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

Student name: Rose Ryan

Submission date: 19 October 2023 (with extension)

Word count = 1,042

Declaration:

The work is my own and has been created with academic integrity.

Table of Contents

1. Background/ context of the business

Turtle Games is a game manufacturer and retailer with a global customer base which manufactures and sells its own products, along with sourcing and selling products manufactured by other companies. Its product range includes books, board games, video games and toys. The company collects data from sales as well as customer reviews. Turtle Games has a business objective of improving overall sales performance by utilising customer trends.

Turtle Games wants to understand:

- how 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 is (if any) between North American, European, and global sales

2. Analytical approach

- The metadata text document file provided (*metadata_turtle_games.txt*) which contained metadata of the two CSV files combined; data quality and reference for the data files, was opened in a text editor and reviewed
- A new Jupyter notebook was created for the project using the Anaconda environment, so Python could be used for the first part of the analysis
- A new R file in R Studio was created for the project in order to use R for the second part of the analysis
- The required libraries for analysis and visualisation were imported into the new Jupyter notebook and R Studio file respectively

- The data sets provided (*turtles_reviews.csv* and *turtle_sales.csv*) were imported into the Jupyter notebook and R Studio environments
- New DataFrames were created to store the data sets
- The DataFrames were sense-checked to determine the metadata, including column names, number of rows and columns, data types, and whether there were any missing values
- The descriptive statistics and metadata of the DataFrames were determined
- Unnecessary columns were dropped from the DataFrames and new DataFrames created where appropriate

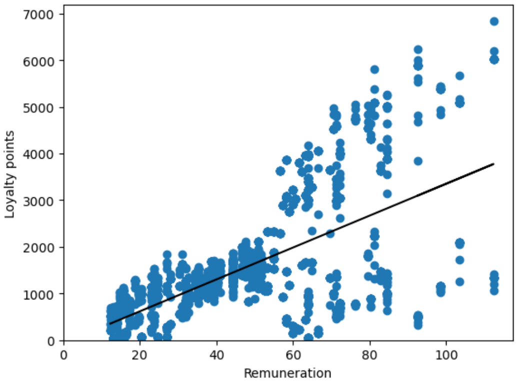
- The correct Python and R libraries were utilised, and functions and variables given names that are intuitive and descriptive
- Detailed and insightful descriptions of code and outputs at each stage were provided using the text markdown feature in Python at each stage
- PEP 8 was adhered to in the Python environment

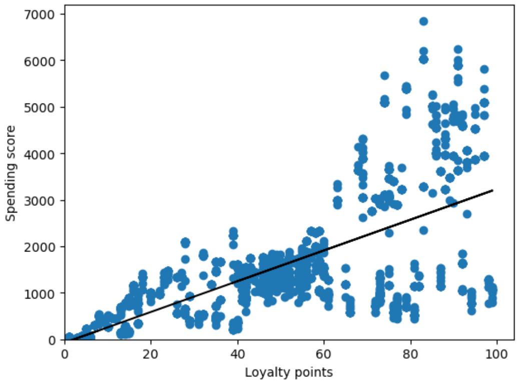
3. Visualisation and insights

Python analysis – customer trends

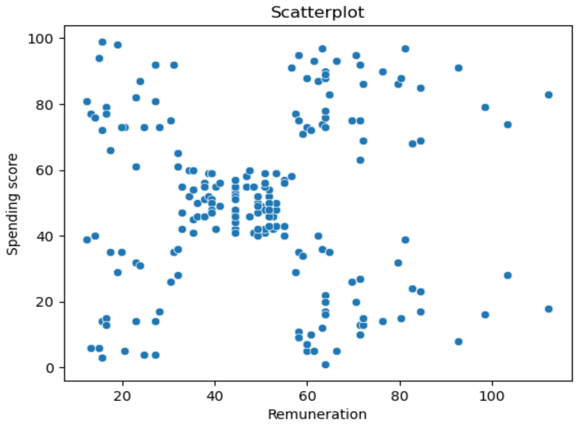
- The Seaborn and Matplotlib libraries were imported to enable visualization of the data contained in the DataFrames
- Visualisations were created in order to identify trends in the data
- When plotting charts, the basic visual design principles concerning chart type, colour, size, resolution, and layout were followed

