Predicting future outcomes – Turtle games

Background and context of the Business Scenario

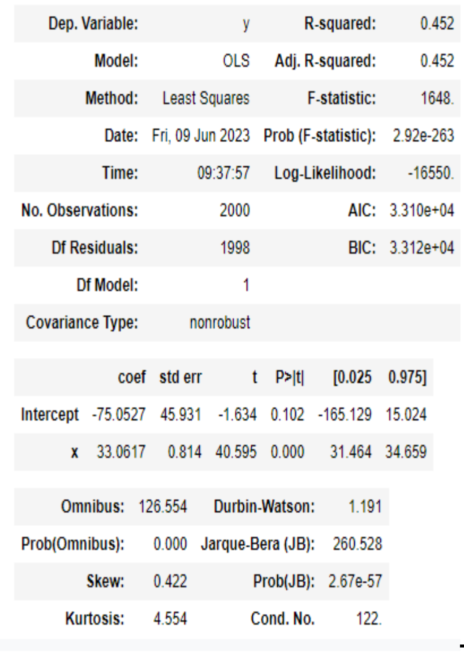
Turtle games is a game manufacturer and retailer operating globally. They produce and sell their own products as well as sourcing and selling products manufactured by other companies.

By collecting data from sales and customer reviews Turtle Games business objective is to improve overall sales performance through utilisation of customer trends. This analysis will aid them in understanding customer trends, insights and determining relationships within the sales data.

1. Make predictions with regression – Loyalty Points.

We applied linear regressions to determine how customers accumulated loyalty points by investigating 3 relationships:
- Loyalty points vs Spending
- Loyalty points vs Remuneration
- Loyalty points vs Age

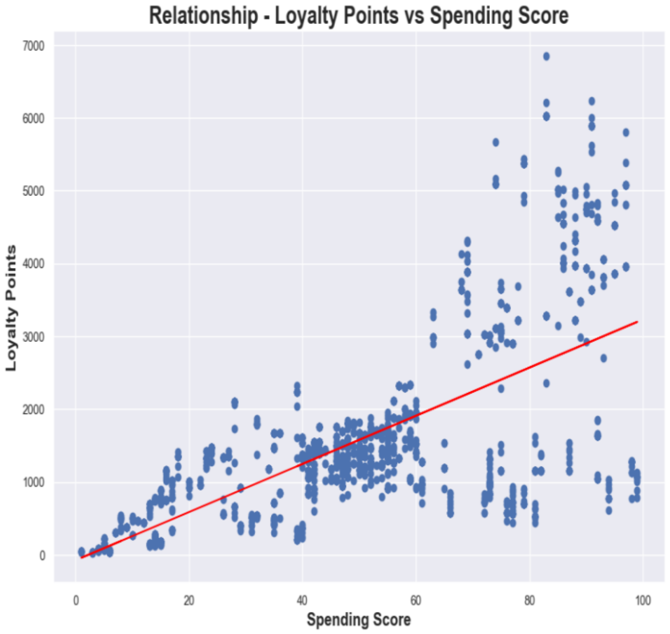
For Loyalty points vs Spending we had an R_2 value of 45.2% suggesting there is some correlation/relationship between the two variables. Even though 54.8% of the variation is unexplained there is an indication that the more a customer spends the likelihood of their loyalty points being higher. Plotting the regression line, we can see a positive linear relationship. Evident in the below figure:



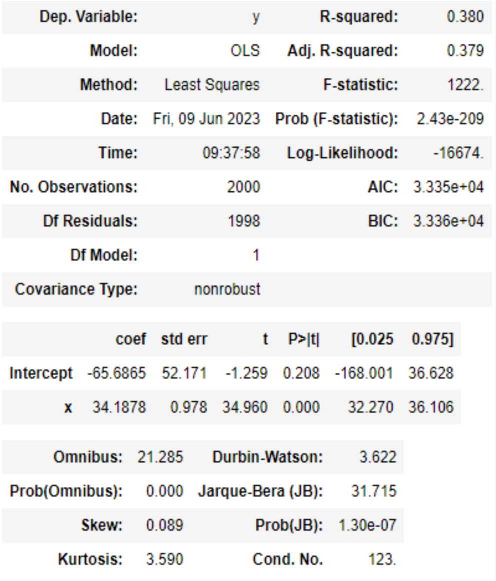
Dep. Variable:	y	R-squared:	0.452
Model:	OLS	Adj. R-squared:	0.452
Method:	Least Squares	F-statistic:	1648.
Date:	Fri, 09 Jun 2023	Prob (F-statistic):	2.92e-263
Time:	09:37:57	Log-Likelihood:	-16550.
No. Observations:	2000	AIC:	3.310e+04
Df Residuals:	1998	BIC:	3.312e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-75.0527	45.931	-1.634	0.102	-165.129	15.024
x	33.0617	0.814	40.595	0.000	31.464	34.659

Omnibus:	126.554	Durbin-Watson:	1.191
Prob(Omnibus):	0.000	Jarque-Bera (JB):	260.528
Skew:	0.422	Prob(JB):	2.67e-57
Kurtosis:	4.554	Cond. No.	122.



