London School of Economics
Data Analytics Accelerator Programme – Summer Cohort CO3

Course:	CO3 Data Analytics Using Python
Assignment:	Predicting future outcomes
Prepared by:	Christos Pieris
Date:	08 January 2023 (Approved Extension)

Content of this report

Submitted deliverables in GitHub repository

1. A Jupyter Notebook (IPYNB)

2. A recording (as a .mp4 file)

3. A PDF report indicating the approach, thought process, results, conclusions, and recommendations (pdf)

1 Background and context of the project

1.1 Project Background – Turtle Games

Turtle Games is a global business in manufacturing and selling their own games, as well as games from other originators to retail customers around the world. Their inventory includes books, toys, board and video games and their primary objective is to improve the sales performance across their entire product range.

1.2 Problem Statement and scope of investigation in this report

This project has been contracted to help Turtle Games to derive insights from their available data that will help them understand how to improve sales performance. While the task can be observed by multiple angles, this investigation considered a range of key items which are summarised below and detailed in the rest of the report.

a) How customers accumulate loyalty points
b) How useful are remuneration and spending scores data
c) Can social data (e.g. Customer reviews) be used in marketing campaigns
d) What is the impact on sales per product
e) The reliability of the data (e.g. Normal distribution, Skewness, Kurtosis)
f) If there is any possible relationship(s) in sales between North America, Europe, and global sales

2 Analytical Approach

2.1 Project development environment:

We are using Python as the language to import, manipulate and interpret the available data and R as the language to perform further statistical analysis. Both languages have been selected for their powerful libraries yet intuitive nature.

The coding environment is split between Junyper Notebook and R Studio. Our collaboration mechanism for development is Gitlab where updates are passed from RStudio and Junyper notebooks in Gitlab repository branches which merge to the base code allowing team members to continually work on the project without disrupting the continuous improvement of the base code.

The Gitlab repository can be accessed by following the link below: https://github.com/KrispyPi/DA301-Predicting-future-outcomes.git

2.2 Available project data

2 .csv files and 1 text file (metadata) provided by the Turtle Games containing details on customer reviews (turtle_revies.csv) and their known sales across geographical regions (turtle_sales.csv).

2.3 Processes taken to explore the data

2.3.1 Importing Libraries:

Importing, manipulating, and visualising data in Python was a vital starting point and it was done by utilising NumPy, Pandas, seaborn and matplotlib's packages. In order to surface customer loyalty, the linear regression function was called from the statsmodel library. Further analysis required K-Means clustering which was done using Python's sklearn and scipy libraries. Lastly, the Python section concluded by performing Natural Language Processing on the available customer reviews provided by Turtle Games. To derive meaning out of otherwise meaningless text, sentiment analysis was constructed using the NLTK toolkit.

Covering the business objectives required further analysis using R where the majority of data importing, wrangling and analysis utilised functions available from R's Tidyverse.

2.3.2 Importing the Data:

The Junyper Notebook and RStudio were used to run Python and R commands to import the files, hence the notebook and files shared the same local directory. Pointing to the files was done utilising the name of the common directory in the Pandas and Tidyverse read csv function which was used to import the files. The result of this operation was a Dataframe, a 2-D table holding the imported data. A second step following the Reviews import in Python was done to remove unnecessary columns, namely the Language and Platform since the both had constant values i.e adding no information. After renaming the Reviews columns a new dataframe and .csv export resulted in what was used in the rest of the project. The theme of using Pandas Dataframes as the medium to import, read, store, and manipulate data was reflected in R, where Sales Data was introduced in dataframes using Tidyverse's read.csv. Finally, subsequent R dataframes were recreated by subsetting the original import, to filter or combine columns and rows as part of the analysis.

2.3.3 Exploring the Data:

With the data imported into data frames, it was possible to start reading into the structure and content of each data set since it important at this initial stage to get an idea of the size and type of our files.

Quickly built rapport with the data sets, by performing the below basic statistics:

1. Printing the first lines to confirm import worked as expected
2. Viewing the information for each data set which gave us the number of columns, rows, the count of each type of data as well as the basic descriptive statistics such as mean values for each column.

2.4 Processes taken to analyse the data

Explored and analysed the data by performing the below steps in Python:

1. Linear Regression to evaluate potential relationships between loyalty points versus each of: age, renumeration and spending scores.

2. Clustering with K-Means to identify groups or sub-sets of data within the entire dataset that can make-up customer segments. The optimal number of clusters was chosen to be k=5 since it seemed to better represent the customer base.
3. Natural Language Processing or NLP to identify the 15 most common words used online by their customers when the discuss their products as well as the most positive and negative customer reviews available to date. NLP can be quite an elaborate task. The data for this task was purely textual and required pre-processing to normalise entries, remove unnecessary words from sentences, tokenise words and assess their frequency before eventually generating the polarity for each review and deriving sentiment scores from each review.

Further explored and analysed the data by performing the below steps in R:

1. Exploratory Data Analysis of Sales Data, by importing the original file and reducing the columns to only those that added value to this part of the analysis. Ranking, Year, Genre, and Publisher were irrelevant to analyse sales performance across different Platforms and geographic regions, so the were dropped. Initial visualisation was performed using scatterplots, histograms and boxplots.
2. Data cleaning and manipulation in R surfaced descriptive statistics such as min, max, mean which were summarised. The impact on sales per product was seen using the group_by and summarise functions which essentially reduced Sales entries to the sum per product. Q-Q Plots and the Shapiro-Wilk test surfaced the normality of the data and correlation functions were called against each Sales column to uncover positive correlation between the two regions and the global sales.
3. Simple and multi-linear regression to determine the correlation between the Sales columns and finally test the models prediction capability by running known test points and observing its ability to predict Global Sales values given the sales observed in two geographic regions.

3 Visualisation and insights

3.1 Overview

3.1.1 Seaborn and pyplot library were used and originally imported into the working environment. The chosen libraries are built on top of the extensive matplotlib library and allows parsing data in Pandas based modules. In R, the ggplot2 package was utilised straight out of the Tidyverse. Qplot was used to quickly plot simple graphs and ggplot for more elaborate results. These were meaningful step to easily produce consistent graphical representations of our data.

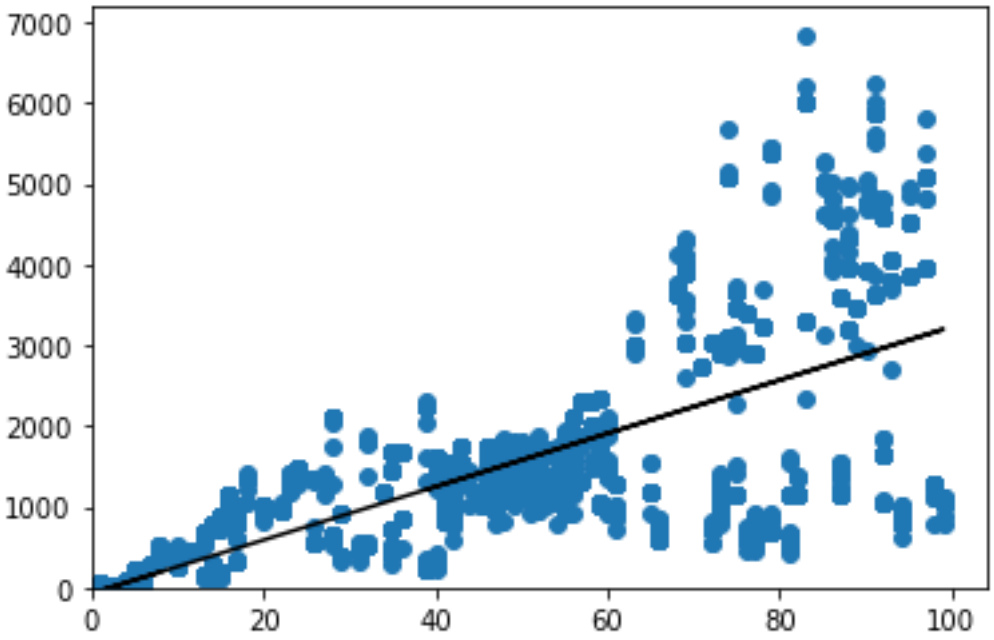
3.1.2 By analysing the data, we identified initial trends, and it was evident that visual representations would help surface them. While the tables that resulted from the analysis provided the facts, it is the interpretation of different facts when put together that this stage achieved.

3.1.3 The project utilised mostly Scatterplots, Histograms and Boxplots. Wordclouds were used for the NLP section which were found very effective and interesting illustrations.