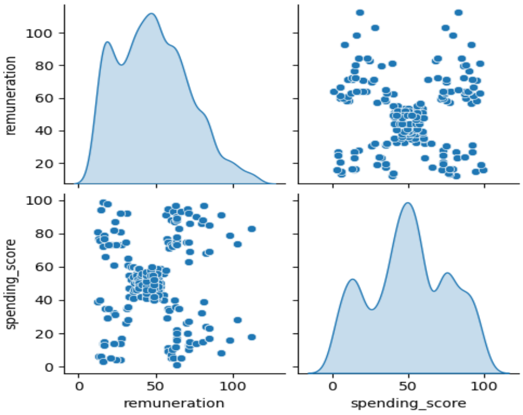
- Spending score vs. loyalty points showed a positive correlation
- Renumeration vs. loyalty also showed a positive correlation
- Age vs. loyalty did not show a correlation

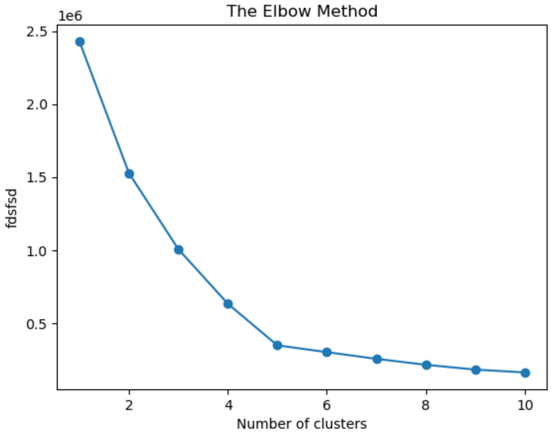


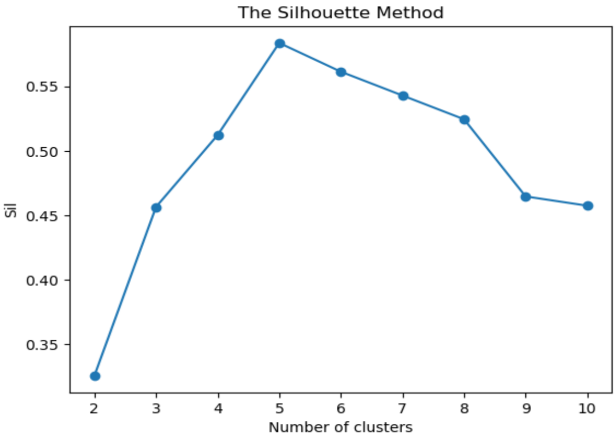


- Remuneration and spending score showed five clusters when plotted in a pairplot, which was also confirmed by both the Elbow and Silhouette methods
- K was set at 5 and K-means clustering was carried out
- This would suggest it would be most effective to target customer groups in five different and independent ways





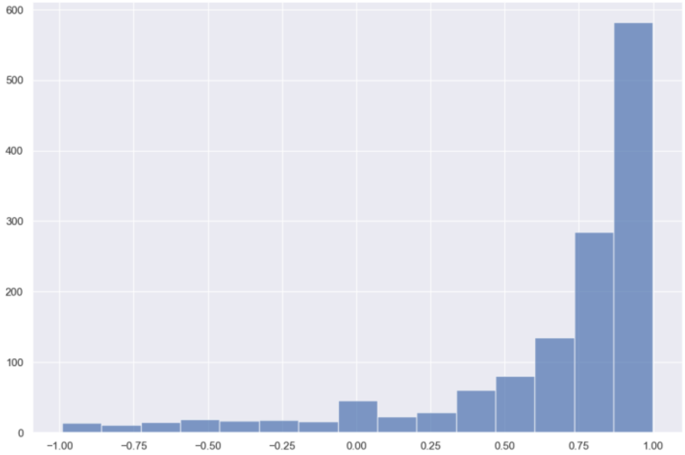




- NLP on the text of online reviews showed the fifteen most common words occurring in reviews (after stopwords had been removed) to be:

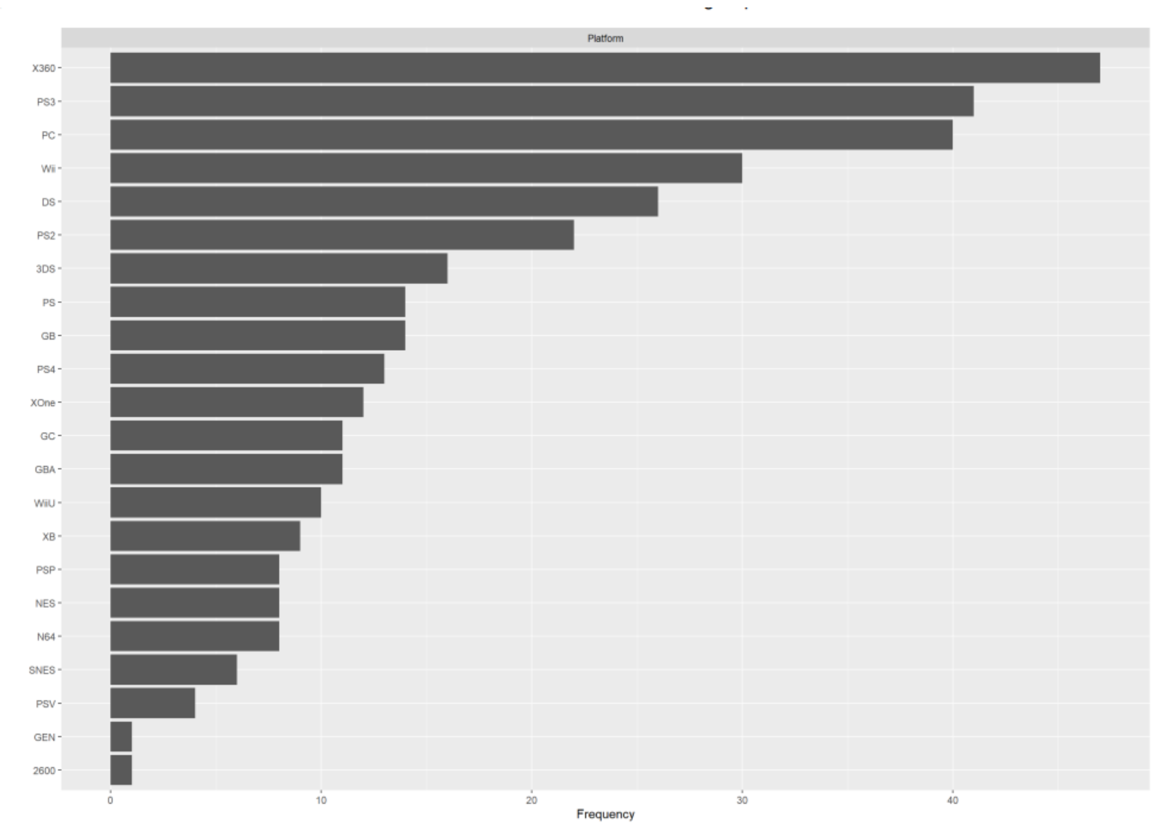
```
('game', 268),
('great', 236),
('fun', 175),
('good', 84),
('love', 70),
('like', 54),
('kids', 48),
('book', 42),
('expansion', 42),
('cute', 40),
('old', 34),
('really', 30),
('set', 29),
('nice', 28),
('one', 28)]
```

- A WordCloud for review text was created after stopwords had been removed which shows an easy to read visual of the most commonly occurring words in the reviews
- The polarity of reviews was calculated and showed a strong positive skew, with 50% of reviews scoring between 0.75 and 1.00.