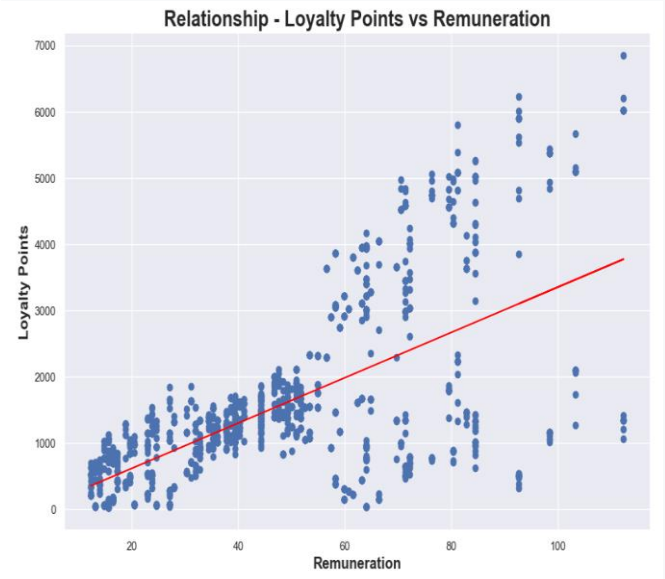
Similarly, this is the case for the correlation/relationship between Loyalty Points vs Remuneration, however less significant with a $R2$ value of 38.0%. Plotting the regression line, we can see a positive linear relationship. Evident in the below figure:



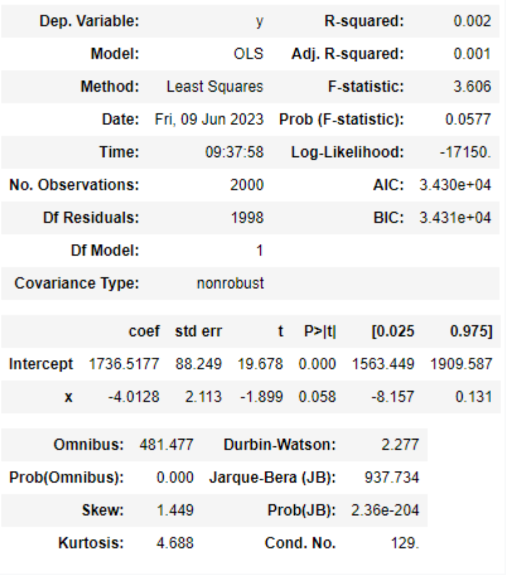
Dep. Variable:	y	R-squared:	0.380
Model:	OLS	Adj. R-squared:	0.379
Method:	Least Squares	F-statistic:	1222.
Date:	Fri, 09 Jun 2023	Prob (F-statistic):	2.43e-209
Time:	09:37:58	Log-Likelihood:	-16674.
No. Observations:	2000	AIC:	3.335e+04
Df Residuals:	1998	BIC:	3.336e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-65.6865	52.171	-1.259	0.208	-168.001	36.628
x	34.1878	0.978	34.960	0.000	32.270	36.106

Omnibus:	21.285	Durbin-Watson:	3.622
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31.715
Skew:	0.089	Prob(JB):	1.30e-07
Kurtosis:	3.590	Cond. No.	123.



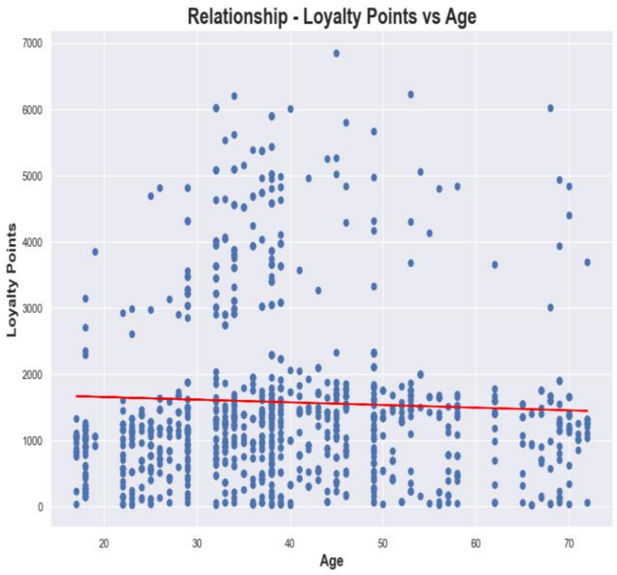
Looking at the correlation/relationship between Loyalty Points vs Age we see there is no relationship between the two variables with a $R2$ value of 0.2%. Plotting the regression line, we can see there is no linear relationship. Evident in the below figure:



Dep. Variable:	y	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	3.606
Date:	Fri, 09 Jun 2023	Prob (F-statistic):	0.0577
Time:	09:37:58	Log-Likelihood:	-17150.
No. Observations:	2000	AIC:	3.430e+04
Df Residuals:	1998	BIC:	3.431e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1736.5177	88.249	19.678	0.000	1563.449	1909.587
x	-4.0128	2.113	-1.899	0.058	-8.157	0.131

Omnibus:	481.477	Durbin-Watson:	2.277
Prob(Omnibus):	0.000	Jarque-Bera (JB):	937.734
Skew:	1.449	Prob(JB):	2.36e-204
Kurtosis:	4.688	Cond. No.	129.

