Introduction

Turtle Games is a game manufacturer and retailer with a global customer base and wants to improve overall sales performance by utilising customer trends. They have an initial set of questions with this in mind:

- How do customers accumulate loyalty points?
- How can groups within the customer base be used to target specific market segments?
- How can social data be used to inform marketing campaigns?
- What impact does each product have on sales?
- How reliable is the data?
- What are the relationships are between North American, European, and global sales?

Analytical approach

First, we viewed the metadata and csv files to understand the kinds of data we were provided with. We then used Python to explore the data further, and investigate the possible relationships between the loyalty points, age, remuneration, and spending scores. The data was cleaned and streamlined at this phase, by checking for missing values, dropping unnecessary columns etc.

For the analysis, we used linear regression (OLS) and the statsmodels functions to evaluate possible linear relationships between loyalty points and age/remuneration/spending scores to determine whether these can be used to predict the loyalty points. The models were plotted onto a graph, as the visualisation helps us to observe any relationships/trends between the variables.

In Python, we also looked for potential distinct groups within the customer base that can be targeted as market segments. We used *k*-means clustering to identify the optimal number of clusters and then apply and plot the data using the created segments.

We also did some Natural Language Processing (NLP) on customer reviews downloaded from Turtle Games website, to understand customer sentiment on its products. We reviewed the sentiment polarity, plotting a histogram of polarity and sentiment scores to visualize overall customer sentiment. We identified the most common words used, the top 20 positive and negative reviews received from the website to glean further insights.

We then switched to R for analysis of the sales dataset. After installing and importing all the necessary libraries, we prepared and explore the data for analysis. We used qplot and ggplot to visualize the data and identify any initial trends or insights.

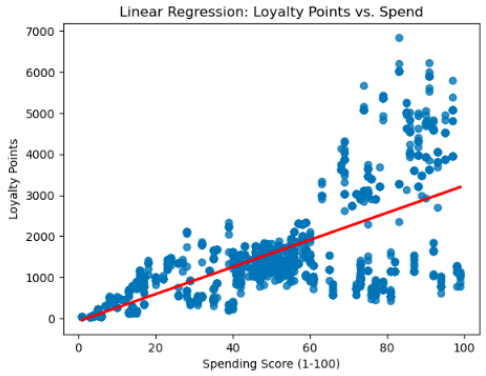
We explored, prepared and explained the normality of the dataset based on plots, Skewness, Kurtosis, and a Shapiro-Wilk test. We also used Pearson's correlation to check for any positive correlations between the sales data. We then created plots for the datasets, with trend lines where applicable.

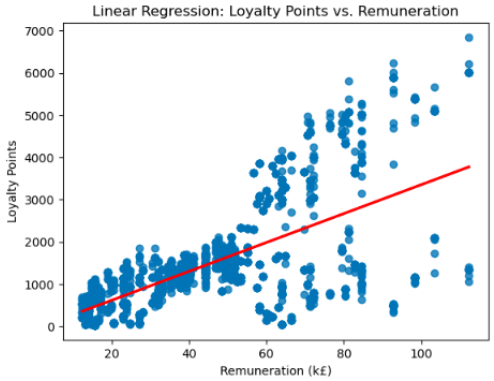
We then investigated possible relationships in the sales data by creating a simple and multiple linear regression model, and then plotted these. We then used the multiple linear regression model to predict global sales based on provided values, comparing the predicted values against observed values in the dataset to check the accuracy and goodness of fit of the model.

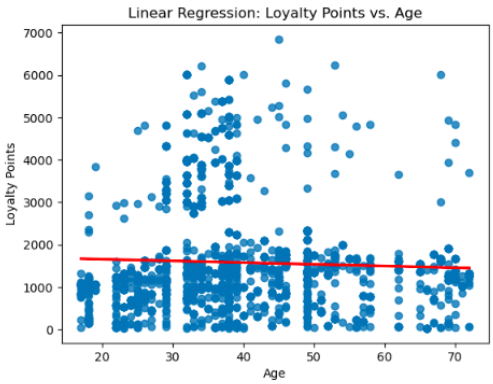
Visualisation and insights

Activity 1

At the start of the analysis, we used linear regression (OLS) and the statsmodels functions to evaluate possible linear relationships between loyalty points and age/remuneration/spending scores. This visualization was chosen as we can clearly see any trends in the data.