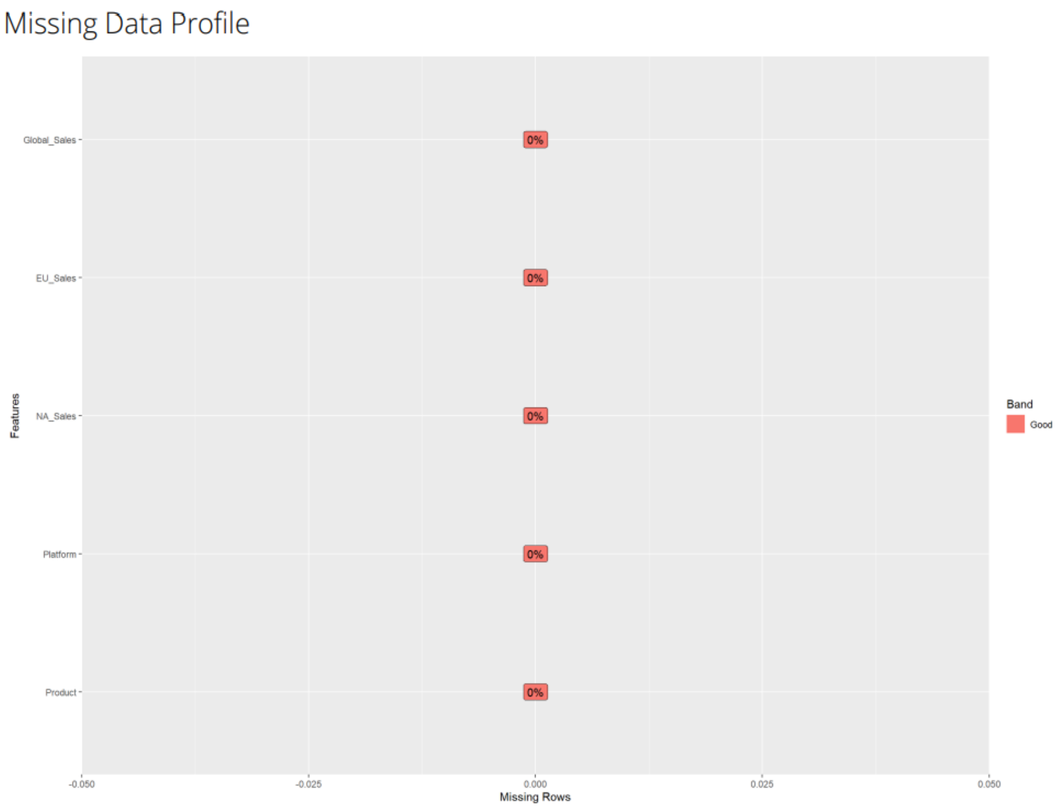
R analysis – how different products affect sales trends

- The top three platforms are:
    1. Xbox 360
    2. PlayStation 3
    3. PC



R analysis – reliability of the data

- No missing data was identified



Descriptive statistics for Products and all regional sales:

```
    Product        Platform            NA_Sales          EU_Sales
 Min.   : 107   Length:352         Min.   : 0.0000   Min.   : 0.000
 1st Qu.:1945   Class :character   1st Qu.: 0.4775   1st Qu.: 0.390
 Median :3340   Mode  :character   Median : 1.8200   Median : 1.170
 Mean   :3607                      Mean   : 2.5160   Mean   : 1.644
 3rd Qu.:5436                      3rd Qu.: 3.1250   3rd Qu.: 2.160
 Max.   :9080                      Max.   :34.0200   Max.   :23.800

  Global_Sales
 Min.   : 0.010
 1st Qu.: 1.115
 Median : 4.320
 Mean   : 5.335
 3rd Qu.: 6.435
 Max.   :67.850
```

Correlation between NA Sales and EU Sales

```
cor(newdata$NA_Sales, newdata$EU_Sales)
0.7055236
```

Interpretation:
This shows a strong correlation between NA and EU Sales, as it is closer to 1 than to -1.

Shapiro-Wilk normality test

```
data:  newdata$Global_Sales
W = 0.6818, p-value < 2.2e-16
```

Interpretation:
The small p-value of 2.2 x 10-16 suggests that the data significantly departs from a normal distribution.

