Lego Set Price Prediction

Multiple Linear Regressions & Neural Network

Team 2
Ji-seong Yang
Jong-hyeok Shin
DSC3006

CONTENTS

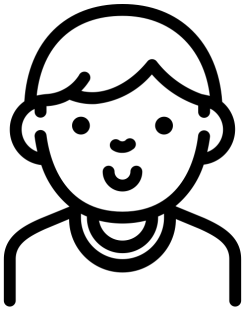
Introduction

Why Lego?

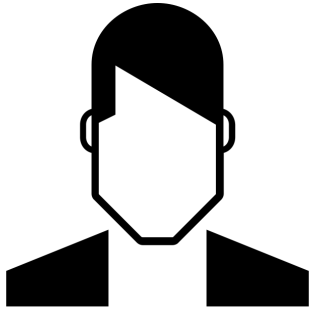
Rising of the Kidults

Kidults?

Kid



Adult

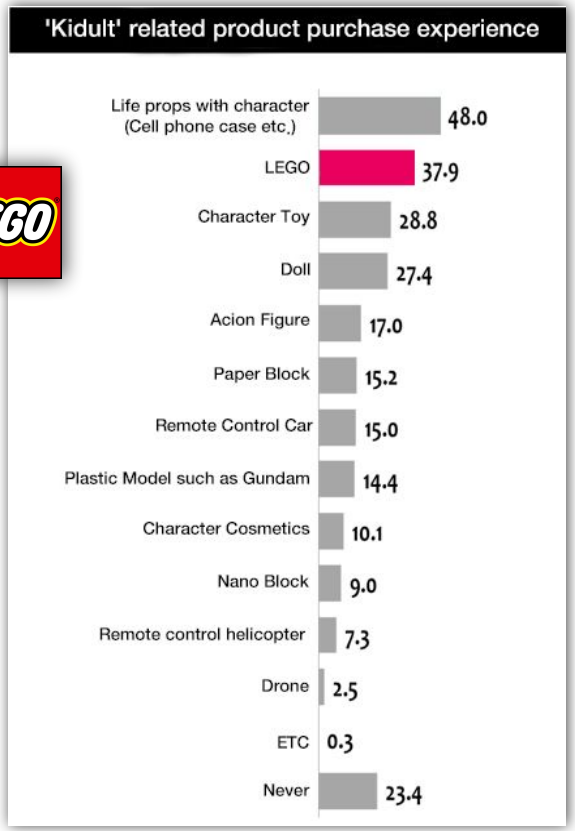


https://www.bbc.com/news/business-45247637

https://toybuzz.org/kidults-market-rockets-to-8-percent-growth-and-continues-expansion/

Adults Playing with Lego Bricks

“Star Wars and LEGO are real favourites for the more mature toy buyer”
- The NPD Group



[TK_201507_NWY1707] 2015 키덜트 족 관련 인식 조사 . TREND MONITOR

Why Are They So Big of a Deal?

NEWS

Lego Is Making Building Sets for Grown Ups

Because adult coloring books weren't enough.

By Amanda Tarlton Nov 20 2018, 1:45 PM

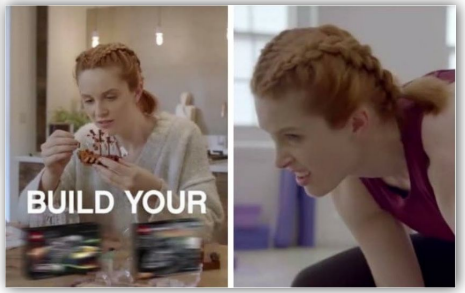
https://www.fatherly.com/news/lego-is-making-building-sets-for-grown-ups/

NEWS ›

Lego invites stressed 'kidult' millennials to discover 'zen' in its bricks

By John Glenday - 29 October 2018 08:40am

https://www.thedrum.com/news/2018/10/29/lego-invites-stressed-kidult-millennials-discover-zen-its-bricks

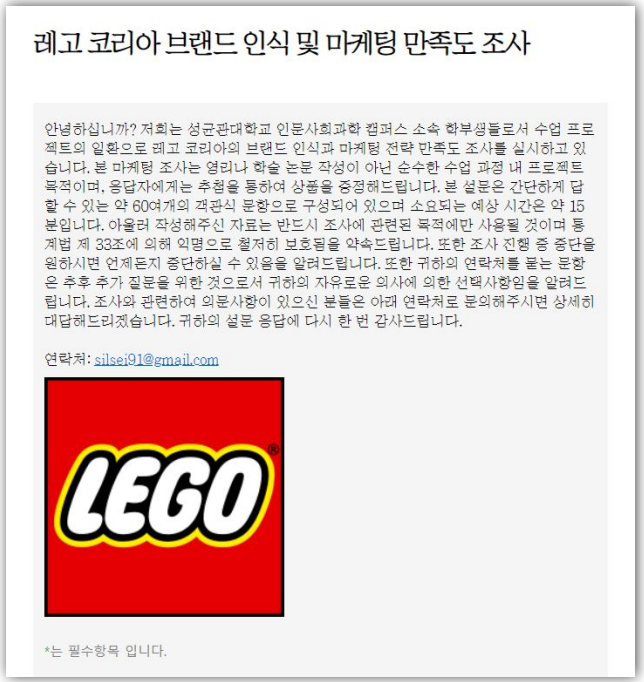


How do Kidults think?

- A marketing survey on Kidults
  - Satisfaction on Lego's marketing efforts
  - Near 100 respondents

94명 응답 응답 별 결과보기 >

주관식 항목의 응답은 최초 50개까지만 표시됩니다.
나머지 응답결과를 확인하려면 '자세히보기'를 이용해 주세요.



레고 코리아 브랜드 인식 및 마케팅 만족도 조사

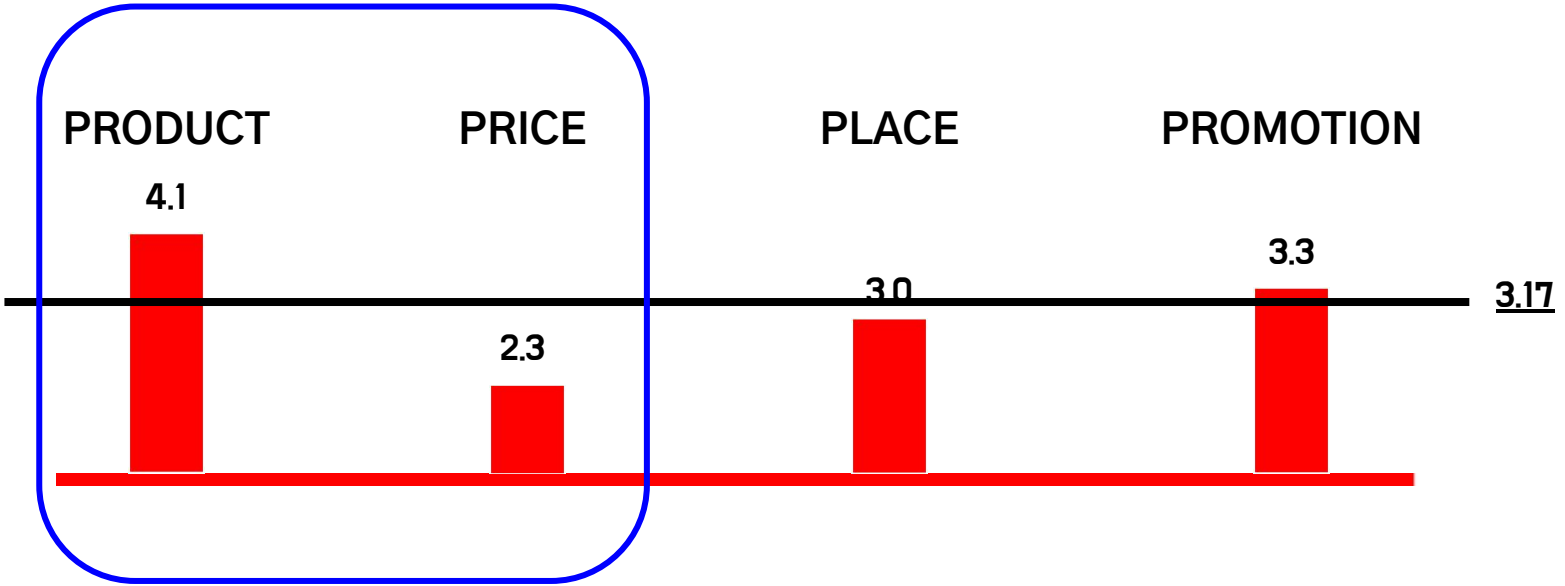
안녕하십니까? 저희는 성균관대학교 인문사회과학 캠퍼스 소속 학부생들로서 수업 프로젝트의 일환으로 레고 코리아의 브랜드 인식과 마케팅 전략 만족도 조사를 실시하고 있습니다. 본 마케팅 조사는 영리나 학술 논문 작성이 아닌 순수한 수업 과정 내 프로젝트 목적이며, 응답자에게는 추첨을 통하여 상품을 증정해드립니다. 본 설문은 간단하게 답할 수 있는 약 60여개의 객관식 문항으로 구성되어 있으며 소요되는 예상 시간은 약 15분입니다. 아울러 작성해주신 자료는 반드시 조사에 관련된 목적에만 사용될 것이며 통계법 제 33조에 의해 익명으로 철저히 보호됨을 약속드립니다. 또한 조사 진행 중 중단을 원하시면 언제든지 중단하실 수 있음을 알려드립니다. 또한 귀하의 연락처를 묻는 문항은 추후 추가 질문을 위한 것으로서 귀하의 자유로운 의사에 의한 선택사항임을 알려드립니다. 조사와 관련하여 의문사항이 있으신 분들은 아래 연락처로 문의해주시면 상세히 대답해드리겠습니다. 귀하의 설문 응답에 다시 한 번 감사드립니다.

