Appendix 1:

Linear regression results interpreation for analysis question 1:
Factors affecting loyalty point accumulation

Age does not appear to influence loyalty point accumulation:

- R Squared is 0.002 which suggests Age explains only 0.2% of loyalty points
- In addition the negative x coeffient of -0.004 shows no / negative changes in age for each loyalty point unit change
- The P-value of 0.577 is greater that the 0.05 significance threshold / level. Therefore it is not possible to reject the null hypthesis that any correlation occured by chance
- The Durbin-Watson test produces a value of 2.129 suggesting no autocorrelation between residuals
- Skew is 0.574 indicating moderate positive / right skewness
- Kurtosis is 2.184 indicating light tailed, playtykurtic distribution
- However as kurtosis is close the perfect (value of 3), it can be inferred that near normal distribution is present
- The standard error is very low producing a value of 0.000236 indicating low dispersal from line of best fit. Therefore model is a good fit
- The t probability value is 0.577. This value is greater than the 0.05 threshold level suggesting that the age variable does not have a significant impact of loyalty point levels

Salary shows correlation with loyalty point accumulation:

- R Squared is 0.380 which means 38% of Salary variable can explain loyalty point levels
- In addition the x coeffient of 0.111 indicates salary increases by £11.10 for each unit loyalty point unit change
- The P-value is extremely low at 2.43e-209, much lower that the 0.05 confidence level - indicating a strong correlation between salary score and loyalty points
- Interestingly, the Durbin-Watson value of 1.461 indicates possible autocorrelation between residuals is present
- Skew is 1.230 which indicates a higher positive right skew
- Kurtosis is 4.357 which indicates a heavy tailed leptokurtic distribution
- The standard error is very low at 0.000318 which indicates that model regression line is a good fit
- The t probability value is zero indicating a very strong correlation between salary and the accumulation of loyalty points

Spending score shows most correlation with loyalty point accumulation:

- R Squared is 0.452 which suggest customer spend 45.2% explains levels of loyalty points.
- The x coefficient is 0.0137 indicating that spend score increases with levels of loyalty points
- The P-value is extremely low at 2.92e-263 - much lower that the 0.05 confidence level - indicating a very strong correlation between spending score and loyalty points.
- The Durbin-Watson test produces a value of 2.599 suggesting no autocorrelation between residuals.
- Skew is 0.768 indicating a moderate positive / right skew.
- Kurtosis is 3.441. This indicates a fairly normal distribution which is slightly heavytailed.
- The standard error is 0.000337. This very low value indicates that regression line is a good fit.
- The probability t value is 0.0137 indicating a strong correlation between salary and the accumulation of loyalty points.

Appendix 2:

K-means clustering methodology within Python for analysis question 2: Determine customer clusters to inform customer segmentation

1. Load and use cleaned data set from linear regression activity

2. Retain relevant variables to assist identification of customer market segments:
Retain variables: gender, salary, spend score and education

3. Explore significance of gender and education variables:
Create exploratory visualisations to inform approach: scatter and pair plots

4. Employ Elbow and Silhouette approaches to establish best fitting cluster size

5. Test and visualise range of cluster sizes

6. Select and deploy optimum k-means cluster model.

7. Final model choice justification: preferred K number is 5

- 5 represents ideal point in both Elbow and Silhouette methods
- When testing 4 clusters the pair plot shows how to separate clusters are merged incorrectly - this is corrected when 5 clusters are used. The 5 cluster number appears to best fit the distribution of dots
- The value of 7 was also tested as on the Silhouette methods shows another point on the plot where a significant change is line plot can be observed. However when plotted, the additional 2 clusters appears to be drawn from existing clusters because they appear more as outliers. However, they appear to better fit with the 5 clusters rather than represent new clusters in their own right.

