Turtle Games – analysis of sales and customer reviews

The gaming industry currently has **2.9 billion gamers** and has **grown exponentially in the past decade** from **$59B to $160B in 2020,** and extortionate growth is expected to continue for the foreseeable future.

Therefore, Turtle Games needs to maximise this opportunity by developing and executing a comprehensive sales strategy.

The aim is to improve sales performance by generating insights through analysing:

- loyalty points
- customer groups
- customer reviews
- product sales
- data reliability, shape and correlation between sales regions

Furthermore, recommendations will be provided to enable further insights to be generated.

Analytical approach

The ***customer reviews*** dataset was loaded into Python to clean, explore and analyse.

Python packages utilised:

- numpy – array manipulation
- pandas – data manipulation
- seaborn – visualisation
- matplotlib – plotting
- statmodels – ordinary least squares modelling
- sklearn – clustering
- nltk – natural language toolkit

I explored the data types, data structure and searched for missing values to provide confidence in subsequent analysis. Unnecessary columns and duplicate observations were removed to aid performance and accuracy.

Loyalty points

Ordinary least squares (OLS) regression methodology was used to model the linear relationship between loyalty points and other independent variables, looking for a predictor of loyalty points.

I then reviewed the descriptive statistics for suitability of fit and generated a series of scatterplots.

Customer groups

In order to support the marketing strategy and categorise customers into groups, I used *k*-means clustering to unite customer remuneration and spending scores.

I used the 'Elbow' and 'Silhouette' methods and plotted the results to investigate and corroborate appropriateness of different cluster numbers of customers.

The plots made it easy to interpret the optimum cluster number. I also reviewed the customer numbers in each cluster, seeking relatively even splits.

Customer reviews

To steer future marketing campaigns, I applied natural language processing (NLP) to customer reviews to analyse sentiment.

I deployed a series of manipulations to standardise the format of the two review columns to facilitate accurate analysis:

- changed to lower-case
- removed punctuation, stopwords and non-alphanumeric characters
- tokenisation

The ***sales*** dataset was loaded into R to clean, explore and analyse.
R libraries utilised:

- tidyverse – facilitates tidy data
- moments – functions to calculate descriptive characteristics
- dplyr – dataframe manipulation

Product sales

I imported the sales dataset and reviewed the descriptive characteristics. I removed superfluous columns and verified the data types. I used scatterplots, histograms and boxplots to aid my exploratory data analysis.
A limitation of these plots is that they can oversimplify or overwhelm so I spent time enhancing them to expose insights:

- utilised colour to add a new dimension
- added lines of best fit with areas of 95% confidence
- amended axies for easier comparison

Data reliability, shape and correlation between sales regions

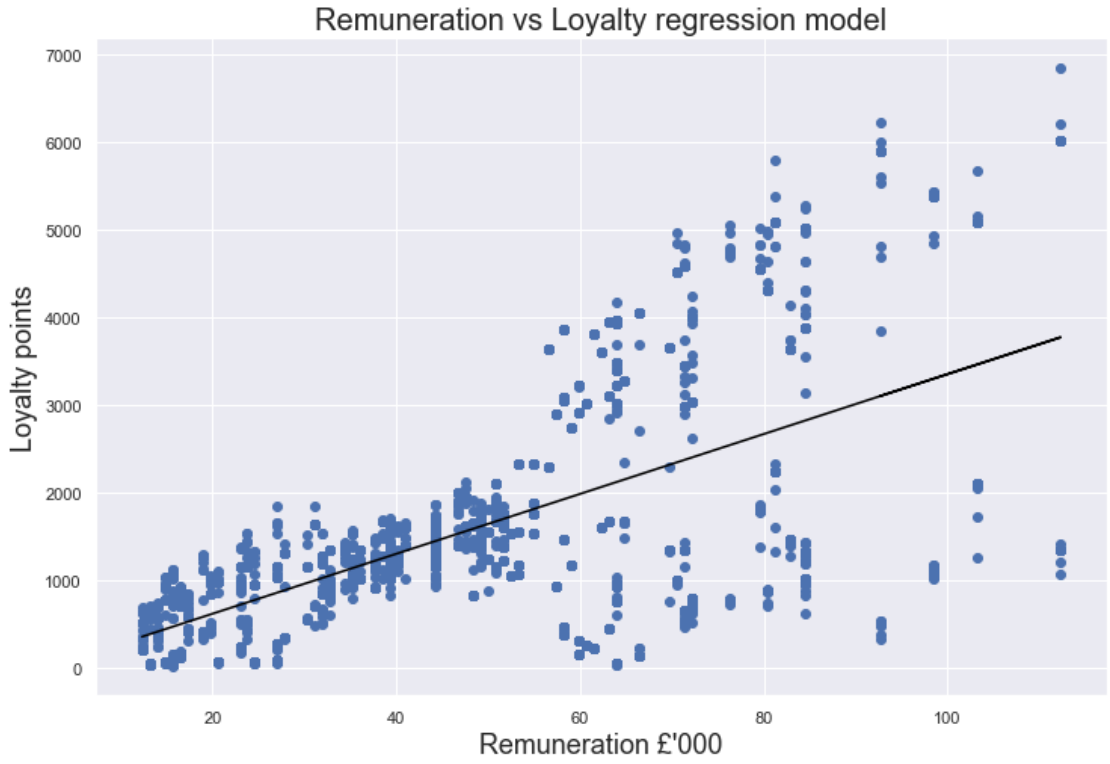
Confidence in the reliability of the data is critical so that the conclusions and recommendations can be trusted.
Therefore, I performed a series of tests:

- reviewed the summary statistics looking for spurious observations
- plotting against a normal distribution for comparison
- performing Shapiro-Wilk tests to test for normality
- used in-built functions to determine skewness and kurtosis
- measured the correlation coefficient between sales regions
- generated a profiling report – an excellent visual overview

Insights

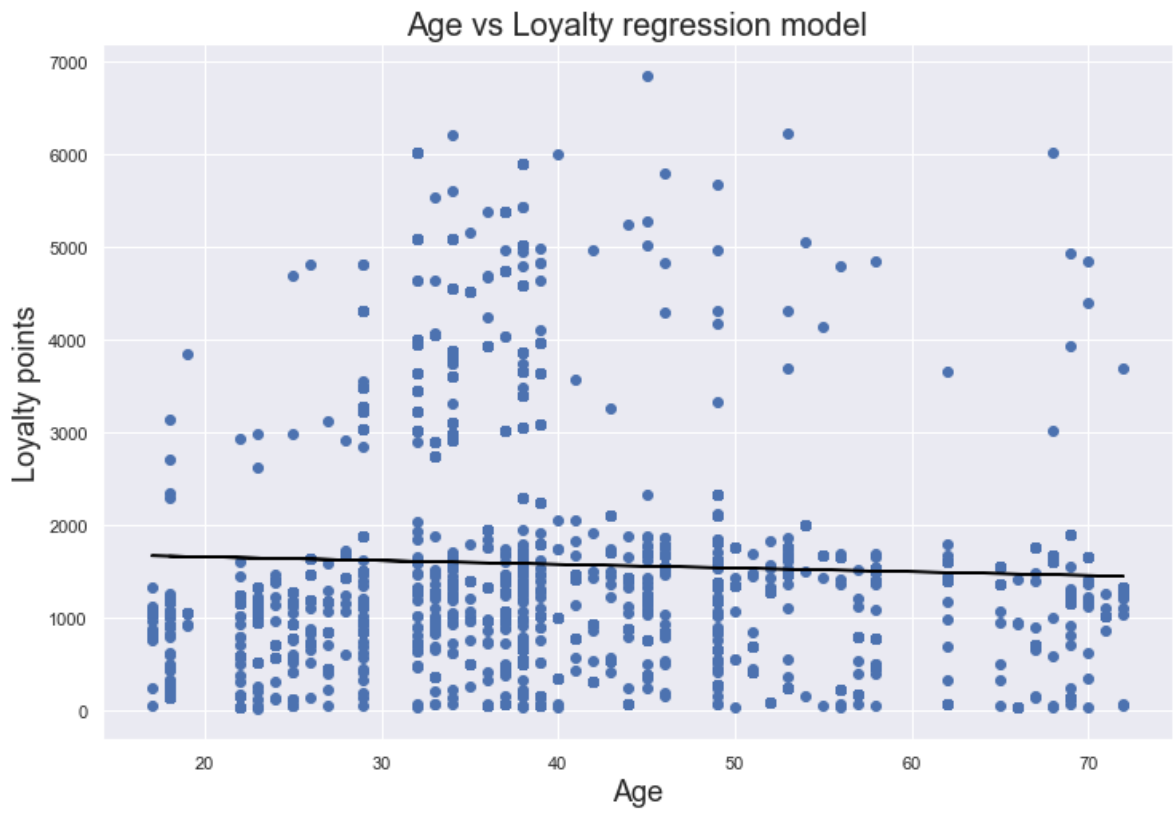
Are there any good indicators of accumulated loyalty points?

