Assignment report:
Predicting future outcomes

Turtle Games case: sales performance and customer trends

Lefteris Touzlatzis
e.touzlatis@gmail.com

Contents

Background information of business case

Turtle Games manufactures and sells a variety of products, including board games and video games among others, to an international customer base. The company's goal is to improve overall sales performance by utilizing customer insights.

Observations and recommendations

- Clear understanding on how loyalty points are accumulated was not concluded. Thus, further investigation on variables like recency, frequency and monetary value of customer spending would be necessary. This information could be easily retrieved by connecting with stakeholders internally.
- Employing k-means clustering method with spending score and remuneration the two variables included in the model, there were 5 clear customer segments identified. As a next step we could explore different (categorical) variables and analyze numerical variables based on these segments to understand better their common characteristics and their homogeneity.
- The output of Natural Language Processing analysis (NLP) demonstrates a slight tendency towards the positive rather negative side. Example words striking out are: fun, good great, love, like, five, stars, expansion. The customer feedback on the positive side is concise with little context, while it is manifested as expressions of excitement (e.g. entertaining, great, awesome). Whereas the top 20 negative reviews and their summaries notably illustrate that many customers are disappointed and find the games difficult, especially for kids. This is an area where product and marketing team should focus to improve experience and perceptions relative to the products.
- Preliminary inspection of sales data in North America (NA), Europe (EU) regions and Globally indicate that there are a few outliers and the frequency distribution is skewed. Interrogation of Sales data for normality and reliability confirm pure results, suggesting that sales data from different regions might be required. Plotting product id's and sales, it is noticeable that products with lower numbers perform better in sales. It's worth investigating the way products are assigned with numbers; smaller number might be older and more established products in the market.
- Including both NA and EU Sales in a multiple linear regression model, almost 97% of the variability in Global Sales is explained suggesting that these two regions are a strong driven of Global Sales.

Customer loyalty points

Reviewing the dataset of customer profiles and reviews, the data types and descriptive statistics look reasonable for the variables look good with no missing values. Categorical variables with single value were omitted.

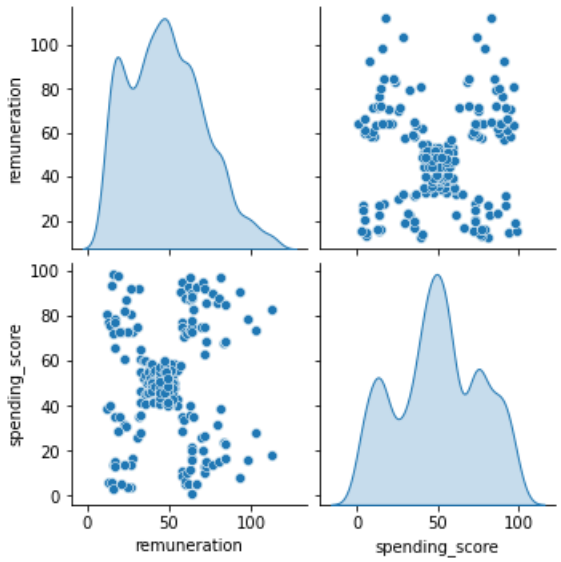
To determine how loyalty points are accumulated, I examined their relationship with three numeric variables of the dataset: spending score, remuneration, age. Three simple linear regressions summaries indicate that none of the 3 variables can predict well the loyalty score. This is due to the fact that R-squared is far below an accepted threshold of 0.60 that could indicate a strong relationship between the independents and dependent variables. That's also visible from the high standard errors as well as the regression plots. Only the

regression Spending VS Loyalty produced the most reasonable results out of the three, but still the R-squared is low (45.2%), standard error with high value with insignificant p-value (>0.05). See Appendix *Fig.1.*

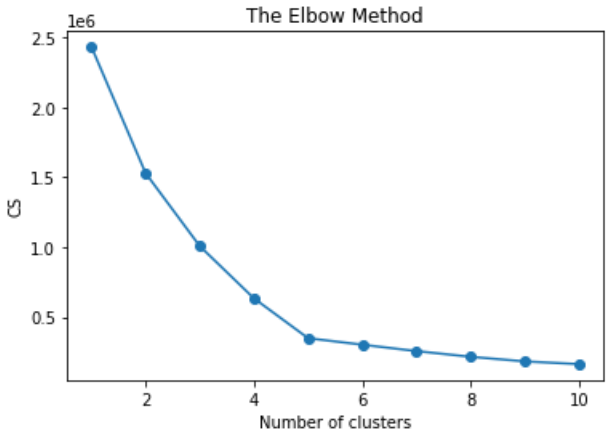
Customer segmentation

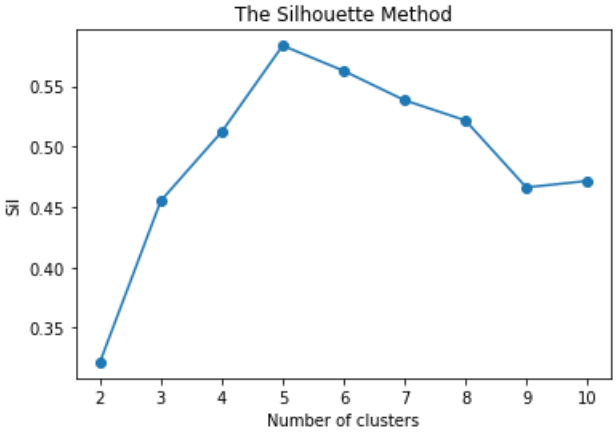
K-means clustering was used to identify segments with similar characteristics based on remuneration and spending score.