Appendix 3: 20 most positive summaries

Review	Polarity	Subjectivity
best gm screen ever	1.00	0.30
wonderful designs	1.00	1.00
perfect	1.00	1.00
theyre the perfect size to keep in the car or a diaper	1.00	1.00
perfect for preschooler	1.00	1.00
awesome sticker activity for the price	1.00	1.00
awesome book	1.00	1.00
he was very happy with his gift	1.00	1.00
awesome	1.00	1.00
awesome and welldesigned for 9 year olds	1.00	1.00
excellent	1.00	1.00
excellent therapy tool	1.00	1.00
the pigeon is the perfect addition to a school library	1.00	1.00
best easter teaching tool	1.00	0.30
wonderful	1.00	1.00
all f the mudpuppy toys are wonderful	1.00	1.00
awesome puzzle	1.00	1.00
not the best quality	1.00	0.30
excellent puzzle	1.00	1.00
the best feedback i can have	1.00	0.30

Appendix 4: 20 most negative summaries

Review	Polarity	Subjectivity
the worst value ive ever seen	-1.00	1.00
boring unless you are a craft person which i am	-1.00	1.00
boring	-1.00	1.00
before this i hated running any rpg campaign dealing with towns because	-0.90	0.70
another worthless dungeon masters screen from galeforce9	-0.80	0.90
disappointed	-0.75	0.75
promotes anger instead of teaching calming methods	-0.70	0.20
too bad this is not what i was expecting	-0.70	0.66
bad qualityall made of paper	-0.70	0.66
at age 31 i found these very difficult to make	-0.65	1.00
small and boring	-0.62	0.70
mad dragon	-0.62	1.00
disappointing	-0.60	0.70
then you will find this board game to be dumb and boring	-0.59	0.63
anger control game	-0.55	0.30
really small disappointed	-0.50	0.57
its uno for the angry	-0.50	1.00
50th anniversary is a sad day for acquire	-0.50	1.00
a disappointing coop game	-0.50	0.55
its also really lame that the doll didnt come with the things she	-0.50	0.75

Appendix 5:

NLP Python Methodology for analysis question 3:
How customer review social data inform marketing campaigns

1. Load and explore the data used within weeks one and two analysis activities:

- Sense-check
- Subset data to focus on relevant variables: review and summary
- Check for missing values

2. Prepare review and summary data for NLP:

- Change to lower case
- Replace punctuation
- Drop duplicates

3. Tokenise and create wordclouds for review and summary variables

4. Frequency distribution and polarity:

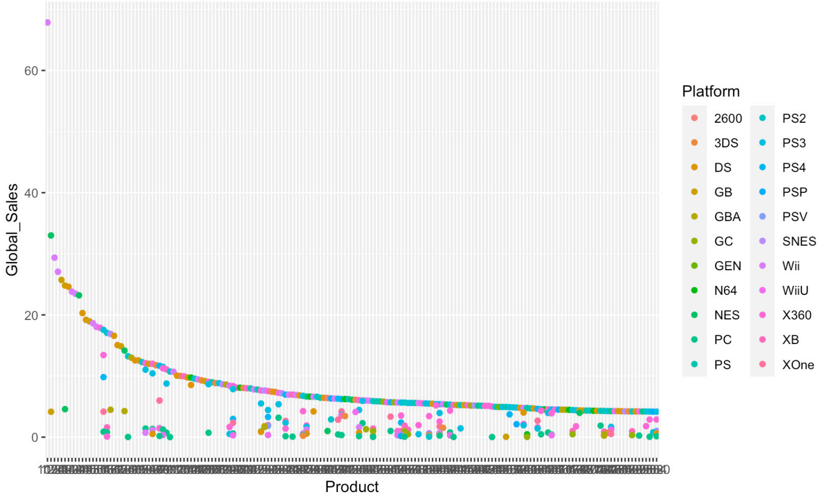
- Create frequency distribution
- Remove alphanumeric characters and stopwords
- Create wordcloud without stopwords
- Identify 15 most common words and polarity