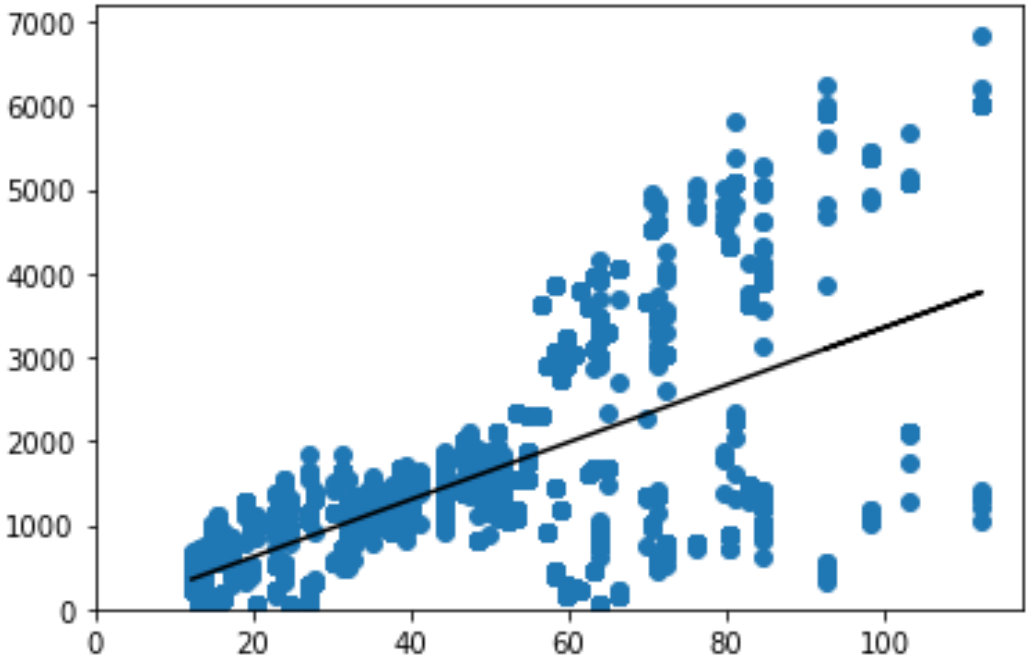
3.2 Linear regressions in Python

At this stage the project identified that both Renumeration and Spend positively correlated with Loyalty so that the more spend or income the higher the indication for Loyalty. On the other hand, Age was less correlated in an inverse direction indicating that with older customers loyalty will tend to suffer. It is expected that Turtle Games will benefit from loyal customers existing in the higher income brackets who are younger as they will be more willing to spend on products.

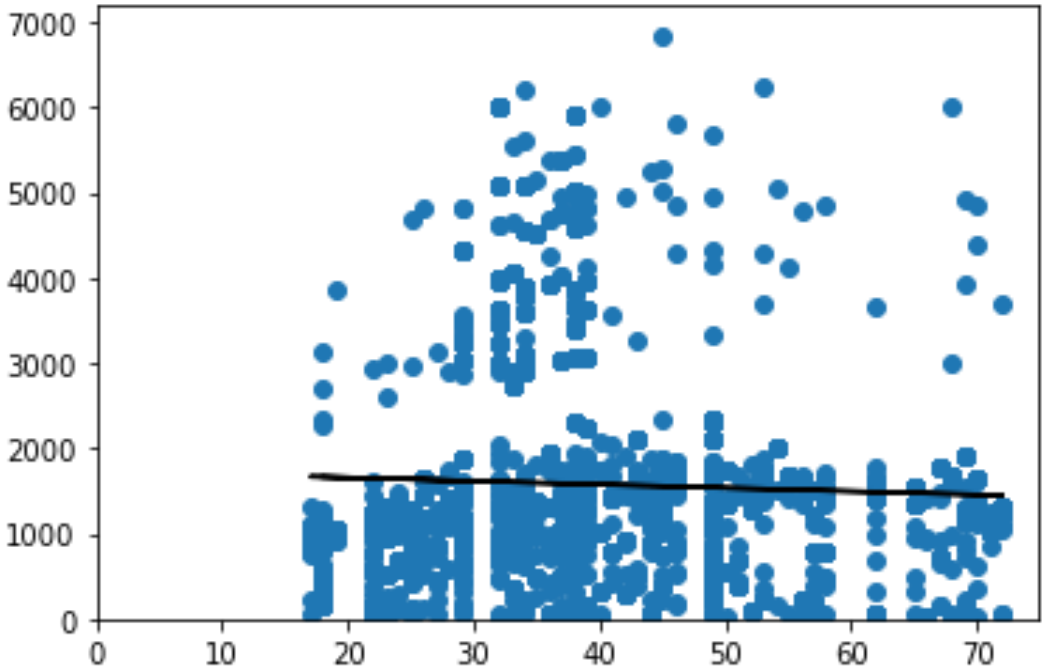
3.2.1 Spending vs Loyalty



3.2.2 Renumeration vs Loyalty



3.2.3 Age vs Loyalty



3.3 Clustering with K-Means in Python

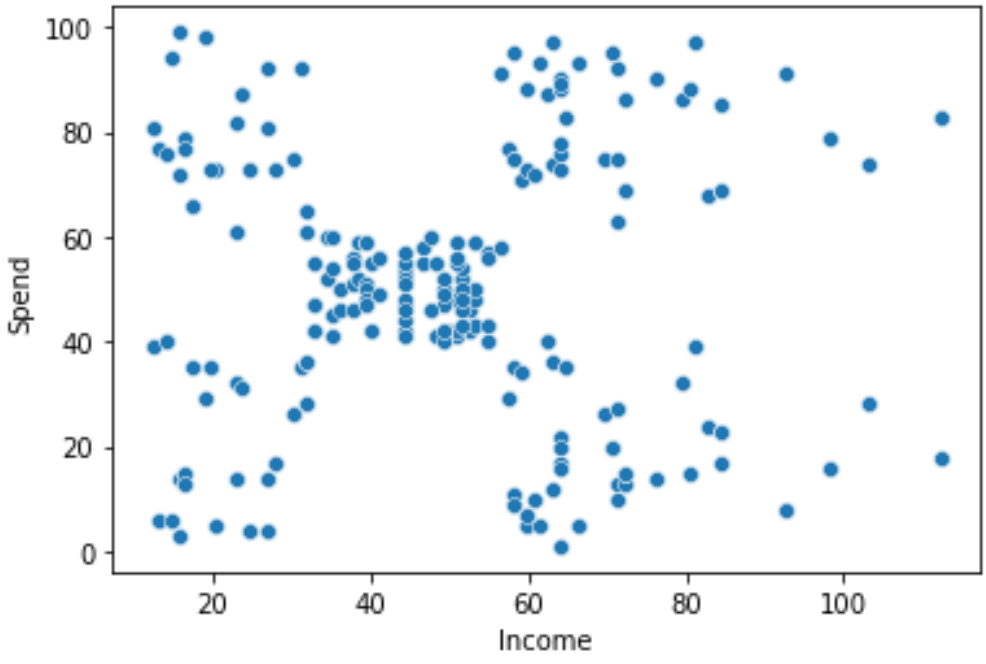
At this stage the customer base was passed through the clustering mechanism enabled by the K-Means clustering in Python. Visually the data points could be seen taking up natural cohorts in the Spend vs Income plot-space. 5 such clusters were chosen as the best fit to describe the customer base. This is an important point to be made against the original business objective since it shines light into the potential segments for Turtle Games to target.

Considering that in an ideal world Turtle Games would have customers spending as much as possible on their products, the resulting segmentation lays the ground for the following questions to be answered.

