DA301 Report: Predicting Future Outcomes

Introduction

Turtle Games (TG) seeks to set pricing, determine product segments, identify customer sentiment and predict sales using data.

Objective

Use data provided to generate insights, supporting visuals and analysis that will achieve TG's objectives.

Assumptions

1. Audience is familiar with statistical concepts but uninterested in the underlying coding.
2. Audience is seeking actionable insights.
3. TG's technologists are familiar with technologies deployed here.

Analysis Approach

After an exploratory data analysis (EDA), I found it absent any qualitative issues. I used Python for analysing the customer profile and social media data and R for the sales data, in keeping with TG's preferences. I used libraries[1] that are available in Python and R for the analytical work to optimise effectiveness and efficiency. I selected libraries that are the most diffused within their respective domains. This ensures that this work is accessible and replicable should TG wish to do so. Wherever possible, I used two alternative methods to interrogate the same data to ensure robustness of analysis.

In a data driven project with an objective to generate actionable insights, it is imperative to communicate as effectively as possible whilst delivering insights into a complex topic. Emphasis was placed on generating stunning visuals[2] using specialist tools that are available within Python and R. Key insights are highlighted in **bold** throughout whilst action points are summarised at the end.

Customer Analysis

I used correlation, linear and multiple linear regression models to explore relationships between the variables within TG's customer data. Only a few variables displayed any strength of connectivity; spending score, age and loyalty points stand out as displayed in the correlation matrix[3] overleaf. Regression analysis is required to meet TG's objective of being able to predict loyalty points. Simple linear regression modelling suggested an almost 0 predictive quality between loyalty points and age. The same relationship with remuneration was stronger; nearly 38% of changes in loyalty points could be determined by changes in remuneration. The strongest relationship was with spending score which explained c.45% of changes in loyalty points.

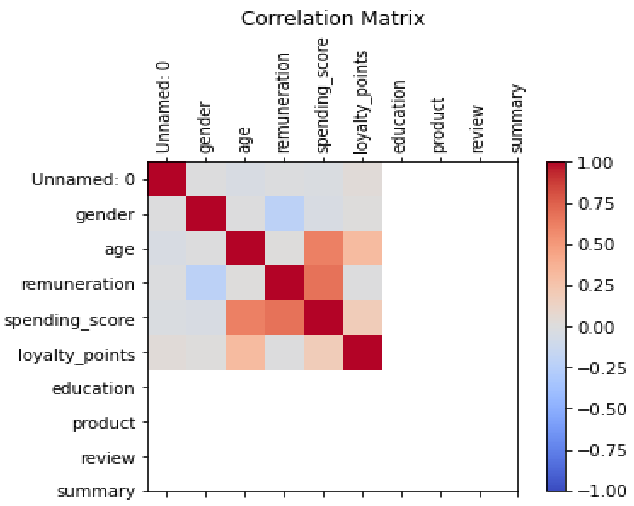
It intuitively follows that customers who spend more likely receive more loyalty points and are better remunerated. It made sense to run a multiple linear regression model incorporating all the variables together to explore these inter-linkages further. This **model was able to accurately predict changes to loyalty points 84% of the time and it incorporated the other**

[1] For more information on libraries in coding and why they are used, refer to https://careerfoundry.com/en/blog/web-development/programming-library-guide/#what-is-a-programming-library

