LSE DA301 Assignment Report

Background

Turtle Games is a game manufacturer and retailer operating globally, which sells its own games as well as games made by other companies. With the goal to improve sales Turtle Games has defined a set of questions pertaining to customer segments, their loyalty program, customer reviews and sales data. In the following report I will use these questions to explore each topic, extracting insights, recommendations, and suggesting further analysis and exploration.

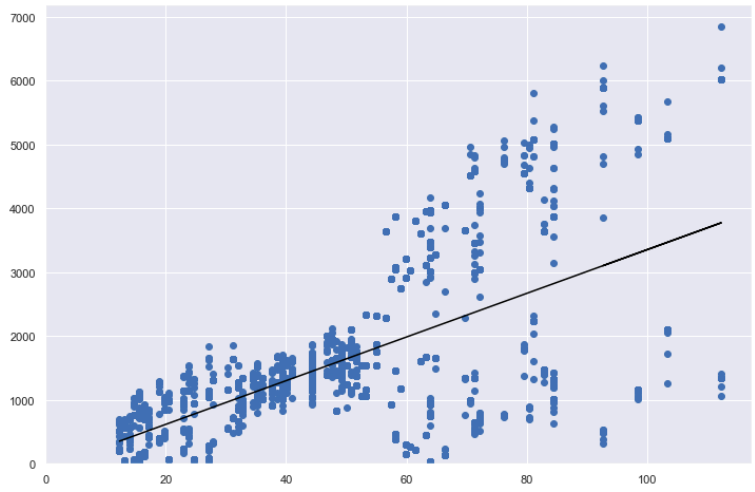
How Customers Accumulate Loyalty Points

I used Python to start exploring the customer reviews data where I found the following:

Average age is 39 and ranges 17 to 72.
Average income is £48K and ranges £12.3K to £112K.
Average spending score is 50 and ranges 1 to 99.
Average loyalty points is 1,578 and ranges 25 to 6,847.

To look at the relationship between Loyalty Points and Remuneration, Spending Score and Age I plotted the data and using the OLS method to add a regression line to determine the relationship between the variables.

Loyalty Points & Remuneration

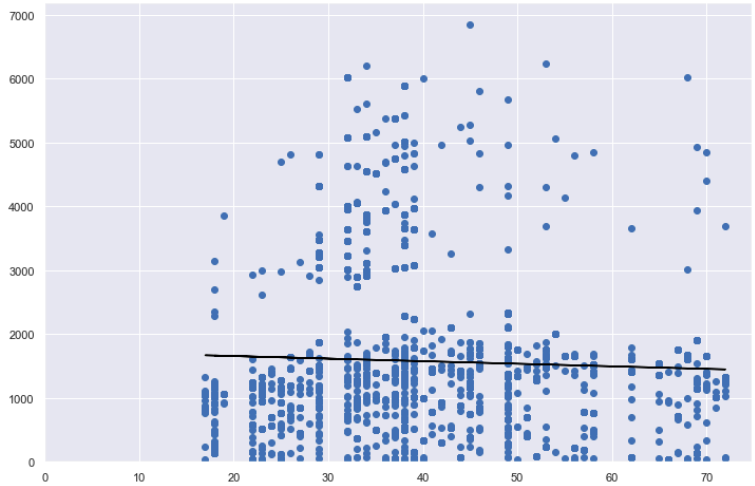


Although there is an overall positive relationship between Loyalty Points and Renumeration, the OLS model shows that Renumeration is only able to explain 38% of the variation of Loyalty Points.

Loyalty Points & Spending Score

Although there is an overall positive relationship between Loyalty Points and Spending Score, the OLS model shows that Spending Score is only able to explain 45% of the variation of Loyalty Points.

