LSEDA301: Data Analytics for Organisational Impact

Predicting future outcomes

Alexandra Van Veijeren

The business problem

Turtle Games is a global retailer and manufacturer of video games, books, board games and toys. The business wants to improve sales performance by utilising trends in customer data. Specifically, Turtle wants to understand:

1. How customers accumulate loyalty points.
2. How customer groups can be used to target specific target markets.
3. How social data can be used to inform social media campaigns.
4. The impact each product has on sales.
5. How reliable their sales data is.
6. If there are any relationships between North American, European and Global Sales.

Analytical approach – Python

The first part of the analysis was done in Python using the Reviews dataset. This dataset contains both numerical and textual data about Turtle Games reviewers.

To prepare the data for analysis, the data was cleaned using basic exploratory analysis. The data had 2000 observations and 11 columns. There were no missing values and no erroneous entries. Columns were renamed for ease of use and unnecessary columns dropped.

The three focus areas for the analysis of the reviews data:

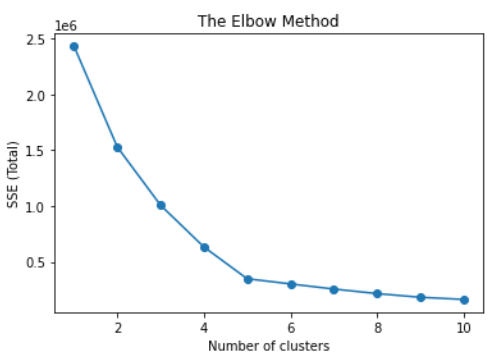
1. Identify predictors of customer loyalty points.
2. Identify groups of customers to inform marketing materials.
3. Identify how social media data can be used to inform marketing campaigns.

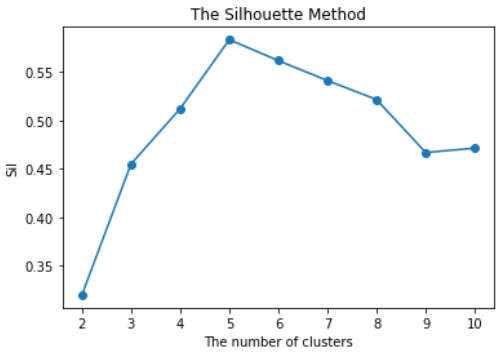
To establish the relationships that numerical data has with customer loyalty, simple linear regression was used. The models were built using the statsmodels library in python. For the three models:

- R-Squared values and p-values of coefficients were analysed to identify the strength of the relationship between the two variables.
- The regression line was plotted on top of the data visualise the fit of each model.

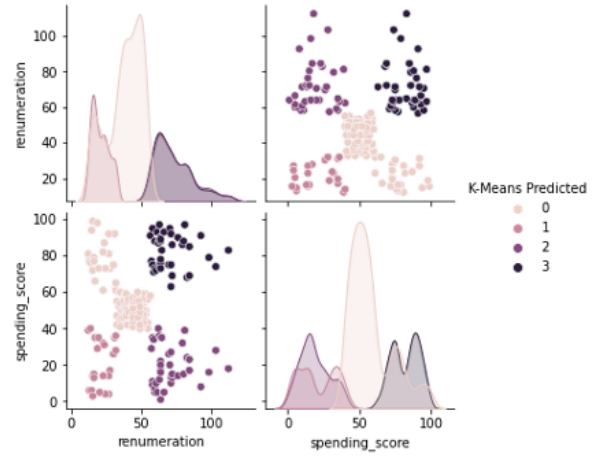
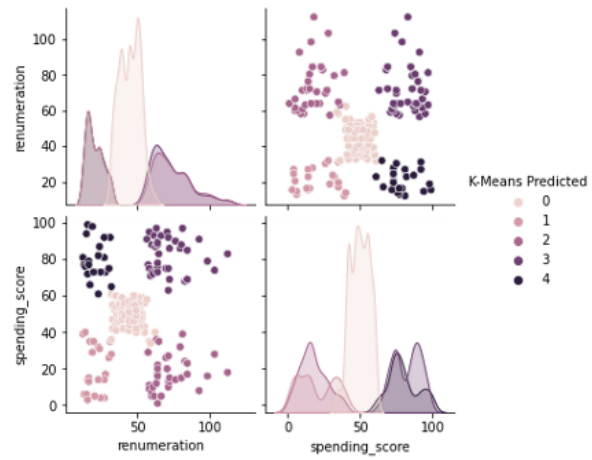
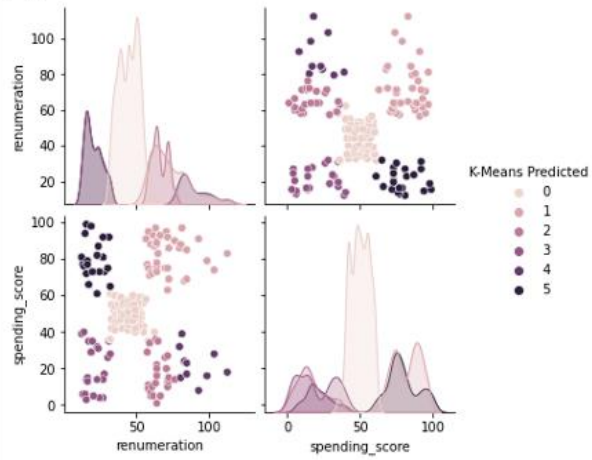
K-means clustering is an unsupervised learning algorithm and was used to identify existing customer clusters within the data. Renumeration and spending score were used as input variables.