Scatterplots shows a pattern of multiple clusters, potentially between 3-5 segments. Pairplots shows the distribution of remuneration (i.e. histogram) is slightly skewed on the right.

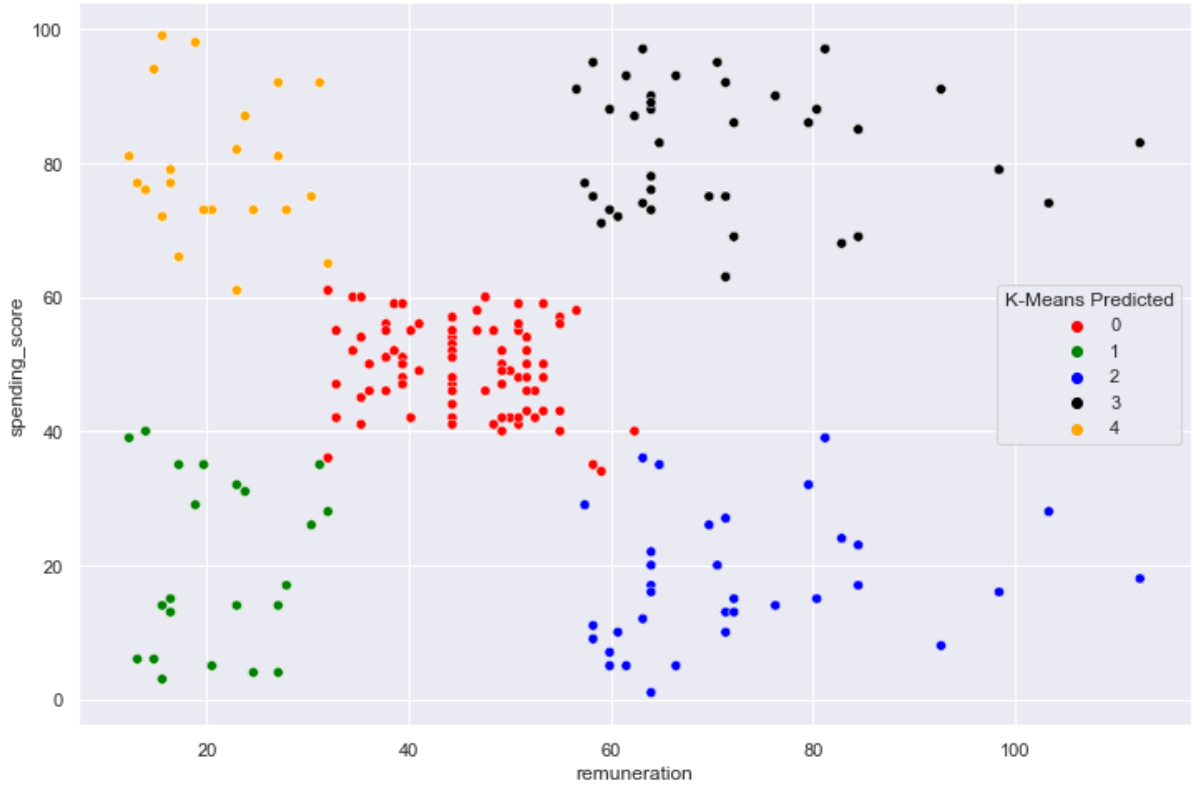


Using two popular methods to determine number of clusters, Elbow indicates the ideal number of clusters is 5 where the line starts flattening, and Silhouette method confirms this number where the mean coefficient value is maximized.





Different number of clusters were evaluated (see Appendix *Fig.2*), but everything leads to the conclusion that 5 is the ideal number of clusters we should pick.



The number of customers falling into these 4 segments are quite balanced, except in the case of cluster 0, where the counts are more than doubled of the rest. However, this cluster seems robust as its observations are unified.

```
0   774
3   356
2   330
1   271
4   269
Name: K-Means Predicted, dtype: int64
```

We could explore further the outliers and see how our model changes in case we remove them. Another point we could take into consideration is how other categorical variables affect the clustering by including them in the visualization.