Loyalty Points & Age

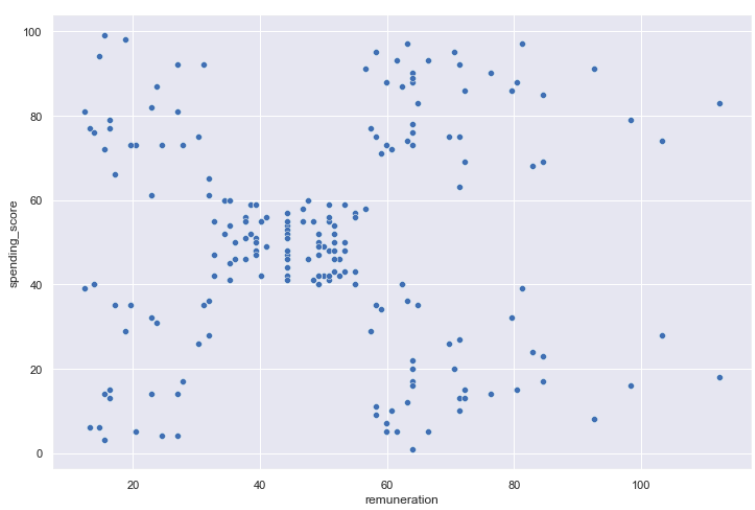


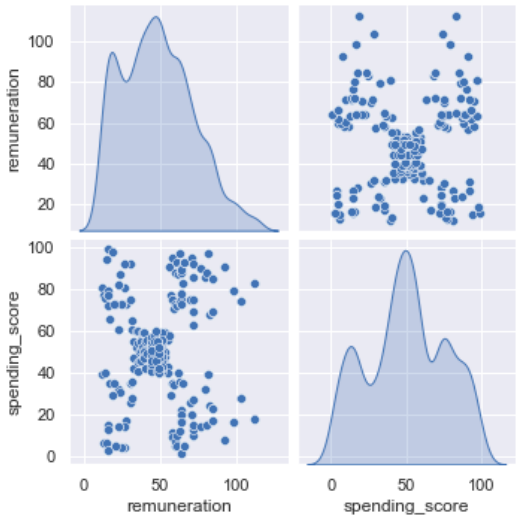
There is no significant relationship between Loyalty Points and Age, the OLS model shows that Age is only able to explain 0.02% of the variation of Loyalty Points which is basically nothing.

Out of these three variables the strongest correlation to Loyalty Points is the Spending Score, which would make sense as usually loyalty points are collected by spending money on purchases.

How Groups Within The Customer Base Can Be Used To Target Specific Market Segments

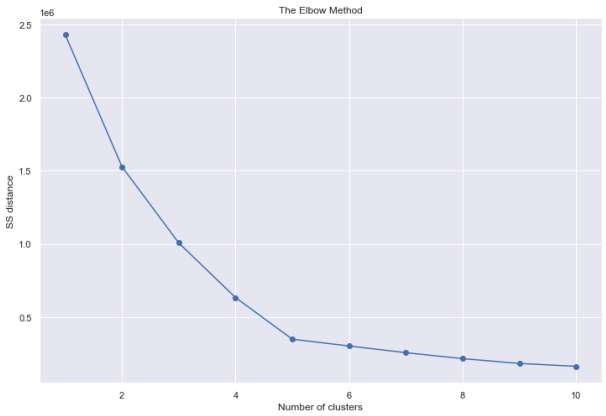
To look at different groups within the customer base I used k-means clustering in Python. I used Remuneration and Spending Score as variables to look at groups based on income and spending levels. To do this I created a data frame containing only the variables of interest and started out by plotting them to see if I could identify any groups visually on a scatter plot and on a pair plot.

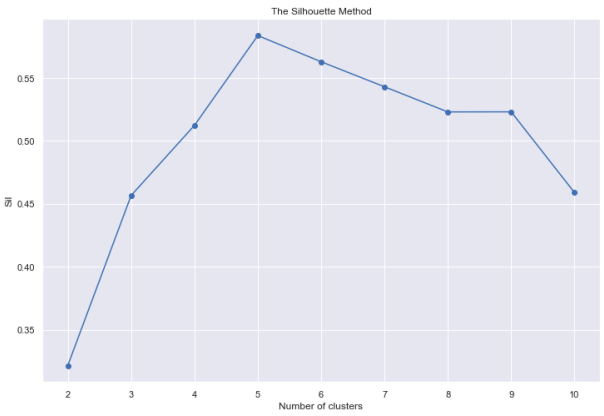




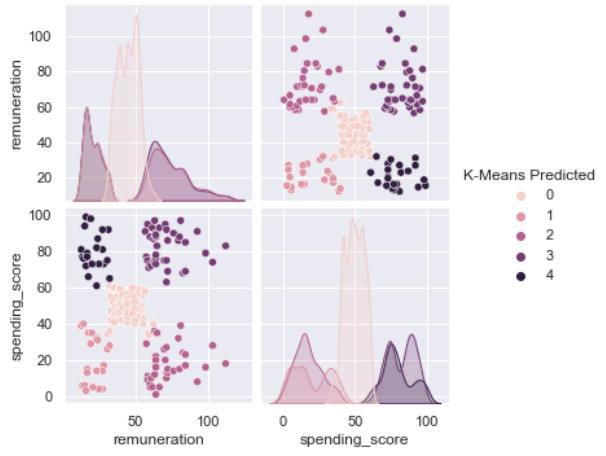
The scatterplots seem to show 5 clear clusters.

