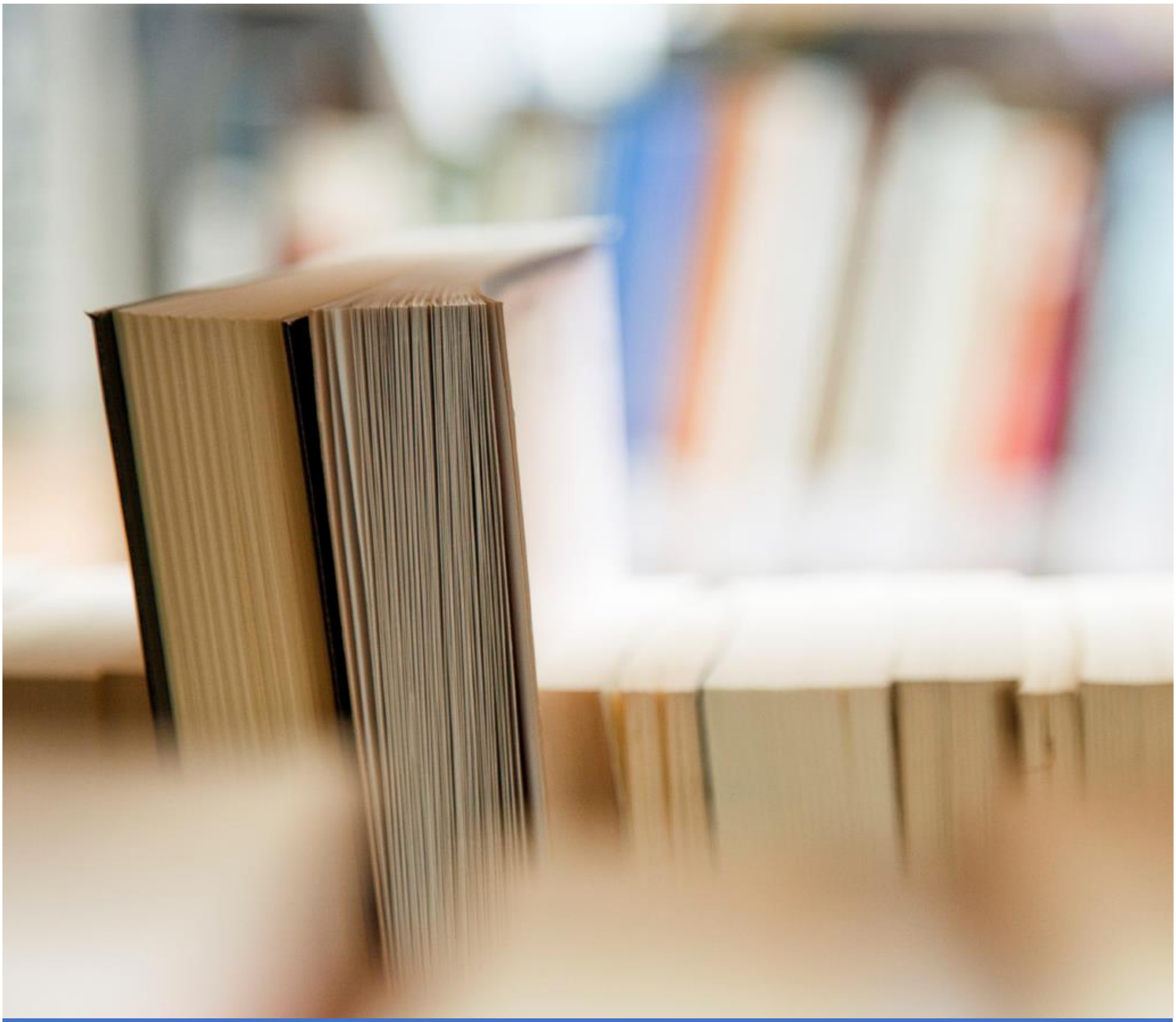


Assignment: Predicting future outcomes

Course 3:

Natasha Sutton

4/17/23

Advance Analytics for Organisational Impact

S. No.	Content	Page No.
1.0	Introduction/background	2
2.0	Analytical Approach	3
3.0	Visualisation and Insights	5
4.0	Patterns and Predictions	15

1.0 Introduction/Background

Our data team were contracted by Turtle Games, global manufacturer and retailer of games to help improve their sales performance by utilising customer trends. The 6 step problem solving process was used along with Problem-funnel to explore areas to focus.

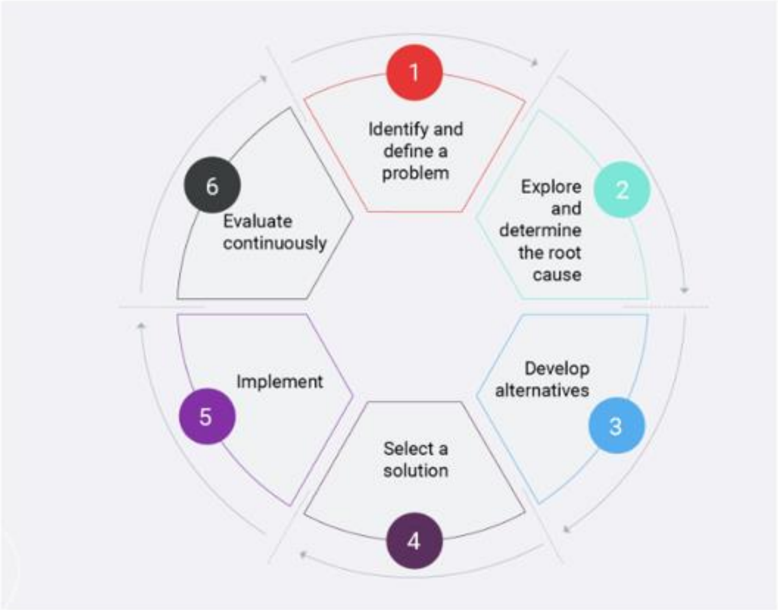
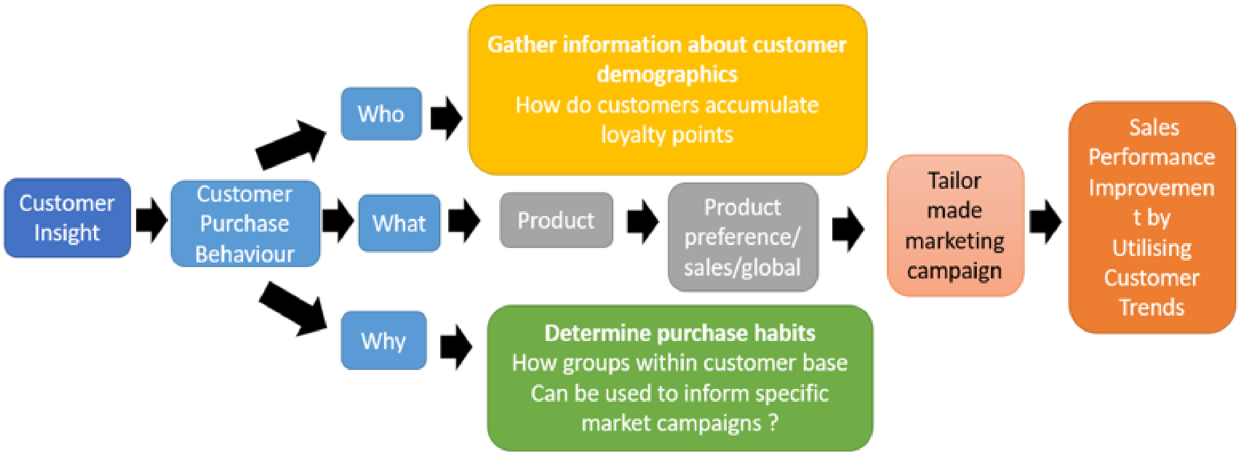


