Turtle Games Sales and Customer Analysis

Background & Context

Turtle Games is a global game manufacturer and retailer, selling its own products and third party products. It sells books, board games, video games and toys. The company collects sales and customer data and has a business objective of improving sales performance by utilizing customer trends. To support this, Turtle would like to understand:

- How customers accumulate loyalty points
- How groups within the customer base can be used to target specific market segments
- How social data can be used to inform marketing campaigns
- The impact that each product has on sales
- How reliable the data is (normality, skewness, kurtosis)
- What the relationship is between European, North American and global sales.

Analytical Approach – Importing and Cleaning

Much of the analytical approach is documented in the accompanying Jupyter notebook and R script and can be followed there, however some key elements are

Customer Review Data

- Data imported into Python from raw .csv file provided by Turtle Games, identified 2000 rows and 11 columns.
- Exploration done to identify no missing values and no duplicates.
- Categorical columns reviewed for numbers of variables and suitableness of labelling. Language and Platform identified only 1 variable per column, so of no use and removed. All others found to be ok.
- Numerical columns reviewed using box plots and histograms with none of it looking normally distributed and only loyalty points having outliers. This was curious as I would normally expect outliers in salary, further review found many duplicated values which again seemed strange, however I had no better data to clean it with so left as is but is on for further clarification with Turtle.
- Cleaned dataframe exported to csv for later use.

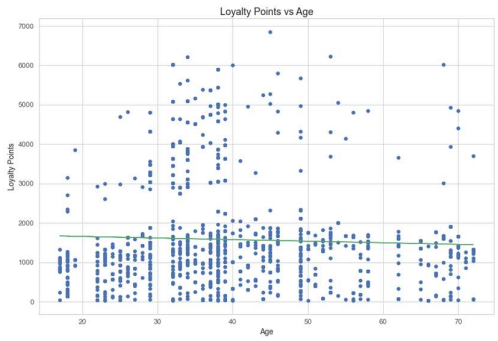
Customer Sales Data

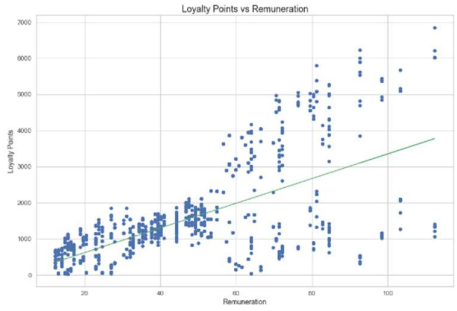
- Data imported into RStudio from raw .csv file provided by Turtle Games, identified 352 rows and 9 columns.
- Exploration done identified 2 missing values in the Year column, on further analysis it was found that these products had entries elsewhere in the data, and the Year column was updated in line with these.
- Columns given as redundant removed (Ranking, Year, Genre & Publisher), though I think there is potential value in the last 3 of those for Analysis.
- For part of the analysis the data was aggregated to total sales at product level (original data is by product by platform).
- Exploratory analysis performed using scatter plots, box plots, histograms with checks on kurtosis and skewness. The sales data was identified as not being normally distributed, positively skewed and leptokurtic. Positive correlation was found between the regional sales, with varying degrees of strength. Detail of this is in the analysis section ‘Is the Data Reliable’.

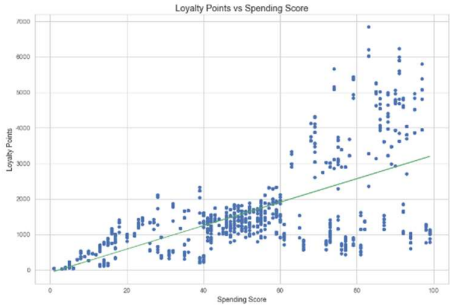
Analytical Approach – Analysis and Insights

How Customers Accumulate Loyalty Points

To understand drivers and relationships of customer loyalty points, statistical analysis was done to look at correlation between loyalty points and other numerical columns. Scatterplots were chosen as the best way to quickly visualize this, to show the relationship between 2 variables and quickly identify a relationship.

