TURTLE GAMES

Understanding Turtle Games Sales and Customer Loyalty
“If you are not taking care of your customer, your competitor will.”

TABLE OF CONTENTS

SUBMITTED BY	Smita Prasad
COURSE	CO4_LSE_DA_301 Advanced Analytics for Organisational Impact
DOCUMENT LENGTH	1173 words (excluding cover)

LSE DA ACCELERATOR TURTLE GAMES

CONTEXT

Turtle Games is a game manufacturer and retailer with a global customer base. The company 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.
To improve overall sales performance, Turtle Games has come up with an initial set of questions which were addressed during this analysis.
While the study intends to focus on the questions posed by the company, it aims to bring further insights into other trends present in the data provided by the company.

ANALYTICAL APPROACH

The dataset includes -:

turtle_reviews.csv – Details on customer gender, age, remuneration, spending score, loyalty points, education, language, platform, review, and summary across products.
turtle_sales.csv – Details of video games sold globally, such as the rank, product, platform, genre, publisher, and their sales across North America, Europe, and worldwide.
metadata_nhs.txt – Details of the data set, data quality, and reference.

The turtle_reviews dataset was analysed using Python. The sales department of Turtle games prefers R to Python, so the sales data was analyzed in R.

EXPLORING THE DATA FOR INSIGHTS IN PYTHON

We begin by loading and exploring the data. The data is loaded in a DataFrame object (pandas library) .Some preliminary steps were taken to sense-check the data and ensure any NA values are handled.
View the data -:

:

	gender	age	remuneration (k£)	spending_score (1-100)	loyalty_points	education	language	platform	product	review	summary
0	Male	18	12.30	39	210	graduate	EN	Web	453	When it comes to a DM's screen, the space on t...	The fact that 50% of this space is wasted on a...
1	Male	23	12.30	81	524	graduate	EN	Web	466	An Open Letter to GaleForce9*:\n\nYour unpaint...	Another worthless Dungeon Master's screen from...
2	Female	22	13.12	6	40	graduate	EN	Web	254	Nice art, nice printing. Why two panels are f...	pretty, but also pretty useless
3	Female	25	13.12	77	562	graduate	EN	Web	263	Amazing buy! Bought it as a gift for our new d...	Five Stars
4	Female	33	13.94	40	366	graduate	EN	Web	291	As my review of GF9's previous screens these w...	Money trap
...	...	...	...	...	...	...	...	...	...	...	...
1995	Female	37	84.46	69	4031	PhD	EN	Web	977	The perfect word game for mixed ages (with Mom...	The perfect word game for mixed ages (with Mom
1996	Female	43	92.66	8	539	PhD	EN	Web	979	Great game. Did not think I would like it whe...	Super fun
1997	Male	34	92.66	91	5614	graduate	EN	Web	1012	Great game for all.....\nKeeps the mind ni...	Great Game
1998	Male	34	98.40	16	1048	PhD	EN	Web	1031	fun game!	Four Stars
1999	Male	32	92.66	8	479	PhD	EN	Web	453	This game is fun. A lot like scrabble without ...	Love this game

Check for missing values -: There are no NA values in the turtle_reviews dataset.

```
# Checking for missing values

df_reviews.isnull().sum()
```

```
gender                  0
age                     0
remuneration (k£)       0
spending_score (1-100)  0
loyalty_points          0
education               0
language                0
platform                0
product                 0
review                  0
summary                 0
dtype: int64
```

View the structure of the DataFrame containing the dataset.

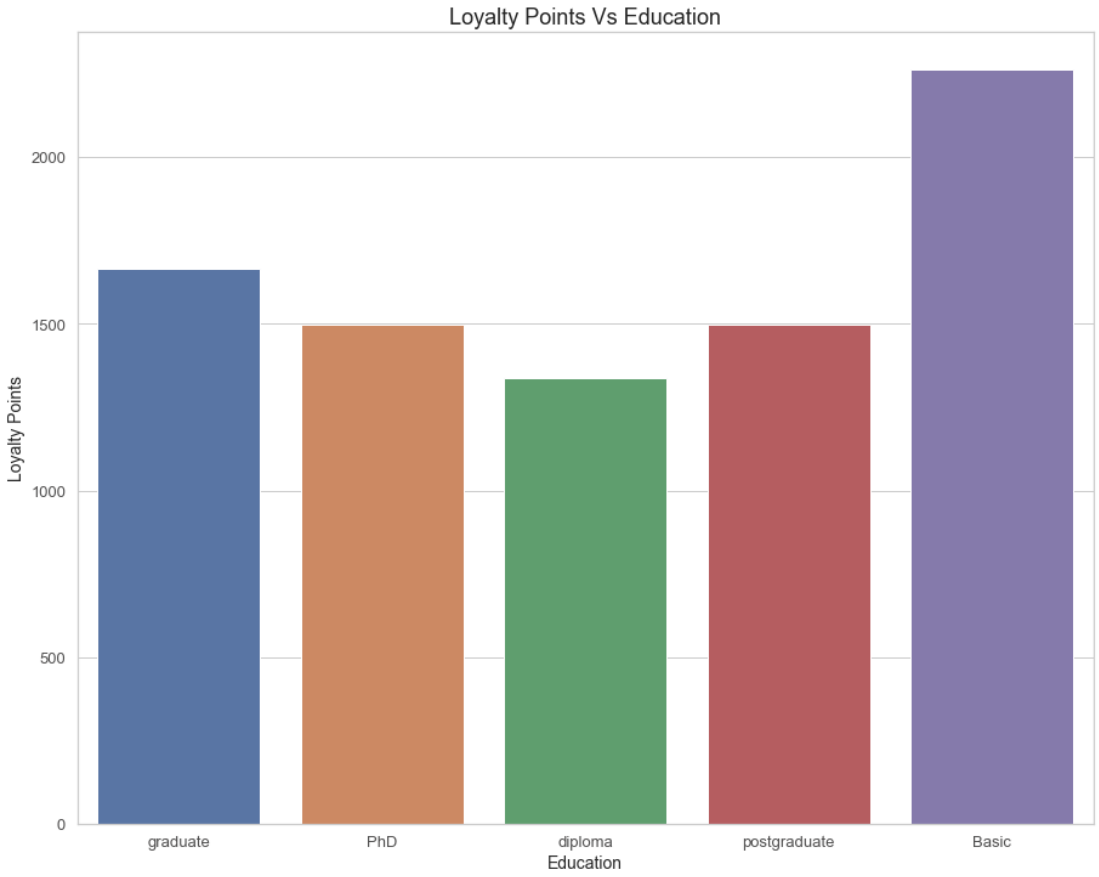
```
# Exploring the data.
df_reviews.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 11 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   gender                  2000 non-null   object
 1   age                     2000 non-null   int64
 2   remuneration (k£)       2000 non-null   float64
 3   spending_score (1-100)  2000 non-null   int64
 4   loyalty_points          2000 non-null   int64
 5   education               2000 non-null   object
 6   language                2000 non-null   object
 7   platform                2000 non-null   object
 8   product                 2000 non-null   int64
 9   review                  2000 non-null   object
 10  summary                 2000 non-null   object
dtypes: float64(1), int64(4), object(6)
memory usage: 172.0+ KB
```

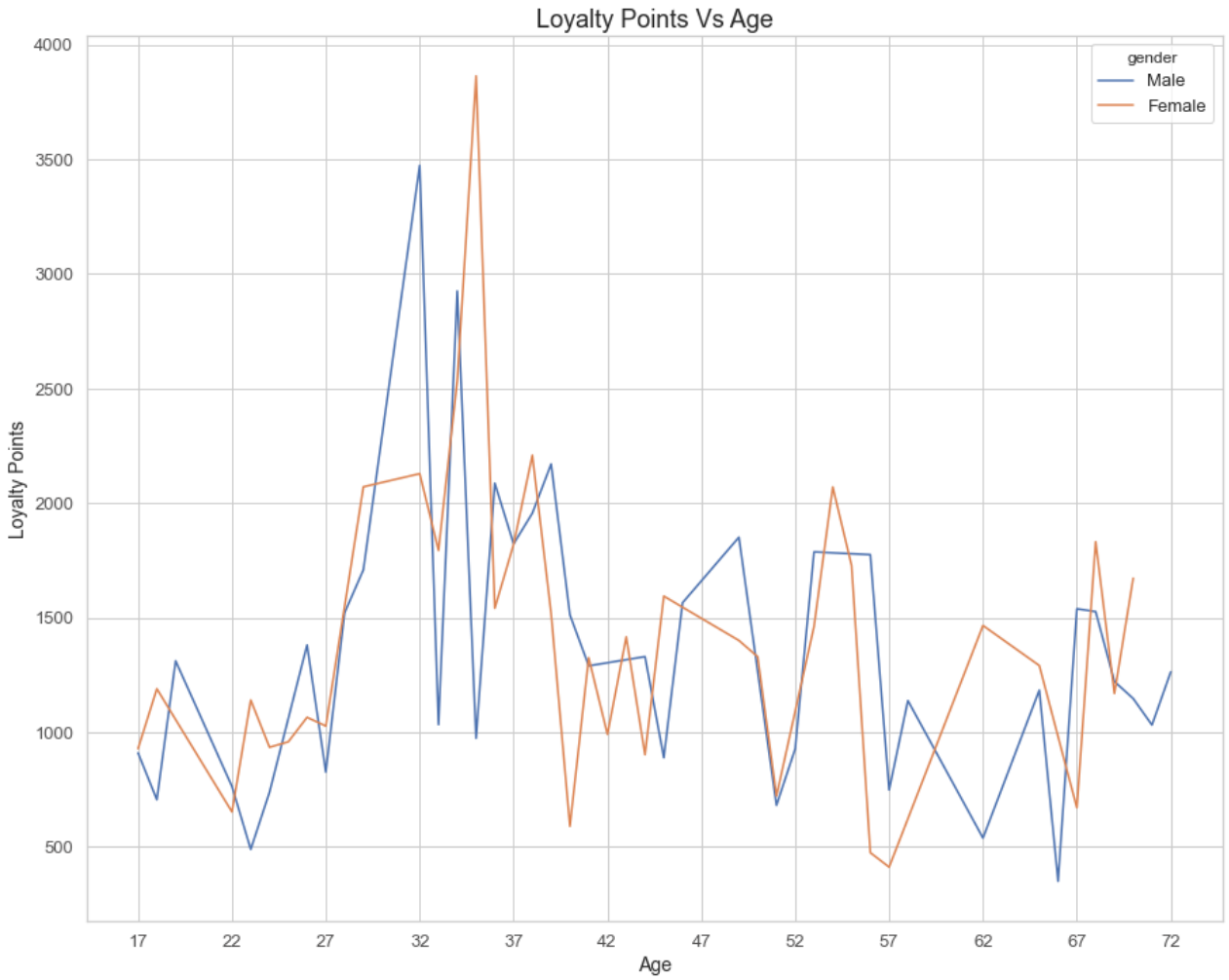
Viewing the DataFrame above it is observed that the columns **language** and **platform** are redundant as they have only one unique value i.e. their values are same for all entries. Therefore, those columns are dropped .

Accumulation Of Loyalty Points

Turtle Games wants to understand how customers accumulate loyalty points. Exploring the data, we observe the following trends in the dataset.