To identify the optimal number of clusters, both the silhouette method and the elbow method were used.





Both methods suggested that five was the optimal number of clusters to use. To verify this, clusters size 4 and 6 were also tested. The below table explains in detail how the optimal number of clusters was chosen.

Pairplot for cluster size = K	Analysis
K = 4 	K = 4 clusters - There are three distinct clusters with little overlap. - Cluster zero does look bigger than the other clusters. In-cluster variation may be larger than what is neccesary. - The kernel distribution for cluster zero for both renumeration and spending score are quite wide. This indicates a large in cluster variation value, as the data is spread out over a long axis. - The goal of clustering is for in-cluster variation to be small. Therefore it would be beneficial to have a larger number of clusters.
K = 5 	- Adding one additional cluster changed the distributions of spending score and renumeration. - Group zero no longer has such a long right-ward tail, implying that the in-cluster variation is smaller. - The clusters themselves are evenly defined.
K = 6 	- Adding an additional cluster did not change the variation of any of the clusters by a large amount.

The last cluster size that brought meaningful change into the division groups was K = 5. Customers characteristics were identified based off the data points cluster allocation.

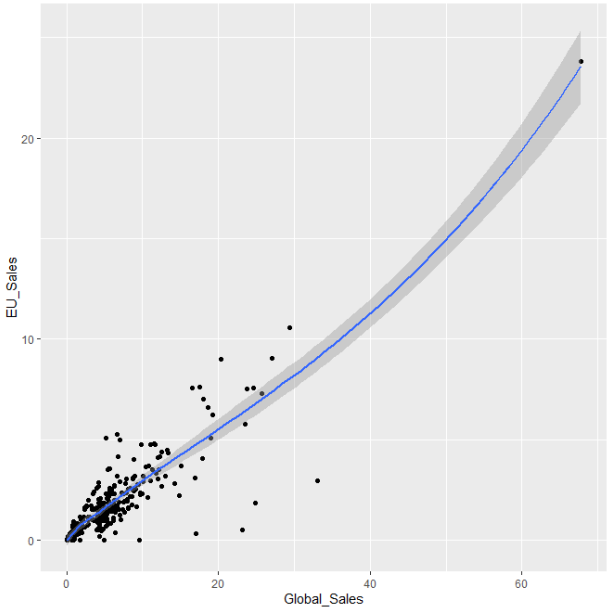
Lastly, to identify how social media can be used to inform marketing campaigns, the reviews data was analysed using WordClouds and VaderSentimentAnalyser.

After removing punctuation, stopwords and tokenizing the words a WordCloud could be generated to identify the top used words in both the reviews and summaries column.

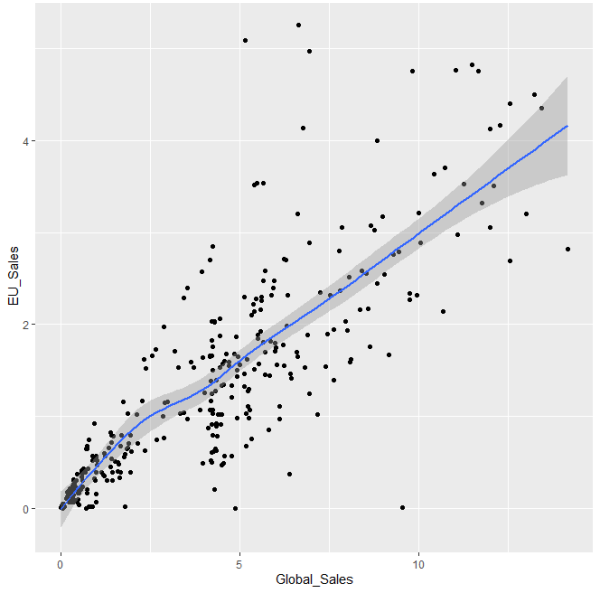
The sentiment of reviews and summaries was then generated using VaderSentimentAnalyser. Histograms of the compound score were built to show the overall sentiment of the reviews data. The most positive and negative reviews were also identified.

The sales data was analysed in R.

The data was prepared using visual exploratory techniques. Outliers within the three sales columns were identified. The goal of this analysis was to get a picture of the true relationship between NA, EU and Global Sales numbers.