Customer reviews analysis

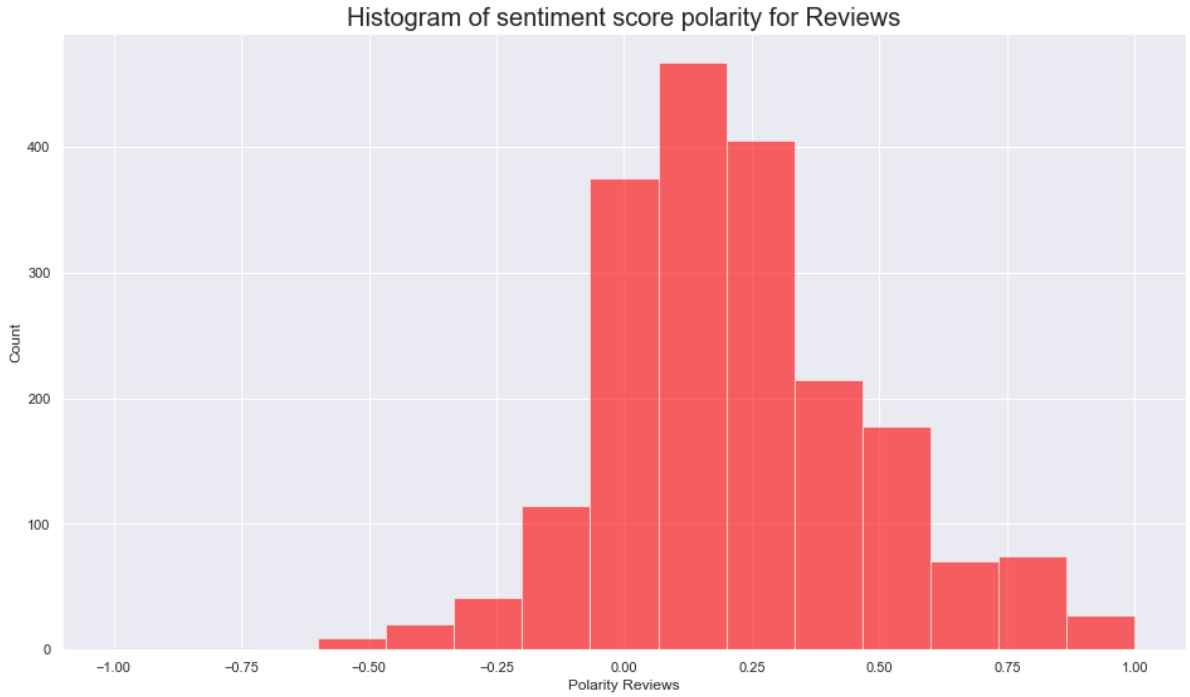
Social data was used for NLP analysis to determine the most common words and positive reviews.

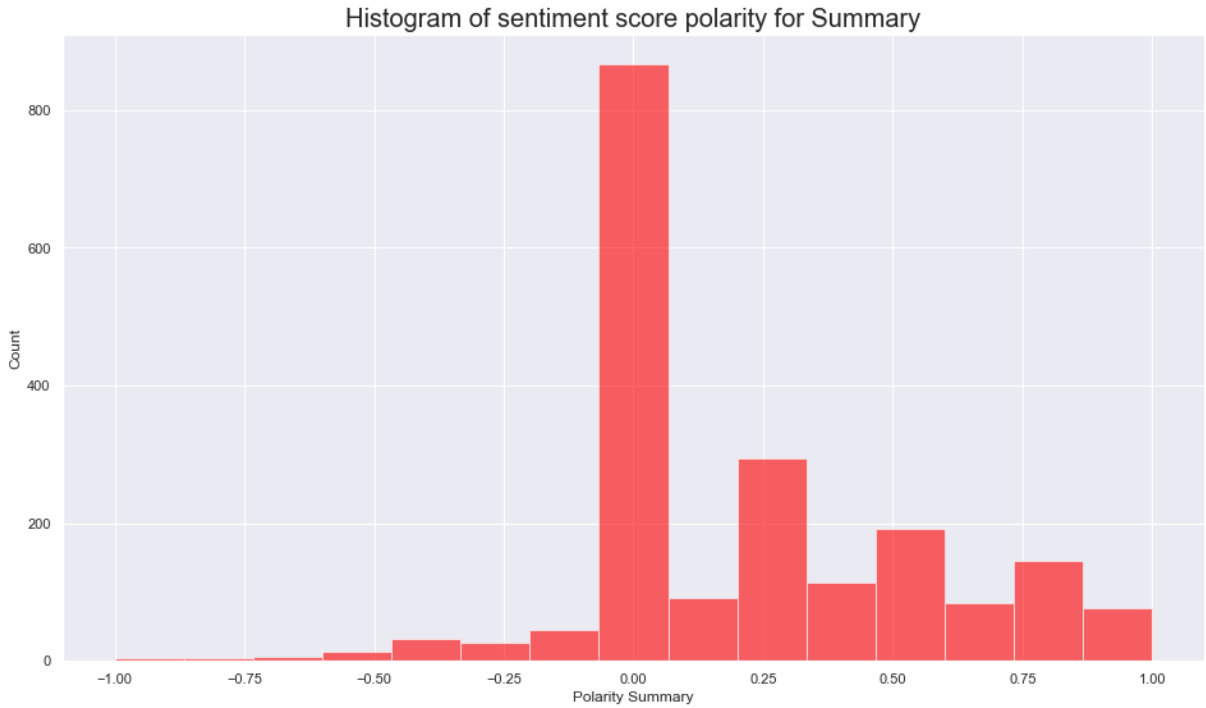
After removing a duplicate review, normalization and tokenization were implemented. Next, Wordclouds and lists of words with the highest frequency were generated, indicating a slight tendency towards the positive.

Wordcloud for Reviews

Wordcloud for Summaries (of the reviews)

To further understand reviews (and summary) sentiment, the polarity scores were explored. Histograms show that sentiment sits slightly more on the positive side for both reviews and summaries.