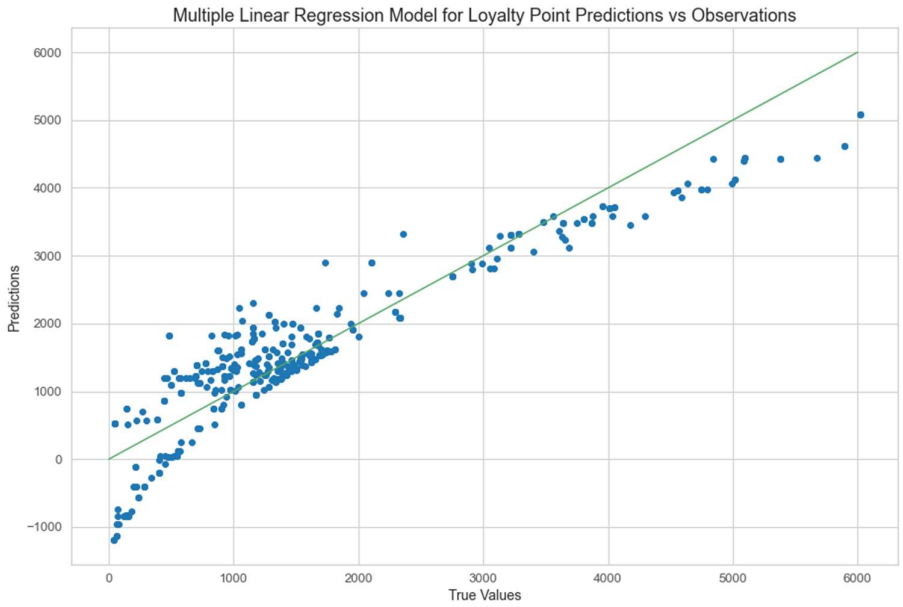






Immediately it was obvious no reliable correlation between Age & Loyalty but there was some positive correlation for the other variables. OLS models were run to show this was relatively weak with r^2 of 0.38 and 0.45 respectively. Visualizing the data this was also showed the “cone” shape of the data, suggesting heteroscedasticity. As this happened reasonably clearly after certain points, I limited Remuneration to 55k and Spending Score to 60. This yielded better results, r^2 of 0.63 and 0.60.

A multiple regression model was created to see if the variables combine for better results. R^2 improved to 0.82. On a predicted vs observed plot on the test data set and the results looked encouraging.



Implications for Turtle, Recommendations and Further Analysis

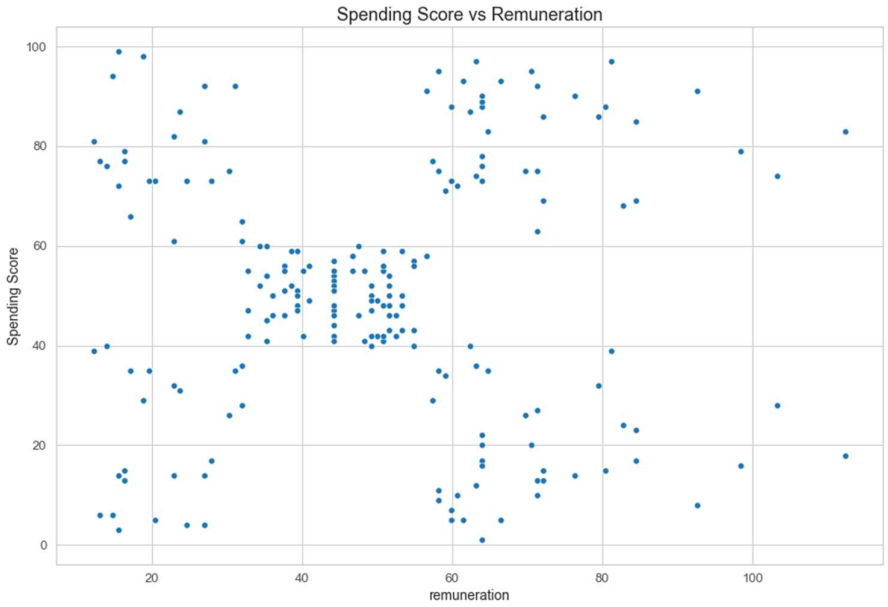
Remuneration and Spending Score combined are a reasonably good explainer for Loyalty Points. These variables can be used to ensure Turtle can predict who its most valuable customers will be. Given the improvement of r^2 when outliers were removed I would like to understand why the data deviated over these points.

