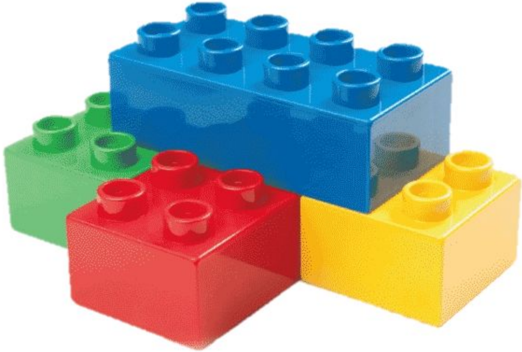


Can we predict the price of a lego set ?

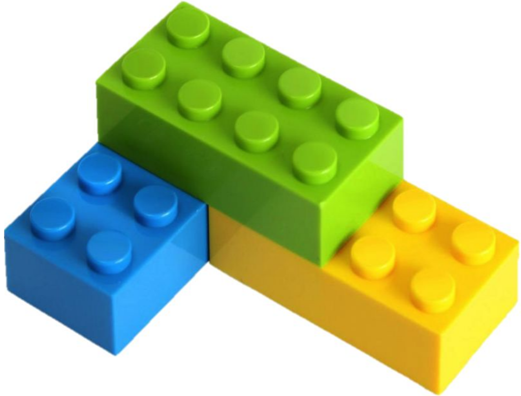


Table of content

Data collection

Collected from: github

Set contains 14 columns and 6000+ rows

	Item_Number	Name	Year	Theme	Subtheme	Pieces	Minifigures	Image_URL	GBP_MSRP	USD_MSRP	CAD_MSRP	EUR_MSRP	Packaging	Availability
0	10246	Detective's Office	2015	Advanced Models	Modular Buildings	2262.0	6.0	http://images.brickset.com/sets/images/10246-1...	132.99	159.99	199.99	149.99	Box	Retail - limited
1	10247	Ferris Wheel	2015	Advanced Models	Fairground	2464.0	10.0	http://images.brickset.com/sets/images/10247-1...	149.99	199.99	229.99	179.99	Box	Retail - limited
2	10248	Ferrari F40	2015	Advanced Models	Vehicles	1158.0	NaN	http://images.brickset.com/sets/images/10248-1...	69.99	99.99	119.99	89.99	Box	LEGO exclusive
3	10249	Toy Shop	2015	Advanced Models	Winter Village	898.0	NaN	http://images.brickset.com/sets/images/10249-1...	59.99	79.99	NaN	69.99	Box	LEGO exclusive
4	10581	Ducks	2015	Duplo	Forest Animals	13.0	1.0	http://images.brickset.com/sets/images/10581-1...	9.99	9.99	12.99	9.99	Box	Retail

Data cleaning

Steps of data cleaning:

- Dropping 8 columns
- Dropping empty rows (~2000)
- Cleaning “themes” (from 22 to 11)
- Replacing Year by Age
- Creating dummies for non-numerical columns
- Checking duplicates
- Converting numerical variables to normal distribution
- Identifying outliers

Data cleaning - Checking duplicates

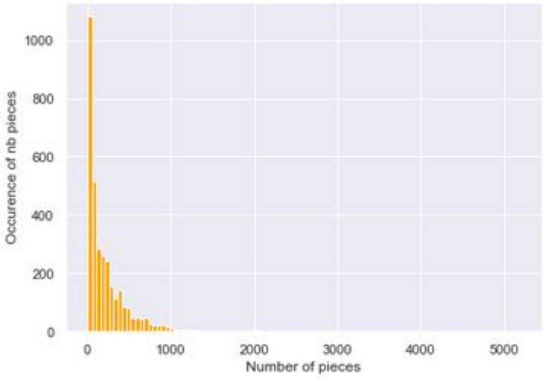
Item Number duplicates (5 first)

	Item_Number	Name	Year	Theme	Subtheme	Pieces	Minifigures	Image_URL	GBP_MSRP	USD_MSRP	CAD_MSRP
296	71008	Classic King	2015	Collectable Minifigures	Series 13	9.0	1.0	http://images.brickset.com/sets/images/71008-1...	2.49	3.99	NaN
297	71008	Sheriff	2015	Collectable Minifigures	Series 13	8.0	1.0	http://images.brickset.com/sets/images/71008-2...	2.49	3.99	NaN
298	71008	Unicorn Girl	2015	Collectable Minifigures	Series 13	6.0	1.0	http://images.brickset.com/sets/images/71008-3...	2.49	3.99	NaN
299	71008	Snake Charmer	2015	Collectable Minifigures	Series 13	7.0	1.0	http://images.brickset.com/sets/images/71008-4...	2.49	3.99	NaN
300	71008	Goblin	2015	Collectable Minifigures	Series 13	7.0	1.0	http://images.brickset.com/sets/images/71008-5...	2.49	3.99	NaN