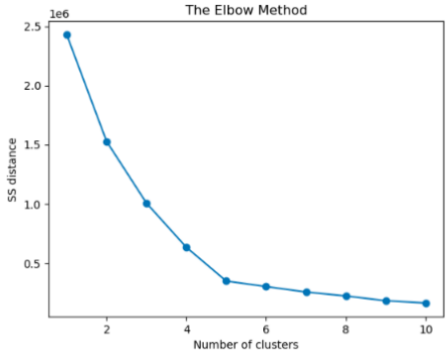


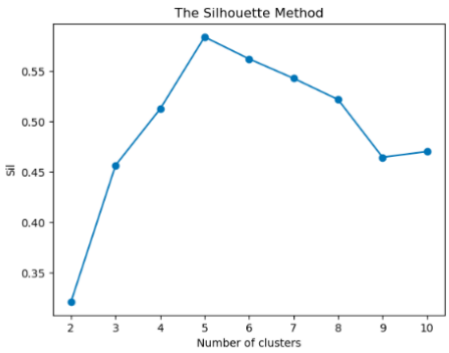


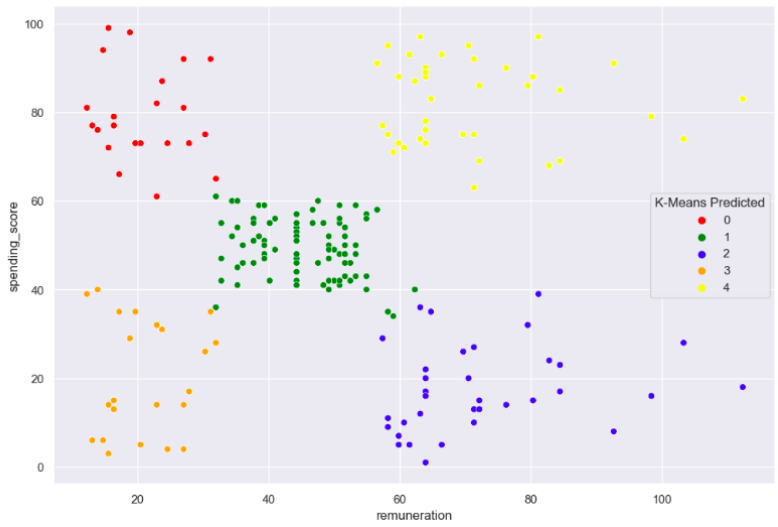
From the plotted models above, it suggests that there is a correlation between loyalty points and spend, remuneration and age. After viewing the R-squared values and the graphs, we can see that spend and remuneration have stronger positive correlation relationship with loyalty, whereas age is not as significant. So, we should focus on these two factors in the next part of the analysis.

Activity 2

Next, we tried to identify potential distinct groups within the customer base that can be targeted as market segments by Turtle Games Marketing team. The Elbow and Silhouette models indicate that k=5` (five clusters) would give the best results (groups) and this seems to be confirmed when plotted onto a scatterplot, as we can observe 5 distinct cluster with varying characteristics.







The market segments are based on remuneration and spending scores, and we can observe that it is not a strictly linear relationship i.e. higher income doesn't equate to higher spending. What we observe is a mix of income/spend behaviours, e.g. high income and lower spend, low income and high spend, etc. This suggests that Turtle Games has varied customer segments with different motivators, and if they are looking to improve sales understanding these motivators will be useful.

The non-linear relationship suggests that there are other factors we haven't considered that are influencing customer behaviour. We would recommend explore what these factors could be to helps us better understand customer behaviours and motivators, which can then inform how Turtle Games can better market their products to them.

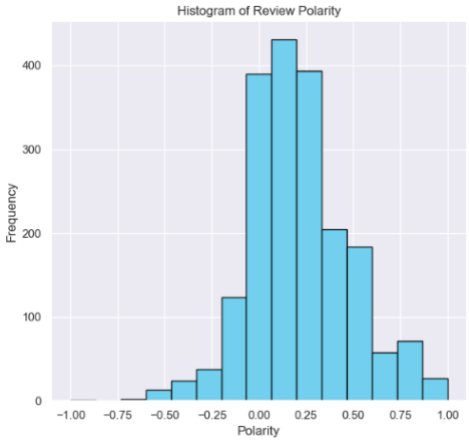
Activity 3

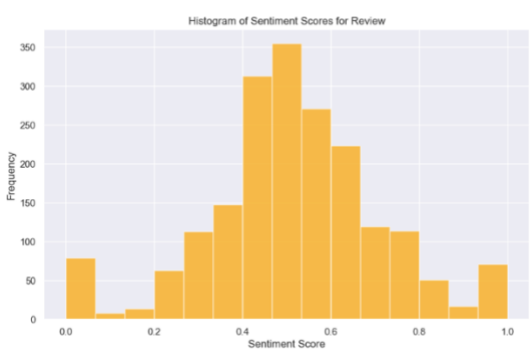
To analyse customer sentiment, we used wordclouds to clearly visualize the cleaned and tokenized data - this highlights the more frequently used words and makes a big impact on audiences.

