Turtle Games
Data Analytics on the Road to Enhanced Sales Performance

Suggested actions towards **Enhanced Sales Performance** are gathered under five separate headings:

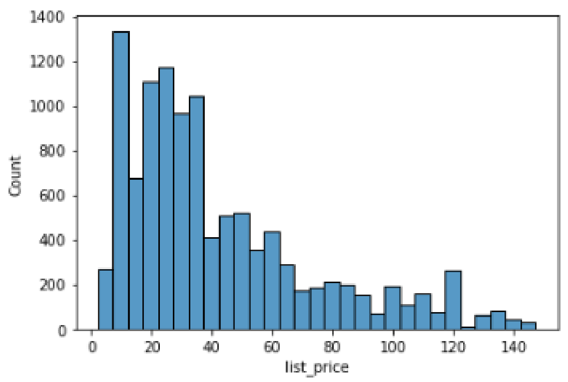
1. **Pricing Model – using linear regression:**

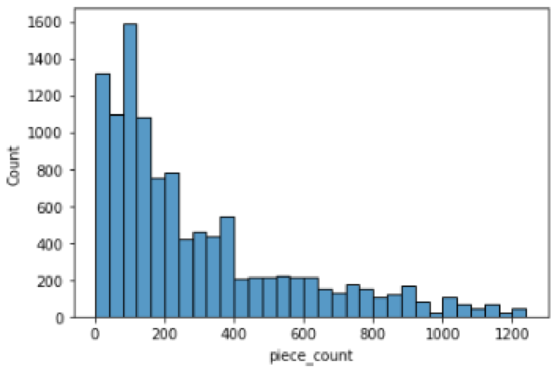
Description and distribution of the data:

Out[5]:

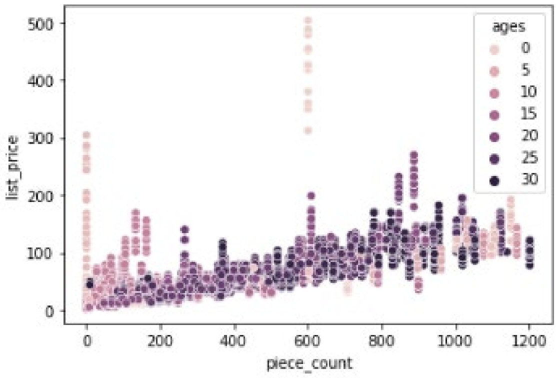
	ages	list_price	num_reviews	piece_count	play_star_rating	review_difficulty	country
count	12261.00000	12261.000000	12261.000000	12261.000000	12261.000000	12261.000000	12261.000000
mean	16.68828	65.141998	14.603050	493.405921	3.709689	1.988826	10.015333
std	8.21868	91.980429	34.356847	825.364580	1.641130	1.787565	6.185450
min	0.00000	2.272400	0.000000	1.000000	0.000000	0.000000	0.000000
25%	11.00000	19.990000	1.000000	97.000000	3.600000	0.000000	4.000000
50%	19.00000	36.587800	4.000000	216.000000	4.400000	2.000000	10.000000
75%	23.00000	70.192200	11.000000	544.000000	4.700000	4.000000	15.000000
max	30.00000	1104.870000	367.000000	7541.000000	5.000000	5.000000	20.000000

Understanding pricing trends:





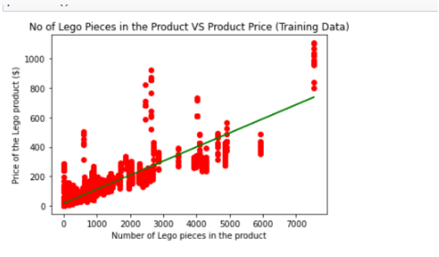
List_price and piece_count data are right skewed.



There seems to be a positive linear correlation between the two data sets with no specific emphasis on 'age'.

Let's predict price applying Linear Regression model to train and test data to allow to simulate how a model would perform on new/unseen data:

- Simple Linear Regression model (predictor variables: number of pieces)




```
# print the R-squared value
print(lr.score(x_train,y_train))
```

```
0.7529271656910888
```

```
# print Intercept and Coefficient
print("Intercept value: ", lr.intercept_)
print("Coffecient value: ", lr.coef_)
```

```
Intercept value:  17.634791702797614
Coffecient value:  [0.09553496]
```

R-squared: Strong R-squared value, as it is higher than 0.7, and it explains almost 75% of the dependent variable.

Coefficient: Each additional lego piece is associated with an increase in the product price of $0,1 (10 cent).

Optimum Price:

In [56]:
```
# Make predictions: price of lego product with 8000 pieces
def calc(slope, intercept, lego_pieces):
    return slope*lego_pieces+intercept

score = calc(0.09553496, 17.634791702797614, 8000)
print(score)
```

```
781.9144717027976
```

- Multiple Linear Regression model (predictor variables: number of pieces and ages)

```
# Checking the value of R-squared, intercept and coefficients
print("R-squared: ", multi.score(x_train, y_train))
print("Intercept: ", multi.intercept_)
print("Coefficients:")
list(zip(x_train, multi.coef_))
```

```
R-squared:  0.7681985466459664
Intercept:  16.98559674920356
Coefficients:

[('piece_count', 0.09569755116044477), ('ages', 0.02987278094702085)]
```

R-squared: Roughly 77% of the variation in Lego prices can be explained using this data set with the pieces and ages variables.

Coefficients: 0.0957 would be the increase in the price of a Lego product with that additional piece. 0.02987 would be the increase in the price of a Lego product with that additional age. These coefficients represent the sensitivity of the dependent variable to unit changes in the respective independent variable.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             list_price   R-squared:                       0.768
Model:                            OLS   Adj. R-squared:                  0.768
Method:                 Least Squares   F-statistic:                 1.422e+04
Date:                Wed, 06 Jul 2022   Prob (F-statistic):               0.00
Time:                        04:18:30   Log-Likelihood:                -44428.
No. Observations:                8582   AIC:                         8.886e+04
Df Residuals:                    8579   BIC:                         8.888e+04
Df Model:                           2
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const          16.9856      1.107     15.341      0.000      14.815      19.156
piece_count     0.0957      0.001    167.909      0.000       0.095       0.097
ages            0.0299      0.056      0.530      0.596      -0.081       0.140
==============================================================================
Omnibus:                     9842.577   Durbin-Watson:                   1.962
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1530257.494
Skew:                           5.830   Prob(JB):                         0.00
Kurtosis:                      67.370   Cond. No.                     2.28e+03
==============================================================================
```

the standard error: the smaller, the better.

T-test statistics: the smaller the standard error, i.e. the more precise the perimeter estimates are, then other things equal, the larger the T values would be.

P-values: the probability of the test statistics value. There's an inverse relationship between the T-value and the P-value. We interpret P-values by comparing them to a significant level 5%. P-value for ‘ages’ is greater than 0.05.

Confidence interval: The confidence interval for 'ages' includes zero. We can conclude that the true coefficient of 'ages' is equal to zero. 'Ages' do not have a statistically significant relationship with ‘price.’

Mean Absolute Error = `21.636188032626283` – MAE is the absolute difference between the actual values and the predicted values. The lower the value, the better is the model’s performance.

Optimum price:

```
# Make predictions: price of lego product with 8000 pieces that are most likely to be purchased by 30 year olds
New_Value1 = 8000
New_Value2 = 29
print ('Predicted Value: \n', multi.predict([[New_Value1 ,New_Value2]]))
```

```
Predicted Value:
 [783.43231668]
```

Conclusion: Number of lego pieces have a statistically significant relationship with the lego price. Age variable doesn't contribute much in explaining the price differences. Other variables can also be added to the model and tested for significance.

2. Analyse customer sentiment reviews:

Out[4]:

	overall	verified	reviewTime	reviewerID	reviewerName	reviewText	summary	unixReviewTime	image
0	2	False	09 22, 2016	A1IDMI31WEANAF	Mackenzie Kent	When it comes to a DM's screen, the space on t...	The fact that 50% of this space is wasted on a...	1474502400	NaN
1	1	False	09 18, 2016	A4BCEVVZ4Y3V3	Jonathan Christian	An Open Letter to GaleForce9*:\n\nYour unpaint...	Another worthless Dungeon Master's screen from...	1474156800	NaN
2	3	True	09 12, 2016	A2EZ9PY1IHHBX0	unpreparedtodie	Nice art, nice printing. Why two panels are f...	pretty, but also pretty useless	1473638400	NaN
3	5	True	03 02, 2017	A139PXTTC2LGHZ	Ashley	Amazing buy! Bought it as a gift for our new d...	Five Stars	1488412800	NaN
4	1	True	02 08, 2017	A3IB33V29XIL8O	Oghma_EM	As my review of GF9's previous screens these w...	Money trap	1486512000	NaN
5	5	True	01 27, 2017	A1J86V48S4KRJE	Cynthia A. Evoniuk	Grandson loves	Five Stars	1485475200	NaN
6	5	False	01 02, 2017	A14J12PRBLGHF4	Amazon Customer	I have bought many gm screens over the years, ...	Best gm screen ever	1483315200	NaN
7	5	True	12 17, 2016	A2UKOWP9ICU416	anon9df0	Came in perfect condition.	Five Stars	1481932800	NaN
8	4	False	12 15, 2016	A2ONKKDETRWT79	Consumer Dad	Could be better but its still great. I love th...	Great but could be even better	1481760000	NaN
9	3	True	12 9, 2016	AK9GN9KZZNTEP	Dallas Gamer Family	My review will mirror others in that this kind...	Another missed opportunity. Not a value add t...	1481241600	['https://images-na.ssl-images-amazon.com/imag...

'ReviewText' column of the above dataframe has been the focus of this part. I've pre-processed data by:

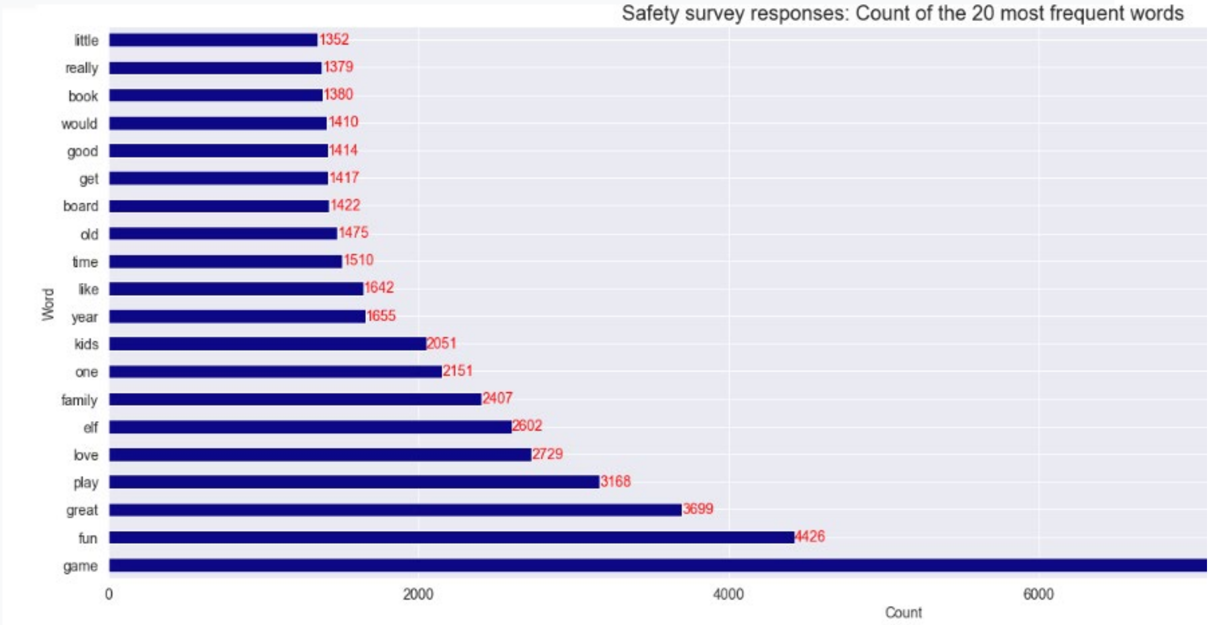
- removing blanks, duplicates, punctuations, unverified comments (assuming that only verified comments are by our customers)
- transforming data to lowercase

Out[50]:

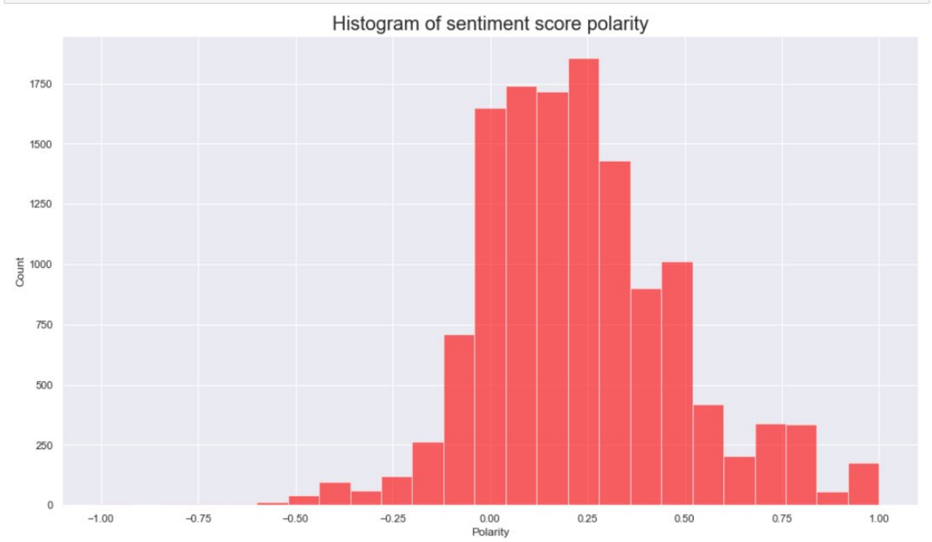
	index	overall	verified	reviewTime	reviewerID	reviewerName	reviewText	summary	unixReviewTime	image	tokens	polarit
0	2	3	True	09 12, 2016	A2EZ9PY1IHHBX0	unpreparedtodie	nice art nice printing why two panels are fill...	pretty, but also pretty useless	1473638400	NaN	[nice, art, nice, printing, why, two, panels, ...	0.11664