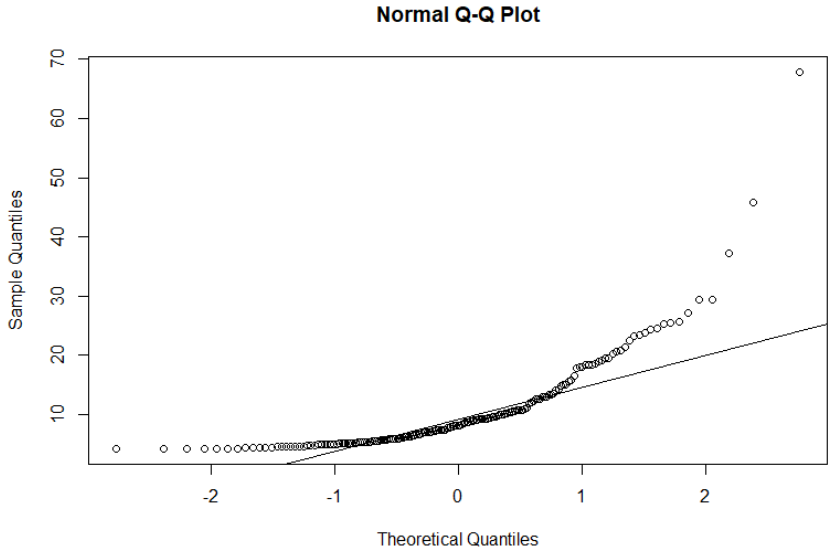
Reading the top 20 negative reviews and summaries through, it's notable that many customers are disappointed and find the games difficult, especially for kids. On the positive side, customer feedback is more concise with less context and manifested as expressions of excitement (see Appendix *Fig.3-5*).

Insights on Products and Sales

Scatterplots show that products can have different popularity between NA and EU as they are scattered, while global sales are more affected by NA as the relationship looks more linear. Frequency distribution of sales are skewed on the right and boxplots show a lot of outliers (see Appendix *Fig.6*).

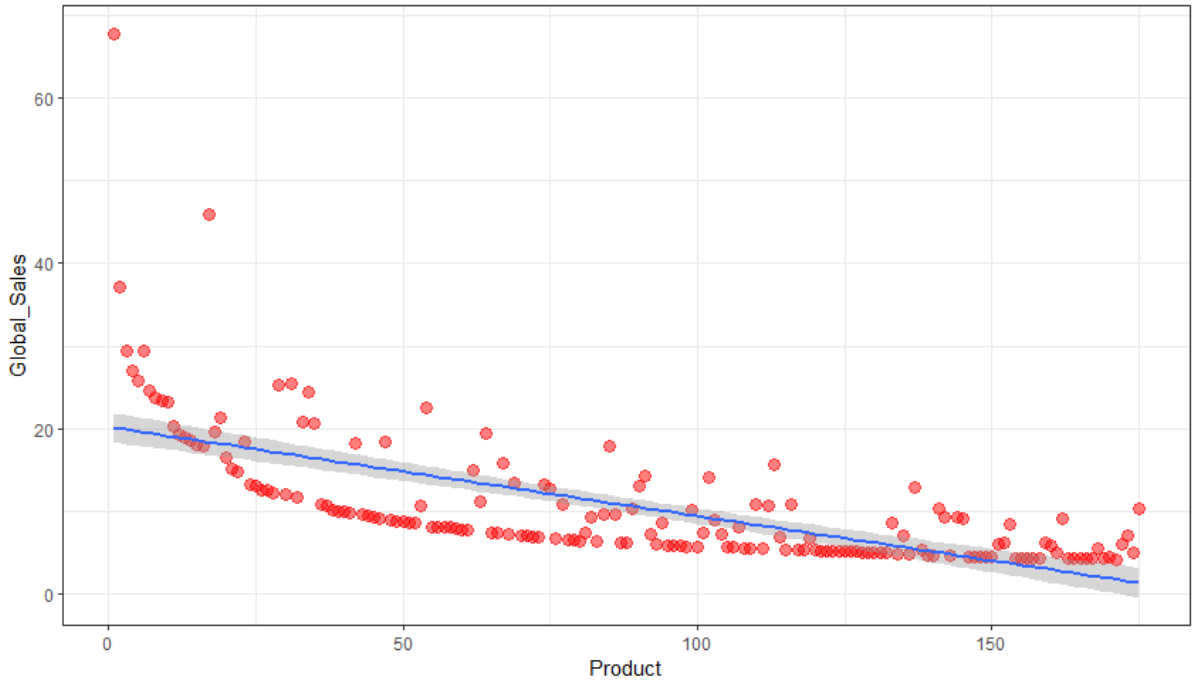
Reliability and manipulation of data

Examination of Sales data was performed and their relationship with products was inspected. Data points in Q-Q plots deviate from the straight line, indicating that we cannot assume normality.



This is also confirmed by the Shapiro-Wilk normality tests, where p-value is much lower than significance level 5%. Skewness and Kurtosis were also outside the accepted values of absolute 1 and 3 respectively.

Plotting the relationship between the Product and Sales, a worth mentioning trend is observed; the products with lower id tend to have higher sales.



Impact of NA and EU sales to Global sales

The correlation matrix between the variables show that Global sales have strong relationship with the sales in the two areas, especially with North America (NA).