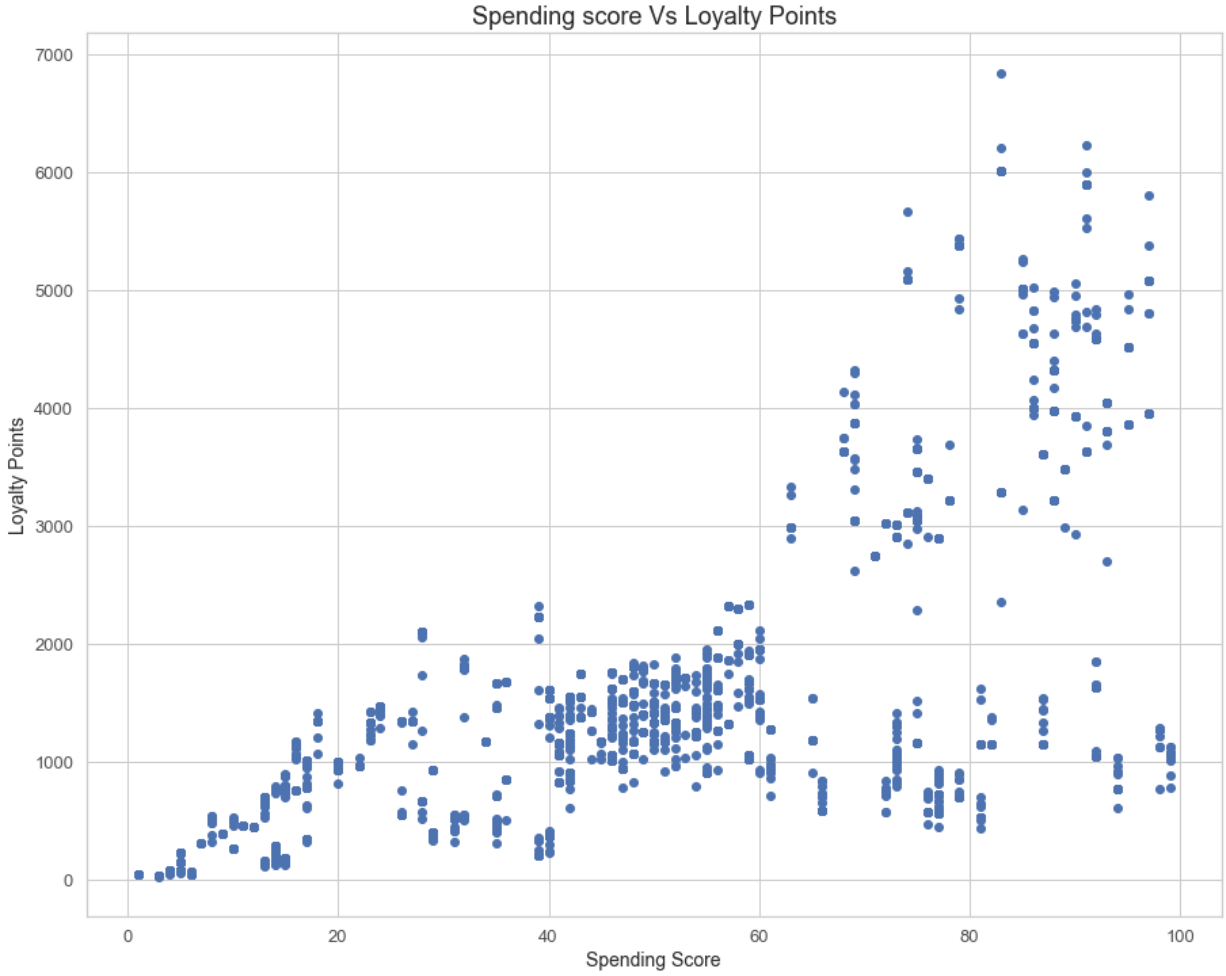
We can see that the customers with Education classified as Basic form a major segment in with high loyalty points.

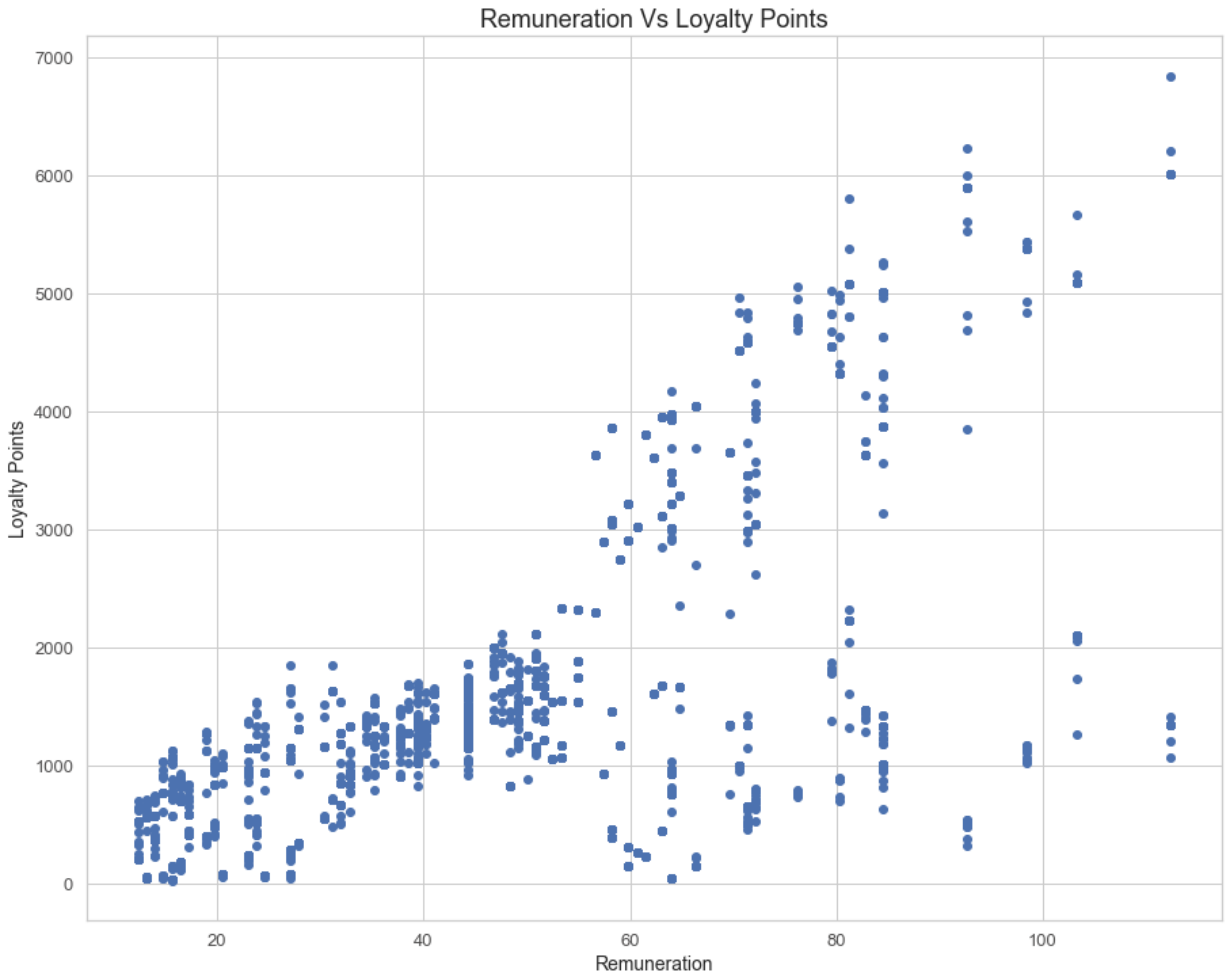


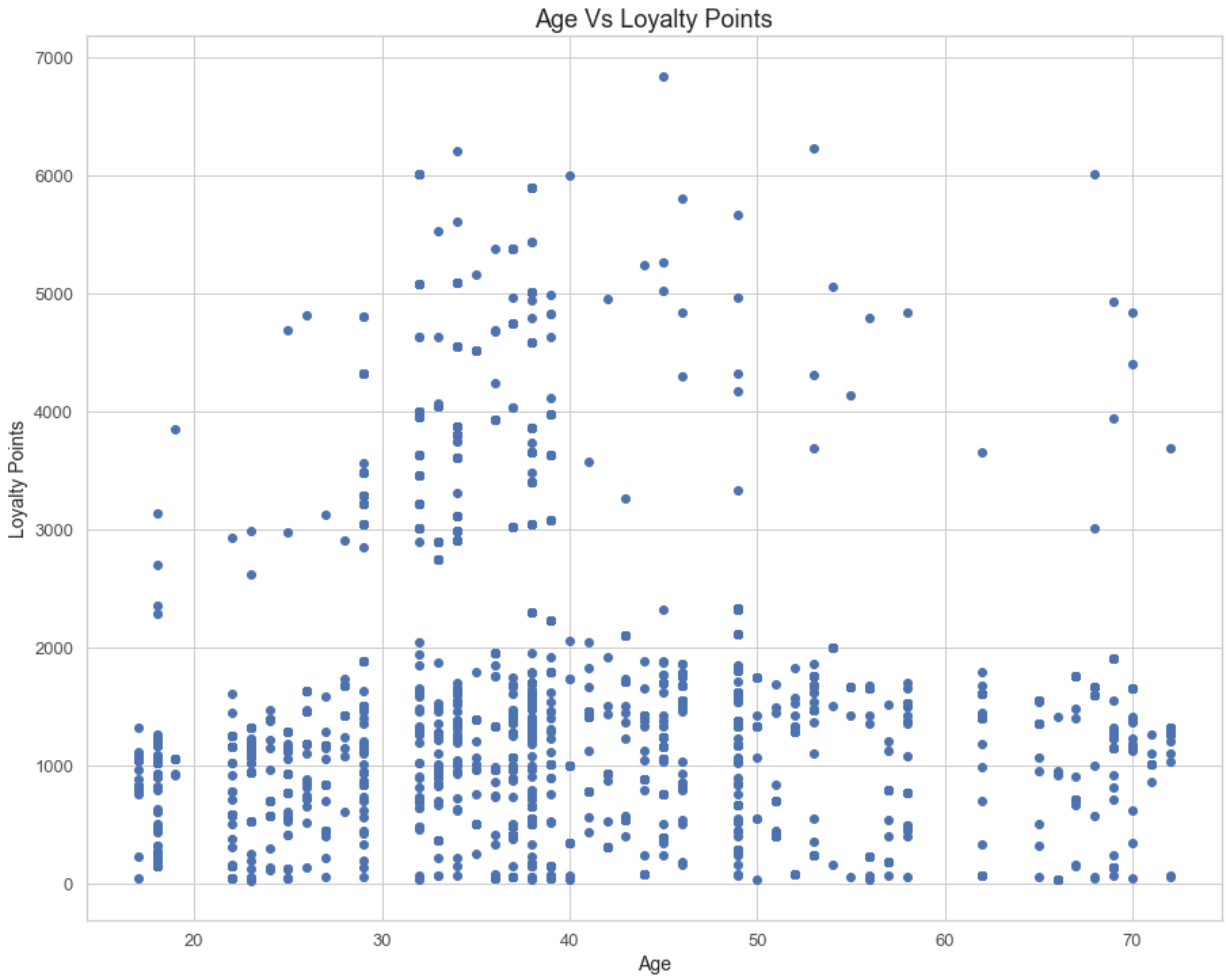
Observations:

- Ages range from 17 to 72
- loyalty points peak for ages 30 and 35.
- highest loyalty points are for females in the age range 30 to 35

To understand the relationship between loyalty points and various variables we conducted a regression analysis. Using simple linear regression models did not yield satisfactory results.




Remuneration Vs Loyalty Points
Loyalty Points
Remuneration
0
1000
2000
3000
4000
5000
6000
7000
20
40
60
80
100



Building a model using multiple regression we were able to build a model that explained 90.7 % of the loyalty points data. However, the model may not explain all the trends in the loyalty points particularly values which are very high or low.

