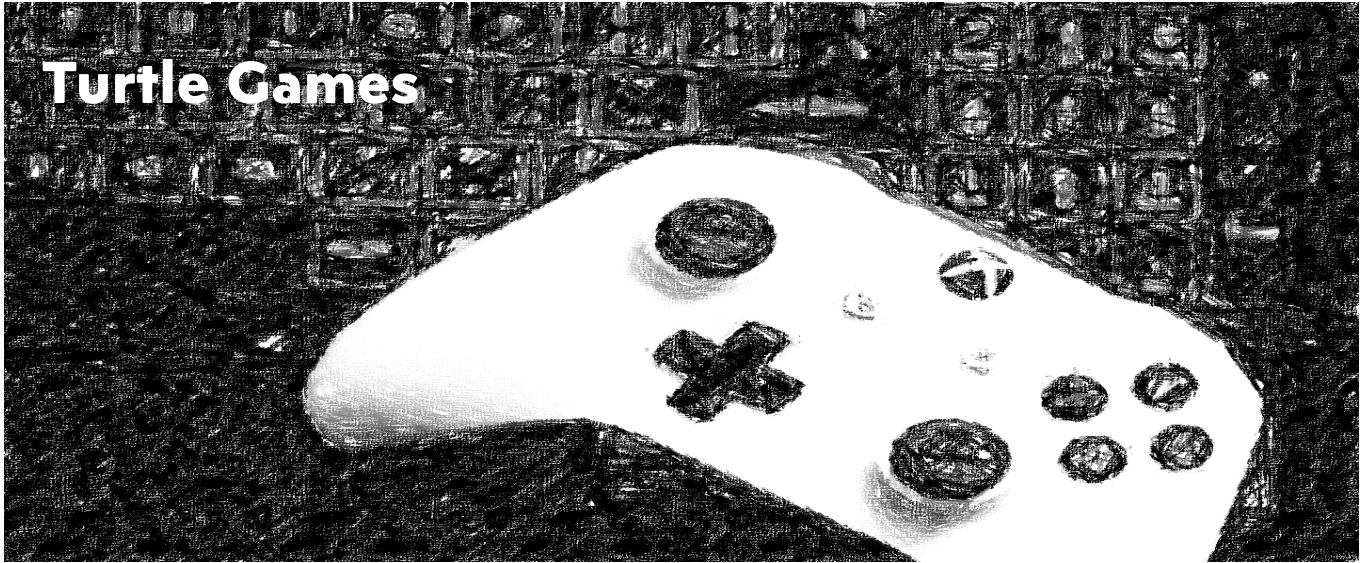


Turtle Games



Assignment Course 3: Predicting Future Outcomes

TURTLE GAMES

Harsh Bhatia

LONDON SCHOOL OF ECONOMICS & POLITICAL SCIENCES | ONLINE DATA ANALYTICS
CAREER ACCELERATOR

Table Of Contents

1.0 BACKGROUND.....	2
2.0 ANALYTICAL APPROACH	2
2.1 CONSUMER BEHAVIOUR & REVIEWS ANALYSIS USING PYTHON:	2
2.1.1 DATA WRANGLING & EXPLORATION:	2
2.1.2 LINEAR REGRESSION:.....	2
2.1.3 MULTIPLE LINEAR REGRESSION:.....	2
2.1.4 GROUPING OF DATA SET IN CLUSTERS	3
2.1.5 ANALYZING REVIEWS USING NLP	3
2.2 TURTLE GAMES SALES DATA ANALYSIS USING R:	4
2.2.1 DATA WRANGLING, EXPLORATION AND VISUALIZATION:.....	4
2.2.2 DATA MODELLING AND PREDICTION:	4
3.0 VISUALIZATIONS AND INSIGHTS	5
3.1 LINEAR REGRESSION ANALYSIS RESULTS:	5
3.2 MULTIPLE LINEAR REGRESSION ANALYSIS RESULTS:	6
3.3 K-MEANS CLUSTERING METHOD:	7
3.4 NLP – REVIEWS ANALYSIS:.....	8
3.5 VISUALIZATION FROM ANALYZING THE SALES DATA:.....	11
3.5.1 SCATTERPLOTS OF SALES BY REGION	11
3.5.2 DISTRIBUTION OF PER PRODUCT SALE PER REGION	12
3.5.3 MOST POPULAR PLATFORMS PER REGION (IN TERMS OF SALES CONTRIBUTION)	13
3.5.4 MOST POPULAR GENRES PER REGION (IN TERMS OF SALES CONTRIBUTION)	14
4.0 PATTERNS & PREDICTIONS:	15
5.0 APPENDIX	17
5.1 TOP 20 POSITIVE REVIEWS.....	17
5.2 TOP 20 NEGATIVE REVIEWS.....	18

1.0 Background

Turtle Games aims to improve its sales performance by analyzing customer loyalty, social data, and sales data across regions. The project has three main goals:

- 1) Understand loyalty point accumulation and its relation to age, spending behavior & remuneration
- 2) Analyze customer reviews to enhance marketing campaigns
- 3) Analyze past sales trends in North America, Europe, and globally and develop a model to forecast future sales.

Python and R will be used to analyze customer reviews and product sales data, respectively, in order to provide valuable insights.