연락처: silsei91@gmail.com

*는 필수항목 입니다.

Marketing Mix Satisfaction

- Satisfaction indexes of “Overall Marketing Strategy of LEGO” (out of five)

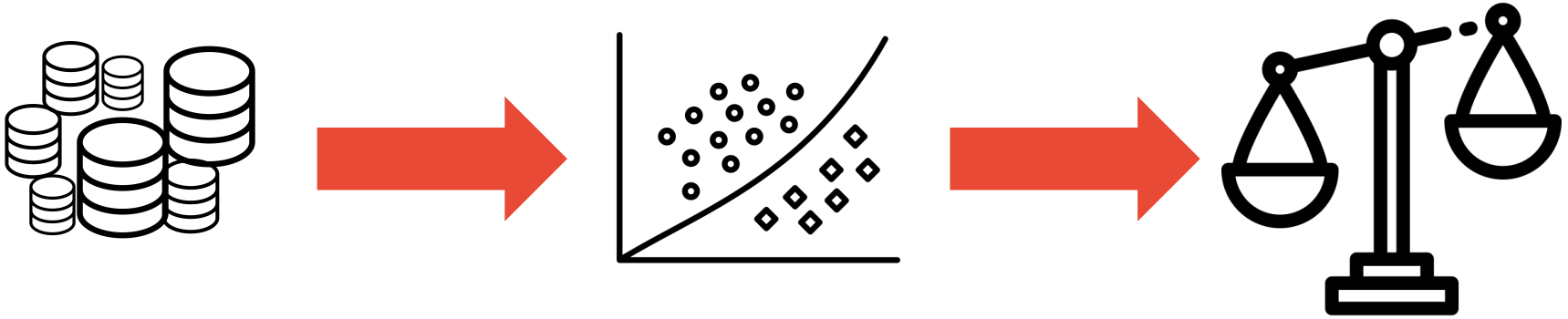


Goal of Project

Is it a right price?

No way, it's a rip-off

Then how much should it be?



- **Decrease the gap between the perceived value to the price**
- **Create a model not or less affected by outliers**

The Dataset

What Do We Have?

The Dataset

- “Are Lego Sets Too Pricey?”, Jonathan Bouchet, Kaggle
  - 12261 rows with 14 columns
  - Does not cover the whole portfolio

	ages	list_price	num_reviews	piece_count	play_star_rating	prod_desc	prod_id	prod_long_desc	review_difficulty	set_name	star_rating	theme_name	val_star_rating	country
0	6-12	29.99	2.0	277.0	4.0	Catapult into action and take back the eggs fr...	75823.0	Use the staircase catapult to launch Red into ...	Average	Bird Island Egg Heist	4.5	Angry Birds™	4.0	US
1	6-12	19.99	2.0	168.0	4.0	Launch a flying attack and rescue the eggs fro...	75822.0	Pilot Pig has taken off from Bird Island with ...	Easy	Piggy Plane Attack	5.0	Angry Birds™	4.0	US
2	6-12	12.99	11.0	74.0	4.3	Chase the piggy with lightning-fast Chuck and ...	75821.0	Pitch speedy bird Chuck against the Piggy Car....	Easy	Piggy Car Escape	4.3	Angry Birds™	4.1	US

The Dataset

Numerical Features

- list_price **[target]**
- num_reviews
- piece_count
- play_star_rating
- star_rating
- val_star_rating
- prod_id **[delete]**

Categorical Features

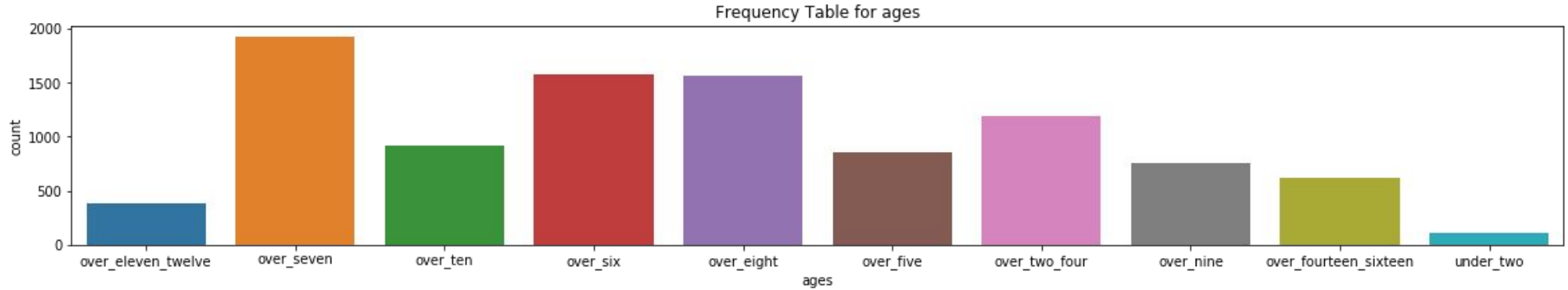
- ages
- theme_name
- review_difficulty
- country
- prod_desc **[delete]**
- prod_long_desc **[delete]**
- set_name **[delete]**

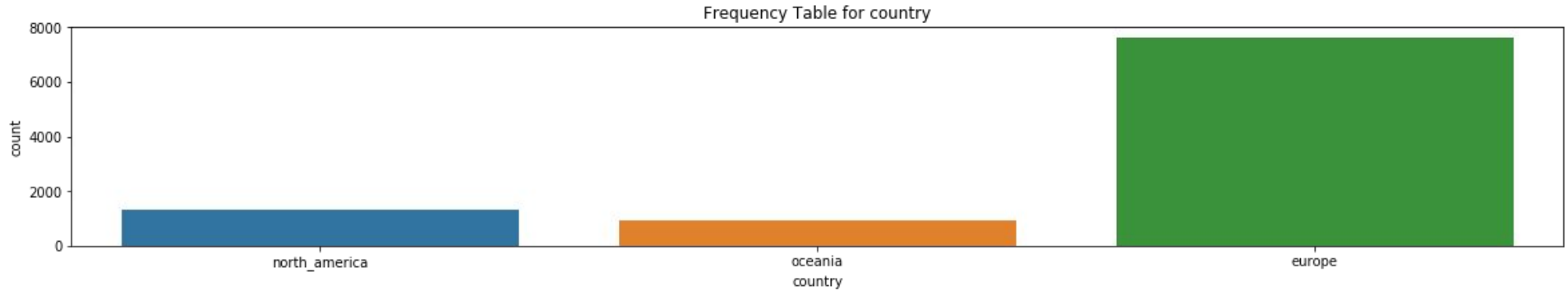
Exploratory Data Analysis (EDA)