Spending score, remuneration and age were all used as independent variables to predict loyalty points and visualised as scatterplots.



Spending score and remuneration both demonstrated cone-shaped spreads suggesting heteroscedasticity. Therefore OLS regression modelling cannot be relied upon and the p-values are untrustworthy.

It can be concluded that spending score and remuneration are not good linear indicators of loyalty points.

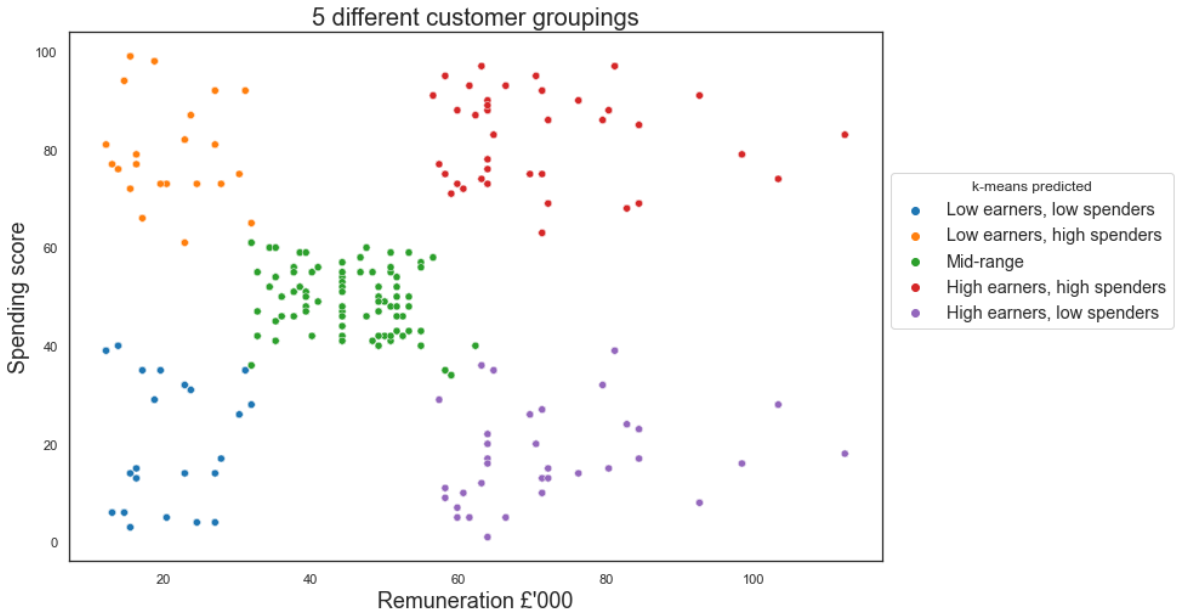


No correlation can be observed between the two variables.

Therefore, using Age and linear regression to model loyalty points is inappropriate.

How can customers be grouped to aid targeted marketing campaigns?

Elbow and Silhouette methods indicated that **five** customer groups is optimum.



Customers could be categorised as:

- Low earners, low spenders - 13.6% of customers
- Low earners, high spenders - 13.5% of customers
    - spending a disproportionate amount with turtle games
- Mid-range - 39% customers form the largest, most dense group
    - Curiously, the range is quite distinct: ~£30-58k remuneration and ~40-60 spending_score with little overlap with other clusters.
- High earners, high spenders - 17.8% of customers
    - can afford to spend a lot with turtle games and do
- High earners, low spenders - 16.5% of customers
    - can afford to spend a lot with turtle games, but don't

The visualisation and data itself do not provide any insight into trends and how the customers and groups are evolving over time.