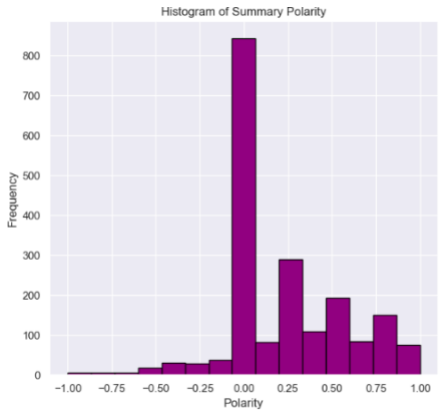
We also extracted the top 15 words used in the review and summary data, which allows us to see the frequency and identify any outliers.

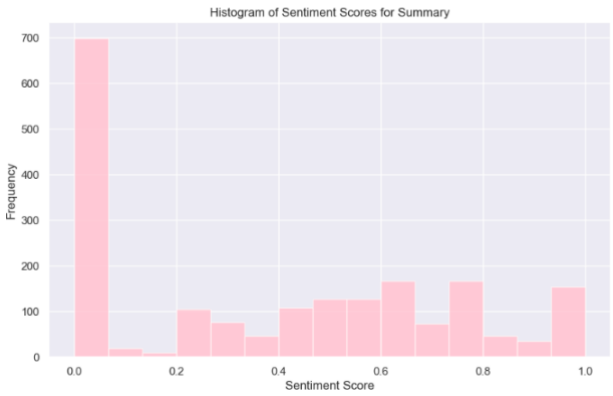
```
Top 15 words in 'summary' column:
stars: 427
five: 342
game: 319
great: 295
fun: 218
love: 93
good: 92
four: 58
like: 54
: 53
expansion: 52
kids: 50
cute: 45
book: 43
one: 38
```

```
Top 15 words in 'review' column:
game: 1671
: 1044
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
```









Based on the histograms displaying the polarity and sentiment scores, it suggests that customers are generally neutral on the product, though the analysis of reviews skew slightly more positive.

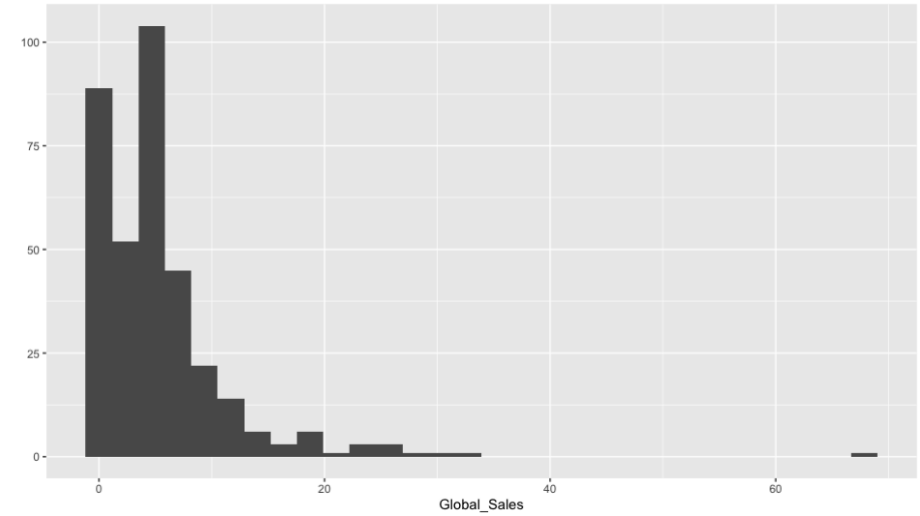
However, the process of preparing the data for NLP could have resulted in some reviews losing the context of their reviews which gives an inaccurate polarity and sentiment score; e.g. one of the Top 20 negative reviews returned seems to actually be a positive review, however due to the presence of certain words meant that it was scored as negative (keeps clients engaged helping develop anger management skills criticism wish cards questions').