See What We Can Do with It

EDA

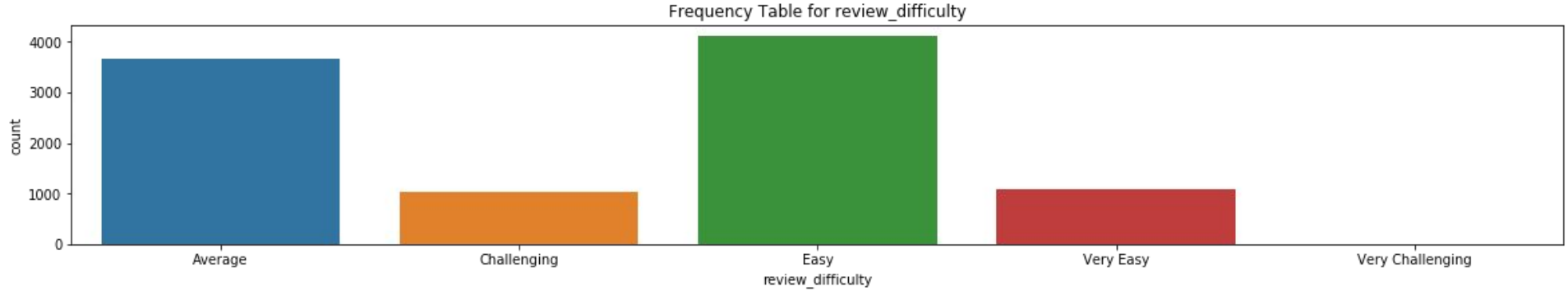
1. Histogram

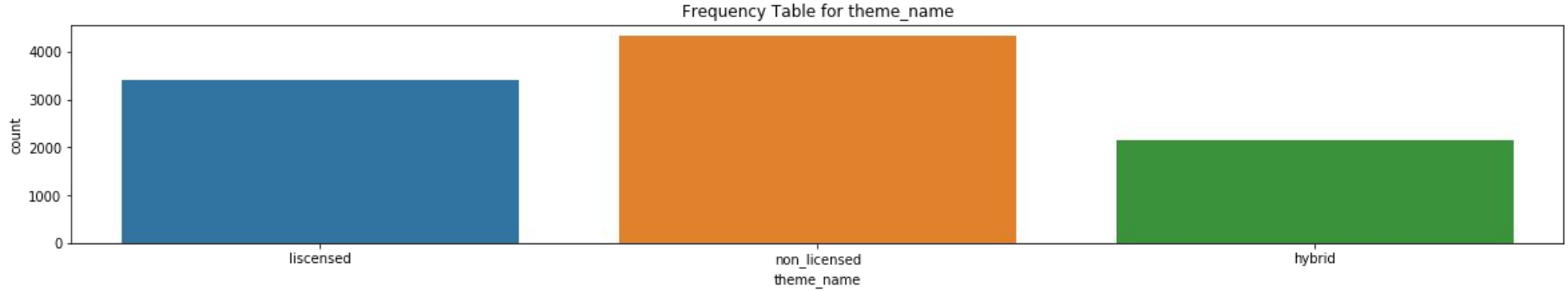




EDA

1. Histogram

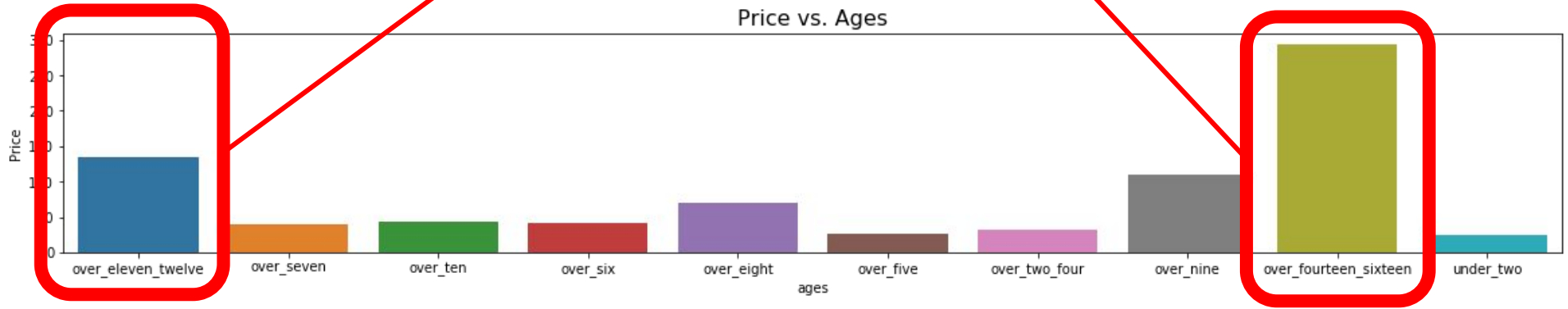


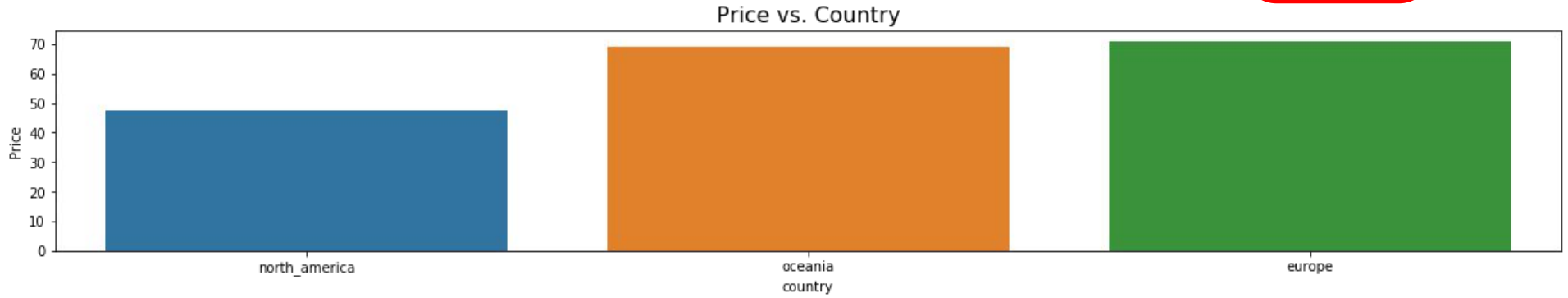


EDA

2. Barplot

11+ 12+ 14+ 16+
→ older consumer
→ more pieces
→ more expensive

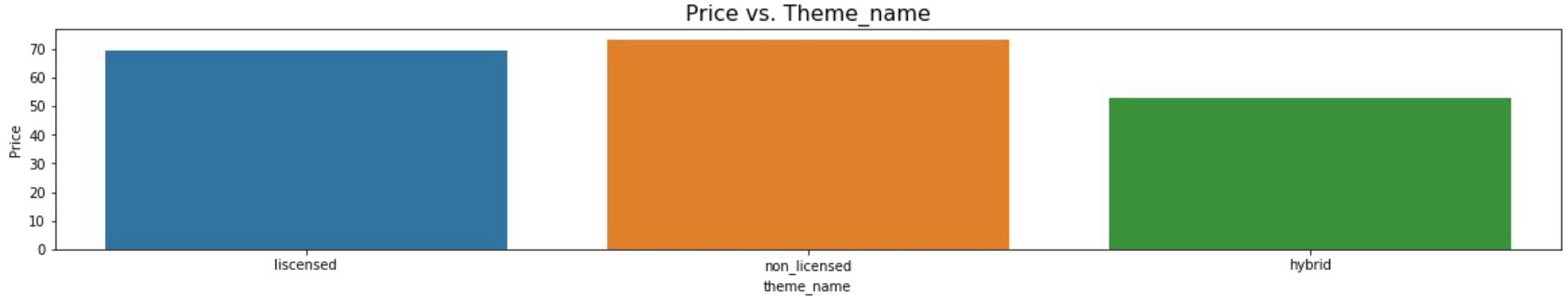




EDA

2. Barplot

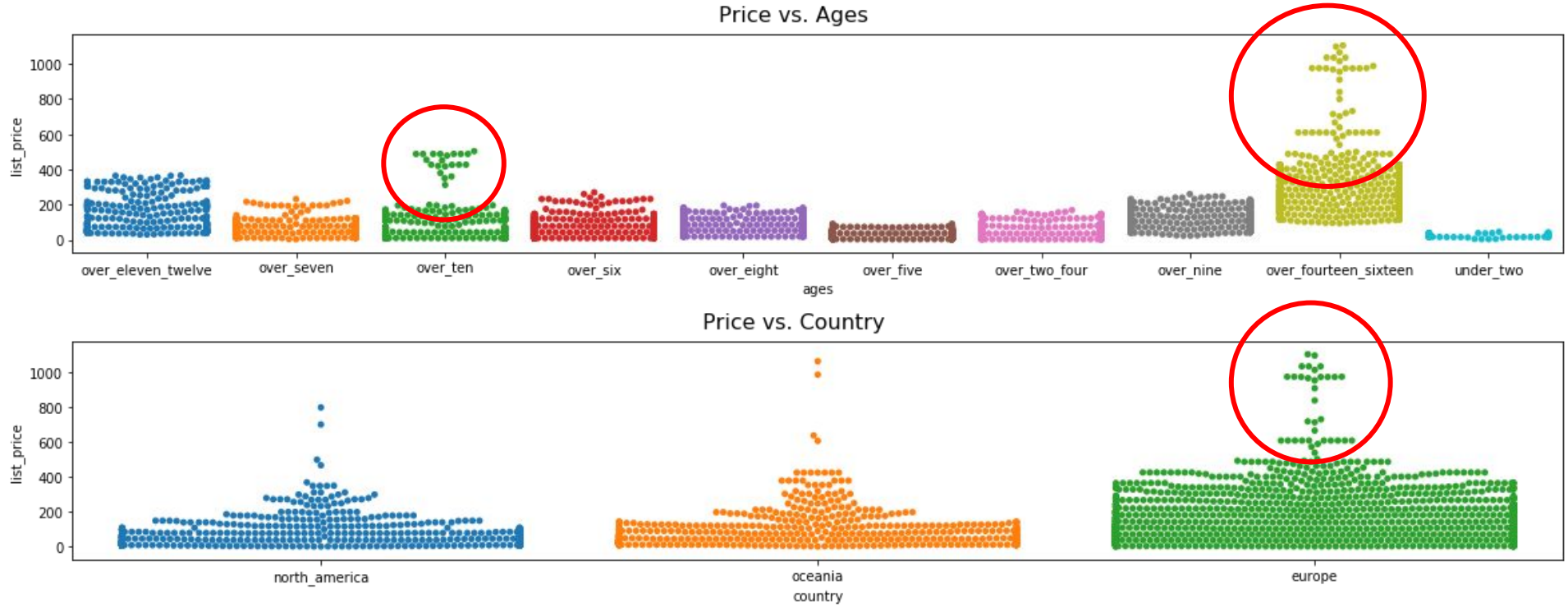
Harder
→ more pieces
→ more expensive



EDA

Outliers

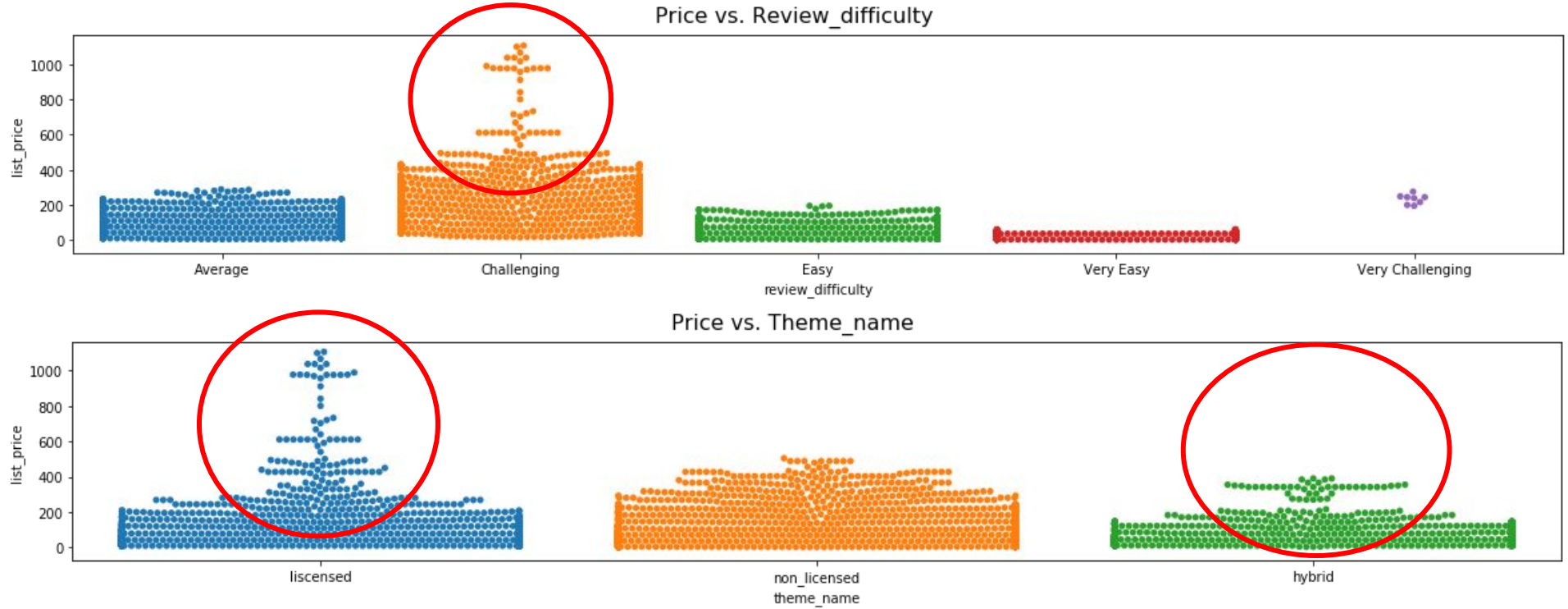
3. Swarmplot



EDA

Outliers

3. Swarmplot



EDA

4. Correlation

piece_count highly correlated with the **list_price**

piece_count correlated with **num_reviews**

ratings correlated with each other

Data Preprocessing

Get it Prepped and Ready!

Remove Unnecessary Features

1. Product Description (*prod_desc & prod_long_desc*)

   - Sets of long free texts which are difficult to handle

Go head to head with the Empire aboard The Arrowhead!

Act out perilous LEGO® Star Wars: The Freemaker Adventures missions with The Arrowhead. This aggressive-looking starship features a lift-off cockpit canopy with space for 3 minifigures and R0-GR inside, a transparent opening dome revealing the ship's removable crystal power source element, 2 spring-loaded shooters and an impressive flip-out battering ram. There's also a service cart with tools and ammo as an extra play starter.