Additionally would like to obtain further data to add to and improve the model, e.g. utilizing gender and education already in the dataset and enable better predictions to be made.

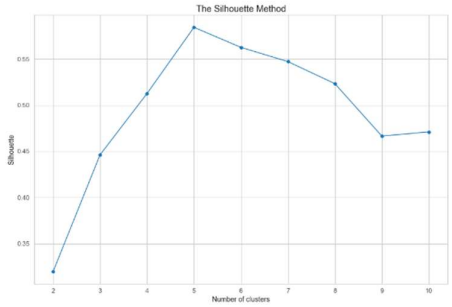
How groups within the customer base can be used to target specific market segments

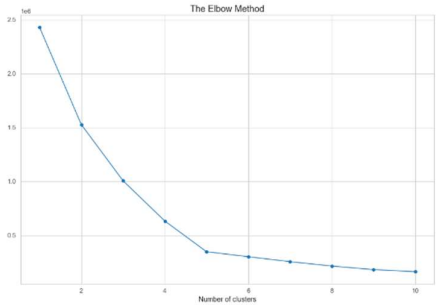
Remuneration and Spending Score were used to see if customers could be clustered into groups, using K-Means Clustering.

Before running the model, a quick scatterplot was used to see if there were any obvious visible clusters.



This initial visualisation indicated potentially 5 clusters. Elbow and silhouette plots were created to validate this observation.

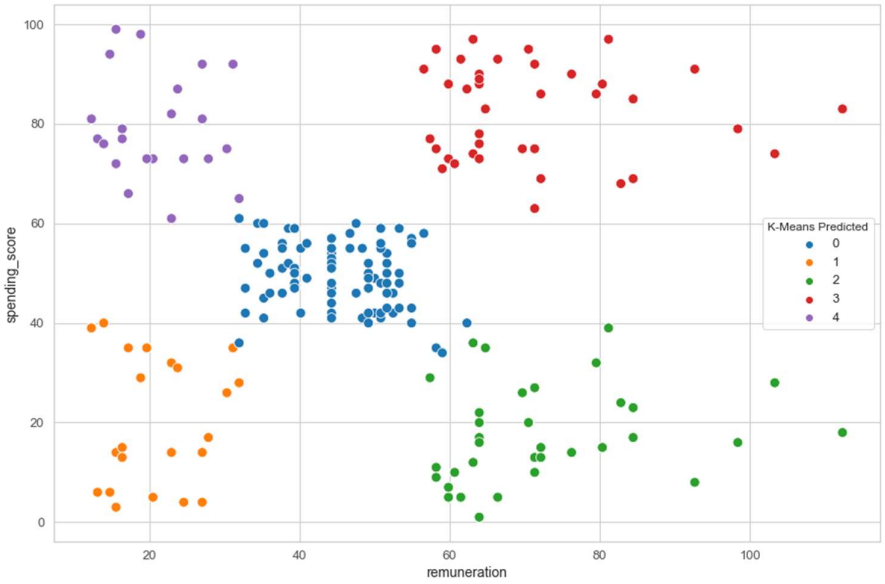




Both methods showed likely clusters would be 5, but analysis was done on 4, 5 and 6. 5 clusters were identified as optimal with the most even split.