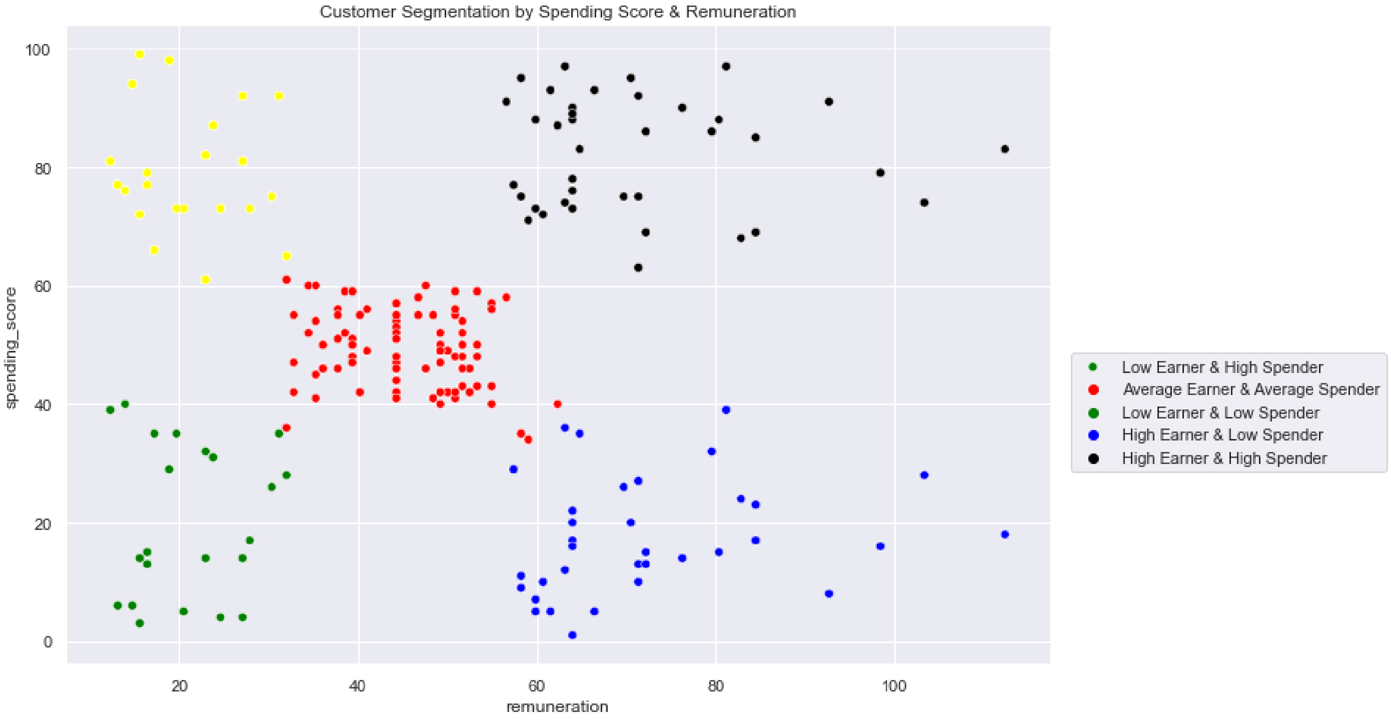
[2] For reasons why visualisations matter, refer to https://www.score.org/blog/3-reasons-liven-your-marketing-visual-content#:~:text=Humans%20respond%20to%20and%20process,to%20the%20brain%20is%20visual.

[3] Correlation Matrices depict the degree of connectivity between variables.

variables together in a manner that is statistically significant. It is important to point out that this does not necessarily establish causation.



Furthermore, I identified groups within the customer base that could be used to target specific market segments by TG's marketing department. I used K-means[4] [5]clustering for this. Two alternative methods both confirmed that **TG's customers are best sub-divided into 5 distinct segments** as depicted below.