2.0 Analytical Approach

2.1 Consumer Behaviour & Reviews Analysis using Python:

2.1.1 Data Wrangling & Exploration:

- We loaded the customer reviews dataset into our Jupyter notebook and created a data frame using `pd.read_csv` function from the pandas library
- We explored the data set statistics and cleaned it by looking for any missing values and dropping unnecessary columns like “Language” and “Platform”

2.1.2 Linear Regression:

- To understand the relationship between loyalty points and other factors like age, remuneration, and spending scores, we carried out simple linear regression between each of these variables
- We used the Ordinary Least Squared (OLS) method for this process by using the in-built `ols()` function.
- As the R-squared values of the individual relationships were not very strong, we went onto creating the multiple regression model to understand their combined effect

2.1.3 Multiple Linear Regression:

- We defined our independent variables and dependent variable separately and used the `LinearRegression()` function to create the multiple linear regression model.
- We trained our model by using the `train_test_split()` function by splitting our data set into 80% training data and 20% kept for testing.
- To confirm that the model is accurate and reliable, we conducted the multicollinearity test using the `vif()` function and tested the heteroscedasticity using the Breusch-Pagan test- `sm.het_breuschpagan()`

2.1.4 Grouping of Data set in Clusters

- Since spending score and remuneration had high impact on loyalty points, we decided to understand the customer groupings based on these factors
- We deployed the K-means clustering method by using the **KMeans** package from **sklearn.cluster** library and **silhouette_score** from **sklearn.metrics** library
- We used the Elbow and Silhouette methods to arrive at the conclusion that we can divide into data set into 5 clusters

2.1.5 Analyzing reviews using NLP

- To further understand our customer sentiment, we downloaded customer reviews from the Turtle Games website and applied various natural language processing (NLP) techniques to find out the 15 most common words and 20 most positive and negative reviews
- Following process was followed:
 - a) Converted all reviews to lower case (**apply(str.lower)**), remove punctuation marks **str.replace('[^\w\s]','')** as well as drop duplicate rows **drop_duplicates()**
 - b) Tokenized each review using the **nlTK library** and **word.tokenize** function
 - c) Removed unnecessary using **english_stopwords library** and **isalnum()** functions
 - d) Created a **WordCloud** for the reviews as well as a frequency distribution using the **FreqDist()** function
 - e) Conducted a polarity and sentiment score analysis of the reviews using the **TextBlob().sentiment** method to understand the overall positive and negative sentiment

Overall, our project provided insights into how customers accumulate loyalty points and their sentiments towards the company, which can be valuable for improving customer experience and retention.

2.2 Turtle Games Sales Data Analysis using R:

2.2.1 Data Wrangling, Exploration and Visualization:

- In this analysis of Turtle Games sales data using R, the Tidyverse library was used for data manipulation, visualization, and modeling
- We imported the data and created a data frame using **read.csv()** function
- Viewed the data using **as_tibble()**, understood the descriptive statistics of the numeric columns using **summary()** function and removed unnecessary columns using the **select()** function
- Quick plots were created using **qplot()** for a better understanding of variable relationships
- To determine the Sales per Product and Sales per Platform, we used the **aggregate()** function to create 2 separate data frames
- We created some more detailed visualizations using the **ggplot()** to derive further insights from the data

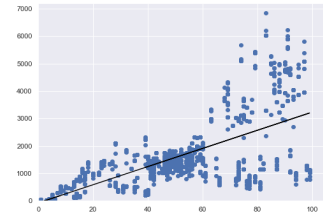
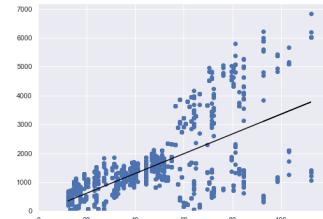
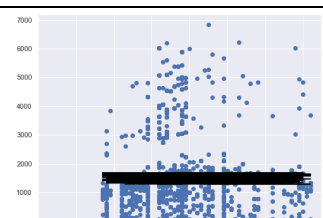
2.2.2 Data Modelling and Prediction:

- To prepare our data for modelling, the normality of the variables was tested using the Shapiro-Wilk test, and the skewness and kurtosis of the dataset were determined
- Since the data was not normal, we transformed the data using the **log()** transformation. Although the log transformation improved normality, the overall fit of the regression model decreased.
- This could mean that the log transformation was not appropriate for this data set and perhaps has introduced some errors. Hence, for further regression model analysis, we continued to use the original data set
- We created the simple linear regression model for all the sales data columns. In all cases, the ab line (line of best fit) was quite away from the actual data showing that the models are not very accurate or reliable
- Then we moved on to conduct a multiple linear regression model to understand the combined effect of NA Sales and EU Sales on the Global Sales

Overall, the analysis provides insights into Turtle Games' sales data, which can be useful for future decision-making.