Figure 1 *Steps in the problem-solving process*

The games manufacturer's goal is to increase sales by consumer insight derived by the data analysis. The analysis would deliver information about the customer demographics (age, gender, income) and customer purchase behaviour (spending score, products) by using loyalty points (reward scheme) as an anchor


Customer Insight
Customer Purchase Behaviour
Who
Gather information about customer demographics
How do customers accumulate loyalty points
What
Product
Product preference/ sales/global
Why
Determine purchase habits
How groups within customer base Can be used to inform specific market campaigns ?
Tailor made marketing campaign
Sales Performance Improvemen t by Utilising Customer Trends

Structured Thinking : The Problem Funnel

Figure 2: The Problem Funnel

.

2.0 Analytical Approach

2.1 Analysis in Python

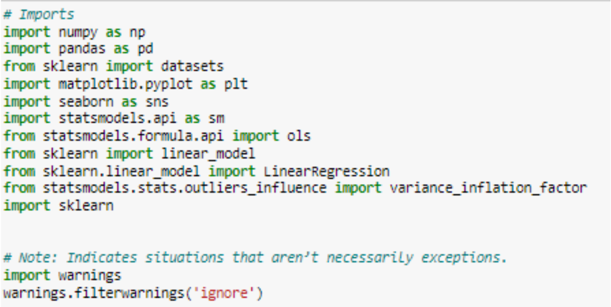
File turtle_reviews.csv .

Validated data: info(), stats describe(), duplicates duplicated(), missing values .isna().sum(). Dropped unwanted columns: language-platform. Columns renamed, .rename(columns={} The cleaned data saved as csv file.

Histogram and Q-Q plot to identify distribution of loyalty data.

2.1.1 Regression

Imported libraries:



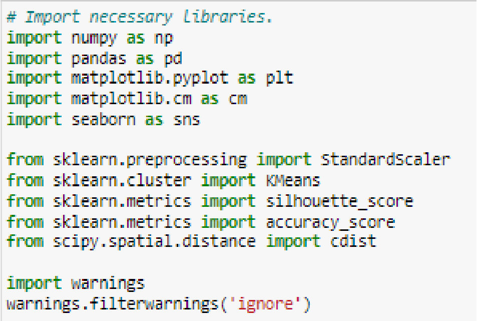
```
# Imports
import numpy as np
import pandas as pd
from sklearn import datasets
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.formula.api import ols
from sklearn import linear_model
from sklearn.linear_model import LinearRegression
from statsmodels.stats.outliers_influence import variance_inflation_factor
import sklearn


# Note: Indicates situations that aren't necessarily exceptions.
import warnings
warnings.filterwarnings('ignore')
```

corr()to visualise correlation between variables: loyalty points, income, spending scores and age. Fitted the ols regression model ols(f, data).fit to determine coefficients, intercept R^2 , p values. The predicted residuals and Q-Q plot checked for distribution for linear regression assumptions[1]. Multicollinearity and Breusch-Pagan test function for homoscedasticity[2]. The data split into test/train and actual vs predicted values assessed.

2.1.2 Clustering

Imported libraries:



```
# Import necessary libraries.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import seaborn as sns

from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.metrics import accuracy_score
from scipy.spatial.distance import cdist

import warnings
warnings.filterwarnings('ignore')
```

Data File:turtle_reviews.csv.

Datafile “turtle_reviews_clean_log.csv” review columns dropped and rearranged. Explored data with info(), describe() functions and visualisation with pairplot(). Identified two variables: Score and Income to provide insight into consumer behaviour and visualised using scatterplot. No correlation with Product variable. Used K-means

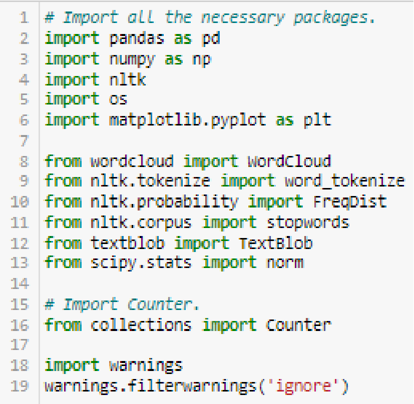
[1] LSE:CO4_LSE_DA_301_ section 6.2.8

[2] Understanding Heteroscedasticity in Regression Analysis - Statology

clustering to group customer into segments based on income and spending scores. K =5 determined via Elbow and Silhoutte methods.

2.1.3 Common Words & Reviews

Targeted marketing campaign



```
# Import all the necessary packages.
import pandas as pd
import numpy as np
import nltk
import os
import matplotlib.pyplot as plt

from wordcloud import WordCloud
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist
from nltk.corpus import stopwords
from textblob import TextBlob
from scipy.stats import norm

# Import Counter.
from collections import Counter

import warnings
warnings.filterwarnings('ignore')
```

The "turtle_reviews_clean_log.csv" dataframe, explored. All columns except review and summary deleted, no missing values identified.

Analysis:

- apply(lambda x: " ".join(x.lower() for x in x.split() to change text to lower case join them for both -Replaced punctuation with .str.replace ('[^\w\s]','')

- Dropped duplicates (39) that appeared in both columns simultaneously with duplicated()

-Tokenised using .apply(word_tokenize) and created word cloud using WordCloud loop: with & without common words using set(stopwords.words('english')).

- Frequently used word determined by FreqDist() importing FreqDist from nltk.probability,

- Sentiment analysis conducted by .apply(generate_polarity) and .apply(generate_subjectivity),

-1 is strongly negative and +1 is strongly positive[3].

- Determined top 20 positive, .nlargest() and negative .nsmallest() statements.

2.2 Analysis in R

North America (NA) Europe(EU)

Libraries imported:tidyverse, reshape2, gridextra, psych, moments, metrics
Data: “turtle_sales.csv”

Data cleaned by removing two rows with missing year data -na.omit(), explored with head(), dim() rows and column size and summary() statistical summary. Ranking column removed.

[3] Positivity Calculation using Vader Sentiment Analyser, from the international journal of Academic Engineering research

The sales data normality investigated with qqnorm() and qqline(). Shapiro.test() to test the normal distribution. Skewness, skewness() and Kurtosis, kurtosis() indicated that sales of three regions were positively skewed with heavy tail.

Sales data grouped by product to identify impact of sales on 175 products. Outliers, high performing products, weren't eliminated.

