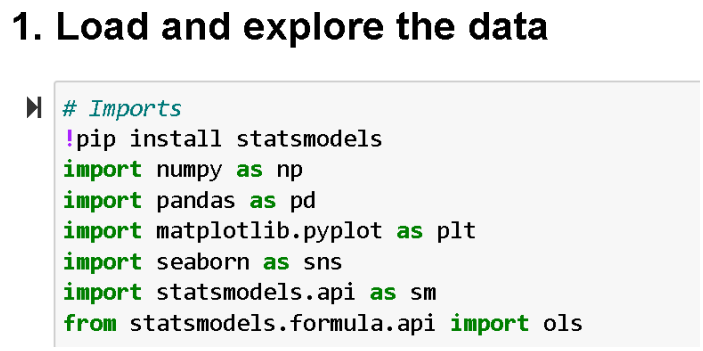

Figure 1

I imported all the datasets into Jupyter Notebook and into Rstudio through reading the csv's (turtle_reviews (reviews)) through the read function through pandas and tidyverse. I viewed all the Dataframes on all of the datasets to check they have imported correctly by printing the head, shape, tail & columns. I checked all the files for any missing values which there were none through isnull().sum(). I also looked at the metadata of the datasets through dtypes() and typeof() which is really important to see the data types of each column. Finally, I then determined the descriptive statistics of them through describe and info. I believe after all this, the datasets were ready to go and ready for analysis. Descriptive statistics were employed to sense-check the data, identify missing values, and derive insights. Redundant columns were removed, and meaningful headings were assigned for clarity. Linear regression analysis, facilitated by the statsmodels library, was employed to evaluate relationships between loyalty points and age/remuneration/spending scores. The analysis of these relationships are shown below in separate subheadings. In 'loyalty_points' was used as the dependent variable.

Spending vs Loyalty

I have defined the variables below and determined the regression model (OLS model) and looked at the summary of the data (Figure 2-1). I made a scatter plot show the relationship between the 2 variables showing a positive correlation (Figure 2-2). I set the X coefficient and the constant to generate the regression table (Figure 2-3). Then I plot the graph with a regression line. (Figure 2-4)

5a) spending vs loyalty

In [11]:
```
df.corr()
# Independent variable.
y = df['spending_score']

# Dependent variable.
x = df['loyalty_points']

# OLS model and summary.
plt.scatter(x, y)
f = 'y ~ x'
test = ols(f, data = df).fit()
test.summary()
```

Out[11]:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.
Date:	Tue, 31 Oct 2023	**Prob (F-statistic):**	2.92e-263
Time:	20:38:27	**Log-Likelihood:**	-8759.4
No. Observations:	2000	**AIC:**	1.752e+04
Df Residuals:	1998	**BIC:**	1.753e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	28.4260	0.685	41.504	0.000	27.083	29.769
x	0.0137	0.000	40.595	0.000	0.013	0.014

Omnibus:	169.397	**Durbin-Watson:**	2.599
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	212.607
Skew:	0.768	**Prob(JB):**	6.81e-47
Kurtosis:	3.441	**Cond. No.**	3.22e+03

Figure 2-1

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.22e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

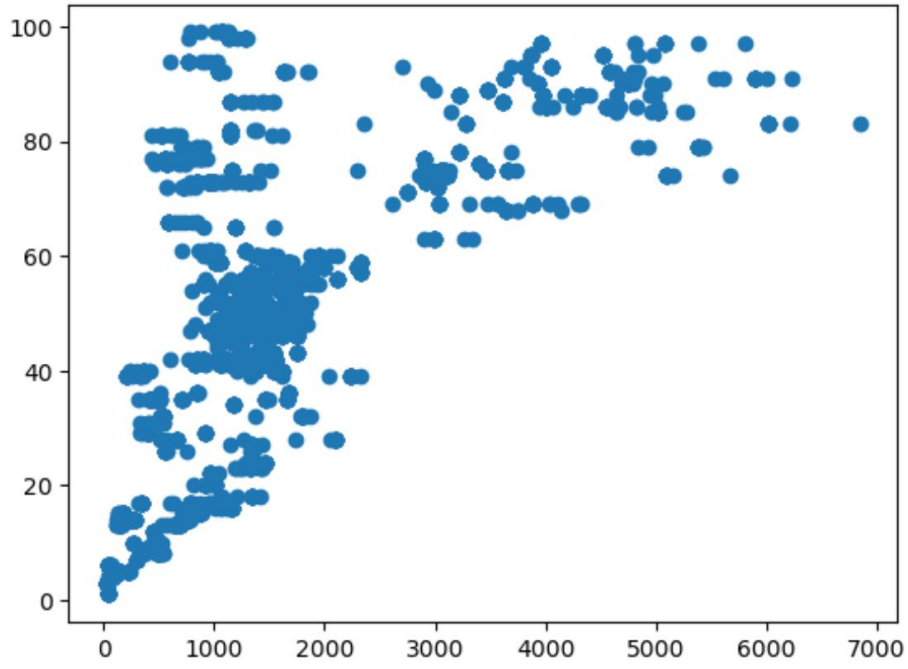

Figure 2-2

In [12]:

```
# Extract the estimated parameters.
print("Parameters: ", test.params)

# Extract the standard errors.
print("Standard errors: ", test.bse)

# Extract the predicted values.
print("Predicted values: ", test.predict())
```

```
Parameters:  Intercept    28.426033
x             0.013671
dtype: float64
Standard errors:  Intercept    0.684905
x            0.000337
dtype: float64
Predicted values:  [ 31.29703545  35.58986696  28.97289101 ... 105.17748592  4
2.75370042
  34.97465225]
```

In [13]:

```
# Set the X coefficient and the constant to generate the regression table.
y_pred = (28.4260) + 0.0137 * df['loyalty_points']

# View the output.
y_pred
```

Out[13]:

```
0        31.3030
1        35.6048
2        28.9740
3        36.1254
4        33.4402
          ...
1995     83.6507
1996     35.8103
1997    105.3378
1998     42.7836
1999     34.9883
Name: loyalty_points, Length: 2000, dtype: float64
```

Figure 2-3

In [14]:

```
# Plot the graph with a regression line.
plt.scatter(x, y)

# Plot the regression line (in black).
plt.plot(x, y_pred, color='black')

# Set the x and y limits on the axes.
plt.xlim(0)
plt.ylim(0)

# View the plot.
plt.show()
```

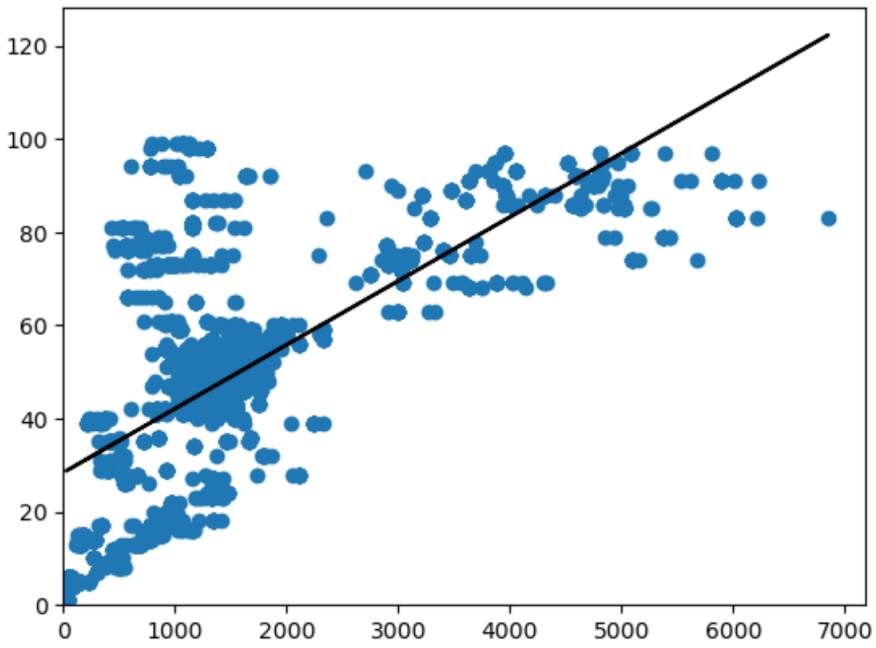

Figure 2-4

Renumeration vs Loyalty

I have defined the variables below and determined the regression model (OLS model) and looked at the summary of the data (Figure 3-1). I made a scatter plot show the relationship between the 2 variables showing a positive correlation (Figure 3-2). I set the X coefficient and the constant to generate the regression table (Figure 3-3). Then I plot the graph with a regression line. (Figure 3-4)

5b) renumeration vs loyalty

In [15]:

```
# Independent variable.
q = df['renumeration']

# Dependent variable.
p = df['loyalty_points']

# OLS model and summary.
plt.scatter(p, q)
e = 'q ~ p'
test_1 = ols(e, data = df).fit()
test_1.summary()
```

Out[15]:

OLS Regression Results

Dep. Variable:	q	**R-squared:**	0.380
Model:	OLS	**Adj. R-squared:**	0.379
Method:	Least Squares	**F-statistic:**	1222.
Date:	Tue, 31 Oct 2023	**Prob (F-statistic):**	2.43e-209
Time:	20:38:28	**Log-Likelihood:**	-8641.8
No. Observations:	2000	**AIC:**	1.729e+04
Df Residuals:	1998	**BIC:**	1.730e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	30.5606	0.646	47.321	0.000	29.294	31.827
p	0.0111	0.000	34.960	0.000	0.010	0.012

Omnibus:	382.801	**Durbin-Watson:**	1.461
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	657.257
Skew:	1.230	**Prob(JB):**	1.90e-143
Kurtosis:	4.357	**Cond. No.**	3.22e+03

Figure 3- 1

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.22e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

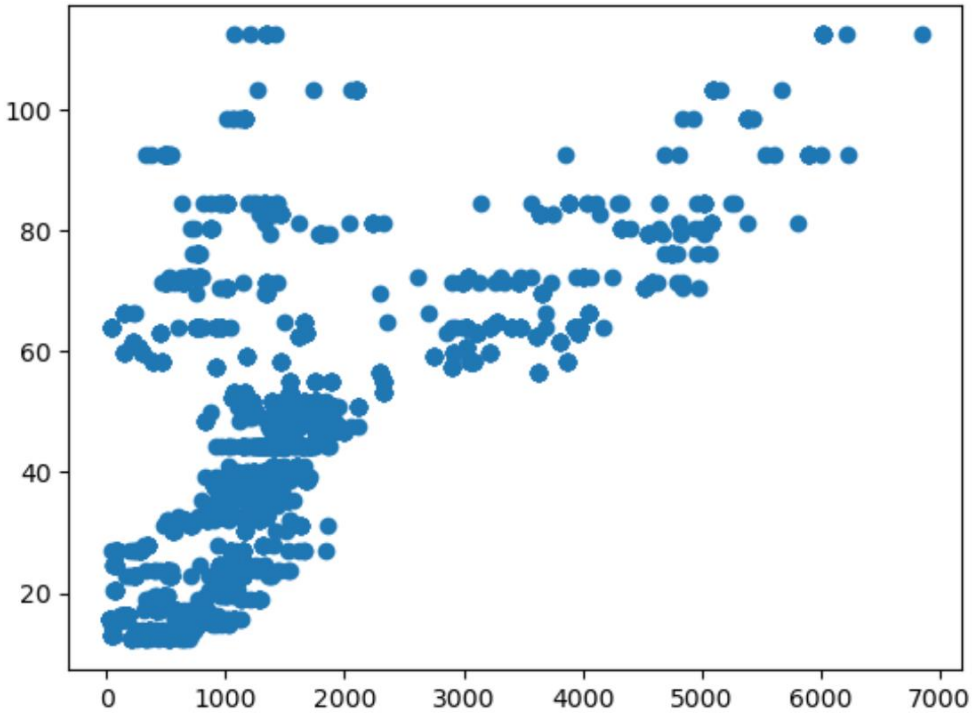

Figure 3- 2

In [16]:

```
# Extract the estimated parameters.
print("Parameters: ", test_1.params)

# Extract the standard errors.
print("Standard errors: ", test_1.bse)

# Extract the predicted values.
print("Predicted values: ", test_1.predict())
```

```
Parameters:  Intercept    30.560555
p             0.011101
dtype: float64
Standard errors:  Intercept    0.645817
p            0.000318
dtype: float64
Predicted values:  [32.89186761 36.3777352  31.00461446 ... 92.88431491 42.194
91551
 35.87816819]
```

In [17]:

```
# Set the the X coefficient and the constant to generate the regression table.
q_pred = (30.5606) + 0.0111 * df['loyalty_points']

# View the output.
q_pred
```

Out[17]:

```
0       32.8916
1       36.3770
2       31.0046
3       36.7988
4       34.6232
         ...
1995    75.3047
1996    36.5435
1997    92.8760
1998    42.1934
1999    35.8775
Name: loyalty_points, Length: 2000, dtype: float64
```

Figure 3-3

In [18]:

```
# Plot graph with regression line.
plt.scatter(p, q)

# Plot the regression line (in black).
plt.plot(p, q_pred, color='black')

# Set the x and y limits on the axes.
plt.xlim(0)
plt.ylim(0)

# View the plot.
plt.show()
```

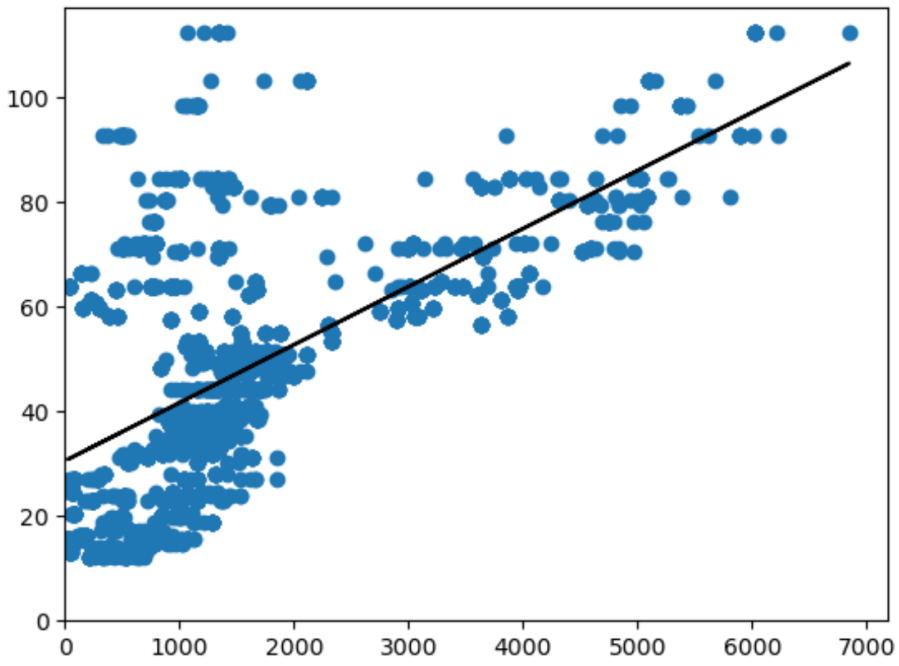

Figure 3-4

Age vs Loyalty

I have defined the variables below and determined the regression model (OLS model) and looked at the summary of the data (Figure 4-1). I made a scatter plot show the relationship between the 2 variables showing a slightly negative correlation (Figure 4-2). I set the X coefficient and the constant to generate the regression table (Figure 4-3). Then I plotted the graph with a regression line. (Figure 4-4)

5c) age vs loyalty

In [19]:

```
# Independent variable.
t = df['age']

# Dependent variable.
s = df['loyalty_points']

# OLS model and summary.
plt.scatter(s, t)
d = 't ~ s'
test_2 = ols(d, data = df).fit()
test_2.summary()
```

Out[19]:

OLS Regression Results

Dep. Variable:	t	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.606
Date:	Tue, 31 Oct 2023	**Prob (F-statistic):**	0.0577
Time:	20:38:32	**Log-Likelihood:**	-8051.8
No. Observations:	2000	**AIC:**	1.611e+04
Df Residuals:	1998	**BIC:**	1.612e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	40.2035	0.481	83.615	0.000	39.261	41.146
s	-0.0004	0.000	-1.899	0.058	-0.001	1.47e-05

Omnibus:	99.357	**Durbin-Watson:**	2.129
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	112.564
Skew:	0.574	**Prob(JB):**	3.61e-25
Kurtosis:	2.814	**Cond. No.**	3.22e+03

Figure 4- 1

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.22e+03. This might indicate that there are strong multicollinearity or other numerical problems.