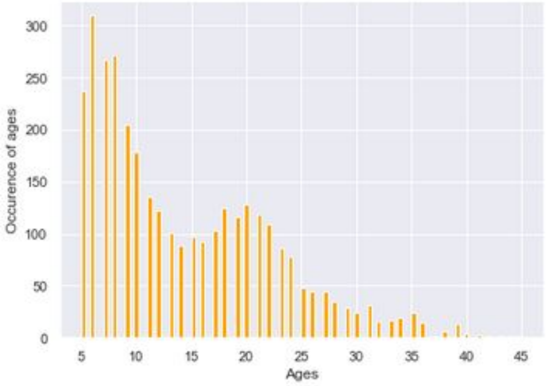
Filter duplicates for the Item Number column ; same Item Number doesn't mean same set, so there are no real duplicates (except numeric one in the end)

Data cleaning - Convert to normal distrib.

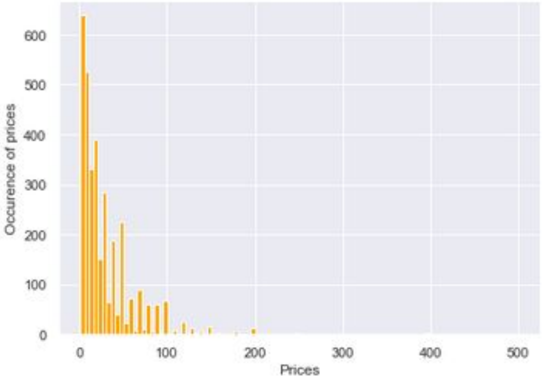
Pieces distribution



Age distribution



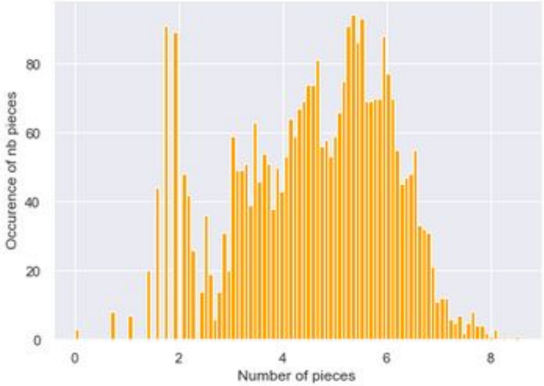
Price distribution



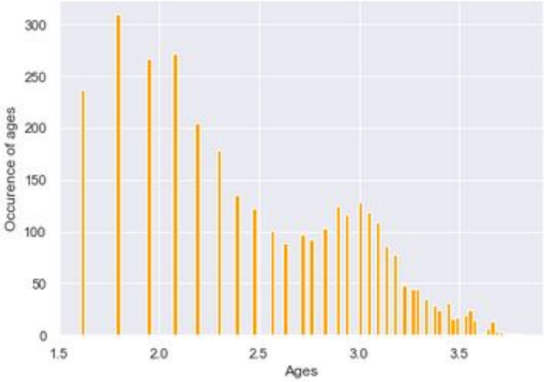
Creating 3 new columns using the boxcox method to convert numerical columns to be normally distributed (using lognormal)

Data cleaning - Convert to normal distrib.