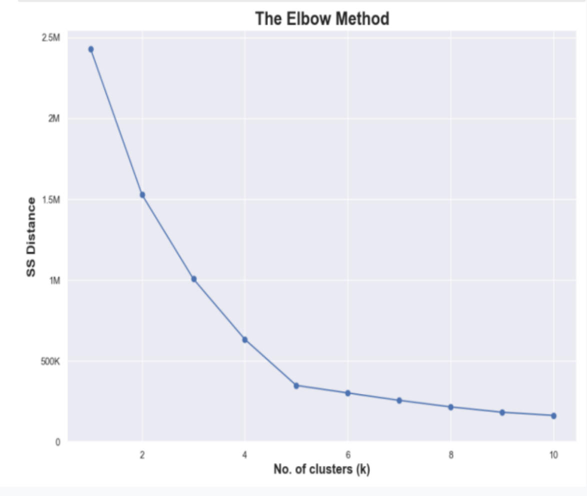


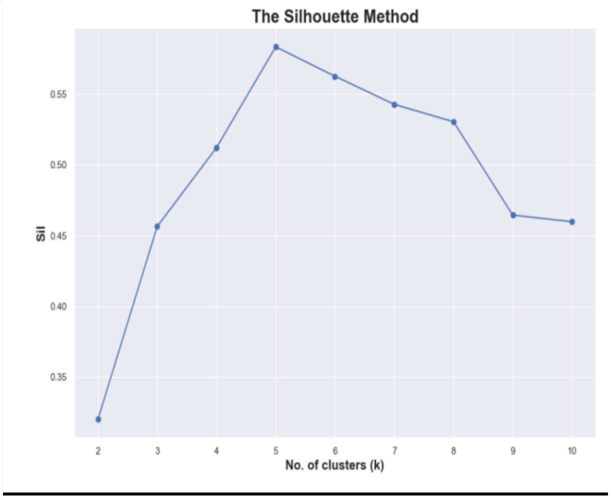
In summary, we could use spending and remuneration against loyalty points, however, the $R2$ values are both below 50% indicating there is some sort of minimal relationship. We can confidently exclude age, as no relationship exists with loyalty points.

2. Make predictions with clusters – k-means.

We applied k-means clustering to determine how useful remuneration and spending scores are in providing data for analysis. With these clusters we would be able to advise which specific customers our marketing department could target.

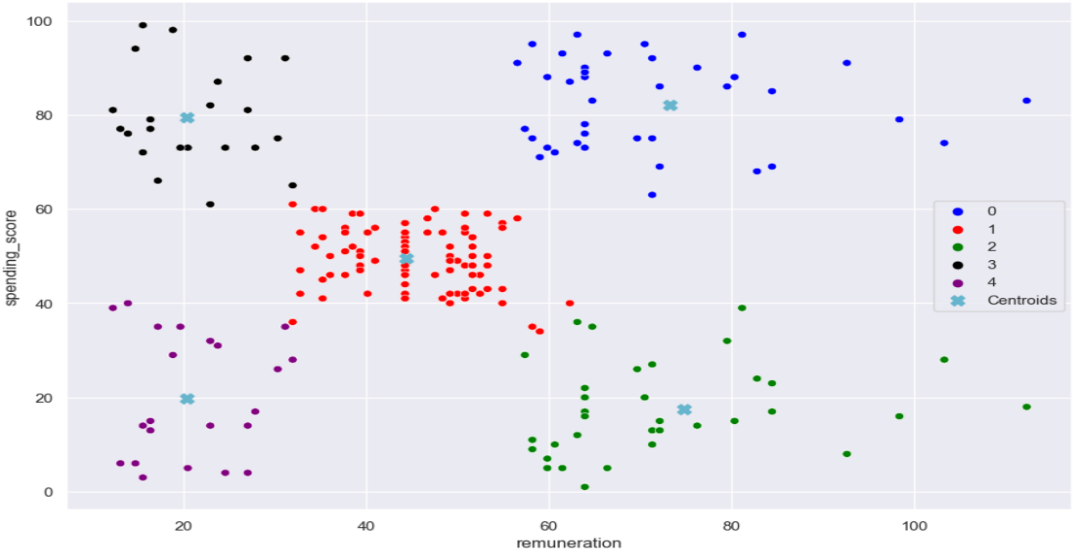
Upon investigating the Elbow and Silhouette method it was determined that the best fit number of clusters was 5. The below shows that the elbow is seen at 5 clusters, this is reiterated when we performed the Silhouette method. As per the below:





Based on the analysis conducted and visualising the clusters, plotting the centroids we can see that (Cluster 1 - red) has the most compact cluster with data points being close to the mean of the cluster, whilst (Cluster 0 - blue) and (Cluster 2 - green) have the most dispersed data points from the mean.

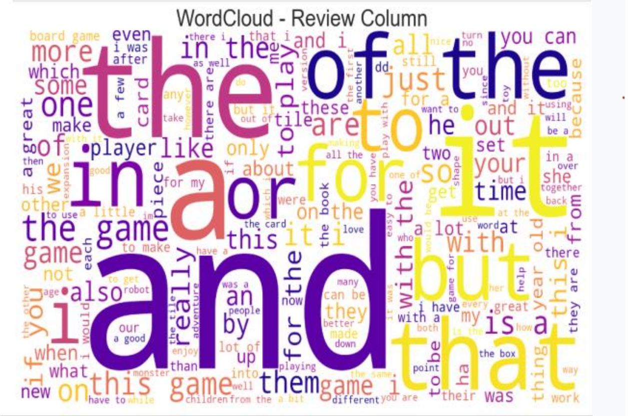
Looking at the clusters more closely there is an opportunity for the marketing team to target the group with high remuneration and a low spending score (Cluster 2 - green), as there is potential for them to spend more at Turtle Games. Perhaps analysing and investigating what worked with high remuneration and high spending score (Cluster 0 - blue) could help this cluster spend more. Evident below:

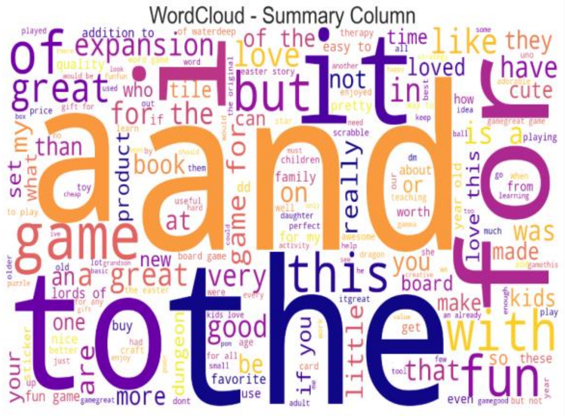


3. Analyse customer sentiments with reviews. – NLP

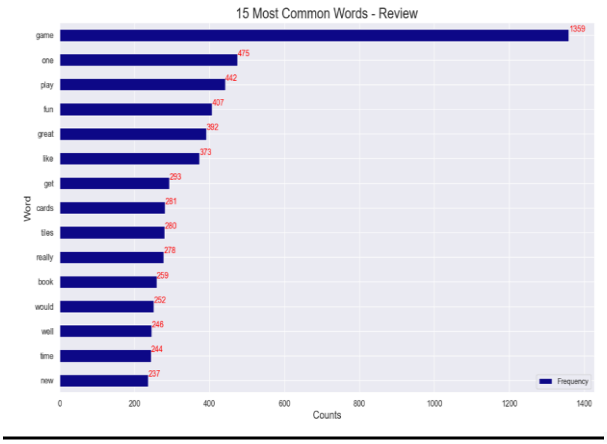
We applied natural language processing (NLP) to determine how social data (e.g., customer reviews) can be used to inform marketing campaigns.

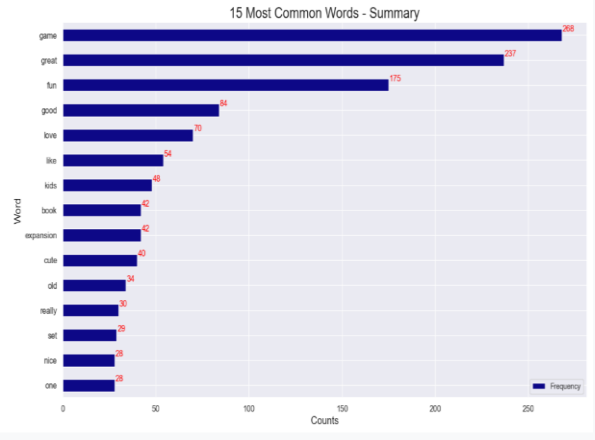
The process we followed was to drop all unnecessary columns to retain the review and summary columns as well as look for any missing values. We then cleaned the data by changing all text to lower case and joining with a space whilst removing any punctuation. Subsequently, we removed the duplicates and were able to tokenise the words and create word clouds. Evident below:



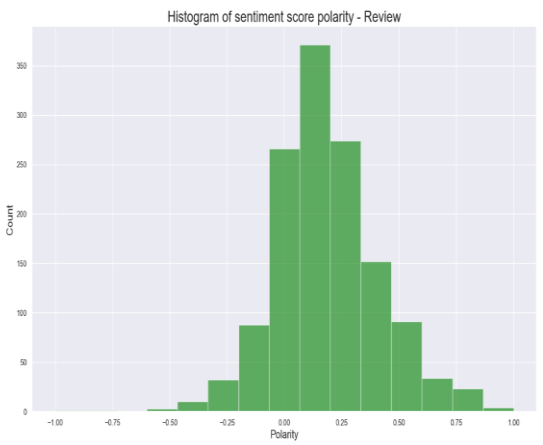


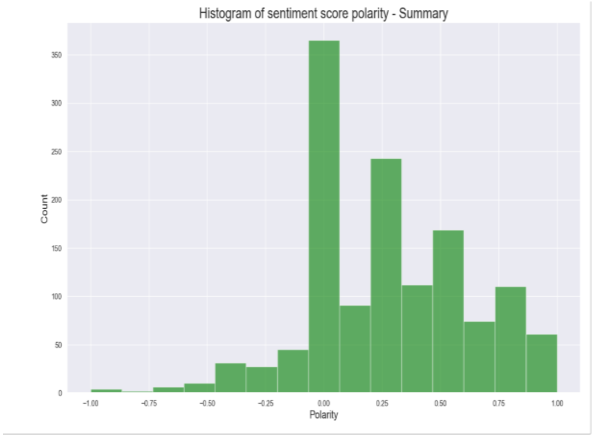
After applying frequency distributions to both the columns we realised that a lot of the words were stop words, so the next step was to remove the alphanumeric characters and stop words. The most common words for each column are evident below and had a neutral or positive sentiment:





We then visualised the sentiment score polarity for both columns. For both columns most of the sentiment score polarity was either neutral or positive. We created histograms for both the columns, it showed that majority of the comments were again neutral or positive. Evident below:

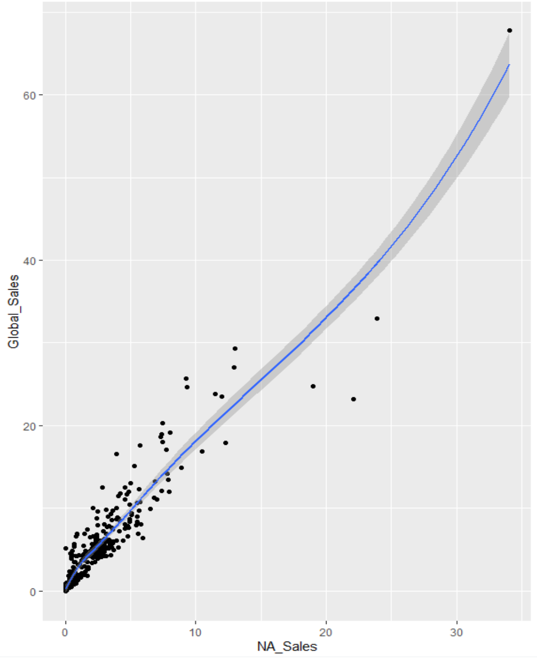


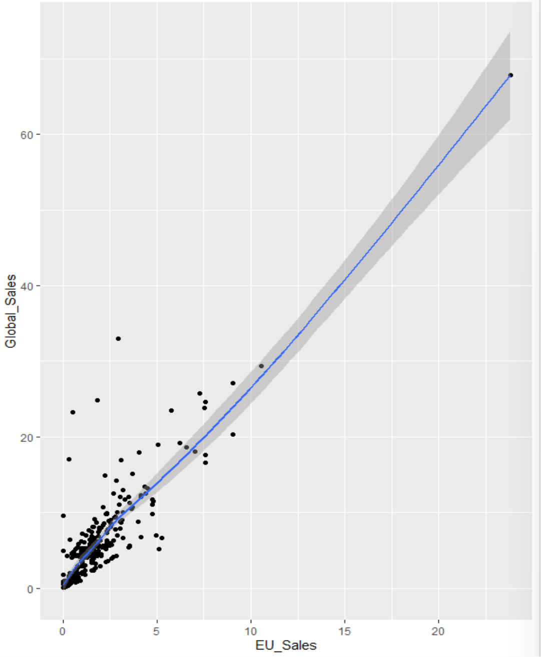


Upon reviewing some of the negative reviews it shows inconsistencies. Such as ‘I bought this as a christmas gift for my grandson it’s a sticker book so how can I go wrong with it’ had a polarity of -0.5 but seems to be a neutral comment. This can suggest that the polarity analysis may misclassify some comments as there are few more neutral or positive comments that have a negative polarity due to having one negative word in the comment, however, the intention is either neutral or positive. Further investigation or analysis will need to be done here.

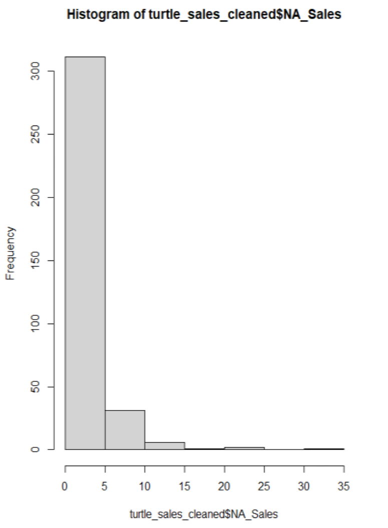
4. Visualising data to get insights.

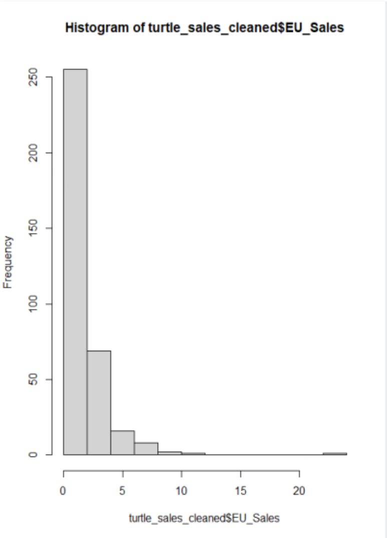
We explored and prepared the data set for analysis by utilising basic statistics and plots. By using the qplot function we were able to determine that there were positive relationships between the 3 regions sales per product, EU, NA and Global. The 2 qplots with the best relationship, best trend line shape, were Global sales vs NA sales and Global sales vs EU sales, as majority of the Global sales come from these two regions. Evident below:

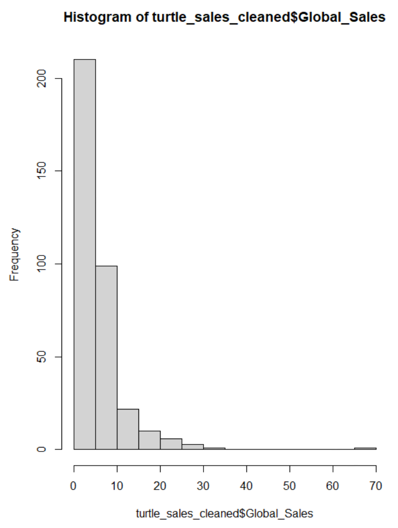




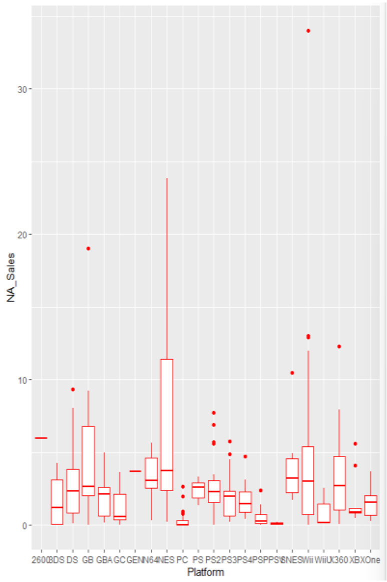
Observing the histograms we saw a common trend for all 3 regions, they were all skewed to the right. An interesting point is that for each histogram the last bin means there is an outlier that is generating a lot of revenue and this will need to be investigated to see if its an anomaly or a way to drive sales for other products. For NA and EU the most density was between the £0 - £5 million bin and for Global it was between the £0 - £10 million bin. Evident below:

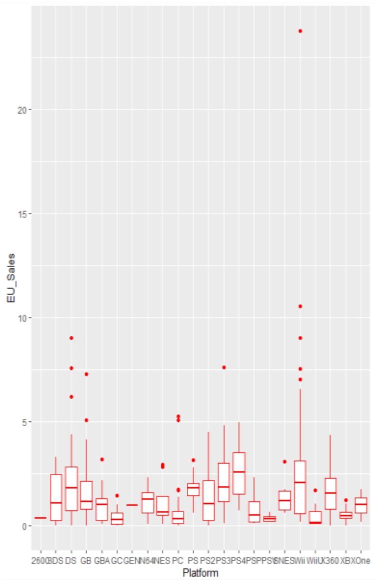


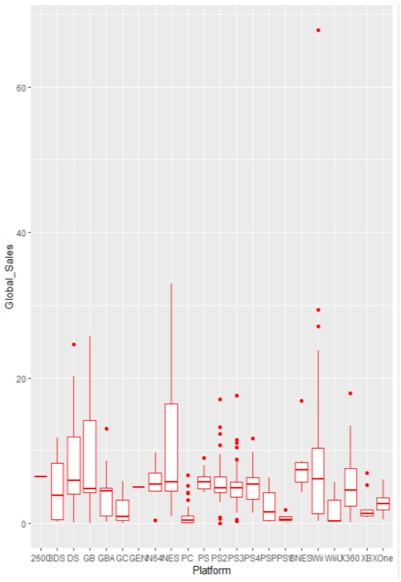




Whilst plotting the boxplots for sales by each platform instead of product we found a common theme across the 3 regions. There were outliers throughout the data set but there was an extreme outlier for the Wii. We would need to investigate why this amount is so extreme. Evident below:



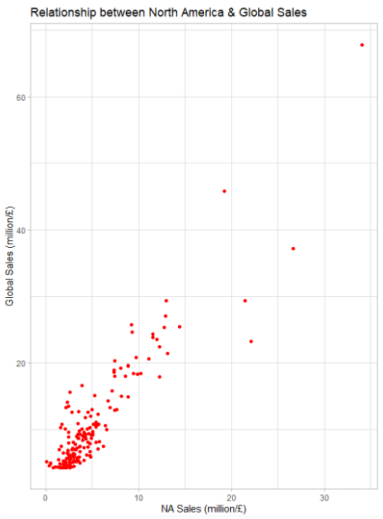


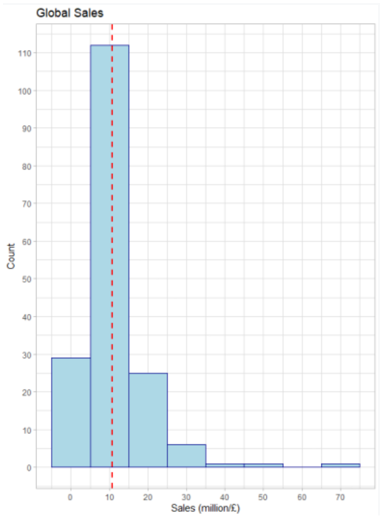


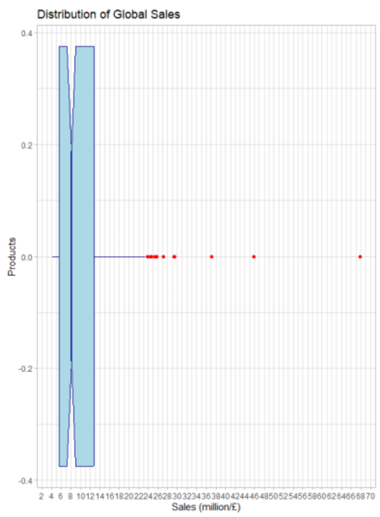
5. Cleaning, manipulating, and visualising the data.

We utilised R, to explore, prepare, and explain the normality of the data set based on plots, Skewness, Kurtosis, and a Shapiro-Wilk test.