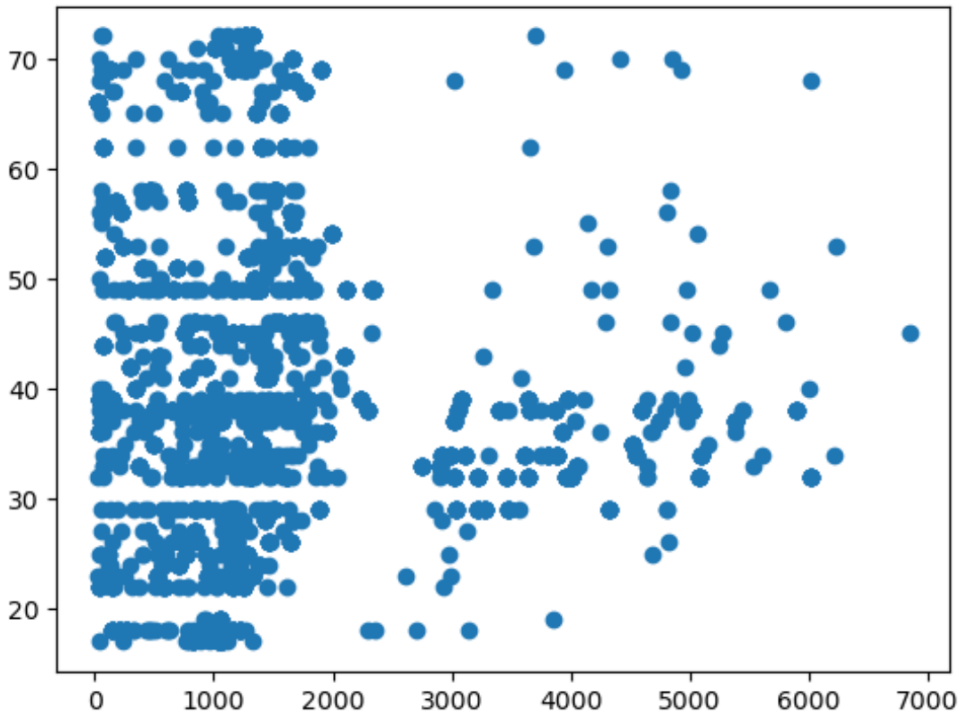

Figure 4-2

In [20]:
```
# Extract the estimated parameters.
print("Parameters: ", test_2.params)

# Extract the standard errors.
print("Standard errors: ", test_2.bse)

# Extract the predicted values.
print("Predicted values: ", test_2.predict())
```

```
Parameters:  Intercept    40.203457
s            -0.000449
dtype: float64
Standard errors:  Intercept    0.480816
s            0.000236
dtype: float64
Predicted values:  [40.10917768 39.96820745 40.18549915 ... 37.683053   39.732
95776
 39.98841019]
```

In [21]:
```
# Set the X coefficient and the constant to generate the regression table.
t_pred = (40.2035) - 0.0004 * df['loyalty_points']

# View the output.
t_pred
```

Out[21]:
```
0       40.1195
1       39.9939
2       40.1875
3       39.9787
4       40.0571
         ...
1995    38.5911
1996    39.9879
1997    37.9579
1998    39.7843
1999    40.0119
Name: loyalty_points, Length: 2000, dtype: float64
```

Figure 4- 3

In [22]:

```
# Plot graph with regression line.
plt.scatter(s, t)

# Plot the regression line (in black).
plt.plot(s, t_pred, color='black')

# Set the x and y limits on the axes.
plt.xlim(0)
plt.ylim(0)

# View the plot.
plt.show()
```

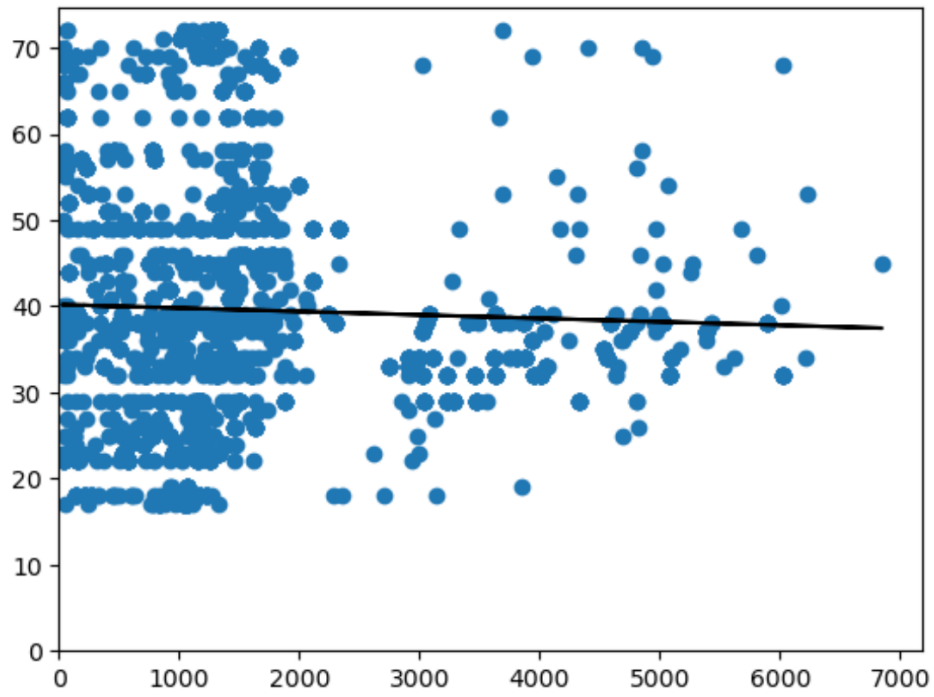

Figure 4- 4

For the NLP task, the Jupyter Notebook was utilised, incorporating pandas for data handling. Prepared the data for NLP by changing to lower case and join the elements in each of the columns (Summary & Review), then tokenised and cleaned. Created a frequency distribution and removed alphanumeric characters and created wordcloud with/without stopwords (Figures 9-9to9-12). Then we used the SentimentIntensityAnalyzer to determine the sentiment score and used TextTholb for the polarity score for both columns. The focus was on identifying the 15 most common words in product reviews/summaries (Figures 8-1&8-2) and extracting the top 20 positive and negative reviews/summaries (Figures 9-1to9-8). I Visualisations such as scatterplots (Figures 5-1&5-2) and pairplots (Figures 6-1&6-2) aided in determining correlations and potential clusters especially in terms of education types and genders.

For NLP, scatterplots and pairplots visually identified clusters, aiding in the interpretation of k-means results. Although there were only 5 education types, it seems that k=5 (five clusters) might give the best results (groups). The five education types are closely related (graduate, postgraduate); therefore, Cluster 0 for both k=4, k=5, k=6 is the largest group. The number of predicted values per

class indicates a better distribution for k=5 than k=4. (See Figures 7-3to7-5 for analysis) The k-means clustering algorithm was applied, and the optimal number of clusters was determined using the Elbow and Silhouette methods. (See Figures 7-1&7-2)

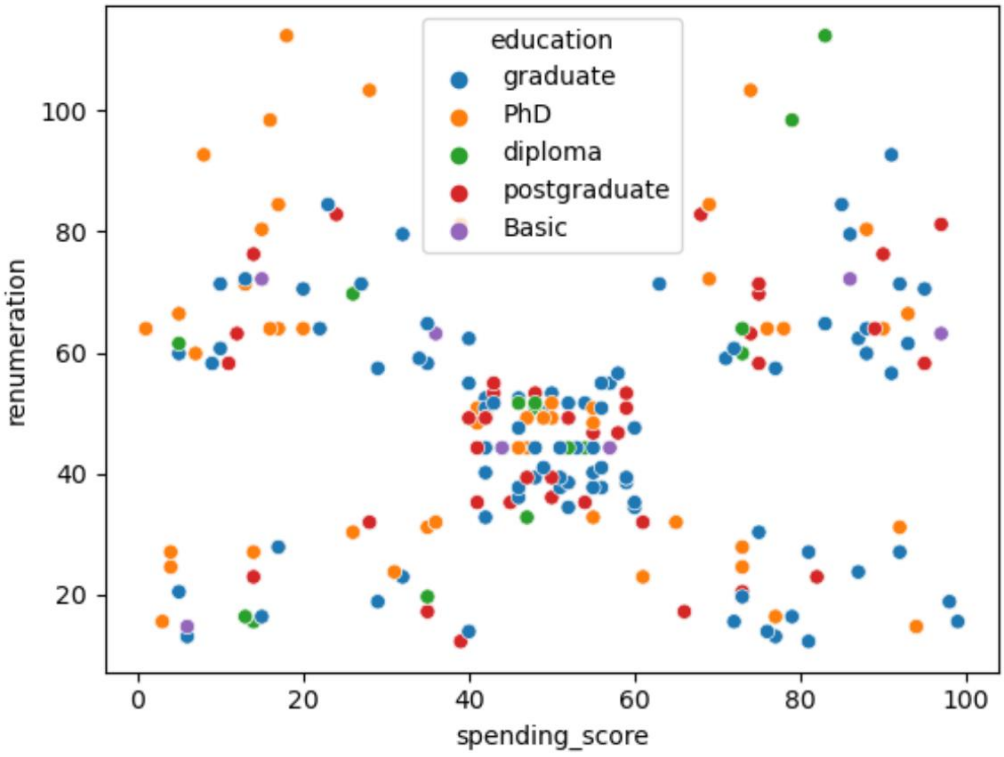

Figure 5-1 (Education)

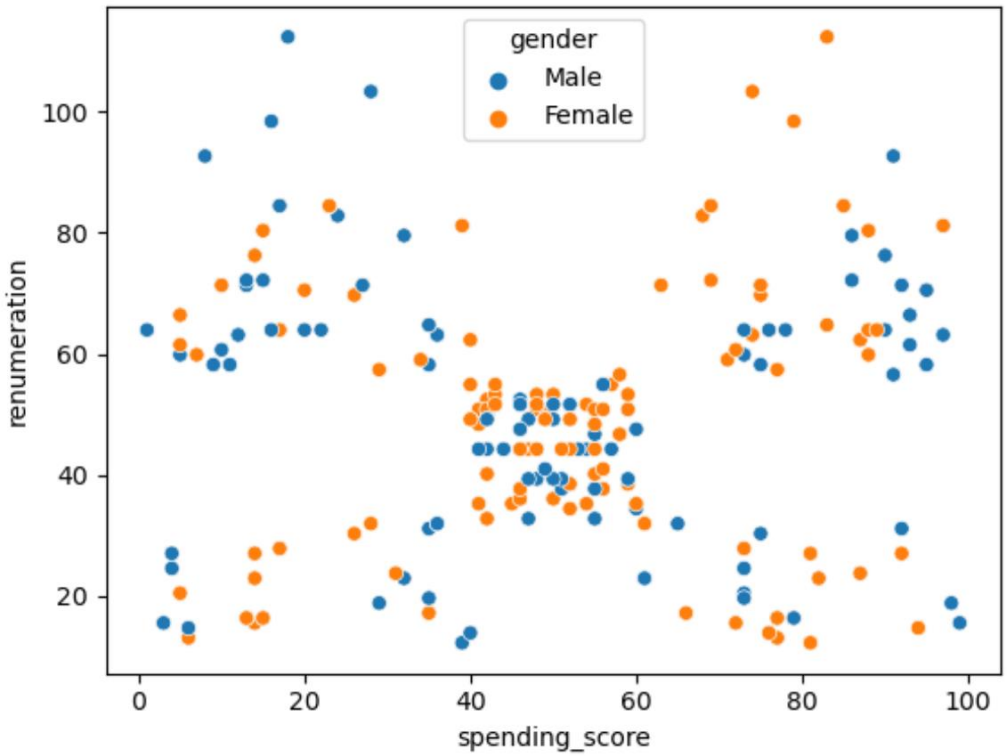

Figure 5-2 (Gender)

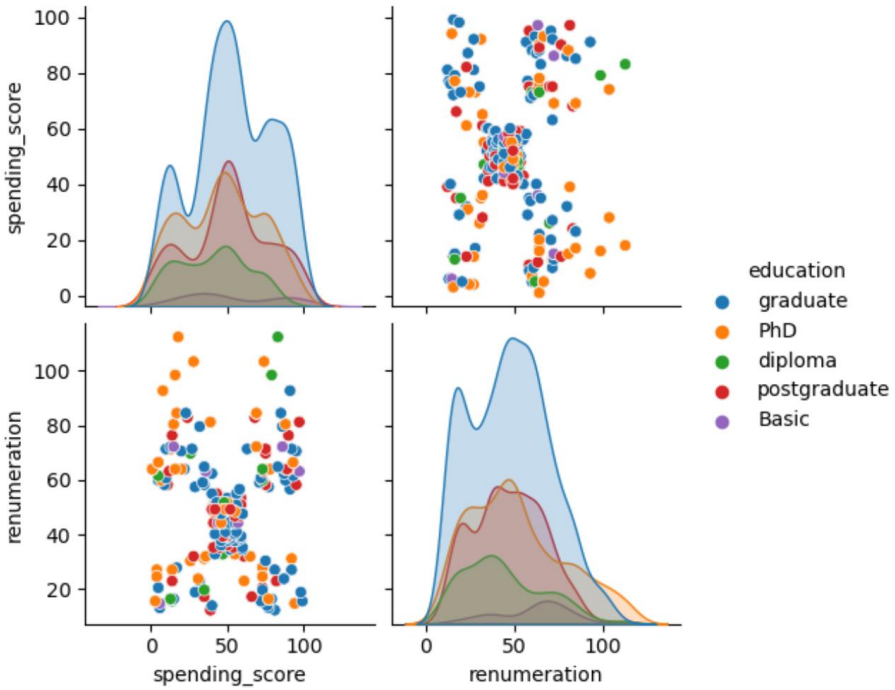

Figure 6-1 (Education)

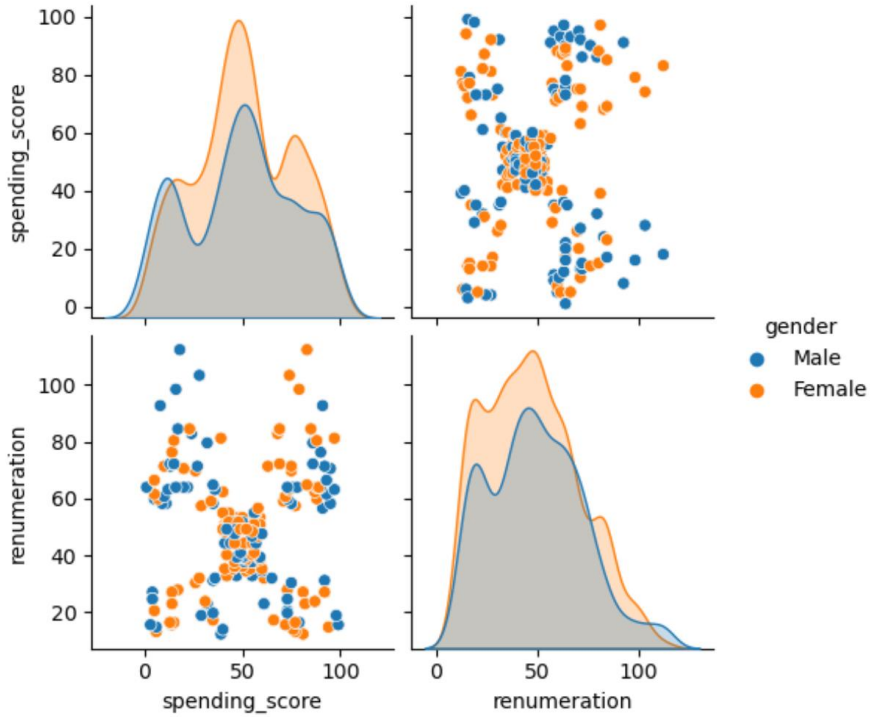

Figure 6-2 (Gender)