	0	1	2	3	4	5
4 clusters	1013.0	280.0	351.0	356.0	0.0	0.0
5 clusters	774.0	271.0	330.0	356.0	269.0	0.0
6 clusters	767.0	356.0	214.0	271.0	123.0	269.0

These clusters fell into 5 categories with similar characteristics



- 0 = mid-spending score, mid-remuneration
- 1 = low-spending score, low-remuneration
- 2 = low-spending score, high-remuneration
- 3 = high-spending score, high-remuneration
- 4 = high-spending score, low-remuneration

Implications for Turtle, Recommendations and Further Analysis

These customer groups could be used for targeted marketing based on their characteristics to give tailored, appropriate adverts (e.g. cat 1 targeted with value propositions and offers, cat 3 targeted with premium content and innovation) and enable Turtle to understand its customer base better.

Further investigation could be done into other characteristics of the customers (types of products ordered, sales on promotion, click-through on mail outs, location etc.) to understand more and enable more focused targeting.

How social data can be used to inform marketing campaigns

A sample of 2,000 reviews and summary comments was analysed using TextBlob library in Python.

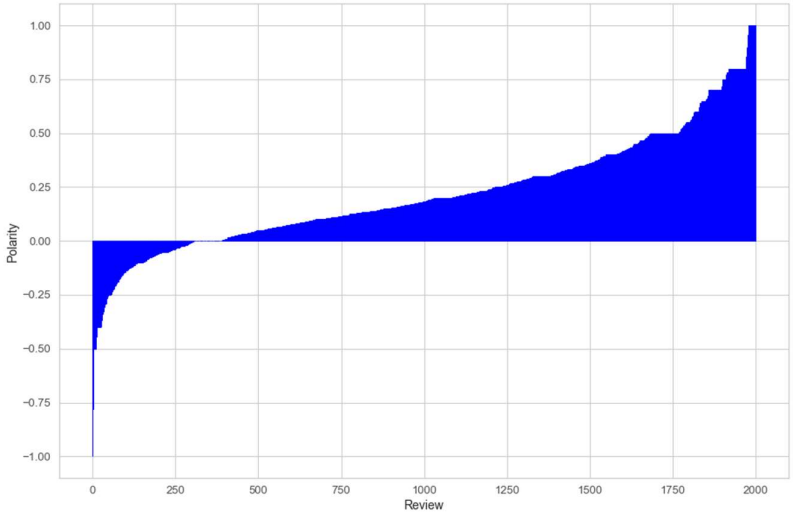
Wordclouds were produced to visualise common words (reviews tokenized and stopwords removed);

