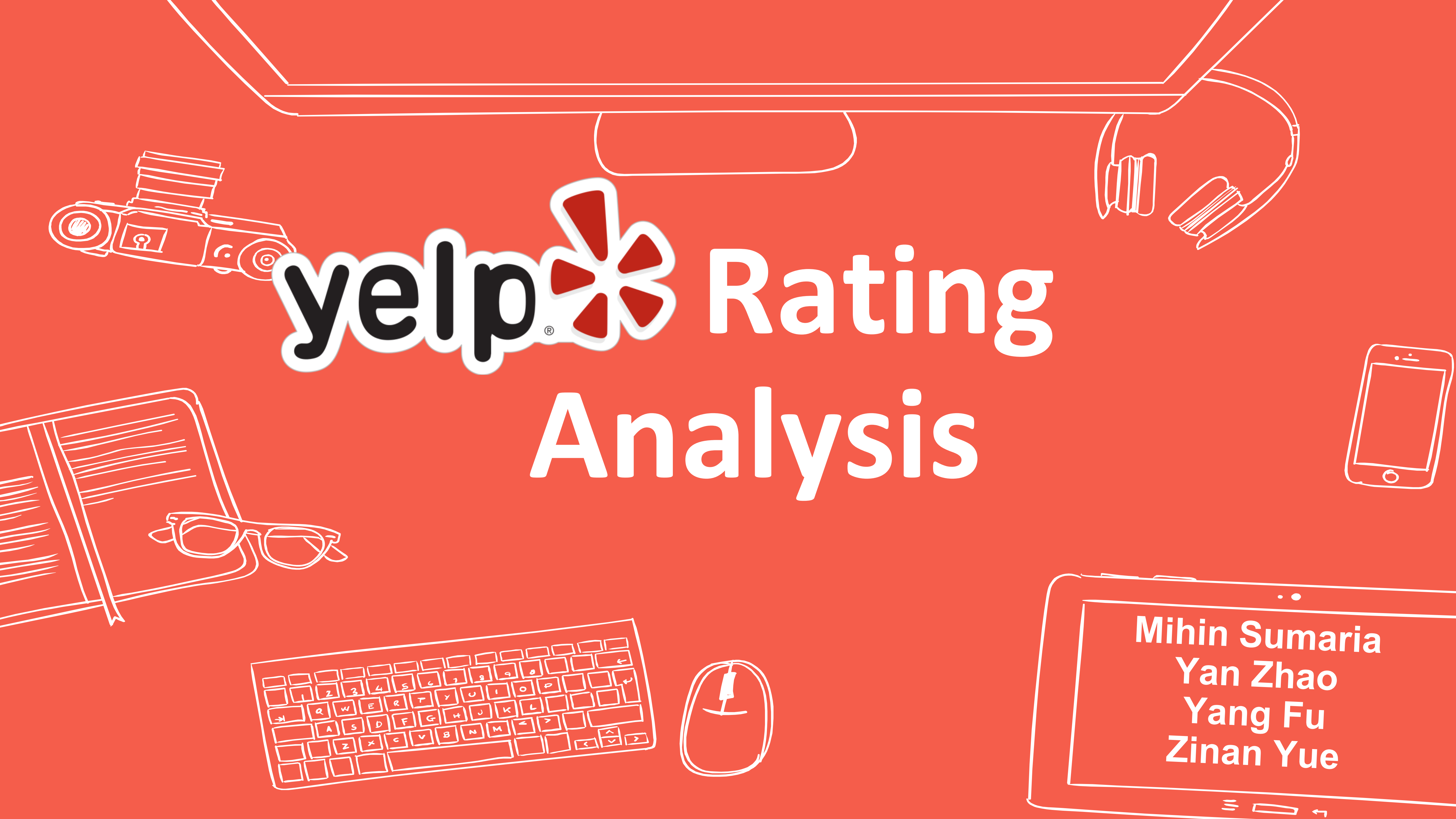
yelp Rating Analysis



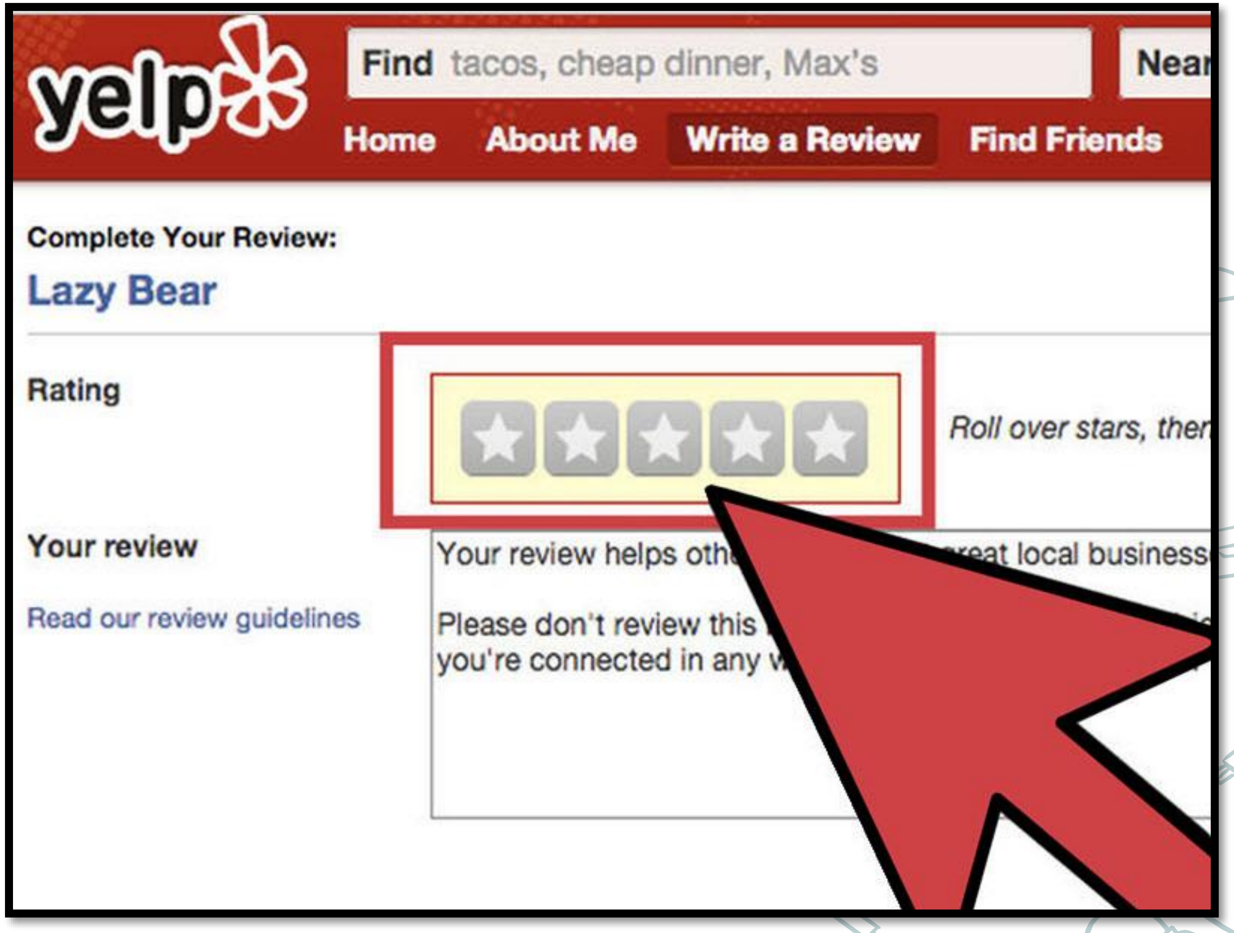
Mihin Sumaria
Yan Zhao
Yang Fu
Zinan Yue

Introduction

- Yelp allows users to review and rate various businesses online.

- A review consists of free-form text and a star rating of 1-5.

- Yelp reviews significantly influence consumers' behaviors.



Complete Your Review:

Lazy Bear

Rating

Roll over stars, then

Your review

Read our review guidelines