The relationship between Global and EU Sales with outliers included.



The relationship between Global and EU Sales with outliers excluded.

With outliers removed, the true relationship appears to be more linear than it is with outliers included. The outliers also make sense in the context of the other variables. Outliers were therefore not excluded.

To identify the impact per product the top and bottom ten products were then found using the arrange() and groupby() functions in R. The proportion of their sales per category (EU, NA) wee found using the mutate() function in R. I focused on the top ten sellers in Europe, North America and globally to gain actionable insights from the large dataset.

To find the relationship between the sales data, the distribution of the data was explored. QQ-Plots and the Shapiro-Wilk test found the data to not be normally distributed. The data is skewed to the right and has very large excess kurtosis due to outliers.

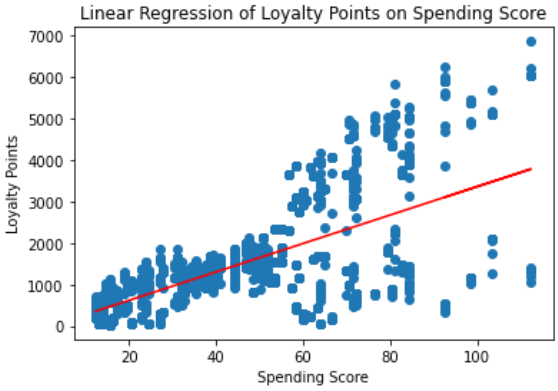
Visualising the relationships between EU, NA and Global Sales suggested that there is a linear relationship between EU and Global and NA and Global Sales. These relationships were estimated using simple and multiple linear regression.

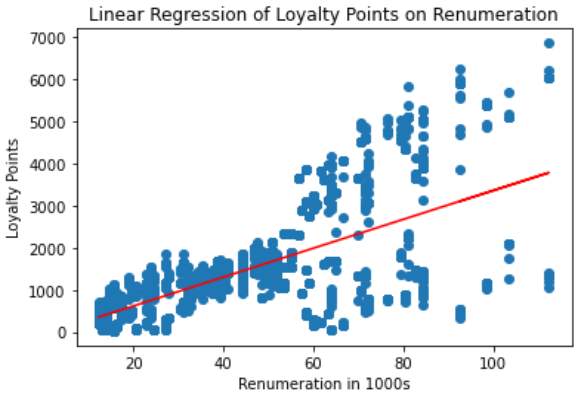
Five models were built and tested. The multiple linear regression without log transformation had the greatest predictive accuracy. The residuals of the model deviated from a normal distribution, but for large enough sample size, one can assume that the distribution of the residuals will tend to zero. For models where predictions are the main goal, this is enough to ignore multicollinearity (Frost, J. n.d.)

Visualisations, insights, and recommendations

How do customers accumulate loyalty points?

The two simple linear regression models had low R-Squared values. High F-statics for both models indicate that there is a significant relationship between renumeration, spending score and loyalty points.





A multiple linear regression validated this hypothesis, with an R-squared value of roughly 82.70%. The above graphs show that both models fit well for low to mid-range loyalty points, but not for high values. Therefore independently, renumeration and spending score could not fully explain how customers accrue loyalty points.

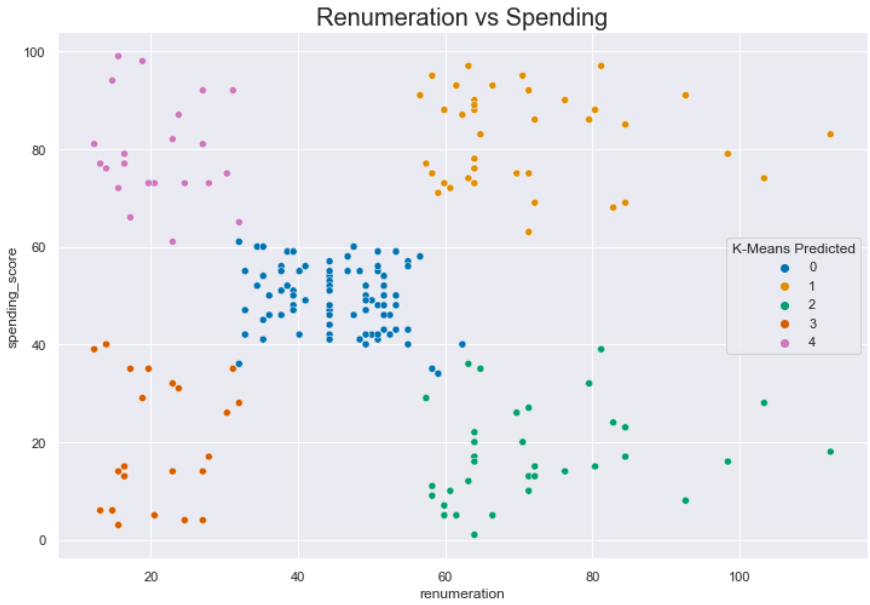
Together, renumeration and spending score can explain how customers generate loyalty points. A multiple linear regression validated this hypothesis, with an R-squared value of roughly 82.70%.

