Predicting future outcomes

Improving overall sales performance for Turtle Games by utilising customer trends.

LSE Data Analytics Career Accelerator
Course 3: Advanced Analytics for Organisational Impact

Jessica Krook

4 May 2023

Background

Turtle Games' strategy is to use customer trends to improve sales performance.

Turtle Games wants insight into:

- Loyalty point accumulation
- Customer groups to target
- Using reviews to inform marketing campaigns
- Impact of products on sales
- Reliability of the data
- Relationships between regional sales

Analytical approach

Initial exploration was done in Python and R:

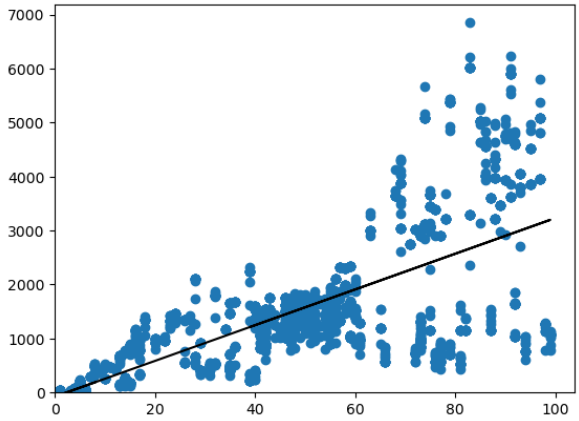
- Loaded data into Data Frames
- Determined metadata and descriptive statistics
- Checked missing values
- Cleaned data by removing irrelevant columns
- Renamed columns so the names are easier to reference

Regression analysis:

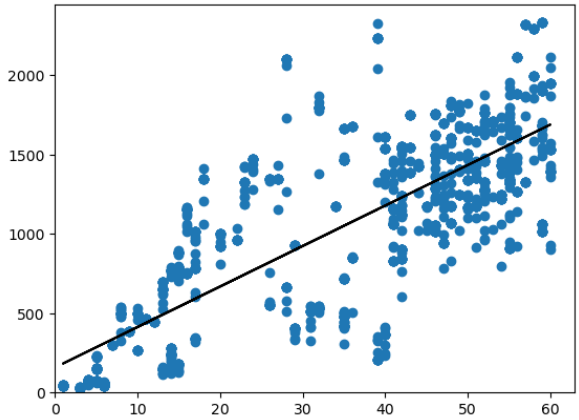
Using Python:

- Dependent variable: loyalty points.
- Independent variables: spending score, age, remuneration
- Created an OLS model for each independent variable.
- Extracted the estimated parameters, standard errors, and predicted values.
- Generated the regression table
- Plotted the linear regression and added a regression line for all three independent variables
- MLR regression analysis: R-squared value is 84, indicating a high level of correlation. We can use this model to accurately predict how the remuneration and spending score will affect loyalty points
- Improving the model: subsetted the data and reran a regression analysis on spending and remuneration. This provided better models but still with only 60% confidence.

Regression of spending score vs loyalty points before subsetting the data:



Regression of spending score vs loyalty points after subsetting the data:



Using R:

- Created linear regression model of:
    - Global and North American Sales
    - Global and European Sales
    - North American and European Sales.
- Plotted a simple linear regression for all three models using qplot
- Created a multiple linear regression model

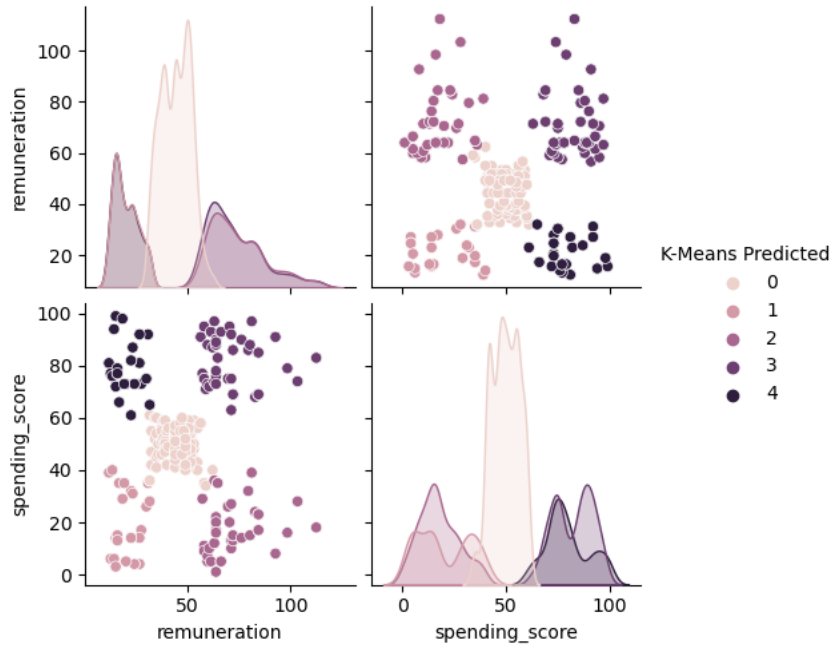
Insights:

- Cannot create good models for predicting customer loyalty points and sales.
- MLR: age and spending score indicates no relationship between the two variables
- MLR: all sales regions indicate a strong relationship between all regions' sales

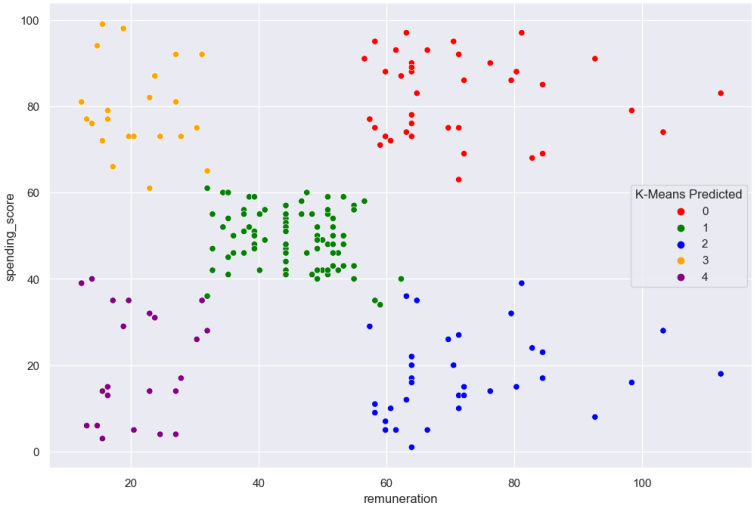
<u>Clustering the data:</u>

- Created new Data Frame including remuneration and spending_score columns
- Created pairplot to check for correlation
- Silhouette and Elbow methods: 5 is the optimal number of clusters
- Checked the usefulness of 4,5,6 by plotting the predicted *k*-means

Visualising 5 clusters:



k-Means predictive model:



Insights:

- By grouping customers into 5 groups, we can better predict what their spending scores will be, based on their remuneration.

<u>Conducting an NLP analysis:</u>

Preparing the data and creating a wordcloud:

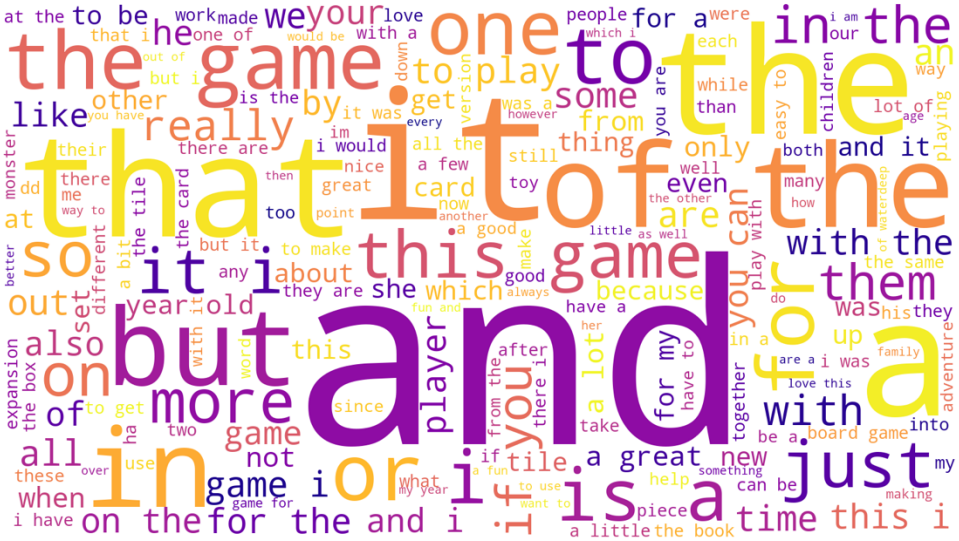
- Created a data frame that includes only review and summary columns.
- For each column:
    - Checked for missing values.

- Lower-cased words and removed punctuation
- Dropped duplicates
- Created a copy of the Data Frame.
- Applied tokenisation on both columns.

- Created a word cloud

Observations:

First word cloud:



- Stop words made it difficult to analyse popular words so I created another wordcloud without them.

Cleaned word cloud:

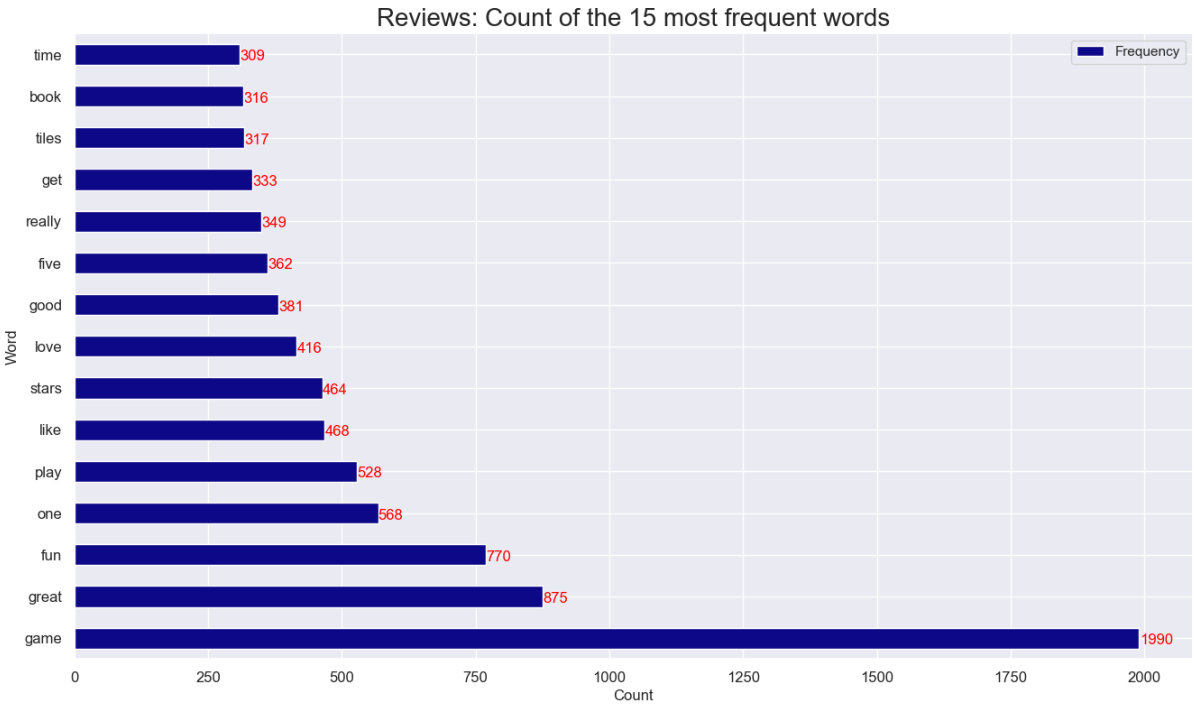
Insights:

- More positive words than negative in the wordcloud.
- Reviewers are mostly positive about their purchases.

- Biggest word: Game
- Also, big positive words: "fun", "five star", "great" and "love".
- Word cloud gives a comprehensive overview of the content in the reviews.
    - Next: gain insights from sentiment analyses.

Frequency distribution:

- Used FreqDist().
- Generated a data frame using Counter for top 15 words
- Created a bar plot to view the count



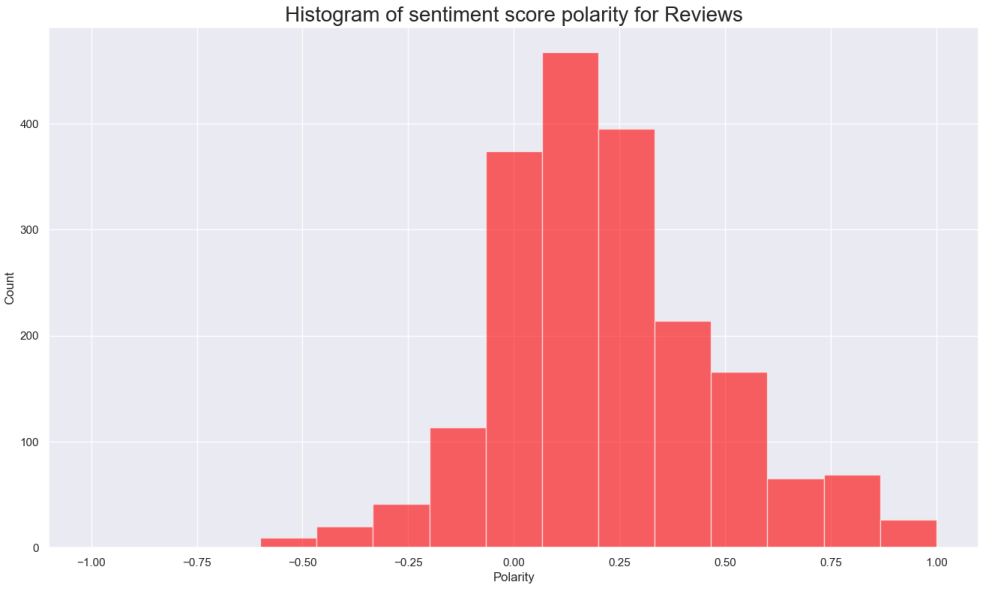
Insights:

- Positive words: great, good, love, like, five (I am assuming referencing 5 stars)
- Questionable word: **one**. Could mean number 1, 1 star or 1 player.
    - Next step: analyse those reviews more closely to see if "**one**" is positive or negative.