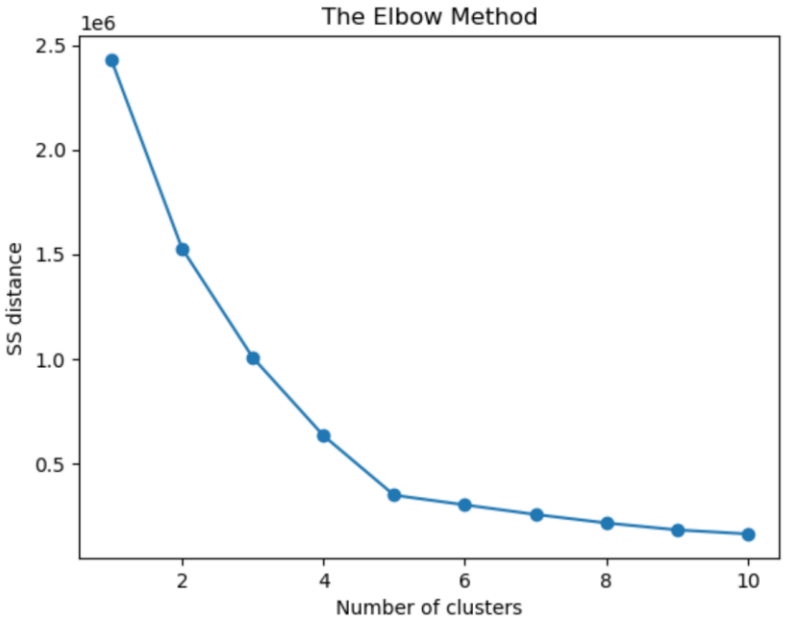

Figure 7-1

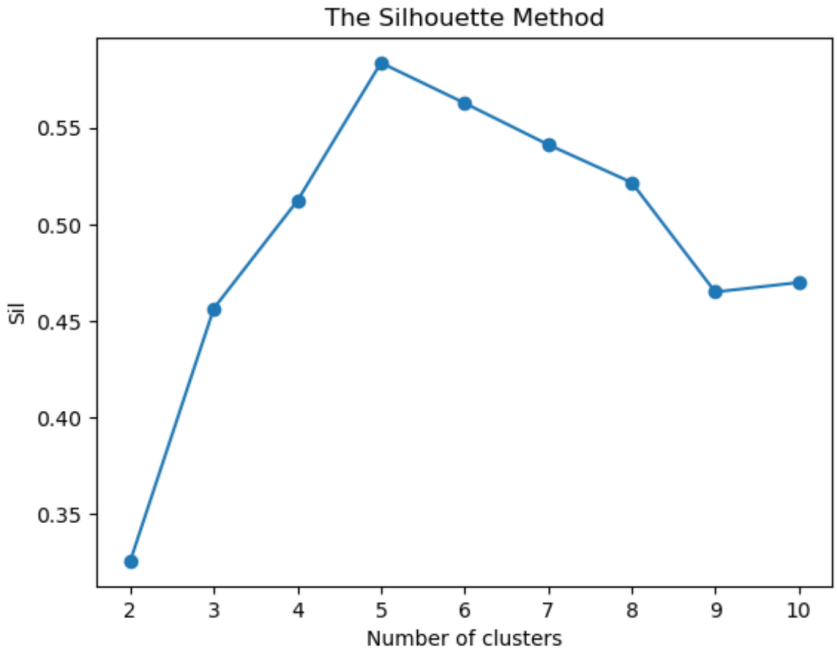

Figure 7-2

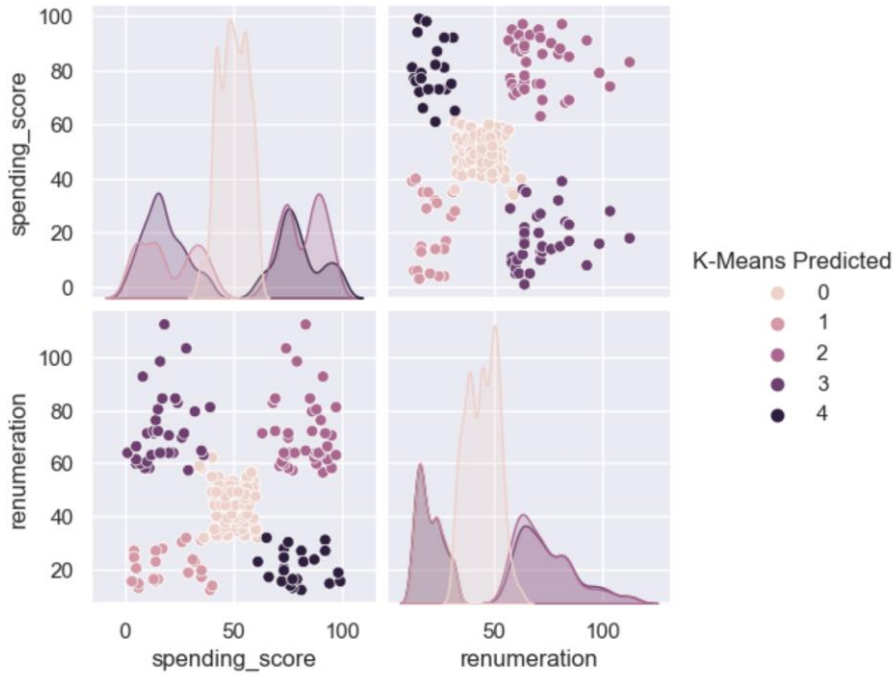

Figure 7-1

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

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

Figure 7-2

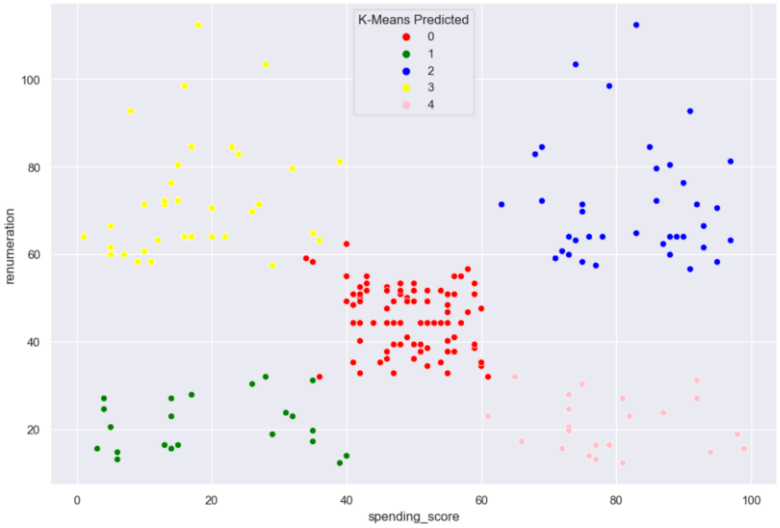

Figure 7-3

4d) Identify 15 most common words and polarity

```
# Determine the 15 most common words.

# Review

# Create a frequency distribution object.
freq_dist_of_words_review_c = FreqDist(df3_r_t_clean)

# Show the five most common elements in the data set.
freq_dist_of_words_review_c.most_common(15)
```

```
20]: [('game', 1663),
      ('great', 574),
      ('fun', 551),
      ('one', 530),
      ('play', 502),
      ('like', 414),
      ('love', 320),
      ('really', 319),
      ('get', 319),
      ('cards', 301),
      ('tiles', 297),
      ('time', 291),
      ('good', 286),
      ('would', 280),
      ('book', 273)]
```

Figure 8-1 (Reviews)

```
# Summary

# Create a frequency distribution object.
freq_dist_of_words_summary_c = FreqDist(df3_s_t_clean)

# Show the five most common elements in the data set.
freq_dist_of_words_summary_c.most_common(15)
```

```
[('stars', 406),
 ('five', 321),
 ('game', 319),
 ('great', 295),
 ('fun', 218),
 ('love', 93),
 ('good', 92),
 ('four', 58),
 ('like', 54),
 ('expansion', 52),
 ('kids', 50),
 ('cute', 45),
 ('book', 43),
 ('one', 38),
 ('awesome', 36)]
```

Figure 8-2 (Summaries)

	index	review	summary	polarity	score
0	1119	here is my review crossposted from boardgamege...	a disappointing coop game	-0.024516	0.5209
1	1559	the one ring is a very innovative rpg set in m...	special dice for a great game	0.146973	0.5209
2	281	i bought this thinking it would be really fun ...	disappointed 9 year old	-0.180952	0.5209
3	459	this is a fun game for kids its basically the ...	this is a fun game for kids	-0.091667	0.5209
4	882	a crappy cardboard ghost of the original hard...	a crappy cardboard ghost of the original hard...	-0.305556	0.5209
5	899	the game is amazing the new version is not car...	50th anniversary is a sad day for acquire	-0.131364	0.5209
6	1069	for those just getting started in the dnd worl...	fun and easy to learn	0.048302	0.5209
7	1148	i dig this its a shame that 56 years ago i was...	enhance your play	0.002811	0.5209
8	1797	the ball of whacks can be a valuable tool for ...	ball of whacks very useful for pain management	0.152927	0.5209
9	347	my 8 yearold granddaughter and i were very fru...	frustating	-0.446250	0.5209
10	506	its really uno type game but anger control stu...	bit dissapointed	-0.058333	0.5209
11	1670	i thought i was getting a new product but the ...	not as expected	0.147273	0.5209
12	2	nice art nice printing why two panels are fil...	pretty but also pretty useless	0.116640	0.5209
13	363	i found that this card game does the opposite ...	promotes anger instead of teaching calming met...	-0.126190	0.5209
14	426	its uno with questions about anger its an okay...	but it gets repetitive and the students start ...	-0.288095	0.5209
15	359	this is horrible the directions are very hard ...	not worth the money	-0.255833	0.5209
16	1332	not a hard game to learn but not easy to win	five stars	0.082292	0.5209
17	852	firstly this game is excellent the previous ed...	one of a kind game design but boy is it made c...	0.020539	0.5209
18	593	if i could give this egg zero stars i would it...	too hard for children or adults to open	-0.236458	0.5209
19	551	my son loves the books but was very disappoint...	poorly made and will require needlethread repa...	-0.074074	0.5209

Figure 9-1 (Reviews – Top 20 Negative)

	sentiment_score
0	-0.9828
1	-0.9821
2	-0.9520
3	-0.9146
4	-0.9052
5	-0.8984
6	-0.8908
7	-0.8829
8	-0.8722
9	-0.8674
10	-0.8668
11	-0.8518
12	-0.8334
13	-0.8179
14	-0.8126
15	-0.8067
16	-0.7946
17	-0.7859
18	-0.7845
19	-0.7468

Figure 9-2 (Reviews – Top 20 Negative)

	index	review	summary	polarity	score
1920	1112	i just bought this game with my 8 year old son...	great game for dad and kids	0.119666	0.5209
1921	586	we were sooo excited to see the easter story e...	we were given the star from afar a few years a...	0.137964	0.5209
1922	926	i was skeptical about castle ravenloft despite...	dungeons and dragons wrath of ashardalon is to...	0.170909	0.5209
1923	1810	monopoly is fun but it takes forever to play t...	our familys favorite game	0.042005	0.5209
1924	1097	i have a wife and kids i dont have time to be...	noobyparttimedders game	0.057018	0.5209
1925	1095	i would recommend this game to any fan of fant...	a very fun boardgame	0.058013	0.5209
1926	1103	who am i middle aged married guy who loves sci...	dd lite for those with limited time	0.118078	0.5209
1927	1575	doctor who the card game was created by martin...	fun card mechanic quirky internal logic	0.050201	0.5209
1928	928	wrath of ashardalon gets everything right for ...	a fast enjoyable board game	0.094880	0.5209
1929	1116	the short short version wrath of ashardalon is...	what it says on the tin	0.085084	0.5209
1930	1073	we own this game as well as castle ravenloft ...	a solid experience for dd light	0.110737	0.5209
1931	1063	if you are a fan of dungeons and dragons or ot...	why havent you bought this already	0.050881	0.5209
1932	1357	lords of waterdeep scoundrels of skullport is ...	take your lords of waterdeep games to a new level	0.101209	0.5209
1933	1295	lords of waterdeep was awesome and scoundrels ...	updated review waterdeep on steroids	-0.011648	0.5209
1934	857	i grew up playing monopoly lots of people did...	acquire the game you should be playing instead...	0.176914	0.5209
1935	358	this kit is awesome my 5year old daughter and ...	easy finished product as cute as the cover	0.330432	0.5209
1936	1666	if you only employ one creativityenhancing res...	an epochal innovation breakthrough	0.171376	0.5209
1937	1570	as a dad of two boys im always on the lookout ...	hobbits love telling tales so do i	0.095337	0.5209
1938	879	whenever i see this game on my shelf i get a d...	acquire review by dads gaming addiction	0.058882	0.5209
1939	1121	disclaimer bought this from a local store paid...	wrath of ashardalon great investment for an av...	0.096860	0.5209

Figure 9-3 (Reviews – Top 20 Positive)

	sentiment_score
1920	0.9962
1921	0.9963
1922	0.9963
1923	0.9964
1924	0.9965
1925	0.9966
1926	0.9967
1927	0.9970
1928	0.9972
1929	0.9974
1930	0.9975
1931	0.9978
1932	0.9984
1933	0.9985
1934	0.9987
1935	0.9991
1936	0.9991
1937	0.9991
1938	0.9994
1939	0.9996

Figure 9-4 (Reviews – Top 20 Positive)

	index	review	summary	polarity	score
0	1119	here is my review crossposted from boardgamege...	a disappointing coop game	-0.024516	0.5209
1	1559	the one ring is a very innovative rpg set in m...	special dice for a great game	0.146973	0.5209
2	281	i bought this thinking it would be really fun ...	disappointed 9 year old	-0.180952	0.5209
3	459	this is a fun game for kids its basically the ...	this is a fun game for kids	-0.091667	0.5209
4	882	a crappy cardboard ghost of the original hard...	a crappy cardboard ghost of the original hard...	-0.305556	0.5209
5	899	the game is amazing the new version is not car...	50th anniversary is a sad day for acquire	-0.131364	0.5209
6	1069	for those just getting started in the dnd worl...	fun and easy to learn	0.048302	0.5209
7	1148	i dig this its a shame that 56 years ago i was...	enhance your play	0.002811	0.5209
8	1797	the ball of whacks can be a valuable tool for ...	ball of whacks very useful for pain management	0.152927	0.5209
9	347	my 8 yearold granddaughter and i were very fru...	frustating	-0.446250	0.5209
10	506	its really uno type game but anger control stu...	bit dissapointed	-0.058333	0.5209
11	1670	i thought i was getting a new product but the ...	not as expected	0.147273	0.5209
12	2	nice art nice printing why two panels are fil...	pretty but also pretty useless	0.116640	0.5209
13	363	i found that this card game does the opposite ...	promotes anger instead of teaching calming met...	-0.126190	0.5209
14	426	its uno with questions about anger its an okay...	but it gets repetitive and the students start ...	-0.288095	0.5209
15	359	this is horrible the directions are very hard ...	not worth the money	-0.255833	0.5209
16	1332	not a hard game to learn but not easy to win	five stars	0.082292	0.5209
17	852	firstly this game is excellent the previous ed...	one of a kind game design but boy is it made c...	0.020539	0.5209
18	593	if i could give this egg zero stars i would it...	too hard for children or adults to open	-0.236458	0.5209
19	551	my son loves the books but was very disappoint...	poorly made and will require needlethread repa...	-0.074074	0.5209

Ne

Figure 9-5 (Summaries – Top 20 Negative)

	sentiment_score
0	-0.9828
1	-0.9821
2	-0.9520
3	-0.9146
4	-0.9052
5	-0.8984
6	-0.8908
7	-0.8829
8	-0.8722
9	-0.8674
10	-0.8668
11	-0.8518
12	-0.8334
13	-0.8179
14	-0.8126
15	-0.8067
16	-0.7946
17	-0.7859
18	-0.7845
19	-0.7468

Figure 9-6 (Summaries – Top 20 Negative)

	index	review	summary	polarity	score
1920	1112	i just bought this game with my 8 year old son...	great game for dad and kids	0.119666	0.5209
1921	586	we were sooo excited to see the easter story e...	we were given the star from afar a few years a...	0.137964	0.5209
1922	926	i was skeptical about castle ravenloft despite...	dungeons and dragons wrath of ashardalon is to...	0.170909	0.5209
1923	1810	monopoly is fun but it takes forever to play t...	our familys favorite game	0.042005	0.5209
1924	1097	i have a wife and kids i dont have time to be...	noobyparttimedders game	0.057018	0.5209
1925	1095	i would recommend this game to any fan of fant...	a very fun boardgame	0.058013	0.5209
1926	1103	who am i middle aged married guy who loves sci...	dd lite for those with limited time	0.118078	0.5209
1927	1575	doctor who the card game was created by martin...	fun card mechanic quirky internal logic	0.050201	0.5209
1928	928	wrath of ashardalon gets everything right for ...	a fast enjoyable board game	0.094880	0.5209
1929	1116	the short short version wrath of ashardalon is...	what it says on the tin	0.085084	0.5209
1930	1073	we own this game as well as castle ravenloft ...	a solid experience for dd light	0.110737	0.5209
1931	1063	if you are a fan of dungeons and dragons or ot...	why havent you bought this already	0.050881	0.5209
1932	1357	lords of waterdeep scoundrels of skullport is ...	take your lords of waterdeep games to a new level	0.101209	0.5209
1933	1295	lords of waterdeep was awesome and scoundrels ...	updated review waterdeep on steroids	-0.011648	0.5209
1934	857	i grew up playing monopoly lots of people did...	acquire the game you should be playing instead...	0.176914	0.5209
1935	358	this kit is awesome my 5year old daughter and ...	easy finished product as cute as the cover	0.330432	0.5209
1936	1666	if you only employ one creativityenhancing res...	an epochal innovation breakthrough	0.171376	0.5209
1937	1570	as a dad of two boys im always on the lookout ...	hobbits love telling tales so do i	0.095337	0.5209
1938	879	whenever i see this game on my shelf i get a d...	acquire review by dads gaming addiction	0.058882	0.5209
1939	1121	disclaimer bought this from a local store paid...	wrath of ashardalon great investment for an av...	0.096860	0.5209

Figure 9-7 (Summaries – Top 20 Positive)

	sentiment_score
1920	0.9962
1921	0.9963
1922	0.9963
1923	0.9964
1924	0.9965
1925	0.9966
1926	0.9967
1927	0.9970
1928	0.9972
1929	0.9974
1930	0.9975
1931	0.9978
1932	0.9984
1933	0.9985
1934	0.9987
1935	0.9991
1936	0.9991
1937	0.9991
1938	0.9994
1939	0.9996

Figure 9-8 (Summaries – Top 20 Positive)

Figure 9-9 (Reviews – Wordcloud – without Stopwords)

Figure 9-10 (Reviews – Wordcloud – with Stopwords)

Figure 9-11 (Summaries – Wordcloud – without Stopwords)

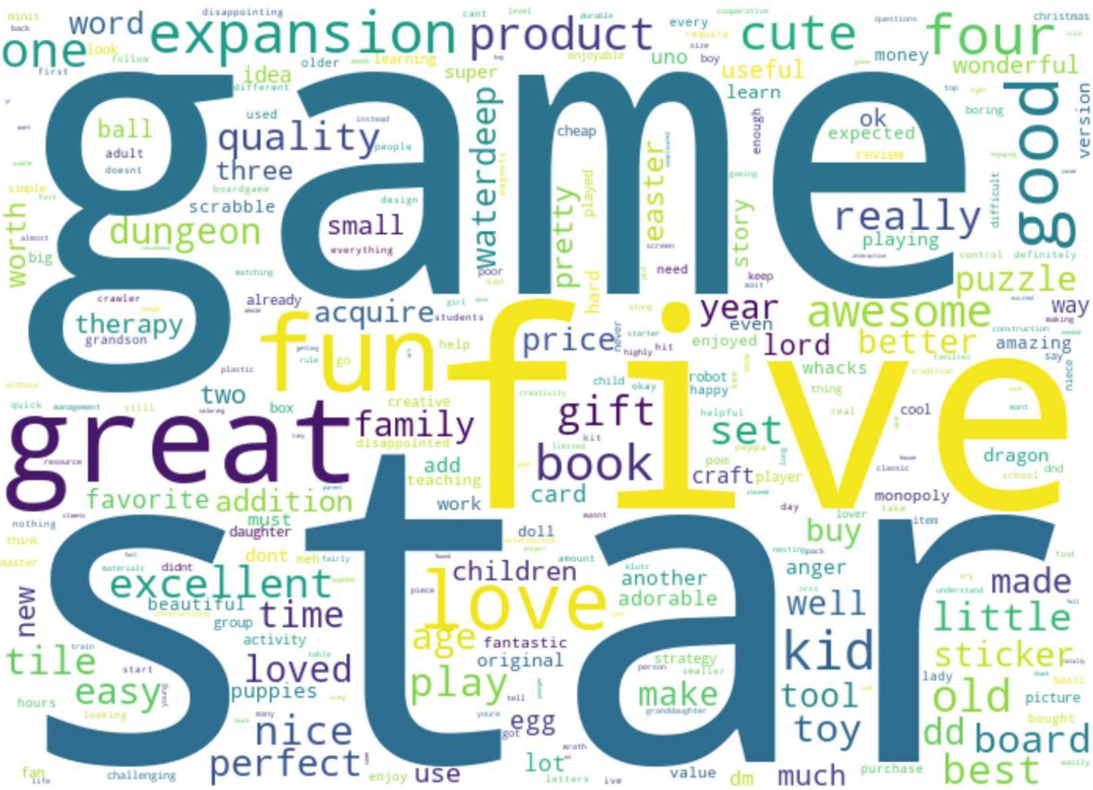

Figure 9-12 (Summaries – Wordcloud – with Stopwords)

