Advanced Analytics for Organisational Impact: Assignment

Predicting Future Outcomes

Rafa Abdelazim
LSE career accelerator

Contents:

Background:

Turtle games is a game manufacturer and retailer with a global customer base. In this report we will use predictive data analysis to improve overall sales performance by utilising customer trends.

Business problems aimed to be solved:

- how do customers accumulate loyalty points
- how groups within the customer base can be used to target specific market segments
- how social data (e.g. customer reviews) can be used to inform marketing campaigns
- the impact that each product has on sales
- how reliable the data is (e.g. normal distribution, skewness, or kurtosis)
- what the relationship(s) is/are (if any) between North American, European, and global sales?

Analytical approach:

Python:

1. **Data Import and Initial Exploration:** In Python, I began by launching Jupyter Notebook and importing essential libraries, such as Pandas, NumPy, Matplotlib, and Statsmodels. I loaded the 'turtle_reviews.csv' dataset, performed data checks, removed redundant columns ('language' and 'platform'), and renamed them for clarity. I saved a clean copy and re-imported it for verification. Linear regression was conducted using Statsmodels to explore potential relationships between 'loyalty points,' 'age,' 'customer pay,' and customer spending.'

2. **Exploratory Data Analysis:** I used linear regression models using stasmodel to understand the effect of age, pay and spending on loyalty points.

3. **K-Means Clustering:** To determine the optimal number of clusters (k), I employed the Silhouette and Elbow methods. Multiple k values were evaluated, and the choice was justified. I used Matplotlib for visualization.

4. **Data Preparation for NLP:** For Natural Language Processing (NLP) tasks, I pre-processed text data in 'review' and 'summary' columns. Libraries such as NLTK and WordCloud assisted in tokenization and word cloud creation. Word frequency distributions and sentiment polarity were calculated.

5. **Sentiment Analysis:** I plotted sentiment histograms and analysed sentiment scores for 'review' and 'summary' columns, using libraries like Matplotlib and TextBlob. The top 20 positive and negative reviews and summaries based on sentiment polarity were printed.

R:

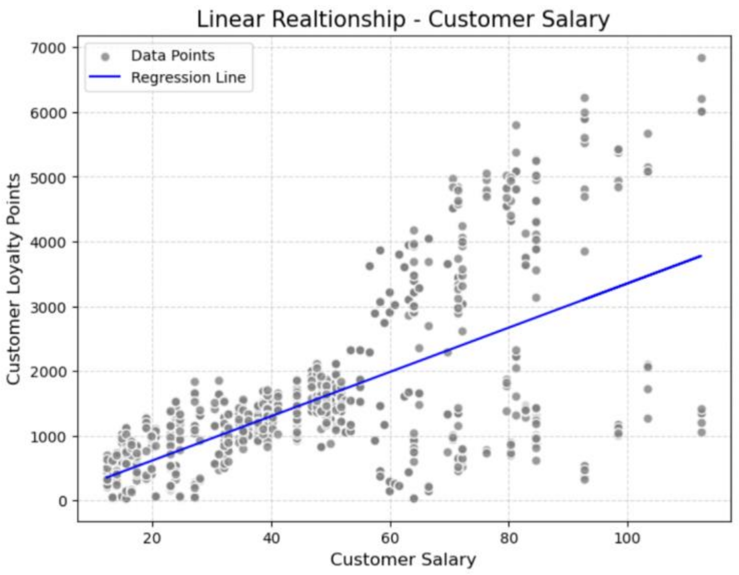
1. **Data Import and Exploration:** I loaded required libraries like 'tidyverse,' and ‘drplyr’, I then imported 'turtle_sales.csv,' and explored the data. I removed redundant columns, summarized the dataset, and created scatterplots, histograms, and boxplots to gain business insights.