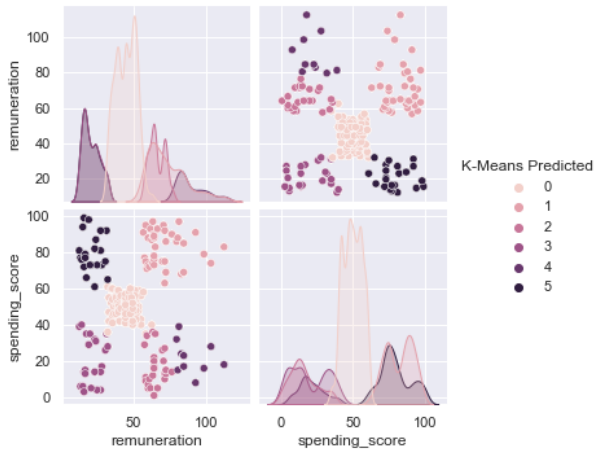
To check how many clusters to use I used the elbow and silhouette methods.

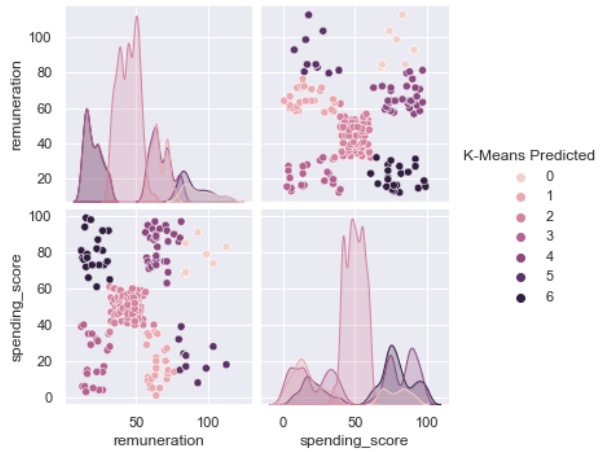




The Silhouette method suggests using 5 clusters, but since the elbow method results are relatively subjective, I tried k-means clustering using 5, 6 and 7 clusters.







Looking at the value counts of each cluster, they are most balanced when using five clusters, although there is one big cluster which is much larger than the rest. When increasing the number of clusters we end up dividing the smaller clusters further into smaller clusters and do not achieve a balance in the value counts. It seems five is the optimal number for K.

I then plotted the 5 clusters.

The plot shows there are five clear clusters of customers:
- High income, low spenders in Blue
- Low income, low spenders in Green
- High Income, high spenders in Purple
- Low income, high spenders in Orange
- Mid income, mid spenders in Red

The biggest cluster of customers is the Red cluster of mid income mid spenders.

As further analysis it would be interesting to investigate which clusters are most profitable, their preferences and shopping habits in order to target them with the most relevant products and offers.

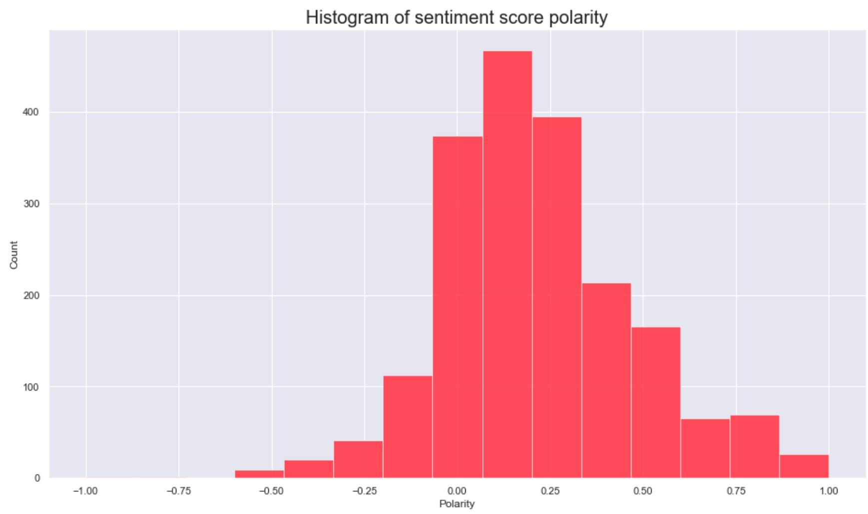
How Customer Reviews Can Be Used To Inform Marketing Campaigns

In Python I used natural language processing tools to analyse customer reviews and the review summaries. To do this it is necessary to clean the data by transforming all the text to lowercase, removing punctuation, removing duplicates, splitting sentences up into single word lists, removing “stop words” such as ‘a’, ‘the’, ‘and’, ‘or’… etc.

After cleaning the data I used a word cloud to identify the most commonly used words.

I then extracted the top 15 most used words and ran a sentiment program to generate the polarity for each word, indicating positive, negative or neutral sentiment. The only word with a negative sentiment score was ‘game’ which in the context of a game company I would classify as neutral.