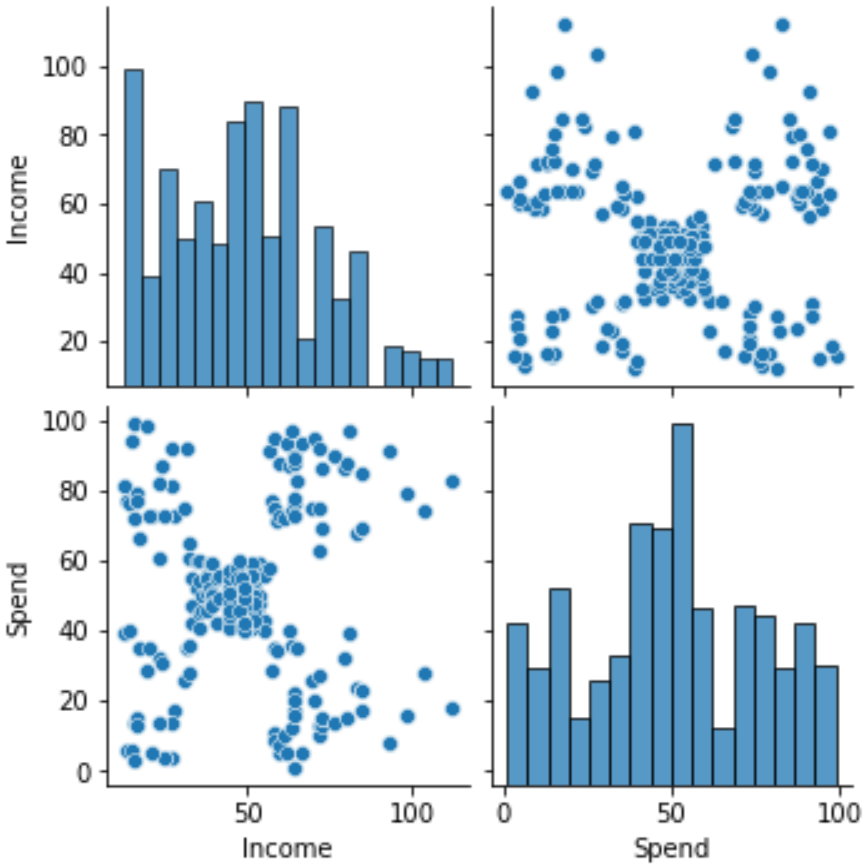
a. What are the characteristics of customers in the highest spending segments?
b. What are we doing as a business to maintain the customers within the highest spending segments?
c. What can we do to sell more items to the customers who are willing to spend more? Cross selling tactics
d. What is the churn of customers within these segments and what can we do to maintain customers within these segments?
e. What can we do to bring more customers over from the lower to higher spending segments.

Answering these questions will allow the business to knowledgeably invest in the right product development and subsequent marketing campaigns that will have the highest likelihood of resulting in the highest possible revenues.

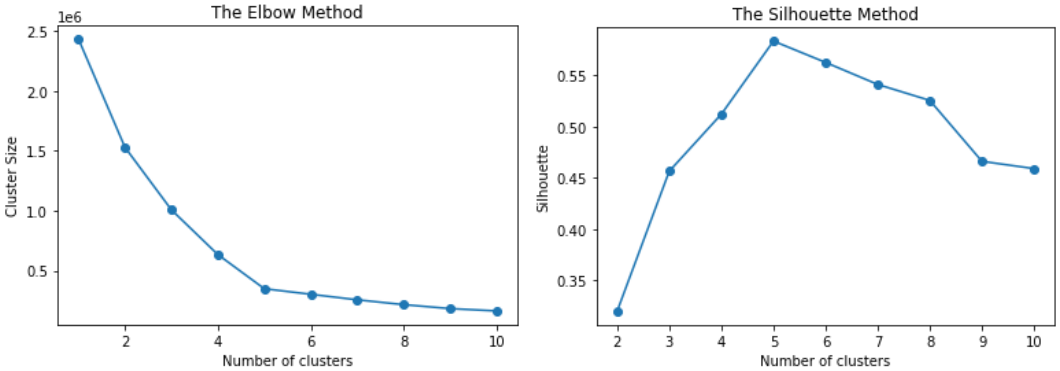
3.3.1 Income vs Spend Scatterplot



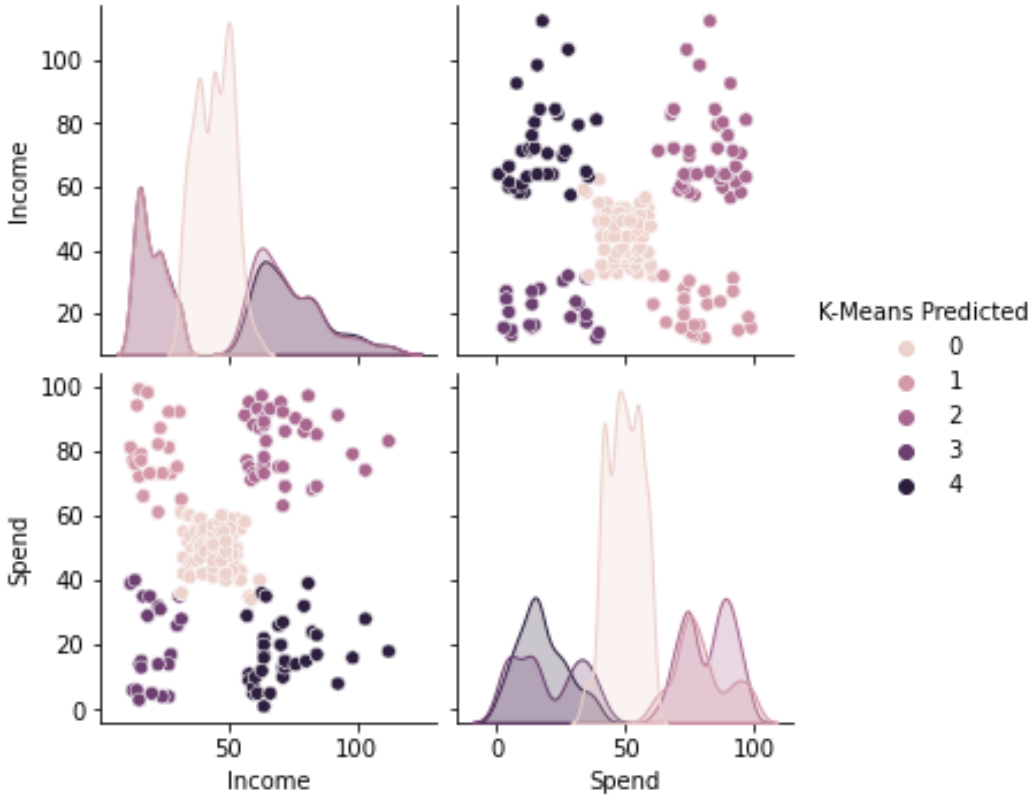
3.3.2 Income vs Spend Pairplot



3.3.3 The Elbow and Silhouette Method



3.3.4 K-Means predicted final pair plot with K = 5

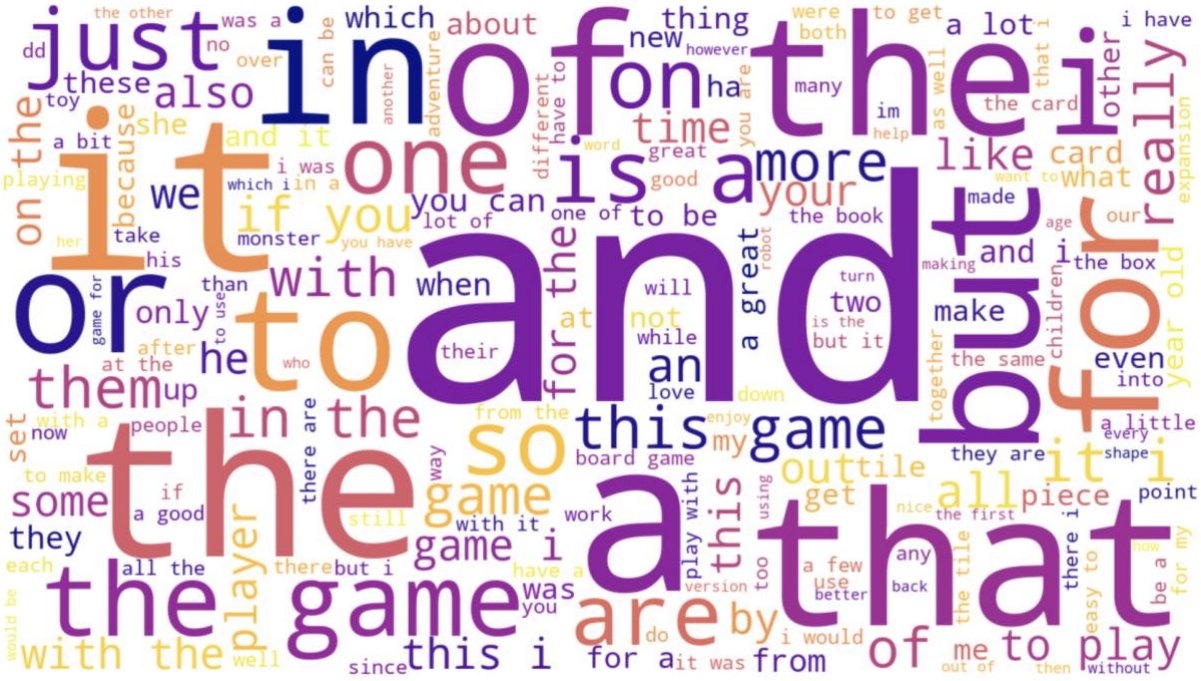


3.4 NLP in Python

The project took on the task to zoom into the available reviews given by Turtle Games customers online. As an overview, some of the most common words used by customers online seem to unsurprisingly be “Game”, “Play”, “Get”, “One” and “Like” and some other not so expected such as “waterdeep”, “monster”.