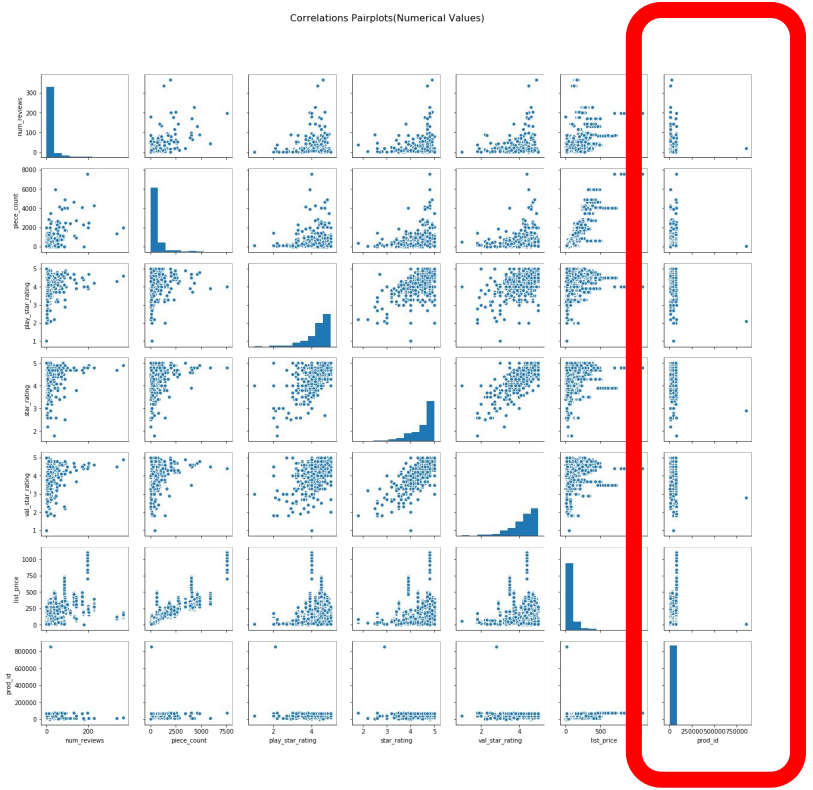
Build the Freemakers' ultimate Kyber-powered rebel starfighter, with space for 3 minifigures, a rotating blade, and spring-loaded shooters
Featured in the LEGO Star Wars: The Freemaker Adventures TV show
Includes Zander, Kordi, Quarrie, a Stormtrooper and a R0-GR figure
LEGO Star Wars building toys are compatible with all LEGO construction sets for creative building
Measures over 2" high, 20" long and 9" wide. Service cart measures over 1" long and 1" wide, and under 1" high
775 pieces – For boys and girls between the ages of 8 and 14 years old

Remove Unnecessary Features

2. Product ID (*prod_id*)

   - Merely a product identifier

   - No meaningful correlation



Remove Unnecessary Features

3. Set Names (*set_name*)

- The name of a specific set

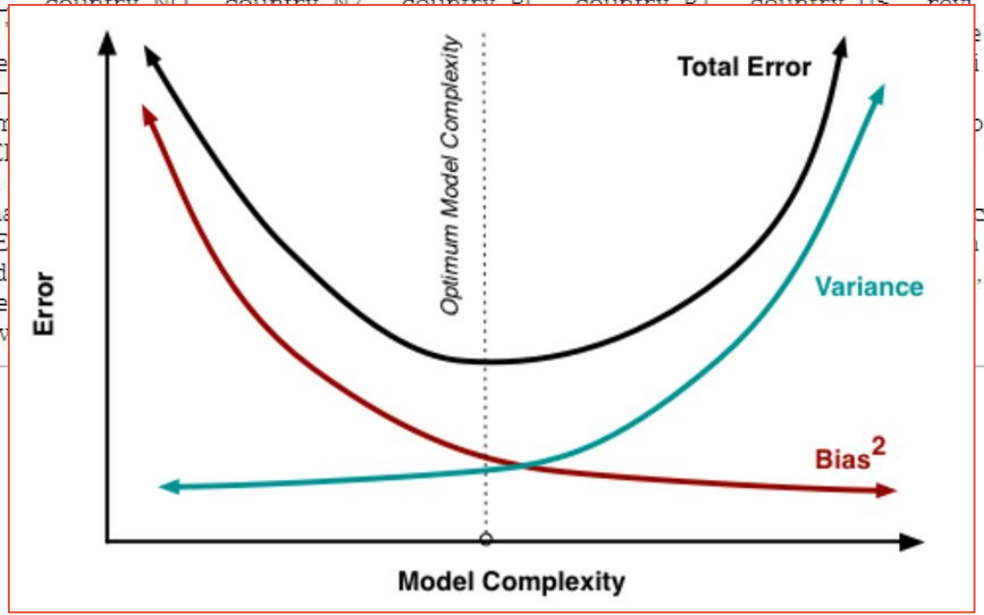
- Extremely strong price-determining factor

- What if a new set is released?

Grouping of Numerous Values

```
In [26]: lego_processed.columns

Out[26]: Index(['ages_10-14', 'ages_10-16', 'ages_11-16', 'ages_12+', 'ages_12-16', 'ages_14+', 'ages_16+', 'ages_1½-3', 'ages_1½-5', 'ages_
2-5', 'ages_4+', 'ages_4-7', 'ages_4-99', 'ages_5+', 'ages_5-12', 'ages_5-8', 'ages_6+', 'ages_6-12', 'ages_6-14', 'ages_7+', 'ages_7-
12', 'ages_7-14', 'ages_8+', 'ages_8-12', 'ages_8-14', 'ages_9+', 'ages_9-12', 'ages_9-14', 'ages_9-16', 'country_AU', 'country_BE',
'country_CA', 'country_CH', 'country_CZ', 'country_DE', 'country_DN', 'country_ES', 'country_FI', 'country_FR', 'country_GB', 'c
ountry_IE', 'country_IT', 'country_LU', 'country_NL', 'country_NO', 'country_NZ', 'country_PL', 'country_PT', 'country_US', 'revi
ew_difficulty_Challenging', 'review_difficulty_Easy',
_Architecture', 'theme_name_BOOST', 'theme_name
ty', 'theme_name_Classic', 'theme_name_Creator 3-
s', 'theme_name_DC Super Hero Girls', 'theme_nam
st Attack', 'theme_name_Disney™', 'theme_name_E
s', 'theme_name_Indoraptor Rampage at Lockwood
e', 'theme_name_LEGO® Creator 3-in-1', 'theme_na
raft™', 'theme_name_Minifigures', 'theme_name_NE
eme_name_Pteranodon Chase', 'theme_name_Speed
'theme_name_THE LEGO® BATMAN MOVIE', 'theme
s', 'piece_count', 'play_star_rating', 'star_rating', 'v
```



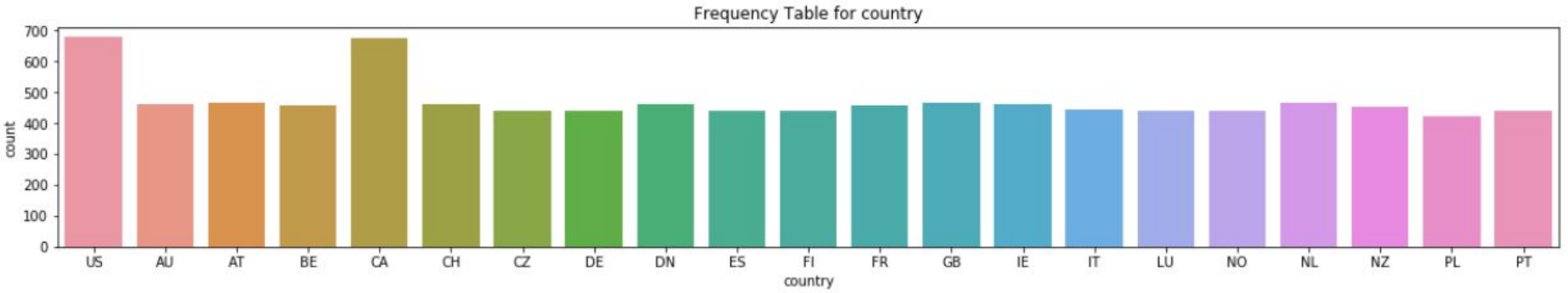
- Model too complex
- Risk of overfitting