	Product	NA_Sales	EU_Sales	Global_Sales
Product	1.0000000	-0.6084875	-0.4886596	-0.6719812
NA_Sales	-0.6084875	1.0000000	0.6209317	0.9162292
EU_Sales	-0.4886596	0.6209317	1.0000000	0.8486148
Global_Sales	-0.6719812	0.9162292	0.8486148	1.0000000

Simple regression between Global and NA Sales gives a high Multiple R-squared, indicating that NA Sales explains 84% of the variability in Global Sales (see Appendix *Fig. 8*). Significant low p-value and positive x coefficient shows there is a positive relationship, where every 1 million sales increase in NA there is a 1.63 million in Global Sales. However, plotting the residuals and line of best fit in the relationship imply the possibility of a nonlinear relationship.

Simple regression between Global and EU Sales shows a weak model with lower Multiple R-squared (72%) than with NA and higher residual values for many of the observations.

Including both the two regions (i.e. NA and EU) as explanatory variables for Global Sales and fitting a multiple regression model, we manage to explain almost 97% of the variability in Global Sales and get a significant contribution from these 2 variables (see Appendix *Fig. 9*). We can still see a slight pattern in the residual, but overall, the model demonstrates adequately that Sales in NA and EU can predict the Global Sales with confidence.

Appendix

Fig.1 Spending score VS Loyalty linear regression

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.
Date:	Tue, 30 Aug 2022	**Prob (F-statistic):**	2.92e-263
Time:	18:29:36	**Log-Likelihood:**	-16550.
No. Observations:	2000	**AIC:**	3.310e+04
Df Residuals:	1998	**BIC:**	3.312e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-75.0527	45.931	-1.634	0.102	-165.129	15.024
x	33.0617	0.814	40.595	0.000	31.464	34.659

Omnibus:	126.554	**Durbin-Watson:**	1.191
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	260.528
Skew:	0.422	**Prob(JB):**	2.67e-57
Kurtosis:	4.554	**Cond. No.**	122.

Fig.2 k-means model with 4 clusters

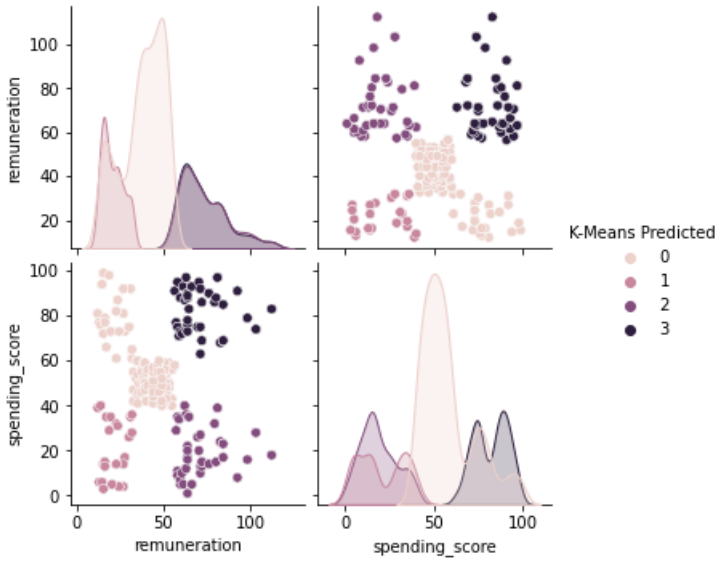


Fig.3 15 most common words for Reviews and Summaries

Review_Word	Frequency
game	1684
great	595
fun	553
one	530
play	502
like	414
love	331
really	319
get	319
cards	301
tiles	297
good	294
time	291
would	280
book	273

Summary_Word	Frequency
stars	465
five	380
game	319
great	295
fun	218
love	93
good	92
four	58
like	54
expansion	52
kids	50
cute	45
book	43
one	38
awesome	36

Fig.4 20 top negative Reviews and Summaries

	neg	neu	pos	compound
difficult	1.000	0.000	0.000	-0.3612
incomplete kit very disappointing	0.538	0.462	0.000	-0.5413
no more comments	0.524	0.476	0.000	-0.2960
a crappy cardboard ghost of the original hard to believe they did this but they did shame on hasbro disgusting	0.487	0.455	0.058	-0.9052
not a hard game to learn but not easy to win	0.470	0.456	0.075	-0.7946
i found the directions difficult	0.455	0.545	0.000	-0.3612
who doesnt love puppies great instructions pictures fun	0.445	0.334	0.221	-0.5207
different kids had red faces not sure they like	0.368	0.632	0.000	-0.4717
got the product in damaged condition	0.367	0.633	0.000	-0.4404
i bought this thinking it would be really fun but i was disappointed its really messy and it isnt nearly as easy as it seems also the glue is useless for a 9 year old the instructions are very difficult	0.362	0.592	0.045	-0.9520
great game poor quality	0.337	0.217	0.446	0.2500
we really did not enjoy this game	0.325	0.675	0.000	-0.4389
not as easy as it looks	0.325	0.675	0.000	-0.3412
hard to put together	0.318	0.682	0.000	-0.1027
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	0.318	0.613	0.069	-0.8674
easytouse great for anger management groups	0.314	0.339	0.347	0.1027
its ok but loses its luster quickly	0.309	0.524	0.168	-0.3291
rather hard for my 11 year old to do alone	0.298	0.702	0.000	-0.3400
smaller than we thought kind of disappointed in it	0.298	0.702	0.000	-0.5256
i really like this game it helps kids recognize anger and talk about difficult emotions	0.287	0.463	0.250	-0.2040