This suggests that a client's salary and spending score work together to be good indicators of how many loyalty points they will generate while they are customers of Turtle Games.

Turtle Games could do another investigation to identify variables to add to the model to improve it, increasing its predictive accuracy.

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

The existing customer base can be broken into five clusters.



The above scatterplot illustrates the salary level and spending score of the clusters. It quickly gives the viewer a summary of two characteristics of each group, as well as a rough idea of the size of each cluster. The below table summarises the average age, salary, and spending score of the above clusters:

Cluster	Av. Age	Size of cluster	Av. Renumertion (Pounds)	Av. Spending Score	Proportion of customers holding tertiary degree or higher
0	42	774	44 418.79	49.53	88.76%
1	36	356	73 240.28	82	87.64%
2	41	330	74 831.21	17.42	86.67%
3	43	271	20 424.35	19.76	80.07%
4	31	269	20 353.68	79.42	95.91%

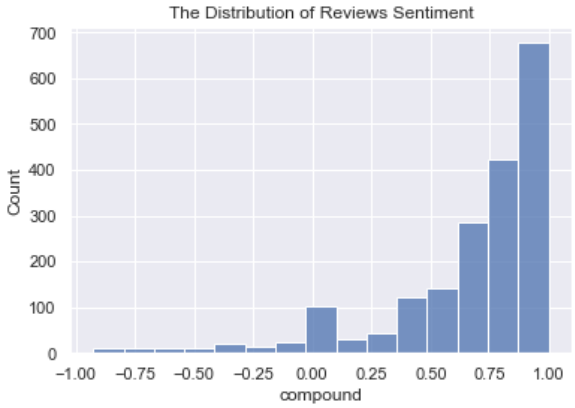
The main differentiating factor for Turtle Games customers is their salary. Most of their customers hold tertiary degrees and can be expected to be between 30 and 40 years old.

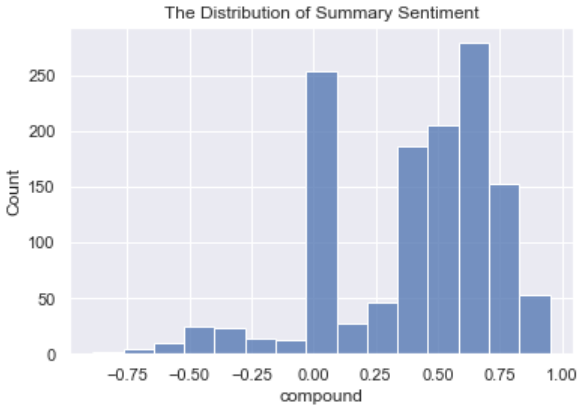
Turtle Games can use this information to target customers holding tertiary degrees who are high earners. The above clusters suggest that a large portion of their customers fall into this bracket already, but do not spend consistently.

The same can be said of customers who are low earners.

How can social data be used to inform marketing campaigns?

Sentiment analysis found the reviews for Turtle Games to be largely positive. Summary sentiment shows a greater number of neutral sentiment. This may be because the summary column attempts to isolate the main message of the review. This may result in a loss of emotive language which could also cause the true sentiment of the review to be missed. However, both distributions show that customers are mostly happy with the products they purchase from Turtle Games.





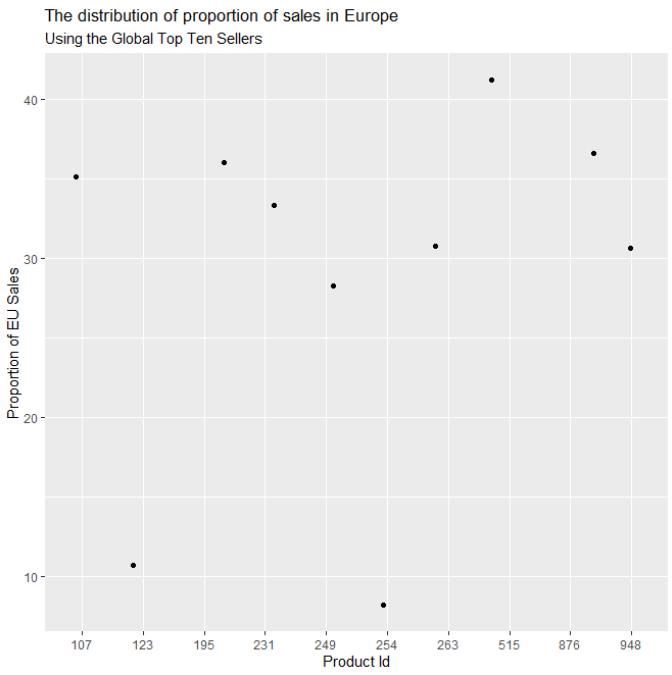
One of the most commonly mentioned words in the summary data was ‘expansion’ and ‘expansion pack’. Turtle Games should review their expansion pack products and try to identify which ones are popular with their customers, as the data indicates that it is getting good reviews online. They could then identify if it is worth marketing that expansion pack further.