1	3	5	True	03 02, 2017	A139PXTTC2LGHZ	Ashley	amazing buy bought it as a gift for our new dm...	Five Stars	1488412800	NaN	[amazing, buy, bought, it, as, a, gift, for, o...	0.57878
2	4	1	True	02 08, 2017	A3IB33V29XIL8O	Oghma_EM	as my review of gf9s previous screens these we...	Money trap	1486512000	NaN	[as, my, review, of, gf9s, previous, screens, ...	-0.31666
3	5	5	True	01 27, 2017	A1J86V48S4KRJE	Cynthia A. Evoniuk	grandson loves	Five Stars	1485475200	NaN	[grandson, loves]	0.00000
4	7	5	True	12 17, 2016	A2UKOWP9ICU416	anon9df0	came in perfect condition	Five Stars	1481932800	NaN	[came, in, perfect, condition]	1.00000
							my review will mirror	Another missed		['https://images-	[my, review,	

Additionally, I've removed the stop words (and, or…) so that the WordCloud includes the context related words of high frequency and relevance.

Number of positive words are significantly high when we check the 20 most frequent words: 'fun', 'great', 'love', 'like', 'good'… Books are highly mentioned as well as board & games (next step: check whether the biagram word 'board game' is among the most frequents). However, we know nothing about the sentiments towards these two words at this point.



When we plot Sentiment Polarity Scores (-1's the lowest, +1's the highest), we see that most comments express a positive sentiment.



If we extract the top 20 positive reviews:

Out[37]:

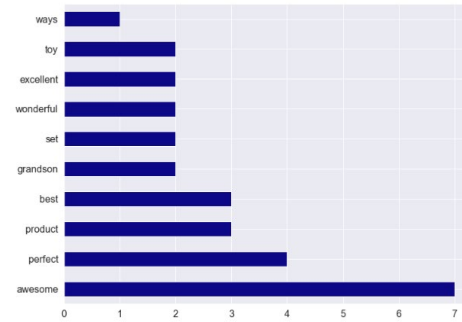
	reviewerID	overall	reviewText	summary	polarity	subjectivity
4	A2UKOWP9ICU416	5	came in perfect condition	Five Stars	1.000000	1.000000
140	A9V7MUGGFFT7R	5	awesome book	Five Stars	1.000000	1.000000
167	A2D0AVXUJVHK1T	5	awesome gift	Five Stars	1.000000	1.000000
444	A273OOTSQQP8ID	5	excellent activity for teaching selfmanagement skills	Five Stars	1.000000	1.000000
471	A3GYWP2LZYRDLI	5	perfect just what i ordered	Five Stars	1.000000	1.000000
533	A1K1J2TG88SOH8	5	wonderful product	Five Stars	1.000000	1.000000
549	A2MW38KK7OMHBX	5	delightful product	Five Stars	1.000000	1.000000
561	A1FWWIJKFY48O	5	wonderful for my grandson to learn the resurrection story	Five Stars	1.000000	1.000000
717	A1ZSF3GAJMDLIJ	5	perfect	Aquire game	1.000000	1.000000
831	A32YPU6CNW8U33	5	awesome	Five Stars	1.000000	1.000000
1009	A2SK2OOZZETTBU	5	awesome set	Five Stars	1.000000	1.000000
1039	AY5402TN448TC	5	best set buy 2 if you have the means	Five Stars	1.000000	0.300000
1048	A3G6DT8C4GZ653	5	awesome addition to my rpg gm system	Five Stars	1.000000	1.000000
1238	AXE5HF9SSP62W	5	one of the best board games i played in along time	Five Stars	1.000000	0.300000
1375	A2ULUNAFBFJSBB	5	my daughter loves her stickers awesome seller thank you	Awesome seller! Thank You	1.000000	1.000000
1513	AOJNB8XQ5EGTL	5	awesome toy	Five Stars	1.000000	1.000000