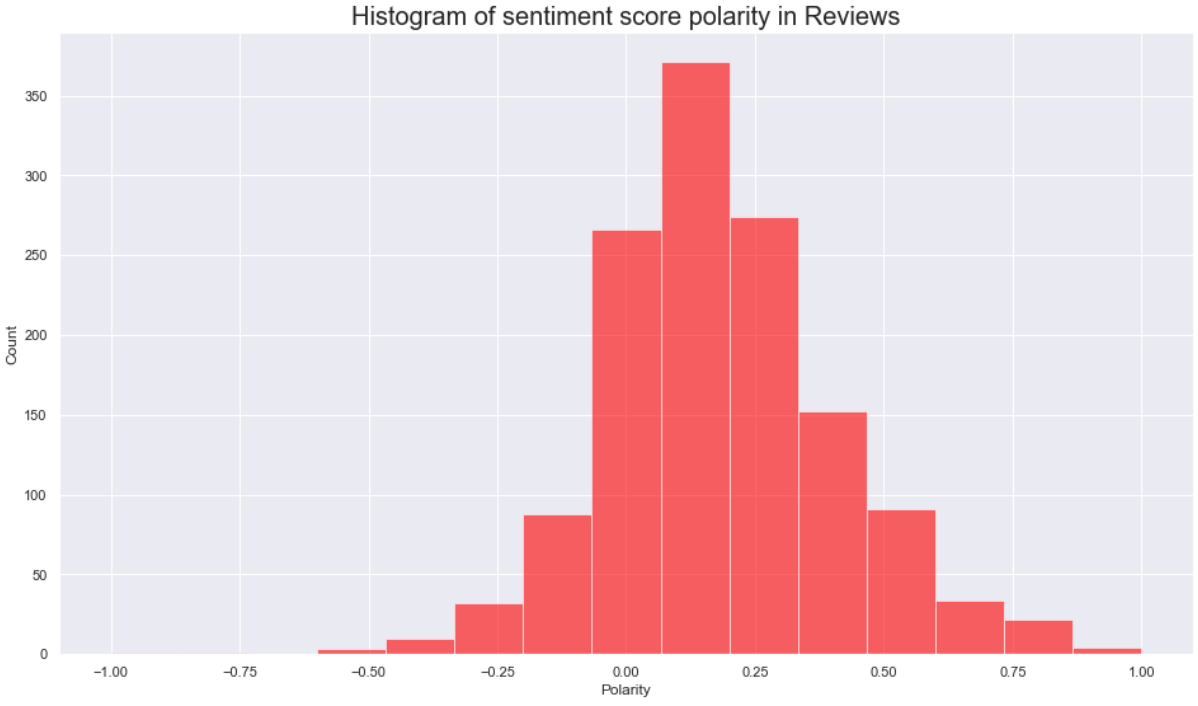
Turtle Games as any other business that trades at this scale was found to be subjected to both positive and negative reviews online. The project identified the top 20 in both directions of sentiment which is expected to help the business identify features related to the customer experience that yielded each. Following this exercise, the business will be in better position to position marketing campaigns that shine on the positive sentiment yielding features while refrain from advertising the opposite.

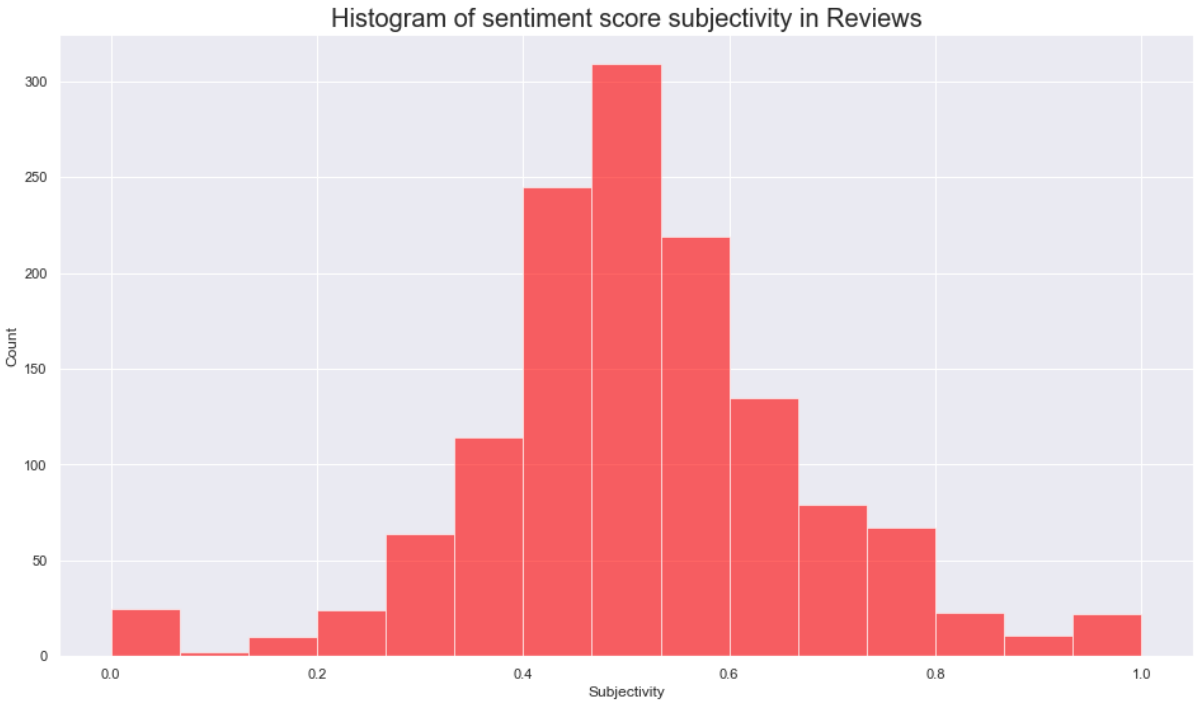
3.4.1 Wordclouds for all Reviews and Summaries



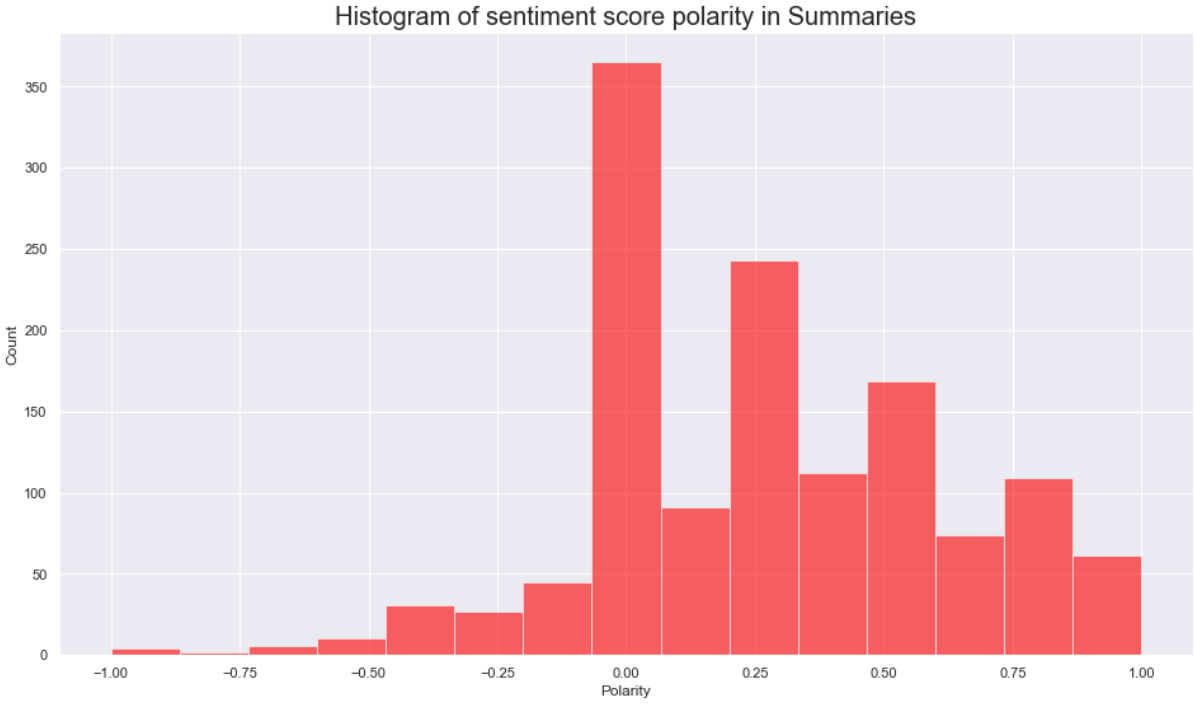
3.4.2 Wordclouds for Reviews and Summaries after stopword extraction