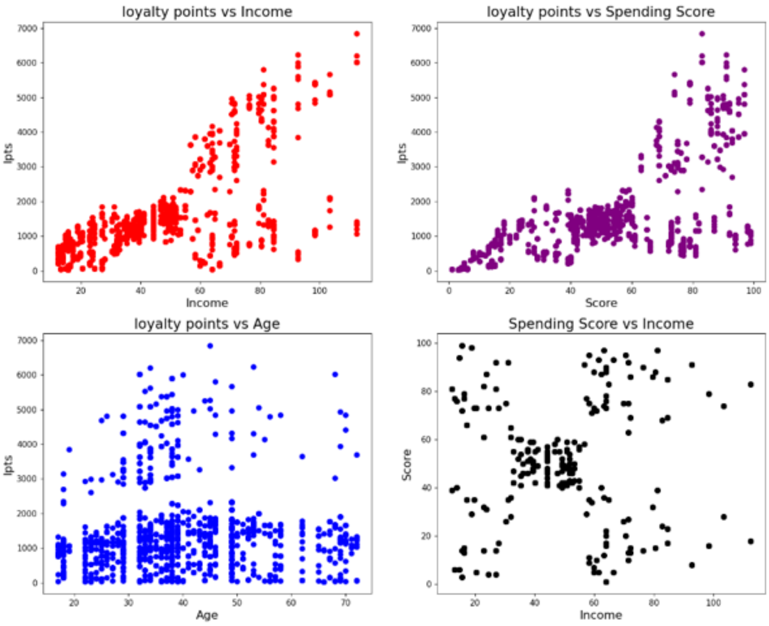
Multiple linear regression conducted on global sales and regional sales with aid of Corplot(). NA sales and EU sales independent variables and Global sales dependent variable. A set of NA and EU values used to test model which produced mean absolute percentage error (MAPE) of 10.3%, 10% is excellent[4].

3.0 Visualisations and insights

3.1 Regression

Correlation plot visualises the relationship between the variables and aids in effective selection of variables for regression analysis. Figure 3 shows poor correlation between age and loyalty points

Figure 3 Correlation Plot for Variables: Loyalty Points, Spending Score, Income, Age



The residual Q-Q plot fig 4, of the MLR analysis of loyalty points and variables: income, score and age, indicates a normal distribution[5]. The plot of actual vs predictive loyalty points showed a good linear correlation above actual loyalty point >500, figure5.

[4] LSE:CO4_LSE_DA_301_ section 6.1.3

[5] Understanding and interpreting Residuals Plot for linear regression

Figure 4 Residual Q-Q plots of MLR

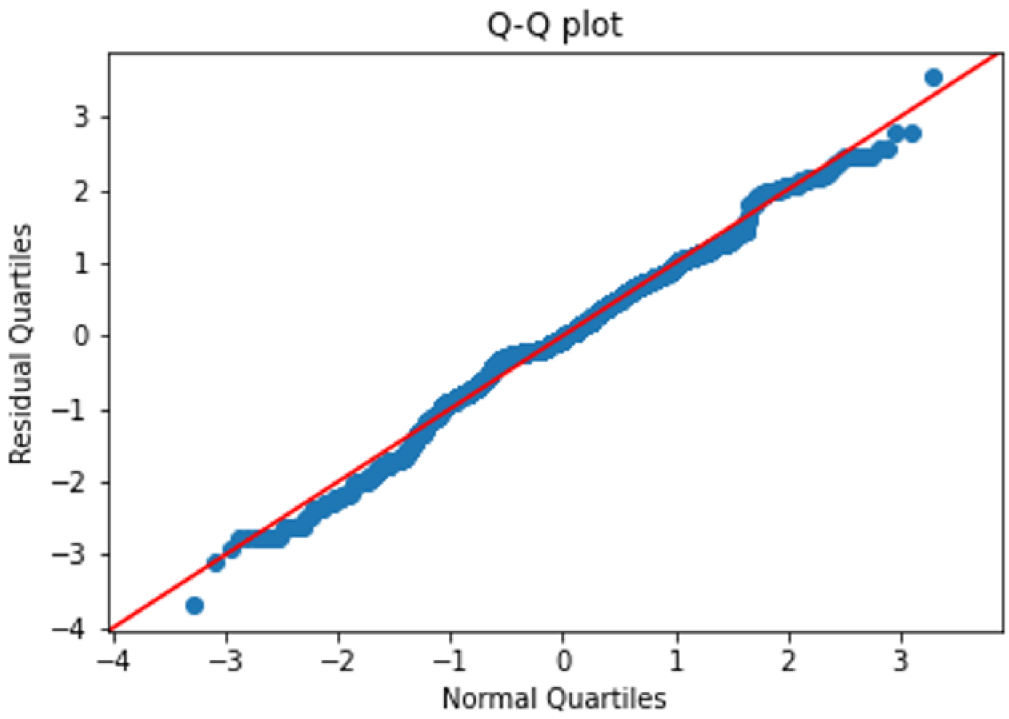
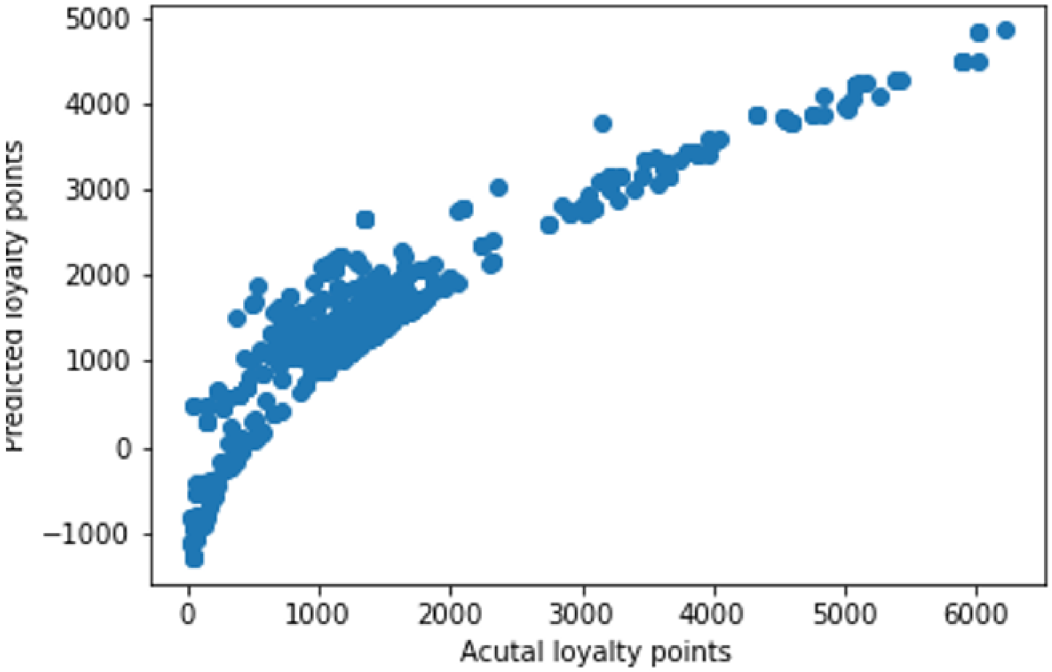


Figure 5 Predicted vs Actual Loyalty Points



3.2 Cluster Analysis

Cluster analysis to understand customer spending habits, figure 6,   5 consumer groups based on income and spend. Table3 defines the cluster groups.