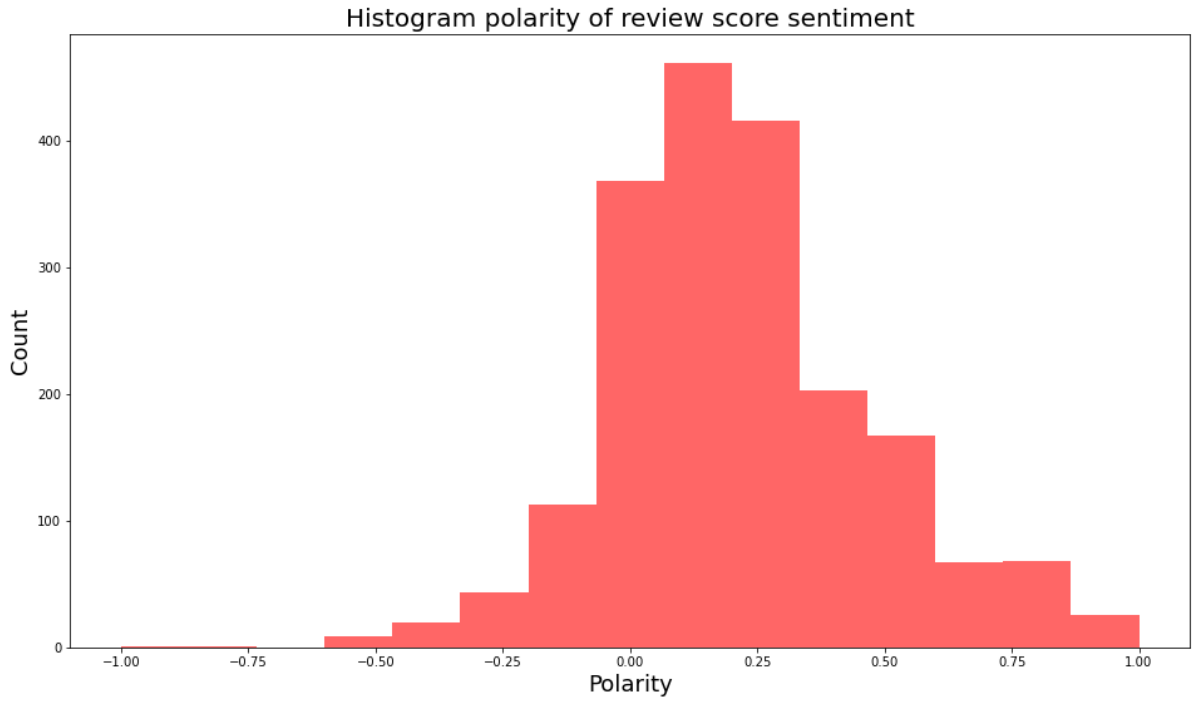
Five clusters is optimum at the moment but this needs to be monitored as it could evolve over time.

How can customer reviews be used to inform marketing campaigns?

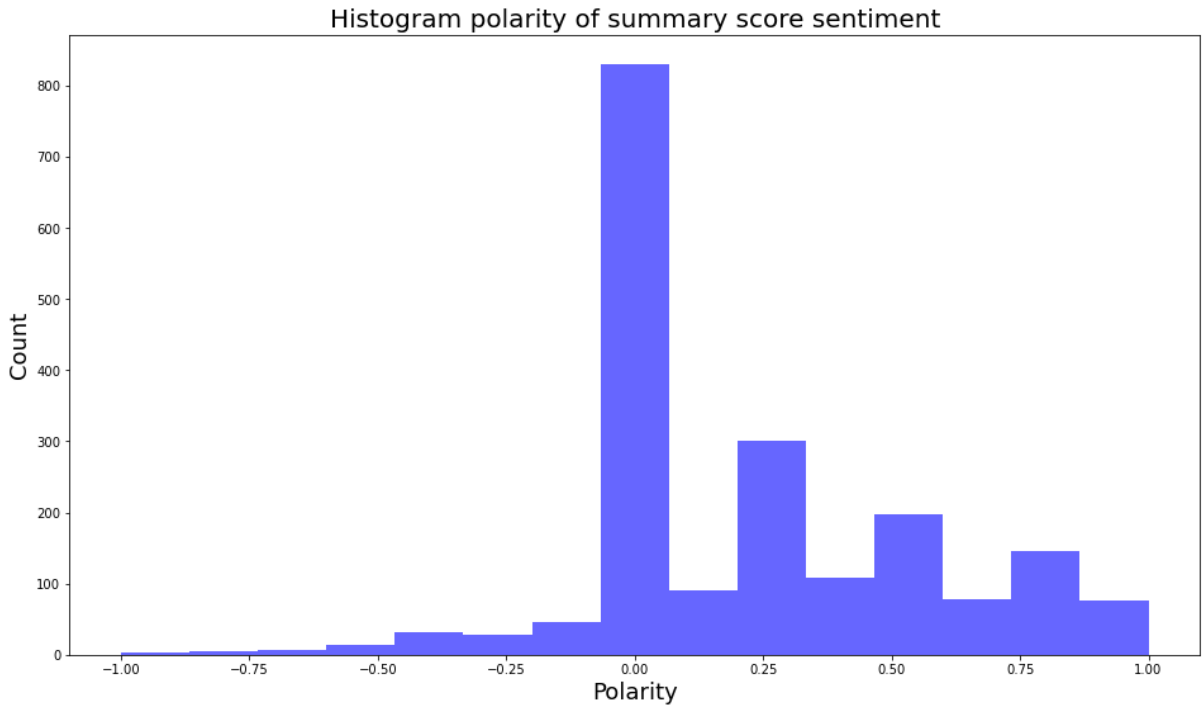
NLP was able to identify that the positive words: great, fun, like, good, are all in the top 15 most-frequent words.