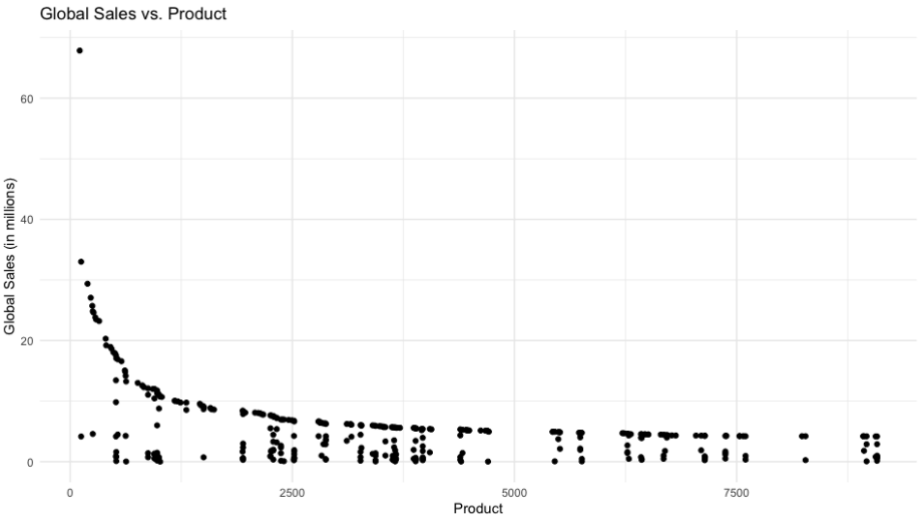
	product	review	review_polarity	review_subjectivity
208	1459	booo unles patient know measure didnt patience neither daughter boring unless craft person	-1.000000	1.000000
526	2849	keeps clients engaged helping develop anger management skills criticism wish cards questions	-0.700000	0.200000
174	5758	sent product granddaughter pompom maker comes two parts supposed snap together create pompoms however parts making unusable cant make pompoms kit useless since sent gift return disappointed	-0.625000	0.475000
182	6504	incomplete kit disappointing	-0.600000	0.700000
538	10281	purchased recommendation two therapists working adopted children children found boring put half way	-0.583333	0.583333
1804	2253	im sorry find product boring frank juvenile	-0.583333	0.750000
364	11056	one staff using game soon dont know well works yet looking cards believe helpful getting conversation started regarding anger control	-0.550000	0.300000
117	2387	bought christmas gift grandson sticker book go wrong gift	-0.500000	0.900000
173	5740	horrible nothing say would give zero stars possible	-0.500000	1.000000
227	231	gift daughter found difficult use	-0.500000	1.000000
230	2173	found directions difficult	-0.500000	1.000000
290	6694	instructions complicated follow	-0.500000	1.000000
301	3525	difficult	-0.500000	1.000000
1524	7533	expensive get	-0.500000	0.700000
1424	2130	one word caution use either expansion mix together items expansion base game thoroughly remove end also symbols differentiating expansion items terribly visible miss first time	-0.487500	0.683333
1058	9560	like wizards coasts game bad think collectible game recommend dd adventure board game mania	-0.480000	0.533333
347	7384	yearold granddaughter frustrated discouraged attempting craft definitely young child difficulty understanding directions disappointed	-0.450000	0.450000
601	977	book bound upside distracting children keep saying readinf upside expensive poor quality	-0.450000	0.650000
476	1577	confusing instructions year olds boring asking question question worded differently	-0.433333	0.666667
74	9597	although isnt much disappointing see small booklet pages stickers read details closely	-0.425000	0.550000

For further analysis, customer sentiment could be mapped in relation to certain products. We could identify the products that require improvement, or conversely the best-reviewed, by the quantity and severity of comments.