R was chosen for tasks involving sales data analysis(turtle_sales.csv). RStudio provided an environment for data exploration, with redundant columns removed (Ranking, Year, Genre, Publisher), and key insights gained through scatterplots(Figures 10-1to10-12), histograms (Figures 11-1to11-3), and boxplots(Figures 12-1to12-3). Normality of the dataset was assessed using Q-Q plots (Figures13-1to13-3), skewness(Figure 14-1), kurtosis(Figure 14-1), and the Shapiro-Wilk test (Figure 15-1). Linear regression models were created to investigate relationships sales data (Correlations – Figure 16-1) , assessing goodness of fit and accuracy. (Figure 17-1 & Figure 18-1)

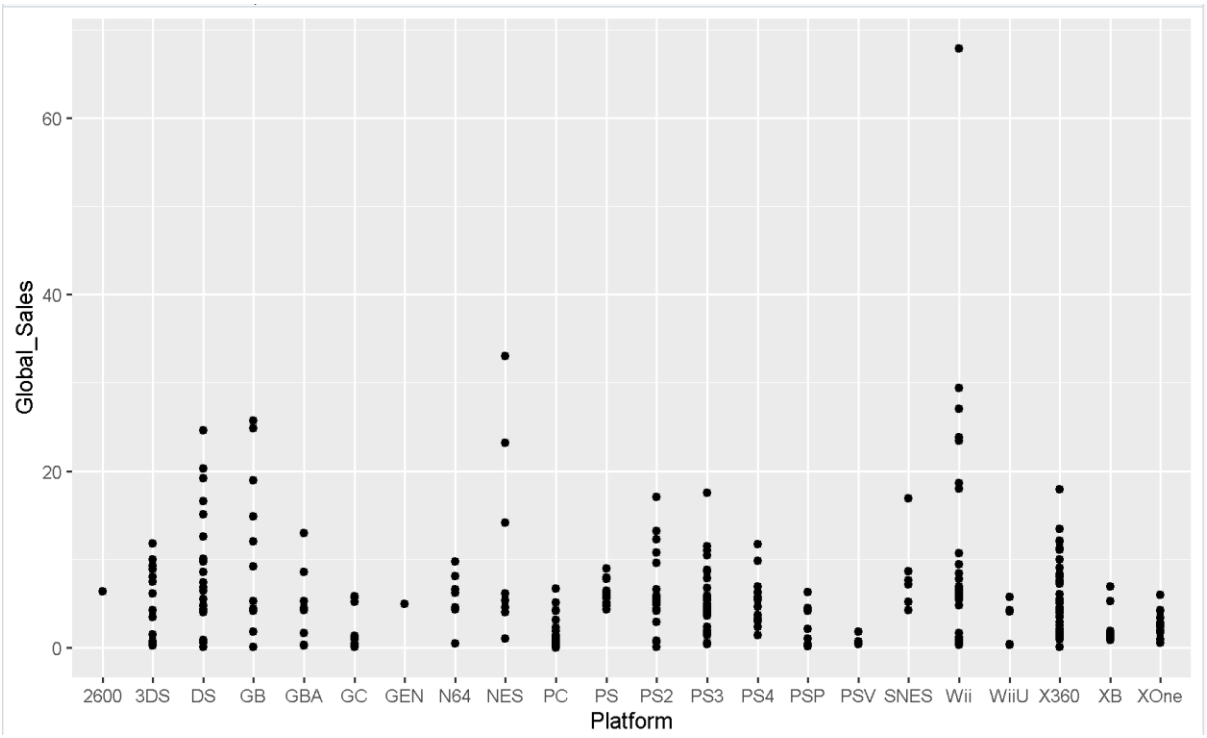

Figure 10-1

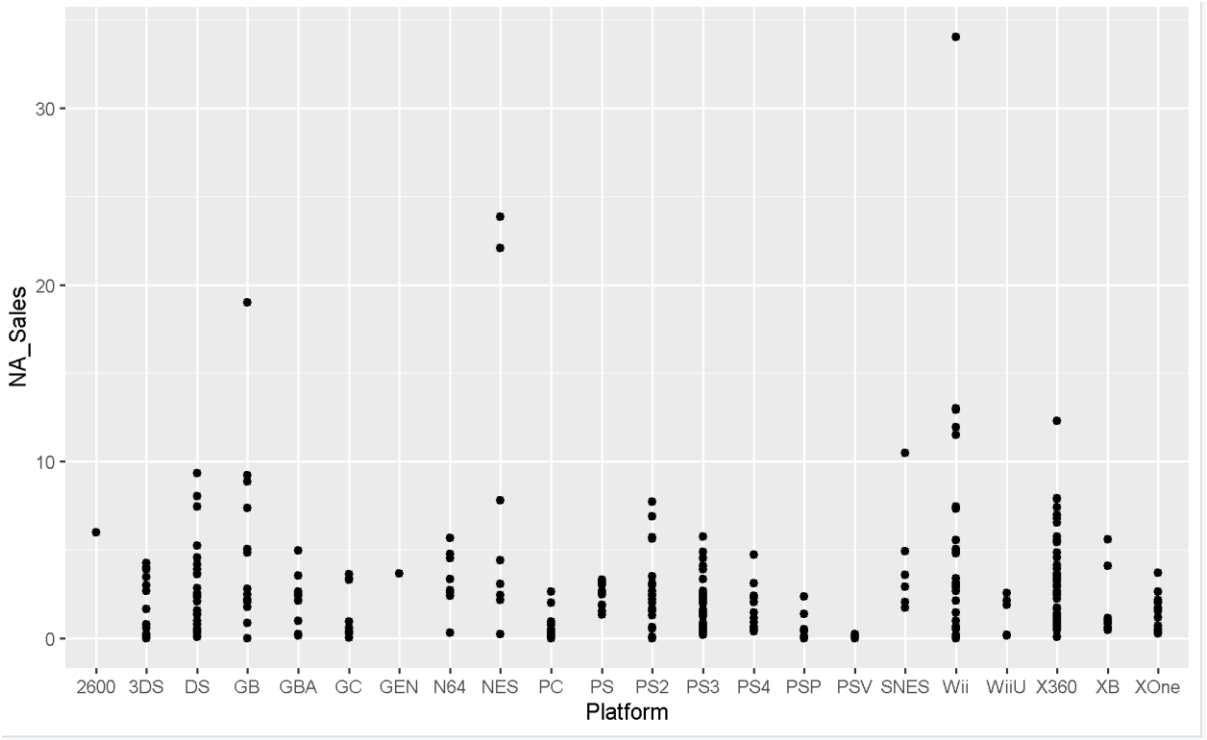

Figure 10-2

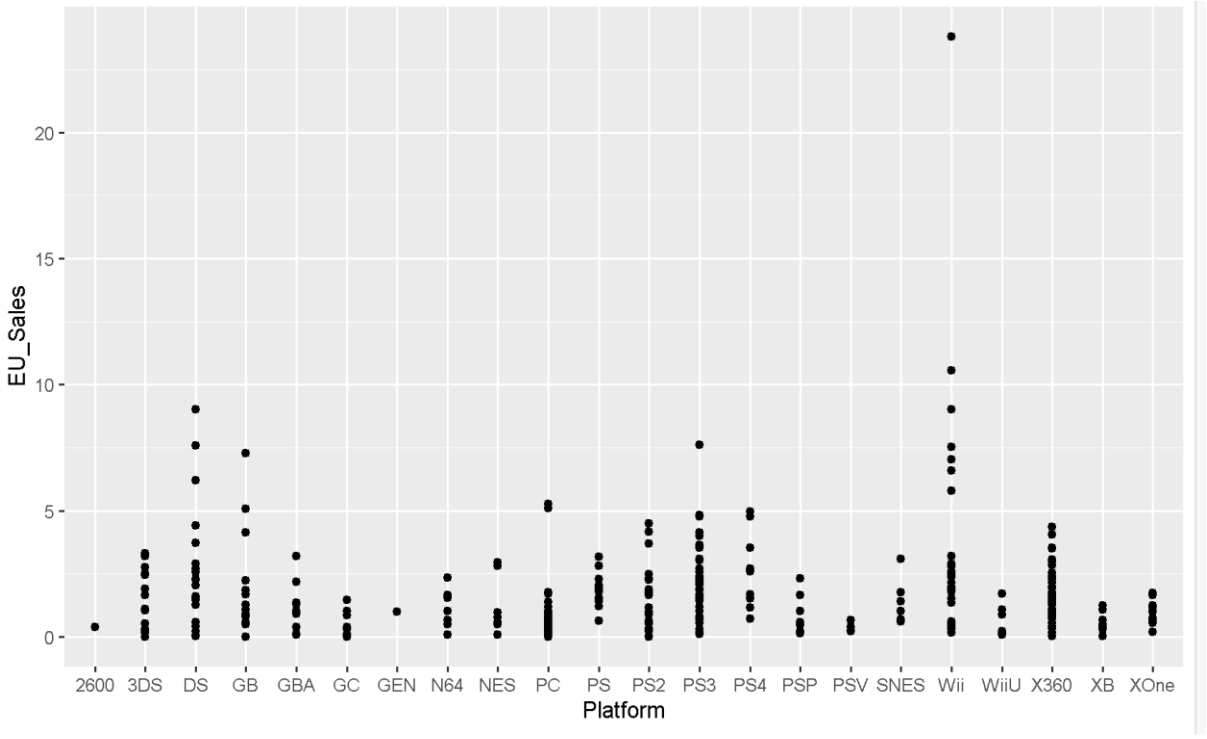

Figure 10-3

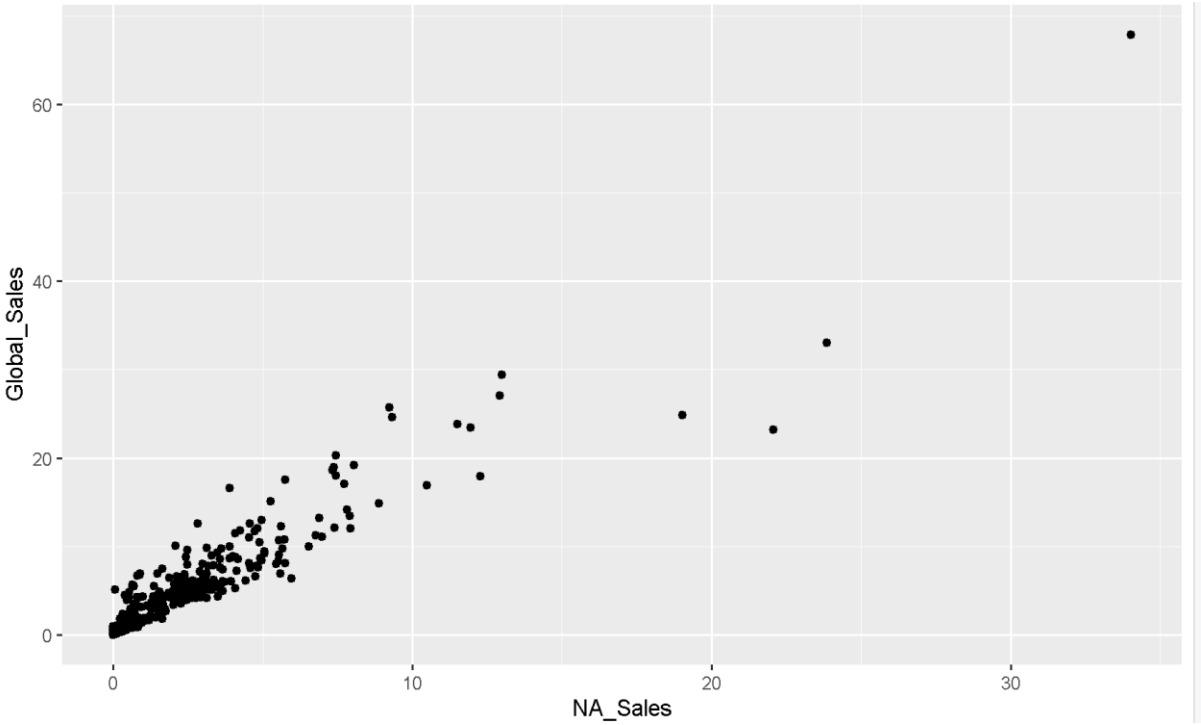

Figure 10-4

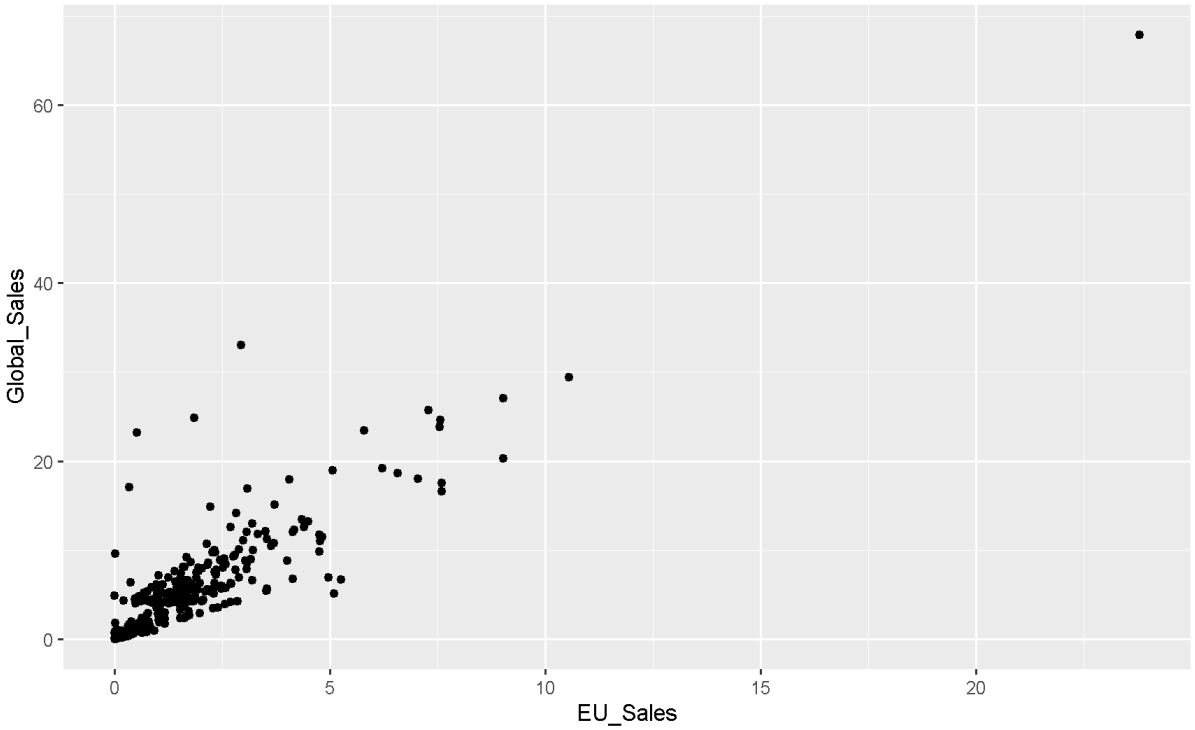

Figure 10-5

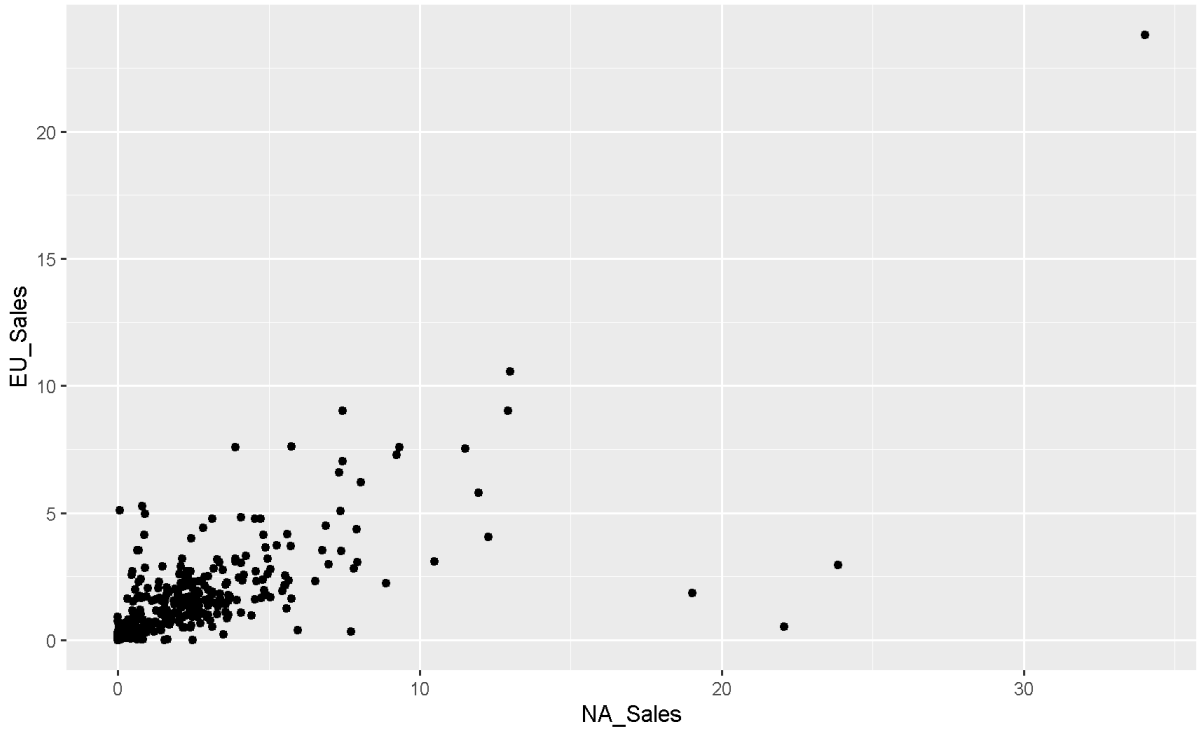

Figure 10-6

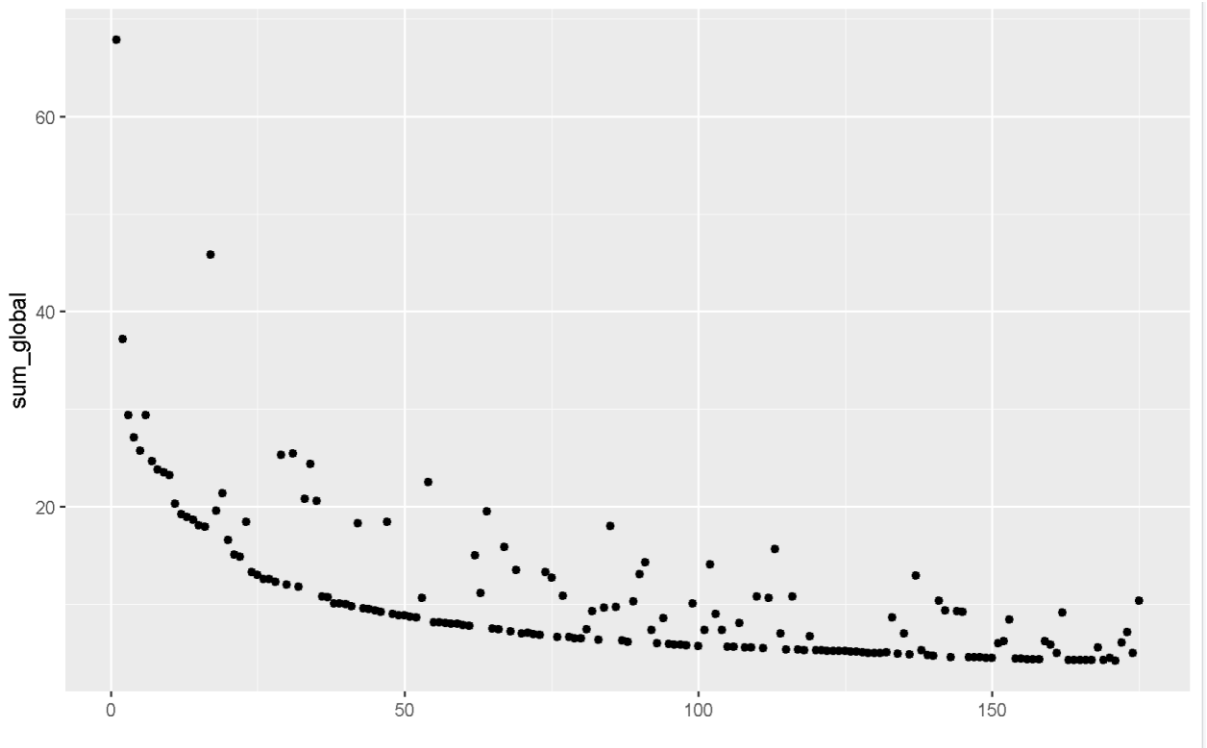

Figure 10-7

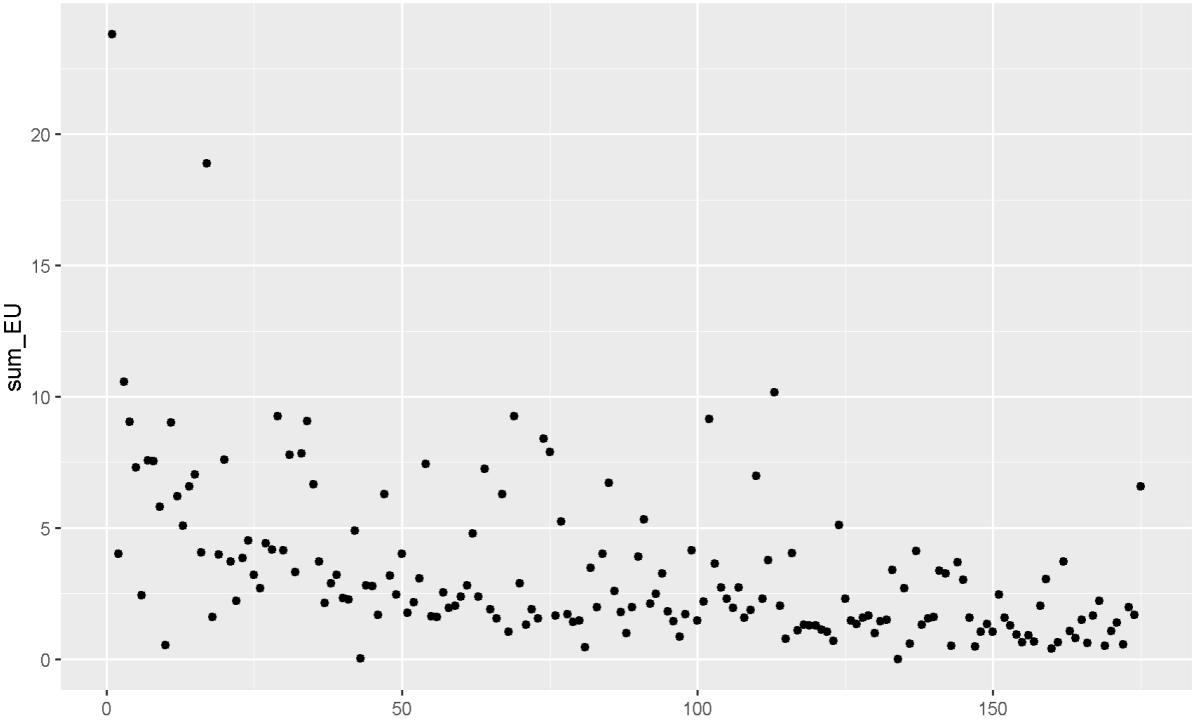

Figure 10-8

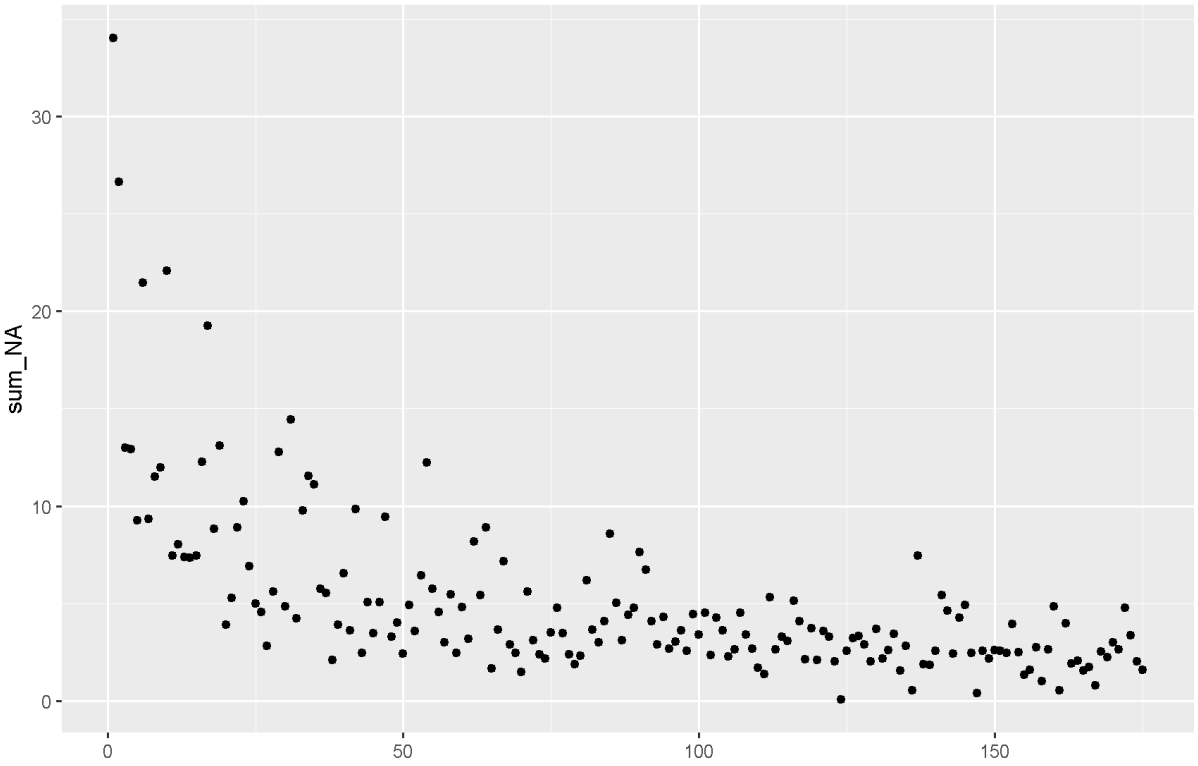

Figure 10-9

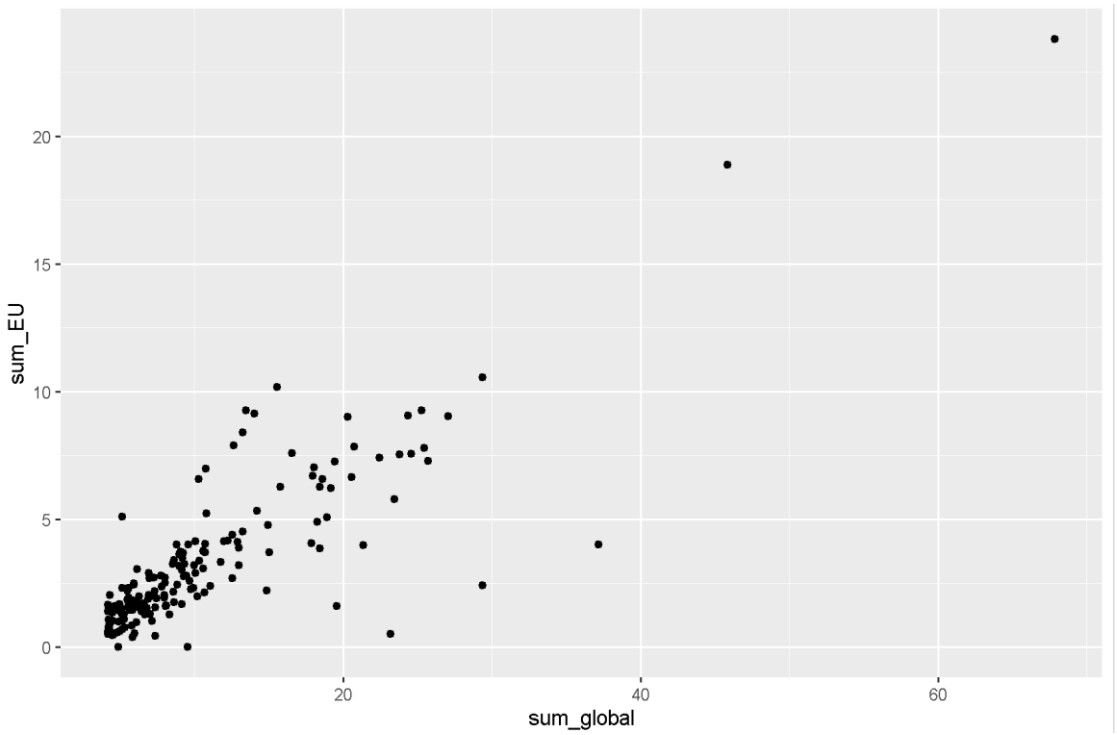

Figure 10-10

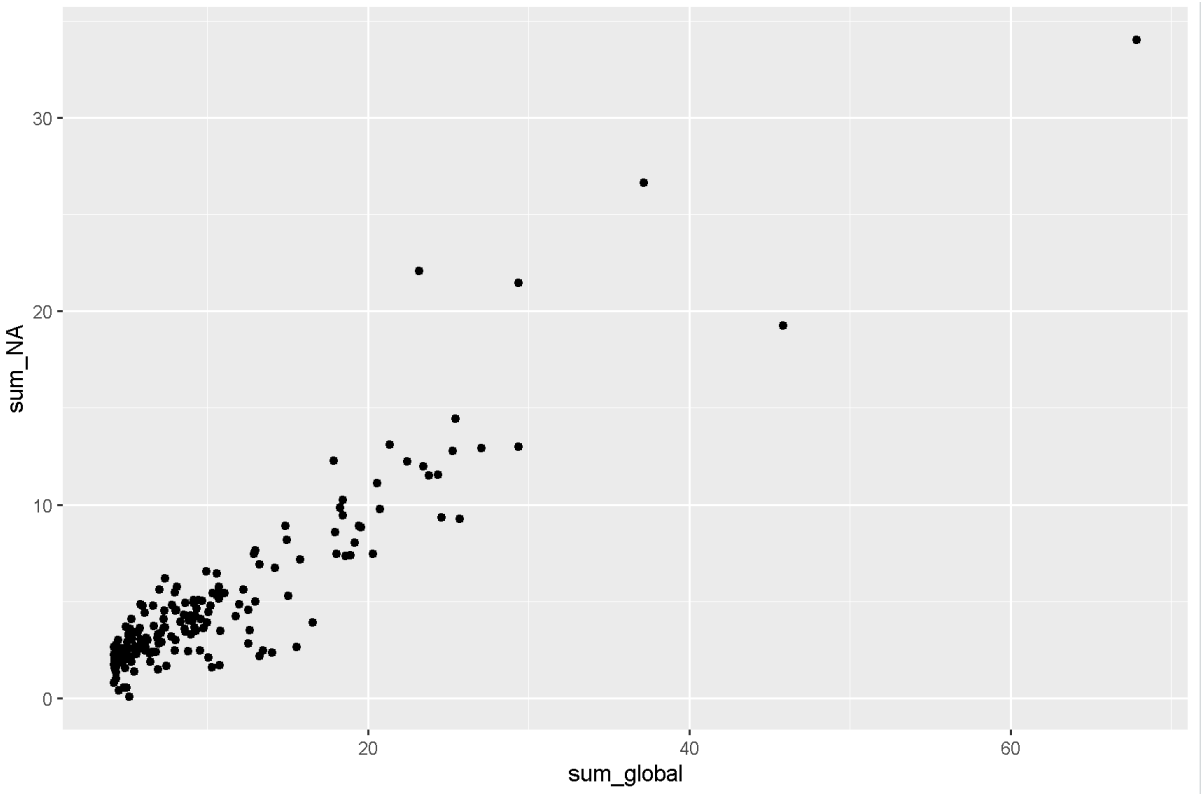

Figure 10-11

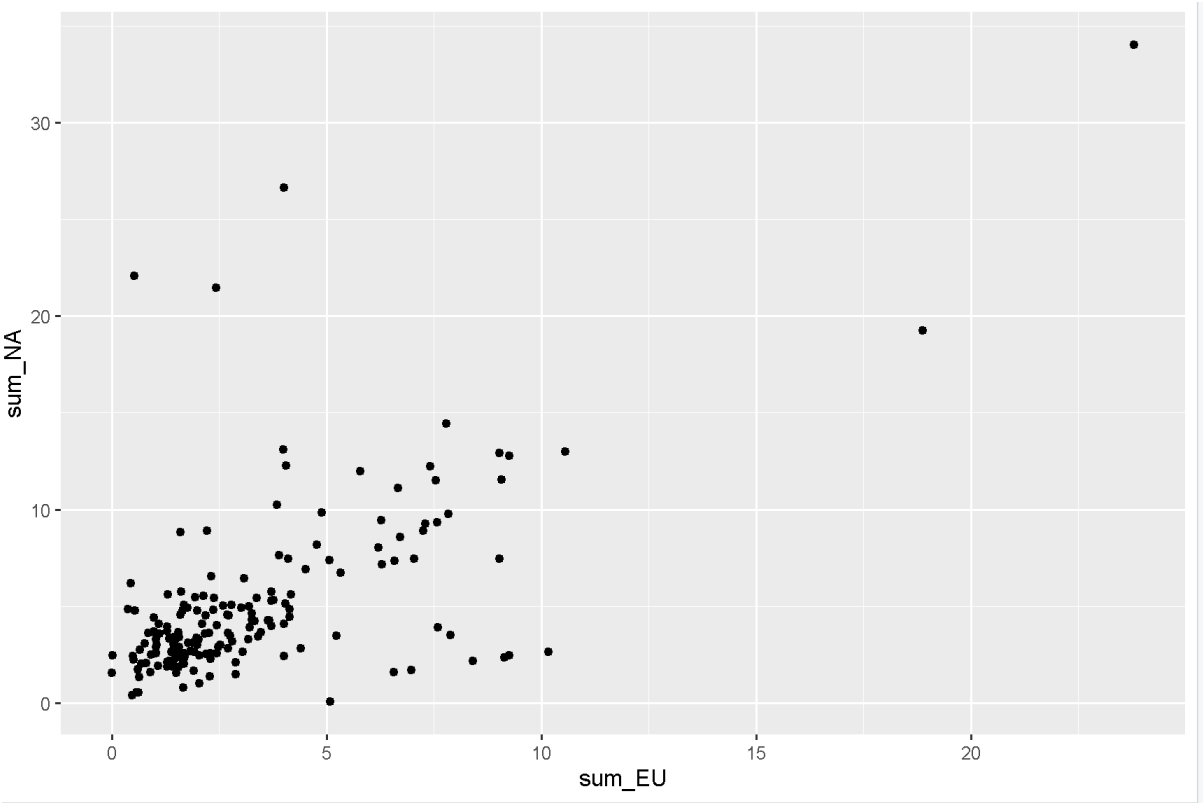

Figure 10-12

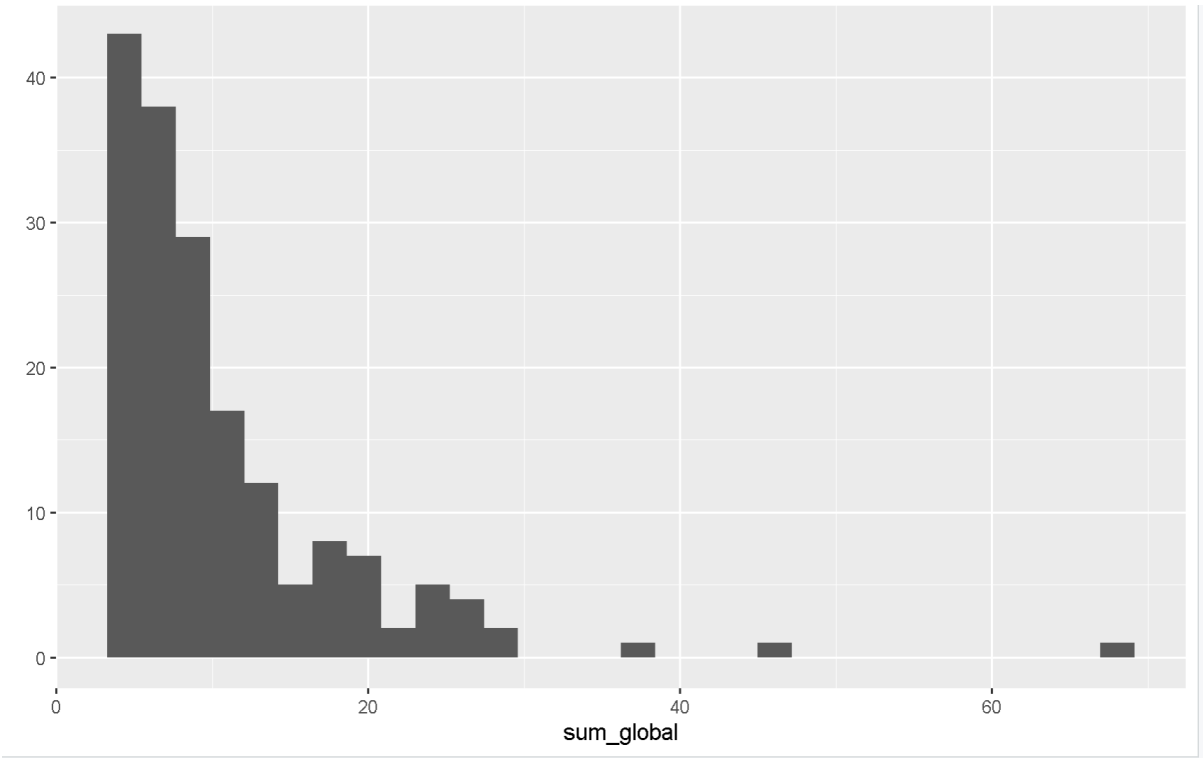

Figure 11-1

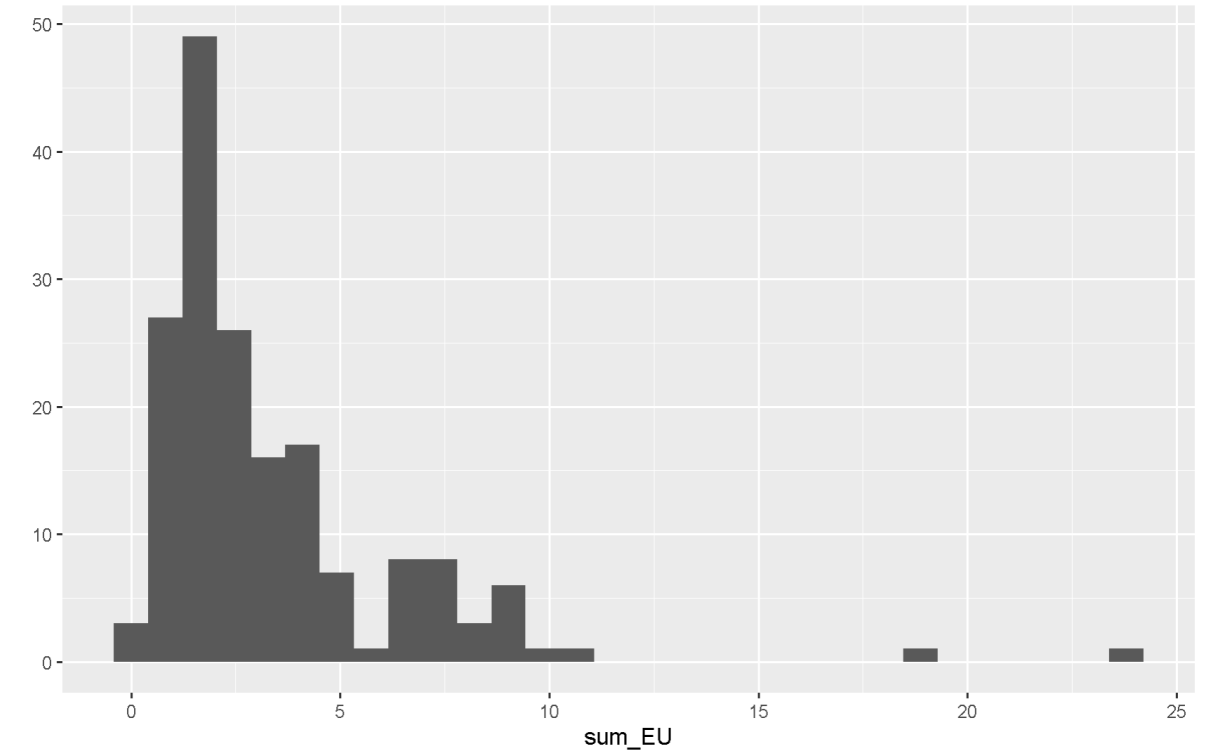

Figure 11-2

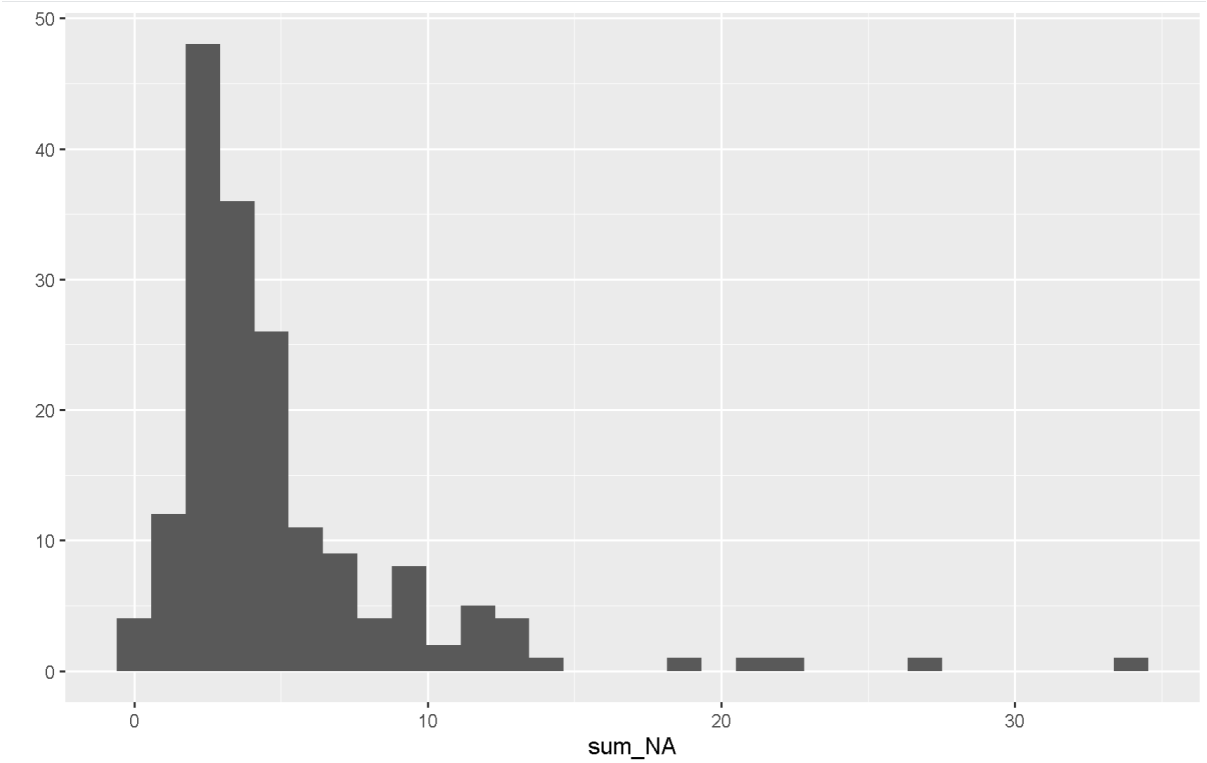

Figure 11-3

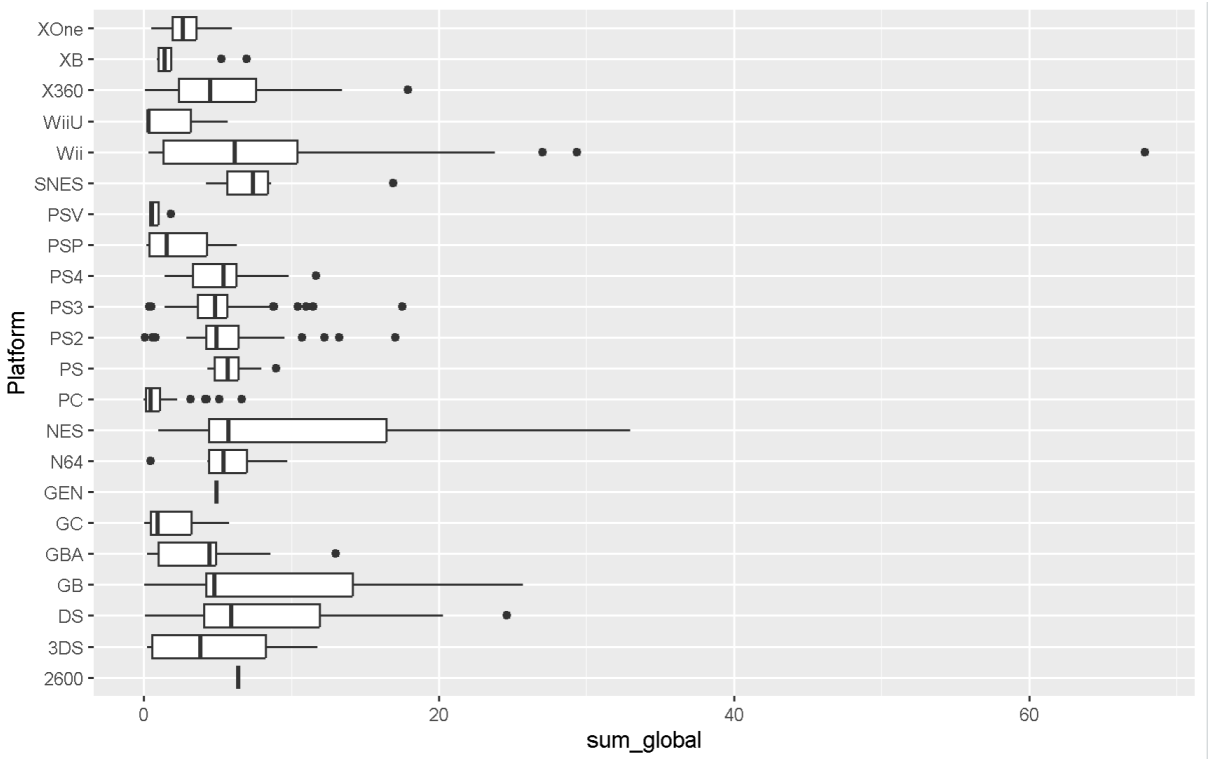

Figure 12-1

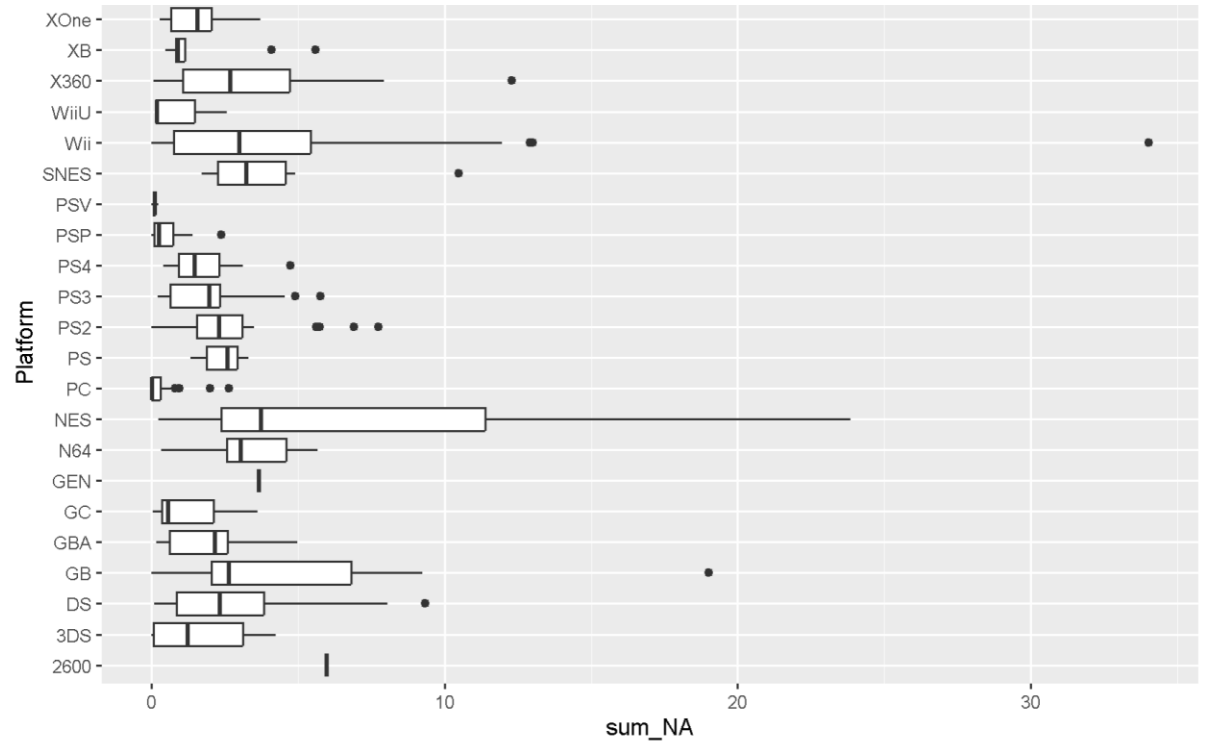

Figure 12-2

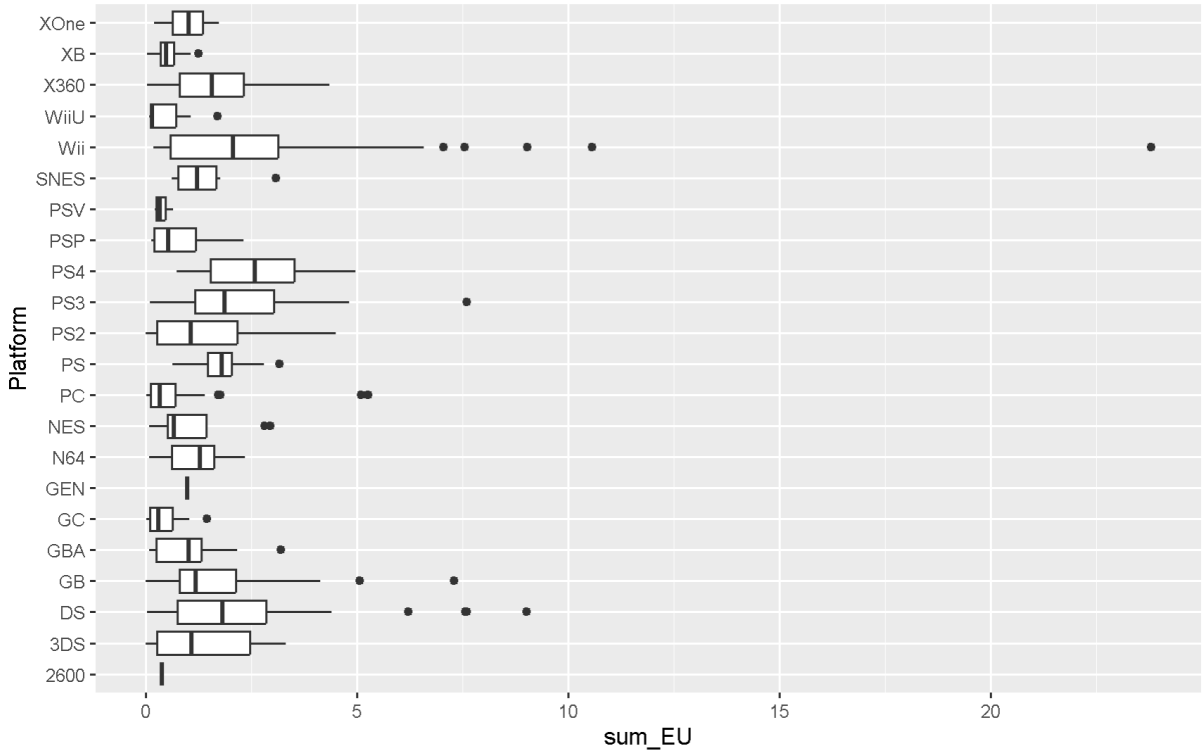

Figure 12-3

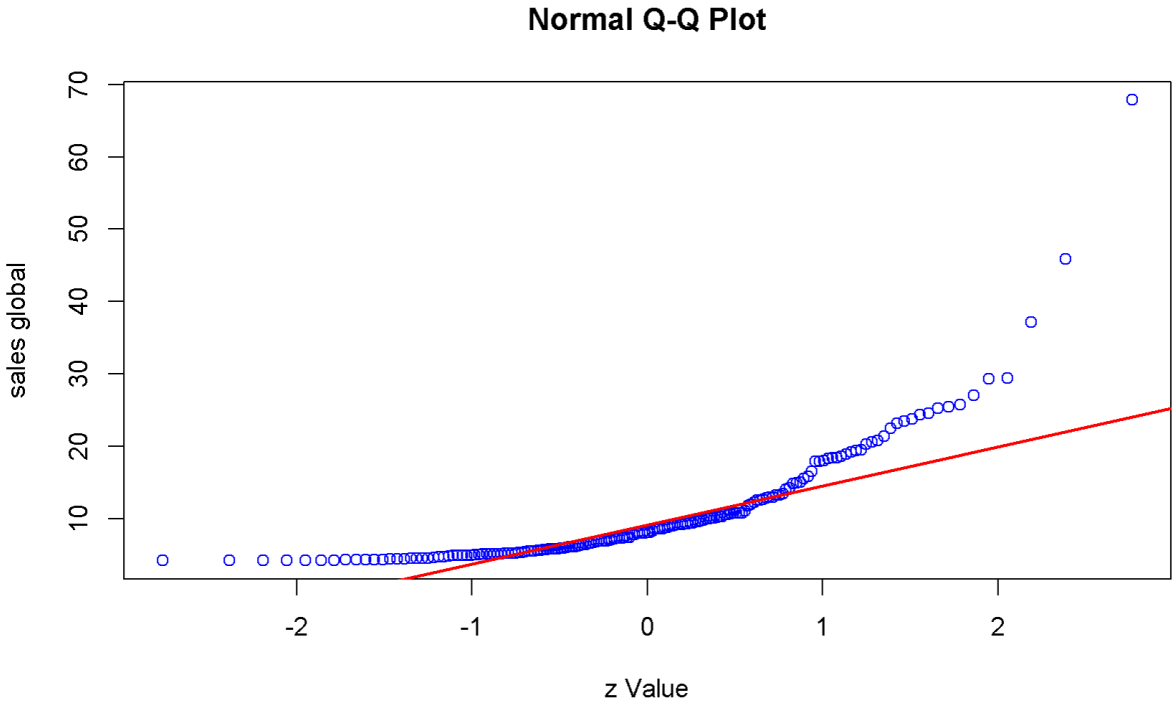

Figure 13-1

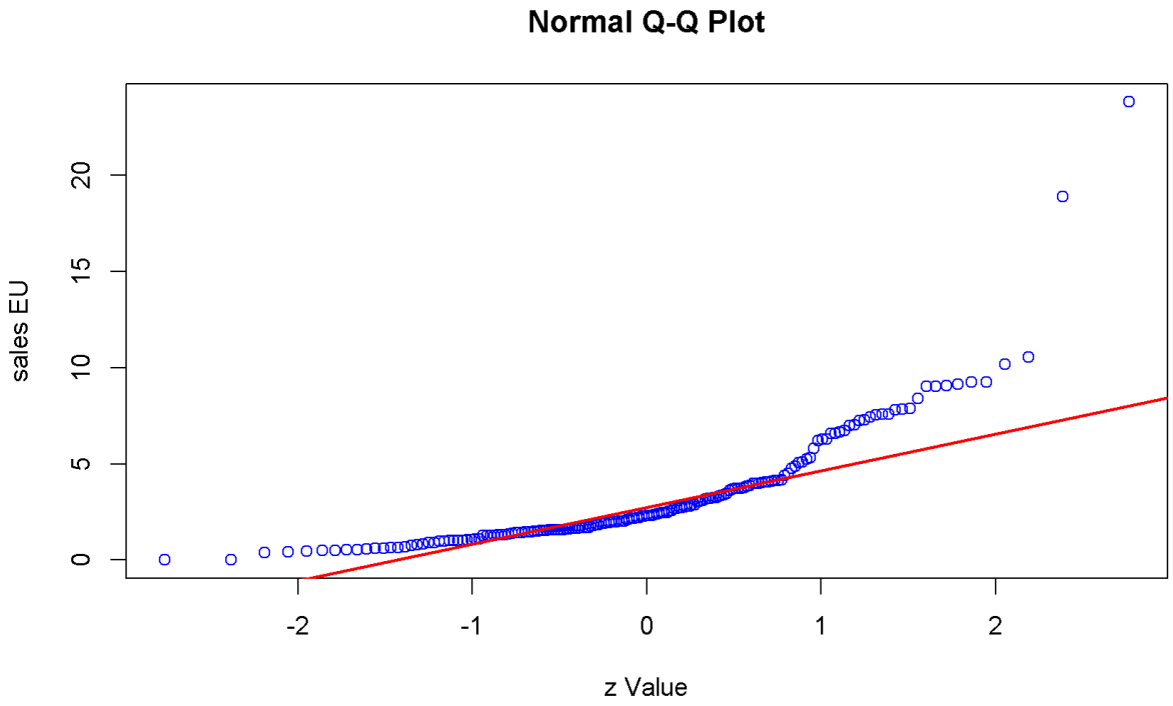