Reviews Wordcloud

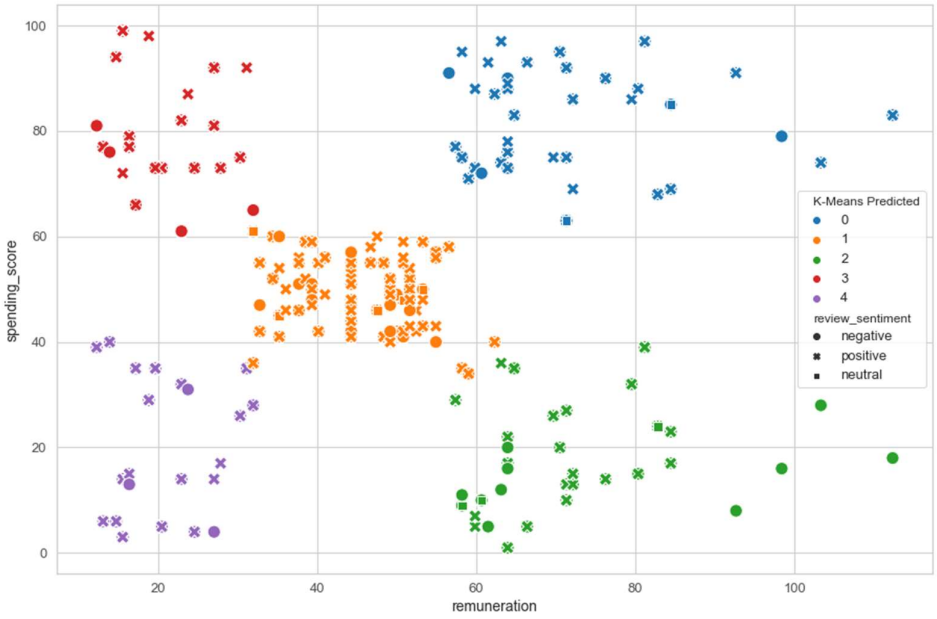
Summary Wordcloud

The most featured words were generally positive. This is good feedback for Turtle because they generally have pleased and satisfied customers and understand recurring themes in comments.

TextBlob was used to generate polarity scores to quantify the sentiment of the customer feedback. For reviews, sentiment was positive - a mean of 0.28 and over 80% positive vs 15% negative. However negative was generally more strongly expressed, illustrated below (larger amount of blue above the curve but the steeper below the x axis).



I also wanted to see if any of the clusters showed more of a tendency to post positively or negatively than the total sample, however all clusters had broadly the same mix of comment polarity.



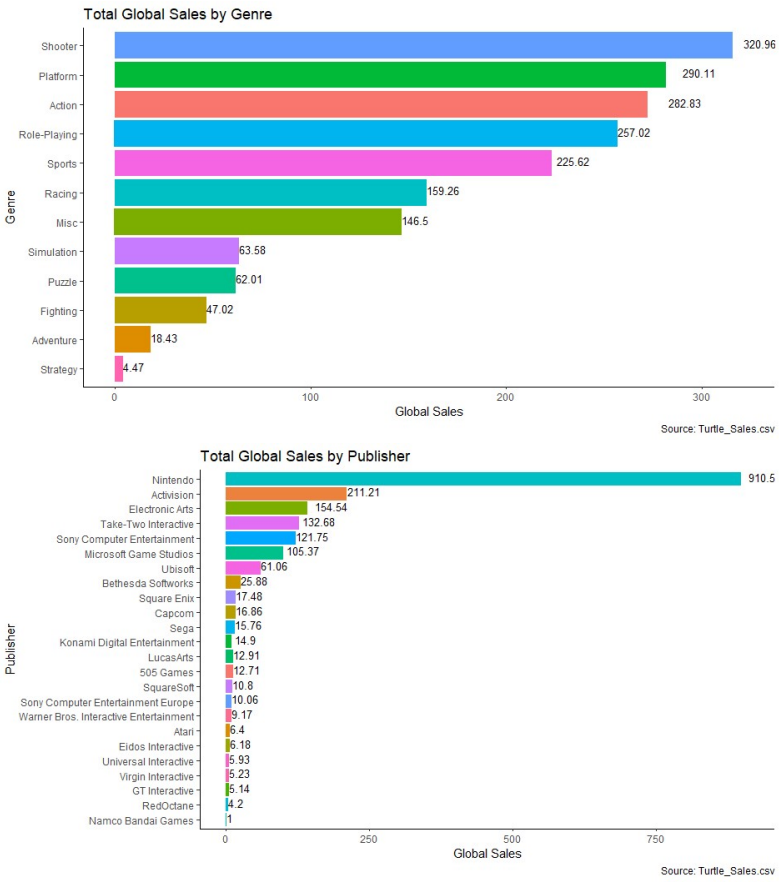
Implications for Turtle, Recommendations and Further Analysis

Turtle can use this data to understand current customer sentiment and to understand any common needs the target markets have – currently customers look to be happy and satisfied. Turtle can get a real-time view of feedback and course-correct accordingly if an issue arises (e.g. games lost in post, damage etc.) or take advantage of positive sentiment by leveraging the relationship with marketing and promotions.