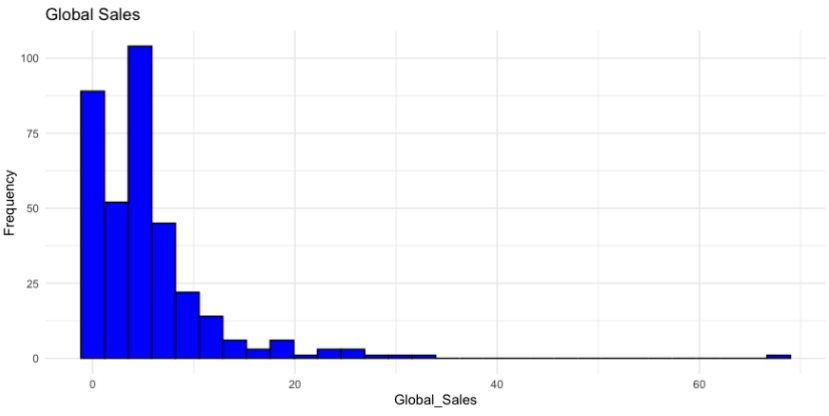
Activity 4

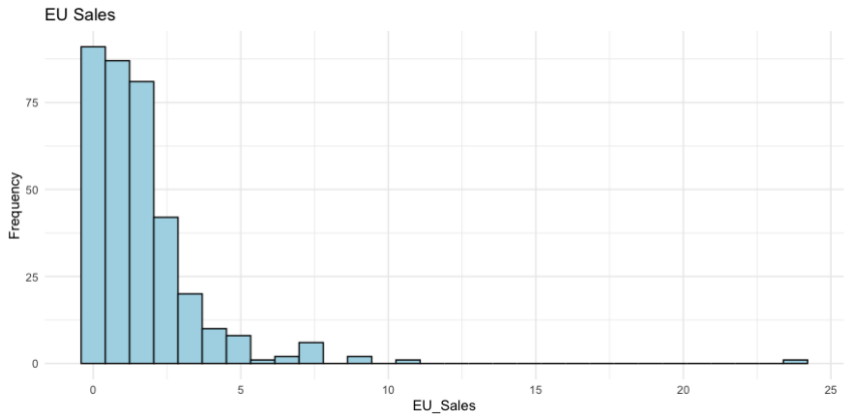
Now we looked to analyse Turtle Games sales dataset, using qplot and ggplot to visualize the data and identify any initial trends or insights.

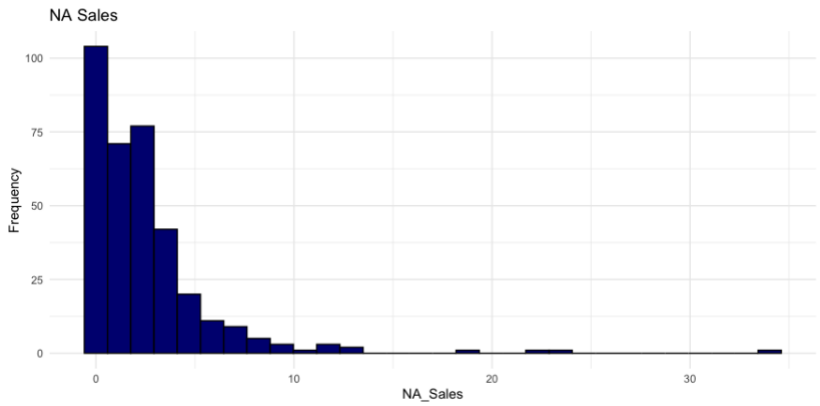




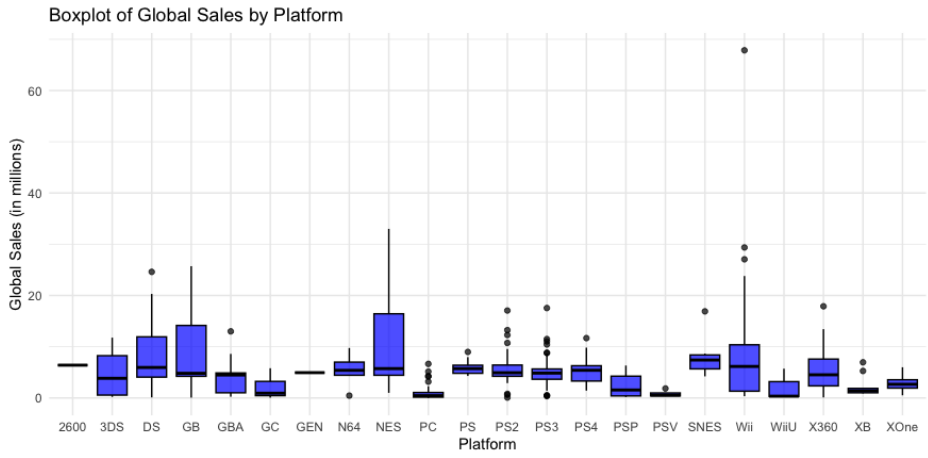
Based on initial analysis, we can see that there are certain products that sell much better than others. We also observe a positive correlation between global, North America (NA) and Europe (EU) sales as they have the same general shape.