Figure 13-2

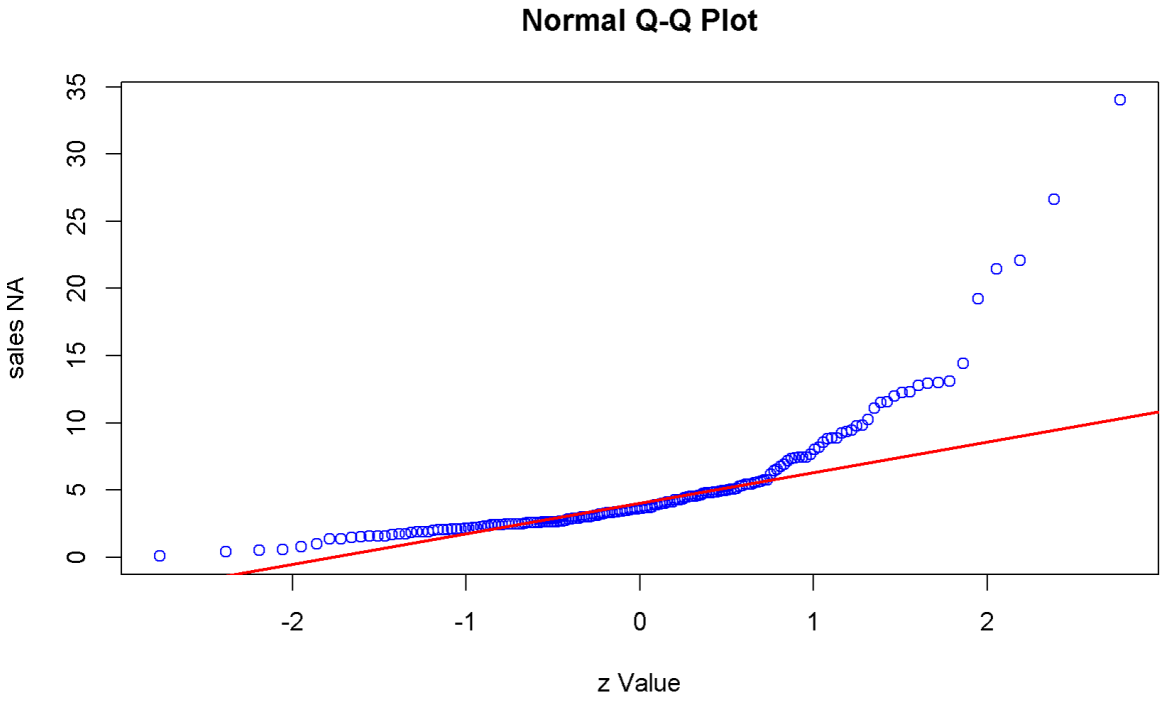

Figure 13-3

```
> skewness(df_s_p$sum_global)
[1] 3.066769
> kurtosis(df_s_p$sum_global)
[1] 17.79072
> skewness(df_s_p$sum_EU)
[1] 2.886029
> kurtosis(df_s_p$sum_EU)
[1] 16.22554
> skewness(df_s_p$sum_NA)
[1] 3.048198
> kurtosis(df_s_p$sum_NA)
[1] 15.6026
```

Figure 14-1

```
> shapiro.test(df_s_p$sum_NA)

        Shapiro-Wilk normality test

data:  df_s_p$sum_NA
W = 0.69813, p-value < 2.2e-16

> shapiro.test(df_s_p$sum_EU)

        Shapiro-Wilk normality test

data:  df_s_p$sum_EU
W = 0.74058, p-value = 2.987e-16

> shapiro.test(df_s_p$sum_global)

        Shapiro-Wilk normality test

data:  df_s_p$sum_global
W = 0.70955, p-value < 2.2e-16
```

Figure 15-1

```
> cor(df_s_p$sum_global, df_s_p$sum_EU)
[1] 0.8486148
> cor(df_s_p$sum_global, df_s_p$sum_NA)
[1] 0.9162292
> cor(df_s_p$sum_EU, df_s_p$sum_NA)
[1] 0.6209317
> round (cor(df_s_p),
+        digits=2)
           Product sum_NA sum_EU sum_global
Product       1.00  -0.54  -0.45      -0.61
sum_NA       -0.54   1.00   0.62       0.92
sum_EU       -0.45   0.62   1.00       0.85
sum_global   -0.61   0.92   0.85       1.00
```

Figure 16-1 (Correlations)

```
model2 <- lm(Product~sum_EU+sum_NA+sum_global, data=df_s_p)
summary(model2)

## 2b) Create a plot (simple linear regression)
# Basic visualisation.

qqnorm(residuals(model2))
qqline(residuals(model2), col='red')

#############################################################################

# 3. Create a multiple linear regression model
# Select only numeric columns from the original data frame.
modelEU <- lm(sum_EU~sum_NA+sum_global, data=df_s_p)
summary(modelEU)
modelNA <- lm(sum_NA~sum_EU+sum_global, data=df_s_p)
summary(modelNA)