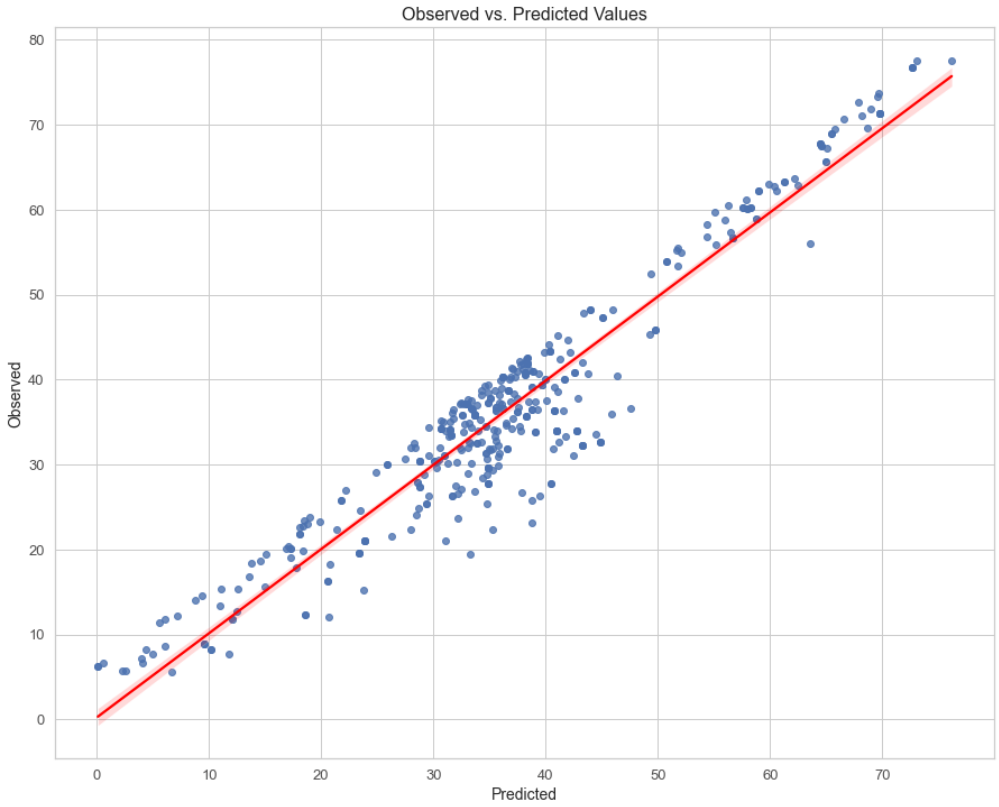
OLS Regression Results

Dep. Variable:	loyalty_points	**R-squared:**	0.907
Model:	OLS	**Adj. R-squared:**	0.907
Method:	Least Squares	**F-statistic:**	6486.
Date:	Sun, 23 Apr 2023	**Prob (F-statistic):**	0.00
Time:	00:22:52	**Log-Likelihood:**	-5938.0
No. Observations:	2000	**AIC:**	1.188e+04
Df Residuals:	1996	**BIC:**	1.191e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-11.3140	0.481	-23.539	0.000	-12.257	-10.371
spending_score	0.4451	0.004	107.289	0.000	0.437	0.453
remuneration	0.4039	0.005	88.542	0.000	0.395	0.413
age	0.1579	0.008	19.799	0.000	0.142	0.174

Omnibus:	352.466	**Durbin-Watson:**	2.493
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	561.569
Skew:	-1.213	**Prob(JB):**	1.14e-122
Kurtosis:	3.925	**Cond. No.**	377.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.



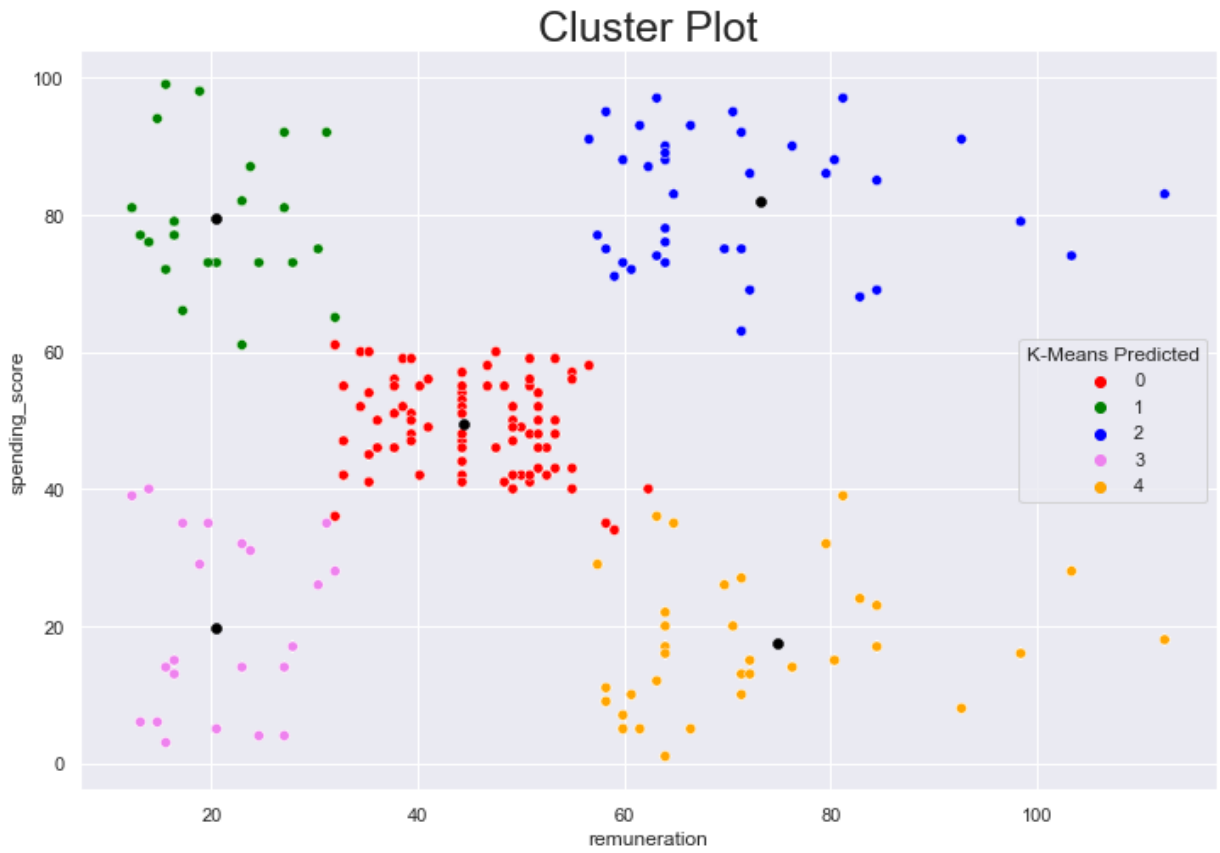
The plot of the Observed Vs Predicted values using the multiple regression model.

Exploring Clusters within the Customer Base

Turtle games wants to identify groups within the customer base that can be used to target specific market segments. Clustering is broadly used to group objects into clusters based on similarity.
We used K-means clustering algorithm to explore groups with the customer dataset to identify groups based on spending scores and remuneration. Using methods such as the Elbow method and Silhouette method we identified 5 main clusters in the dataset as following-:

CLUSTER NUMBER	REMUNERATION	SPENDING-SCORE
0 .	44.42	49.53
1	20.35	79.42
2	73.24	82.01
3	20.42	19.76
4	74.83	17.42

This can be visualized as shown below -:



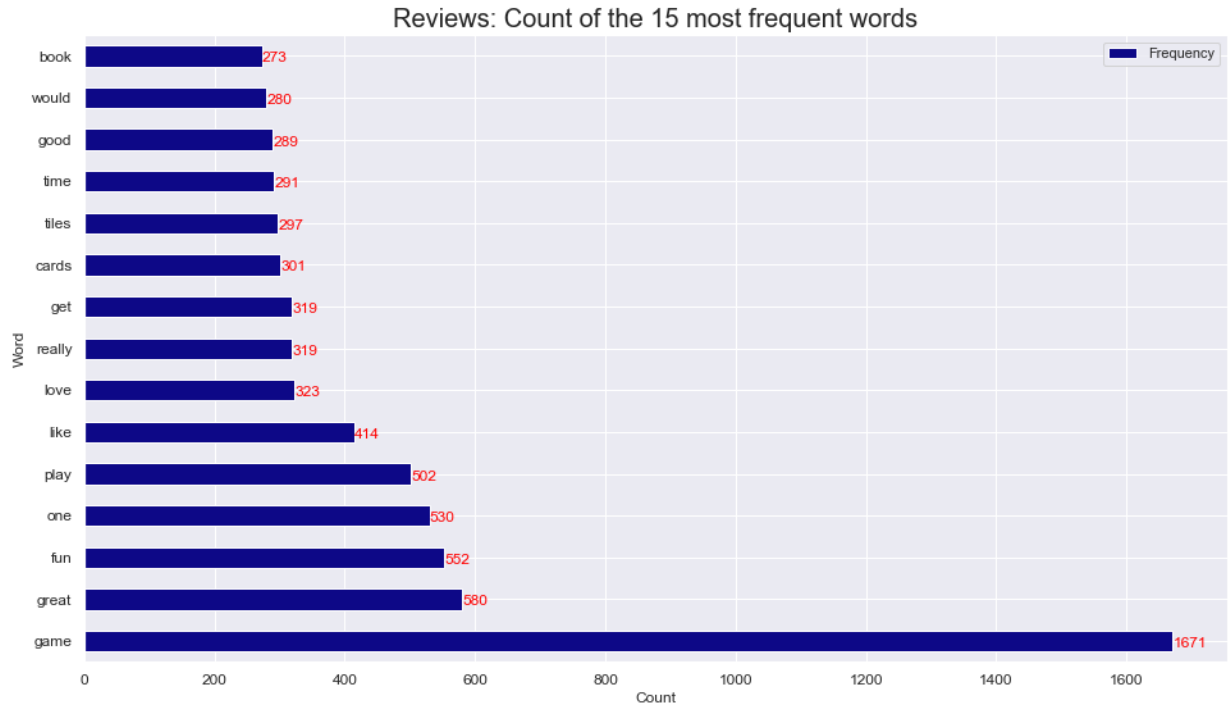
Analyzing Customer Reviews using NLP

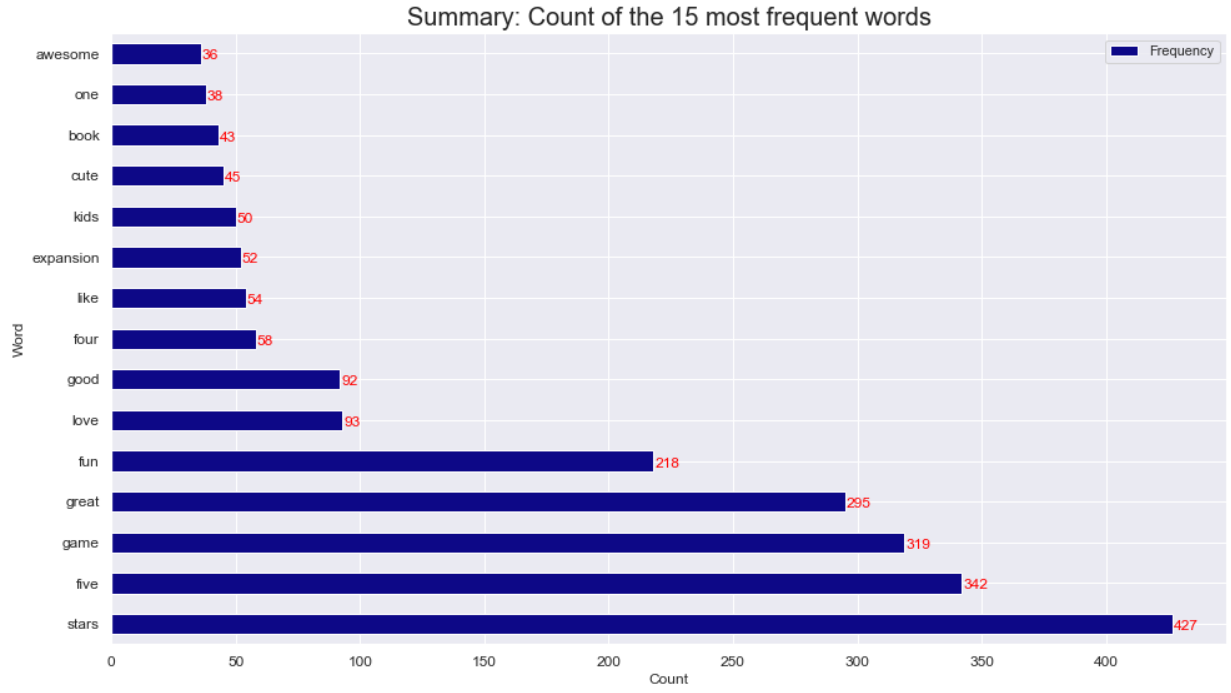
We used natural language processing techniques to analyze the customer reviews and to identify keywords, topics etc.