I then applied the same sentiment scoring to each review and plotted the results in a histogram, which showed that reviews tend to be more positive than negative.

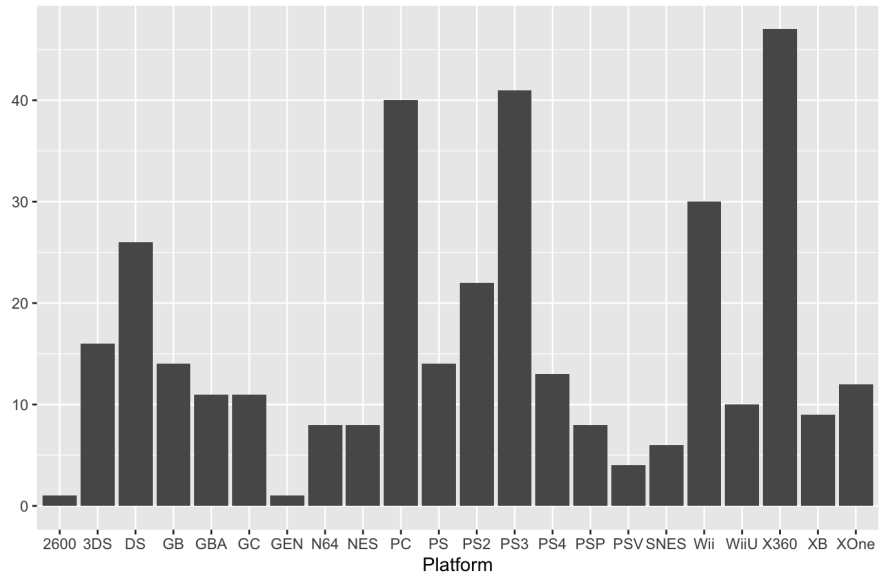


Overall review sentiments are positive. When looking as summary sentiments a large portion is neutral, however when looking at the top 20 positive reviews we see most summaries are 'five stars' but have a sentiment score of 0 (neutral) indicating the sentiment analysis is not able to recognise five stars as a positive comment. Having noted this, it is best to focus on the sentiment scores of reviews rather than summaries.

What Is The Impact That Each Product Has On Sales

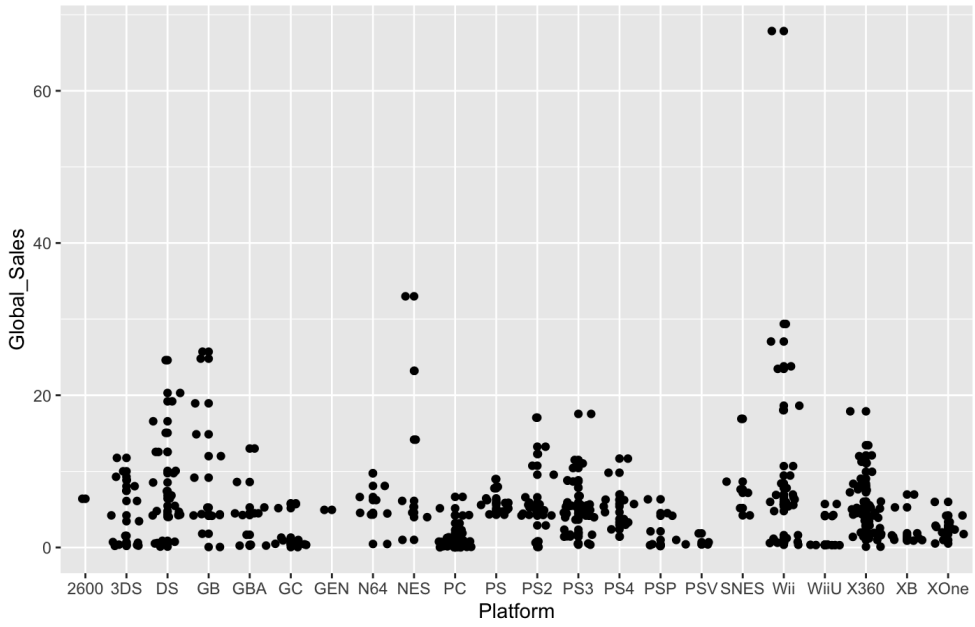
To explore the sales data I used R, starting with loading the sales csv file, and removing columns I was not interested in looking at 'Ranking', 'Year', 'Genre', and 'Publisher'. I then plotted the data using different visualisations.