1518	A1WM6M903KL9RQ	3	it is the best thing to play with and also mind blowing in some ways	Three Stars	1.000000	0.300000
1524	A3BEWPNW57XTTY	5	excellent toy to simulate thought	Five Stars	1.000000	1.000000
1734	AL93VA74KNH9W	5	perfect for tutoring my grandson in spelling	tutoring	1.000000	1.000000
1944	A82PXKARYAWL	5	very happy with this product	Five Stars	1.000000	1.000000

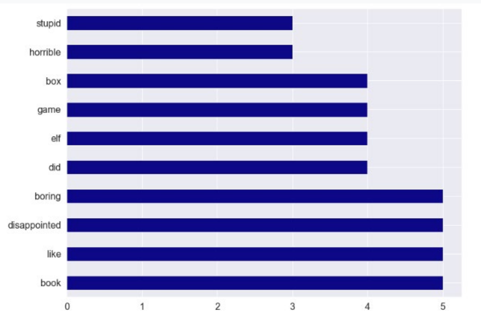
and the bottom 20 negative reviews:

Out[38]:

	reviewerID	overall	reviewText	summary	polarity	subjectivity
180	A3SCMMOUFRA9VK	1	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	1.000000
1802	A28APXX53Y3OBG	1	kids did not like it thought it was boring	Not so much fun	-1.000000	1.000000
2899	A29ZPOASXZI493	1	some of the suggestions are disgusting	One Star	-1.000000	1.000000
7365	A1NA67C1C1ESRB	1	awful we did not receive what was advertised we paid 30 for the boxes set with book we got the elf in a bag without the book	Not What Was Advertised	-1.000000	1.000000
7045	A3S8TI3M8BCBRA	3	was the elf on the shelf but it didnt have the dvd i was very disappointed	Three Stars	-0.975000	0.975000
8483	A35OX0453C1M70	1	i havent even taken it out of the box yet but its already falling apart i contacted customer service and never even got a response i am very disappointed in this product	Poor quality. Falling apart in multiple places.	-0.975000	0.975000
8144	A3A522DVPJNI4D	2	cliche and stupid i should not drink and amazon	Hahaha. Ho Ho Ho.	-0.800000	1.000000
8256	AUBU47RORRSMB	1	just stupid	One Star	-0.800000	1.000000
155	AWUPAM7C4GTWZ	1	incomplete kit very disappointing	INCOMPLETE KIT!	-0.780000	0.910000
12488	A2DRLFCLO4WWBY	4	i like this product for my daughter she is into the bad kitty book collection so it was an added bonus	Good Kitty	-0.700000	0.666667
3748	A2QP0VYB6DEGTB	2	ordered for my sons birthday opened it up today to play and the board is damaged before we even take it out of the box the game is already falling apart very disappointed	Damaged board out of box	-0.687500	0.687500
3779	A6FB3CH3GBD4	1	id like to upload a photo of the condition of the game boxit looks like its been used as a soccer ball 2 corners of the box are smashed in and on is even ripped how am i supposed to give this as a gift without it looking like i bought this on clearance very disappointed	:(	-0.687500	0.687500
10548	AW39RO1HLML9A	1	horrible and incomplete flash cardsdo not buy not helpful i was too late to return them	One Star	-0.650000	0.800000
10104	A27ZZ950XCUQJ	1	boring did i mention boring well its boring pass on this one there are a lot better games out there	Boring	-0.625000	0.875000