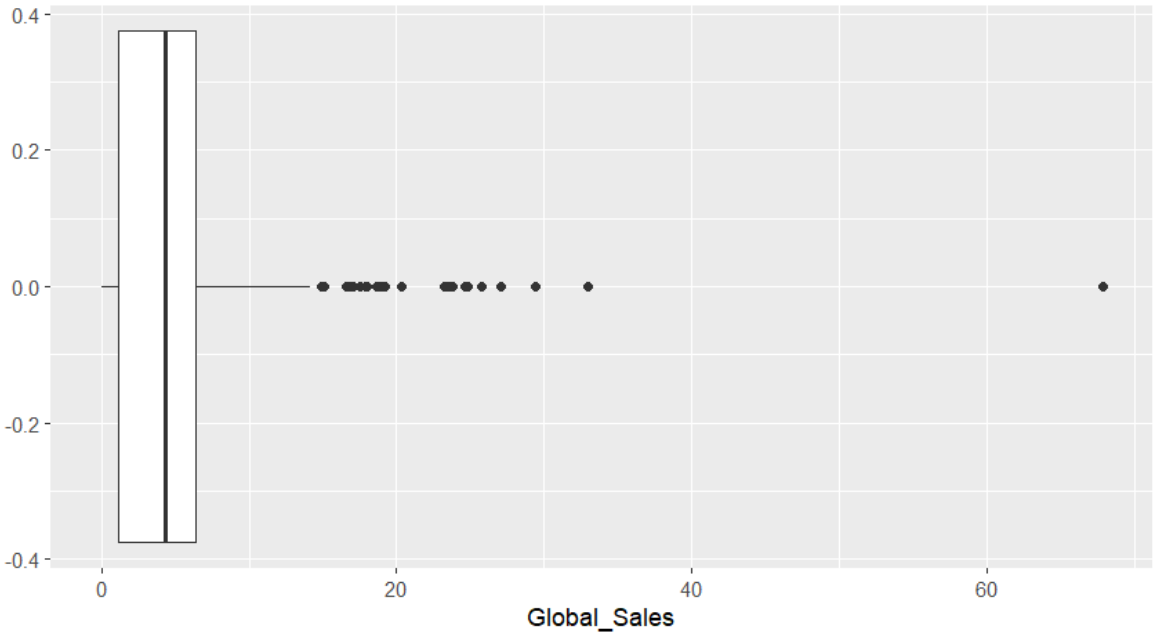
	neg	neu	pos	compound
disappointed	1.000	0.000	0.0	-0.4767
boring	1.000	0.000	0.0	-0.3182
disappointing	1.000	0.000	0.0	-0.4939
frustrating	1.000	0.000	0.0	-0.4404
meh	1.000	0.000	0.0	-0.0772
defective poor qc	0.857	0.143	0.0	-0.7184
not great	0.767	0.233	0.0	-0.5096
mad dragon	0.762	0.238	0.0	-0.4939
no 20 sided die	0.753	0.247	0.0	-0.7269
damaged product	0.744	0.256	0.0	-0.4404
faulty product	0.697	0.303	0.0	-0.3182
money trap	0.697	0.303	0.0	-0.3182
nothing special	0.693	0.307	0.0	-0.3089
wimpy magnets	0.655	0.345	0.0	-0.2263
anger control game	0.649	0.351	0.0	-0.5719
box totally destroyed	0.636	0.364	0.0	-0.5413
really small disappointed	0.628	0.372	0.0	-0.5233
a disappointing coop game	0.615	0.385	0.0	-0.4939
da bomb game	0.615	0.385	0.0	-0.4939
very weak game	0.615	0.385	0.0	-0.4927

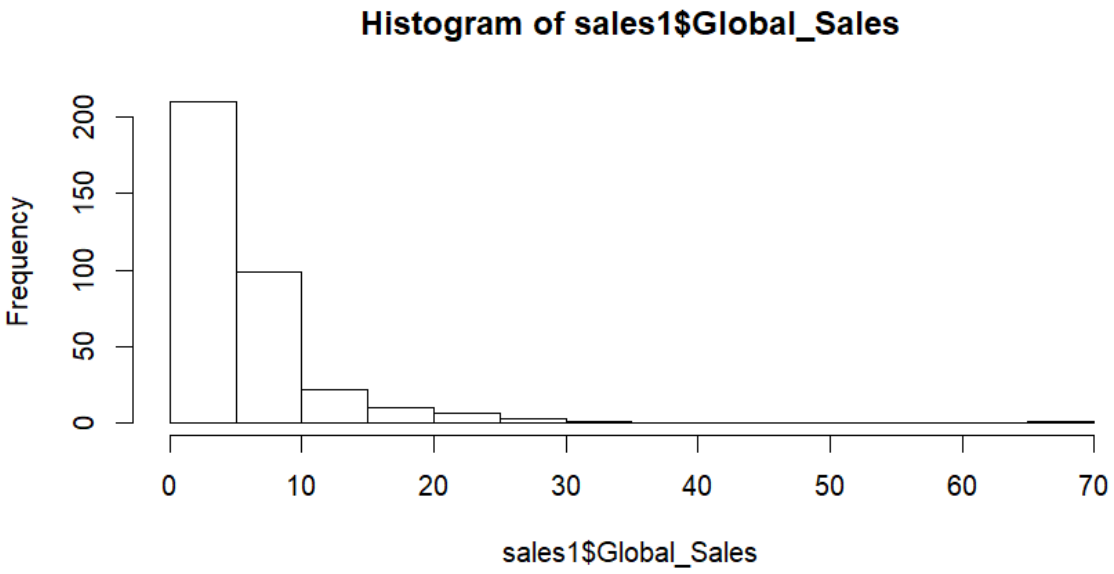
Fig.5 20 top positive Reviews and Summaries