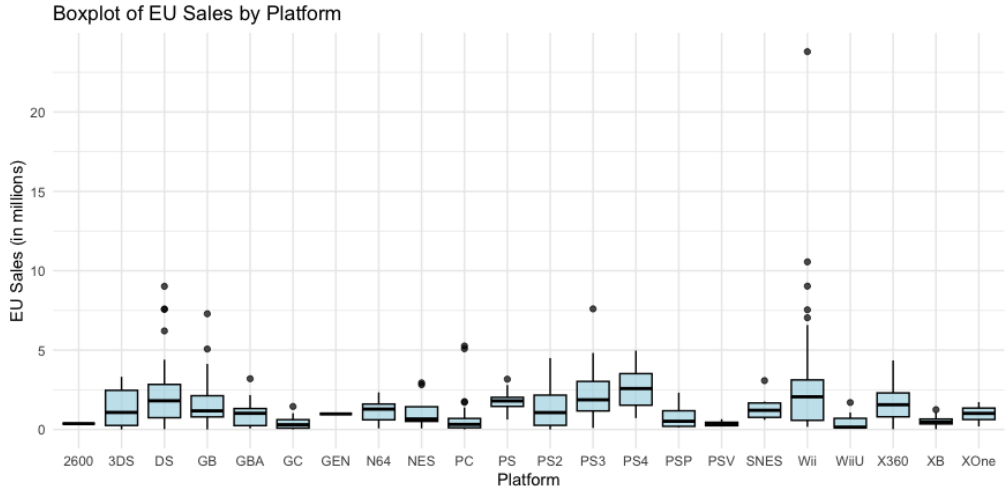


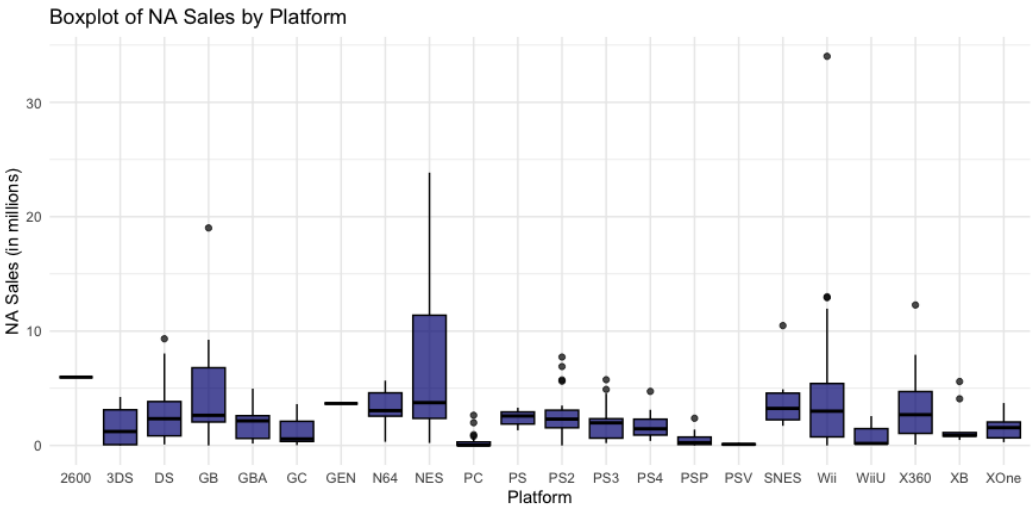




We also observed that products launched certain platforms perform significantly better than others. However, there is variation based on region.