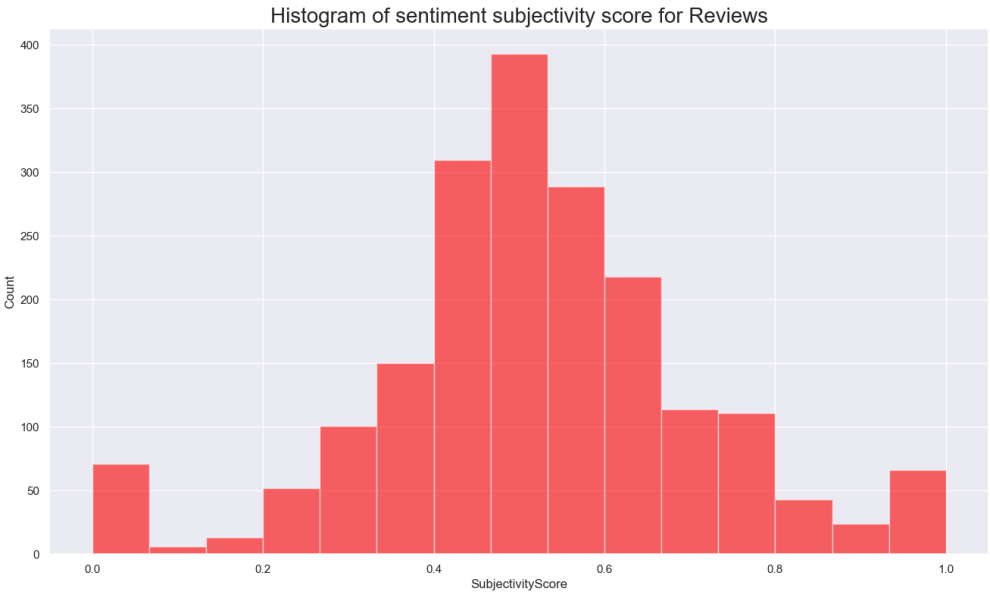
Determine polarity:

- Used generate_polarity and generate_subjectivity to calculate the sentiment polarity of reviews and summary
- Created histograms of polarity for reviews and summary.
- Used nsmallest and nlargest to identify 20 most positive and negative reviews and summaries

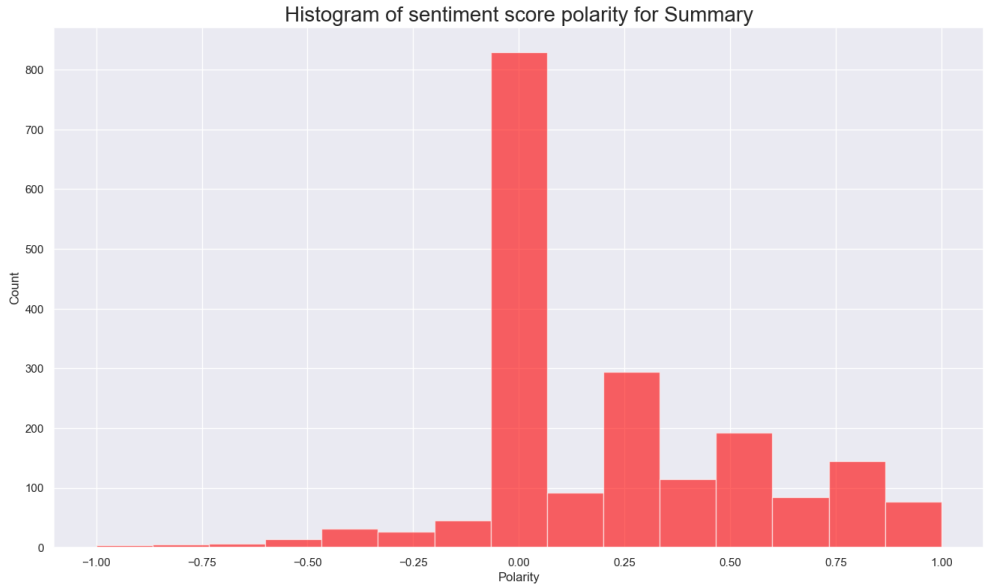
Histogram insights:



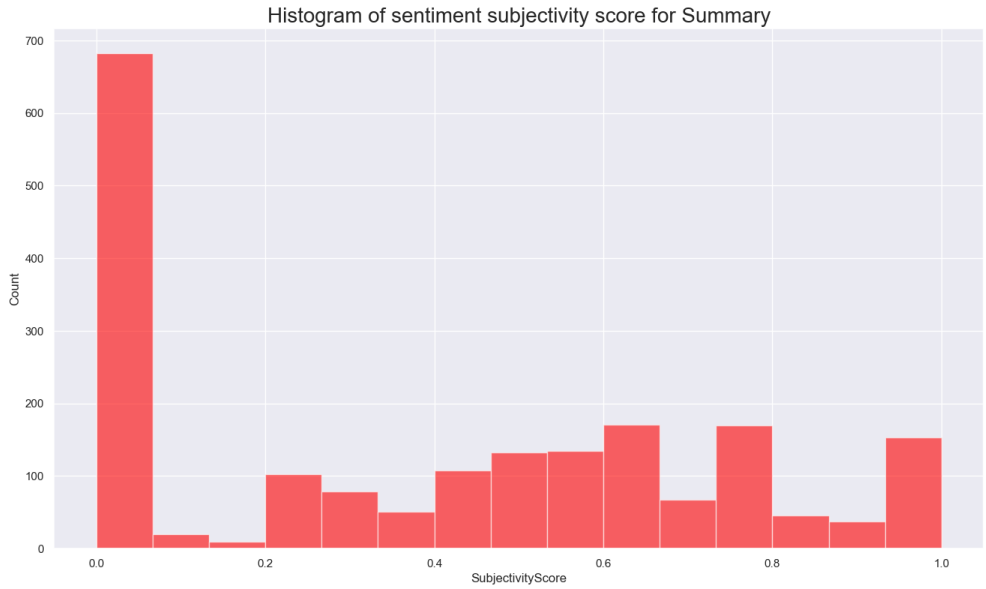
Reviews polarity: neutral to positive, there are few negative scores



Reviews subjectivity: falls around 0.5 – the reviews have an equal mix of subjectivity and objectivity. This is normal as people can get emotional when they are writing reviews and do not stay objective about purchases.