Grouping of Values

1. Country Names (*country*)

   - Regrouping by the continents where each country belongs

   - Europe, North America, and Oceania

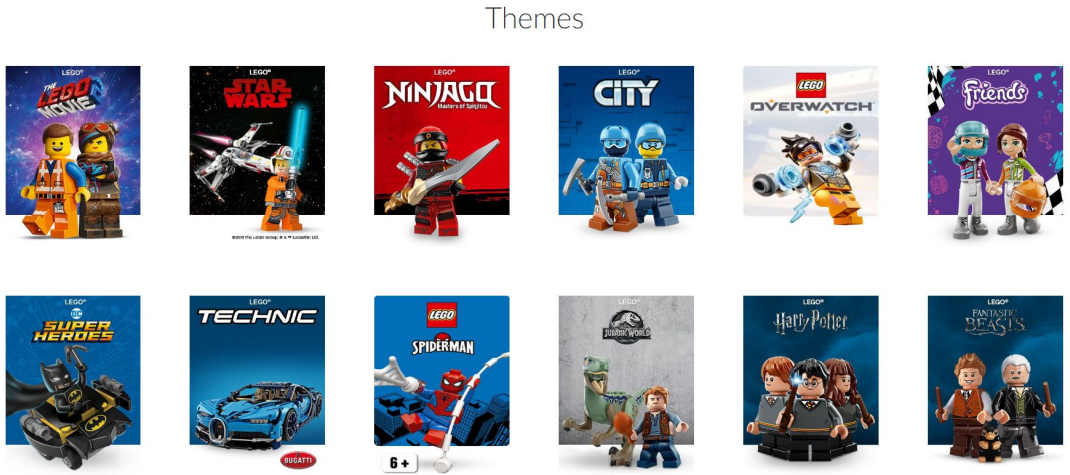


Grouping of Numerous Values

2. Theme Names (*theme_name*)

- Broader than individual set names

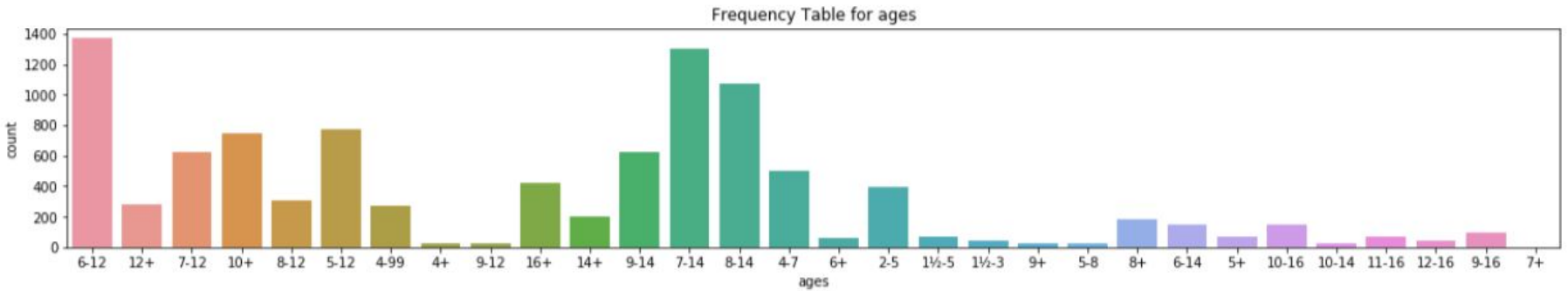
- e.g. ‘licensed’, ‘non-licensed’, ‘hybrid’



Grouping of Numerous Values

3. Age Range (*ages*)

- Difficult to reclassify due to inconsistency of criteria
- Set based on the boundary value of the range
- e.g. ‘under_two’, “over_two_four”, “over_ten” etc



Categorical Value Encoding

- One-hot Encoding

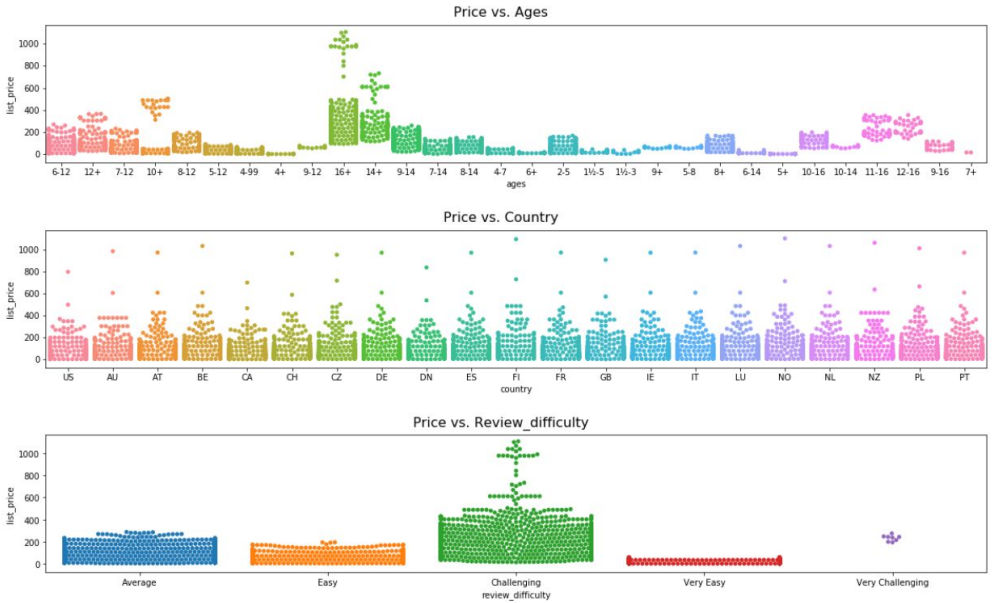
e.g. review_difficulty

review_difficulty_Challenging
review_difficulty_Easy
review_difficulty_Very Challenging
review_difficulty_Very Easy

Outlier Removal and Standardization

- Detected in the EDA stage: sklearn, ‘IsolationForest’

- Scaling: sklearn, ‘StandardScaler()’



Multi-collinearity Problem

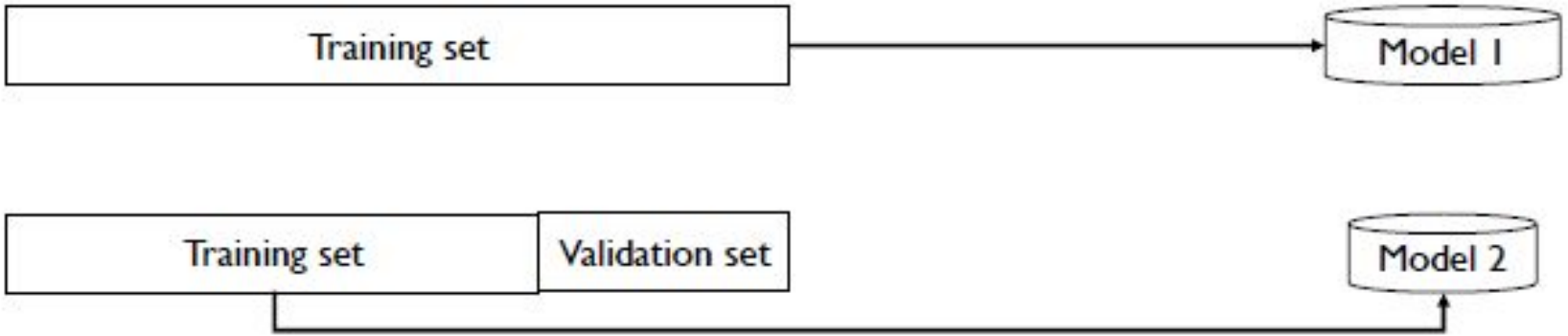
- Increases the variance of the model estimator
- Variance Inflation Factor (VIF)
  - Before & After scaling
  - Collinearity fair when VIF < 10

	VIF Factor	features
0	1.816666	num_reviews
1	2.104650	piece_count
2	70.382928	play_star_rating
3	146.832094	star_rating
4	94.723226	val_star_rating

	VIF Factor	features
0	1.488600	num_reviews
1	1.497245	piece_count
2	1.614229	play_star_rating
3	2.605206	star_rating
4	2.191309	val_star_rating

Data Split (Hold-Out)

- Train : Validation : Test = 60% : 20% : 20%



Model Fitting and Evaluation

Get Back to the Real Job

Ordinary Least Squares (OLS)

- Full Model with **94** features before grouping values
- High R square and Significant p-value

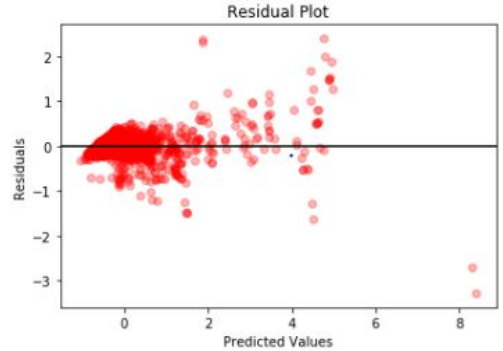
OLS Regression Results

Dep. Variable:	list_price	**R-squared:**	0.921
Model:	OLS	**Adj. R-squared:**	0.920
Method:	Least Squares	**F-statistic:**	1268.
Date:	Fri, 21 Dec 2018	**Prob (F-statistic):**	0.00
Time:	00:41:40	**Log-Likelihood:**	-1499.9
No. Observations:	9910	**AIC:**	3180.
Df Residuals:	9820	**BIC:**	3828.
Df Model:	90		
Covariance Type:	nonrobust		