	neg	neu	pos	compound
entertaining	0.0	0.0	1.0	0.4404
fun gift	0.0	0.0	1.0	0.7351
ok	0.0	0.0	1.0	0.2960
cool	0.0	0.0	1.0	0.3182
great	0.0	0.0	1.0	0.6249
fantastic	0.0	0.0	1.0	0.5574
satisfied thanks	0.0	0.0	1.0	0.6908
awesome	0.0	0.0	1.0	0.6249
satisfied	0.0	0.0	1.0	0.4215
nice	0.0	0.0	1.0	0.4215
cute	0.0	0.0	1.0	0.4588
outstanding	0.0	0.0	1.0	0.6124
loved loved loved	0.0	0.0	1.0	0.9136
fine	0.0	0.0	1.0	0.2023
fun	0.0	0.0	1.0	0.5106
good	0.0	0.0	1.0	0.4404
gift	0.0	0.0	1.0	0.4404
perfect	0.0	0.0	1.0	0.5719
thx	0.0	0.0	1.0	0.3612
super fun	0.0	0.0	1.0	0.8020

	neg	neu	pos	compound
wonderful	0.0	0.0	1.0	0.5719
awesome	0.0	0.0	1.0	0.6249
nifty	0.0	0.0	1.0	0.4019
brilliant	0.0	0.0	1.0	0.5859
thanks	0.0	0.0	1.0	0.4404
good	0.0	0.0	1.0	0.4404
great helper	0.0	0.0	1.0	0.7579
great gift	0.0	0.0	1.0	0.7906
great	0.0	0.0	1.0	0.6249
super cute	0.0	0.0	1.0	0.7845
fun	0.0	0.0	1.0	0.5106
ok	0.0	0.0	1.0	0.2960
perfect	0.0	0.0	1.0	0.5719
precious	0.0	0.0	1.0	0.5719
pretty cool	0.0	0.0	1.0	0.6705
love	0.0	0.0	1.0	0.6369
beautiful	0.0	0.0	1.0	0.5994
useful	0.0	0.0	1.0	0.4404
wow	0.0	0.0	1.0	0.5859
great fun	0.0	0.0	1.0	0.8126

Fig.6 Scatterplots, histograms and boxplots on Sales data


0.4
0.2
0.0
-0.2
-0.4
0
20
40
60
Global_Sales


Histogram of sales1$Global_Sales
Frequency
0
50
100
150
200
0
10
20
30
40
50
60
70
sales1$Global_Sales