Skewness

```
skewness(newdata$Global_Sales)
4.045582
```

Interpretation:
The positive skewness value (greater than 0) shows that distribution is skewed to the right.

Kurtosis

```
kurtosis(newdata$Global_Sales)
32.63966
```

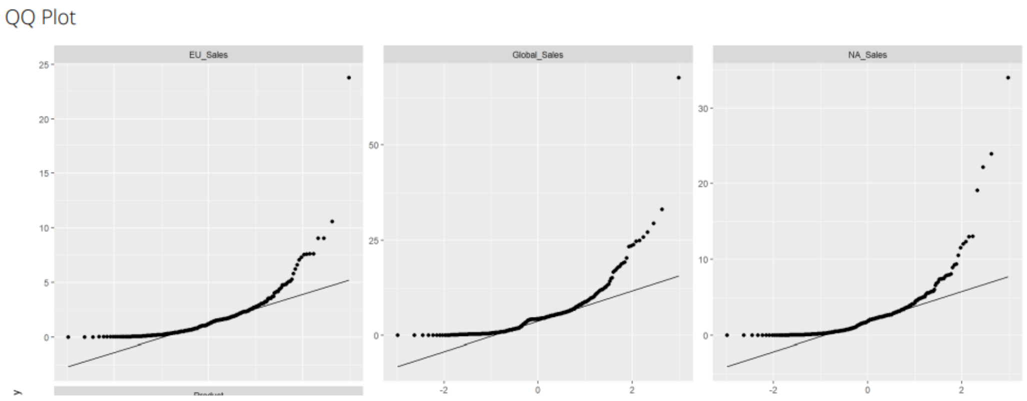
Interpretation:
The very high positive kurtosis value indicates that the distribution has heavier tails and a more peaked central region, as well as outlying data points.

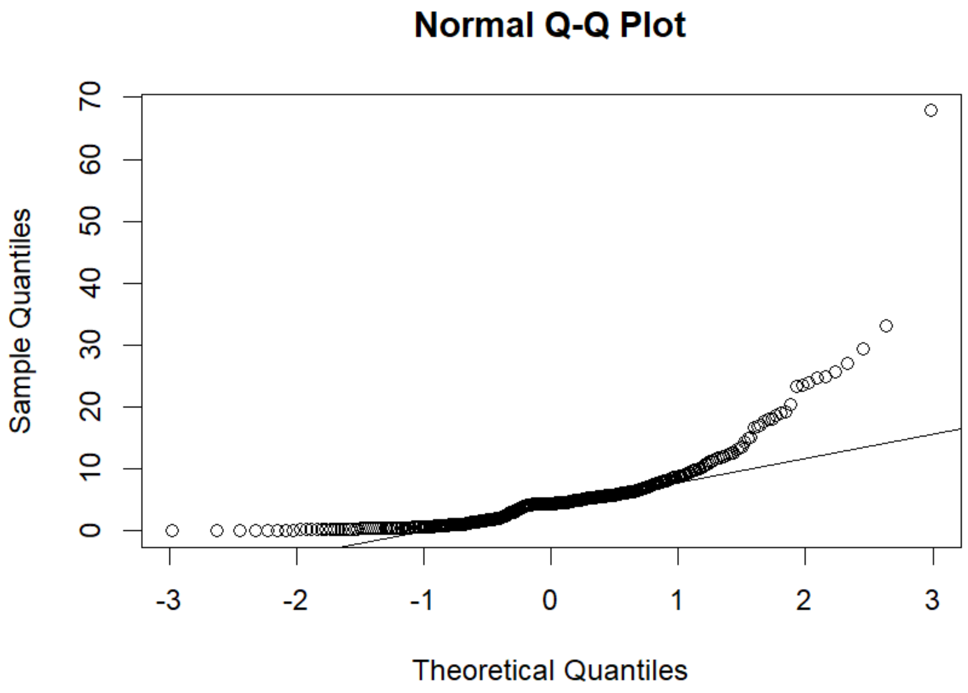
QQPlot

Interpretation:

The curve of the plot infers that the data is skewed, with all data points lying either on or above the theoretical normal line.