```

Figure 17-1

```
# Creating a data frame
Predict_a <- data.frame(sum_EU = c(23.80), sum_NA =c(34.02), sum_global=c(63.3))
Predict_b <- data.frame(sum_EU = c(1.56), sum_NA =c(3.93), sum_global=c(6))
Predict_c <- data.frame(sum_EU = c(0.65), sum_NA =c(2.73), sum_global=c(5))
Predict_d <- data.frame(sum_EU = c(0.97), sum_NA =c(2.26), sum_global=c(4))
Predict_e <- data.frame(sum_EU = c(0.52), sum_NA =c(22.08), sum_global=c(26))
# Predicts the future values

# a) NA_Sales_sum of 34.02 and EU_Sales_sum of 23.80.
predict(modelEU, newdata = Predict_a, interval = 'confidence')
predict(modelNA, newdata = Predict_a, interval = 'confidence')
# b) NA_Sales_sum of 3.93 and EU_Sales_sum of 1.56.
predict(modelEU, newdata = Predict_b, interval = 'confidence')
predict(modelNA, newdata = Predict_b, interval = 'confidence')
# c) NA_Sales_sum of 2.73 and EU_Sales_sum of 0.65.
predict(modelEU, newdata = Predict_c, interval = 'confidence')
predict(modelNA, newdata = Predict_c, interval = 'confidence')
# d) NA_Sales_sum of 2.26 and EU_Sales_sum of 0.97.
predict(modelEU, newdata = Predict_d, interval = 'confidence')
predict(modelNA, newdata = Predict_d, interval = 'confidence')
# e) NA_Sales_sum of 22.08 and EU_Sales_sum of 0.52.
predict(modelEU, newdata = Predict_e, interval = 'confidence')
predict(modelNA, newdata = Predict_e, interval = 'confidence')
```

Figure 18-1

Visualisation and insights:

The selected visualizations were tailored to address Turtle Games' objectives effectively. I created all my visualisations in relation to the Turtle Games wants to find out. I created subheadings for each of the visualisations and I have described my rationale for the selected visualisations under each of them.