Figure 6

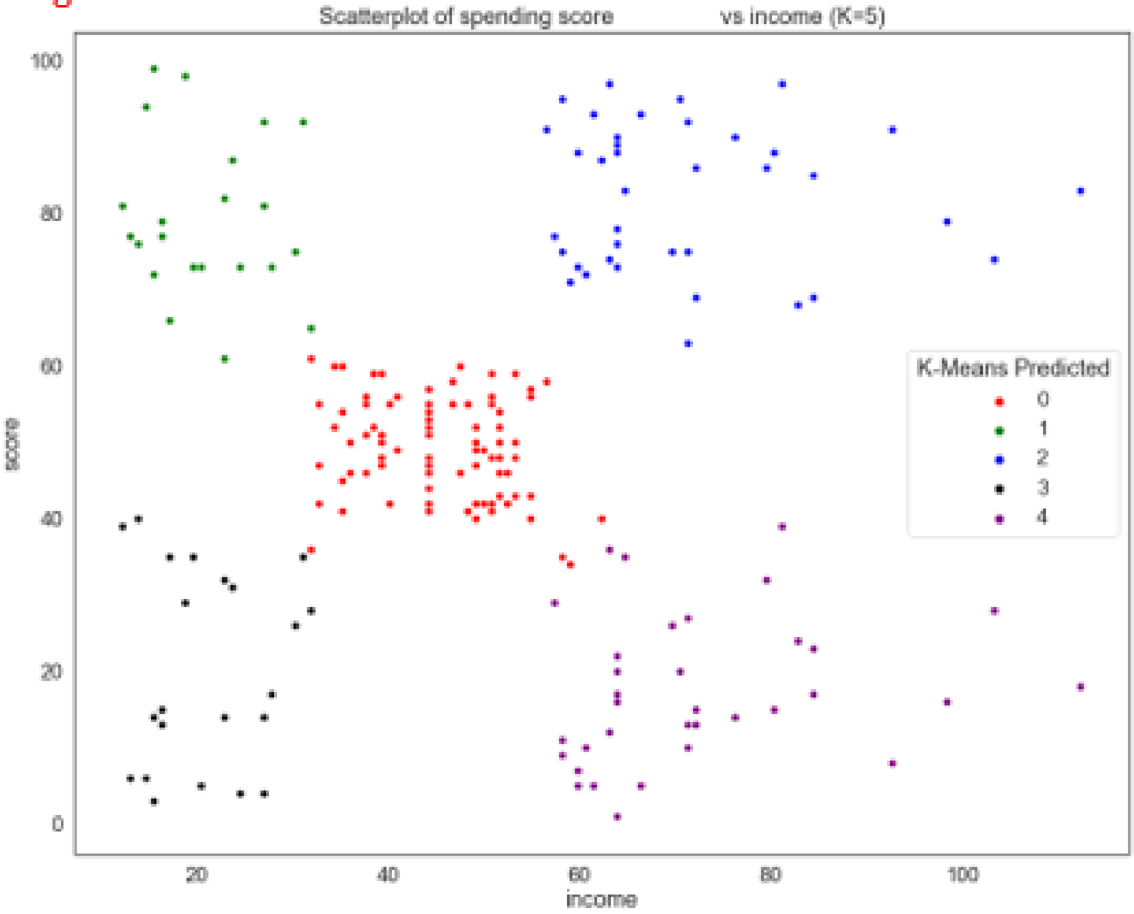


Table 3 Consumer Clusters based on Earning and Spending Behaviour

Cluster no	Income	Spending Score
Cluster 3	Low	Low
Clusters 1	Low	High
Cluster 0	Average	Average
Cluster 4	High	Low
Cluster 2	High	High

3.3 The Consumer Review

Identified top frequently used words could be visualised in a word cloud as shown in the table4 below. Figures 7/8 show ascending frequency of the popular 15 words used in customer reviews which were all positive.

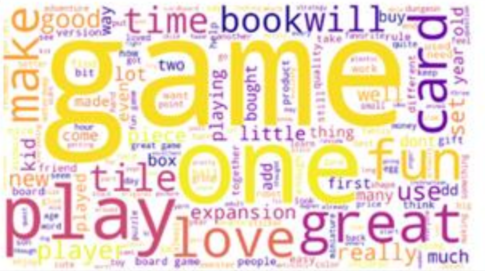
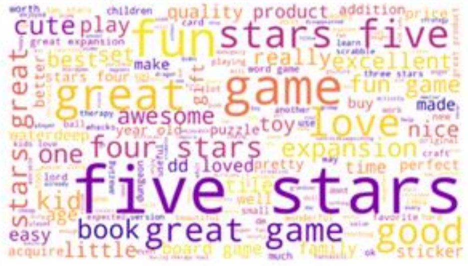
Table 4: Word Cloud Generation		
Analysis	Review	Summary
Word Cloud		