12224	AA4CAMGYC7M4Z	1	had no idea the extent you have to go through to put this together hundreds and i mean hundreds of pieces that dont snap together it will take my teen age son and i months to put this stupid thing together horrible plan horrible	It will take my teen age son and I months to put this stupid thing together	-0.622500	0.737500
7290	A34RRLNN518VTN	1	i received a small paperback bookfor 3000 the picture shows an elf hardcover book and box that it all comes in very disappointed for the student we bought this for	VERY DISAPPOINTED for the student we bought this for	-0.612500	0.687500
4405	A3UNLN0MDO3V0I	5	want to hate your friends and family get this game	Five Stars	-0.600000	0.650000
11288	A373S7PHHOP35X	1	piece of crap game caused a fight in my house	One Star	-0.600000	0.600000
3066	A31AJ70ZUX1U1H	1	disappointment	Disappointing	-0.600000	0.400000
7165	A103JMWFQ7X8ZS	2	i was unhappy with it because the elf wasnt with it the book wasnt even wrapped	Two Stars	-0.600000	0.900000

we can then use them to generate a document-term matrix in order to extract positive/negative features from written text:





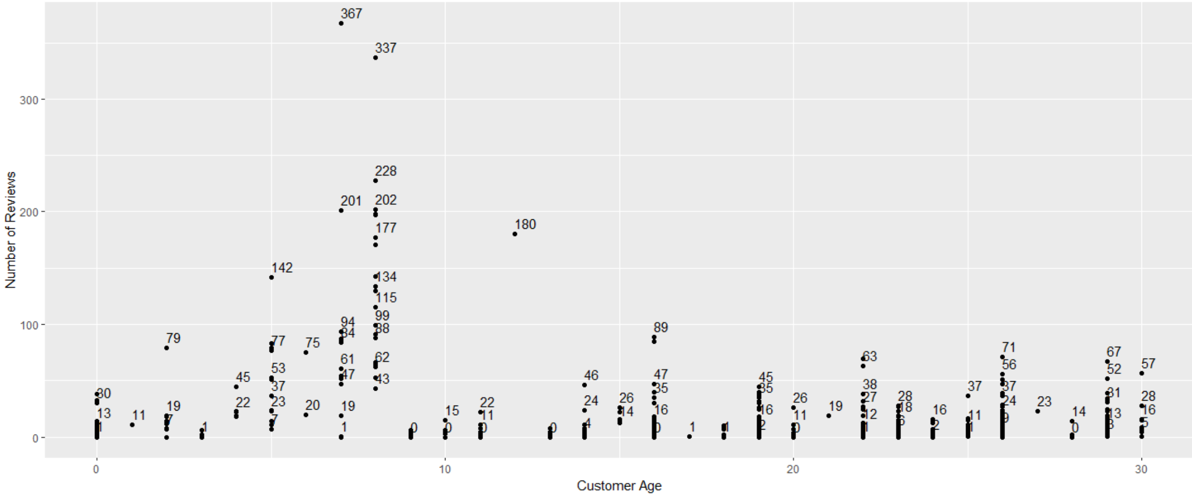
For further investigation:

- The positive popular words might be implying that:
    - toys and sets are popular
    - grandparents prefer to buy the products for grandsons
- The negative popular words might be implying that:
    - books and box games are not as satisfactory and interesting
      or
    - the boxes of the products (ie packaging) are causing disappointment

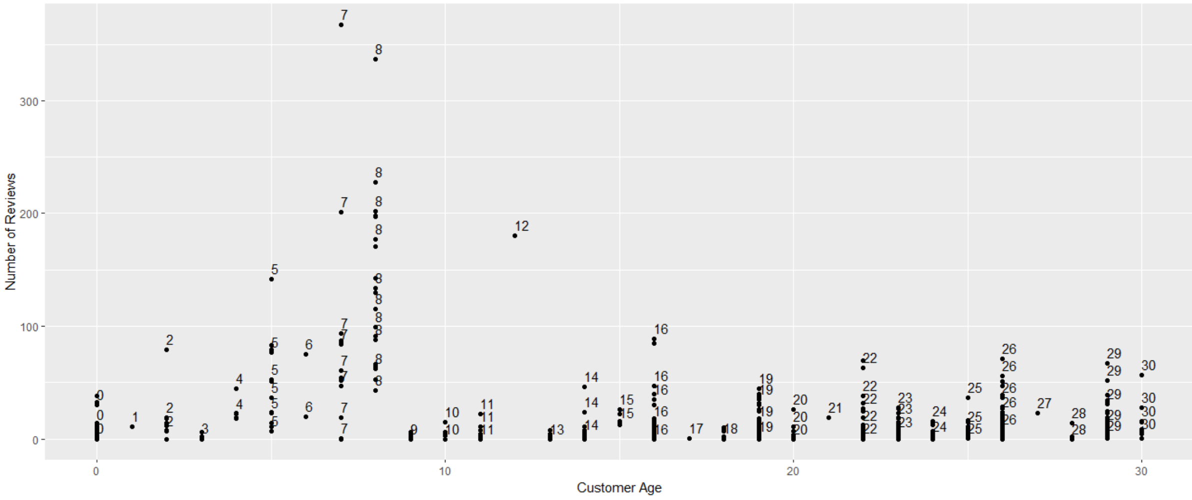
3. Analyse customer behaviour:

- The customer group that will most likely leave a review on the product they've purchased:

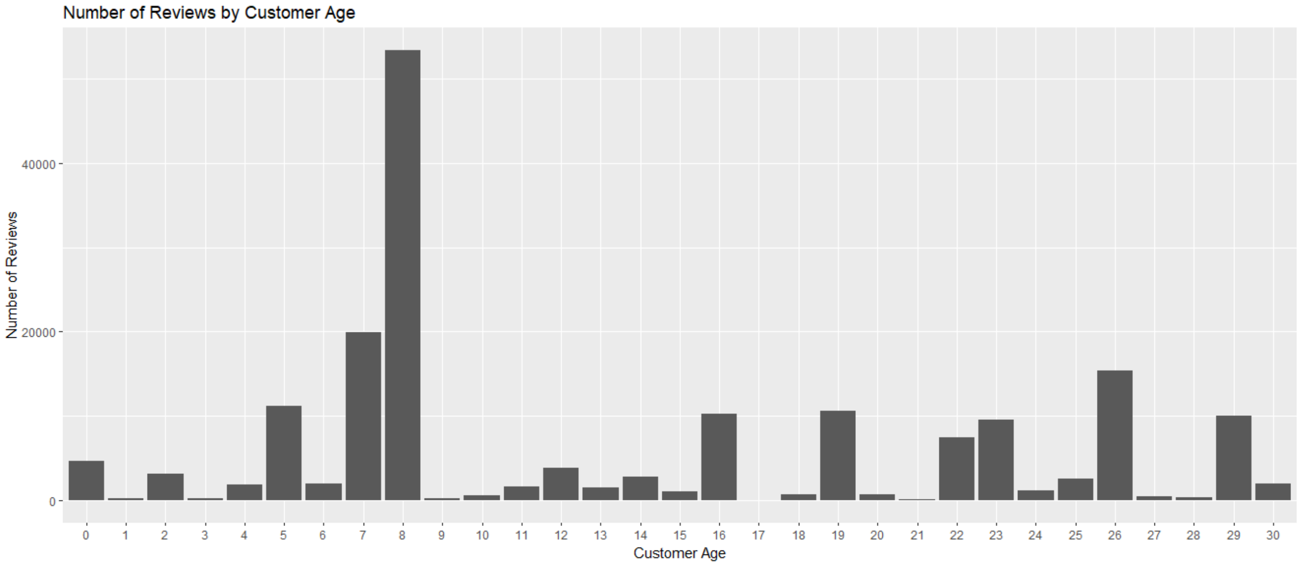
Number of Reviews by Customer Age



Customer Age corresponding to Number of Reviews

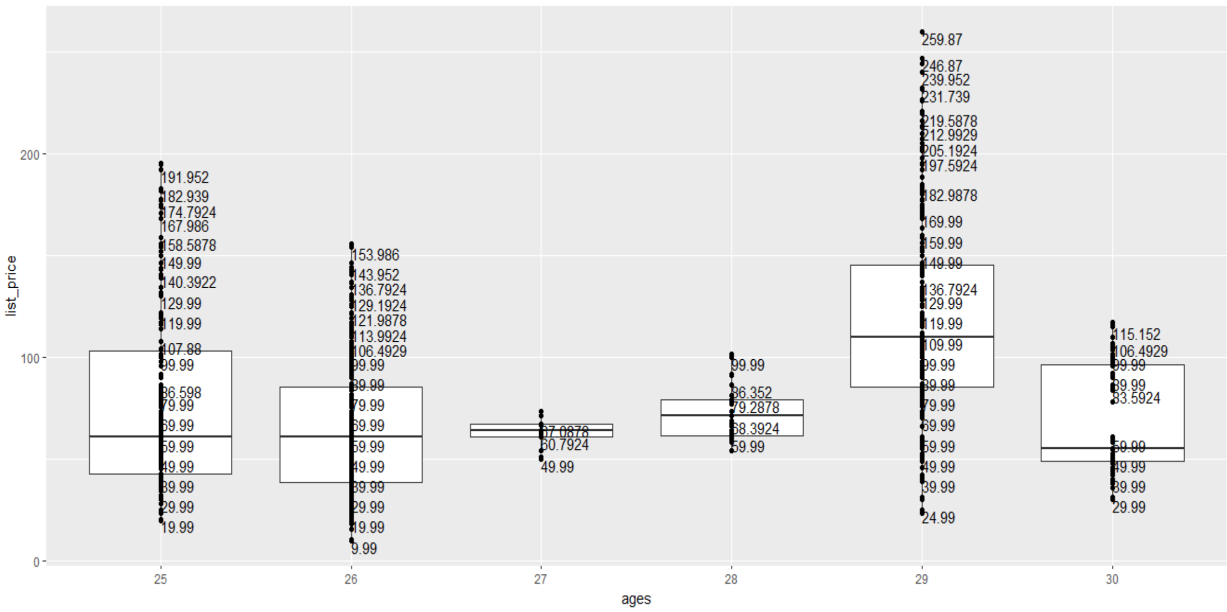


The above two plot will give an idea on "the age group that submits many reviews" (looks like 7–8-year-olds). However, these plots might under-state the total number of reviews, as scatter plot does not give a hint on the overlapping data points.



This chart shows us the customer group of age 8 has left the maximum number of reviews up until now (above 50,000). The second highest is the group of age 7 who has left less than half of the comments of 8-year-olds.

- The most popular, expensive Lego set purchased by customers who are at least 25 years old:



The group of 29-year-olds paid $259.87 for a Lego set.

Now that we have some additional insight on customer behaviour:

- We can analyse the comments made for 8-year-old products and understand the areas of improvement
- We could introduce new fancy products to attract the attention of 29-year-olds.

4. Predict the global sales (in millions) for the next financial year – data preparation

“