Summary polarity: neutral to positive, there is very little negative sentiment in the reviews



Summary Subjectivity: data is highly objective; we should use this to get a real reflection of the sentiment of the customers

Positive and negative review and summaries insights:

- Negative reviews: reviews in the bottom 20 already start leaning positive after the first 15 reviews. Few negative reviews, I would suggest reaching out to the customers to possibly rectify their problems.
- Negative summary: words disappointed and disappointing multiple times. We need to investigate the reviews of these summaries to see why customers were disappointed.

- Positive reviews: Use the positive reviews that are objective for accurate predictions of what products will sell.

- Positive summary: High subjectivity score - therefore we should read the full reviews. e.g.: perfect for a preschooler could indicate the product is easy and it may have been marketed to an older person and the reviewer was being ironic.

Recommendations:

- Look at products that have positive and objective reviews to predict which products will sell.
- Online reviews are important factors that influence sales.
- We also need to look at how negative reviews impact sales and if we should carry less products with negative reviews

<u>Visualisation Design and interpretation:</u>

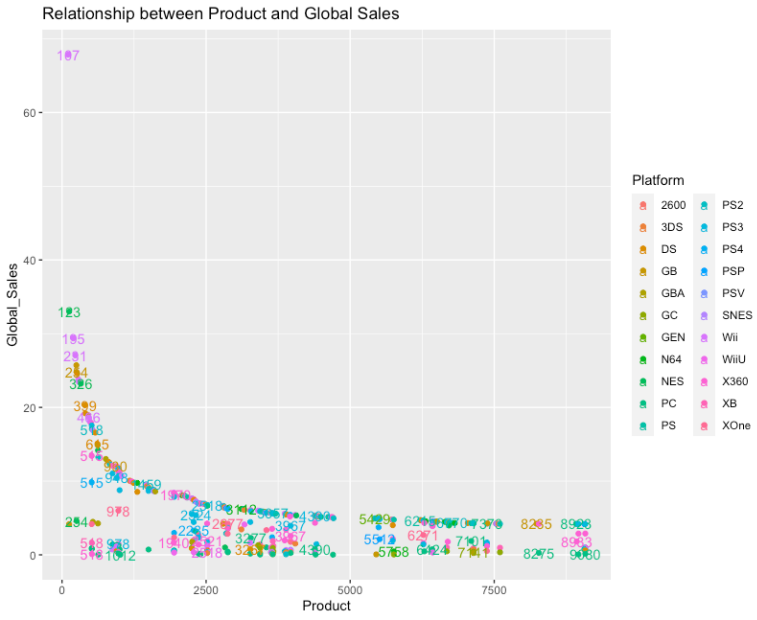
Preparing data for visualisations:

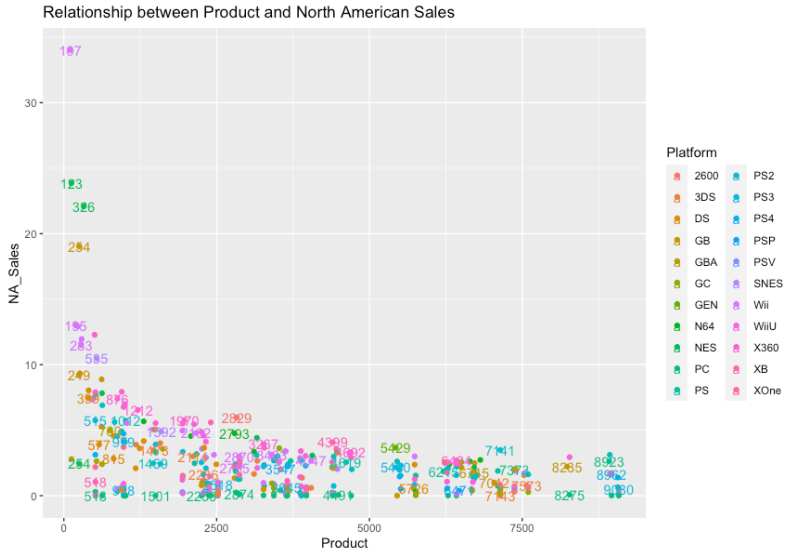
- Imported tidyverse, ggplot2 and ggrepel packages
- Loaded the data
- Subsetted the data frame to only keep relevant columns
- Viewed the descriptive statistics