Number of games by platform



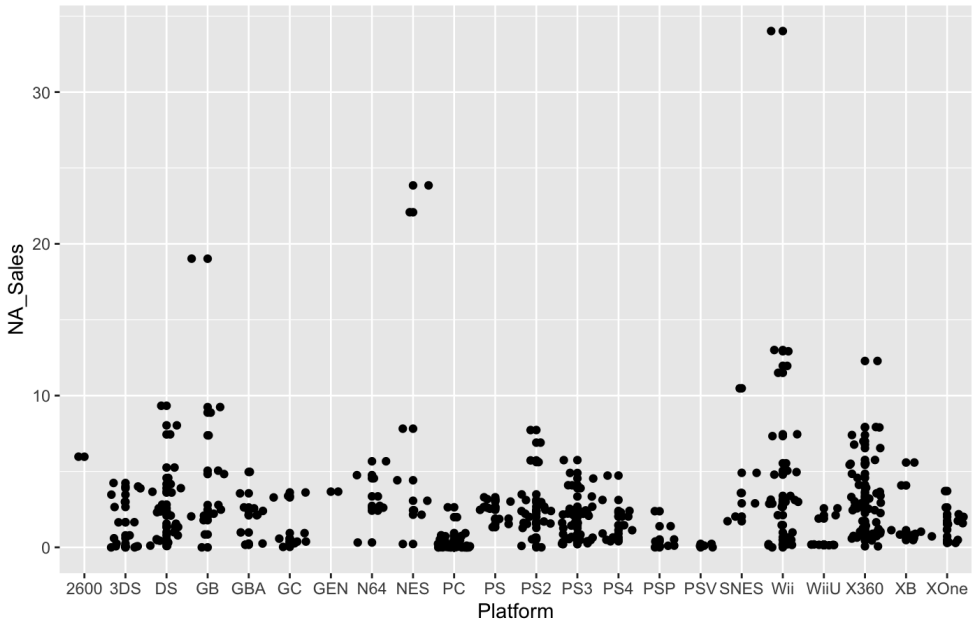
X360 seems to be the most popular console, as it has the greatest number of games, followed by PS3, PC and Wii.

Sales per game by platform

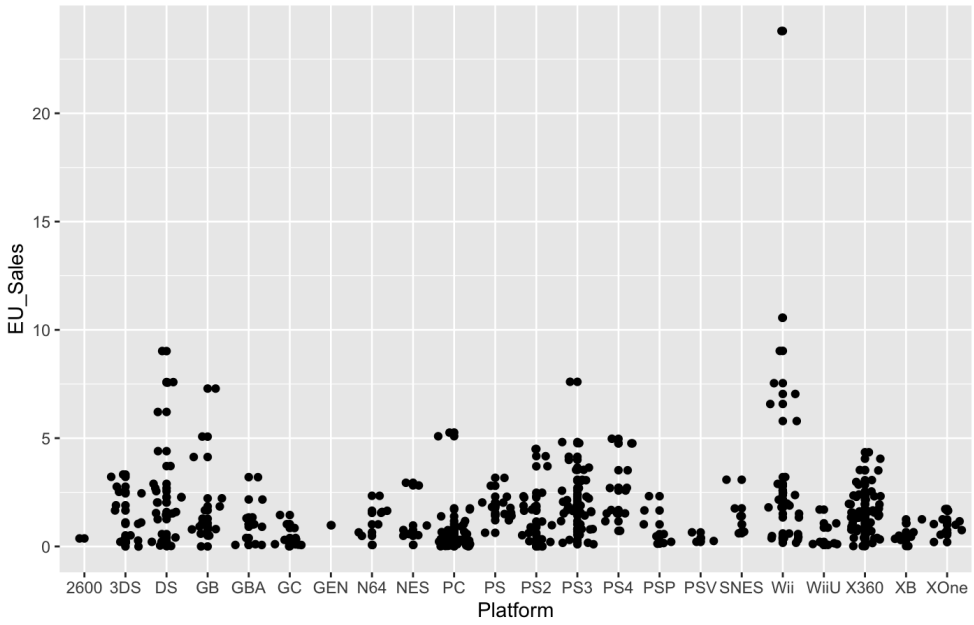


The majority of games seem to fall between 0 and 20 million pounds in sales, while two Wii games generated just under 70 million pounds.