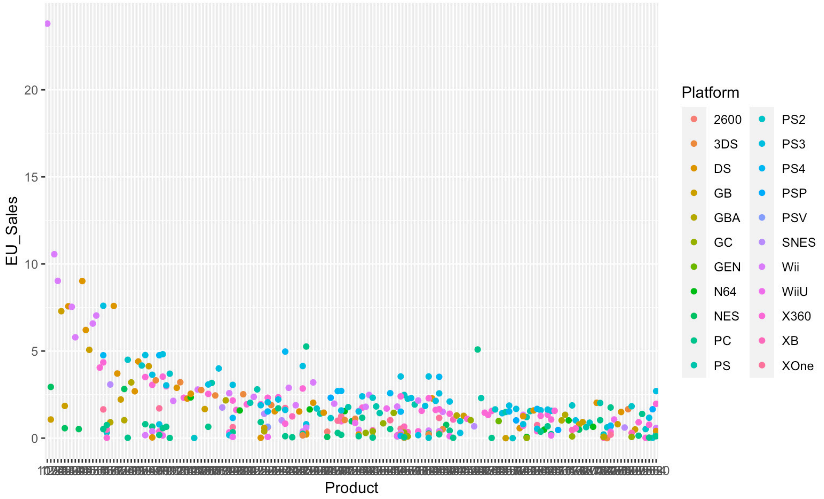
5. Determine sentiment using the Textblob method for both review and summary variables:

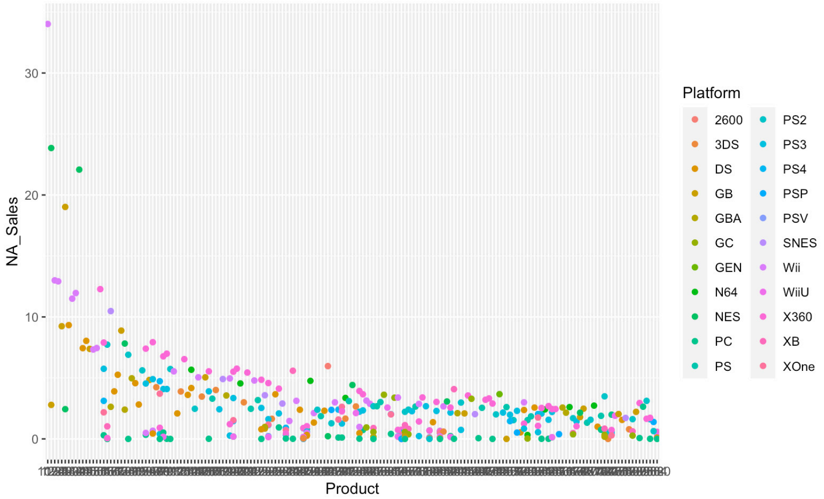
- Establish polarity and subjectivity scores
- Plot histograms of polarity and subjectivity

6. Identify and print the top 20 positive and negative reviews and summaries respectively

Appendix 6: Exploratory sales scatter plots







Appendix 8: Shapiro-Wilk result interpretation

Global Region

- P-value: 2.2e-16
- Less than 0.05, therefore normal distribution is not assumed
- Skewness is 3.066769
- Data is highly skewed to the right
- Kurtosis: 17.79072
- As this value is much greater than 3, data is leptokurtic and very heavy tailed
- Standard Deviation: 8.129224
- Highest SD is observed within the Global sales region

Europe Region

- P-value: 2.987e-16
- Less than 0.05, therefore normal distribution is not assumed
- Skewness is 2.886029
- Data is highly skewed to the right
- Kurtosis: 16.22554
- As this value is much greater than 3, data is leptokurtic and very heavy tailed
- Standard Deviation: 3.083948
- Lowest SD is observed within the Europe region

North America Region

- P-value: 2.92e-16
- Less than 0.05, therefore normal distribution is not assumed
- Skewness is 3.048198
- Data is highly skewed to the right
- Kurtosis: 15.6026
- As this value is much greater than 3, data is leptokurtic and very heavy tailed
- Standard Deviation: 4.556351