When summary reviews were analysed, it can be concluded that “four stars” and “five stars” were prevalent.

Histograms were used to visualise sentiment scores:



Overall review sentiment is clearly very positive, with no reviews with a negative sentiment below -0.6.



The summary review seems to have a dampening effect as the highest ‘count’ around zero is much greater.

It appears the lower word count of the summary reviews results in a less precise calculation and a grouping around positive 0.25 intervals, e.g. 0, 0.25, 0.5.

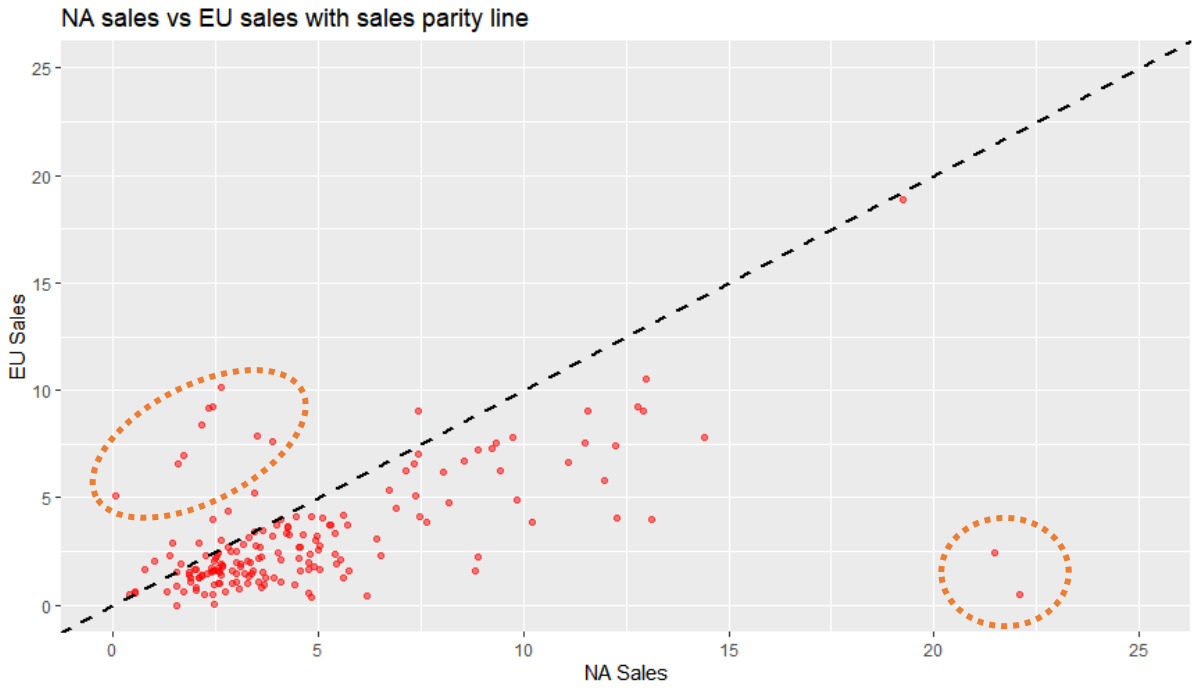
```
Average Customer review length:  57
Average Summary review length:  5
```

We should expect the sentiment score of the two review types to be strongly aligned, but this is not true:

This suggests that the summary reviews cannot be trusted for their accuracy.