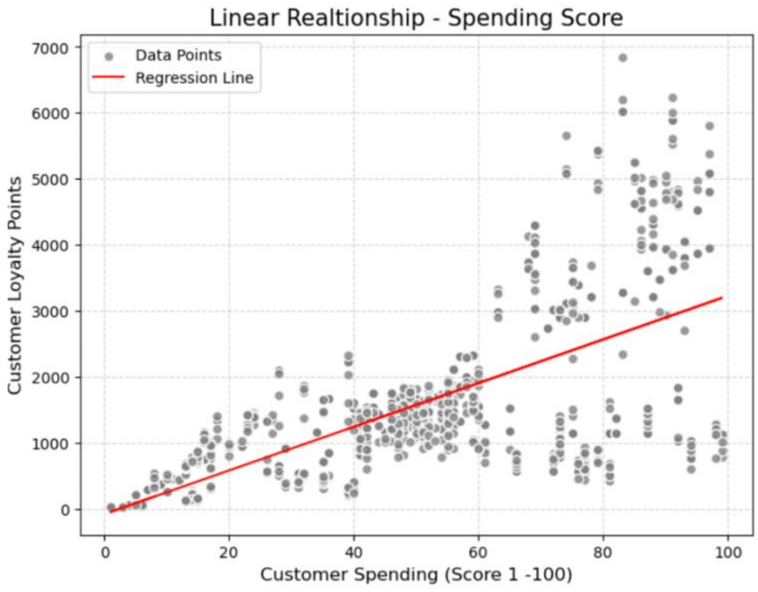
2. **Descriptive Statistics and Distribution:** I calculated min, max, and mean values and summarized the data frame. I then aggregated the data using group_by and created visualisations to gain insights. I then checked for normality using Q-Q plots and a Shapiro-Wilk test. Skewness and kurtosis were assessed, and correlations between sales columns were explored.

3. **Linear Regression Analysis**: I conducted simple and multiple linear regression, viewed outputs, and visualized results. Predictions were made for global sales based on provided values, comparing them to observed data

Visualisation and insights:

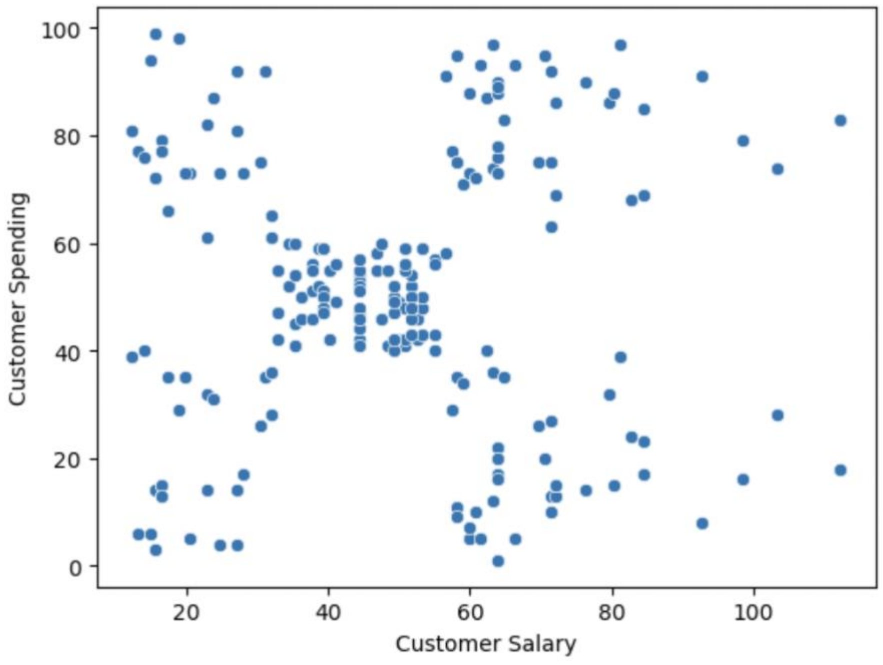
1. How customers accumulate loyalty points:



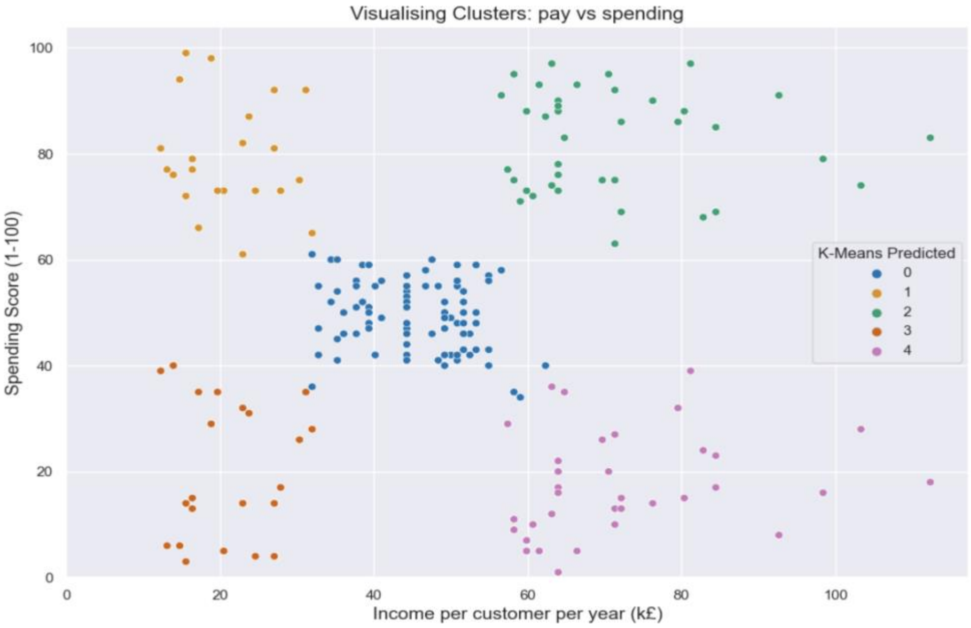


Linear models demonstrate customer salary and customer spending have a positive linear relationship to customers accumulating loyalty points, however there is no correlation between customer spending and customer salary therefore reduces the chance of multiple collinearities in an MLR model.

In both linear models we can see as the dependant and independent variables increase the data spreads further away from the regression line. This suggests that there may be other factors influencing the relationship



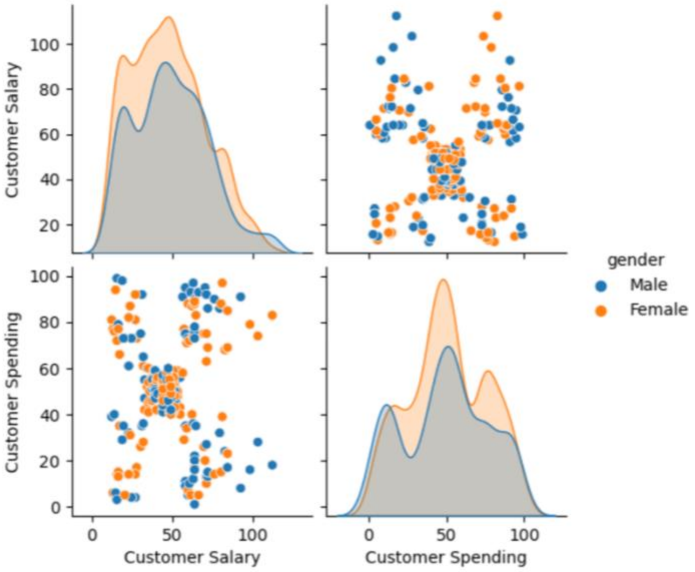
2. How groups within the customer base can be used to target specific market segments



Using the elbow method and silhouette method we determined that there are 5 cluster groups within customers.
These groups can be described as:

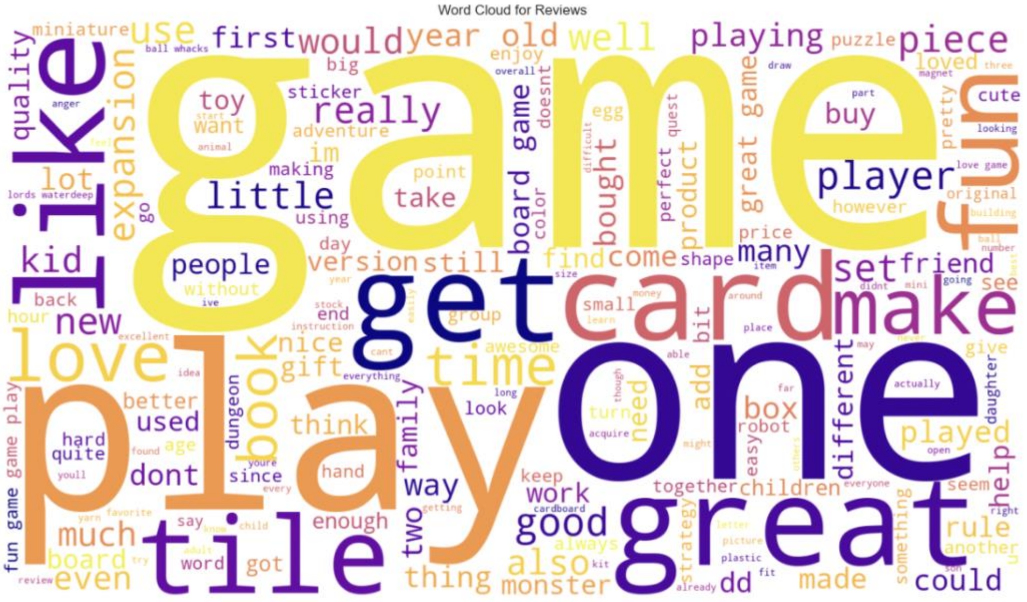
- 0 - Customers with average income (£40 -60k) and medium spending scores (40-60)
- 1 - Customer with low income (less than £30k) and high spending scores (>60)
- 2 – Customers with high incomes (>£60k) and high spending score (>60)
- 3 - Customers with low income (less than £30k) and low spending score (<40)
- 4 - Customers with high income (>£60k) and low spending score (<40)

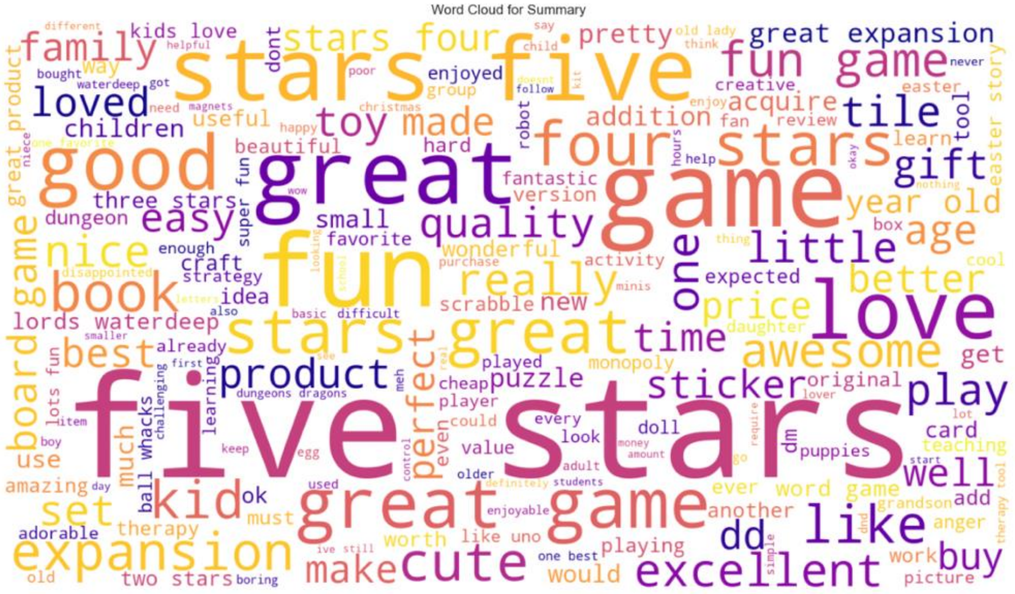
This pair-plot also indicates females are higher earners and spenders than males.