Appendix 9:

Sales region correlations

Europe and North America sales:

- R = 0.6209317
- Pearsons correlation coefficient is greater than zero
- Therefore positive association between variables

Europe and Global sales:

- R = 0.8486148
- Pearsons correlation coefficient is near to 1
- Therefore strong positive association between variables

North America and Global sales:

- R = 0.9162292
- Pearsons correlation coefficient is almost 1
- Therefore very strong positive association between variables

Strength of correlation between sales variables can be ranked as:

# 1. North America and Global - strongest correlation
# 2. Europe and Global - second strongest correlation
# 3. Europe and North America - third, moderate correlation

Appendix 10:

Linear regression results

Linear regression: Europe sales by product

European Sales: R-squared explains 20.47% of variability
P-value is 3.25e-10: less than 0.05 so Europe sales are significant
Pr(>|t|) is 3.25e-10: less than 0.05 so model fit is good

Linear regression: North America sales by product

NAmerica Sales: R-squared explains 29.54% of variability
P-value is 7.688e-15: less than 0.05 so Europe sales are significant
Pr(>|t|) is 2e-26: less than 0.05 so model fit is good

Linear regression: Global sales by product

# Global Sales: R-squared explains 36.74% of variability
# P-value is 2.2e-16: less than 0.05 so Europe sales are significant
Pr(>|t|) is 2e-16: less than 0.05 so model fit is good

Log transformations

Europe changes after log transformation

R-squared increases from 20.47% to 23.81%
P-value lowers from 3.25e-10 to 1.585e-10
Pr(>|t|) lowers from 3.25e-10 to 1.58e-10

NAmerica changes after log transformation

# R-squared increases from 29.54% to 39.1%
# P-value lowers from 7.688e-15 to 2.2e-16
Pr(>|t|) lowers from 2e-26 to 2e-16

Global changes after log transformation

R-squared increases from 36.74% to 54.85%
P-value remains unchanged at 2.2e-16
Pr(>|t|) remains unchanged at 2e-16

Appendix 11:

Product and region correlation results

	Product	Europe	NAmerica	Global
Product	1.0000000	-0.4524737	-0.5435505	-0.6061376
Europe	-0.4524737	1.0000000	0.6209317	0.8486148
NAmerica	-0.5435505	0.6209317	1.0000000	0.9162292
Global -	0.6061376	0.8486148	0.9162292	1.0000000

Multiple Linear regression results

Model A Global Sales
Adj R-squared: 97.09% of global sales explained by all variables
P-value is 2.2e-16: all variables are significant
Pr(>|t|) is 8.24e-130: less than 0.05 so model fit is good

Model B Europe
Adj R-squared: 77.97% of global sales explained by Product and Europe sales
P-value is 2.2e-16: variables are significant
Pr(>|t|) is 2e-16: less than 0.05 so model fit is good

Model C North America
Adj R-squared: 85.44% of global sales explained by Product and NAmeric sales
P-value is 2.2e-16: variables are significant
Pr(>|t|) is 2e-16: less than 0.05 so model fit is good

Appendix 12:

Prediction results

Prediction A

Actual global value is 67.80
Predicted values:
fit: 66.3587
lwr: 64.71258
upr: 68.00482

Prediction B

Actual global value is 7.4
Predicted values:
fit: 7.514566
lwr: 7.212602
upr: 7.81653

Prediction C

# Actual global value is 4.32
# Predicted values:
# fit: 4.245235
# lwr: 3.873852
# upr: 4.616618

Prediction D

Actual global value is 6.12
# Predicted values:
# fit: 7.428072
# lwr: 7.140741
# upr: 7.715403

Prediction E
# Actual global value is 23.2
# Predicted values:
# fit: 26.54879
# lwr: 25.37628
# upr: 27.7213