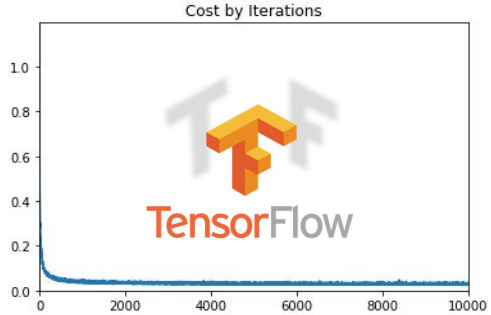
Neural Network (Tensorflow)

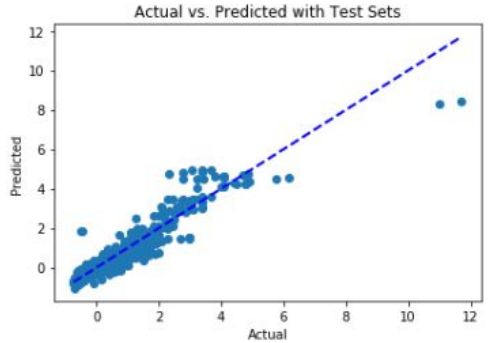
- What about the Neural Network?

epoch = 10000, dropout rate = 0.7, 2 hidden layers



MSE: 0.082122125





MSE: 0.0331

Plain MLR (MSE: 0.08)

(?, 94) (22, 50) ⋯ (50, 50) (50, 1)

NN (MSE: 0.03)

Ordinary Least Squares (OLS)

- Reduced model with **22** features after grouping values
- Still high R square and significant p-value

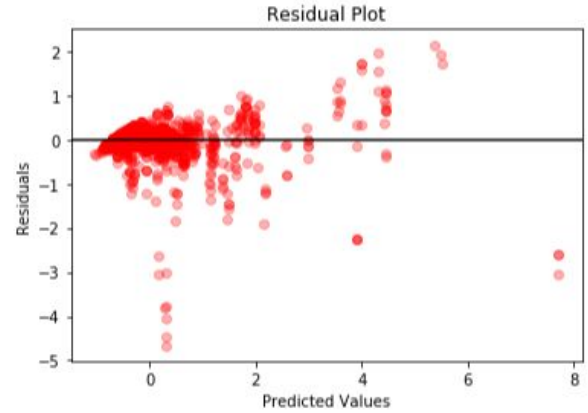
OLS Regression Results

Dep. Variable:	list_price	**R-squared:**	0.853
Model:	OLS	**Adj. R-squared:**	0.853
Method:	Least Squares	**F-statistic:**	2610.
Date:	Fri, 21 Dec 2018	**Prob (F-statistic):**	0.00
Time:	01:01:25	**Log-Likelihood:**	-4552.2
No. Observations:	9904	**AIC:**	9148.
Df Residuals:	9882	**BIC:**	9307.
Df Model:	22		
Covariance Type:	nonrobust		

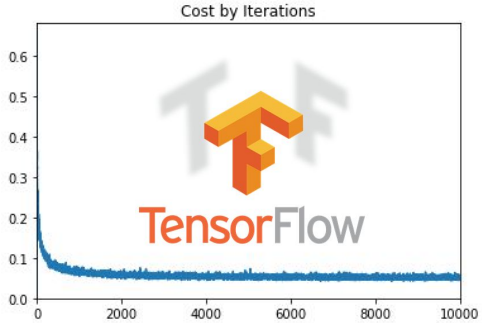
Neural Network (Tensorflow)

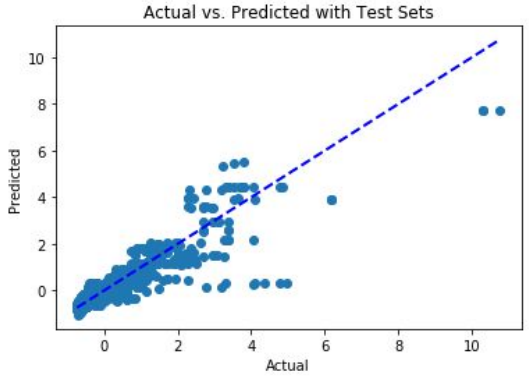
- What about the Neural Network?

epoch = 10000, dropout rate = 0.7, 2 hidden layers



MSE: 0.15949753





MSE: 0.0492

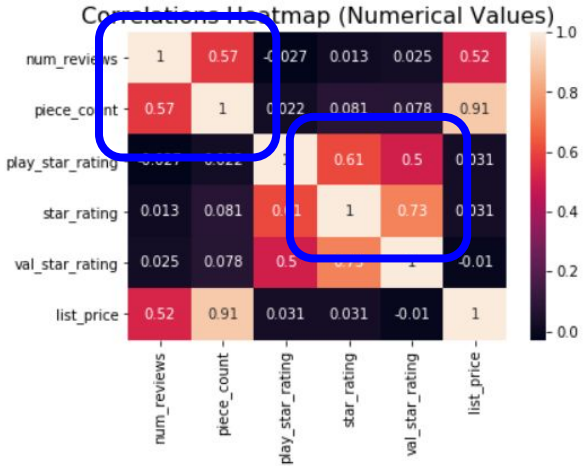
Plain MLR (MSE: 0.16)

(?, 22) (22, 50) ⋯ (50, 50) (50, 1)

NN (MSE: 0.05)

Remaining Problems

- Not bad. But,
  - Despite the sample size, maybe try a simpler one?
  - and also collinearity still remaining...



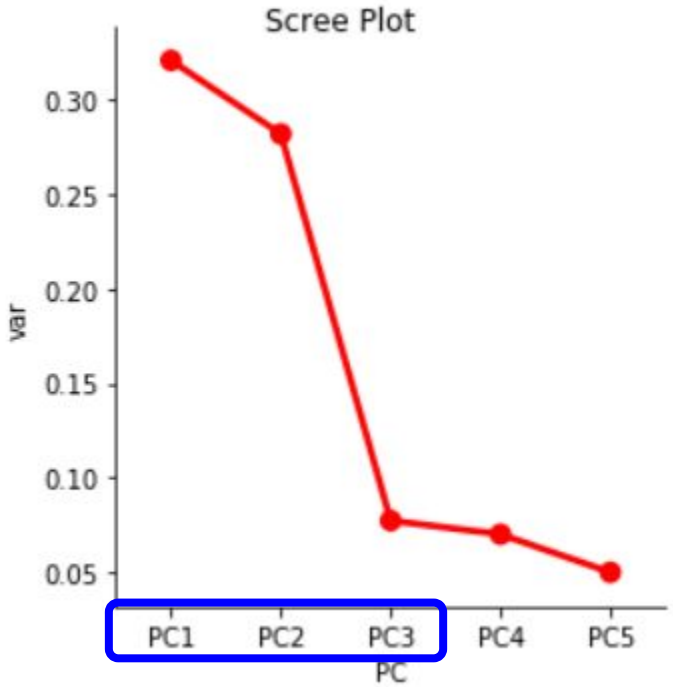
	coef	std err	t	P>|t|	[0.025	0.975]
ages_over_eleven_twelve	-0.1115	0.023	-4.887	0.000	-0.156	-0.067
ages_over_five	0.1019	0.017	5.901	0.000	0.068	0.136
ages_over_fourteen_sixteen	-0.2356	0.031	-7.501	0.000	-0.297	-0.174
ages_over_nine	0.0636	0.016	3.884	0.000	0.032	0.096
ages_over_seven	0.0456	0.012	3.886	0.000	0.023	0.069
ages_over_six	0.0497	0.013	3.802	0.000	0.024	0.075
ages_over_ten	0.1188	0.018	6.685	0.000	0.084	0.154
ages_over_two_four	0.2101	0.014	14.868	0.000	0.182	0.238
ages_under_two	0.1635	0.037	4.390	0.000	0.090	0.236
country_north_america	-0.1445	0.011	-12.789	0.000	-0.167	-0.122
country_oceania	0.0003	0.013	0.021	0.983	-0.026	0.026
review_difficulty_Challenging	0.1719	0.017	9.874	0.000	0.138	0.206
review_difficulty_Easy	-0.1052	0.010	-10.921	0.000	-0.124	-0.086
review_difficulty_Very Challenging	0.6023	0.139	4.346	0.000	0.331	0.874
review_difficulty_Very Easy	-0.1636	0.015	-10.616	0.000	-0.194	-0.133
theme_name_liscensed	0.1101	0.011	10.468	0.000	0.089	0.131
theme_name_non_licensed	-0.0812	0.009	-8.671	0.000	-0.100	-0.063
num_reviews	0.0259	0.006	4.655	0.000	0.015	0.037
piece_count	0.9177	0.007	124.446	0.000	0.903	0.932
play_star_rating	0.0825	0.006	13.723	0.000	0.071	0.094
star_rating	-0.0252	0.007	-3.788	0.000	-0.038	-0.012
val_star_rating	-0.1083	0.006	-18.607	0.000	-0.120	-0.097