3.0 Visualizations and Insights

3.1 Linear Regression Analysis Results:

Independent Variable	R-Squared	P-value	Intercept	Slope	Plot
Spending	0.45	0	-75.05	33.06	 A scatter plot showing the relationship between Spending (x-axis, 0 to 100) and Loyalty Points (y-axis, 0 to 7000). The data points are blue dots, and a black regression line shows a positive correlation. The line starts near the origin and rises steadily.
Remuneration	0.38	0	-65.68	34.19	 A scatter plot showing the relationship between Remuneration (x-axis, 0 to 100) and Loyalty Points (y-axis, 0 to 7000). The data points are blue dots, and a black regression line shows a positive correlation. The line starts near the origin and rises steadily.
Age	0.002	0.058	1736.52	-4.01	 A scatter plot showing the relationship between Age (x-axis, 0 to 70) and Loyalty Points (y-axis, 0 to 7000). The data points are blue dots, and a black regression line shows a very slight negative correlation. The line is nearly horizontal, starting at a high y-intercept and slightly declining.

- Spending causes 45% while Remuneration causes 38% of the variability in Loyalty Points respectively. Both variables show a moderate positive correlation with Loyalty points. Since the p-value is less than 0.05 in both cases, they are statistically significant.
- Age causes only 0.1% of the variability in Loyalty points. Also, the p-value is greater than 0.05, hence it is not significant.

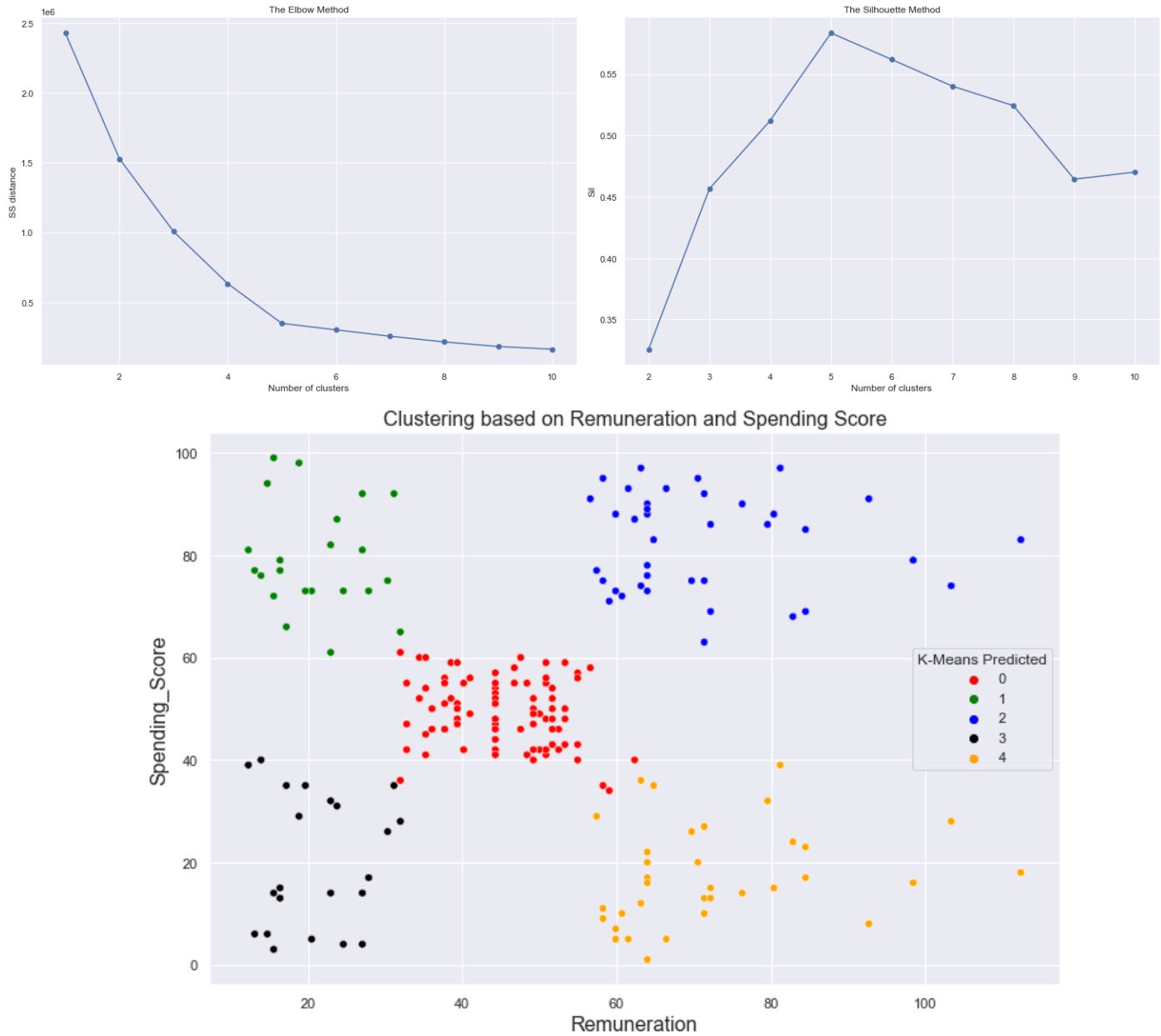
3.2 Multiple Linear Regression Analysis results:

Variable	Coefficient	P-value	R-Squared
Spending Score	33.97	0	0.842
Remuneration	34.25	0	
Age	11.01	0	

- The model showed a strong performance with R-squared value of 0.842 showing that the combination of spending, remuneration & age causes 84.2% variability in the loyalty points. Though all 3 factors are statistically significant, Remuneration & Spending have a stronger impact compared to Age.
- The multiple regression model does not have any multicollinearity or heteroscedasticity, thus we can conclude that model is stable and has high accuracy.
- Since the multiple regression model shows a more stronger impact on Loyalty points compared to the simple linear regression models of each factor with loyalty points, we suggest to use the multiple regression model to predict the loyalty points accumulation of users based on their spending, remuneration and age.