Further work I would like to see in this area would be to use another library (vadersentiment) to validate the scores (some comments using TextBlob were given negative polarity when clearly positive). The review data comes with product code so this too could be overlaid to look at sentiment at product level to give further focused insight.

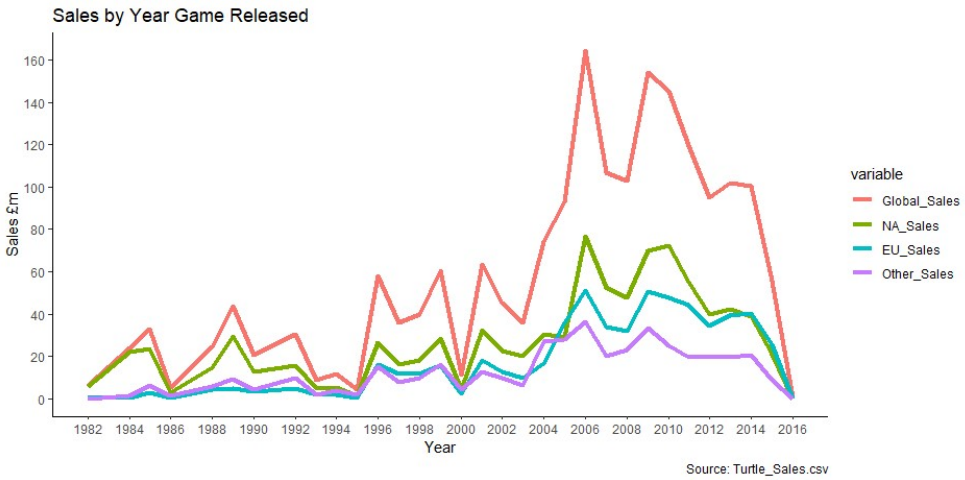
The impact that each product has on sales

To understand the impact of types of product on sales, I used bar charts as they quickly and clearly showed to top performers by Genre and Publisher



From Global Sales of £1.878m, 73% come from the top 5 Genres and Nintendo games account for almost half of total sales with 82% of total sales coming from the top 5 Publishers.

Also, when looking at sales by year (of release for each game) we can see that recent releases have not achieved as high sales as the peaks in the mid-00s (assume 2016 is not a full year yet) and NA Sales in particular have dropped significantly since 2010, so much so that EU sales for recent release games are higher.



Implications for Turtle, Recommendations and Further Analysis

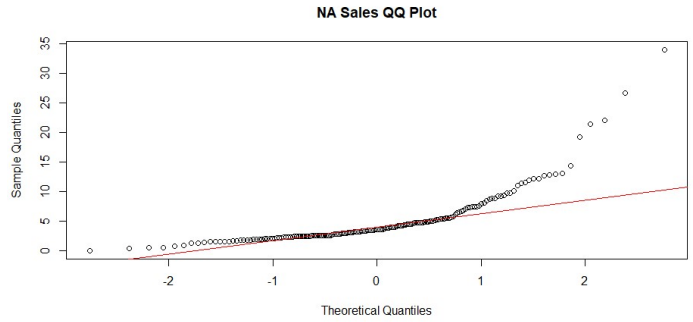
This type of information should be used by Turtle to ensure they can predict sales based on Genre and Publisher, and manage good relationships with publishers to maintain good supply, negotiate favourable margins and align marketing spend.