Your review helps oth reat local business

Please don't review this

you're connected in any w

Image Courtesy: https://www.wikihow.com/Find-and-Write-a-Business-Review-on-Yelp

Problem to solve: Review Rating Prediction

- Predict the star ratings for businesses using users’ reviews and users’ personal information.

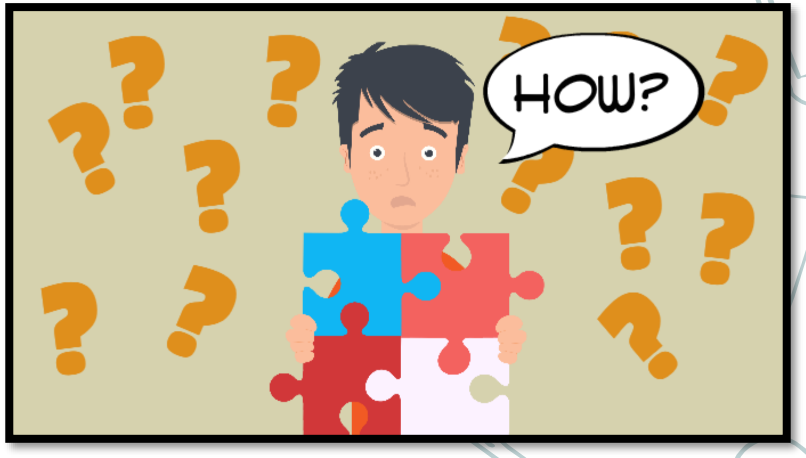


Image Courtesy: https://oliverbenn.com/the-10-step-process-to-solve-any-problem-2