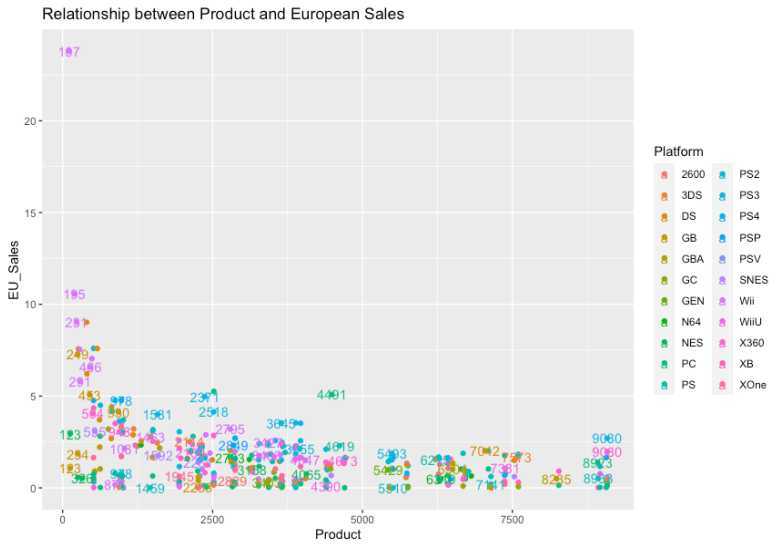
Creating visualisations:

Scatterplots:

- Scatterplots determining the relationship between:
    1. Each Sales Region and Product
        a. Colour: Platform to gauge products per platform
        b. Label: Product – ID most popular products





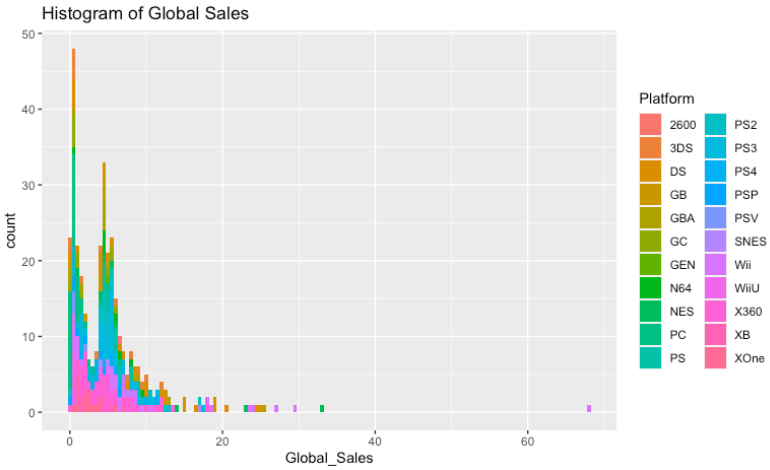


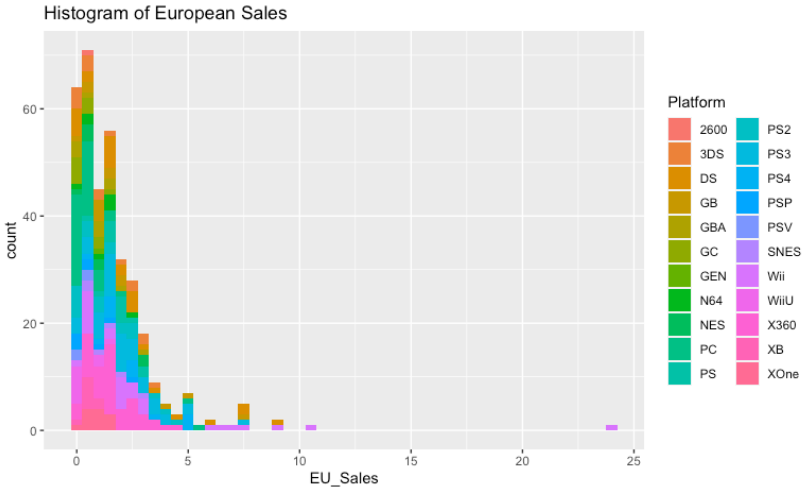
Insights:

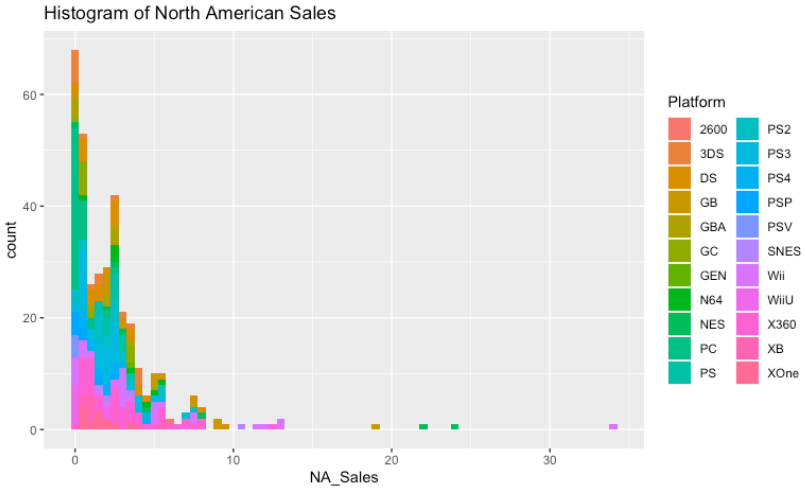
- All regions' most popular product: 107
- Product 123's popularity in North America pushes its popularity in Global Sales, even though it is not easily visible in EU sales
    - There should be two strategies for the different regions as the popularity of products is different
- However, product 195 is popular in both regions meaning products are popular in both regions
    - Going forward: analyse popular products and categorise them to predict future sales of similar products.