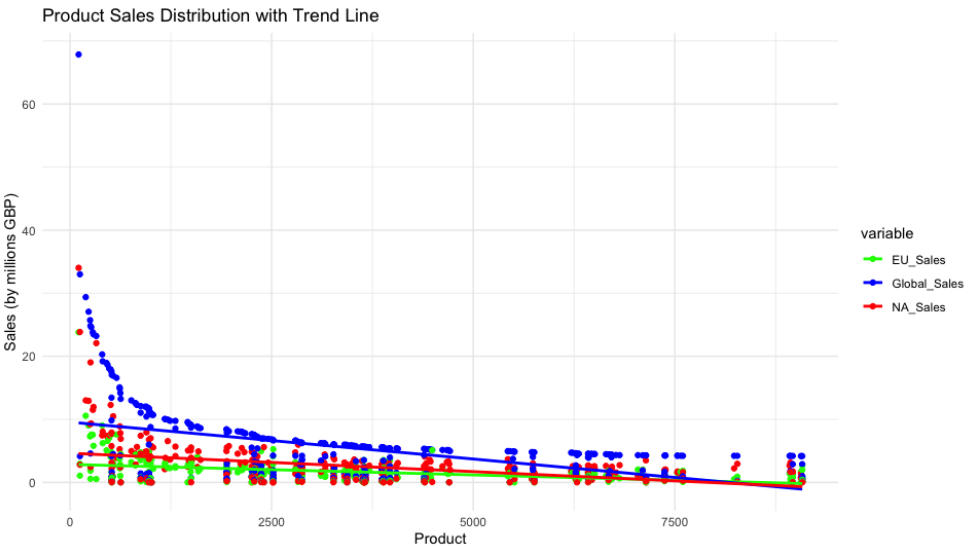




Activity 5

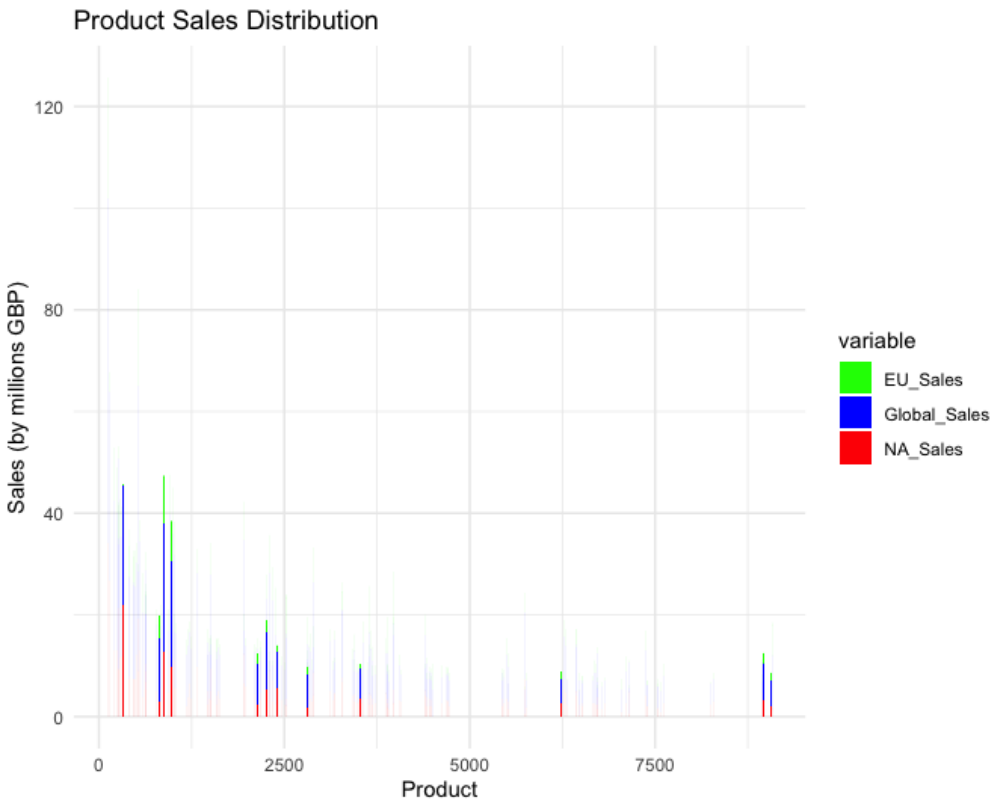
After checking the skewness and kurtosis of the datasets, we used a mix of scatterplots, histograms and boxplots to review and determine insights into the data set.

We reconfirmed that there is a positive correlation between global sales, EU sales and NA sales, and from the trendline in the Product Sales Distribution scatterplot that there is a general downward trend of sales to higher product IDs.



We would want to explore further what is the trait for products with higher ID numbers, e.g. are they newer products, more expensive etc. to identify what could be the main factor influencing this trend.

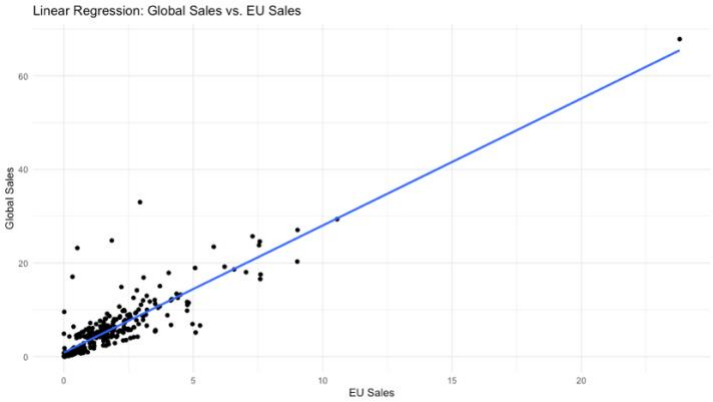
We can also see from the scatterplot and stacked barplot that certain products are more popular in certain regions. We would want to make note of that to optimise sales and marketing of the products.