North America Sales



Europe Sales

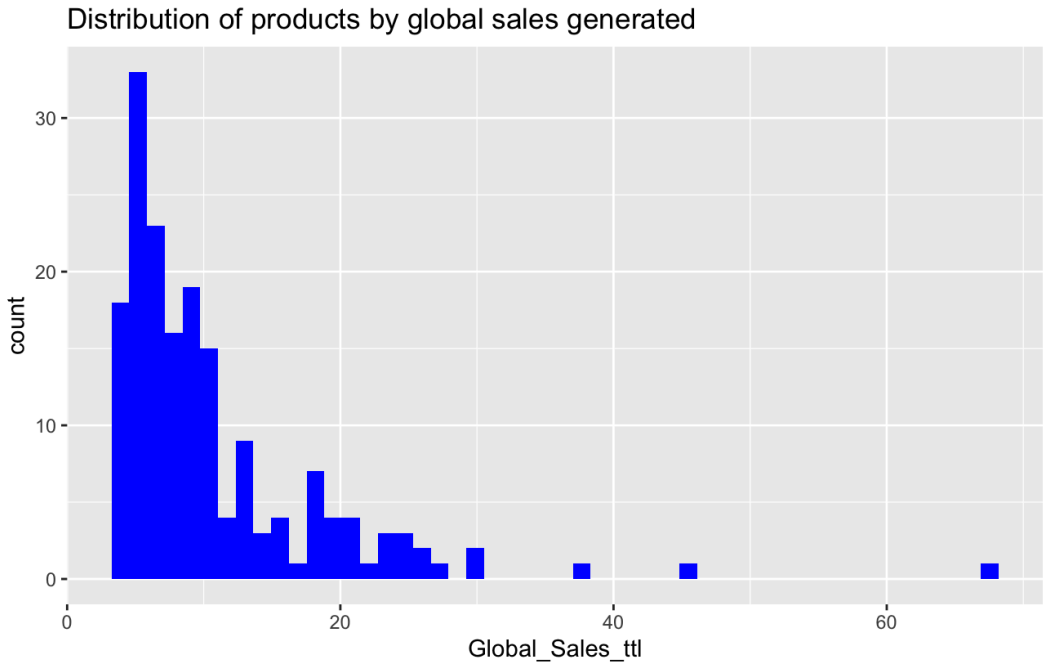


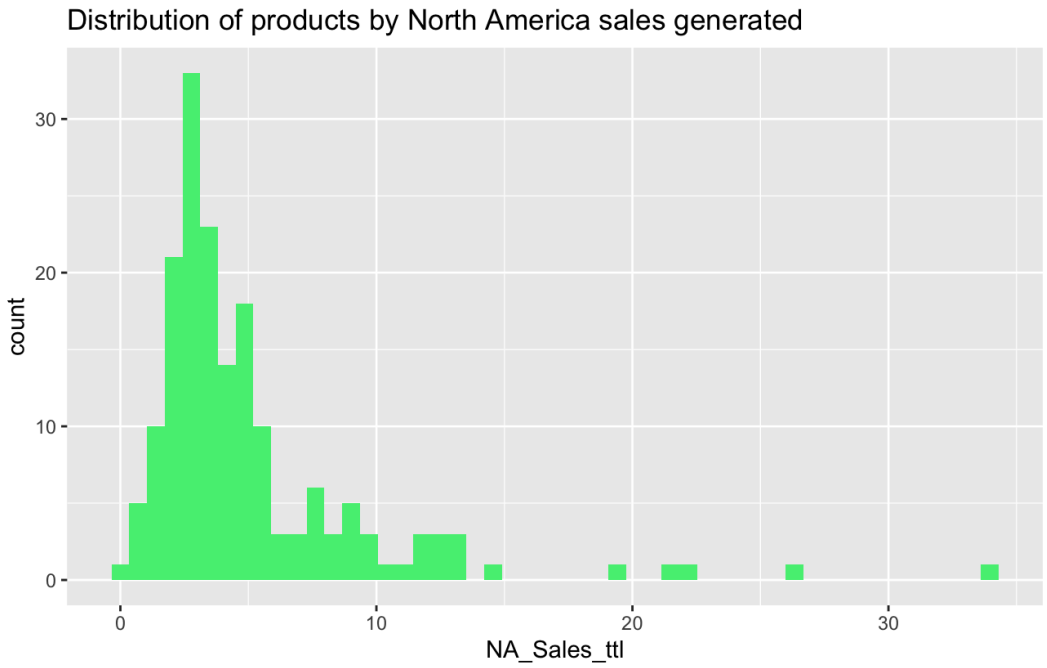
The two highest selling Wii games lead sales in both North America and Europe, however see greater sales in North America generating just under 35M each and 25M each in Europe.