Figure 7

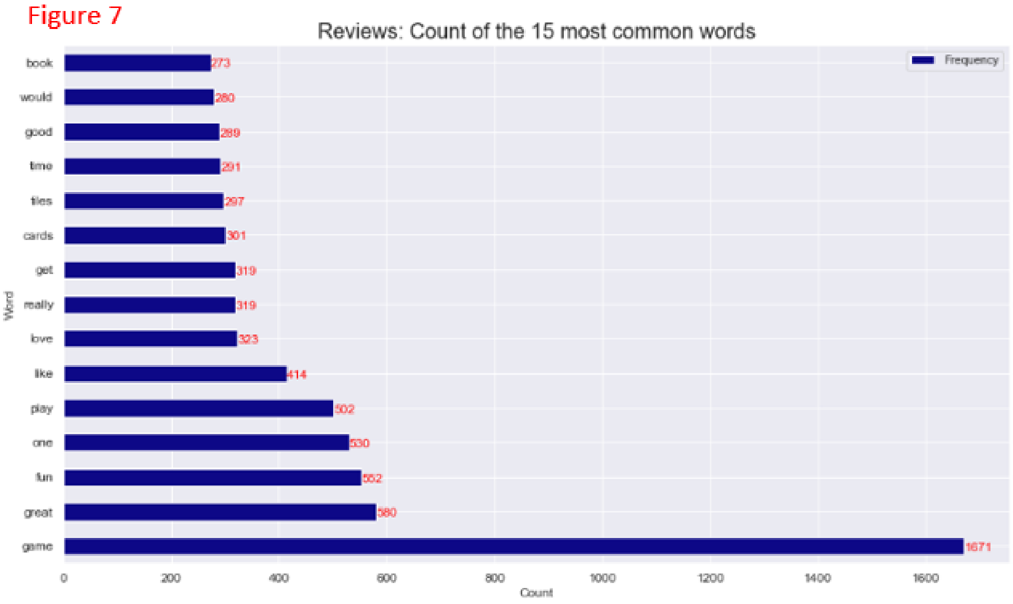
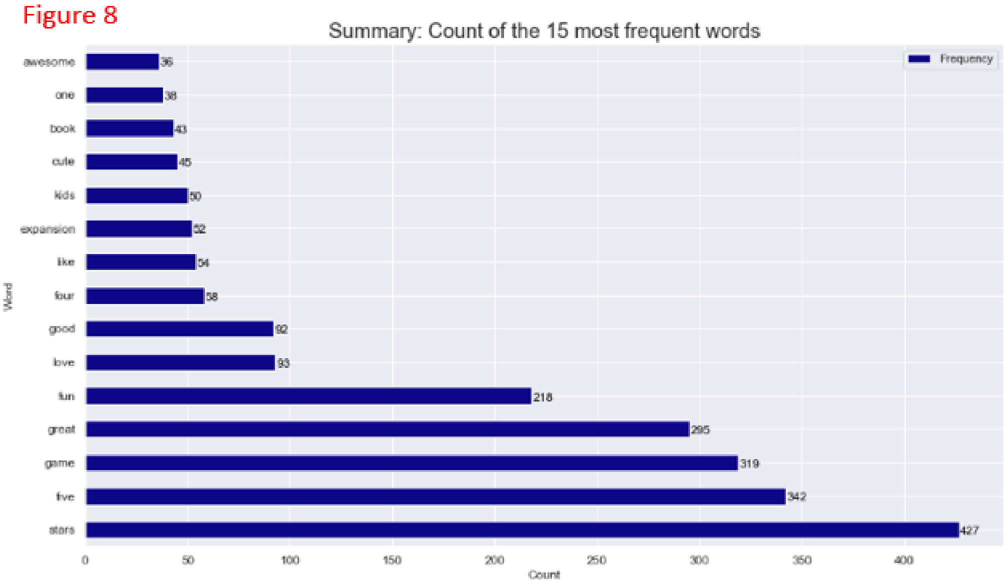
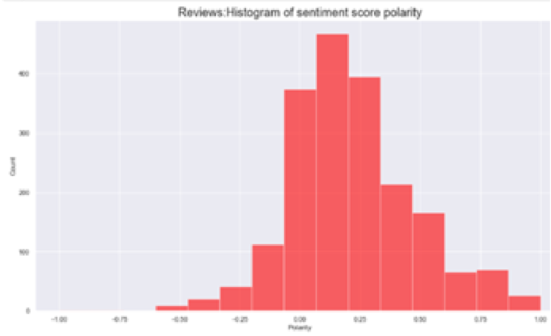
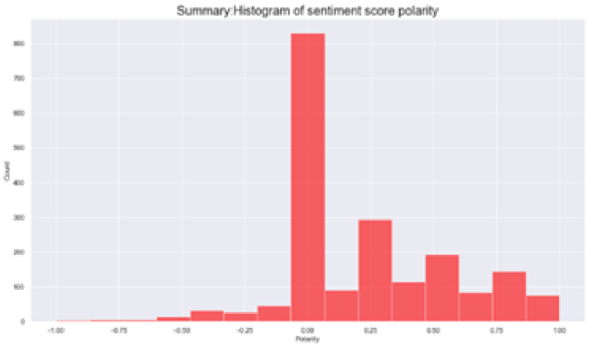
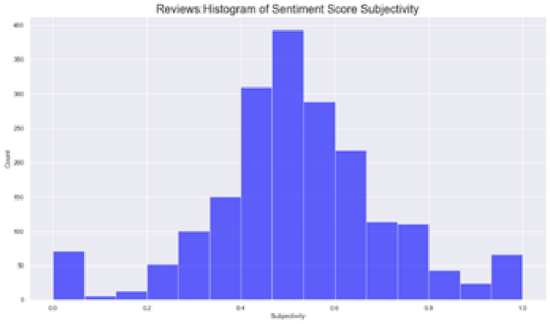
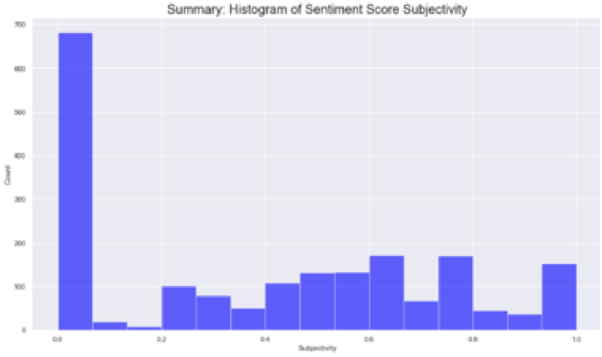


Figure 8



Sentiment analysis on customer review and summary are summarise in table5 below:

Table 5: The Reviews and Summary – Polarity and Subjectivity illustration

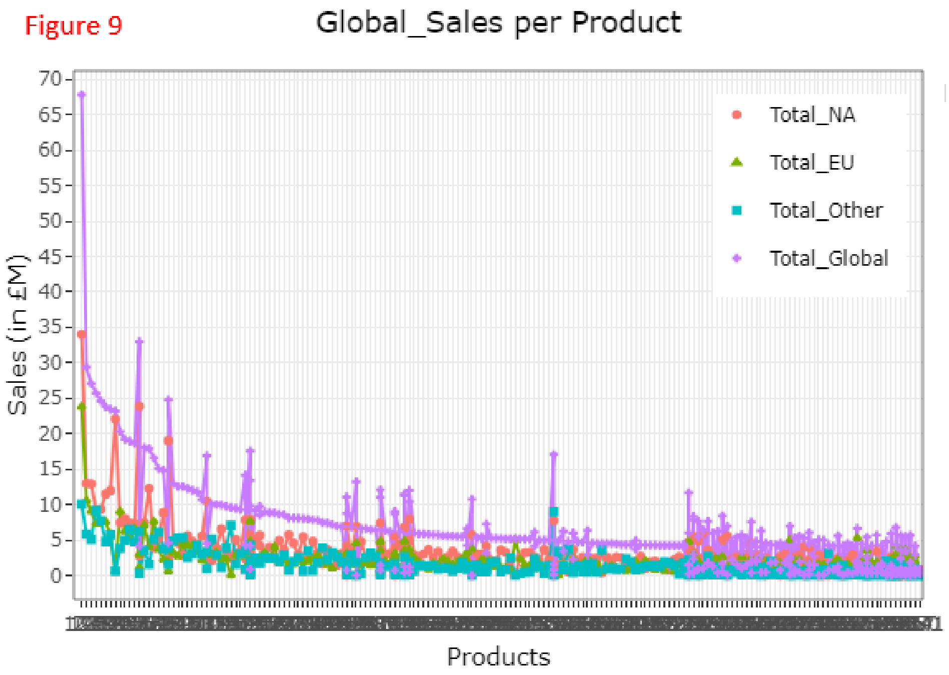
	Reviews	Summary
Polarity Score	 Reviews:Histogram of sentiment score polarity 	 Summary:Histogram of sentiment score polarity
	polarity with highest between 0-0.25, neutral to positive sentiment.	The highest frequesncy is at 0.0, neutral sentiment
Subjectivity Score	 Reviews:Histogram of Sentiment Score Subjectivity 	 Summary: Histogram of Sentiment Score Subjectivity
	subjectivity highest between 0.4-0.6, suggesting a reviews were balanced	Skewed towards right, indicating most comment were objective.