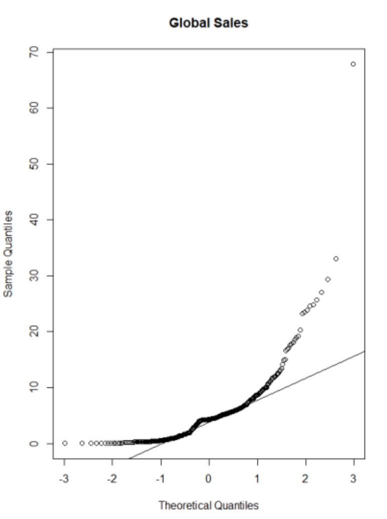
Firstly, by using the group by function we were able to group data based on product and determine the sum per product. We created some plots based on the new grouped data. Below are the plots for the Global sales. Evident below:

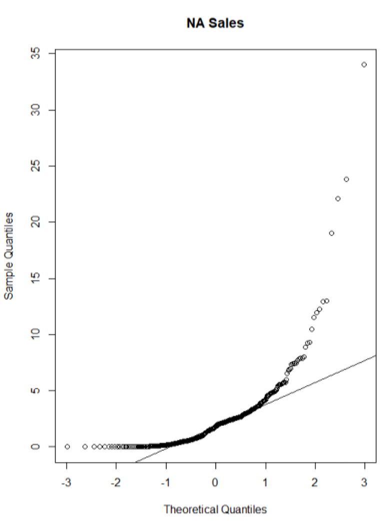


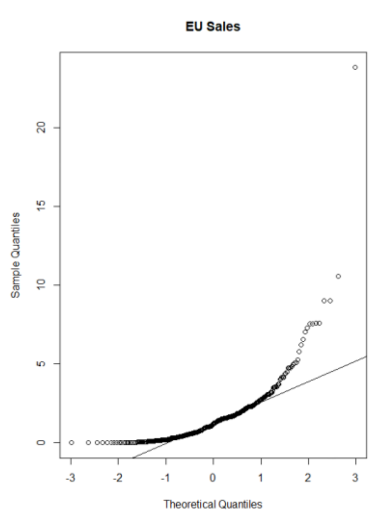




We then created Q-Q plots for each of the three regions with a line of best fit. Evident below:



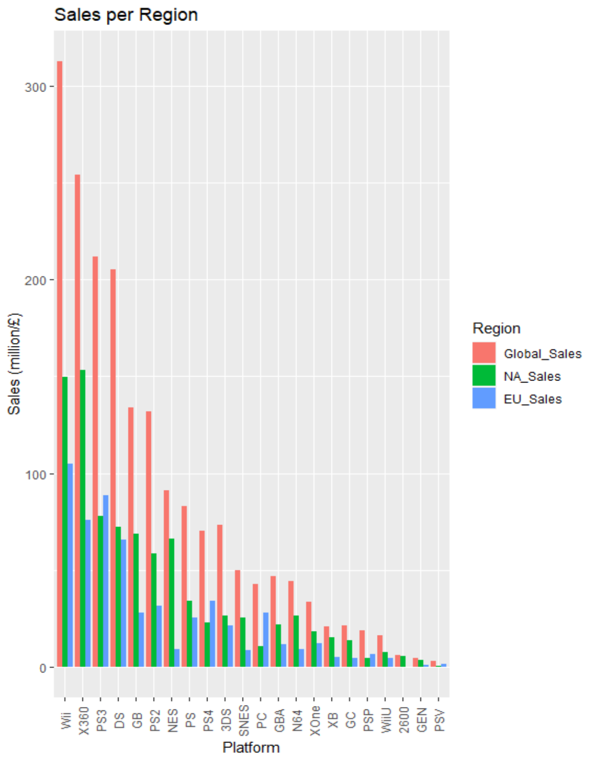


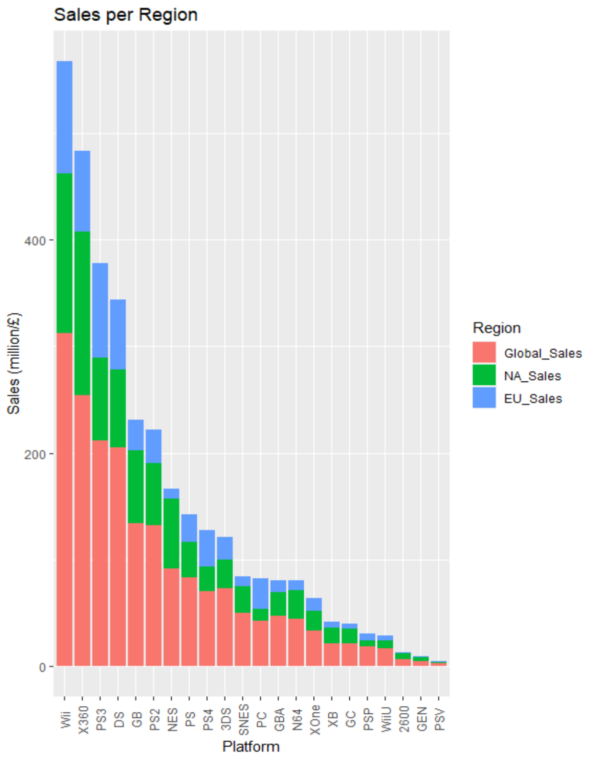


After conducting the Shapiro-Wilk test with very small p values we could confirm that the data is not normally distributed we conducted the skewness and kurtosis for each of the regions. Each region had a skewness of almost 3 meaning the data sets are extremely skewed as the values are greater than 1. Similarly, with kurtosis close to 16 it suggests high kurtosis with heavy tails and more outliers.

Determining the correlation all regions had positive correlations. NA and Global had the highest positive relationship of 91.62%. Global and EU have a strong positive correlation of 84.86% and EU and NA had a positive relationship of 62.09%. Based on this it seems that products that are popular in NA would be popular with EU as well helping the sales team.