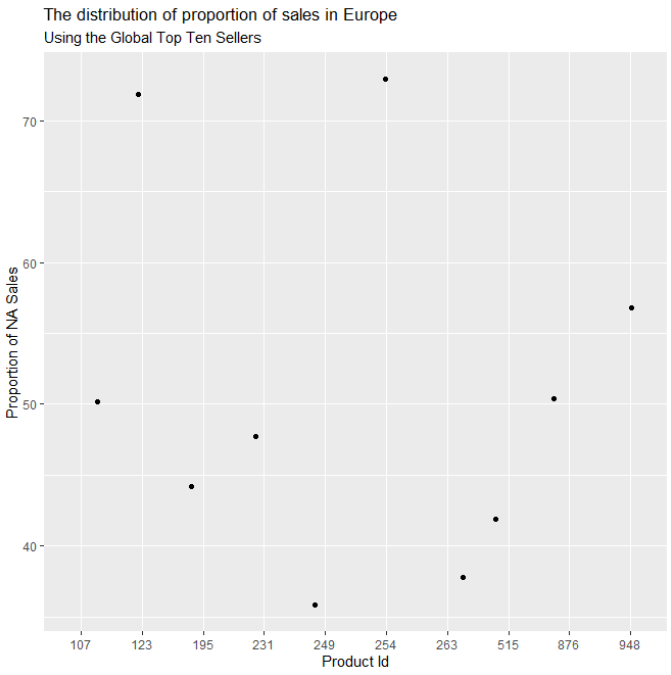
The reviews with the highest negative overall sentiment showed that many customers were disappointed in their products. Some of those reviews were relating to the quality of the product – referencing the packaging or the instructions. Turtle Games could send out a survey to customers asking them about their experience with the quality of the products they have purchased. This can form a part of its own marketing – Turtle Games trying to engage with their customers and build up a good rapport with them.

Since the reviews tend towards positive sentiment, Turtle Games could confidently include customer satisfaction in their marketing materials and share these reviews.

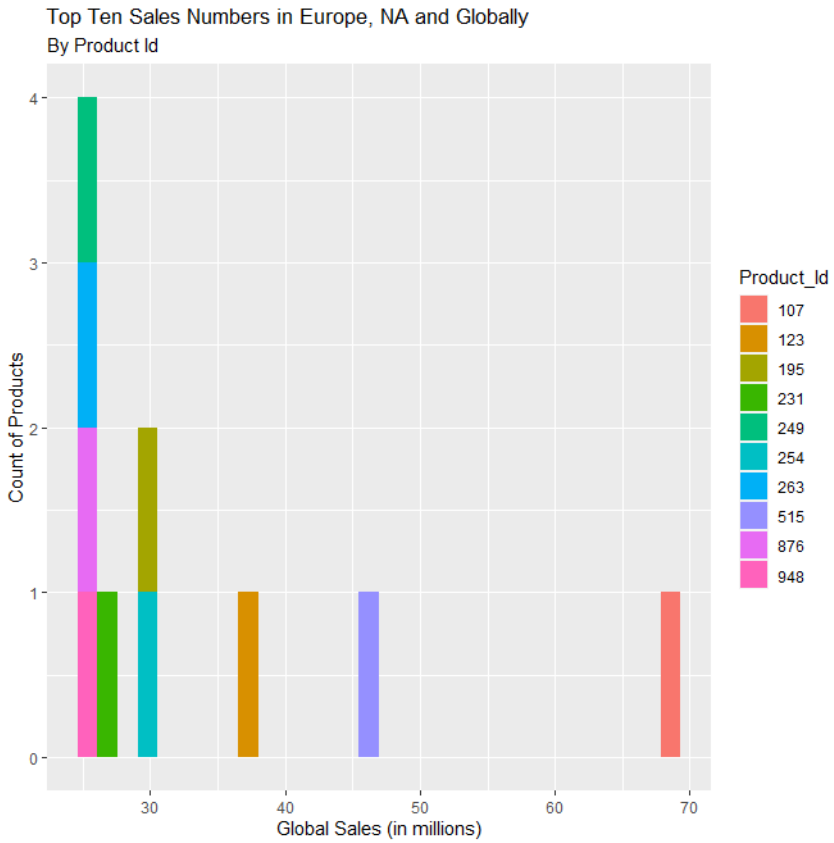
What impact does each product have on sales?

The following graphs show the distribution of proportion that European Sales and North American Sales make of the global top ten seller for Turtle Games. The viewer can immediately get a sense of the range of the proportions that the different regions make up of the total sales, and they are able to identify the impact that a product had in a region.