Data Collection & Processing

- Extract data from review.json and user.json and merge them into a dataframe.
  - text, cool, funny, useful, user_id
  - fans, user_cool, user_funny, user_id

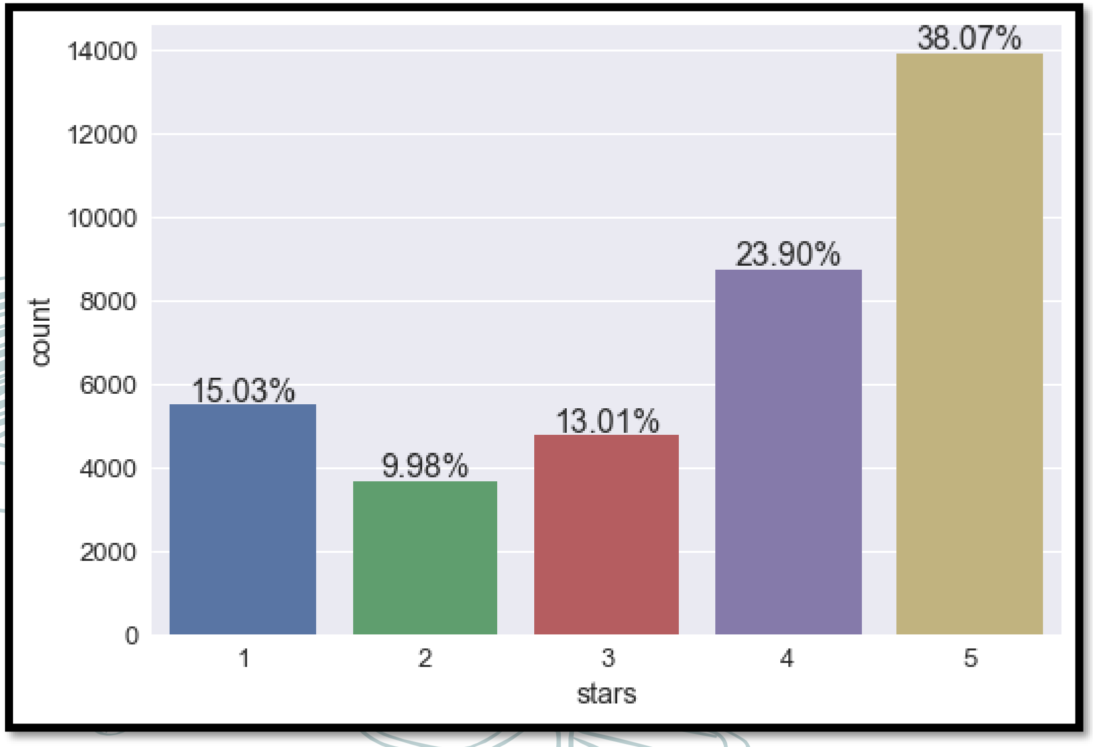
- The review.json has over 5 million records
  - Our laptops were not happy at all
  - Categorize reviews by State using business.json
  - Size down to the State of Illinois: approx. 35000 records