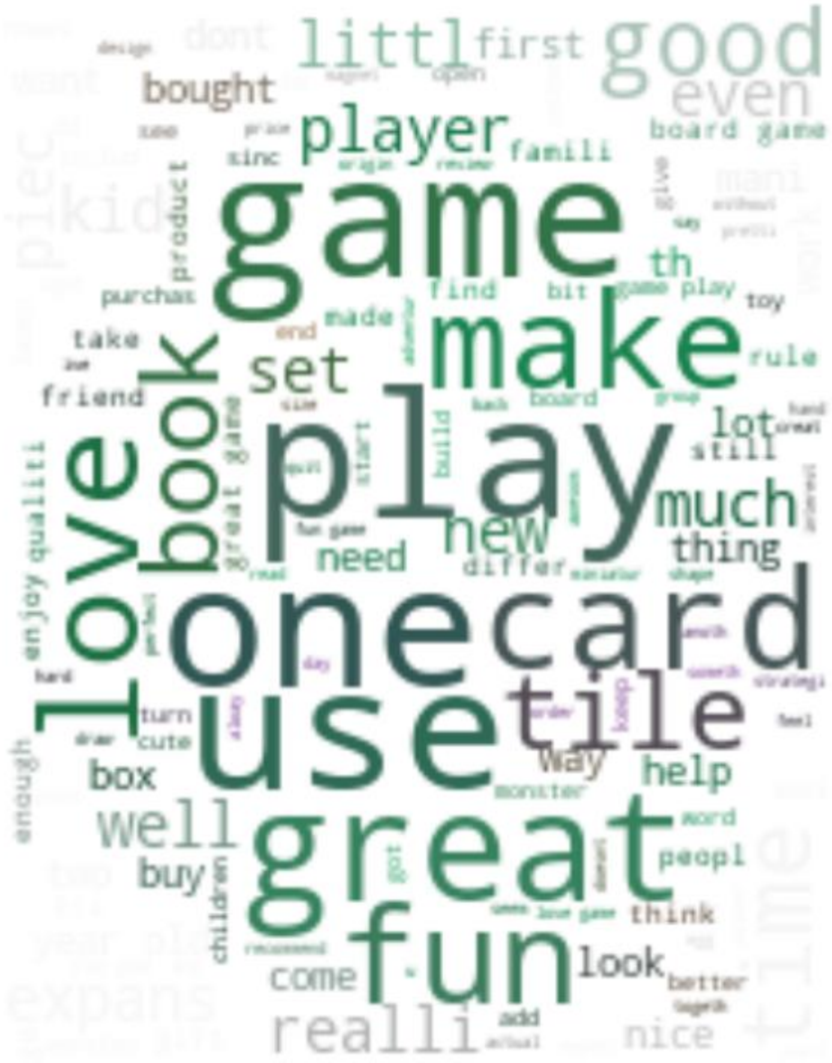


[4] For fuller explanation of how this works, refer to https://www.analyticsvidhya.com/blog/2020/10/a-simple-explanation-of-k-means-clustering/

[5] It is a statistical method to identify groups within a larger population that are related through shared characteristics.

Sentiment Analysis

Customer reviews downloaded from TG’s website were used to steer the marketing department’s approach to future campaigns. After data wrangling, I created the below Word Cloud[6] to summarise the most common words used in TG’s customers reviews[7].



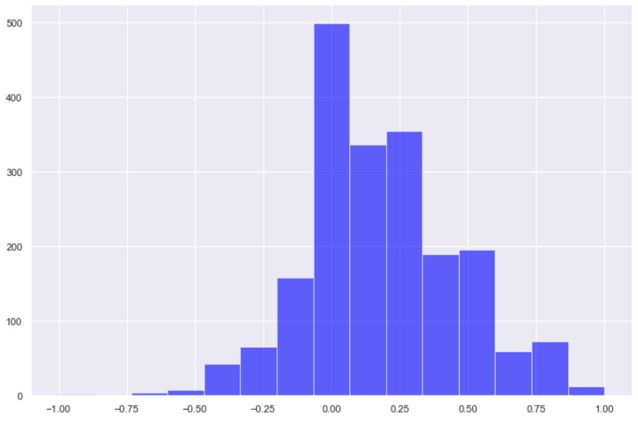
From this, I can say that **customers are broadly positive about the products they buy from TG.** There are no negative words whilst words such as fun, great, love and good are prominent and all convey positive sentiments.

Further, I investigated the sentiments expressed in TG’s customer reviews. Using two libraries, I got somewhat different results[8]. I found that most **customer reviews were neutral although there were more positive reviews than negative reviews**; displayed in the charts overleaf.