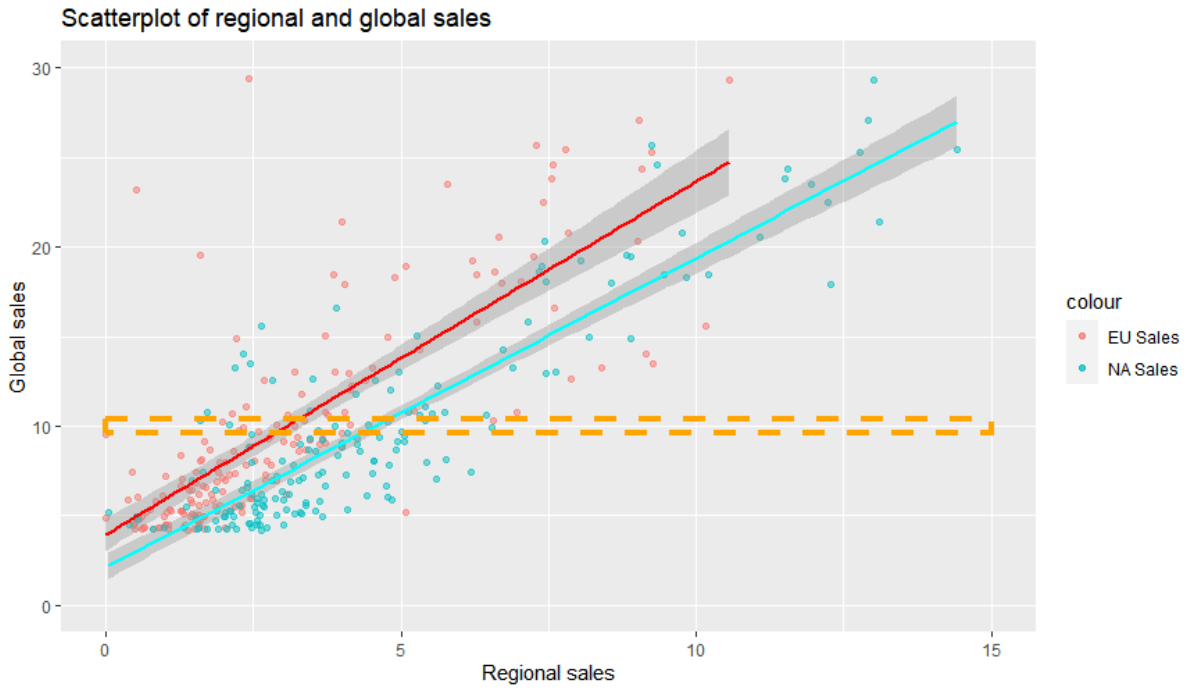
To guide marketing campaigns, further analysis could utilise NLP to identify themes in subsetted data, e.g. by genre, and align to key words and review sentiment.

How do sales differ by product?



You can see that the overwhelming majority of products sell more in NA than EU (because they are below the dotted parity line).

There are a few notable extreme outliers, highlighted above, where product sales in one region have been much higher than the other.

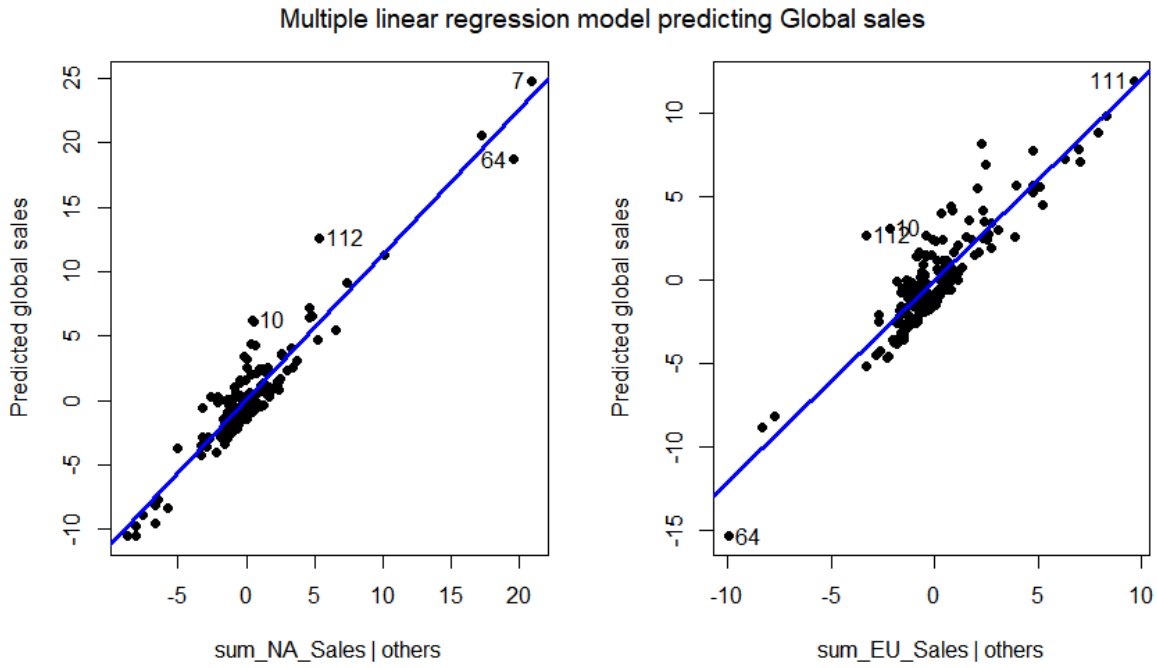


The red and blue lines represent linear regression models, predicting how regional sales correspond to global sales. There is medium positive correlation but with substantial variation across products.