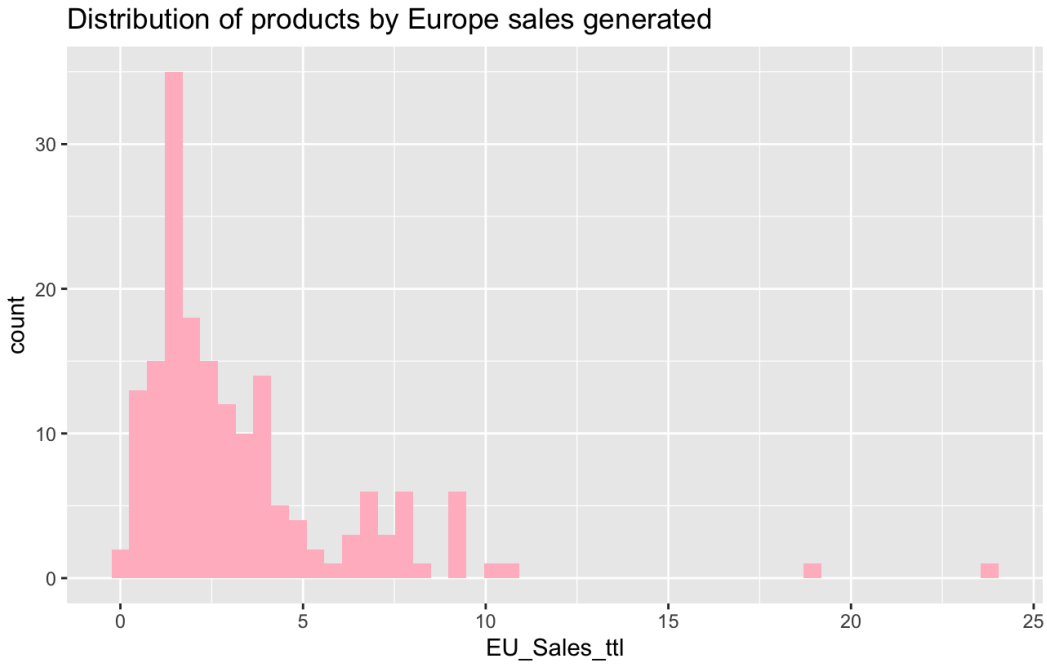
North America has more games that generated over 10M in sales, while Europe seems to only have 4 games that went over 10M in sales.

It would be interesting to see the overall sales mix taken by each region and the sales mix by region for each platform, to see if certain platforms are preferred in certain regions.

I plotted histograms of sales by product per region.

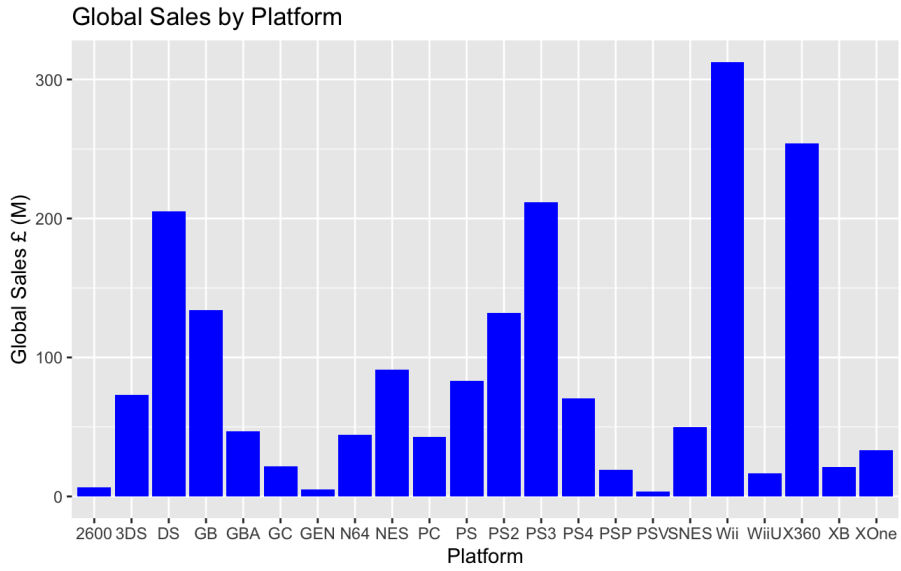

Distribution of products by global sales generated
count
30
20
10
0
0
20
40
60
Global_Sales_ttl