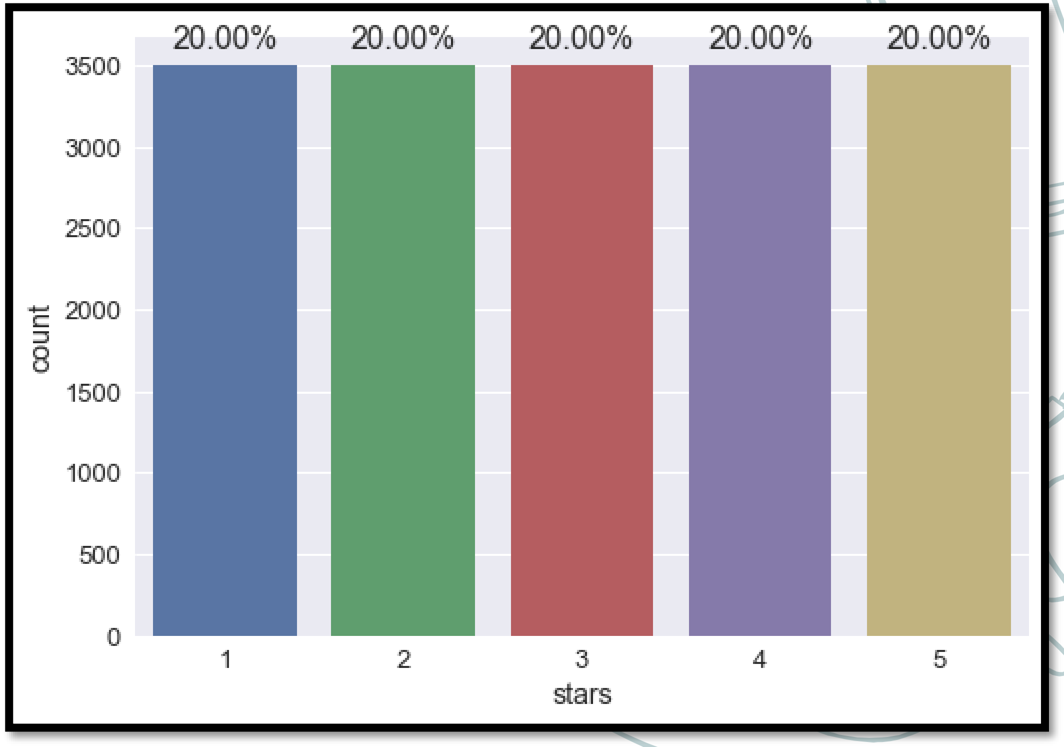
Sentiment Analysis on reviews

- Textblob library - processing the review in common natural language
- Polarity and Subjectivity value returned.
- **Subjectivity:** Float value range [0.0, 1.0], 0.0 is most objective and 1.0 is most subjective
- **Polarity:** Float value range [-1.0, 1.0], -1.0 is most negative and 1.0 is most positive
- Added the Subjectivity and Polarity columns into the dataframe

Data Processing

- Class Imbalance
- Sample 3,500 for each rating.





Data Processing

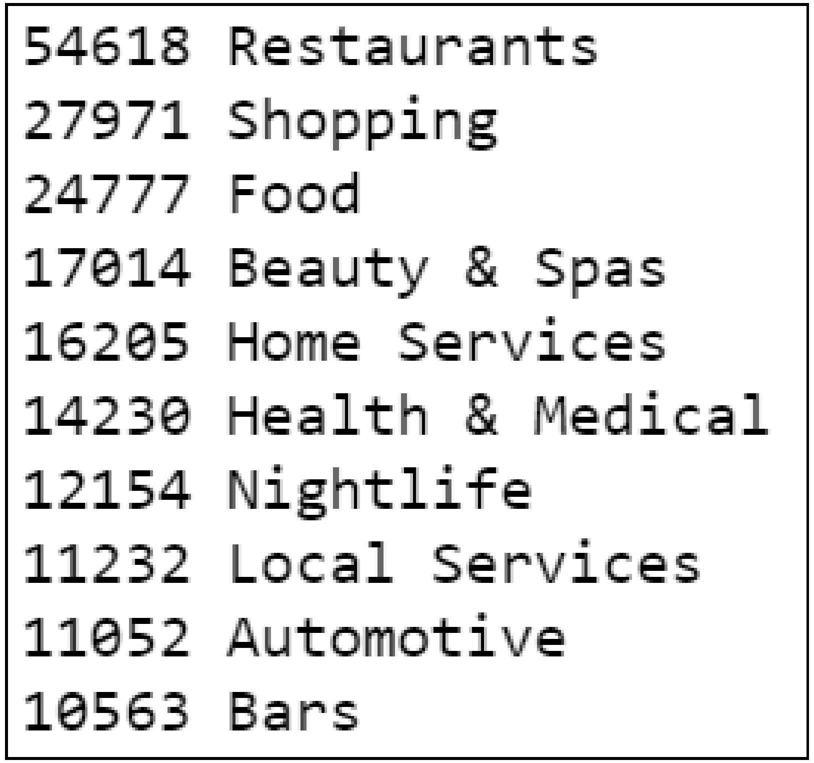
- Split dataset as training data and test data, 25% Testing 75% Training.

- Predictors that we used 'cool', 'funny', 'useful', 'fans', 'user_cool’, 'user_funny', 'user_useful', 'polarity', 'subjectivity’

- Response variable 'stars'