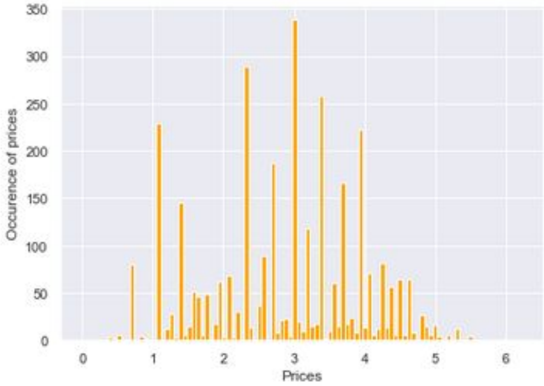
Pieces distribution



Age distribution



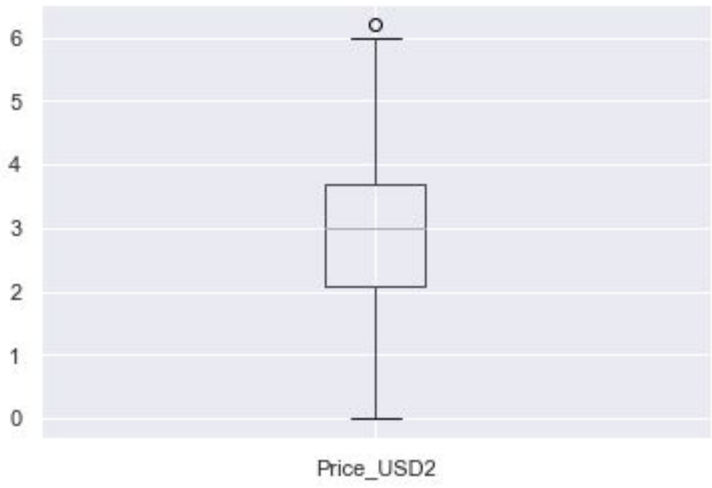
Price distribution



Creating 3 new columns using the boxcox method to convert numerical columns to be normally distributed (using lognormal)

Data cleaning - Identifying outliers

Price outliers



Number of pieces outliers

Creating 2 new boolean columns for price and number of pieces outliers (1/0)

Regression analysis

OLS Regression Results

Dep. Variable:	Price_USD2	**R-squared:**	0.832
Model:	OLS	**Adj. R-squared:**	0.831
Method:	Least Squares	**F-statistic:**	821.1
Date:	Fri, 01 May 2020	**Prob (F-statistic):**	0.00
Time:	12:32:29	**Log-Likelihood:**	-2080.8
No. Observations:	3346	**AIC:**	4204.
Df Residuals:	3325	**BIC:**	4332.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0549	0.052	-1.057	0.290	-0.157	0.047
Outliers_Piece	1.9922	0.263	7.561	0.000	1.476	2.509
Theme_City	0.2164	0.033	6.584	0.000	0.152	0.281
Theme_Duplo	1.2444	0.033	37.149	0.000	1.179	1.310
Theme_Friends	0.1227	0.052	2.381	0.017	0.022	0.224
Theme_Ninjago	0.1325	0.046	2.872	0.004	0.042	0.223
Theme_Other	0.2508	0.021	12.002	0.000	0.210	0.292
Theme_Star Wars	0.1470	0.032	4.531	0.000	0.083	0.211
Availability_Promotional	-0.3523	0.070	-5.001	0.000	-0.490	-0.214
Availability_Retail	-0.1994	0.041	-4.874	0.000	-0.280	-0.119
Availability_Retail - limited	-0.2087	0.049	-4.284	0.000	-0.304	-0.113
Availability_Unknown	-1.4513	0.454	-3.196	0.001	-2.342	-0.561
Packaging_Box	-0.6102	0.038	-16.083	0.000	-0.685	-0.536
Packaging_Box with backing card	-0.3705	0.121	-3.051	0.002	-0.609	-0.132
Packaging_Bucket	-1.2445	0.136	-9.122	0.000	-1.512	-0.977
Packaging_Not specified	-0.4219	0.053	-7.977	0.000	-0.526	-0.318
Packaging_Other	-0.4740	0.119	-3.980	0.000	-0.707	-0.240
Packaging_Plastic box	2.2445	0.180	12.478	0.000	1.892	2.597
Packaging_Polybag	-0.9616	0.062	-15.562	0.000	-1.083	-0.840
Packaging_Tub	-0.9542	0.128	-7.462	0.000	-1.205	-0.704
Pieces2	0.7318	0.007	102.953	0.000	0.718	0.746

Dropping:

- 4 Themes
- 2 Availability
- 1 Packaging
- Price Outlier
- Age

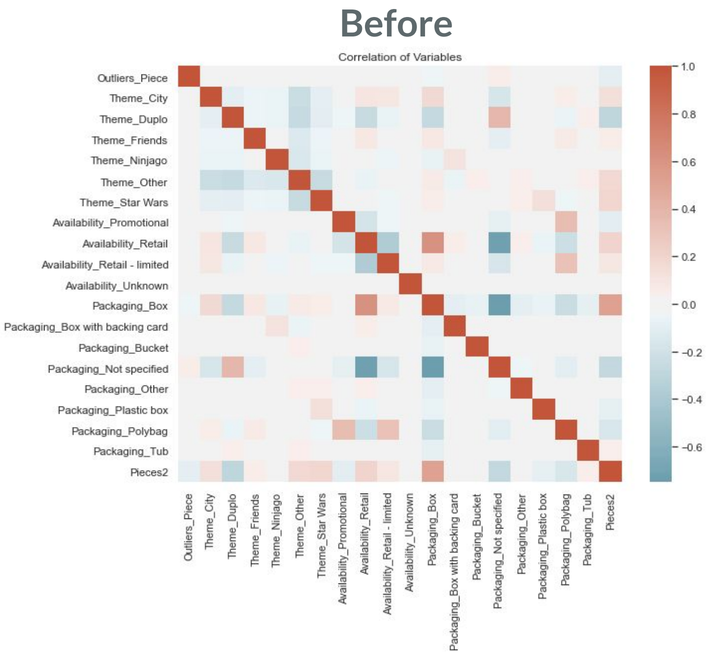
R^2= 0.832

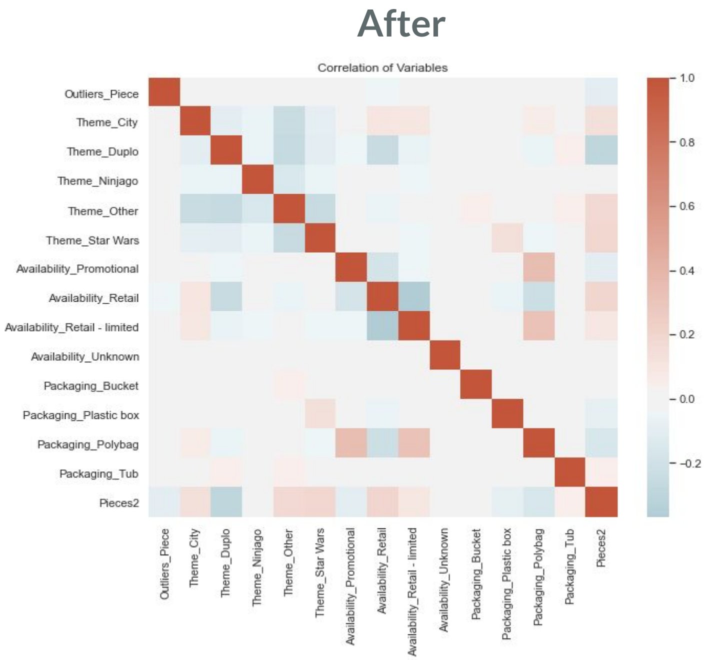
Const pvalue > 0.05

Checking assumptions

1. Multicollinearity
2. Linearity
3. Autocorrelation
4. Homoscedasticity
5. Exogeneity of residuals

Multicollinearity





Dropping:

- 4 Packaging
- 1 Theme

Conclusion: after cleaning the columns, there is no multicollinearity

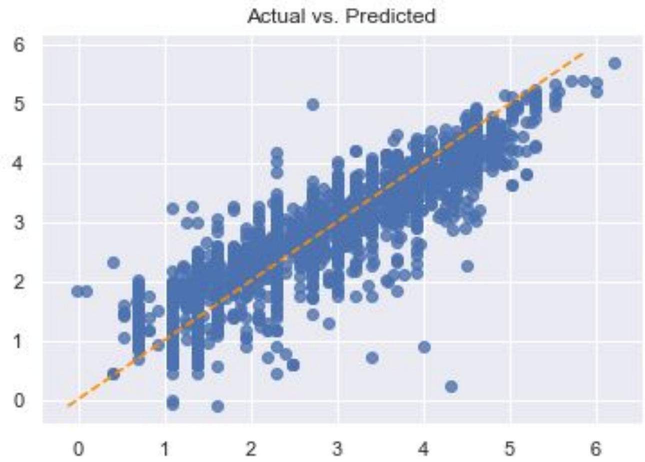
Linearity

```
Parameters that are most likely VIOLATE linearity assumption and their correlation with Price_USD2

Series([], Name: Price_USD2, dtype: float64)


Parameters that are most likely FOLLOW linearity assumption and their correlation with Price_USD2

Pieces2    0.827721
Name: Price_USD2, dtype: float64
```