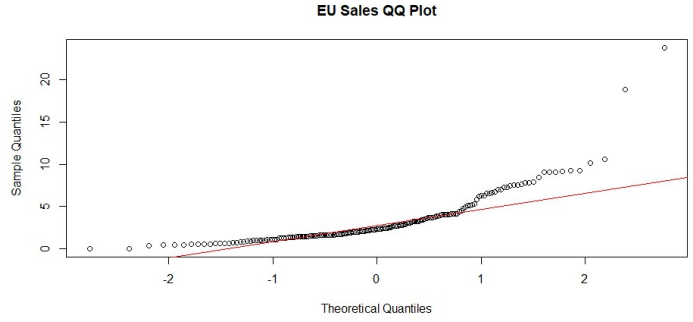
Turtle should also seek to understand why sales from recently released have not performed as well and what has caused the sales decline in NA region.

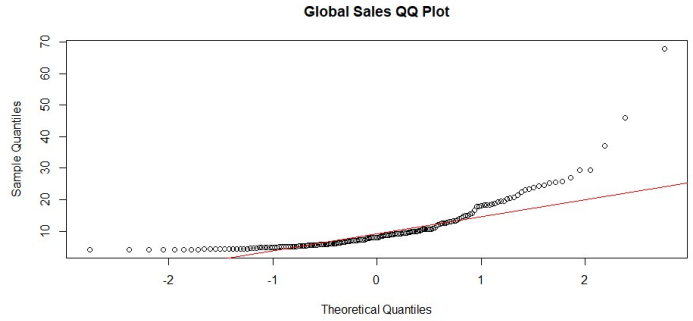
How reliable the data is (normality, skewness, kurtosis)

In order to make reliable recommendations based on data, the data itself should be of good quality.

To determine normality, QQ plots were used.







All three sales columns were similar – showing a good deal of variance from the red line in the plot, thus signalling non-normality.

Further to this, Shapiro-Wilks tests showed;

```
> # EU
> shapiro.test((sales_1_grouped$EU_Sales))

        Shapiro-Wilk normality test

data:  (sales_1_grouped$EU_Sales)
W = 0.74058, p-value = 2.987e-16

> # NA
> shapiro.test((sales_1_grouped$NA_Sales))

        Shapiro-Wilk normality test

data:  (sales_1_grouped$NA_Sales)
W = 0.69813, p-value < 2.2e-16

> # Global
> shapiro.test((sales_1_grouped$Global_Sales))

        Shapiro-Wilk normality test

data:  (sales_1_grouped$Global_Sales)
W = 0.70955, p-value < 2.2e-16
```

The low p-values (<0.05) showing the data is not normally distributed and a rejection of the H0 hypothesis of normal distribution.

Finally, Skewness and Kurtosis tests showed;

```
> # EU
> skewness(sales_1_grouped$EU_Sales)
[1] 2.886029
> kurtosis(sales_1_grouped$EU_Sales)
[1] 16.22554
> # NA
> skewness(sales_1_grouped$NA_Sales)
[1] 3.048198
> kurtosis(sales_1_grouped$NA_Sales)
[1] 15.6026
> # Global
> skewness(sales_1_grouped$Global_Sales)
[1] 3.066769
> kurtosis(sales_1_grouped$Global_Sales)
[1] 17.79072
```

All columns were skewed to the right (most of the data falling to the right of the peak, mean higher than medians and a long tail). Kurtosis showed the data was leptokurtic with high peaks (lots of data points concentrated round fewer values) with longer tails and more extreme values.

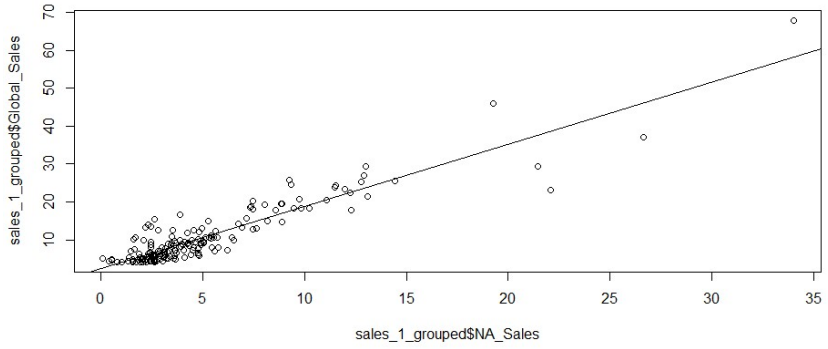
Implications for Turtle, Recommendations and Further Analysis