<u>Histograms</u>:

2. Global Sales
3. EU Sales
4. NA Sales







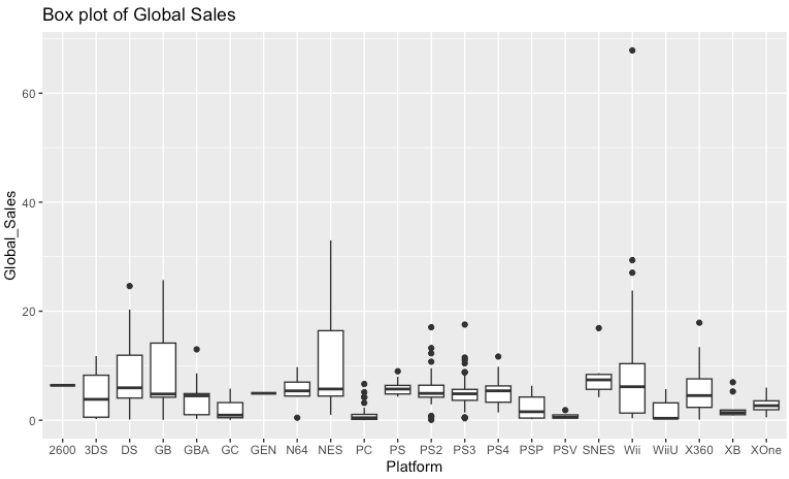
Insights:

- Most products sit in the same price category with some outliers.
- Wii products are an outlier in price, they are more expensive in each sales category

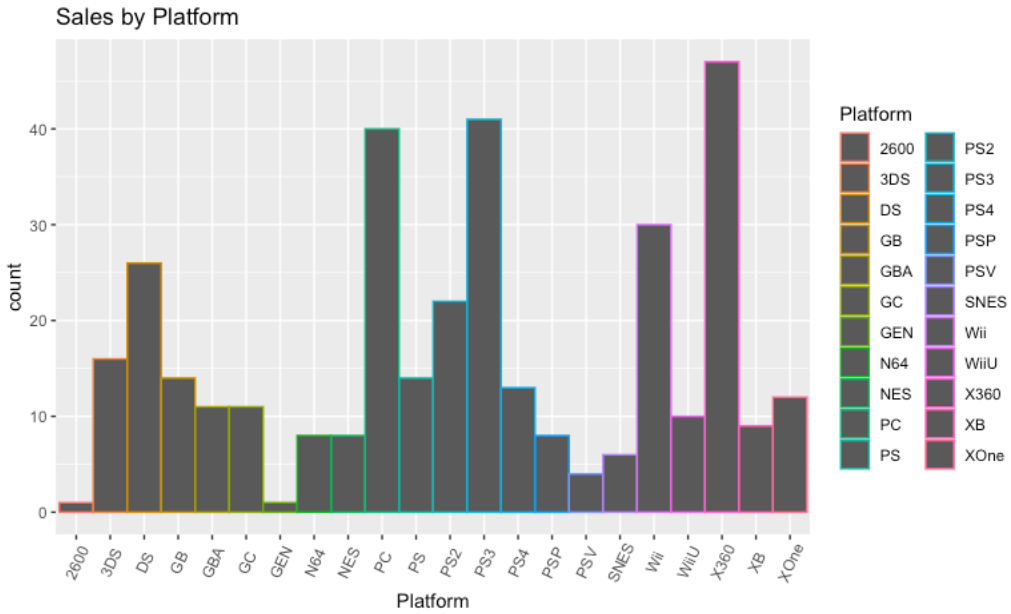
<u>Boxplots:</u>

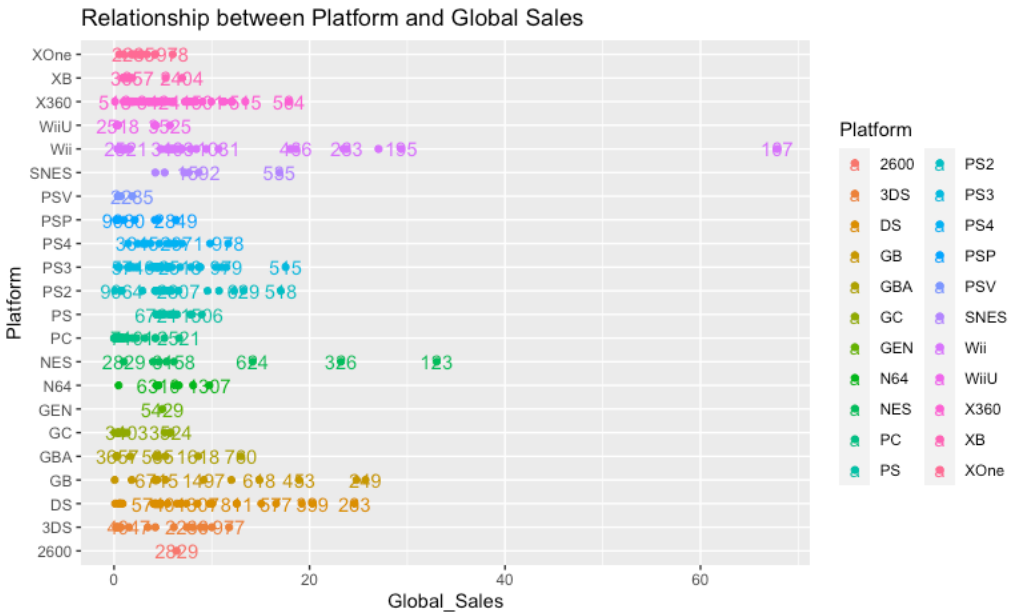
- Grouped the data by platform
- Created boxplots that look at regional sales grouped by platform

```
#Grouping the Data By Platform
# Group by sum of multiple columns
agg_tbl <- sales2 %>% group_by(Platform) %>%
  summarise(sum_nasales = sum(NA_Sales),
            sum_eusales= sum(EU_Sales),
            sum_gsales= sum(Global_Sales),
            count_product = sum(Product > 0),
            .groups = 'drop')
agg_tbl
```