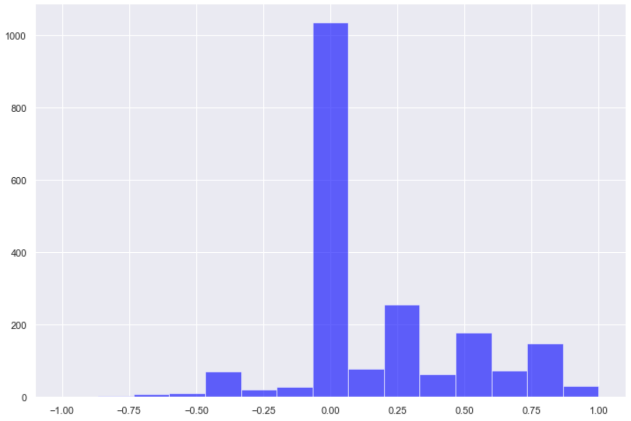
[6] For more details on Word Clouds and why and how they work, refer to https://www.participoll.com/what-is-a-word-cloud-and-why-should-i-use-one/

[7] World Clouds are great to display the most frequent words and to quickly identify the relative sentiment of such words through their meaning.

[8] As one library was trained more to look for emojis and social media content, I landed on the other as being more suited for analysing the more text heavy content of the customer reviews and summaries presented here.

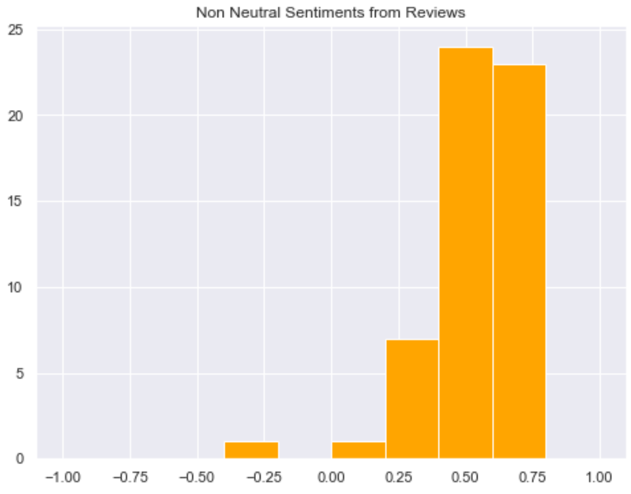


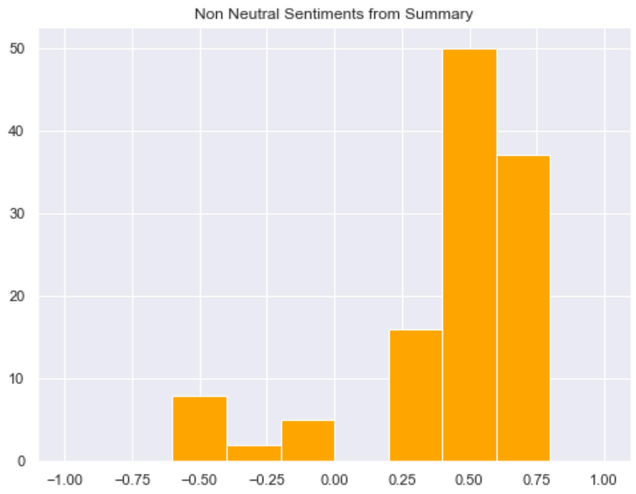
Histogram of sentiment score polarity of reviews



Histogram of sentiment score polarity of summary

I noted that the strongest of the negative sentiments were less polarised than the best of the positive sentiments. We can conclude that **the unhappiest of TG's customers feel less negatively about TG than the positivity felt by their happiest customers.** The charts below show that there are many more strongly positive reviews in both absolute numbers and strength of sentiment.