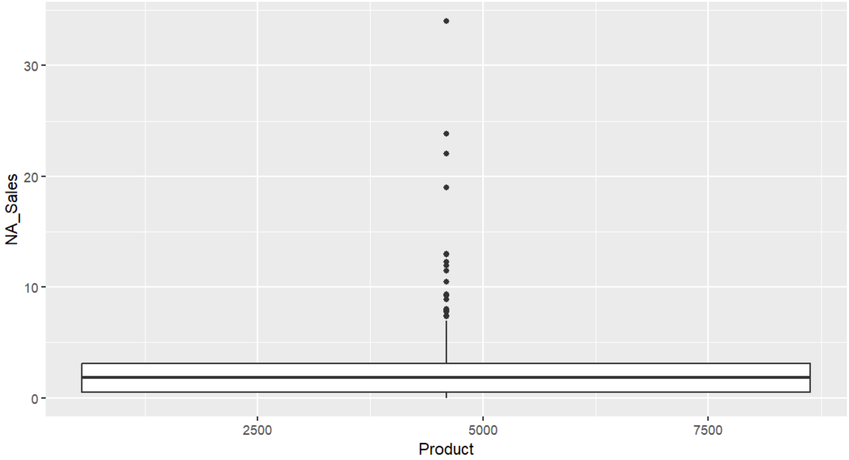
They also all show a positive skew to the data, with some extreme outliers for *NA_Sales*.