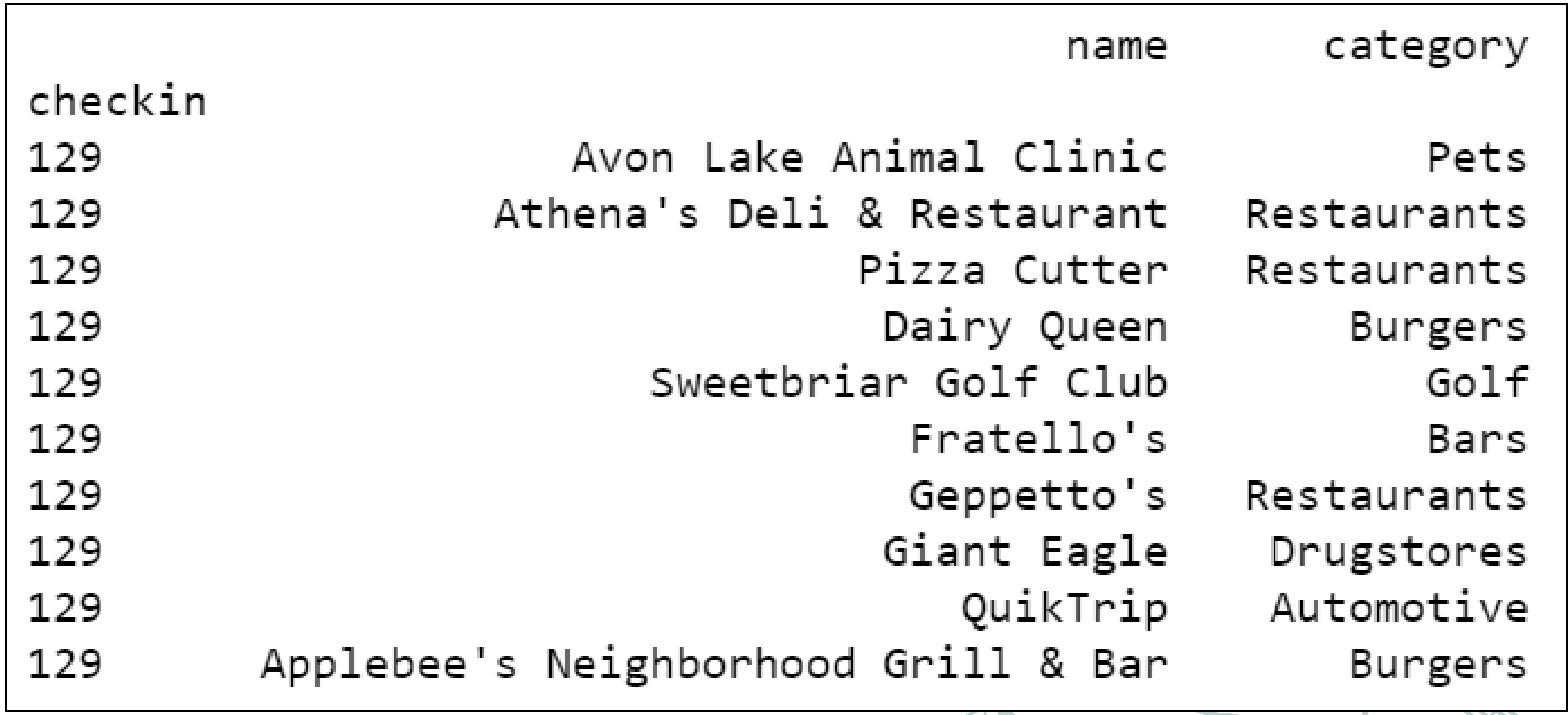
Data Exploration

Top 10 Categories

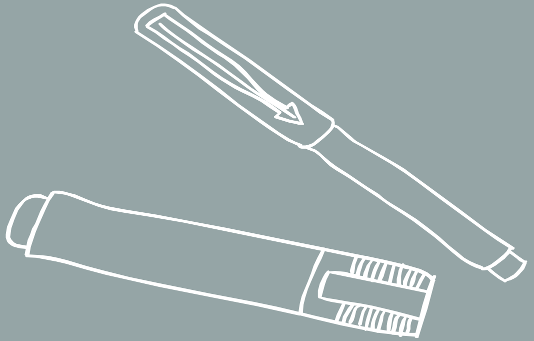
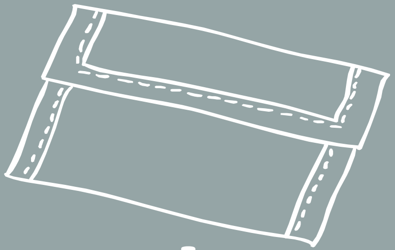
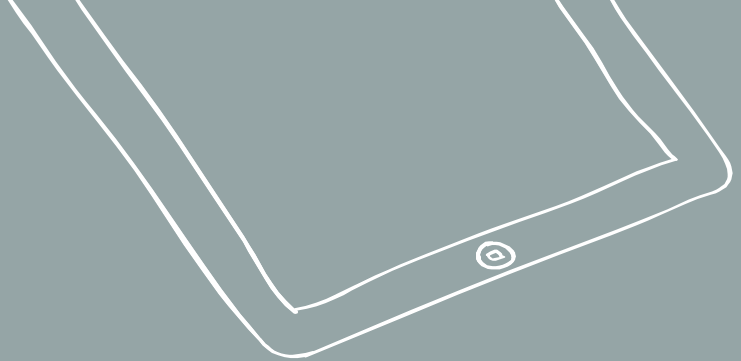


```
54618 Restaurants
27971 Shopping
24777 Food
17014 Beauty & Spas
16205 Home Services
14230 Health & Medical
12154 Nightlife
11232 Local Services
11052 Automotive
10563 Bars
```

Top 10 Business Objects



```
                                            name     category
checkin
129                       Avon Lake Animal Clinic         Pets
129                     Athena's Deli & Restaurant  Restaurants
129                                  Pizza Cutter  Restaurants
129                                  Dairy Queen      Burgers
129                           Sweetbriar Golf Club         Golf
129                                    Fratello's         Bars
129                                     Geppetto's  Restaurants
129                                   Giant Eagle  Drugstores
129                                      QuikTrip   Automotive
129         Applebee's Neighborhood Grill & Bar      Burgers
```