	review	summary	rev_tokens	summary_tokens
0	when it comes to a dms screen the space on the...	the fact that 50 of this space is wasted on ar...	[when, it, comes, to, a, dms, screen, the, spa...	[the, fact, that, 50, of, this, space, is, was...
1	an open letter to galeforce9 your unpainted mi...	another worthless dungeon masters screen from ...	[an, open, letter, to, galeforce9, your, unpai...	[another, worthless, dungeon, masters, screen,...
2	nice art nice printing why two panels are fill...	pretty but also pretty useless	[nice, art, nice, printing, why, two, panels, ...	[pretty, but, also, pretty, useless]
3	amazing buy bought it as a gift for our new dm...	five stars	[amazing, buy, bought, it, as, a, gift, for, o...	[five, stars]
4	as my review of gf9s previous screens these we...	money trap	[as, my, review, of, gf9s, previous, screens, ...	[money, trap]

Frequent words in the summary column viewed as a wordcloud-:


five stars
great
game
fun
love
four stars
great game
stars
great
stars five
good
expansion
little
waterdeep
year old
perfect
loved
nice
book
cute
kid
toy
tile
best
excellent
one
quality
awesome
set
sticker
family
really
age
play
fun
game
made
product
easy
time
puzzle
gift
buy
make
better
pretty
acquire
worth
addition
children
monopoly
word game
three stars
two stars
super fun
great expansion
stars four
board game
dungeon
price
useful
well


Reviews: Count of the 15 most frequent words
Frequency
book 273
would 280
good 289
time 291
tiles 297
cards 301
get 319
really 319
love 323
like 414
play 502
one 530
fun 552
great 580
game 1671
Word
0 200 400 600 800 1000 1200 1400 1600
Count

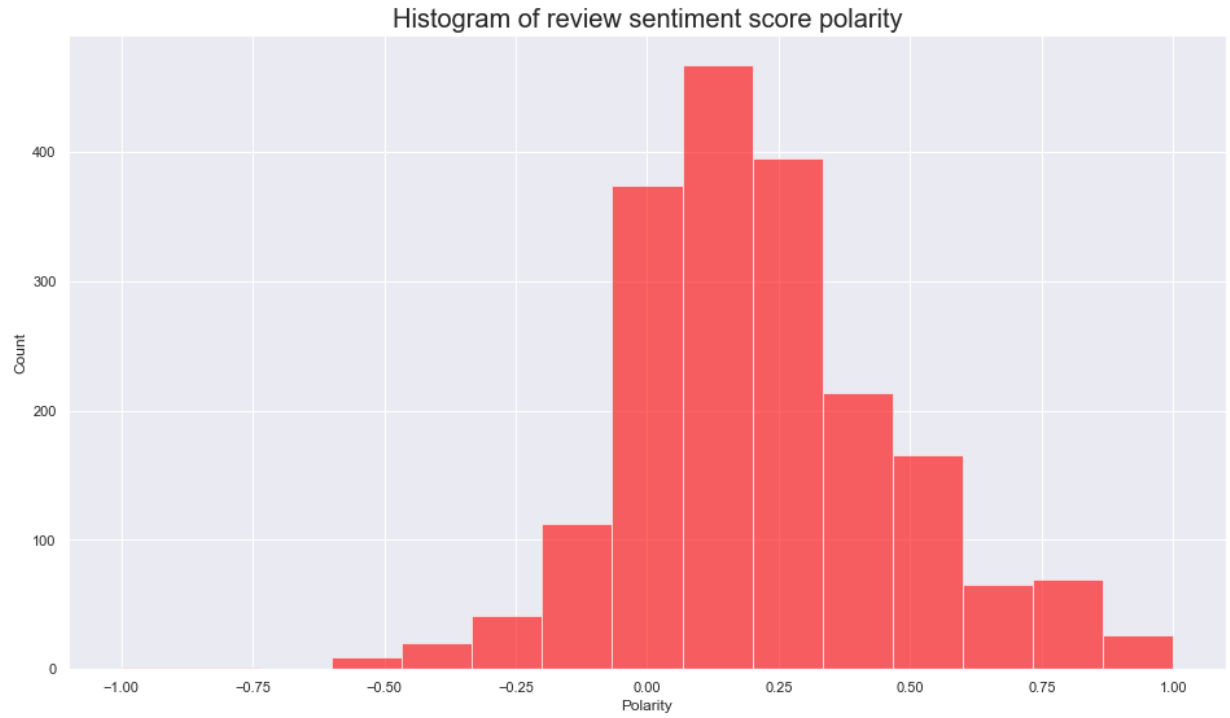

Summary: Count of the 15 most frequent words
Frequency
awesome 36
one 38
book 43
cute 45
kids 50
expansion 52
like 54
four 58
good 92
love 93
fun 218
great 295
game 319
five 342
stars 427
Word
0 50 100 150 200 250 300 350 400
Count

We then classified the reviews using a library called Textblob which supports sentiment classification in Python. When a sentence is passed into Textblob it gives two outputs, which are polarity and subjectivity. Polarity is the output that lies between [-1,1], where -1 refers to negative sentiment and +1 refers to positive sentiment.

top 20 negative reviews

	summary	summary_polarity
208	boring unless you are a craft person which i am	-1.000000
21	the worst value ive ever seen	-1.000000
829	boring	-1.000000
1166	before this i hated running any rpg campaign d...	-0.900000
1	another worthless dungeon masters screen from ...	-0.800000
793	disappointed	-0.750000
1620	disappointed	-0.750000
144	disappointed	-0.750000
631	disappointed	-0.750000
363	promotes anger instead of teaching calming met...	-0.700000
885	too bad this is not what i was expecting	-0.700000
890	bad qualityall made of paper	-0.700000
178	at age 31 i found these very difficult to make	-0.650000
518	mad dragon	-0.625000
101	small and boring	-0.625000
1804	disappointing	-0.600000
1015	disappointing	-0.600000
1115	disappointing	-0.600000
805	disappointing	-0.600000
1003	then you will find this board game to be dumb ...	-0.591667

Viewing the histogram of both review and summary polarity scores we notice that the histogram is skewed positively indicating more positive reviews than negative ones.


Histogram of review sentiment score polarity
Count
0
100
200
300
400
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Polarity

top 20 positive reviews

	summary	summary_polarity
1028	one of the best	1.0
1935	excellent	1.0
815	one of the best games ever	1.0
1630	awesome learning tool	1.0
163	he was very happy with his gift	1.0
1170	best orcs from wotc	1.0
1388	awesome expansion	1.0
1078	perfect gift	1.0
140	awesome sticker activity for the price	1.0
1171	awesome	1.0
647	wonderful	1.0
651	all f the mudpuppy toys are wonderful	1.0
1488	the perfect gift for preschool construction fans	1.0
1083	best dungeon crawler	1.0
657	awesome puzzle	1.0
980	the best among the dd boardgames	1.0
1230	awesome addition to our dd antics	1.0
161	awesome book	1.0
1417	wonderful and	1.0
1454	awesome expansion	1.0

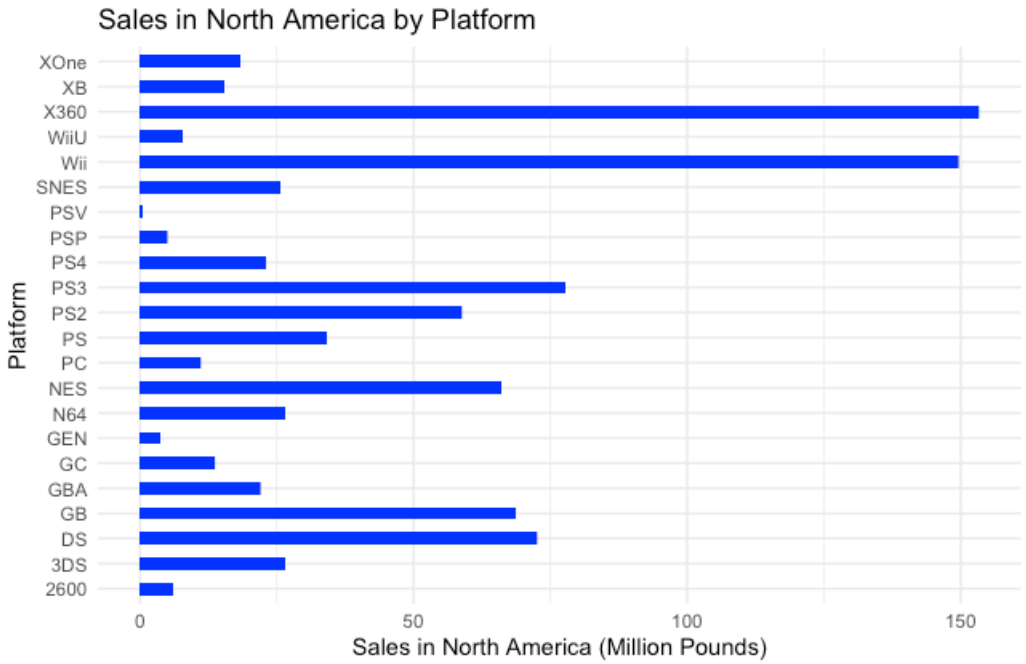
ANALYSIS OF SALES DATA IN R

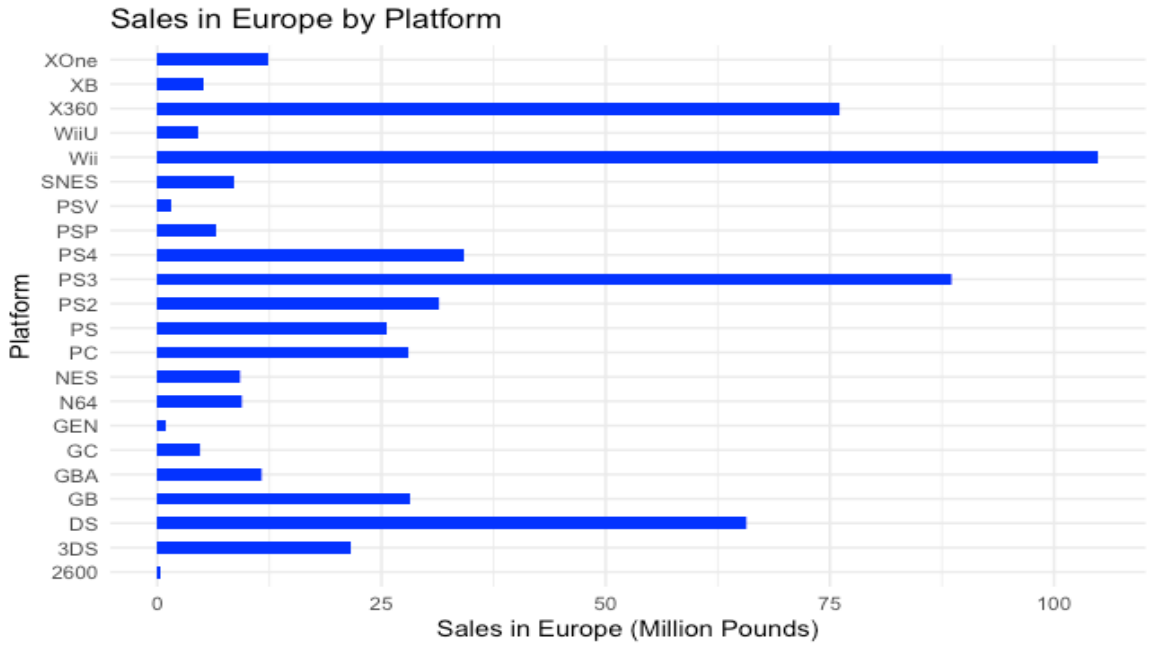
Due to its unmatched packages for data exploration and experimentation, R is favoured by statisticians, academics, and data analysts (e.g. financial industry). The sales team preferred R for the analysis.
We first loaded the dataset turtle_sales.csv into Rstudio and sense checked the data.
Checking NA values, we noticed two rows have NA values in the Year column.