A key insight is highlighted by the orange section. It demonstrates that if global sales are £10m, then NA is likely to contribute ~£4m and EU ~£3m, i.e. less than NA.

Due to global sales literally being the sum of all regional sales, a **multiple linear regression** model should prove effective.

The associated t-values of NA and EU sales are both highly significant in predicting global sales. Adjusted r^2 of 0.966 demonstrates near perfect correlation.



Both scatterplots above demonstrate a strong positive association between the predictor variable and the response variable (global sales).

Therefore, as expected, NA and EU sales are an excellent predictor of global sales.

If further analysis could determine a strong correlation between non-sales variable(s) and sales, then this would be hugely beneficial to predict sales, e.g. platform and genre.

Certain games have been released more than once. It would be interesting to understand the relationship between number of releases and cumulative sales.

How reliable is the data?

	NA Sales	EU Sales	Global Sales
Measures of central tendency			
Minimum	0.00	0.00	0.00
Median	1.82	1.17	4.32
Mean	2.52	1.64	5.34
Maximum	34.02	23.80	67.85
Shapiro-Wilk normality test			
p-value	<0.001	<0.001	<0.001
Measures of shape			
Skewness	3.05	2.89	3.07
Kurtosis	15.60	16.23	17.79
Correlation coefficients			
NA sales	1	0.62	0.92
EU sales	0.62	1	0.85
Global sales	0.92	0.85	1

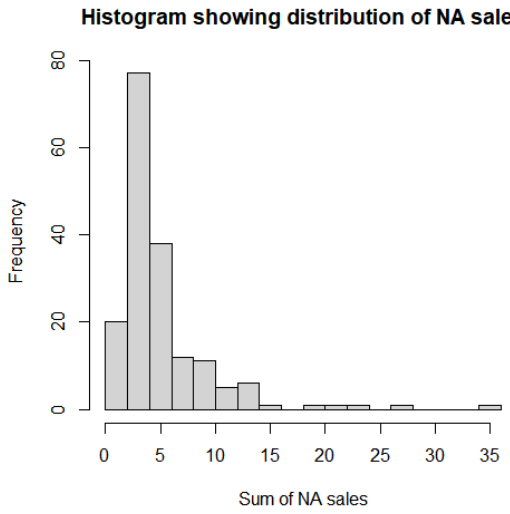
The three sets of sales share similar distribution shapes, with a considerable right skew and a non-normal distribution, as verified when plotted against a normal distribution on a ‘QQ-plot’ and by the Shapiro-Wilk test significantly being <0.05.

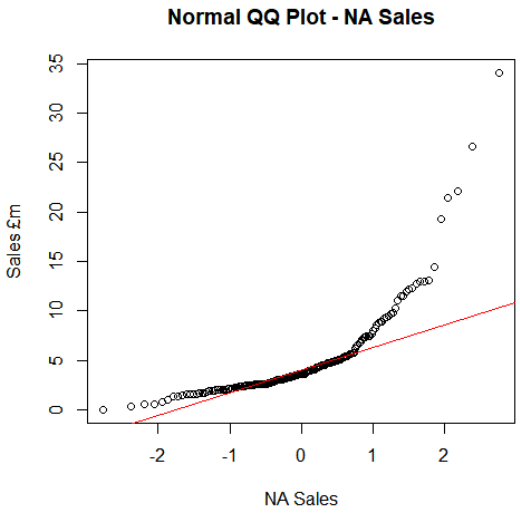
This distribution shape is to be expected as maximising sales is the priority.

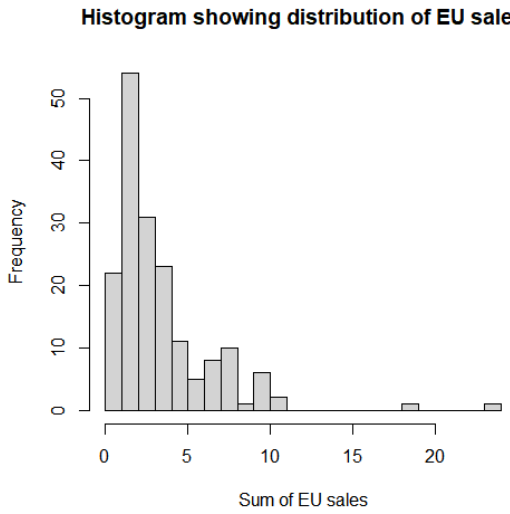
Kurtosis >3 signifies a high peaked distribution corresponding to most games selling under £20m globally.