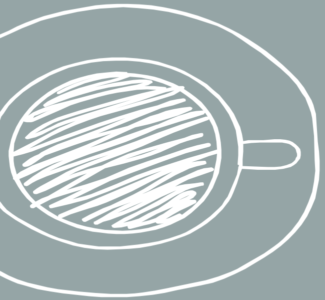




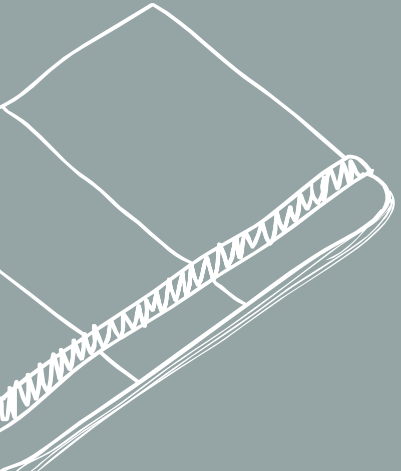
Methods

Linear Regression

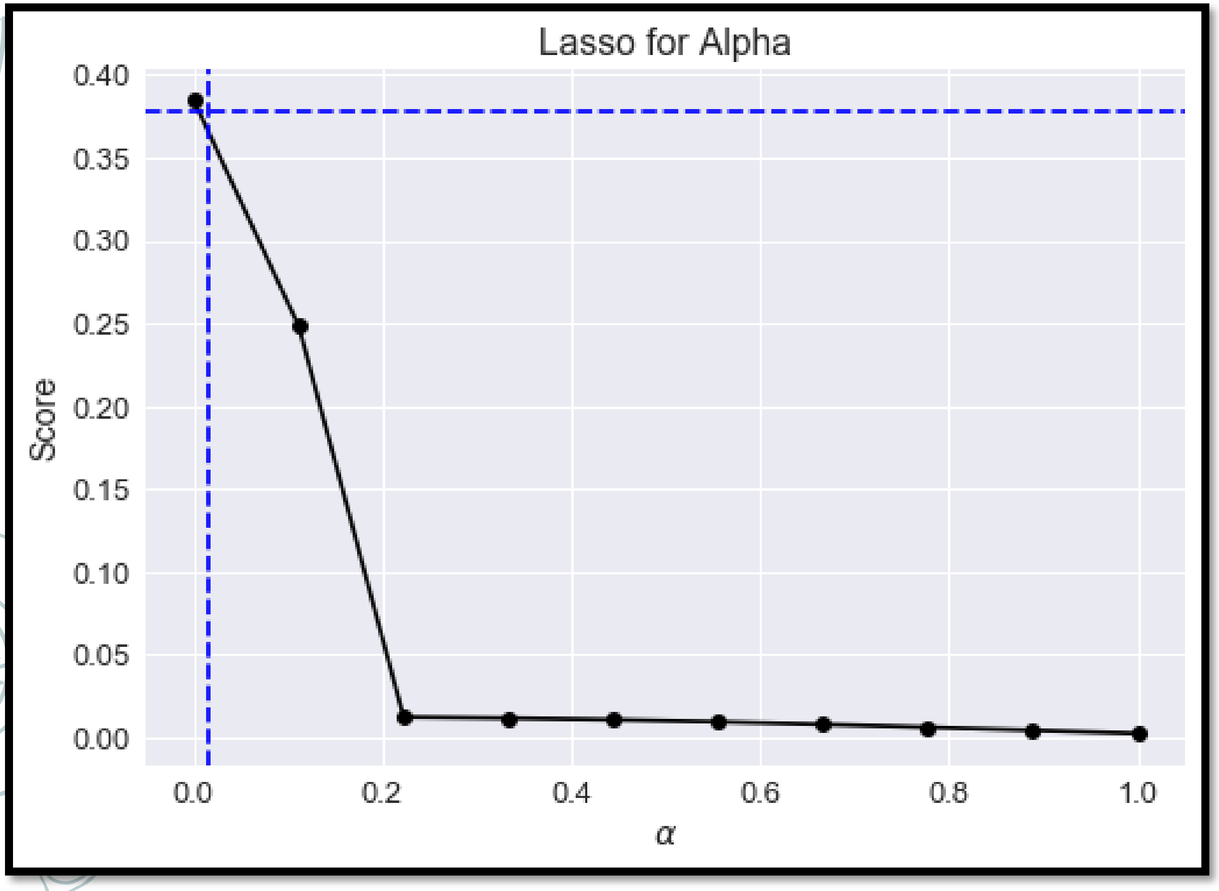
LASSO

Polynomial Regression





Lasso Regression