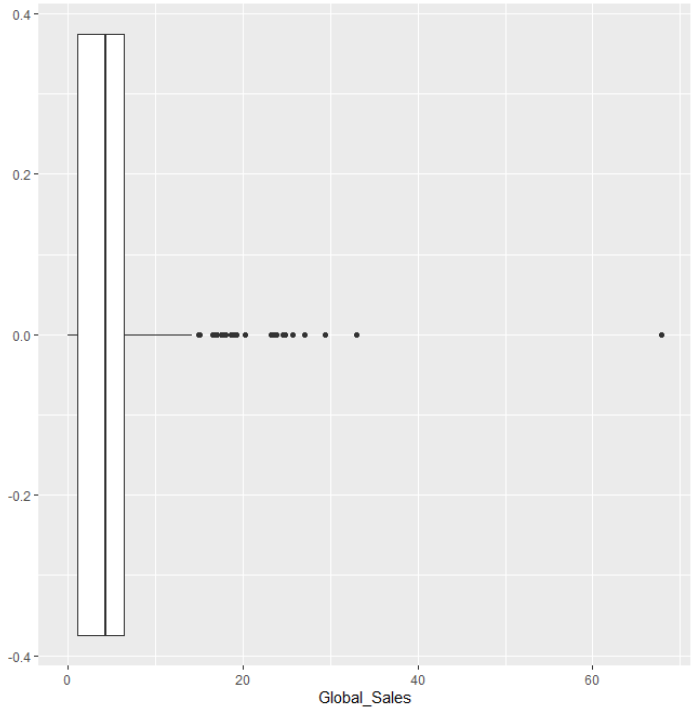
North American Sales made up a greater proportion of global sales than European sales, as demonstrated by the range of values. Below global sales are coloured by Id.

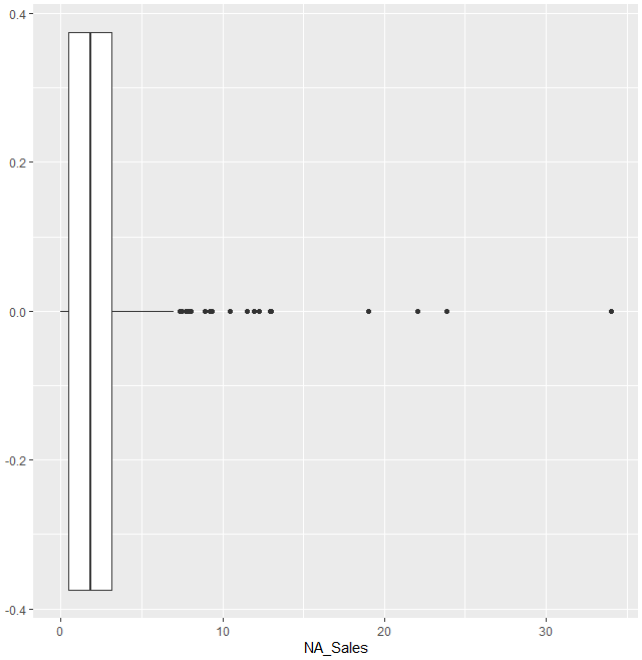


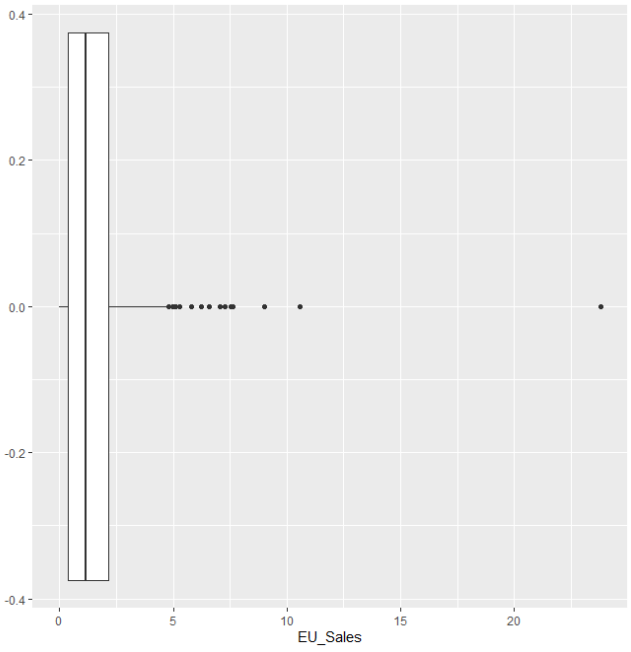
Products 123 and 254 both performed well in North America but very poorly in Europe. For Turtle Games, a further investigation into why these products sold so differently in the two regions would be useful.

How reliable is the data?