3.4 Sales and product correlation - R statistics

Impact of Sales per product

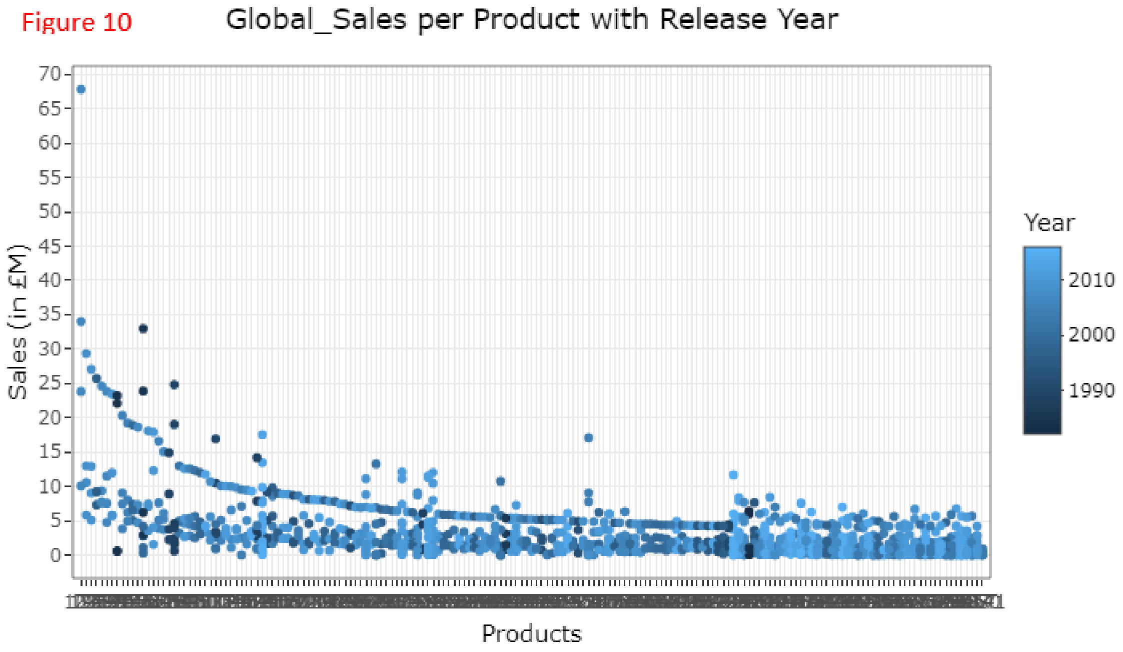
Sales of products by regions shown in figure 9 for 175 products, show outliers which were high performers.

Figure 9



By adding the year as a variable figure 10, indicate high performing products had been released a while.

Figure 10



Relationship between sales and Genre

Figure 11 illustrate popularity of genre in descending order and variation of sales. Figure 12 shows the genre preference varies by region: Shooter in NA, and Role-Playing in other regions

Figure 11

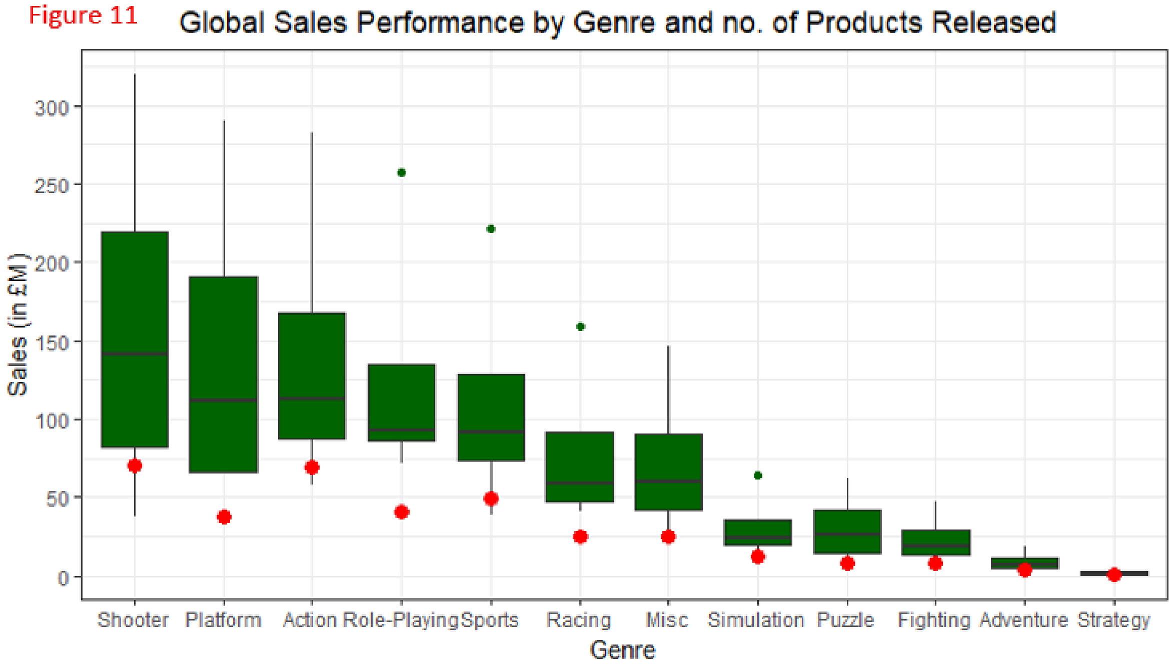
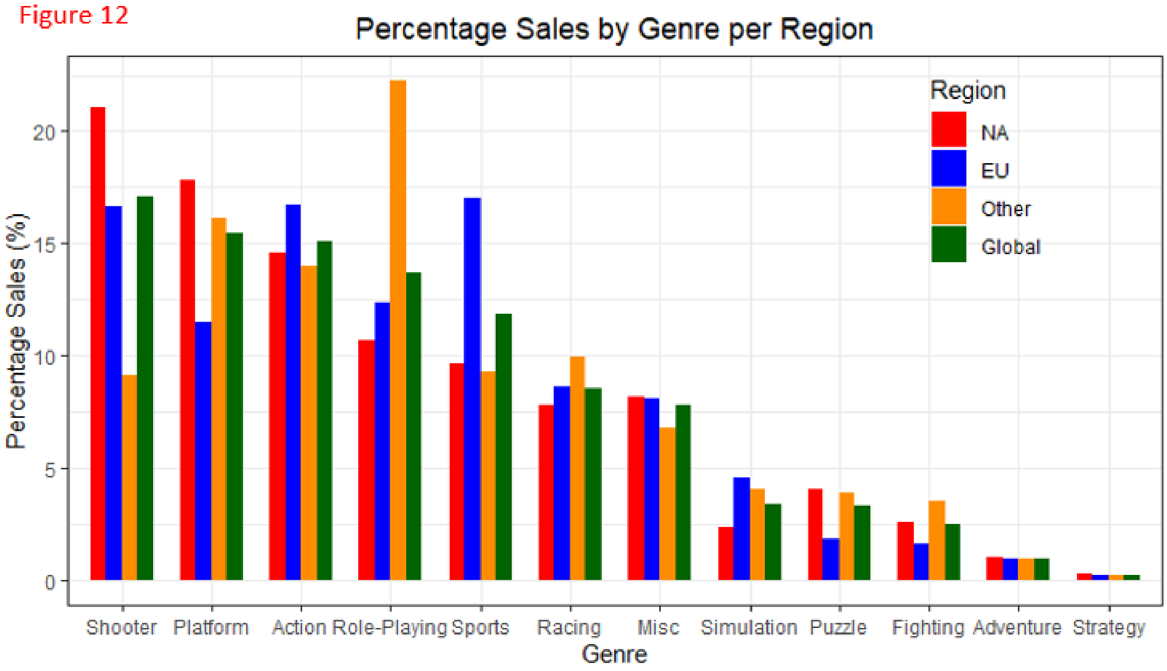


Figure 12



Relationship between Sales and Platform

The platform popularity in descending order shows Wii top performer globally and in NA, Figures 13 & 14. The platform preference also varies by regions.