3.3 K-Means clustering method:

Elbow and Silhouette methods to arrive at the conclusion that we can divide into data set into 5 clusters

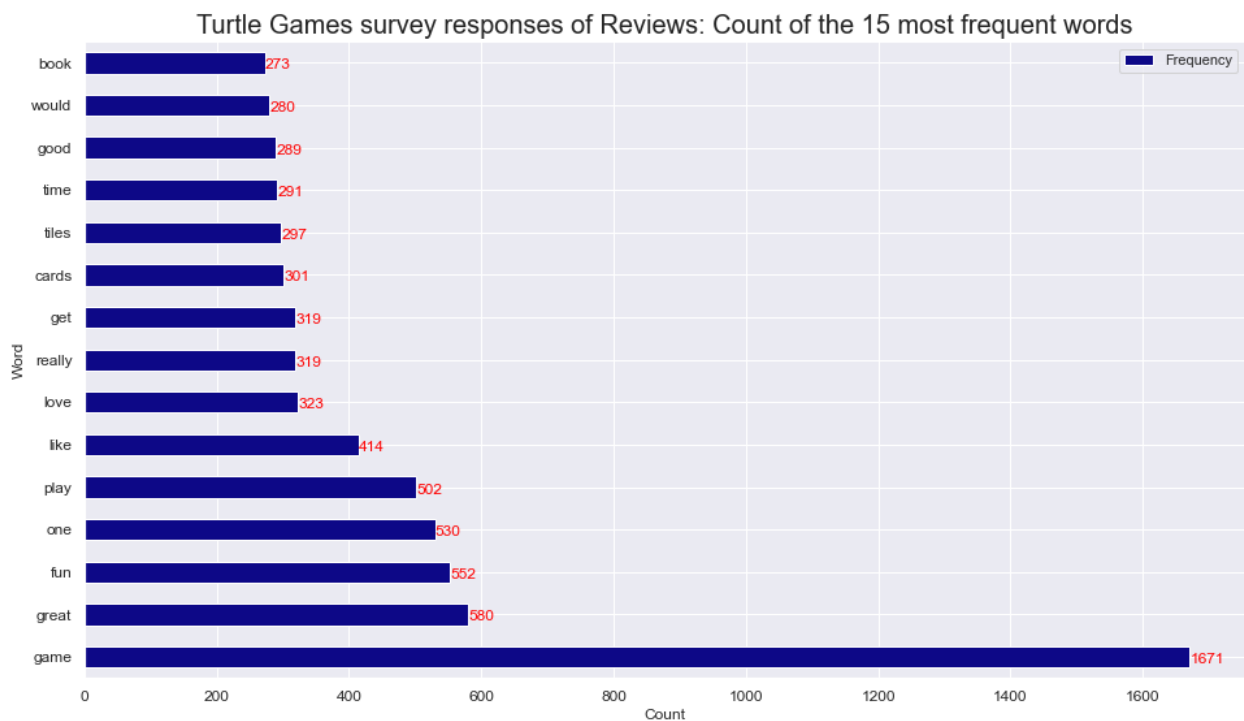


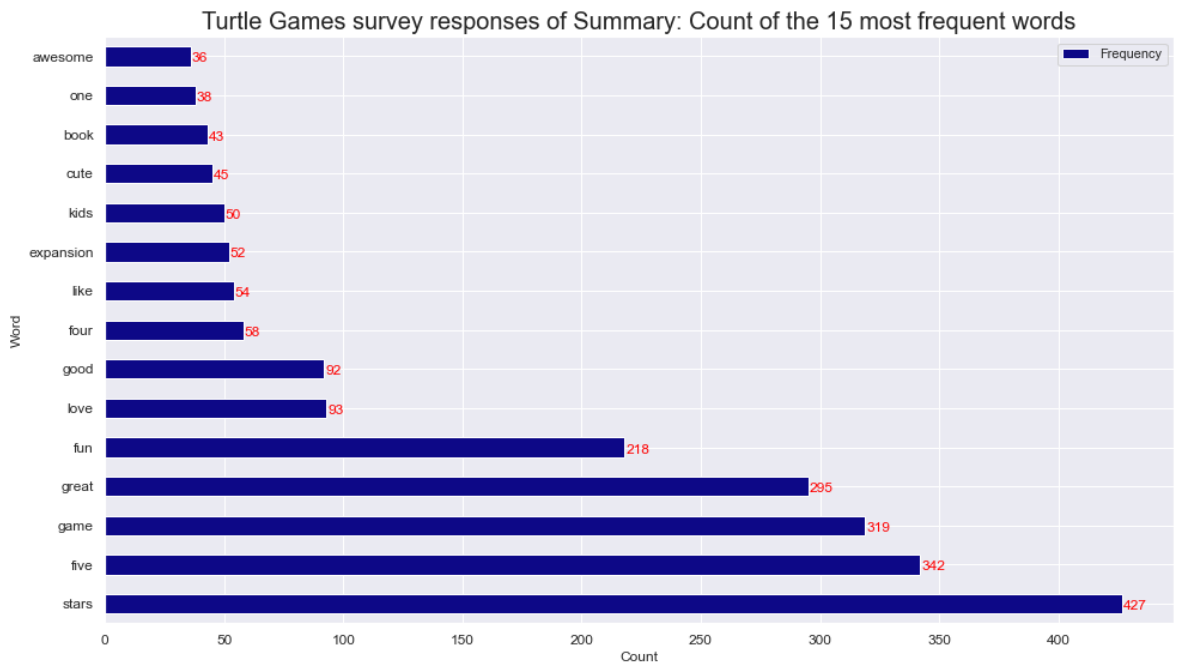
3.4 NLP – Reviews Analysis:

1. Word cloud of the reviews to spot frequent words:



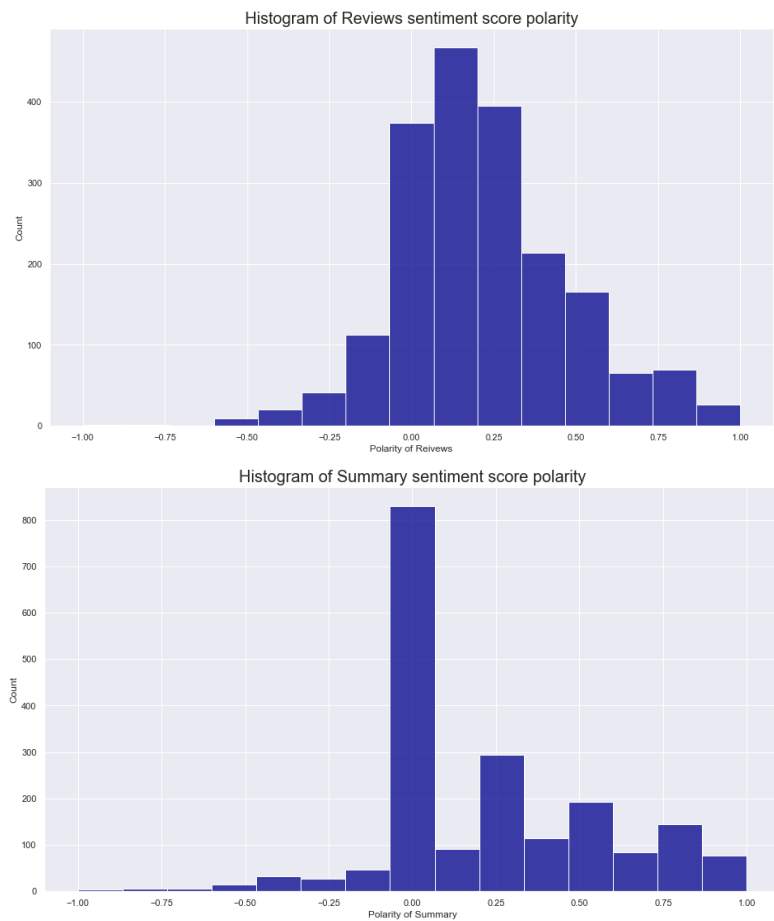
- ## 2. Frequency distribution of reviews and summary:





3. Sentiment & Polarity score

Overall, we can see from the histogram plot that most of the reviews and summary are closer to neutral but there is more distribution towards the positive side, indicating that in general the sentiment is positive.