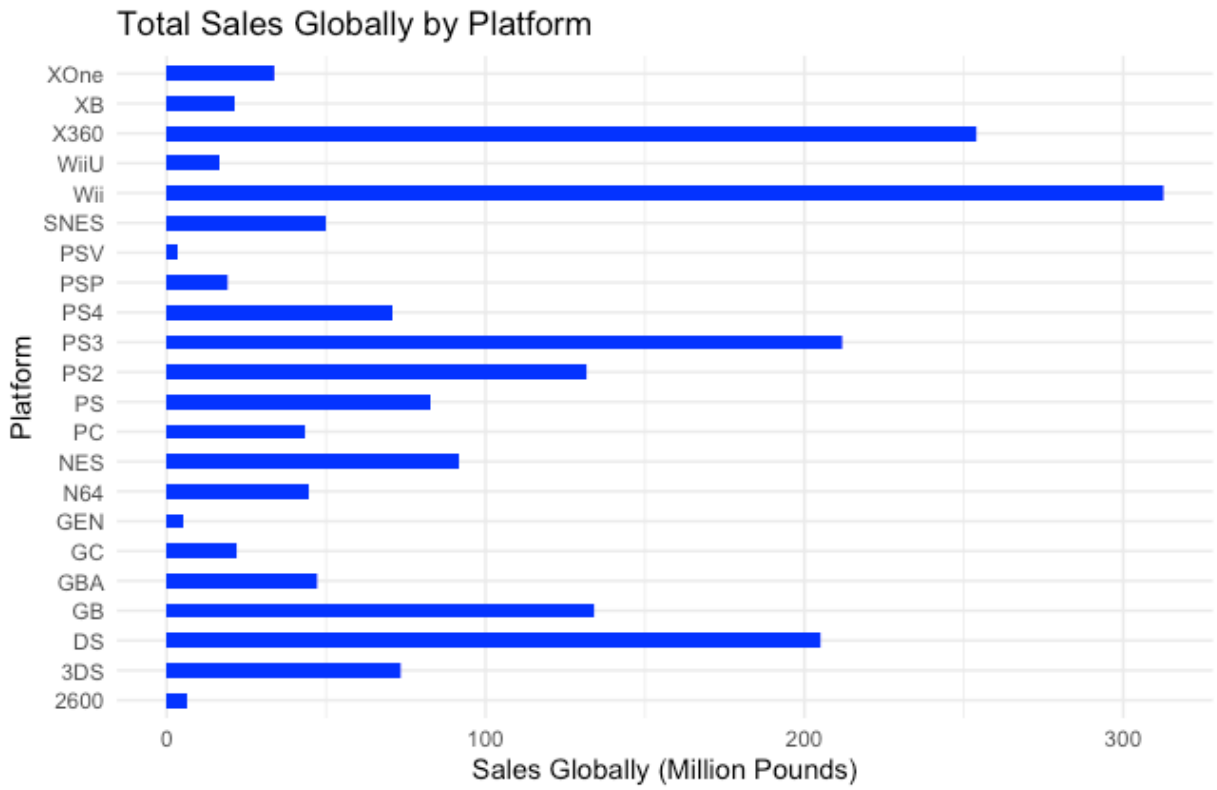
```
    Ranking Product Platform Year   Genre       Publisher NA_Sales EU_Sales
180     180    7141      PS2   NA  Sports Electronic Arts     3.49     0.21
258    1128     948       PC   NA Shooter      Activision     0.48     0.66
    Global_Sales
180         4.29
258         1.34
```

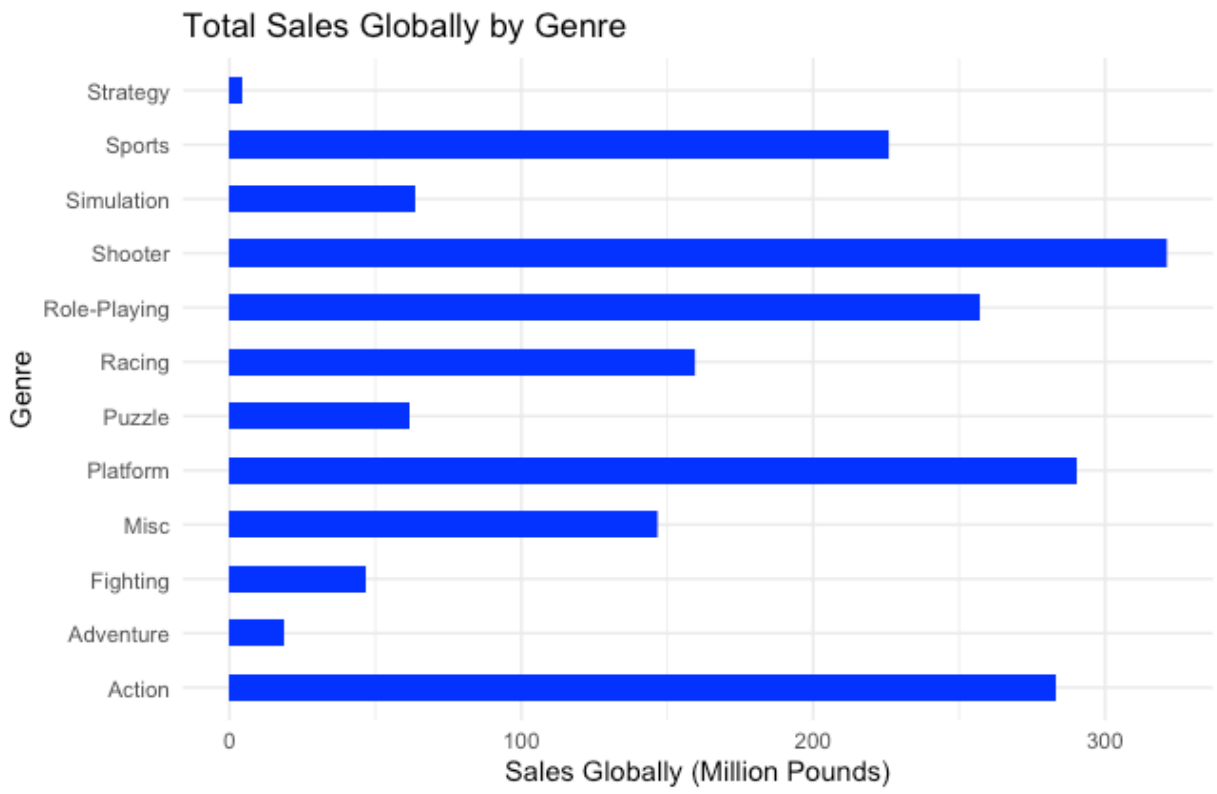
As the data in other columns like the sales values seem valid, we did not remove the rows.

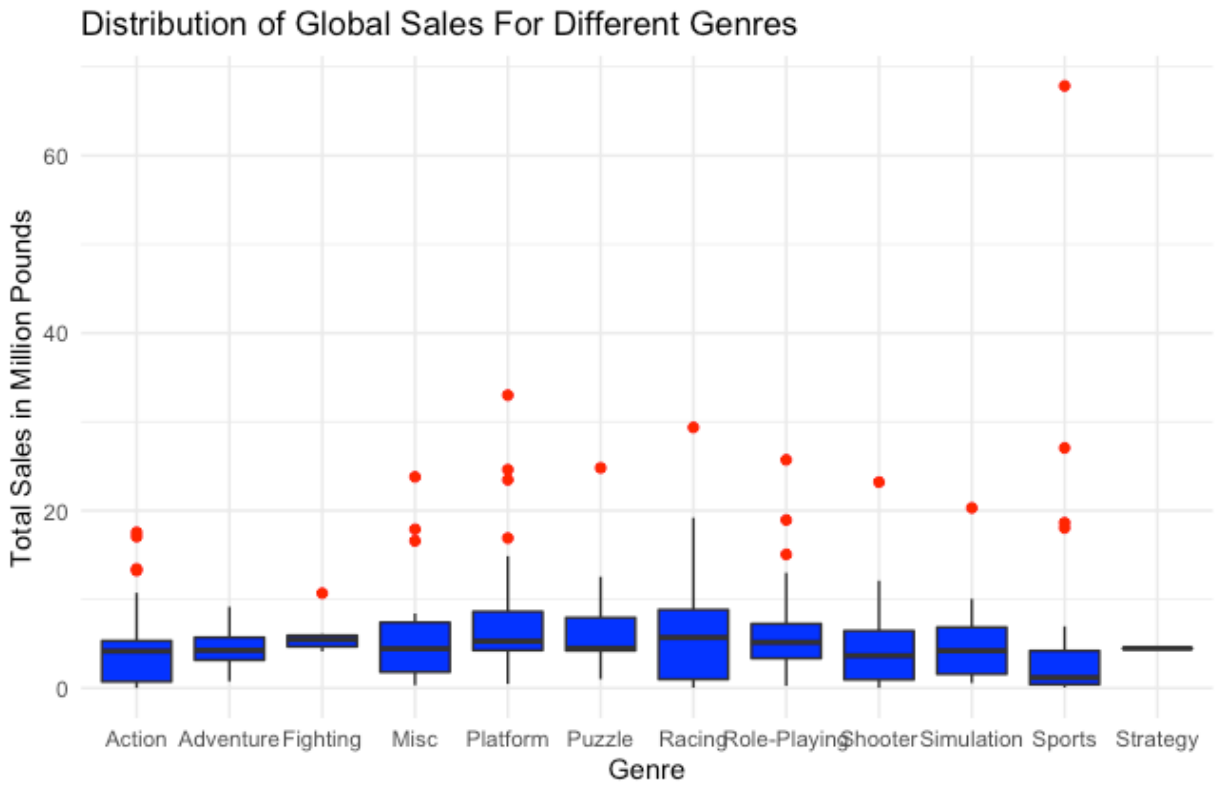
Exploring the Data, we observed the following trends visualized below-:

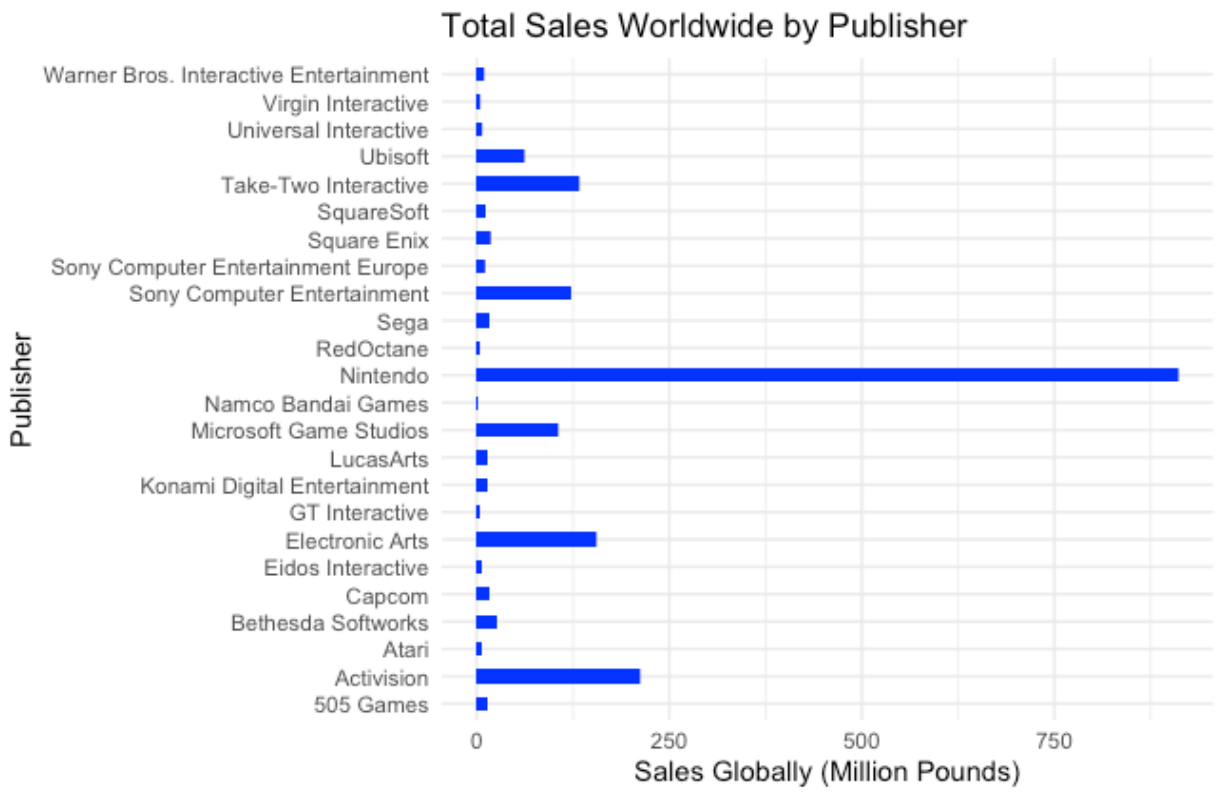





Total Sales Globally by Platform
XOne
XB
X360
WiiU
Wii
SNES
PSV
PSP
PS4
PS3
PS2
PS
PC
NES
N64
GEN
GC
GBA
GB
DS
3DS
2600
Platform
0
100
200
300
Sales Globally (Million Pounds)

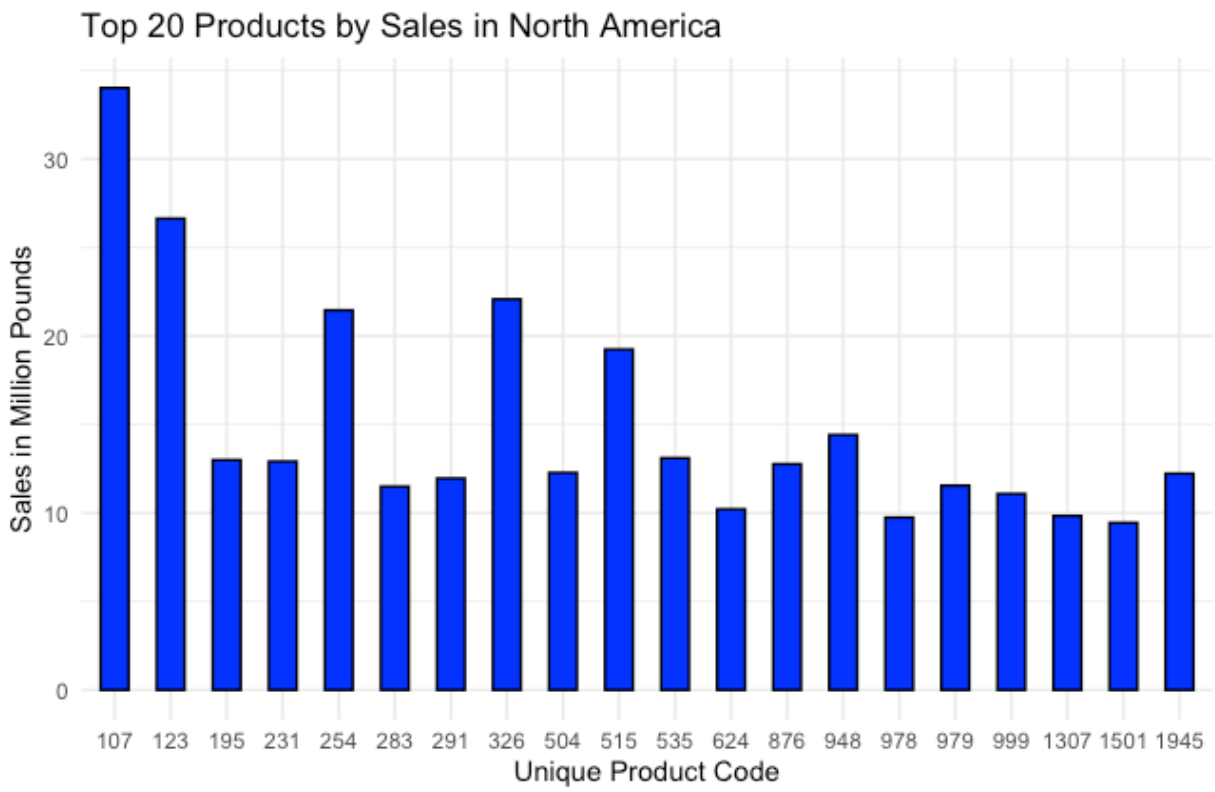

Total Sales Globally by Genre
Strategy
Sports
Simulation
Shooter
Role-Playing
Racing
Puzzle
Platform
Misc
Fighting
Adventure
Action
Genre
0
100
200
300
Sales Globally (Million Pounds)

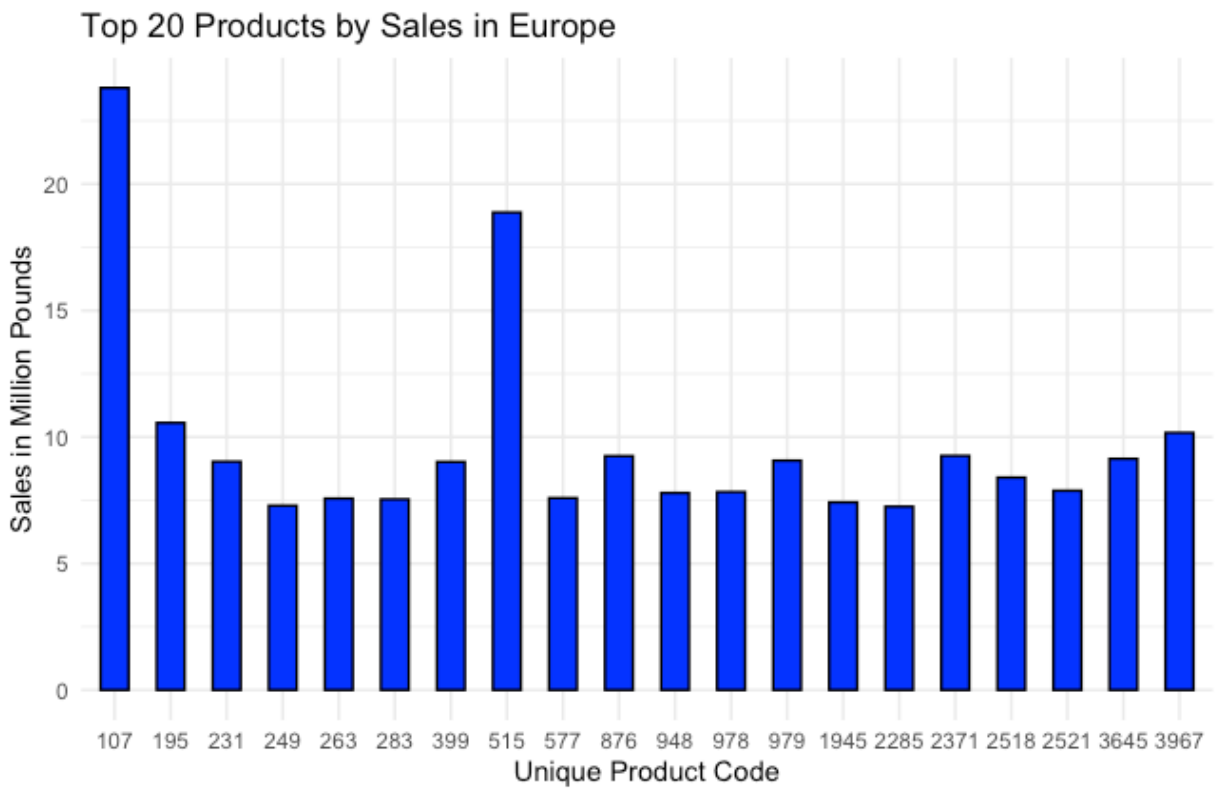

Distribution of Global Sales For Different Genres
Total Sales in Million Pounds
60
40
20
0
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy
Genre

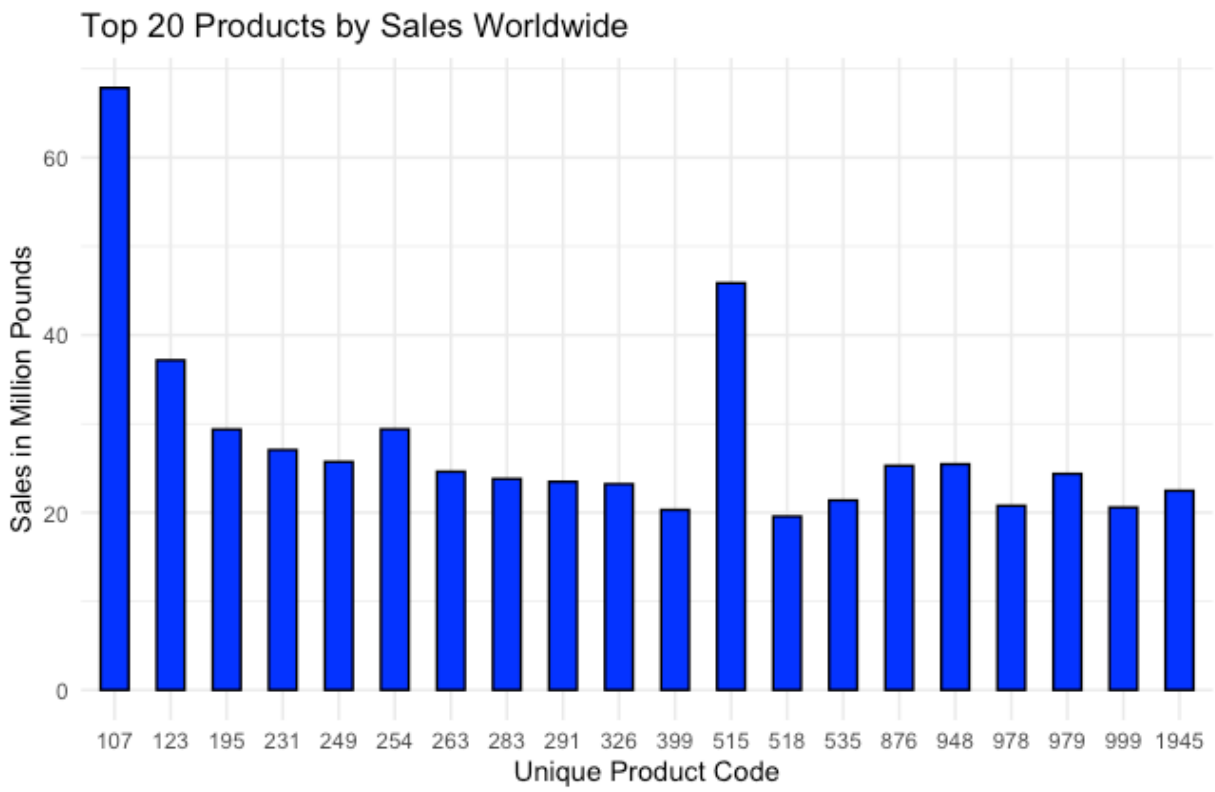


The Impact That Each Product Has on Sales

We aggregated the data by product to gain insight into how the products impact sales.
The graphs below depict the top twenty products sold in North America, Europe and globally by turtle games.