Figure 13

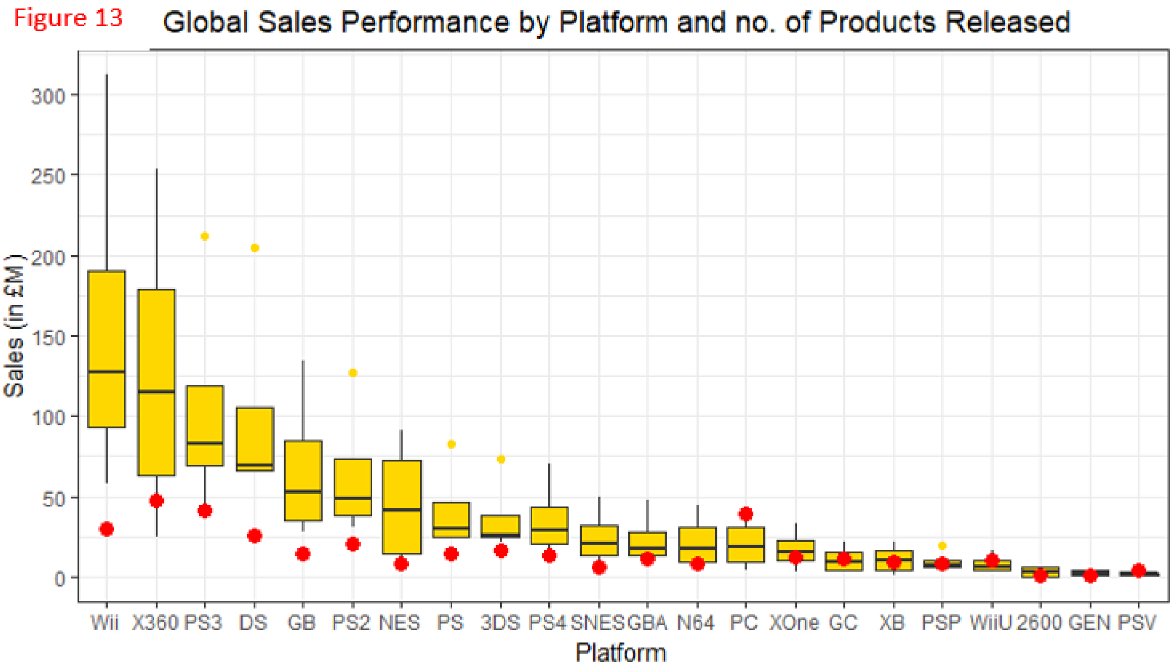
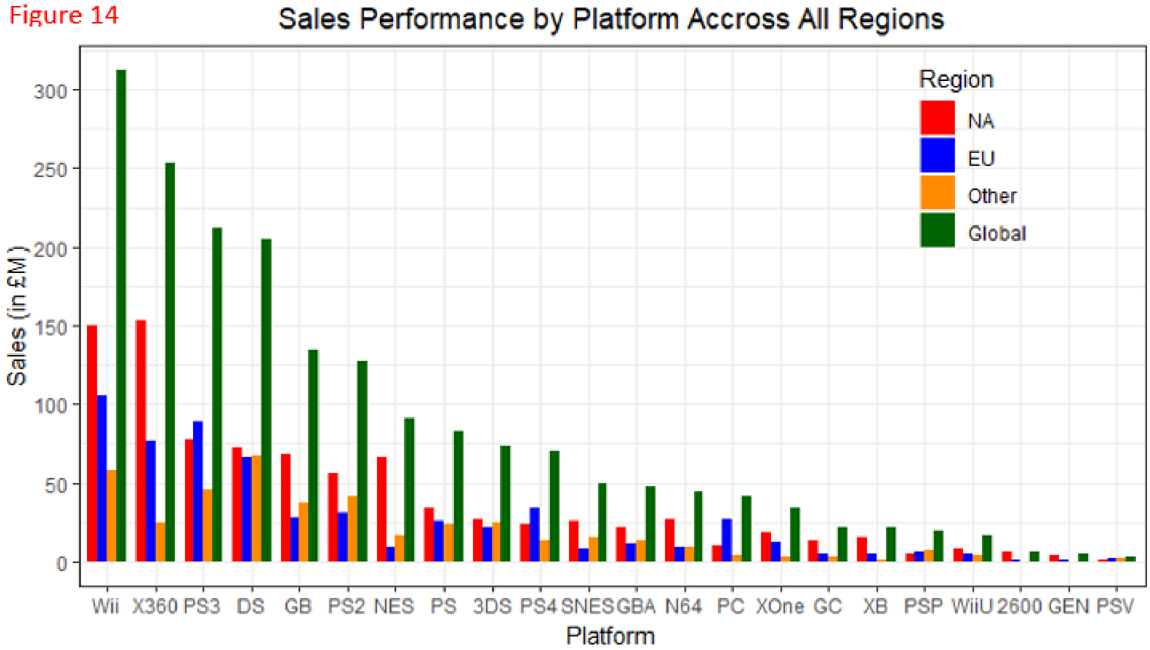


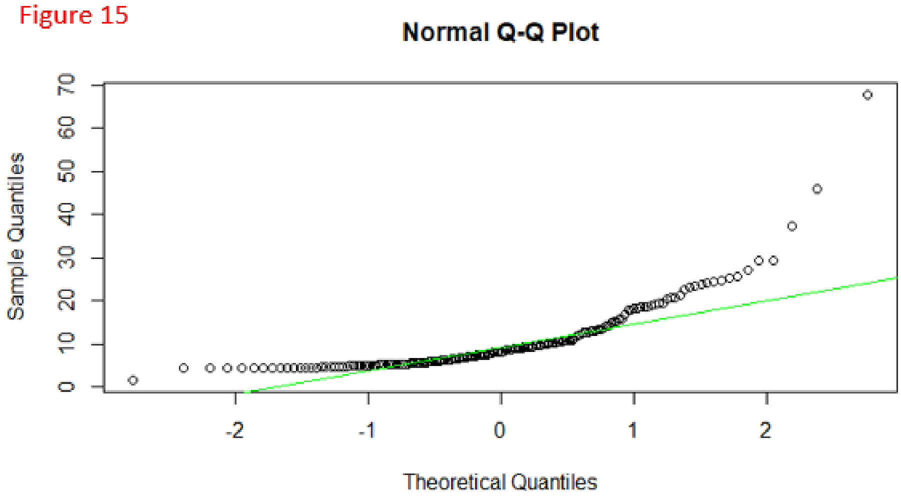
Figure 14



Reliability of data

Data visualised to check for normality qqnorm and qqline plots. Plot shows data on the line but not tail ends. NA and EU sales data follow the trend as in Figure15 with asymmetry and heavy tail, positive skewness.