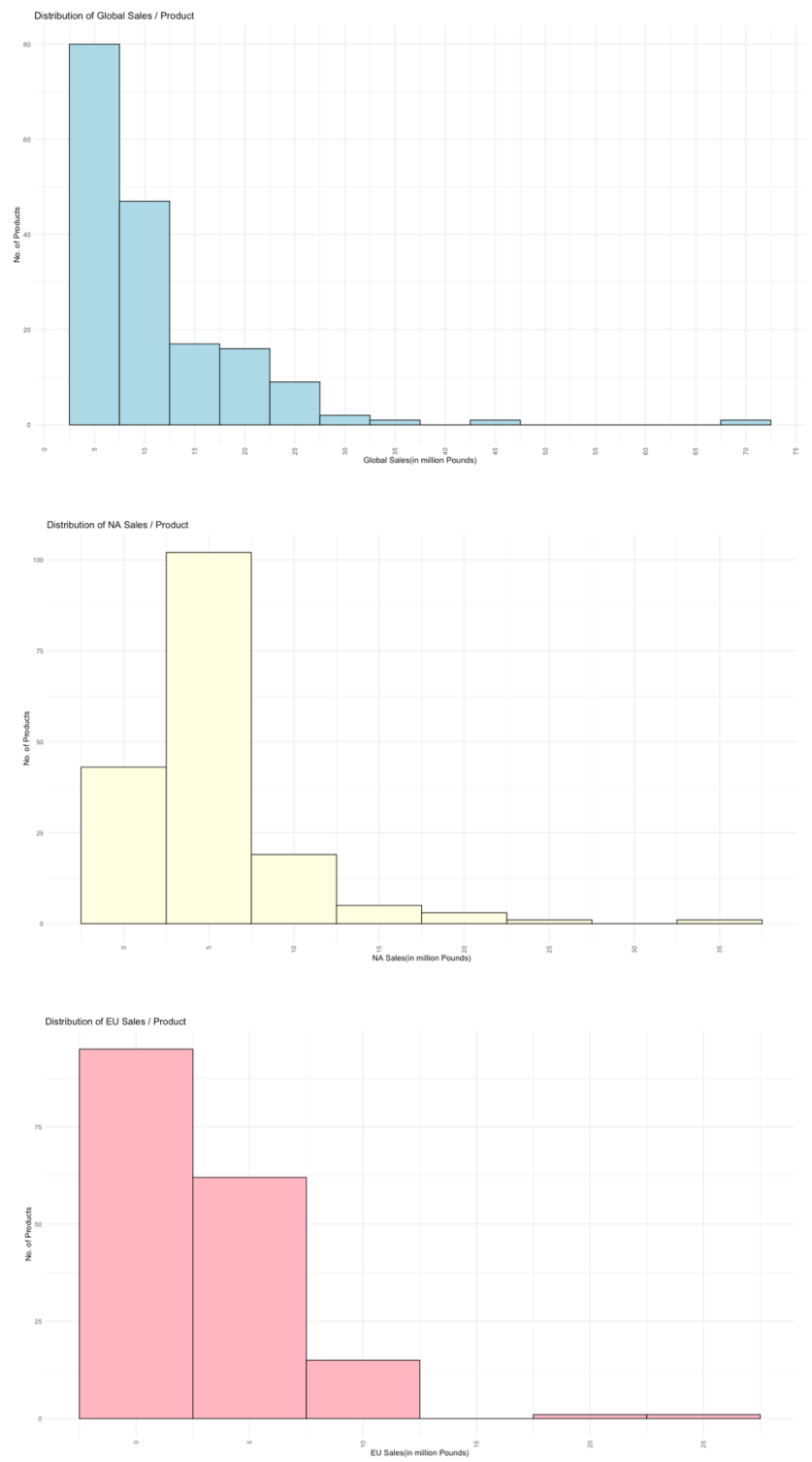


3.5 Visualization from analyzing the Sales Data:

3.5.1 Scatterplots of sales by region

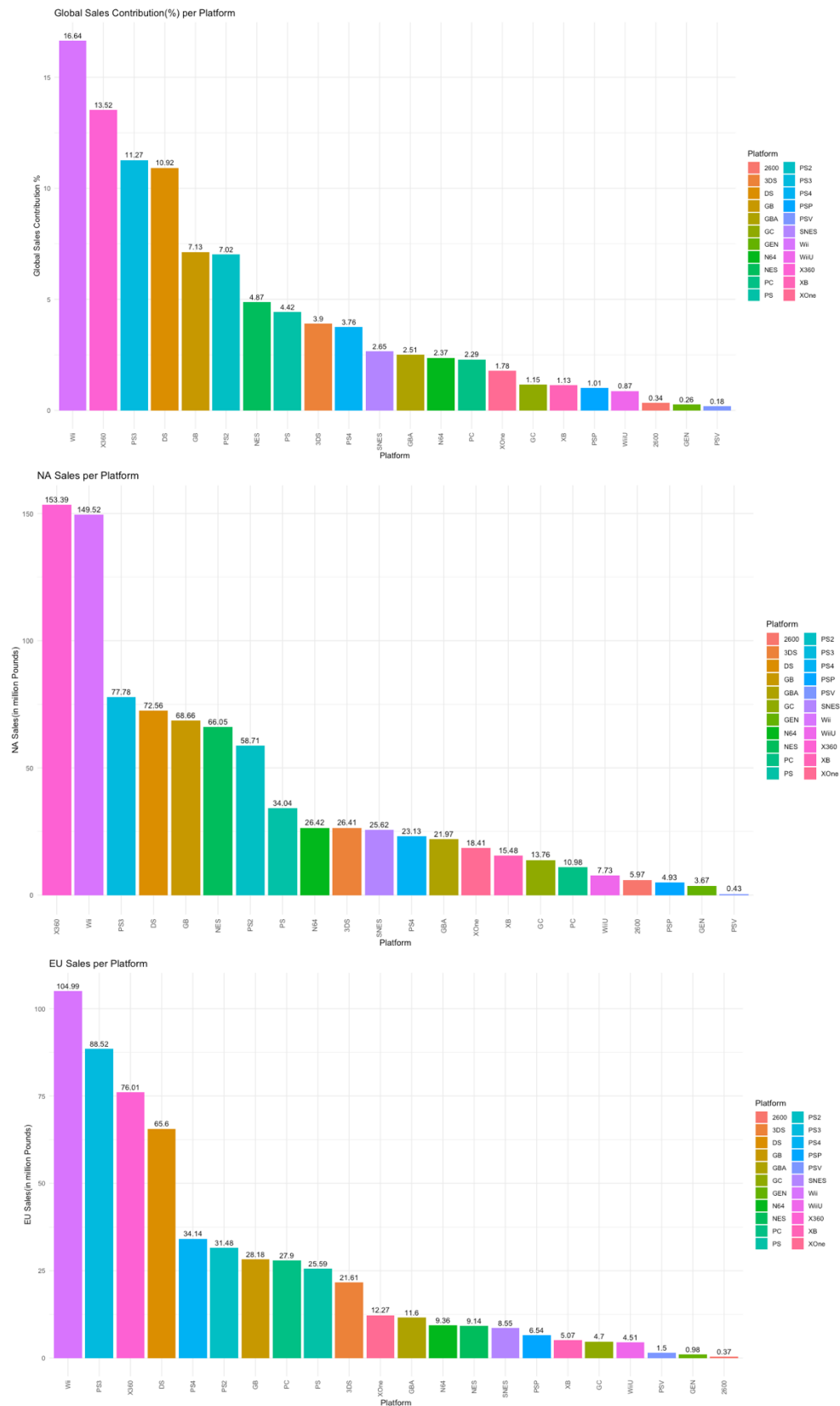


3.5.2 Distribution of Per Product Sale per region