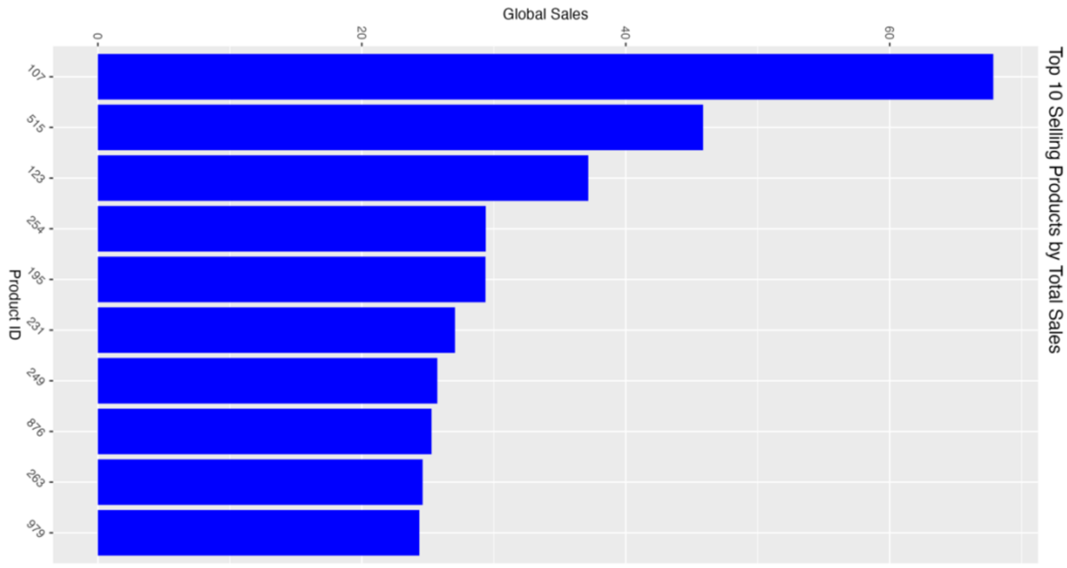
3. How social data can be used to inform marketing campaigns

The reviews and summary wordcloud show positive words like ‘five stars’, ‘fun’ ‘like’ and ‘great’, which overall highlights a positive sentiment towards turtle games. It also shows ‘game’ as most used in reviews which shows that games (video or board games) are more popular products compared to toys and books.





4. Impact each product has on sales

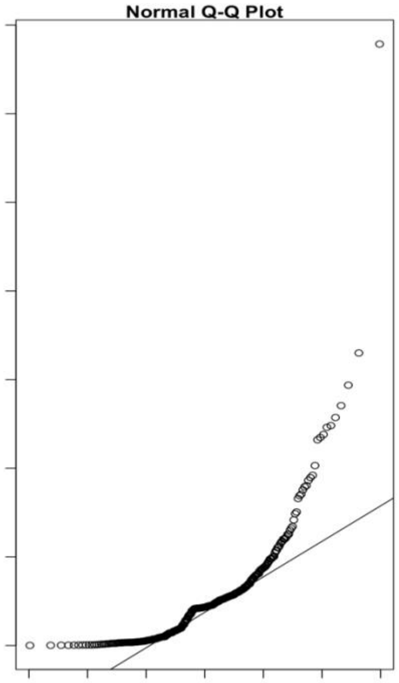


This bar plot shows the top 10 selling products. With product “107” being relatively higher than the second selling product “515”, making around £67m of total global sales (3.6%), while product 515 makes around £45m (2.4%) and product “123” making around £38m (2%). With the rest of the top selling ranging from around £25-30m.