Top 20 Products by Sales in North America
Sales in Million Pounds
0
10
20
30
107 123 195 231 254 283 291 326 504 515 535 624 876 948 978 979 999 1307 1501 1945
Unique Product Code

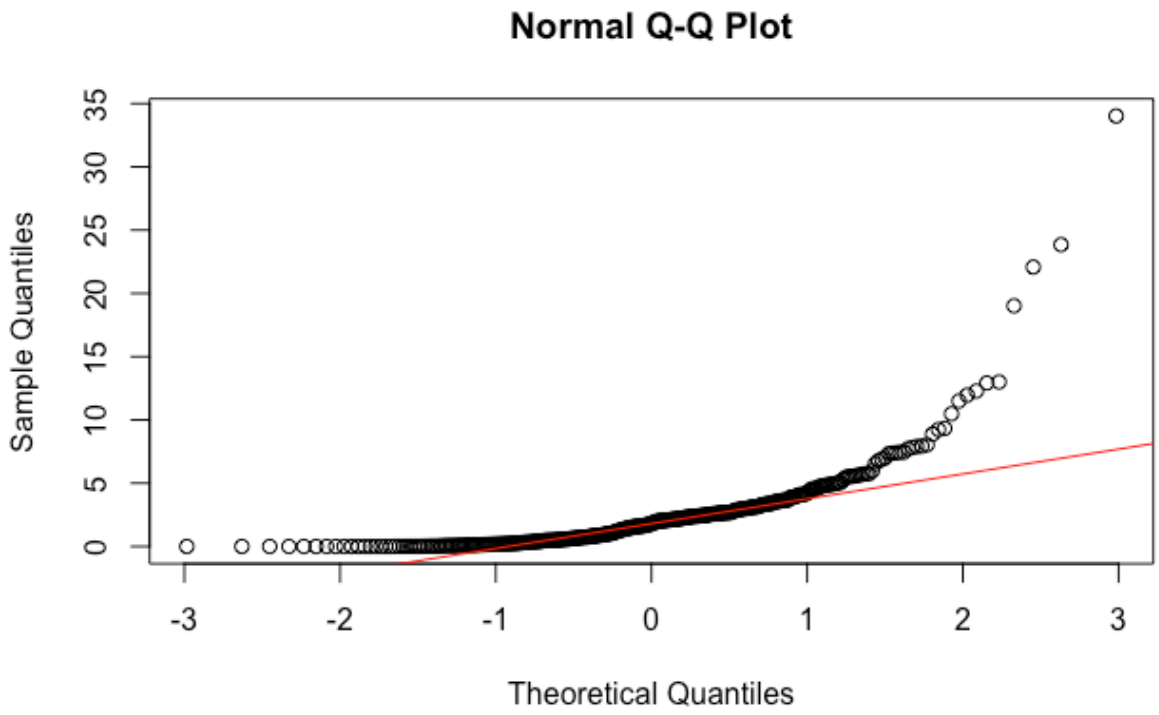

Top 20 Products by Sales in Europe
Sales in Million Pounds
0
5
10
15
20
107 195 231 249 263 283 399 515 577 876 948 978 979 1945 2285 2371 2518 2521 3645 3967
Unique Product Code



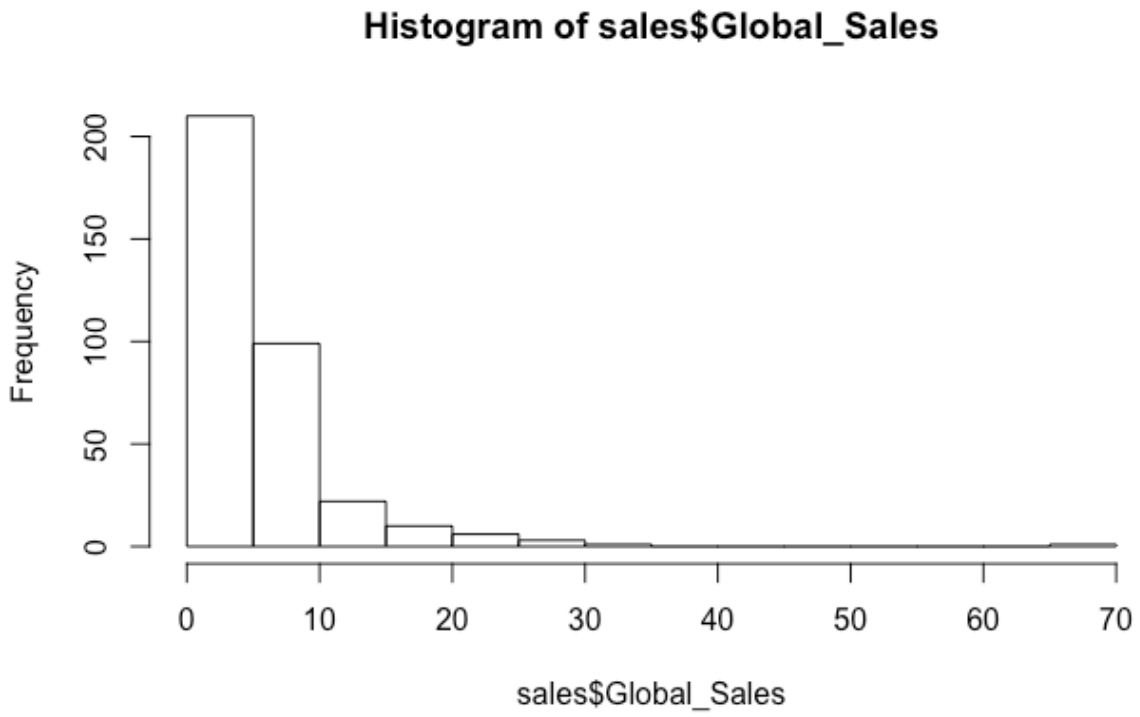
Exploring The Properties Of The Sales Data

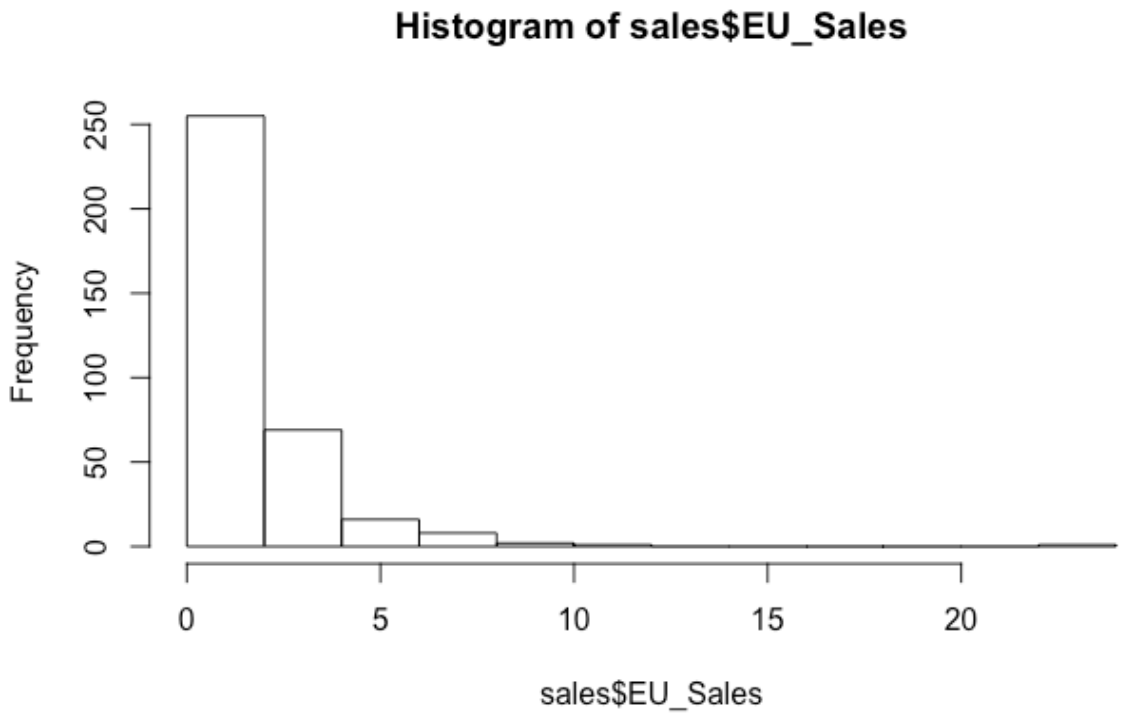
The sales department of Turtle games wants to understand the reliability of the data extracted from its website.

We first tested the data for normality using the Shapiro-wilk's test. The sales data for North America, Europe and Global sales data were not normally distributed as also evident by their Q-Q plots and histograms.

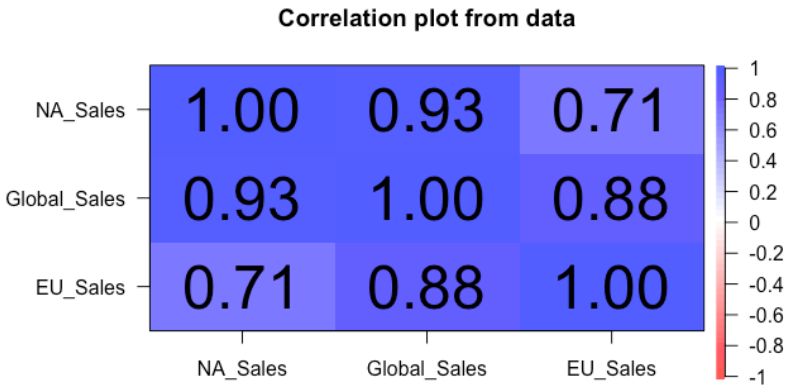


Q-Q plot for NA


Histogram of sales$Global_Sales
Frequency
0
50
100
150
200
0
10
20
30
40
50
60
70
sales$Global_Sales

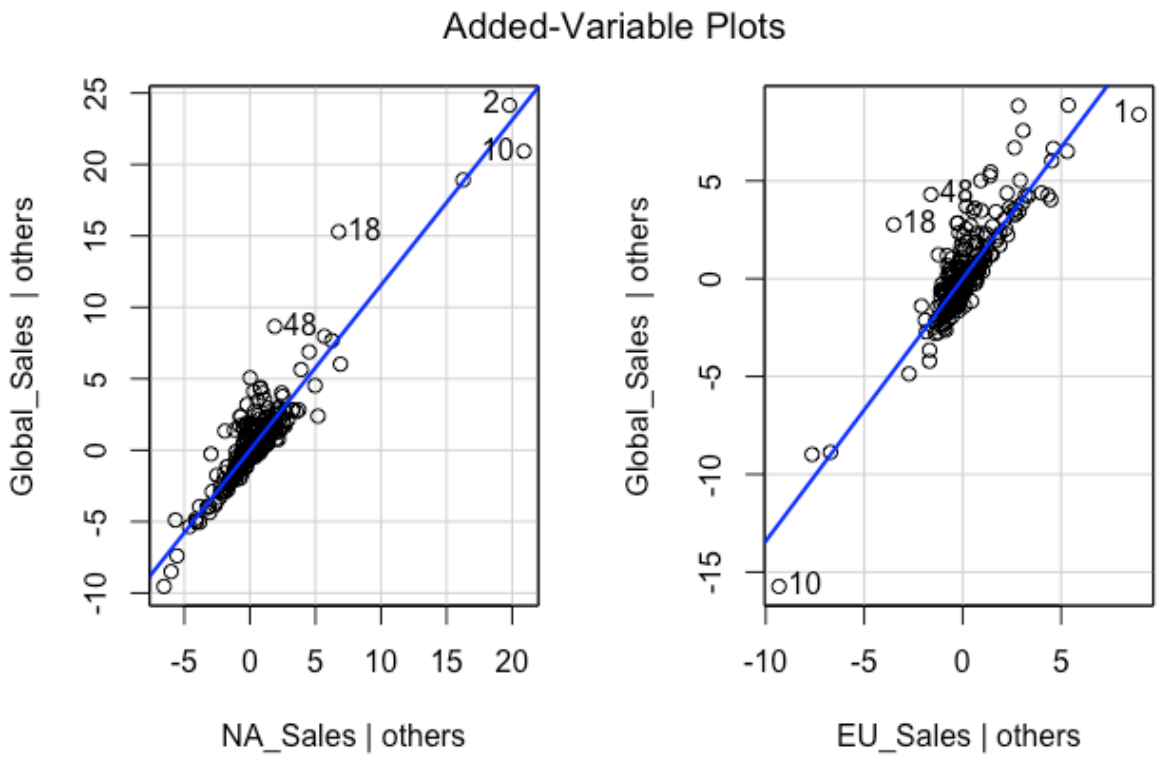


We next checked the skewness to assesses the extent to which a variable's distribution is symmetrical. The North America (NA), Europe (EU) and Global sales columns had data with positive skewness. In a positive skew, the outliers will be present on the right side of the curve indicating entries with high sales in this case.
We then tested the data for correlation. As expected, the NA and EU Sales columns are highly correlated to the Global Sales since global sales includes EU and NA Sales.