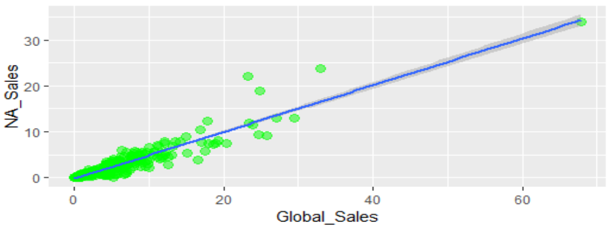


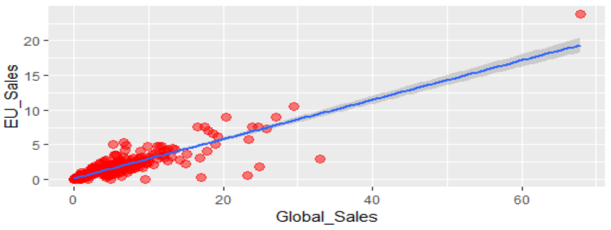


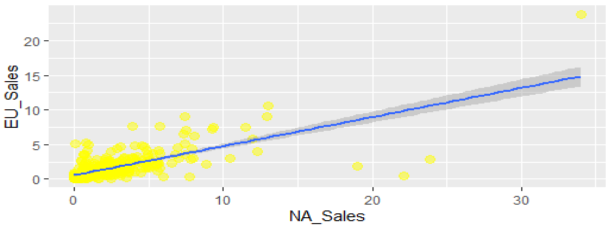
The top 20 positive and negative reviews are identified and presented in the accompanying Jupyter Notebook along with the 15 most common words used in online product reviews. They support the Word Cloud and Sentiment Analysis presented thus far.

Sales Data Analysis

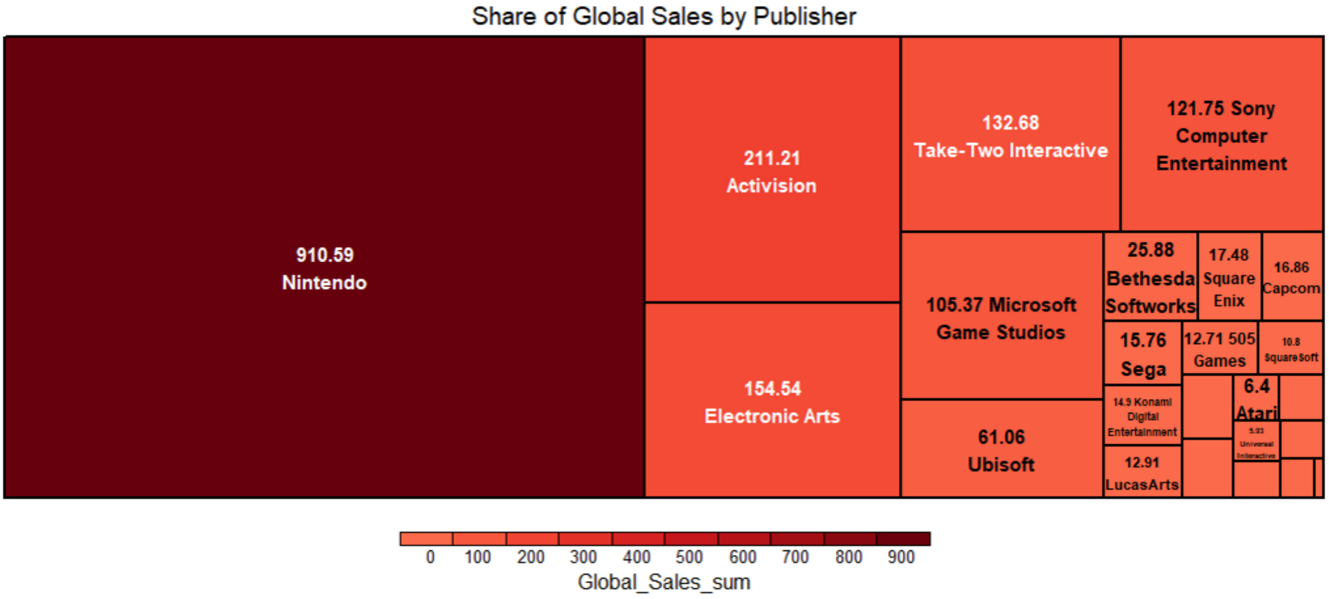
Due to TG's preference, analysis related to sales was performed in R. EDA displayed robust data quality and the likelihood of strong linear relationships between sales from the regions and Total Global Sales (GS) as shown below.