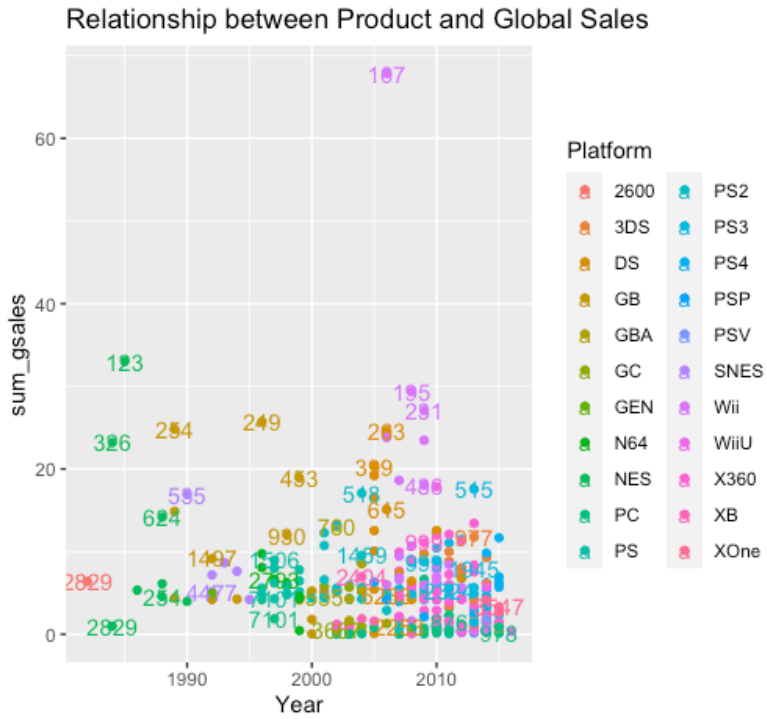
Plots that give insight into how to best craft a marketing strategy:





Insights and next steps:

- X360: most products
- Highest sales by platform: Wii.
    - We should have a marketing campaign to push X360 products to drive their sales up.
    - Evaluate if we should continue to sell products from platforms PSV and 2600 as their global sales are low.



Insights:

- Graph that best suits the data set – shows the sum of global sales of each product relative to the year it was released
- Shows trends of popular games and links them to a platform
- Use graph to gauge the popularity of games relative to their platform and predict which games and platforms will sell
- There are more products in the higher band of sales (20-40) that belong to the Wii platform

Determining normality with R:

- For all sales regions:
    - Q-Q plots: all sales have products that stray from the reference line, therefore the data is not normal
    - Shapiro-Wilk test: p-value: =<0.05. data is not normally distributed.
    - Skewness: high positive skewness
    - Kurtosis: data is highly skewed and is not normally distributed
    - Correlation check (r-values):
        a. Global and North American Sales: 0.9 - strong positive correlation
        b. Global and European Sales: 0.9 - strong positive correlation
        c. North American and European Sales: 0.7 - moderate positive correlation
            - Moderate to strong positive correlation between Sales regions
        d. Product and Global Sales: -0.4409046
        e. Product and North American Sales: -0.4047865
        f. Product and European Sales: -0.3894246
            - low negative correlation between Products and Sales values

<u>Final Insights and Recommendations:</u>

- Loyalty point accumulation:
    - There is no reliable model to predict loyalty point accumulation – next steps would be to collect more data on customers and rerun predictive models
- Customer groups to target:
    - customers with remuneration bands of 10-30 and 55-120. Investigate the popular products and create different strategies for each group. Gather more data about each group to understand why they have high spending scores
- Using reviews to inform marketing campaigns:
    - Look at products with positive and objective reviews to predict which products will sell.
    - Look at negative reviews' impact of sales
- Impact of products on sales:
    - Most products in inventory: X360 Platform
    - Highest sales: Wii

        - Start a marketing campaign to push X360 products to drive their sales up
- Reliability of the data:
    - The data is not normally distributed and there is little correlation between products and sales. The data is not reliable to make marketing strategy or predict sales based on product trends
- Relationships between regional sales:
    - Strong relationship between regional sales, we can use regions sales to predict other region's sales.

Wordcount: 1459