How customers accumulate loyalty points?

Linear regression plots illustrated the relationships between loyalty points and age/remuneration/spending scores, providing a clear visualization of potential predictive capabilities. (See figures 2-4, 3-4- 4-4). Per these visualisations as spending scores and renumeration increases, customer loyalty points increase, showing a positive correlation. And as age increases, customer loyalty points very very slightly decrease but this is a very weak correlation and therefore the relationship is inconclusive.

How groups within the customer base can be used to target specific market segments?

Per the pair plots visualisations (See figures 6-1&6-2) , you can see that graduates is a market segment that Turtle Games should target as the spending scores and renumeration are significantly higher compared to the other education types. Also females have a higher spending scores and renumeration compared males. So female graduates is specific segment that Turtle Games should target

How social data (e.g. customer reviews) can be used to inform marketing campaigns?

Per the histograms (figures 19-1to19-4), you can see the sentiment scores & polarity scores are mostly positive for reviews and summary. (See Figures 19-5&19-6 for the average polarity and sentiment scores) And per the word clouds that 'fun' and 'great' are showing as the most common words used. Additionally following the top 20 positive and negative reviews based on sentiment and polarity provides insights into what Turtle Games does best/worst. (Figures 9-1to9-8) Therefore these providing areas to improve upon in terms of meeting customer needs, therefore being able to market better to their customers. The average polarity of summary