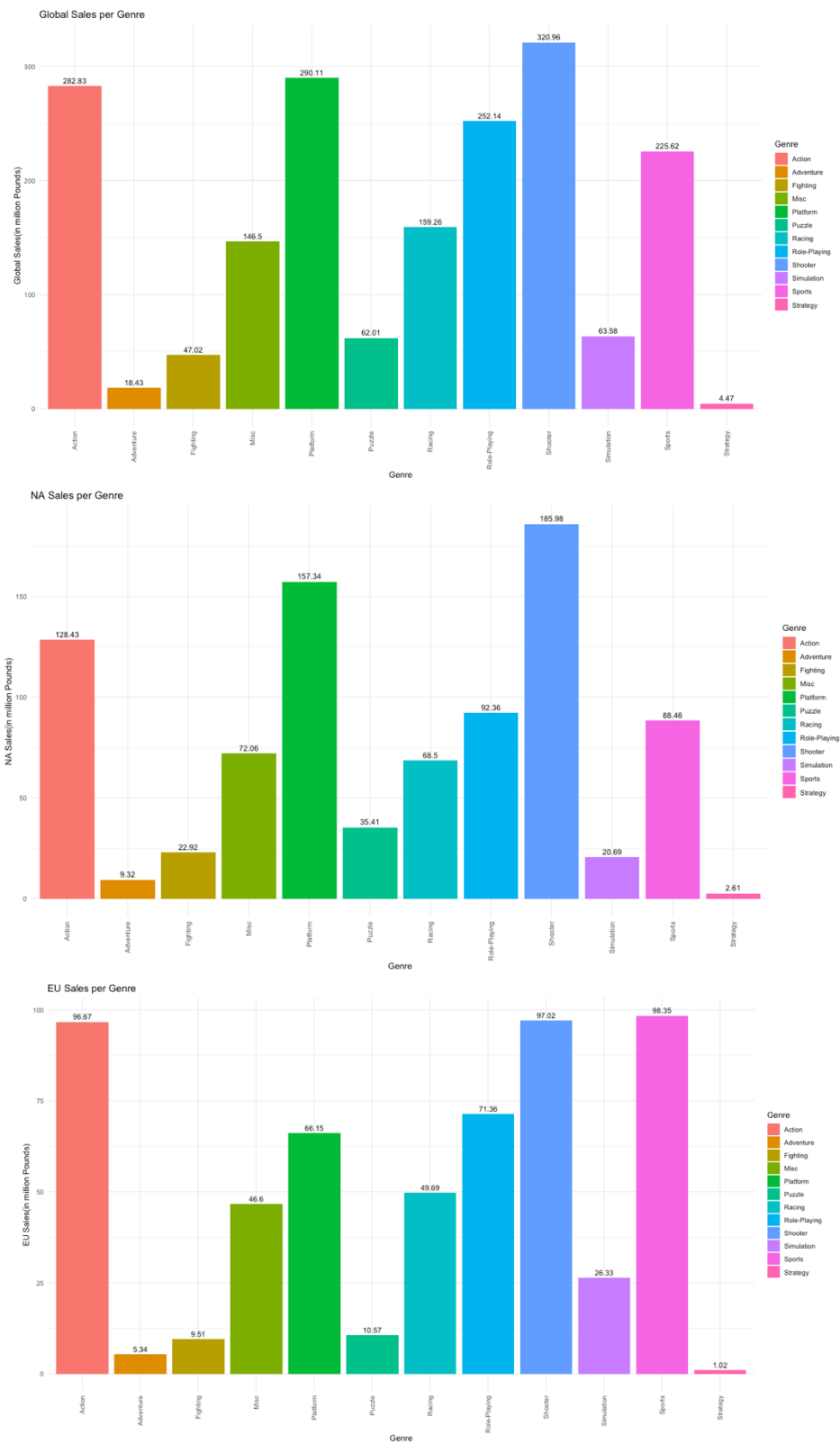


3.5.3 Most Popular Platforms per region (in terms of sales contribution)

We plotted bar charts of sales per platform in each region to understand which platforms are popular.