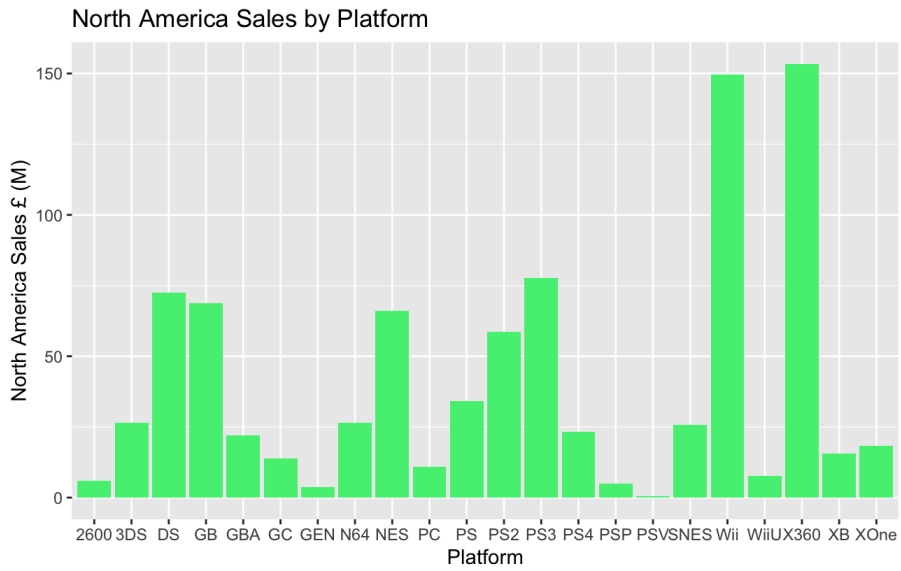

Distribution of products by North America sales generated
count
30
20
10
0
0
10
20
30
NA_Sales_ttl

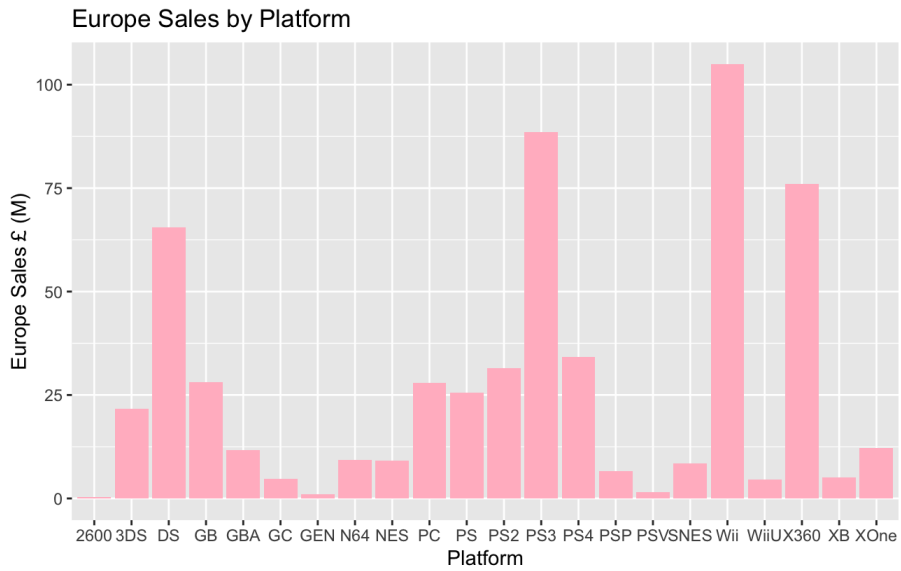

Distribution of products by Europe sales generated
count
30
20
10
0
0
5
10
15
20
25
EU_Sales_ttl

Across regions, but especially in Europe, there is a significant number of products which generate relatively low sales, while fewer products are able to generate over £10M in sales each. It would be interesting to further explore what sales mix% these low performing product represent, look into the cost of manufacturing and/or sourcing these low performing products and determine if it is worth keeping all these products in the assortment, or if it might be more profitable to invest in the top performing products.

I plotted bar charts of sales by platform per region.







Best performing platform in Europe is Wii, followed by PS3, X360 and DS.
Best performing platform in North America is X360, closely followed by Wii and then PS3, and DS.
Globally the top platform is Wii, followed by X360, PS3 and DS. Although the same 4 platforms appear at the top of sales across regions, the trends are slightly different trends, for example X360 is not nearly as popular in Europe as it is in North America. This suggests it is worth customising assortment and marketing by region.

How Reliable Is The Data

To determine if the data is normally distributed I used the Shapiro-Walk test in R which showed all 3 sales data columns are not normally distributed. All three showed positive skewness and kurtosis higher than 3, meaning the data has longer tails than normal distribution. These more extreme values suggest there are certain games which perform extraordinarily well and should be considered in any sales and marketing planning.

What Are The Relationships Between North American, European, And Global Sales

In R I used the Pearson Correlation to determine correlations between the sales data and found that they are all positively correlated, with the strongest correlation being between North America and Global sales suggesting North America follows global trends pretty closely, while Europe trends are more unique.

I used linear regression to identify the relationship between the sales data. Simple linear regression confirmed what we found from the Pearson Correlation, so I used multiple linear regression to use both North America and Europe sales as predictors of global sales. The regression model showed that the best model with the highest adjusted r-squared value the multiple linear regression model using both region sales which is able to explain 87% of global sales.

I tested the multiple linear regression model and found that the model predicted global sales that are fairly close to the actual observed figures, however 4 out of 5 predicted figures were slightly bigger than the the observed sales.

This can be used to test new products by launching them in North America (requiring a smaller investment than a full global launch), and testing out the performance as we can infer the global sales will follow a similar trend.