```
# A tibble: 16,598 x 9
    Rank Name                        Platform Year  Genre        Publisher NA_Sales EU_Sales Global_Sales
   <int> <chr>                       <chr>    <chr> <chr>        <chr>        <dbl>    <dbl>        <dbl>
 1     1 Wii Sports                  Wii      2006  Sports       Nintendo      41.5    29.0          82.7
 2     2 Super Mario Bros.           NES      1985  Platform     Nintendo      29.1     3.58         40.2
 3     3 Mario Kart Wii              Wii      2008  Racing       Nintendo      15.8    12.9          35.8
 4     4 Wii Sports Resort           Wii      2009  Sports       Nintendo      15.8    11.0          33
 5     5 Pokemon Red/Pokemon Blue    GB       1996  Role-Playing Nintendo      11.3     8.89         31.4
 6     6 Tetris                      GB       1989  Puzzle       Nintendo      23.2     2.26         30.3
 7     7 New Super Mario Bros.       DS       2006  Platform     Nintendo      11.4     9.23         30.0
 8     8 Wii Play                    Wii      2006  Misc         Nintendo      14.0     9.2          29.0
 9     9 New Super Mario Bros. Wii   Wii      2009  Platform     Nintendo      14.6     7.06         28.6
10    10 Duck Hunt                   NES      1984  Shooter      Nintendo      26.9     0.63         28.3
# … with 16,588 more rows
> 
```

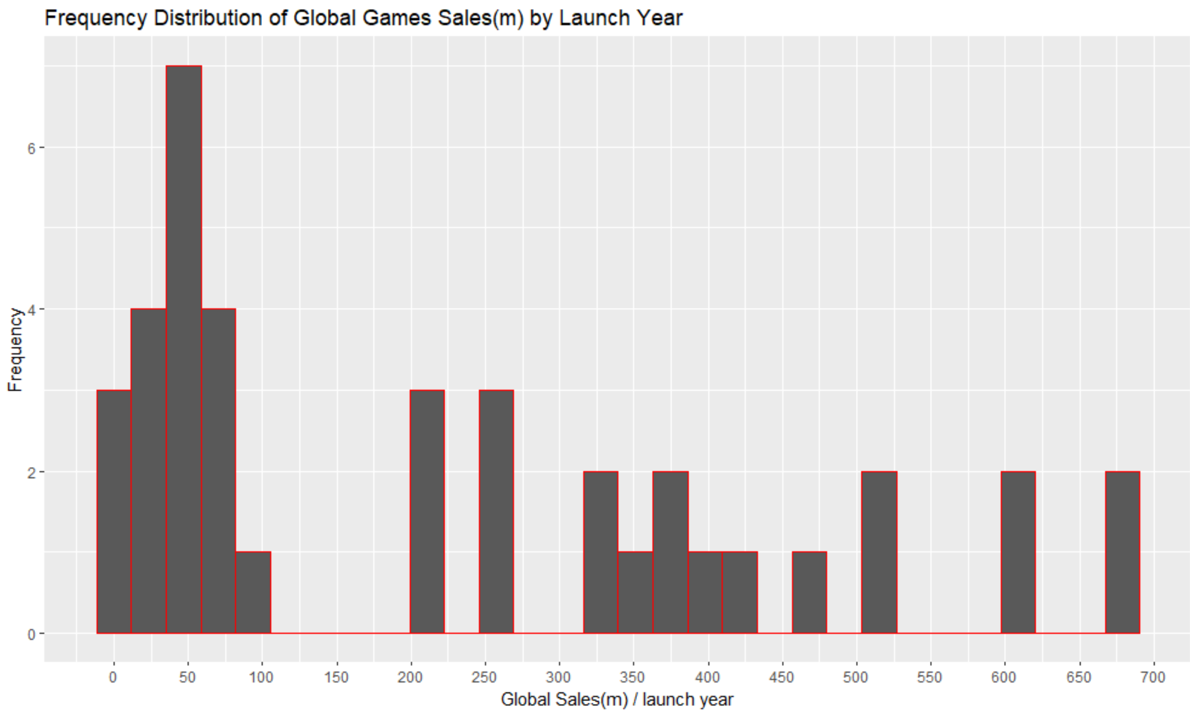
- Understand and wrangle the above summarised data set
    - Convert Year data type to integer
    - Identify and filter 271 missing values