Conclusion: There is linearity

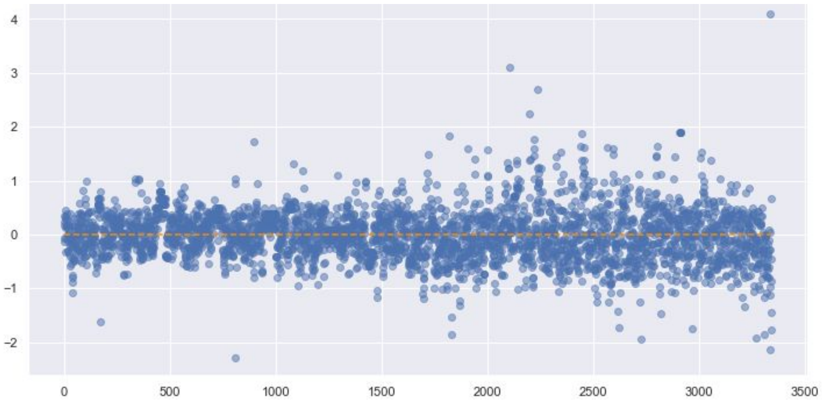
Autocorrelation & Homoscedasticity

Autocorrelation

```
Performing Durbin-Watson Test
-------------------------------------
Durbin-Watson: 1.1109496025957153
Signs of positive autocorrelation

Assumption not satisfied
```

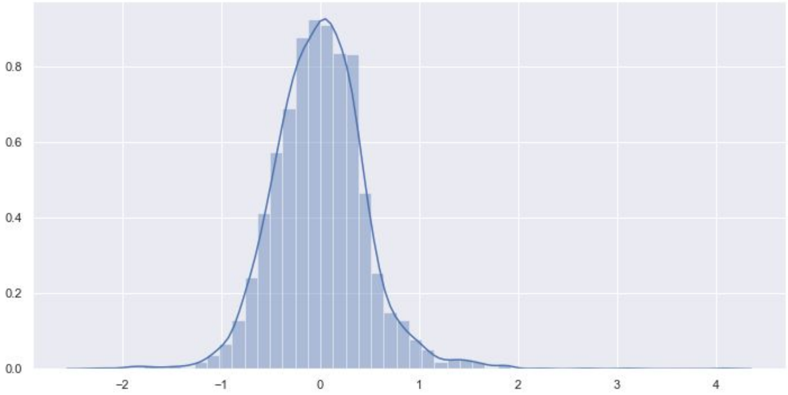
Homoscedasticity



Conclusion: there is a positive autocorrelation and potentially homoscedasticity

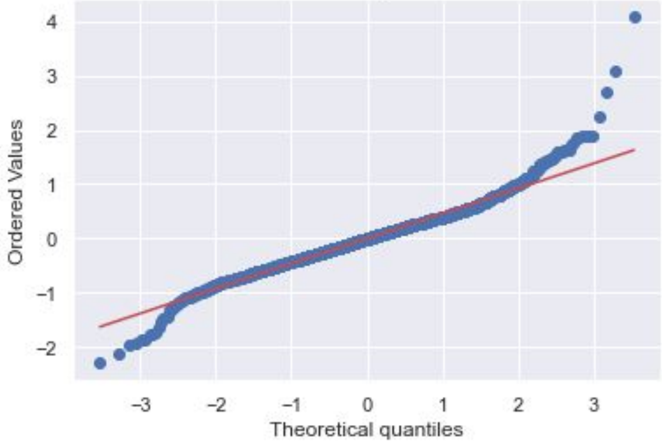
Exogeneity of residuals

Plotting Residuals



Using the Anderson-Darling test for normal distribution
p-value from the test - below 0.05 generally means non-normal:
0.0

Probability Plot



Conclusion: there is exogeneity of residuals as they don't really follow a normal law