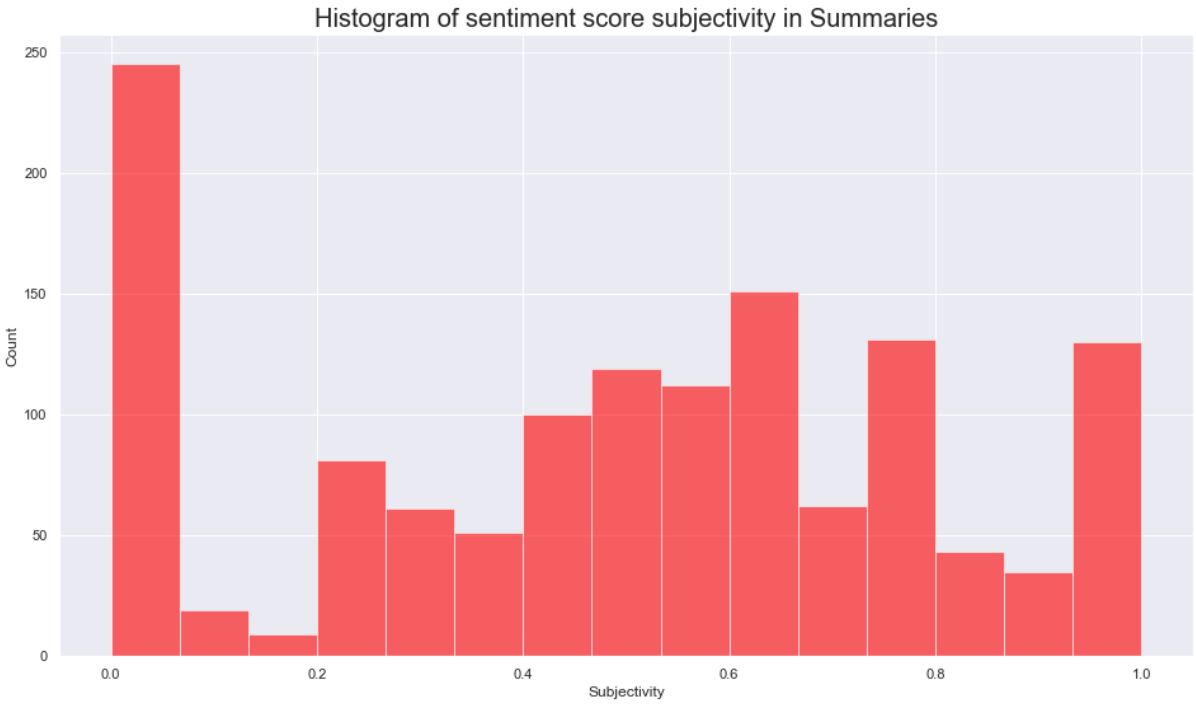
3.4.3 Sentiment score Polarity and Subjectivity distribution in Reviews





3.4.4 Sentiment and Subjectivity Score distribution in Summaries

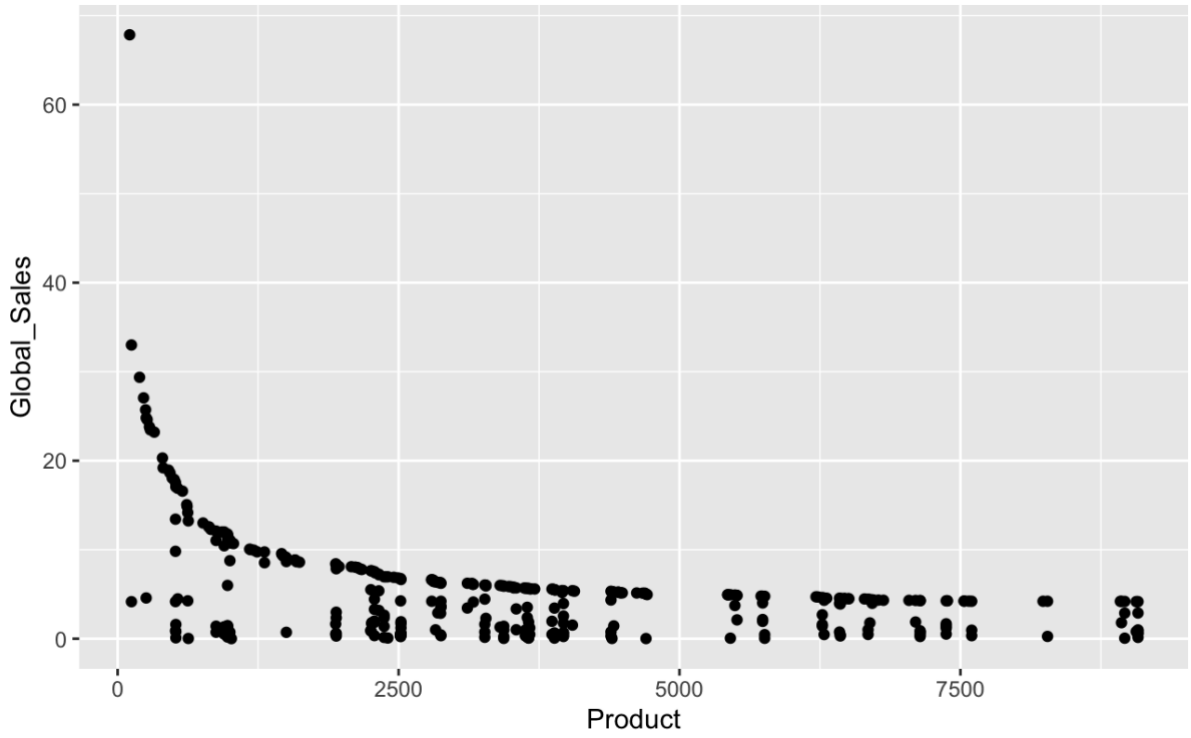



Histogram of sentiment score subjectivity in Summaries
Count
Subjectivity
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0