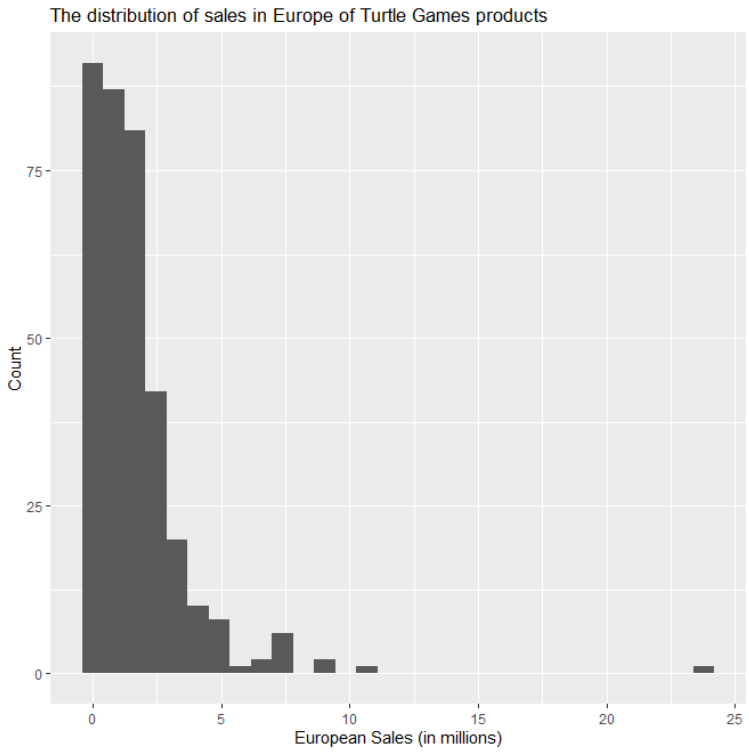
The Shapiro-Wilk tests on European, Global and North American sales showed that none of the columns are normally distributed. As the below boxplots illustrate, the data is positively skewed. All three columns have large kurtosis values due to the large outliers in the right tails.







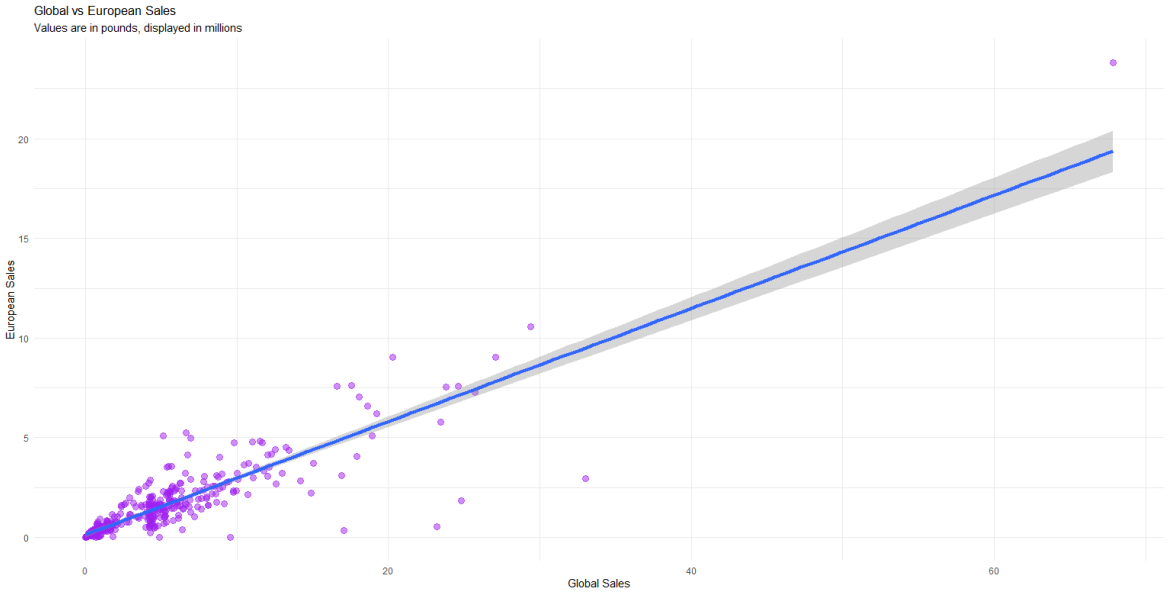
The data is not unreliable because it is not normally distributed. The data may tend towards a different distribution. Looking at the distribution of EU sales, one could hypothesize that the data follows a Chi-Square distribution and perform statistical inference and modelling with this in mind.