Checking assumptions - Results

1. Multicollinearity ✓
2. Linearity ✓
3. Autocorrelation ✗
4. Homoscedasticity ?
5. Exogeneity of residuals ✗

Final Linear Regression

OLS Regression Results

Dep. Variable:	Price_USD2	**R-squared:**	0.818
Model:	OLS	**Adj. R-squared:**	0.818
Method:	Least Squares	**F-statistic:**	1000.
Date:	Fri, 01 May 2020	**Prob (F-statistic):**	0.00
Time:	12:34:29	**Log-Likelihood:**	-2207.9
No. Observations:	3346	**AIC:**	4448.
Df Residuals:	3330	**BIC:**	4546.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2128	0.032	-6.634	0.000	-0.276	-0.150
Outliers_Piece	1.7995	0.273	6.589	0.000	1.264	2.335
Theme_City	0.1490	0.033	4.464	0.000	0.084	0.214
Theme_Duplo	1.1268	0.033	33.882	0.000	1.062	1.192
Theme_Ninjago	0.2087	0.047	4.427	0.000	0.116	0.301
Theme_Other	0.1795	0.021	8.752	0.000	0.139	0.220
Theme_Star Wars	0.1045	0.033	3.168	0.002	0.040	0.169
Availability_Promotional	-0.5304	0.064	-8.285	0.000	-0.656	-0.405
Availability_Retail	-0.2851	0.020	-14.042	0.000	-0.325	-0.245
Availability_Retail - limited	-0.3115	0.036	-8.562	0.000	-0.383	-0.240
Availability_Unknown	-1.6455	0.470	-3.504	0.000	-2.566	-0.725
Packaging_Bucket	-0.6451	0.136	-4.738	0.000	-0.912	-0.378
Packaging_Plastic box	2.5505	0.181	14.119	0.000	2.196	2.905
Packaging_Polybag	-0.4642	0.054	-8.552	0.000	-0.571	-0.358
Packaging_Tub	-0.3307	0.126	-2.617	0.009	-0.579	-0.083
Pieces2	0.6784	0.006	105.979	0.000	0.666	0.691

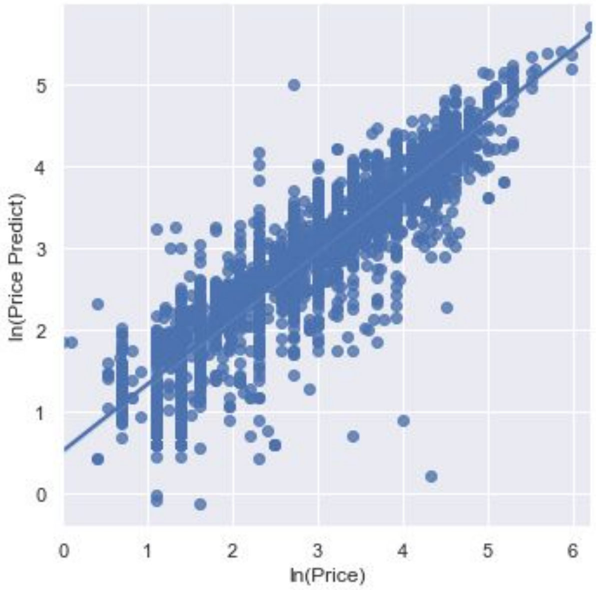
R^2= 0.818

Const pvalue = 0.00

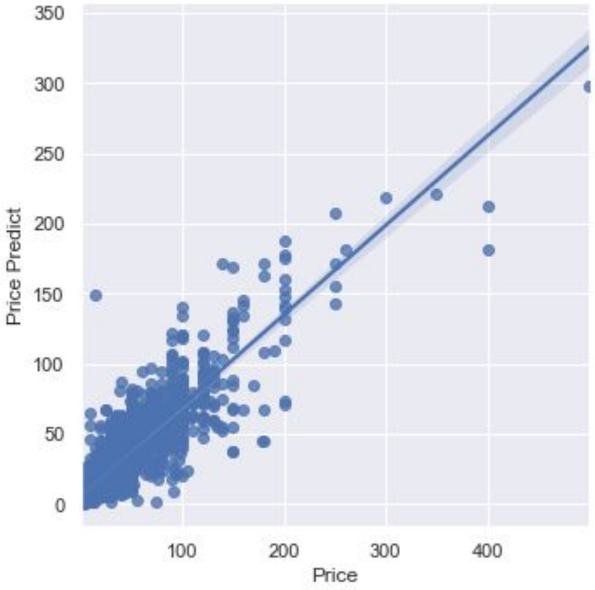
Model equation

ln(y)= -0.21+1.8*Outliers_pieces+ β1*Themes+β2*Packagings+β3*Availability+0.68*Number_pieces

With boxcox



After



Conclusion

Obstacles:

- Time constraint
- Understanding of mathematical concepts
- Running the model a lots of times

Improvements:

- Finding a dataset with more variables
- Have more rows