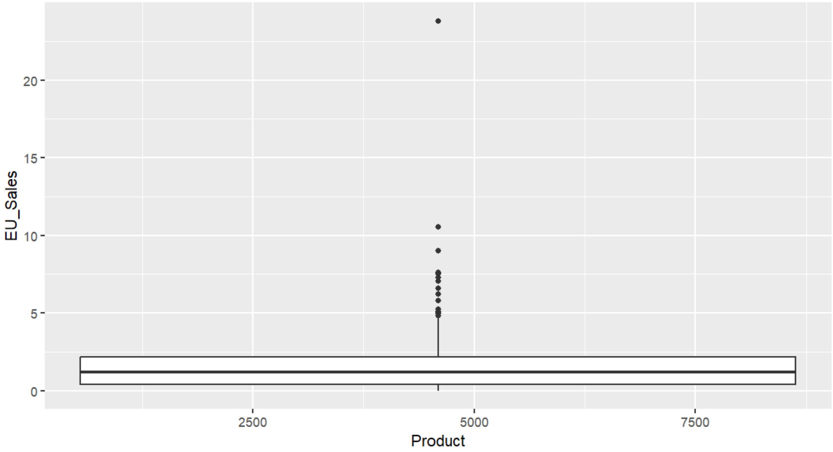


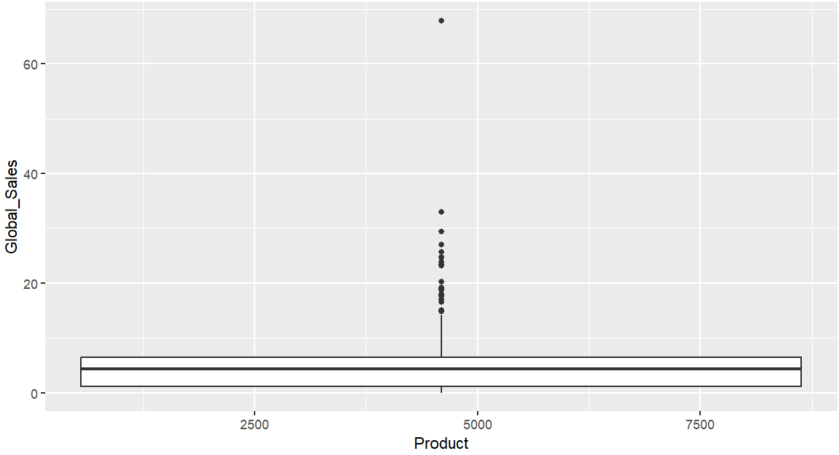


Boxplots

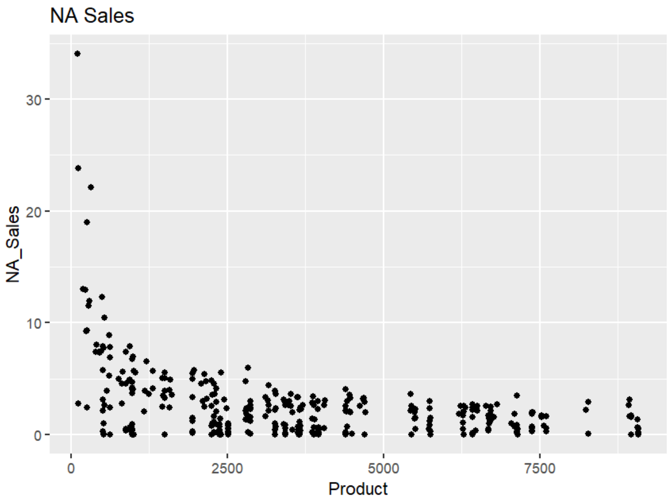
- Boxplots for all three sales regions show similar trends

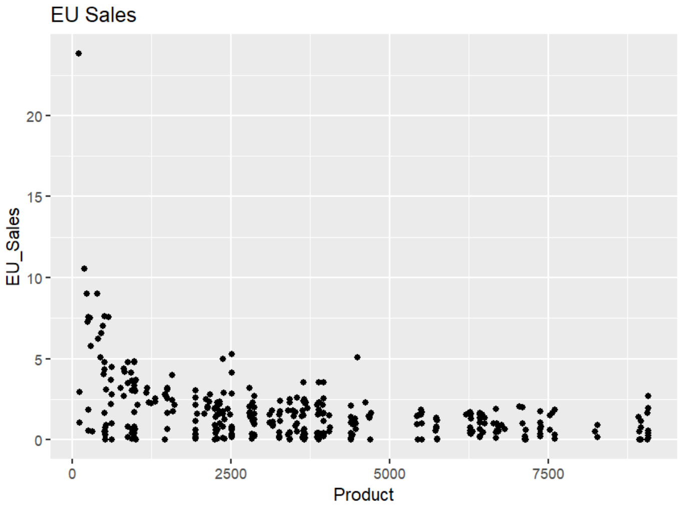


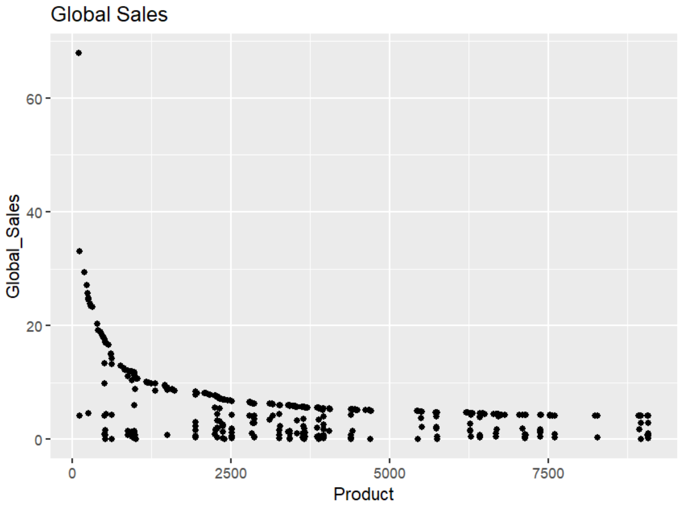




Linear regression







4. Patterns and predictions

- When targeting specific customer groups, this should be approached in five different ways, based on the distribution within the customers for renumeration and loyalty points
- Customers with lower spending should be targeted more in marketing in order to attempt to increase their spend within the environment
- The platforms with lower-performing figures could benefit from improved marketing and advertising
- Reviews are mostly positive, based on the WordCloud, fifteen most commonly occurring words, and polarity, so there would not be a great deal of work needed to be done to improve upon this for the online presence of Turtle games