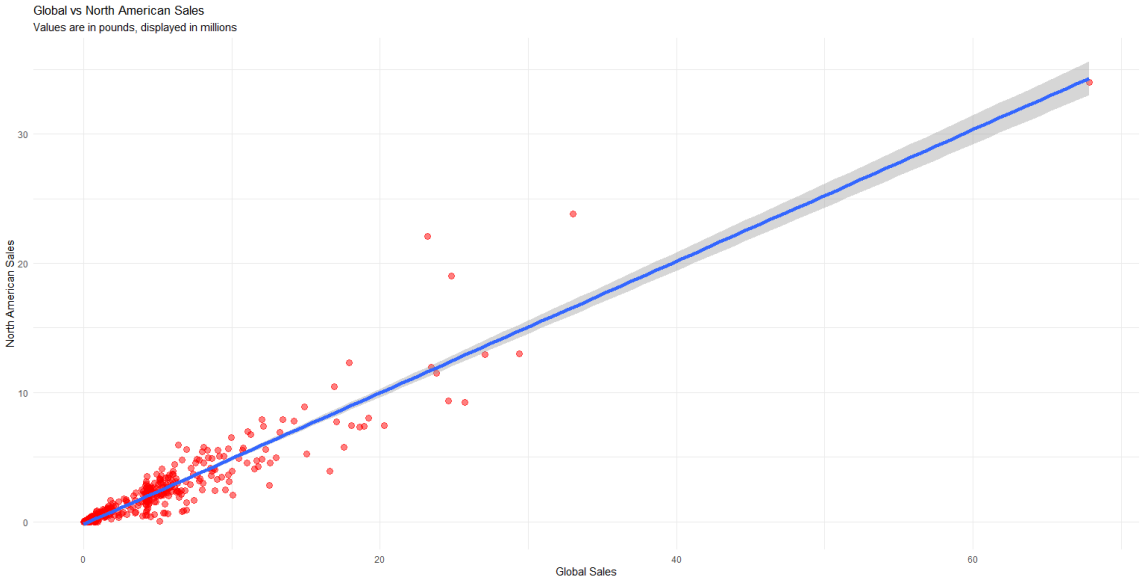


The data is also not required to be normally distributed to build linear and multiple linear regression models.

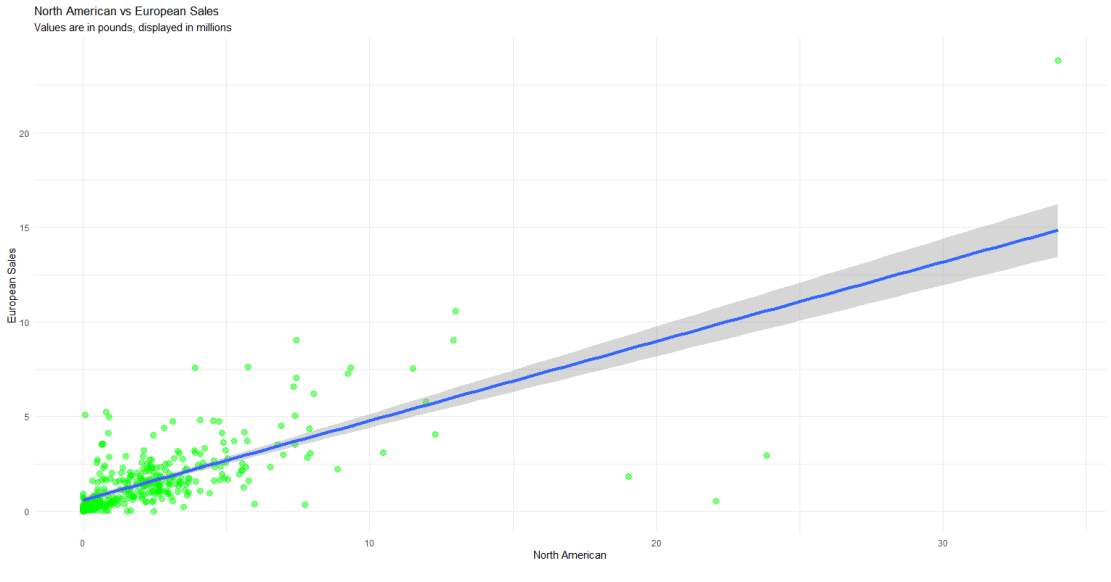
What relationships are there between North American, European and Global Sales?

Scatter plots of North American and European Sales against Global Sales respectively show that sales in both regions have a linear relationship with Global Sales.





European Sales and North American Sales are also positively correlated, but their relationship is not as linear nor as strong as with Global Sales.



The data can be best modelled by a multiple linear regression:

$$Y = 0.22175 + 1.34197x_{eu} + 1.15543x_{NA} + \varepsilon$$

This model had the highest adjusted R^2 value and the closest predicted value out of the five models considered.

Therefore, there exists a positive linear relationship between Global, European, and North American sales.

I do not think that this data is accurately modelled by a linear relationship. After using a log transformation, the residuals still deviated for normality. Turtle Games should model the data using another regression technique – such as second-degree polynomial.

References

Frost, J. (n.d.). *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions*. Retrieved from Statistics by Jim: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/#:~:text=Multicollinearity%20occurs%20when%20independent%20variables%20in%20a%20regression,you%20fit%20the%20model%20and%20interpret%20the%20results.