    - Transform the data for consistency (convert Genre to lower-case, merge Platform and Genre)
- Aggregate the data, and form a data table that includes Sales figures aggregated by launch years:

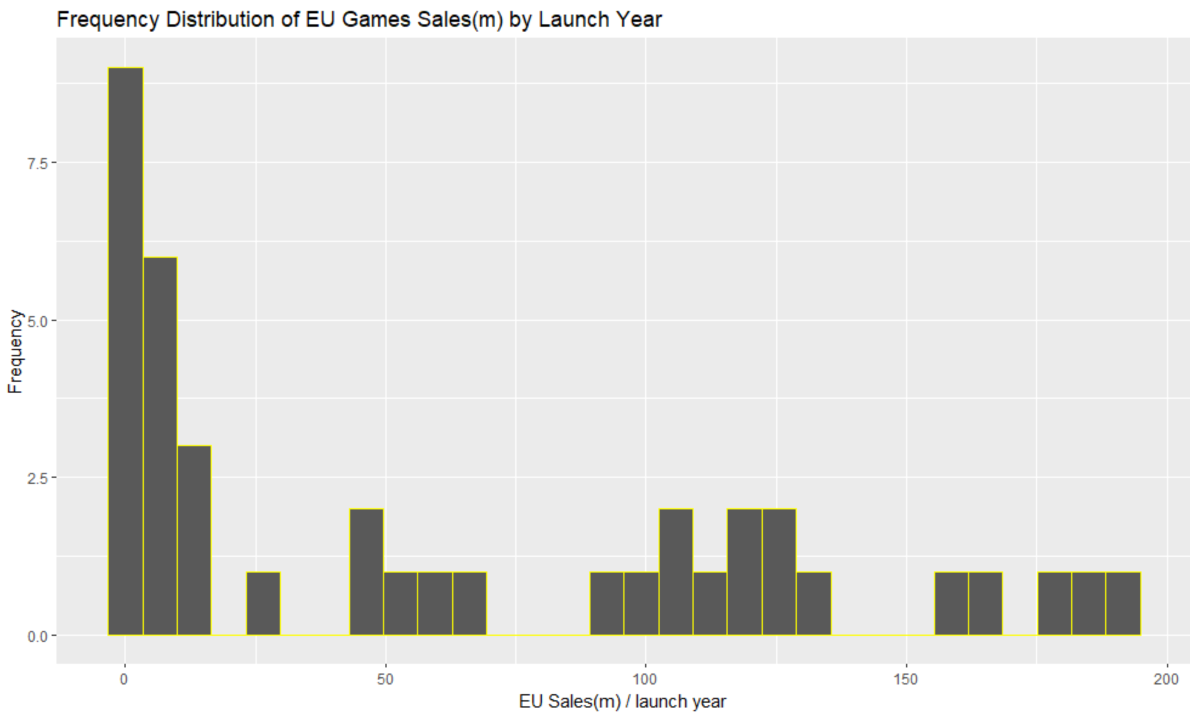
Filter

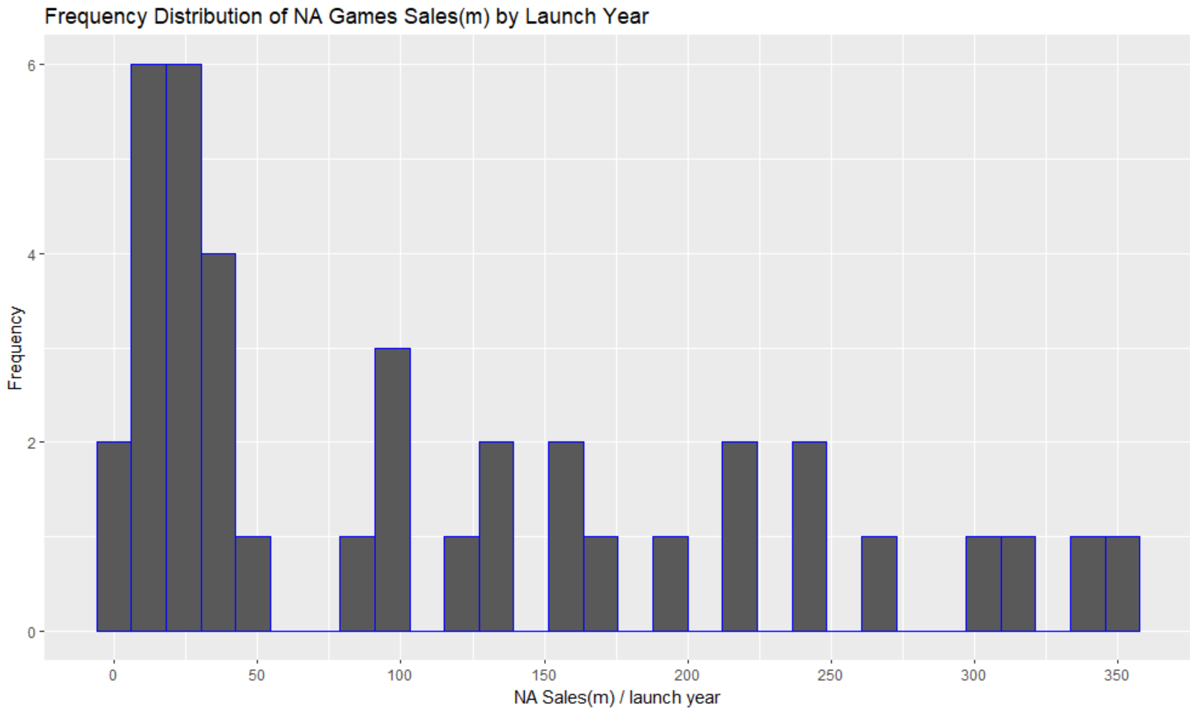
	Year	Global_Sales	EU_Sales	NA_Sales
1	1980	11.38	0.67	10.59
2	1981	35.77	1.96	33.40
3	1982	28.86	1.65	26.92
4	1983	16.79	0.80	7.76
5	1984	50.36	2.10	33.28
6	1985	53.94	4.74	33.73
7	1986	37.07	2.84	12.50
8	1987	21.74	1.41	8.46
9	1988	47.22	6.59	23.87
10	1989	73.45	8.44	45.15
11	1990	49.39	7.63	25.46
12	1991	32.23	3.95	12.76
13	1992	76.16	11.71	33.87
14	1993	45.98	4.65	15.12
15	1994	79.17	14.88	28.15
16	1995	88.11	14.90	24.82
17	1996	199.15	47.26	86.76
18	1997	200.98	48.32	94.75
19	1998	256.47	66.90	128.36
20	1999	251.27	62.67	126.06
21	2000	201.56	52.75	94.49
22	2001	331.47	94.89	173.98
23	2002	395.52	109.74	216.19
24	2003	357.85	103.81	193.59
25	2004	419.31	107.32	222.59
26	2005	459.94	121.94	242.61
27	2006	521.04	129.24	263.12
28	2007	611.13	160.50	312.05
29	2008	678.90	184.40	351.44
30	2009	667.30	191.59	338.85
31	2010	600.45	176.73	304.24
32	2011	515.99	167.44	241.06
33	2012	363.54	118.78	154.96
34	2013	368.11	125.80	154.77
35	2014	337.05	125.65	131.97
36	2015	264.44	97.71	102.82
37	2016	70.93	26.76	22.66
38	2017	0.05	0.00	0.00
39	2020	0.29	0.00	0.27

Showing 1 to 39 of 39 entries, 4 total columns

- Visualise the data to understand the trends between the variables:
  - The skewness



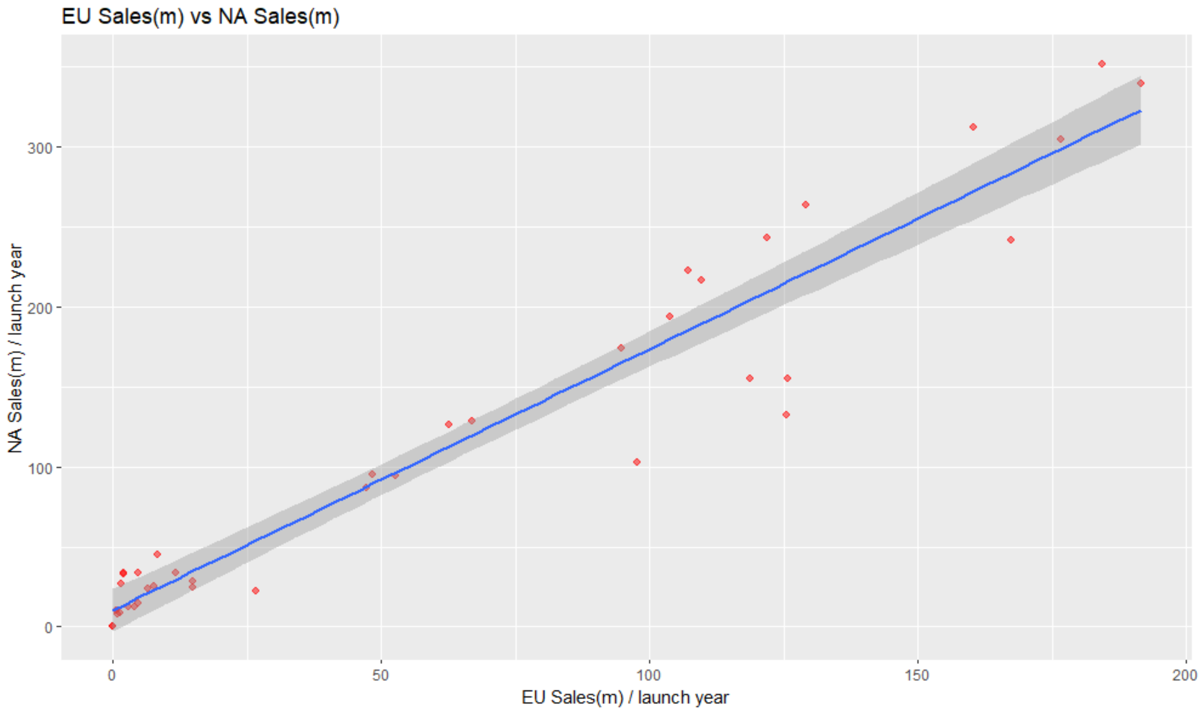