Principal Component Analysis (PCA)

- Let's try again with few important variables!

  - Optimal number of PC: 2

  - But, we'll take **3**.



OLS Again

- The Simplest Model

OLS Regression Results

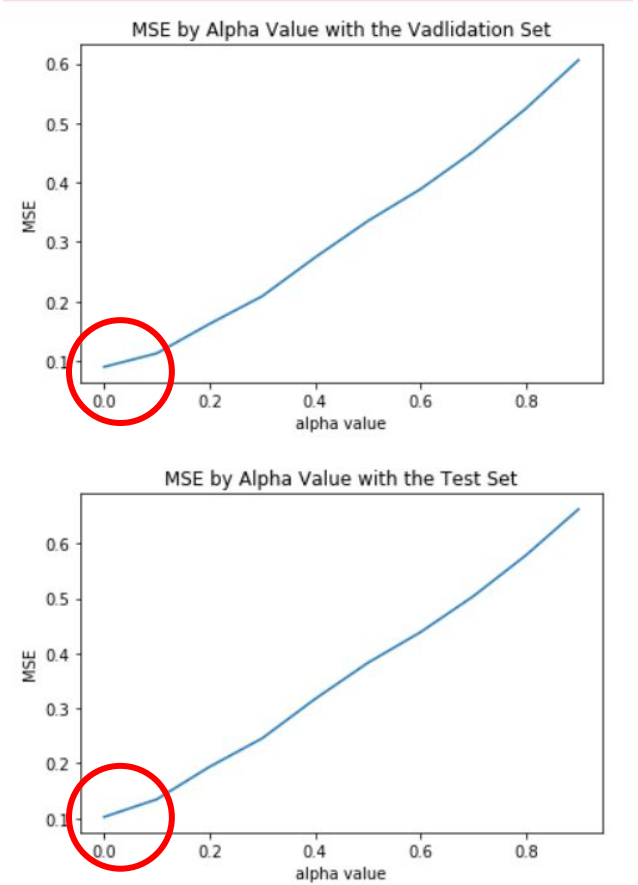
Dep. Variable:	list_price	**R-squared:**	0.909
Model:	OLS	**Adj. R-squared:**	0.908
Method:	Least Square	**F-statistic:**	3.278e+04
Date:	Fri, 21 Dec 20 8	**Prob (F-statistic):**	0.00
Time:	01:04:59		
No. Observations:	9903	**AIC:**	4425.
Df Residuals:	9900	**BIC:**	4447.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PC1	0.5556	0.002	290.073	0.000	0.552	0.559
PC2	0.2022	0.002	98.922	0.000	0.198	0.206
PC3	-0.2592	0.004	-66.339	0.000	-0.267	-0.252

Omnibus:	3369.407	**Durbin-Watson:**	1.197
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	116611.084
Skew:	0.975	**Prob(JB):**	0.00
Kurtosis:	19.697	**Cond. No.**	2.04

Least Absolute Shrinkage and Selection Operator (LASSO)

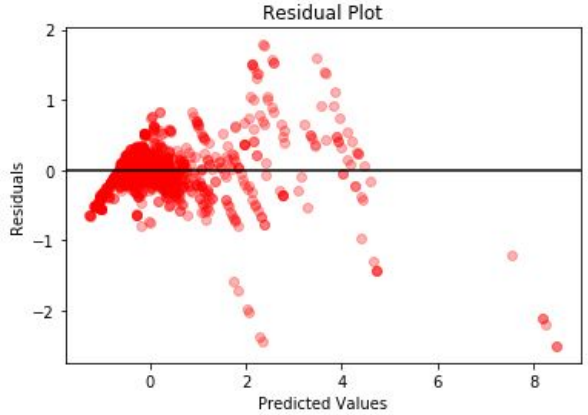
- Regularization?
- Best performs where alpha = 0
- No weight constrainment needed



NN Again

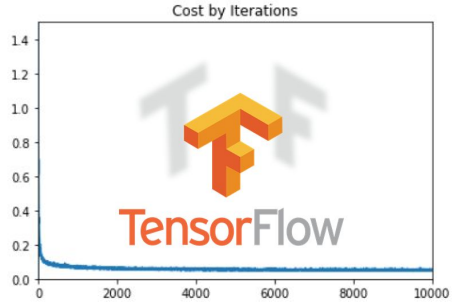
- What about the Neural Network?

epoch = 10000, dropout rate = 0.7, 2 hidden layers

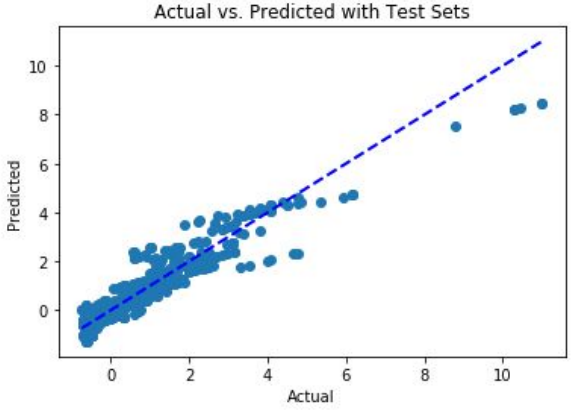


MSE: 0.1022482

Plain MLR (MSE: 0.1)



(?, 3) (3, 50) ··· (50, 50) (50, 1)



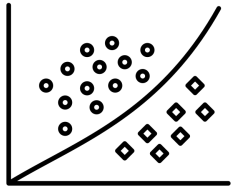
MSE: 0.0521

NN (MSE: 0.05)

Model Summary

- Efficiency– Accuracy Tradeoff
  - Both Model better with many features
  - However, NN takes time

Number of features	plain MLR	NN
94	0.0821	0.0331
22	0.1595	0.0492
3	0.1022	0.0521

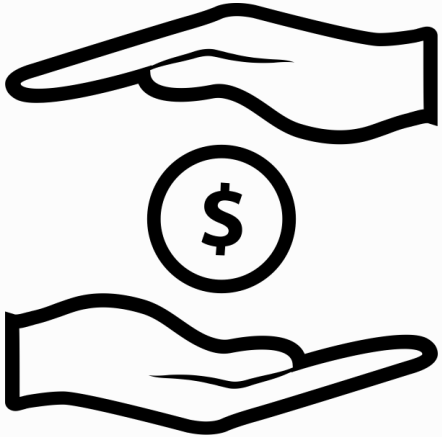


Conclusion

What are the findings and lessons?

What we can do with this model?

- Setting a basis of LEGO set price
  - Customer-oriented viewpoint
- Beneficial to both the customers and the manufacturer
- Applies to all consumer goods
- LR is useful, and powerful combined with NN



Limitations

- Incomplete nature of the dataset
  - Does not cover all product lines
  - Does not cover all the countries
  - Does not cover other features
    - e.g.) Costs, inventories
- Better modeling with better dataset

Thank You

Merry Christmas