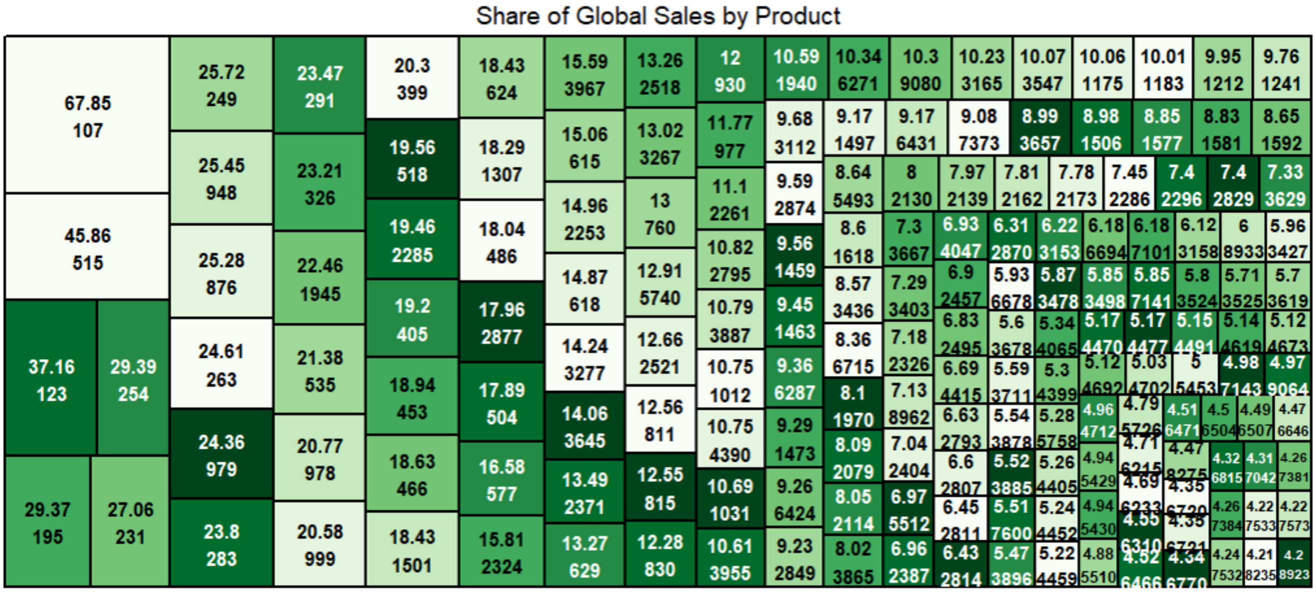




I noted that **the sales data is incomplete**; the data provided from the regions do not add up to GS. During EDA, I discovered very strong skew in the data. With further analysis, I confirmed that the sales data is not normally distributed. Subsequent analysis showed that the **revenue from sales is very concentrated amongst a handful of the products and publishers stocked** by TG. However, tests such as model errors **confirmed that the data was suitable for regression modelling**.