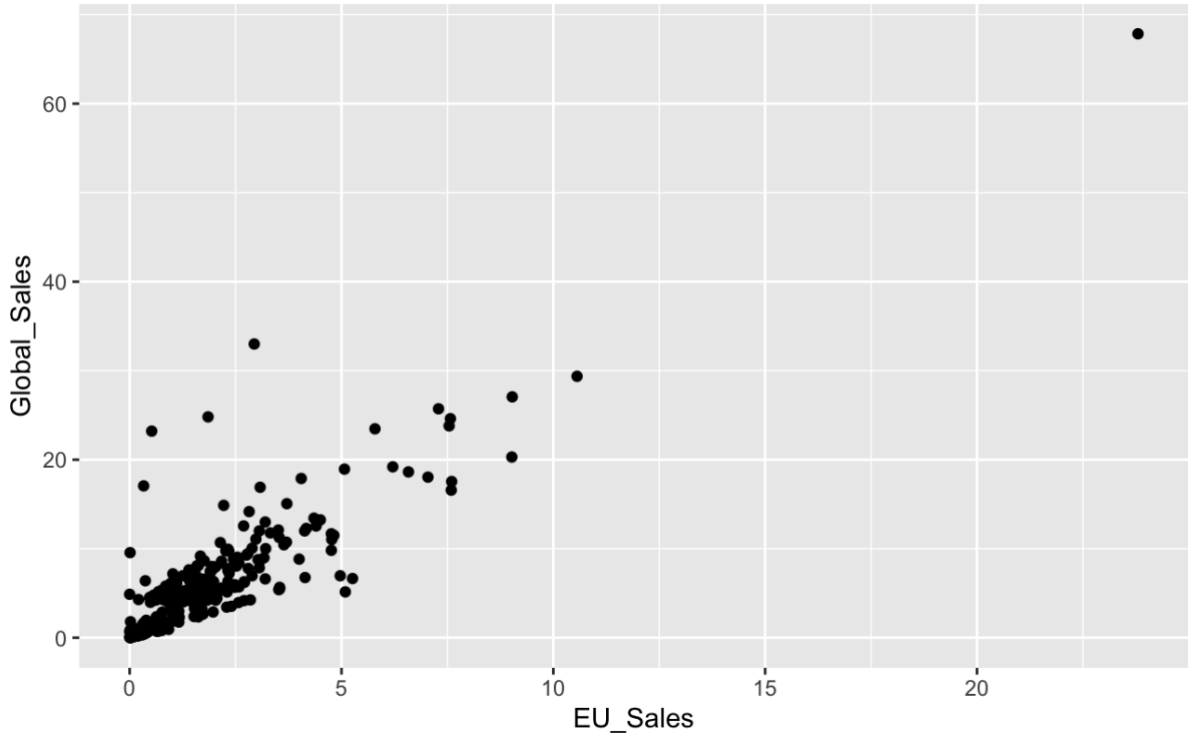
3.5 EDA in R

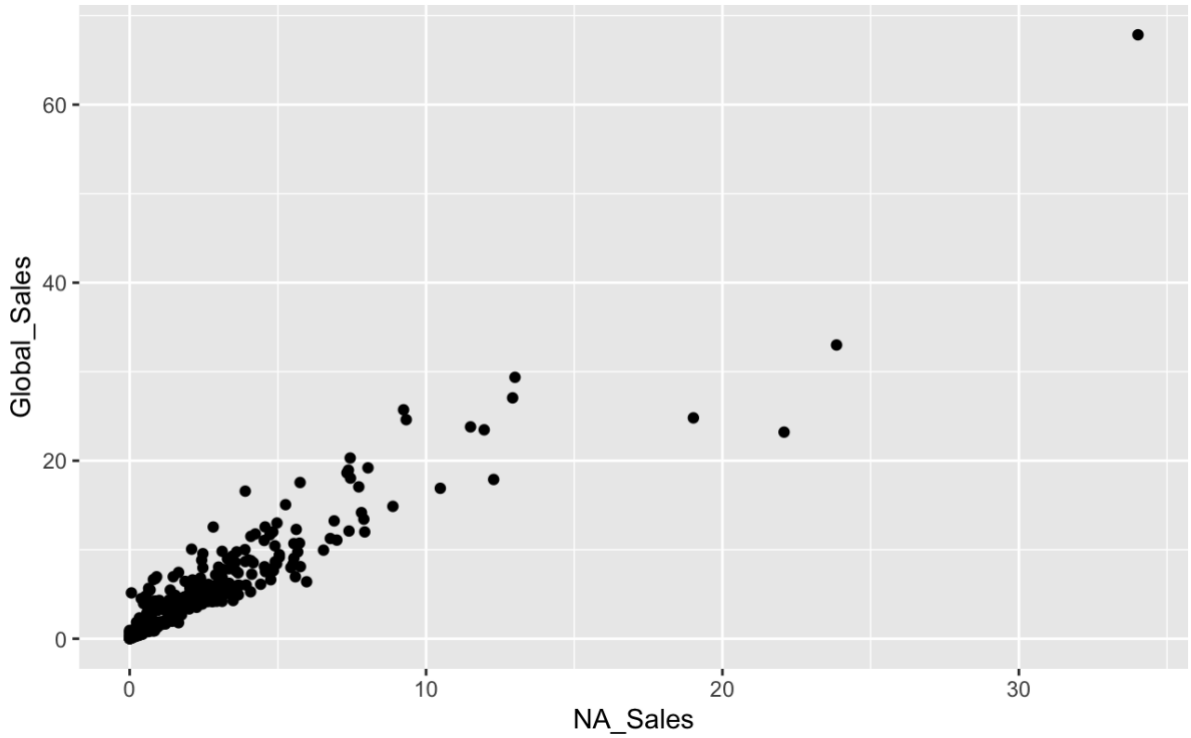
As an overview of this stage, the impact of individual products in Global Sales was found to be described almost as exponential, with a small amount of product contributing with high figure sales and the majority of the remainder products almost equally contributing to sales.

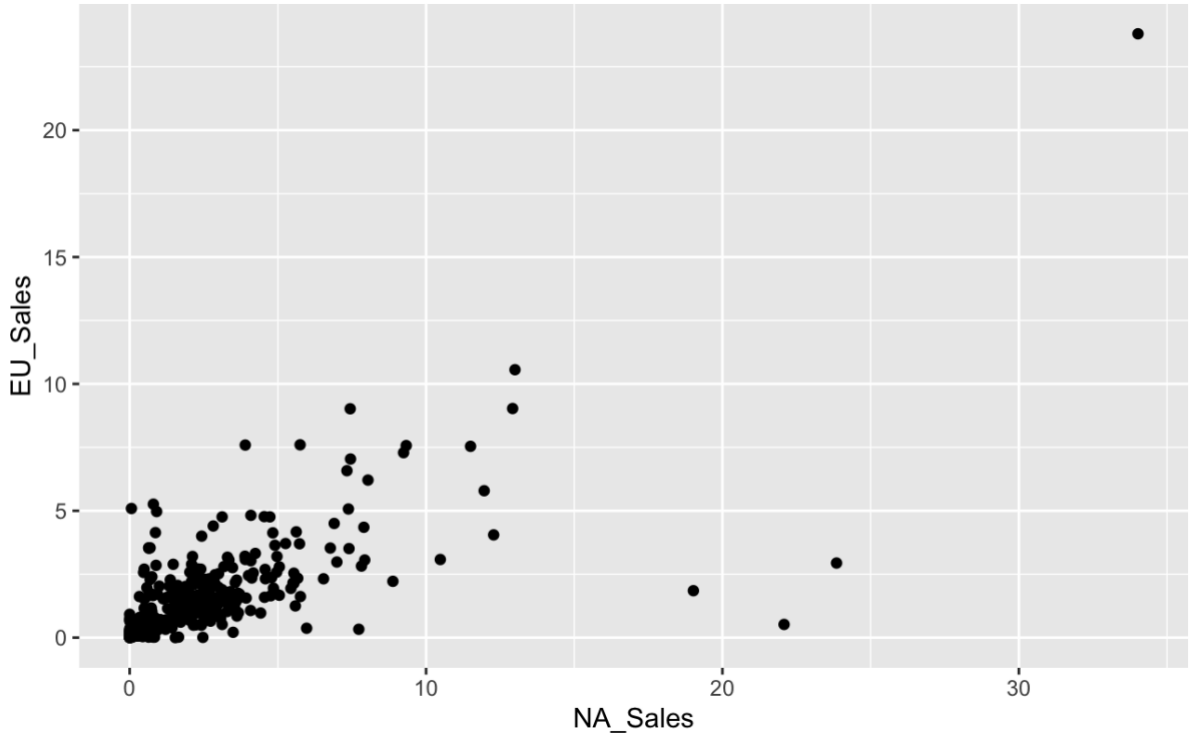
3.5.1 Scatterplots out of Sales Data, looking into NA, EU and Global Sales.



Furthermore, at this stage it was observed that Sales across regions and Globally are positively correlated so that more sales in EU will potentially mean the same in NA and vice versa.