```
> skewness(Global_Sales_by_year$Global_Sales)
[1] 0.6954635
> kurtosis(Global_Sales_by_year$Global_Sales)
[1] 2.211808
> skewness(EU_Sales_by_year$EU_Sales)
[1] 0.6041092
> kurtosis(EU_Sales_by_year$EU_Sales)
[1] 1.953936
> skewness(NA_Sales_by_year$NA_Sales)
[1] 0.8040209
> kurtosis(NA_Sales_by_year$NA_Sales)
[1] 2.377007
```

In the case of our data set, the sales calls data are positively skewed. Moreover, positive kurtosis means that the distribution is more peaked and have fatter tails.

- <u>What's the correlation between EU Sales and NA Sales (the variables that will help us predict Global Sales):</u>



EU Sales and NA Sales look strongly correlated.
We need to be cautious about multicollinearity, keeping in mind the below:

> **"Multicollinearity affects the coefficients and p-values, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit statistics. If your primary goal is to make predictions, and you don't need to understand the role of each independent variable, you don't need to reduce severe multicollinearity."**

https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

5. Predict the global sales (in millions) for the next financial year – regression models

```
> cor(Sales_by_year)
                    X      Year Global_Sales  EU_Sales  NA_Sales
X           1.0000000 0.9996409    0.5890087 0.6565908 0.5339382
Year        0.9996409 1.0000000    0.5794317 0.6469072 0.5249655
Global_Sales 0.5890087 0.5794317    1.0000000 0.9858848 0.9923798
EU_Sales    0.6565908 0.6469072    0.9858848 1.0000000 0.9637526
NA_Sales    0.5339382 0.5249655    0.9923798 0.9637526 1.0000000
```

This overview shows us how all the variables are correlated.
according to Evans' classification, above 0.80 is a very strong correlation.
It is good to go with EU Sales and NA Sales to predict Global Sales.

When we run our model the summary statistics comes as follows:

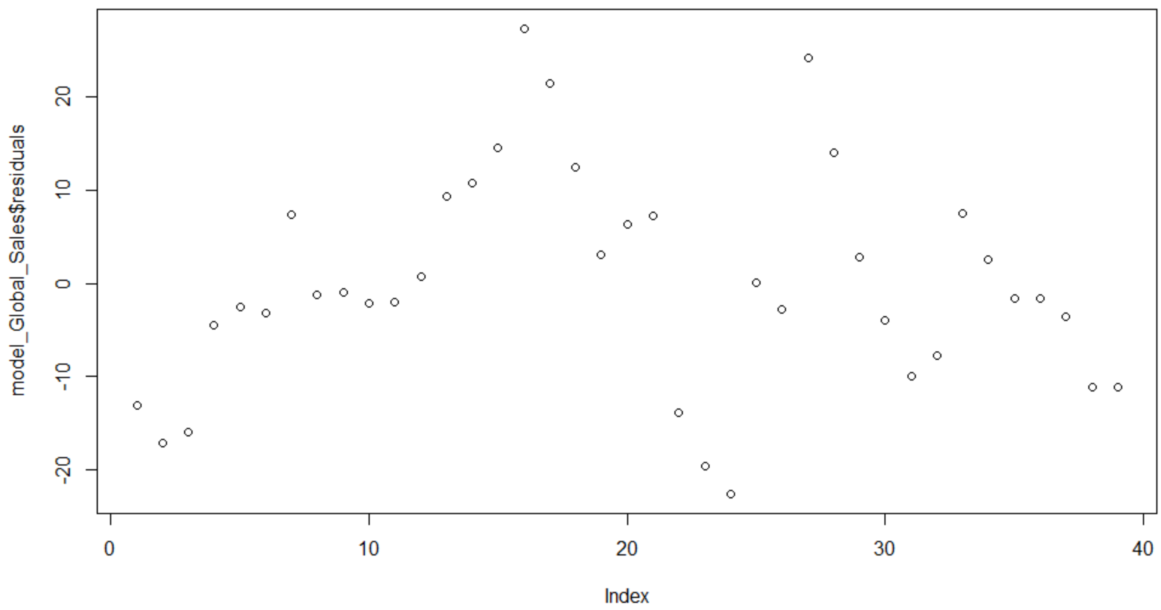
```
Call:
lm(formula = Global_Sales ~ EU_Sales + NA_Sales, data = Sales_by_year)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.653  -6.134  -1.595   7.333  27.415 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  11.1457     2.7604   4.038  0.00027 ***
EU_Sales      1.3795     0.1135  12.151 2.67e-14 ***
NA_Sales      1.1682     0.0671  17.409  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.9 on 36 degrees of freedom
Multiple R-squared:  0.997,	Adjusted R-squared:  0.9969 
F-statistic:  6030 on 2 and 36 DF,  p-value: < 2.2e-16

> 
```

The multiple R-squared for this model is 0.997, while the adjusted R-squared for this model is 0.9969.

Let’s look at the significance of the explanatory variables in the coefficients table. We can see that both EU_Sales and NA_Sales are very significant.

These figures show that we are on track.

We would also want to look at the residuals from this model:



There is no pattern in these residuals. They look like white noise.

One final step before we estimate the Global Sales for the next financial year would be testing the model:

17	1996	199.15	47.26	86.76
18	1997	200.98	48.32	94.75
19	1998	256.47	66.90	128.36
20	1999	251.27	62.67	126.06
21	2000	201.56	52.75	94.49

When we input the EU Sales and NA Sales of the products launched in 2000 to the model, we get 194.2972. The actual figure is 201.56. So, the model works fine.

Now, we can start our forecasting.
Assumption: EU Sales would increase 4% and NA Sales would increase 6% next year (2023)
If the run the model with these figures we get $8833.497(m) as the result (compared to $8820.36(m) sales of this year.)

As a next step it would be best to revisit the assumption about EU_Sales and NA_Sales 2023 figures and create regression models for these two, to predict them with more confidence.