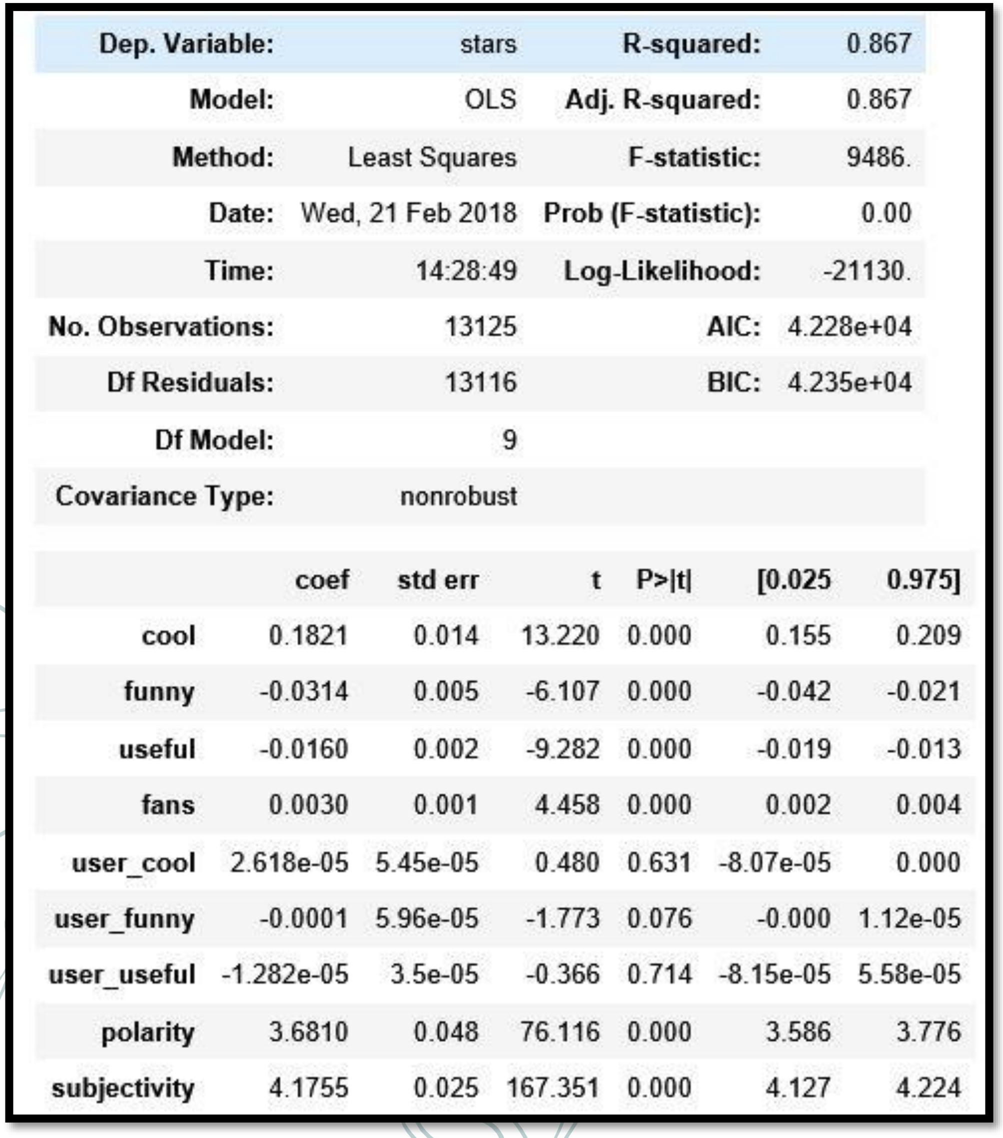
'Best alpha’ = 0.0164

'Best coefficient’:
[1.55840546e-01, -3.90019839e-02, -2.01887390e-02, 1.62620499e-03, 7.15103908e-05, -1.18584846e-04, -2.86495571e-05, 3.40648298e+00, 0.00000000e+00]