The sales data for Turtle is not normal, skewed and showed positive kurtosis. There is further work that could be done to aim for more normality, such as removal of outliers, transformation (sqrt, log) and checking the data to be completely free from error. Implications will depend on the impact on any predictive models of the data shape, I expect impact will be small as we build a predictive model using the three data sets that have similar shapes to each other.

What the relationship is between European, North American and Global sales.

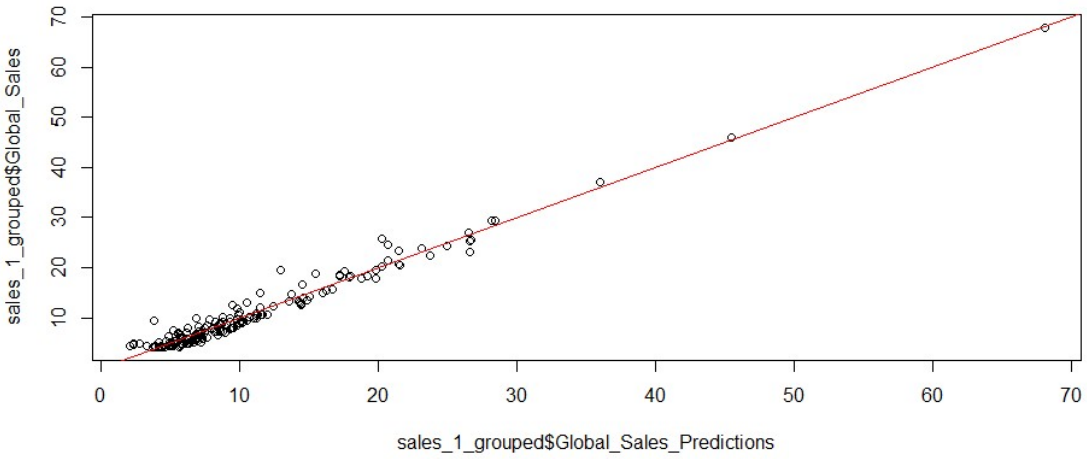
The ability to predict sales based on independent variables will be useful to Turtle games when forecasting and planning future releases.

Linear Regression models were created to look at EU v NA, EU v Global and NA v Global. EU sales were found to be a poor predictor of NA sales and a moderate predictor of Global Sales. NA Sales to Global sales was the best fit with an r^2 of 0.84, which I chose to visualise on a scatter plot with best line as it is easier to interpret than an r^2 in isolation;



For every £1m of NA sales, it could be predicted global sales would be £1.6m and that this would be reasonably accurate.

Multiple regression using EU and NA yielded better results with a r^2 of 0.97 showing that sales from these two regions explain the global sales very well. Given the good performance of the model, I added predictions for every product in the aggregated dataframe to compare to the observed value to check the accuracy, again a scatterplot was best to visualise prediction vs observed with a best fit line of 100% accuracy.



MAPE of the model is 12.8%, meaning the average prediction is wrong by 12.8% so this model can be considered good.

Implications for Turtle, Recommendations and Further Analysis

Using this model, Turtle games could reasonably reliably predict the global sales for a product based on the NA and EU sales. NA & EU sales make up about 75% of Global sales, so this is perhaps not surprising that they are a good predictor of Global Sales. Further work should be done to try and understand how EU and NA sales could be predicted, based on other variables such as genre, game type, game reviews, marketing spend, seasonality, regional sales etc. That kind of model would have more value as a predictor of sales and be useful for them in terms of forecasting future revenue.