Figure 19-1

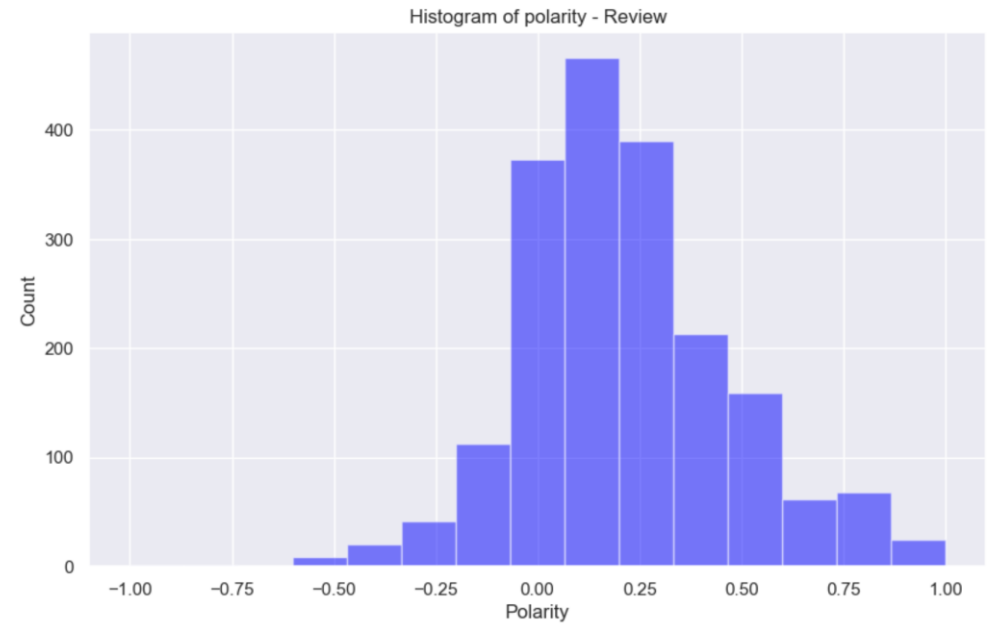

Figure 19-2

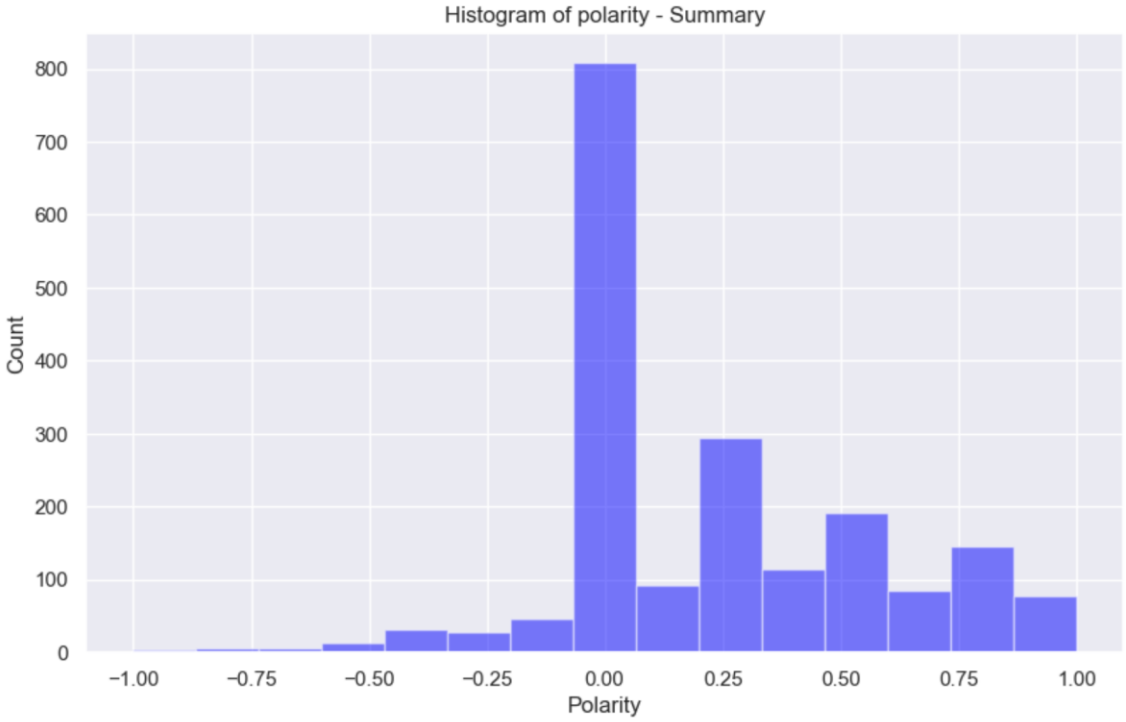

Figure 19-3

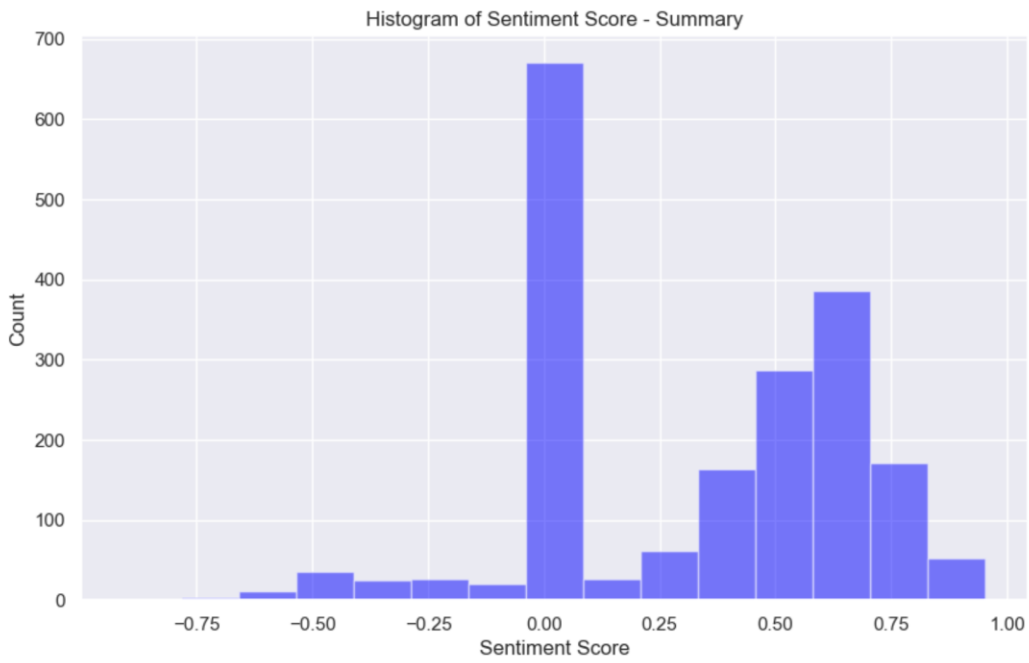

Figure 19-4

	polarity	sentiment_score
count	1940.000000	1940.000000
mean	0.210761	0.643193
std	0.259221	0.392889
min	-1.000000	-0.982800
25%	0.044472	0.535350
50%	0.175000	0.790600
75%	0.350000	0.911025
max	1.000000	0.999600

Figure 19-5 – Average scores – Reviews

```
sample_summary.describe()
```

	polarity	sentiment_score
count	1940.000000	1940.000000
mean	0.226100	0.316243
std	0.338521	0.344831
min	-1.000000	-0.905200
25%	0.000000	0.000000
50%	0.100000	0.440400
75%	0.500000	0.624900
max	1.000000	0.952400

Figure 19-6 – Average scores – Summary

The impact that each product has on sales

Per figures 12-1-12-3 the boxplots shows that the various gaming platforms have very varied sales based on region, shown by NES having significantly long whiskers in North America compared to the EU & global sales.

How reliable the data is? (e.g. normal distribution, skewness, or kurtosis)?

Per Figures 14-1&15-1, the skewness values suggest a right-skewed distribution due to the positive value. The kurtosis values are quite high, indicating heavy tails or more extreme values in the distribution compared to a normal distribution. The p-values are extremely small, much smaller than the significance level, which suggests strong evidence against a null hypothesis. Per Figures 13-1to13-3, the points deviate from the line, especially at the tails, it suggests departure from normality and therefore not normally distributed. Please see figures 20-1to20-6 which shows the linear regression models normal Q-Q plots which also share the same conclusions as made above with figures 13-1to 13-3.

What the relationship(s) is/are (if any) between North American, European, and global sales?

Per figure 16-1 showing the correlations between North American, European, and global sales, it shows that global sales has a strong positive correlation between European sales and North American Sales (0.85 and 0.92 respectively). Therefore as global sales increase then North American European sales will likely increase as a result and visa versa. However, an increase sales in Europe doesn't always translate to sales in North America, it is still a positive correlation but as strong compared to global sales (0.62 respectively).

```
Call:
lm(formula = Product ~ sum_EU + sum_NA + sum_global, data = df_s_p)

Residuals:
    Min      1Q  Median      3Q     Max
-2876.3 -1331.2  -532.2  1406.4  6065.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5638.41     243.67  23.140  < 2e-16 ***
sum_EU        451.86     128.79   3.508 0.000576 ***
sum_NA        337.30     115.10   2.930 0.003847 **
sum_global   -498.54      95.59  -5.215 5.26e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1868 on 171 degrees of freedom
Multiple R-squared:  0.4107,    Adjusted R-squared:  0.4004
F-statistic: 39.72 on 3 and 171 DF,  p-value: < 2.2e-16
```

Figure 20-1

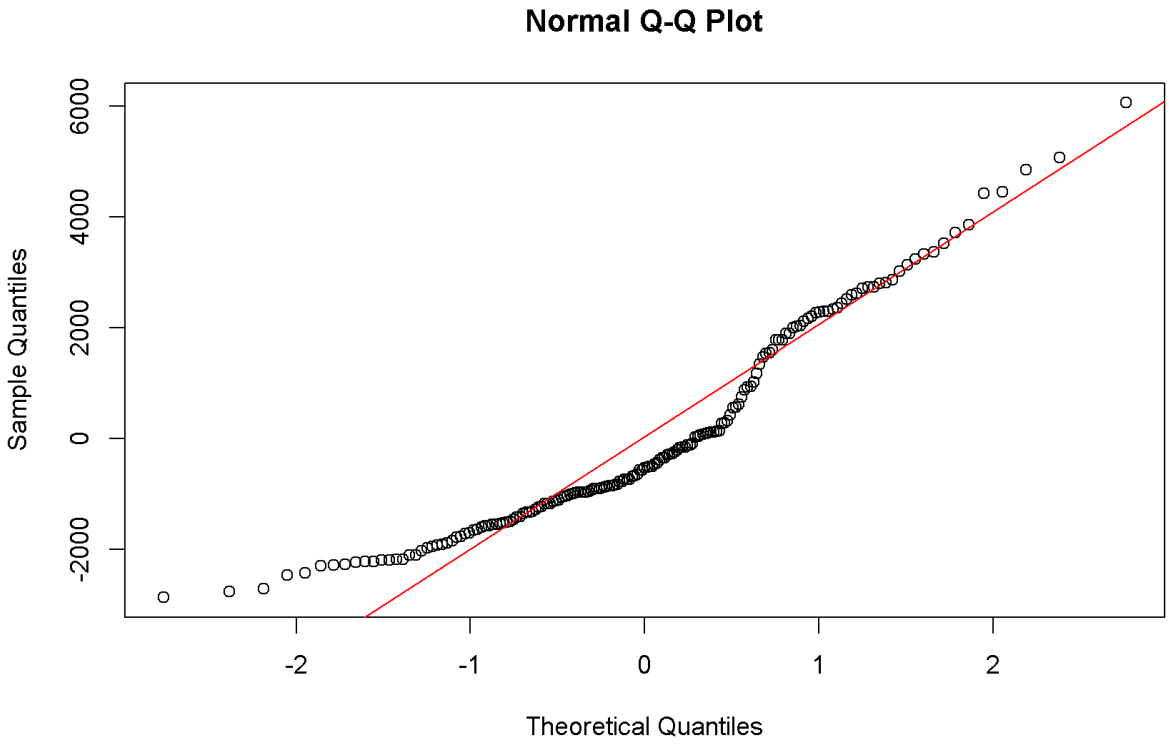

Figure 20-2

```
Call:
lm(formula = sum_EU ~ sum_NA + sum_global, data = df_s_p)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0541 -0.4622  0.1874  0.7434  2.3991

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.44509    0.14021  -3.175  0.00178 **
sum_NA      -0.66028    0.04592 -14.378  < 2e-16 ***
sum_global   0.66101    0.02574  25.682  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.106 on 172 degrees of freedom
Multiple R-squared:  0.8729,    Adjusted R-squared:  0.8714
F-statistic: 590.7 on 2 and 172 DF,  p-value: < 2.2e-16
```

Figure 20-3

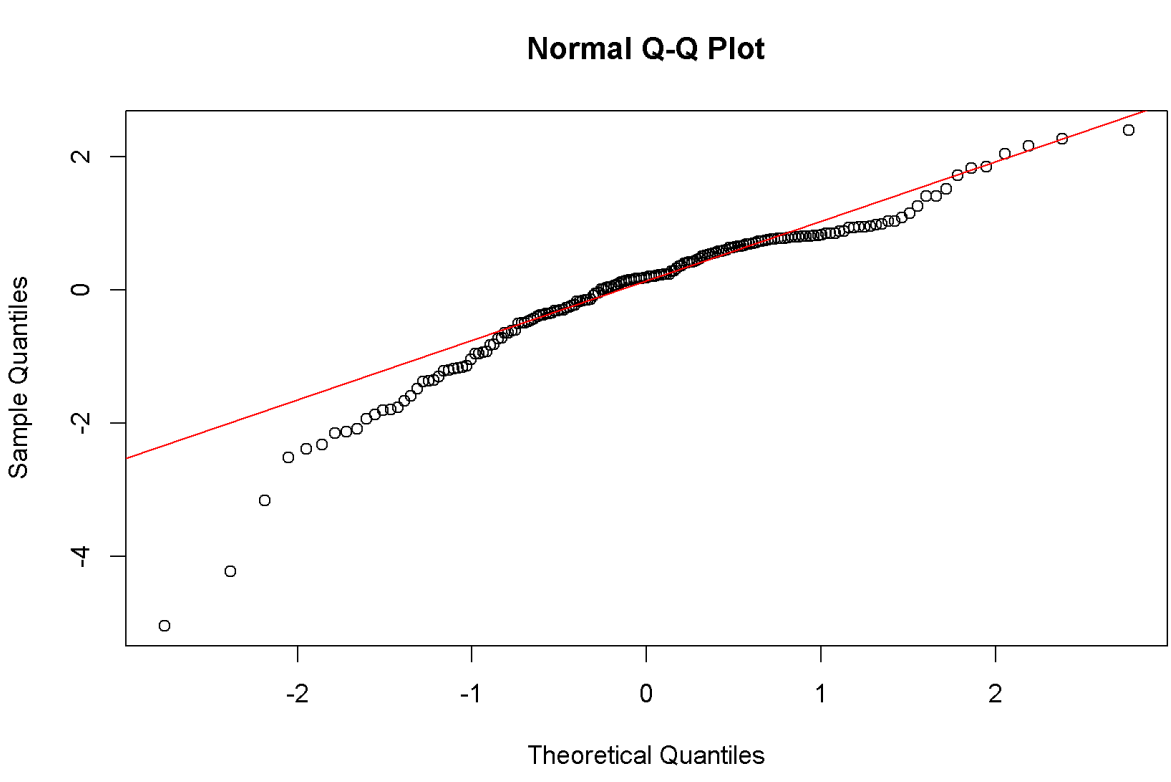

Figure 20-4

```
Call:
lm(formula = sum_NA ~ sum_EU + sum_global, data = df_s_p)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5257 -0.5242  0.1496  0.8119  4.9856

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.57221    0.15541  -3.682  0.00031 ***
sum_EU      -0.82671    0.05750 -14.378  < 2e-16 ***
sum_global   0.77969    0.02181  35.745  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.237 on 172 degrees of freedom
Multiple R-squared:  0.9271,    Adjusted R-squared:  0.9263
F-statistic:  1094 on 2 and 172 DF,  p-value: < 2.2e-16
```

Figure 20-5

\

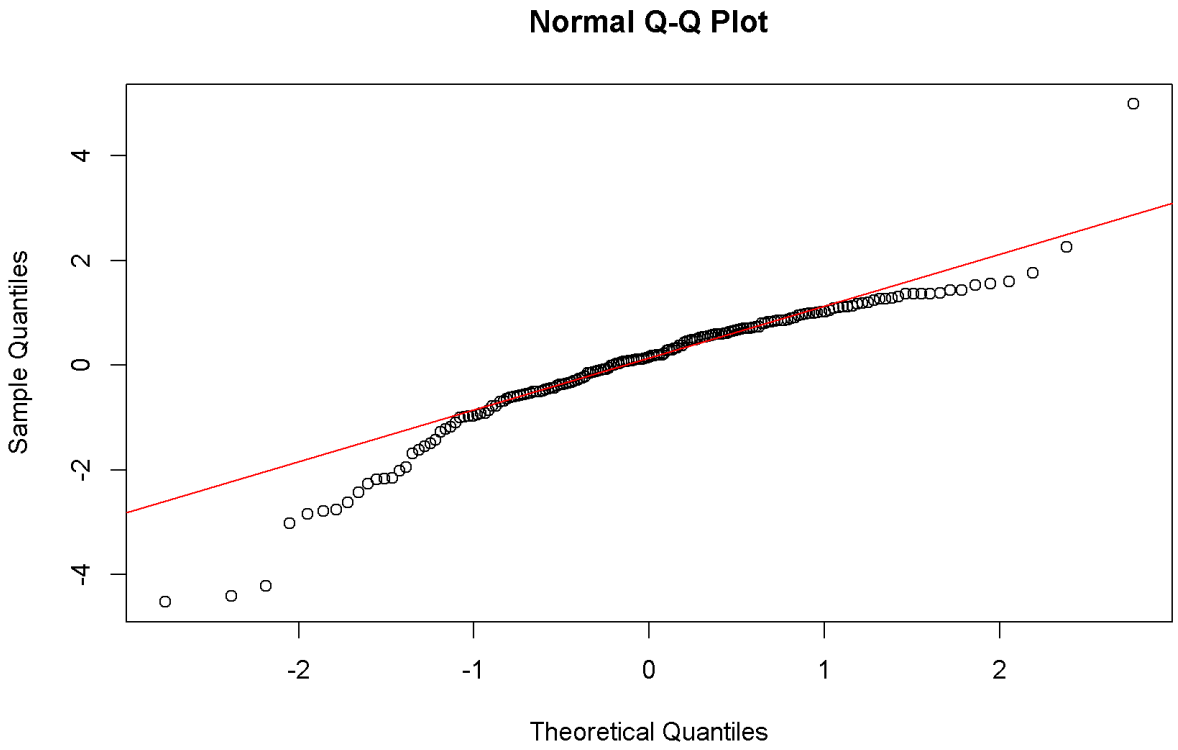

Figure 20-6

```
> # a) NA_Sales_sum of 34.02 and EU_Sales_sum of 23.80.
> predict(modelEU, newdata = Predict_a, interval = 'confidence')
       fit      lwr      upr
1 18.93437 17.83715 20.03159
> predict(modelNA, newdata = Predict_a, interval = 'confidence')
       fit      lwr      upr
1 29.10612 27.82868 30.38355
> # b) NA_Sales_sum of 3.93 and EU_Sales_sum of 1.56.
> predict(modelEU, newdata = Predict_b, interval = 'confidence')
       fit       lwr     upr
1 0.926096 0.7017121 1.15048
> predict(modelNA, newdata = Predict_b, interval = 'confidence')
      fit      lwr      upr
1 2.81623 2.600969 3.031491

> # c) NA_Sales_sum of 2.73 and EU_Sales_sum of 0.65.
> predict(modelEU, newdata = Predict_c, interval = 'confidence')
       fit       lwr      upr
1 1.057415 0.8478759 1.266954
> predict(modelNA, newdata = Predict_c, interval = 'confidence')
       fit      lwr      upr
1 2.788854 2.544715 3.032992
> # d) NA_Sales_sum of 2.26 and EU_Sales_sum of 0.97.
> predict(modelEU, newdata = Predict_d, interval = 'confidence')
        fit       lwr       upr
1 0.7067328 0.4841957 0.9292698
> predict(modelNA, newdata = Predict_d, interval = 'confidence')
       fit     lwr      upr
1 1.744619 1.50388 1.985358

> # e) NA_Sales_sum of 22.08 and EU_Sales_sum of 0.52.
> predict(modelEU, newdata = Predict_e, interval = 'confidence')
       fit      lwr      upr
1 2.162329 1.259071 3.065588
> predict(modelNA, newdata = Predict_e, interval = 'confidence')
       fit      lwr      upr
1 19.26973 18.31101 20.22845
```

Patterns and predictions:

Patterns in customer behaviour and potential predictions emerged from the analyses. Linear regression highlighted significant relationships between loyalty points and renumeration/spending variables.

The k-means clustering identified distinct customer segments based on remuneration and spending scores, providing valuable insights for targeted marketing towards female graduates as mentioned previously.

Sales data exploration revealed patterns in regional sales, guiding potential marketing strategies. The normality assessment indicated the distribution characteristics of sales data, aiding in predicting future outcomes. Linear regression models provided insights into the impact of different variables on global sales, enabling data-driven decision-making.

The data analyses conducted for Turtle Games offer valuable insights into customer behaviour, product impact on sales, and opportunities for targeted marketing. The combination of Python and R allowed us to explore diverse aspects of the business, providing a comprehensive understanding of the data landscape. Further exploration could involve deeper sentiment analysis of customer reviews

and refining predictive models for more accurate sales forecasts. Overall, the findings support Turtle Games in making data-driven decisions to enhance sales performance.