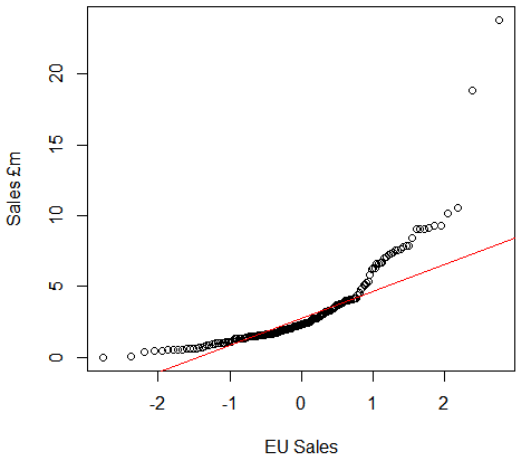
NA sales correlate most with global sales at 92%, which aligns with NA sales being the largest contributor of regional sales.

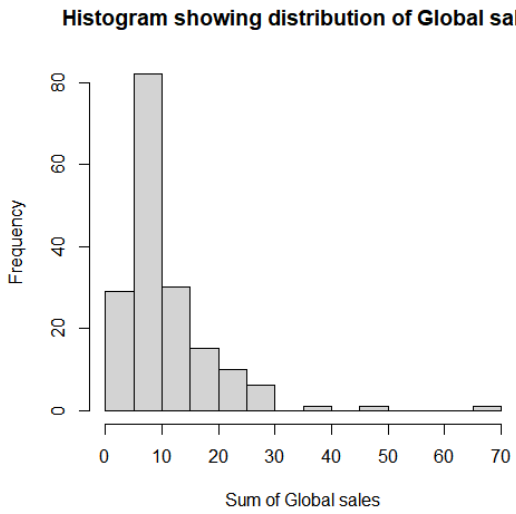
Overall, we can take great confidence that the correlation between specified variables is a true reflection and can therefore have certainty around our conclusions.

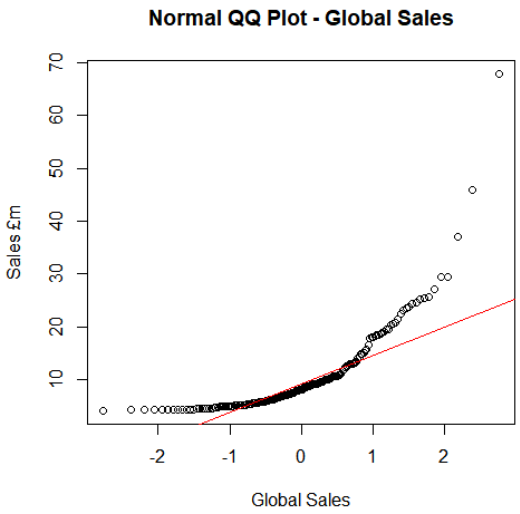












The red lines indicate a normal distribution for comparison.

Recommendations

Loyalty points should be closely aligned to spending_score, as both are dependent upon customer spending. Review spending_score process for suitability as it should be able to predict loyalty points, enabling better decision-making by stakeholders.

Consider rewarding customer who reach certain **loyalty** thresholds. This should encourage brand loyalty.

Increase engagement in 'high earners, low spenders' **customer group**. Targeted communication of new releases, offers and bundles.

Reward loyalty of 'high earners, high spenders'. Invites to launch events and marketing questionnaires should prioritise this group.

Monitor mid-range customers who, depending on circumstances, could move into one of the other groups.

Be sensitive of 'low earning' groups. Ethical decision-making is key here. Turtle Games do not want to incur reputational damage via mis-targeted marketing.

Correlate **customer reviews** with **product sales** and analyse. Customers will be willing to pay higher amounts for well-reviewed games.

Correlate **customer reviews** with **regional sales** and analyse. Customers will have different game preferences depending on where they live. Marketing strategies should reflect this.

The **summary review** currently offers no value. Consider replacing with a quantitative scale, e.g. 1-5 stars.

Corroborate **reviews** with external sources to generate further insight.

Include production and overhead costs in **product profitability** evaluations. Turtle Games will have profit margin targets and a portfolio of products with different margins that need to be managed.

Align **reviews to sales** so that pricing can be optimised, customers will be willing to pay a higher price for products that are comparable to well-reviewed products.

Perform this analysis with the absent **region**(s). 22% of global sales are not assigned to a region. This insight will align macro sales strategy.