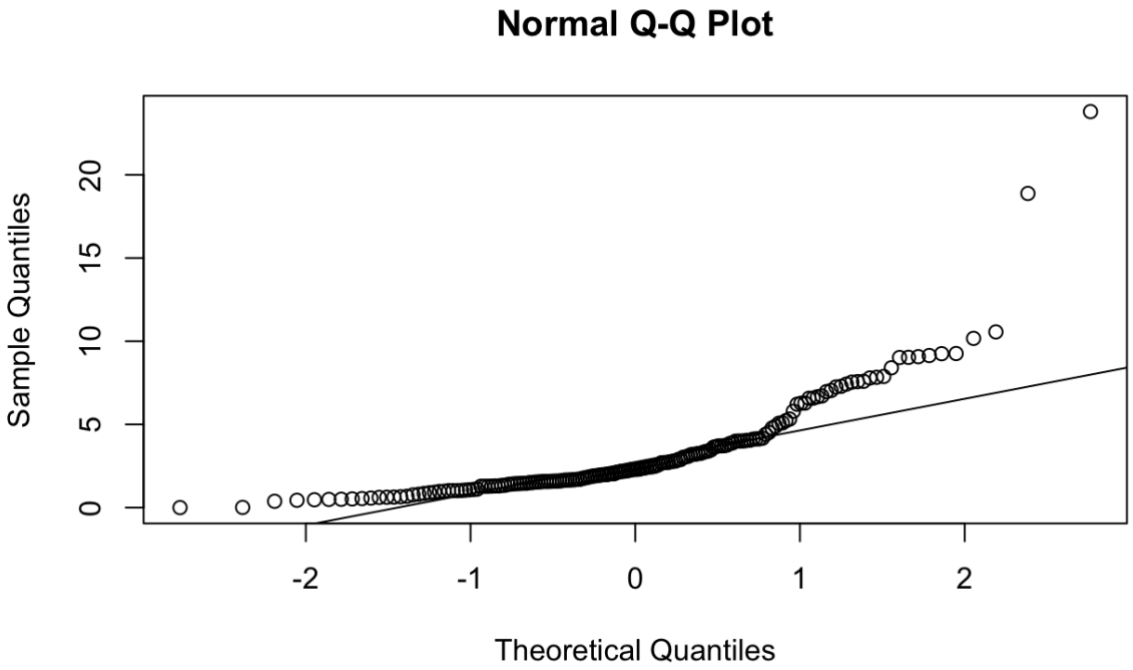

Global_Sales
60
40
20
0
0
5
10
15
20
EU_Sales


Global_Sales
60
40
20
0
0
10
20
30
NA_Sales

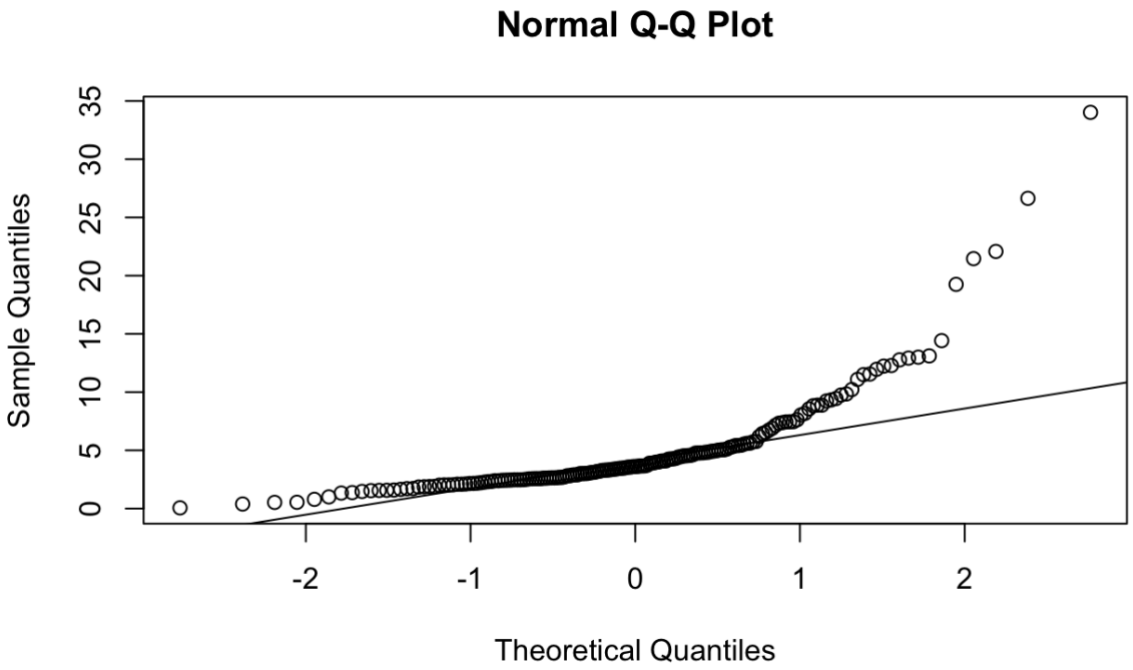

EU_Sales
NA_Sales
0
5
10
15
20
0
10
20
30

3.6 Q-Q Plots for each Sales Data column

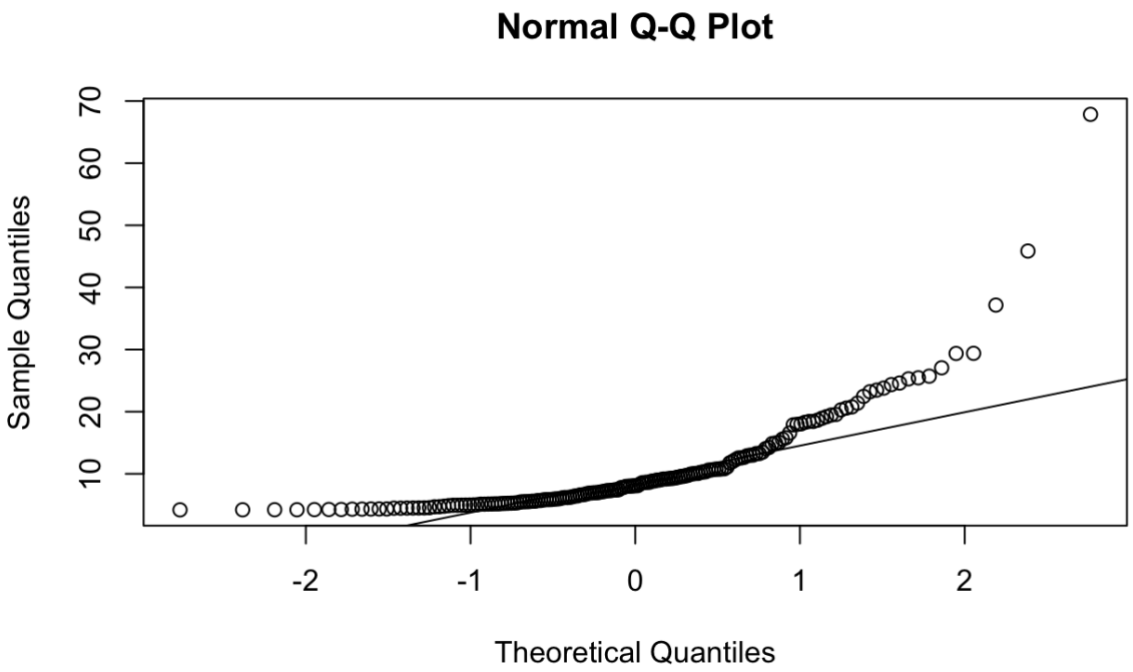
3.6.1 EU Sales data



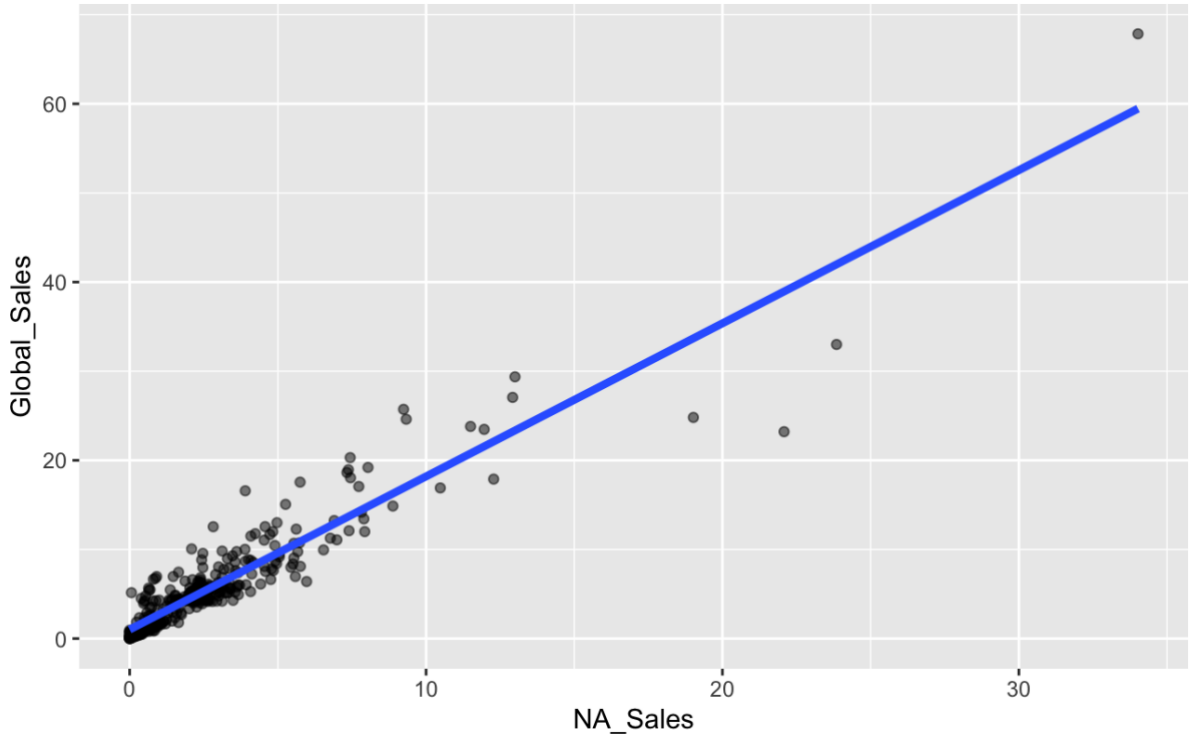
3.6.2 NA Sales Data



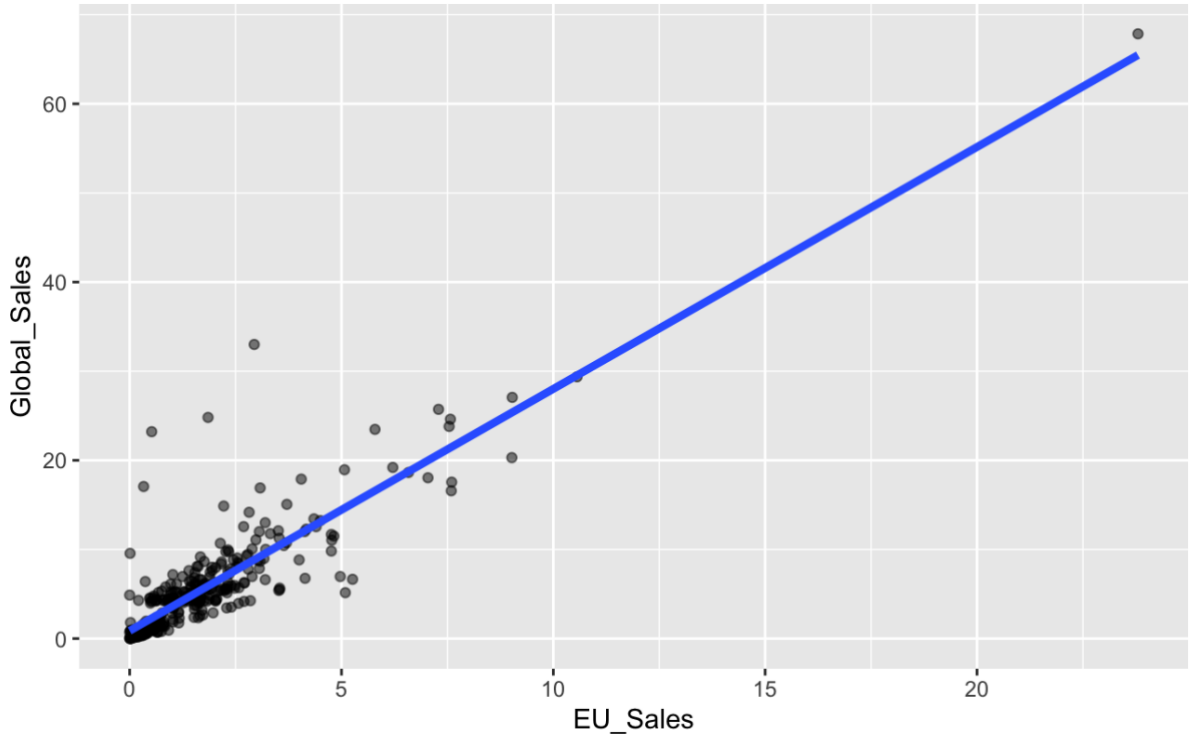
3.6.3 Global Sales Data



3.6.4 NA vs Global Sales



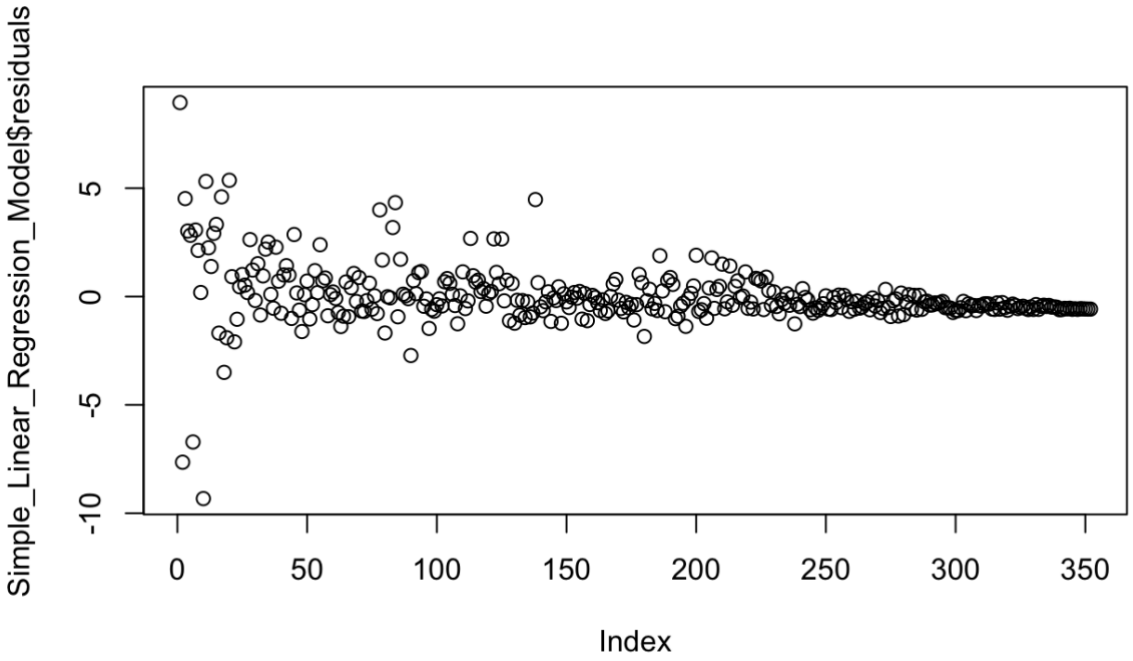
3.6.5 EU vs Global Sales



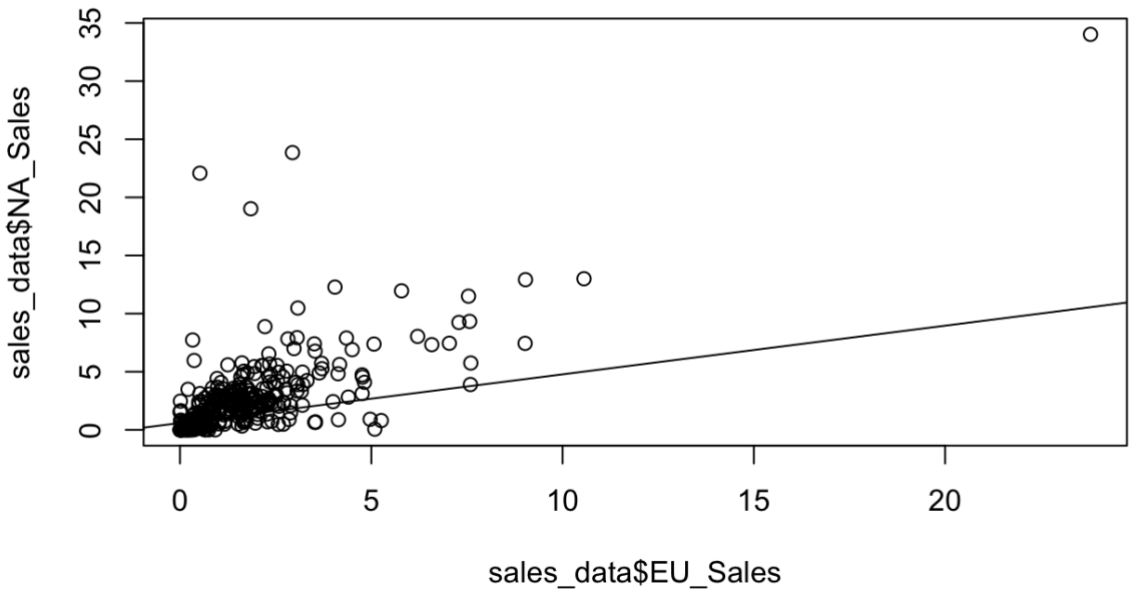
3.7 Linear Regression in R

At this stage of the project, it was possible to use data from EU and NA regions to predict the sales outcomes Globally. This was done using two variable linear regression function in R. The model was fited on the available numerical data and it was tested using a range of known data points from EU and NA Sales to predict the outcomes in Global Sales.

3.7.1 Simple Linear Regression Residuals



3.7.2 EU vs NA Relationship plot



4 Summary

The project consumed the available historical data from Turtle Games customers, their reviews and the overall sales per product, platform, and geography. Data exploration and initial analysis was done using both Python and R that identified key trends such as:

1. Loyalty tends to grow in customers who earn more and consequently spend more, though age seems to be loosely inversely correlated.
2. A small number of products seem to be the highflyers contributing to Sales more than the bulk of their products which seem to be equally contributing. This shines light to the products worth prioritising and creates room to consider reducing part of the product range while investing in their best performing products. This will potentially have a play in cost vs benefit and is worth exploring further.

Furthermore, Python's powerful libraries enabled the project to perform clustering and natural language processing. The first indicated:

1. 5 customer segments which differ in terms of spend vs income, an important input when it comes to customer segmentation and targeting during marketing campaigns.
2. Flagging the highest spending segmentations as the targeted area, the business can deploy tactics that will push eligible customers to enter and stay in these areas.

At the same time, NLP was used to identify:

1. The most frequently used words being in the realm of expectation, signalling to the business that their brand around games is strong.
2. Customer review sentiment was identified and tracked down to allow the business to recover the features that result in certain customer experience. This is a powerful tool to be taken in by the business to shine on their best features and improve on the not so good ones.

Sales data allowed the project to construct a R-based model which showed that:

1. EU and NA Sales are positively correlated with Global Sales, indicating that what sells in one geographic region does so in another
2. Also that the business can predict Global Sales expectations by looking into historical data from the two other regions.