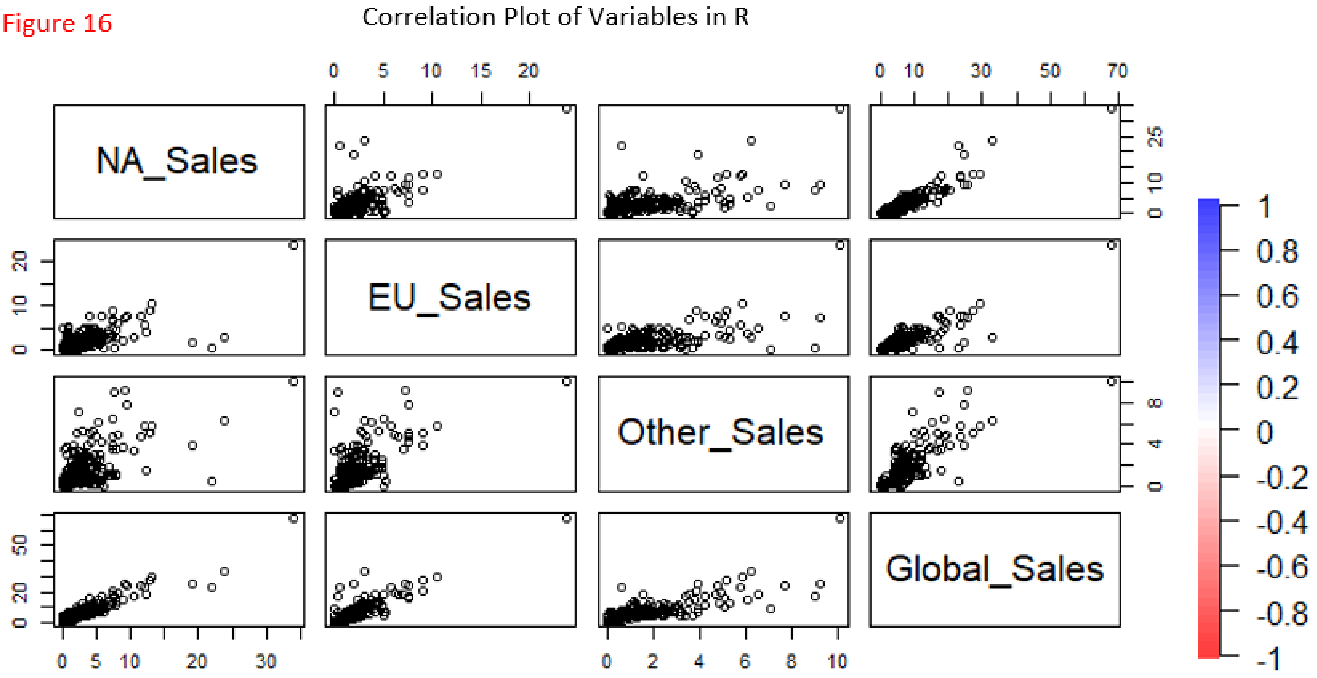
Figure 15



Correlation between sales of NA, EU and Global Sales

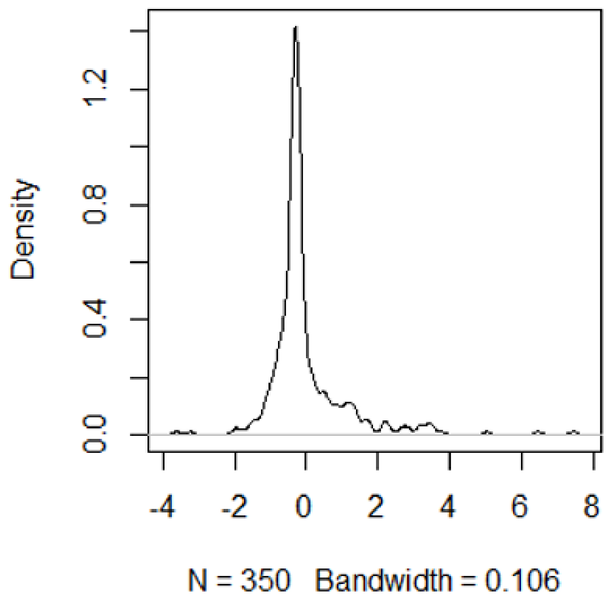
Correlation plot between variables shown in figure 16 and 17 for MLR model.

Figure 16



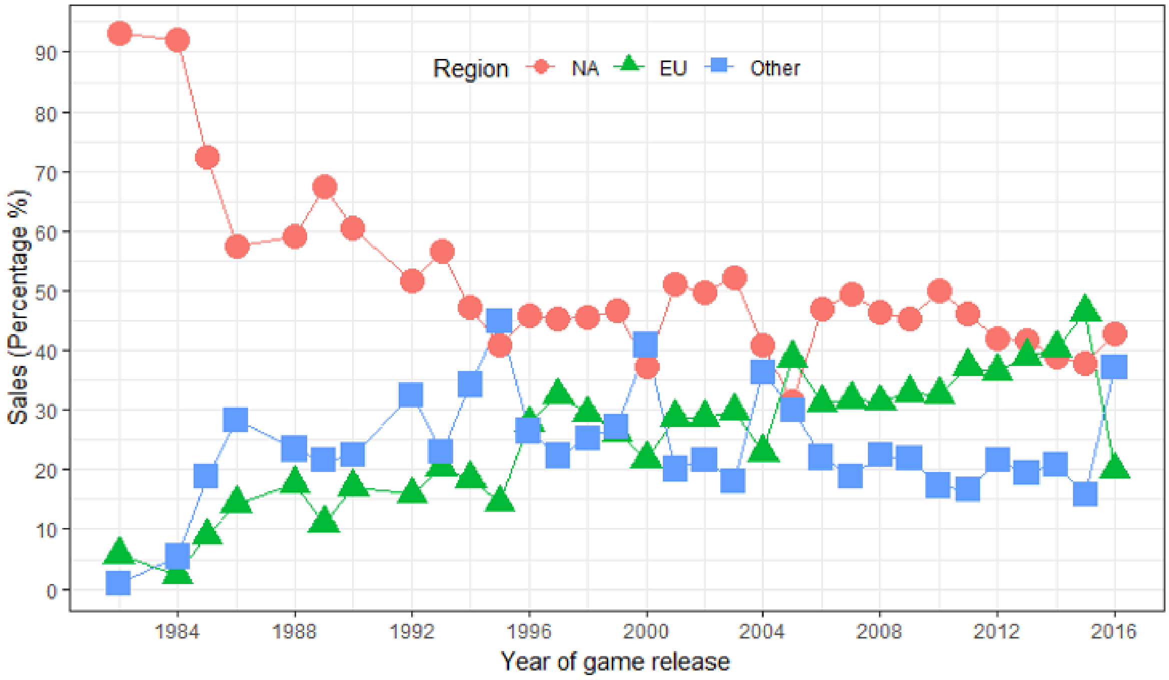
The plot density of the residual of MLR model on sales showed a bell shaped curve with positive skewness, figure 18 and a MAPE of 10.3% which considered acceoptable. However, since data isn’t normal distribution, it cannot be used to make any sales predictions.

Figure 18 **density.default(x = modelSales4.res**



Following Chart-figure19 shows the increasingly sales in EU and other regions by year.

Figure 19 Percentage of Global Sales per Region with Game Release Year



4.0 Patterns and predictions

1. MLR model to predict loyalty points although had good correlation cannot be used due to heteroscedasticity.
2. MLR model to predict global sales showed high correlation with NA and EU sales but the data was not normally distributed thus cannot not be used.
3. The cluster analysis showed 5 consumer groups based on their earning and spending behaviour. Marketing can be targeted towards high earners.
4. Further market research to understand preferences and behaviours high earners is recommended.
5. Customers in different regions prefer different types of products so marketing campaign needs to be streamlined accordingly.
6. Customer reviews showed generally sentiment towards the product are positive.
7. Marketing needs to focus on product quality, making products and instructions user friendly.
8. Real time, yearly sales data is a better metric to understand product performance not product release year.
9. The sales by product release year show a decline towards the end as new products had just been released.
10. Initially global sales were dependent on NA and EU sales but the sales on other regions are steadily growing.
11. Any predictions made using models from this data would be obsolete as the latest release year was 2016