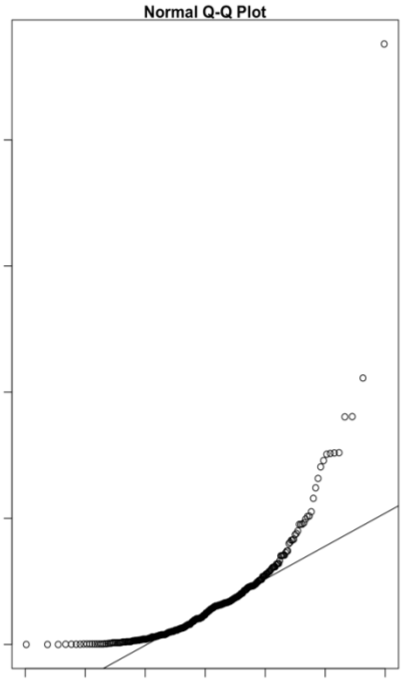
5. How reliable the data is

All sales do not appear to follow normal distribution, with skewness >1 indicating positively skewed to the right and leptokurtic distribution (kurtosis > 3) which shows a heavy-tailed distribution and suggests extreme values or outliers.

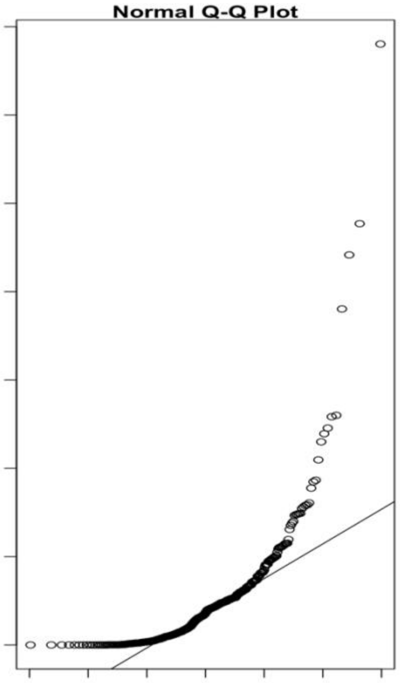
Global



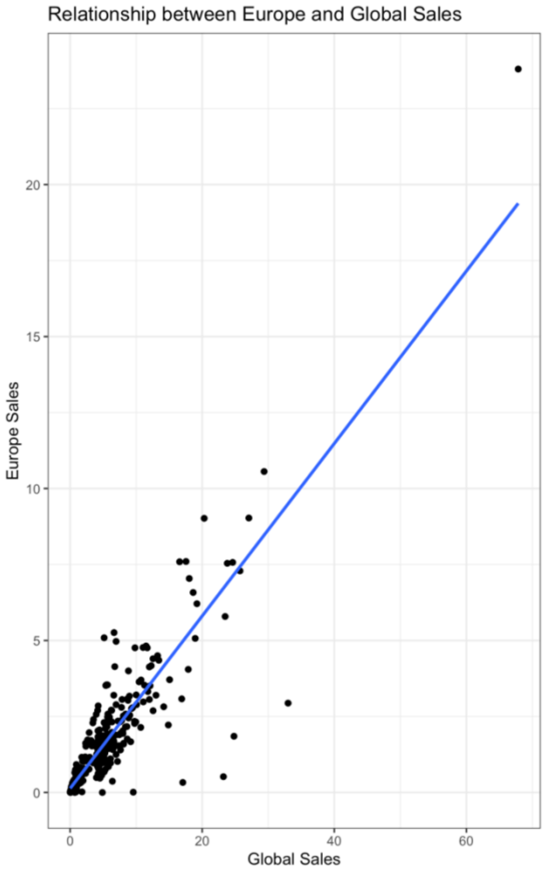
North America

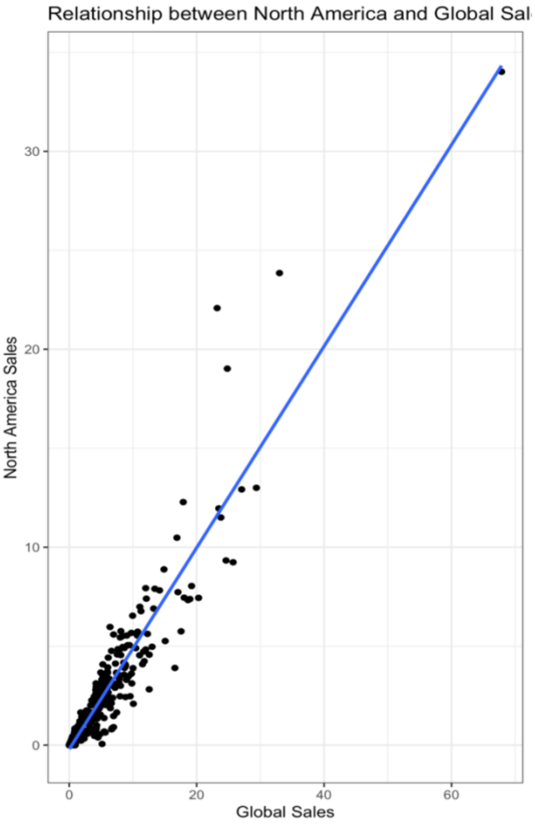


Europe

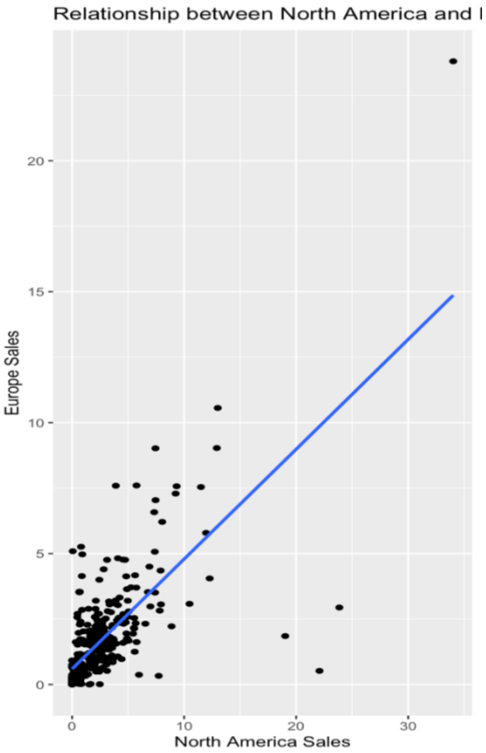


6. What the relationship(s) is/are (if any) between North American, European, and global sales?





There is a positive linear relationship between Europe and North America sales and Global Sales, with correlation coefficient of 0.94 (NA) and 0.88 (EU) which is very close to 1, indicating an almost perfect correlation.



However, the linear relationship is not as strong between Europe and North America, but still portrays a positive linear relationship with a correlation coefficient of 0.7.

Patterns and Predictions:

- Customer spending and salary influences the accumulation of loyalty points.
- Linear model suggests that spending contributes to 45% of variance in loyalty points.
- Coefficient value suggests that if spending score goes up by 1, loyalty points increase by 33.
- Linear model suggests that salary contributes to 38% of variance in loyalty points.
- Coefficient value suggests that if salary score goes up by £1k, loyalty points increase by 34.
- There are 5 clusters when analysing customer pay and spending relationship, marketing strategies can be tailored depending on the group.
- Games (video or board games) are the most popular products sold by turtle games.
- Review sentiment score indicates positive polarity with most words being neutral or close (0-0.125).
- Summary sentiment score indicates slight positive polarity, but neutral comments are significantly high
- Further analysis to investigate reviews and feedback on specific products.
- Top 10 selling products out of 175 products make up to 18% of total Global sales (£337 million)
- Product “107”, “515” and “123” are the highest selling globally.
- Action, shooter and sports are the most popular genres.
- X360, PS3 and PC are the most popular platforms.
- Nintendo, Electronic Arts and Activision are the highest selling publishers
- North America contributes the most to Global sales
- North America and Europe make up 97% of Global sales, therefore can confidently predict Global sales.