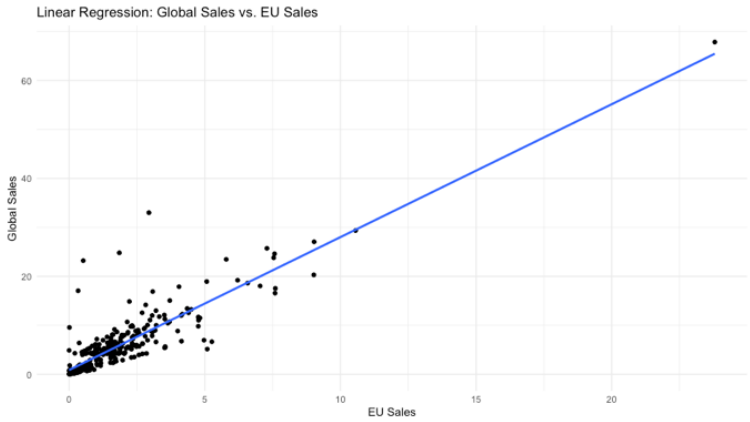


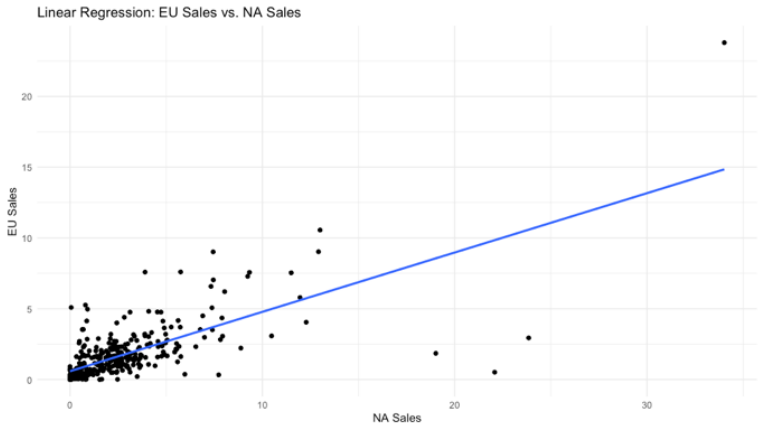
Activity 6

By creating a simple and multiple linear regression model, we can investigate any possible relationship in the sales data. The multiple linear regression model indicates a positive correlation between global sales, EU and NA Sales.

```
> # Print the correlation matrix
> print(cor_matrix)
                Product   NA_Sales   EU_Sales Global_Sales
Product       1.0000000 -0.4047865 -0.3894246   -0.4409046
NA_Sales     -0.4047865  1.0000000  0.7055236    0.9349455
EU_Sales     -0.3894246  0.7055236  1.0000000    0.8775575
Global_Sales -0.4409046  0.9349455  0.8775575    1.0000000
```







Generally, the predicted values suggest a relatively accurate predictive model. However, there is variation in the differences between observed and predicted values, i.e. some predictions have a larger difference. This could be due to other factors not considered in the model (e.g. seasonality)

```
> # Print the filtered results
> print(filtered_results)
    NA_Sales_sum EU_Sales_sum Observed_Global_Sales Predicted_Global_Sales
1          34.02        23.80                67.85              71.468572
10         22.08         0.52                23.21              26.431567
97          4.42         0.97                 6.12               6.630482
99          3.93         1.56                 6.04               6.856083
142         2.90         1.56                 5.01               5.665985
176         2.73         0.65                 4.32               4.248367
196         3.12         0.52                 4.20               4.524530
211         2.26         0.97                 3.53               4.134744
237         0.23         0.65                 1.85               1.359781
266         0.41         0.52                 1.01               1.393302
276         0.30         0.52                 0.89               1.266205
284         0.01         0.65                 0.73               1.105585
286         0.01         0.65                 0.72               1.105585
>
```

We have generally good confidence in the model to predict sales based on the variables used. Our recommendation would be to consider using additional models to predict futures sales based on other factors. As an example, sales based on a time series can be useful to view any seasonal trends and assist in optimising product supply and marketing during any peak periods.