Further exploration into looking at sales per platform we had to melt the data to be able to use at as a fill when creating a grouped and stack bar plot to better visualise the sales. See below:



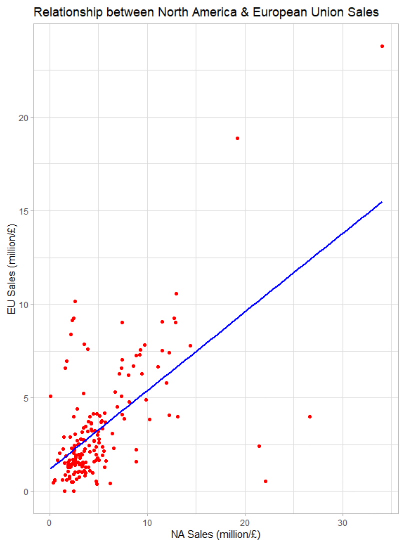


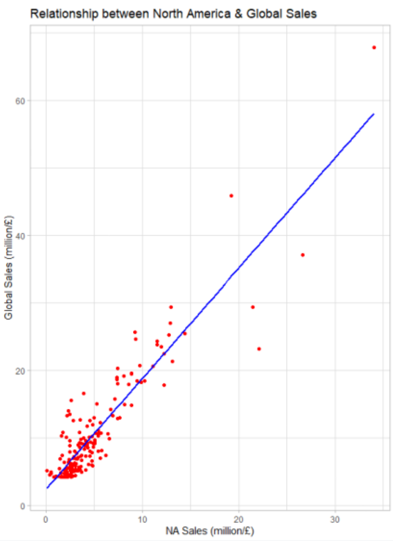
The 5 best global selling platforms are the Wii, X360, PS3, DS and GB with the Wii having the most sales with £312.56 million. The 5 worst global selling platforms are the PSP, WiiU, 2600, Gen and PSV with the PSV being the lowest selling platform with £3.34 million. The 3 most popular platforms in NA are in order of X360, Wii and PS3 for EU the 3 most popular platforms in order of Wii, PS3 and X360. The 3 least popular platforms in NA are PSP, Gen and PSV. The 3 least popular platforms in EU are PSV, Gen and 2600. Questionnaires could be sent out to understand why customers prefer certain platforms to try promoting sales.

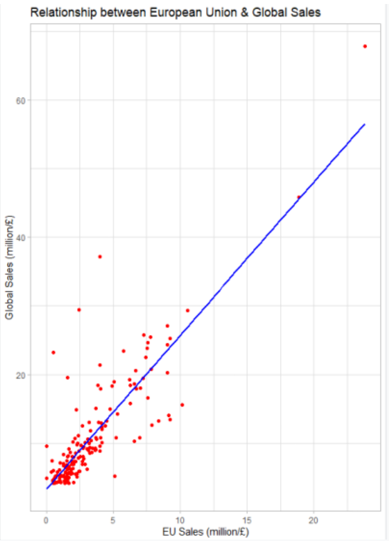
6. Making recommendations to the business.

We investigated any possible relationship(s) in the sales data by creating a simple and multiple linear regression model.

Initially we created linear regression models to see correlations between the regions. We compared each region to each other. After constructing the models, we used the summary function to extract the R-squared values. Values observed were EU ~ NA 38.56%, Global ~ EU = 72.01% and GL ~ NA 83.95%. On the relationships below we can clearly see that Global sales vs the other regions the trend line is closer to 45-degree angle and that the points are closer to the trend line.

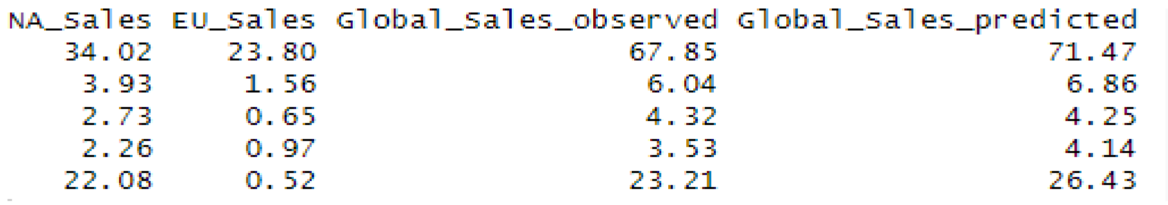






Alternatively, we used a multiple linear regression to determine the predicted Global sales by using the EU_Sales and NA_Sales observed values. We then used the summary function on the model to determine the R-squared value which was 96.68% meaning the variation on Global sales are due to the NA and Eu sales. The value is very close to 100% which means we can reliably utilise this model to predict global sales.

We then used our multiple linear regression model to compare our predicted values to the observed values and the following results were seen:



```
NA_Sales EU_Sales Global_Sales_observed Global_Sales_predicted
   34.02    23.80                 67.85                  71.47
    3.93     1.56                  6.04                   6.86
    2.73     0.65                  4.32                   4.25
    2.26     0.97                  3.53                   4.14
   22.08     0.52                 23.21                  26.43
```

We can see the predicted values are quite close to the observed values further indicating that we can confidently use this model for future predictions of Global sales by inputting the NA and EU sales.

Sources used:

https://www.researchgate.net/profile/Doaa-Mohey-El-Din/publication/313843559_Negative_Polarity_Levels_for_Sentiment_Analysis/links/58aa093692851cf0e3c6baa4/Negative-Polarity-Levels-for-Sentiment-Analysis.pdf

https://stackoverflow.com/questions/47000494/how-to-add-mean-and-mode-to-ggplot-histogram

http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization

https://www.geeksforgeeks.org/how-to-create-a-grouped-barplot-in-r/