3.5.4 Most popular Genres per region (in terms of sales contribution)



4.0 Patterns & Predictions:

Based on our data analysis process and various visualizations, we would like to highlight the following patterns, trends/insights for further action:

- Marketing team can use the grouping shown in the K-means clustering to better target their campaigns. By implementing such a targeted approach, the marketing campaigns can be more effective to move customers to higher spending groups, ultimately increasing sales revenue:
 - Low Remuneration and Low Spending
 - High Remuneration and Low Spending
 - Medium Remuneration and Medium Spending
 - Low Remuneration and High Spending
 - High Remuneration and High Spending
- Analysis of the reviews shows us that the most common words are mostly positive. Moreover, the sentiment plot was also more biased towards the positive side. From analyzing the top negative reviews and looking at the Word Cloud, we were able to see some areas where improvement can be made. For e.g. some of the customers found the game to be difficult as the instructions were not clear. A more detailed analysis of the negative reviews is required to understand more areas for improvement
- Scatterplots of the regional sales based on products shows us a trend of positive correlation between the sales per region. Also, we can see the older products tend to have higher sales. This could be due to them selling for more years.
- There is also a slight difference in the most selling products across EU and NA - for e.g. Products 123, 254 & 326 have sales ranging from 26.64, 21.46, 22.08 million pounds in NA, however in EU, their sale is only 4.01, 2.42, 0.52 million pounds respectively. This could be due to the popularity of certain platforms in each region, however the phenomenon can be explored in detail in a separate study.
- Globally, the sale of most products ranges between 0-20 million pounds. In NA, the sale of most products ranges between 0-10 million pounds while in EU, the sale of most products ranges between 0-5 million pounds
- In terms of platforms, Wii, X360, PS3 form the top 3 platforms in terms of sales contribution globally where the top 5 platforms contribute to 60% of the Global Sales. There is a slight difference in the popularity of the platforms across EU and NA which can be further explored to market products accordingly:

GLOBAL	NA	EU
<ul style="list-style-type: none">•1. Wii•2. X360•3. PS3•4. DS•5. GB	<ul style="list-style-type: none">•1. X360•2. Wii•3. PS3•4. DS•5. GB	<ul style="list-style-type: none">•1. Wii•2. PS3•3. X360•4. DS•5. PS4

- From the Genre histograms, we can see the popular genres in each region. We could promote the game genres which are popular in each region to improve sales:

GLOBAL	NA	EU
<ul style="list-style-type: none">•1. Shooter•2. Platform•3. Action•4. Role Playing•5. Sports	<ul style="list-style-type: none">•1. Shooter•2. Platform•3. Action•4. Role-Playing•5. Sports	<ul style="list-style-type: none">•1. Sports•2. Shooter•3. Action•4. Role-Playing•5. Platform

- The Multiple Linear Regression model for predicting Global Sales based on NA & EU sales has a R-squared value of 0.97 i.e. 97% of the variability in Global Sales is due to NA & EU sales. Data used for the model was not normal so the model might not be very reliable. However, there is no multicollinearity and no heteroscedasticity which makes it stable. Hence, we can use this model but with some caution.

5.0 Appendix

5.1 Top 20 positive reviews

	review	summary	polarity_review
	came in perfect condition	five stars	1.000000
	awesome book	five stars	1.000000
	awesome gift	five stars	1.000000
	excellent activity for teaching selfmanagement skills	five stars	1.000000
	perfect just what i ordered	five stars	1.000000
	wonderful product	five stars	1.000000
	delightful product	five stars	1.000000
	wonderful for my grandson to learn the resurrection story	five stars	1.000000
	perfect	acquire game	1.000000
	awesome	five stars	1.000000
	awesome set	five stars	1.000000
	best set buy 2 if you have the means	five stars	1.000000
	awesome addition to my rpg gm system	five stars	1.000000
	its awesome	five stars	1.000000
	one of the best board games i played in along time	five stars	1.000000
	my daughter loves her stickers awesome seller thank you	awesome seller thank you	1.000000
	this was perfect to go with the 7 bean bags i just wish they were not separate orders	five stars	1.000000
	awesome toy	five stars	1.000000
	it is the best thing to play with and also mind blowing in some ways	three stars	1.000000
	excellent toy to simulate thought	five stars	1.000000

5.2 Top 20 Negative Reviews

review	summary	polarity_review
booo unles you are patient know how to measure i didnt have the patience neither did my daughter boring unless you are a craft person which i am not	boring unless you are a craft person which i am	-1.000000
incomplete kit very disappointing	incomplete kit	-0.780000
im sorry i just find this product to be boring and to be frank juvenile	disappointing	-0.583333
one of my staff will be using this game soon so i dont know how well it works as yet but after looking at the cards i believe it will be helpful in getting a conversation started regarding anger and what to do to control it	anger control game	-0.550000
i bought this as a christmas gift for my grandson its a sticker book so how can i go wrong with this gift	stickers	-0.500000
this was a gift for my daughter i found it difficult to use	two stars	-0.500000
i found the directions difficult	three stars	-0.500000
instructions are complicated to follow	two stars	-0.500000
difficult	three stars	-0.500000
expensive for what you get	two stars	-0.500000
i sent this product to my granddaughter the pompom maker comes in two parts and is supposed to snap together to create the pompoms however both parts were the same making it unusable if you cant make the pompoms the kit is useless since this was sent as a gift i do not have it to return very disappointed	faulty product	-0.491667
my 8 yearold granddaughter and i were very frustrated and discouraged attempting this craft it is definitely not for a young child i too had difficulty understanding the directions we were very disappointed	frustating	-0.446250
i purchased this on the recommendation of two therapists working with my adopted children the children found it boring and put it down half way through	hmmm	-0.440741
very hard complicated to make these	one star	-0.439583