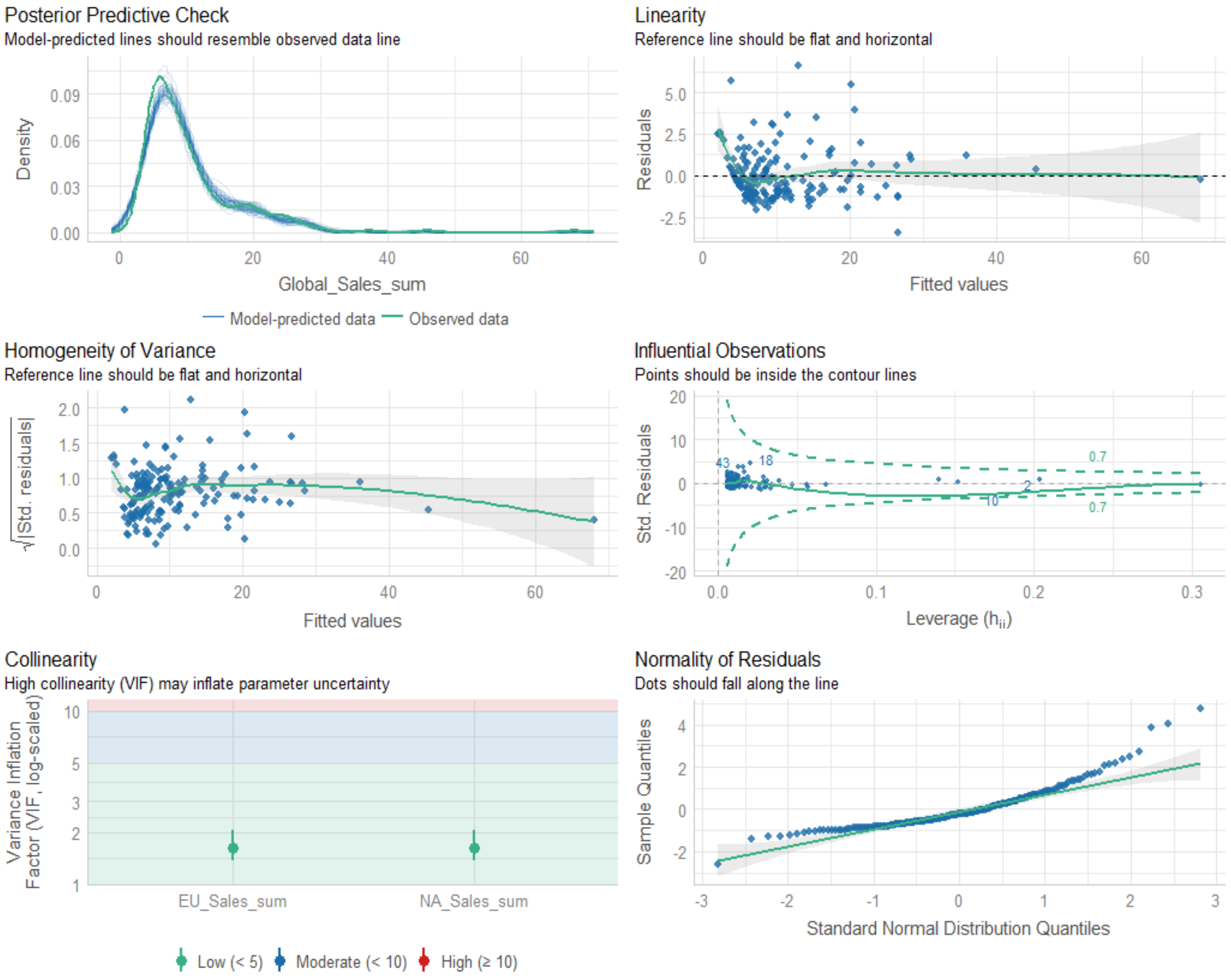


Of the 175 products sold by TG, 25% account for 50% of the sales revenues globally; similar concentrations apply to regional data. Global concentrations are summarised for publishers above and products below.



Linear regression models between the regional and global sales data produced mixed results for predictive strength and with very poor accuracy. Data transformation techniques failed to improve results.

I decided to run a multiple linear regression model which produced very promising results with minimal predictive errors. This aligns to the fact that for the most part, the three variables are extensions of each other. **The model cannot in the real world predict future sales.** It is useful, however, to determine the likely impact of changes to one of the variables to the remaining. A summary of the outputs - below - from tests carried out to ensure the model's accuracy confirmed that the model was acceptable. A table was created to display predicted versus observed data for ease of user reference.



Finally, I constructed similar models against groupings by Genre, Platform and Publisher. The latter produced the best results although high collinearity somewhat limit the model's usefulness.

Recommended Actions

1. Target Higher Earners through marketing campaigns; they outspend others.
2. Explore how best to shift the large number of neutral customers into stronger sentiments.
3. Explore the possibility of trimming the range of products and publishers stocked, their relationship with customer sentiment and spending
4. Sales department should link the data gathering process with their colleagues in marketing.
5. Provide further data for analysis of customer's spending patterns versus their sentiment, set pricing and predict sales.

1189 words excluding titles, headers and references.