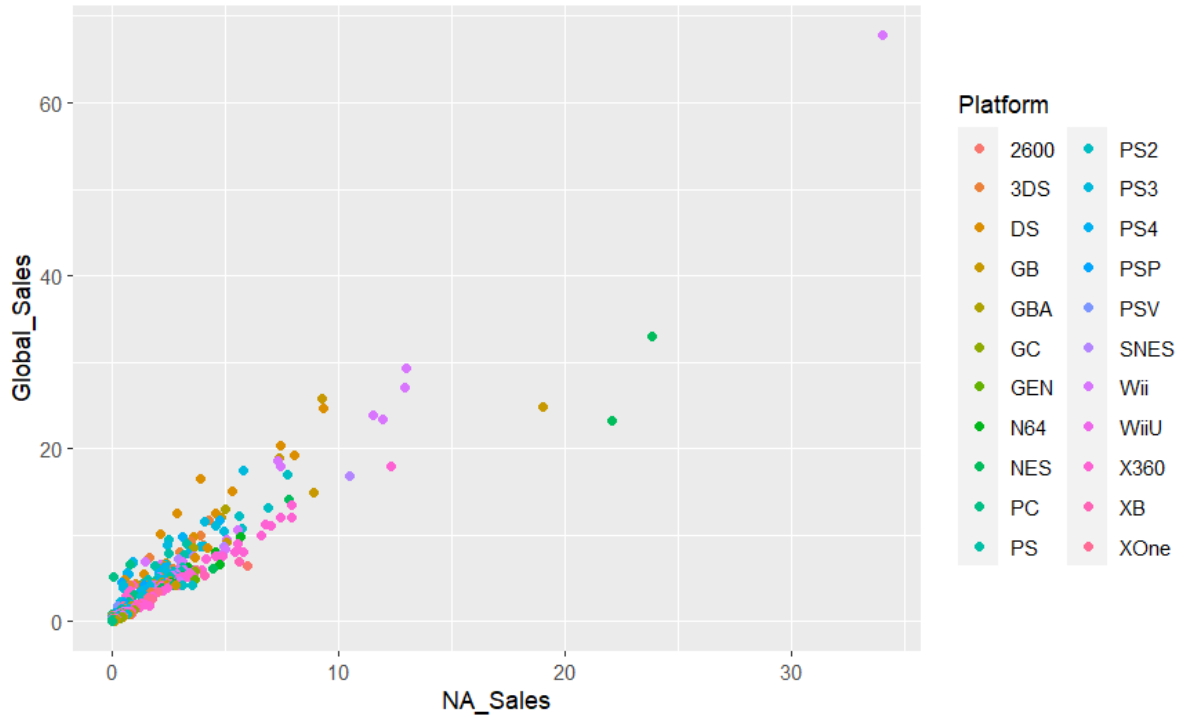


Fig.7 Shapiro-Wilk normality test on Sales

data: product_sales$Global_Sales

W = 0.70955, p-value < 2.2e-16

data: product_sales$NA_Sales

W = 0.69813, p-value < 2.2e-16

data: product_sales$EU_Sales

W = 0.74058, p-value = 2.987e-16

Fig.8 Skewness and Kurtosis of Sales

skewness(product_sales$Global_Sales)

[1] 3.066769

kurtosis(product_sales$Global_Sales)

[1] 17.79072

Fig.8 Simple Linear Regression results

```
Call:
lm(formula = Global_Sales ~ NA_Sales, data = product_sales)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3417  -1.8198  -0.5933   1.4322  11.9345

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.45768    0.36961   6.649 3.71e-10 ***
NA_Sales     1.63469    0.05435  30.079  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.266 on 173 degrees of freedom
Multiple R-squared:  0.8395,    Adjusted R-squared:  0.8385
F-statistic: 904.7 on 1 and 173 DF,  p-value: < 2.2e-16
```

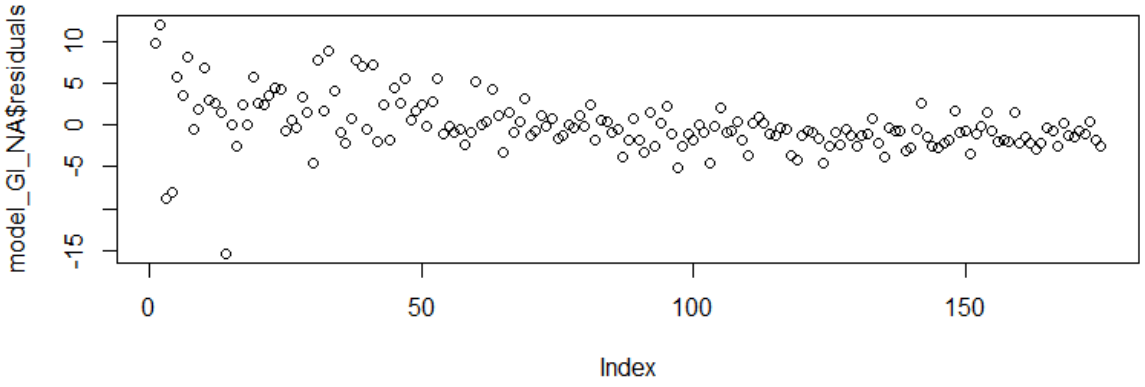


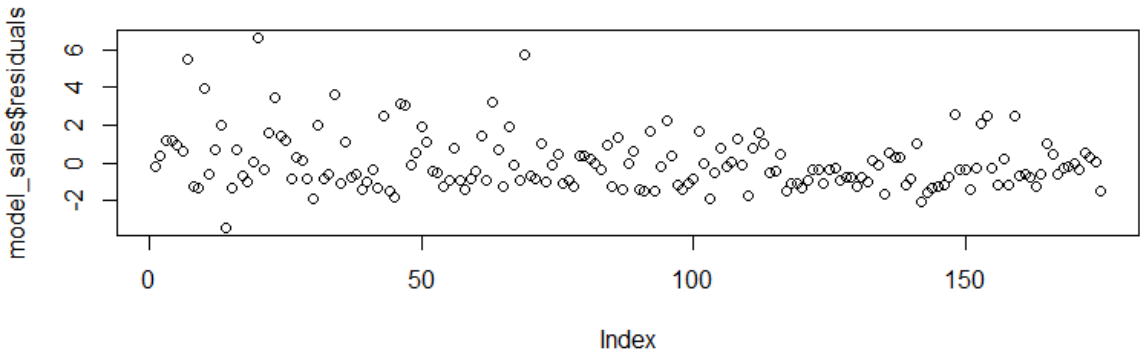
Fig.9 Multiple Linear Regression

```
Call:
lm(formula = Global_Sales ~ NA_Sales + EU_Sales, data = product_sales)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4156 -1.0112 -0.3344  0.6516  6.6163

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.04242    0.17736   5.877 2.11e-08 ***
NA_Sales     1.13040    0.03162  35.745  < 2e-16 ***
EU_Sales     1.19992    0.04672  25.682  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.49 on 172 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.9664
F-statistic:  2504 on 2 and 172 DF,  p-value: < 2.2e-16
```


model_sales$residuals
-2
0
2
4
6
0
50
100
150
Index