The relationship(s) between North American, European, and global sales

After testing for correlation between the North America, Europe and Global sales we further investigated any possible relationship(s) in the sales data by creating a simple and multiple linear regression model.
The multiple linear regression model yielded better results overall than the simple linear regression models.
We can make predictions of Global Sales using the equation where y is the Global Sales value -:
y = 0.22175 +1.15543*NA_Sales + 1.34197*EU_Sales
The Adjusted R-squared of the model is 0.9685 or 96.8 % indicates a good model. However, the residuals from the model are not normally distributed indicating model does not explain all trends and behaviour in the data.

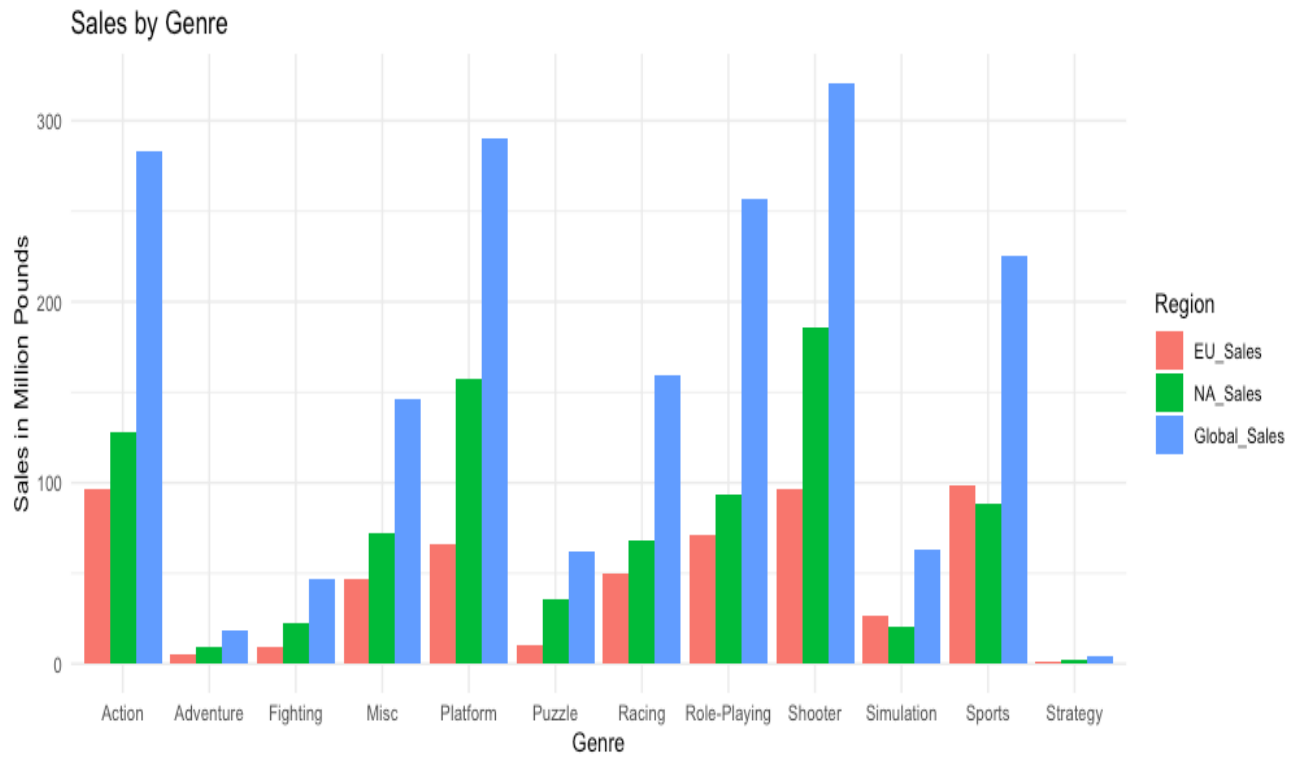


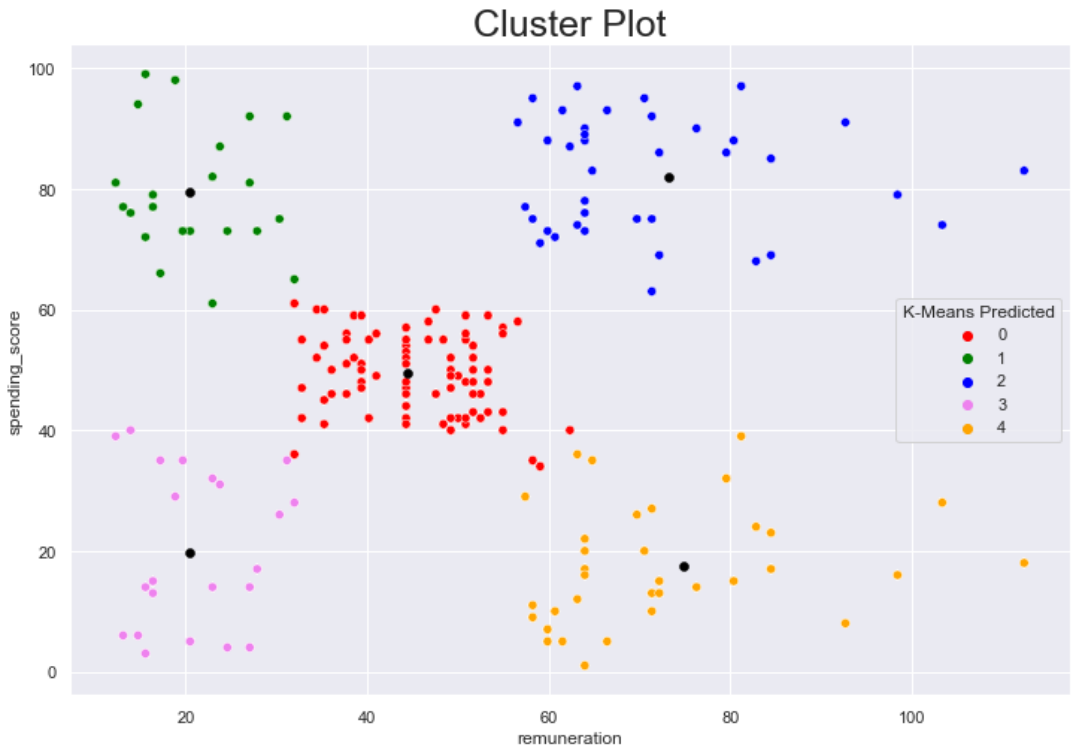
SUMMARY OF INSIGHTS AND RECOMMENDATIONS

- Product with code 107 is the highest selling product in each region. This accounts for almost 3.6 % of sales worldwide.

```
> filter(sales,sales$Product==107)
  Ranking Product Platform Year  Genre Publisher NA_Sales EU_Sales Global_Sales
1       1     107      Wii 2006 Sports  Nintendo    34.02     23.8        67.85
> # total sales percentage worldwide
> round(max(sales$Global_Sales)/sum(sales$Global_Sales),3)*100
[1] 3.6
```

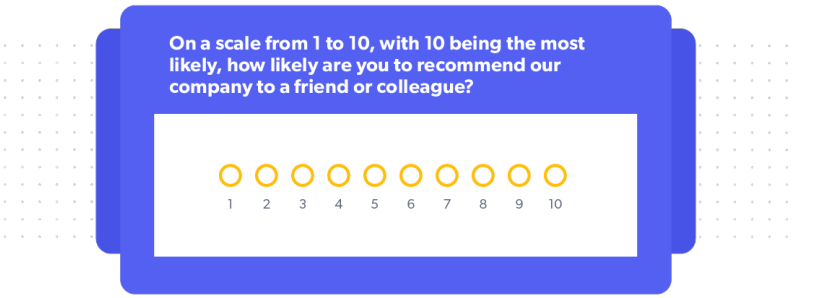
-Nintendo has the highest sales overall as well

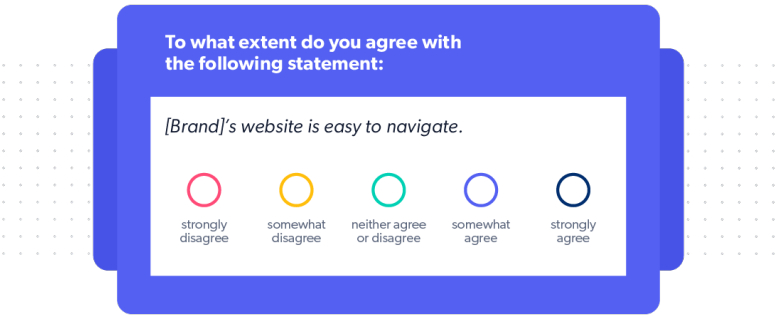

Sales by Genre
Sales in Million Pounds
300
200
100
0
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy
Genre
Region
EU_Sales
NA_Sales
Global_Sales

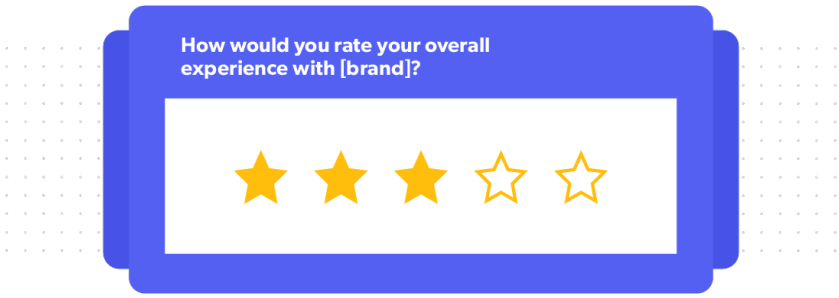


Recommendations -:

- use popular genres in each region to design marketing campaign.
- In Cluster Plot (above) cluster 4 is a high renumeration – low spending score cluster. Use clusters to develop customized campaigns.
- consider building a dashboard for regular monitoring of sales.
- consider conducting a survey to understand customer loyalty and their experience better and more objectively using format given below.


On a scale from 1 to 10, with 10 being the most likely, how likely are you to recommend our company to a friend or colleague?
1 2 3 4 5 6 7 8 9 10


To what extent do you agree with the following statement:
[Brand]'s website is easy to navigate.
strongly disagree
somewhat disagree
neither agree or disagree
somewhat agree
strongly agree


How would you rate your overall experience with [brand]?