'Non-zero coefficient number is ‘ = 8

Mean_error = 5.2803564442

Linear Regression – All variables



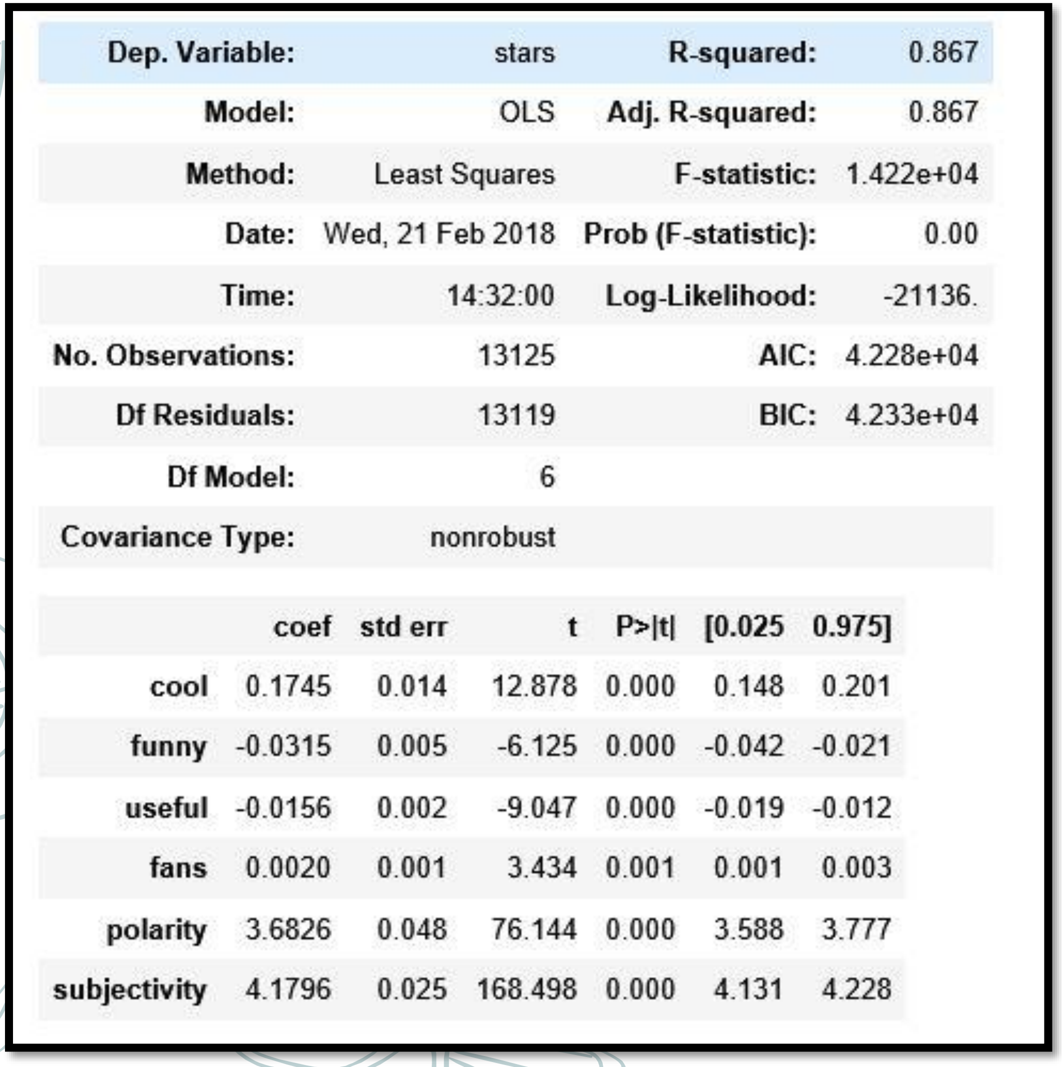
Dep. Variable:	stars	**R-squared:**	0.867
Model:	OLS	**Adj. R-squared:**	0.867
Method:	Least Squares	**F-statistic:**	9486.
Date:	Wed, 21 Feb 2018	**Prob (F-statistic):**	0.00
Time:	14:28:49	**Log-Likelihood:**	-21130.
No. Observations:	13125	**AIC:**	4.228e+04
Df Residuals:	13116	**BIC:**	4.235e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
cool	0.1821	0.014	13.220	0.000	0.155	0.209
funny	-0.0314	0.005	-6.107	0.000	-0.042	-0.021
useful	-0.0160	0.002	-9.282	0.000	-0.019	-0.013
fans	0.0030	0.001	4.458	0.000	0.002	0.004
user_cool	2.618e-05	5.45e-05	0.480	0.631	-8.07e-05	0.000
user_funny	-0.0001	5.96e-05	-1.773	0.076	-0.000	1.12e-05
user_useful	-1.282e-05	3.5e-05	-0.366	0.714	-8.15e-05	5.58e-05
polarity	3.6810	0.048	76.116	0.000	3.586	3.776
subjectivity	4.1755	0.025	167.351	0.000	4.127	4.224

‘User_cool’, ‘User_funny’ & ‘User_useful’ are not statistically significant.

Relatively good R-square

Linear Regression – Improved



Dep. Variable:	stars	**R-squared:**	0.867
Model:	OLS	**Adj. R-squared:**	0.867
Method:	Least Squares	**F-statistic:**	1.422e+04
Date:	Wed, 21 Feb 2018	**Prob (F-statistic):**	0.00
Time:	14:32:00	**Log-Likelihood:**	-21136.
No. Observations:	13125	**AIC:**	4.228e+04
Df Residuals:	13119	**BIC:**	4.233e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
cool	0.1745	0.014	12.878	0.000	0.148	0.201
funny	-0.0315	0.005	-6.125	0.000	-0.042	-0.021
useful	-0.0156	0.002	-9.047	0.000	-0.019	-0.012
fans	0.0020	0.001	3.434	0.001	0.001	0.003
polarity	3.6826	0.048	76.144	0.000	3.588	3.777
subjectivity	4.1796	0.025	168.498	0.000	4.131	4.228

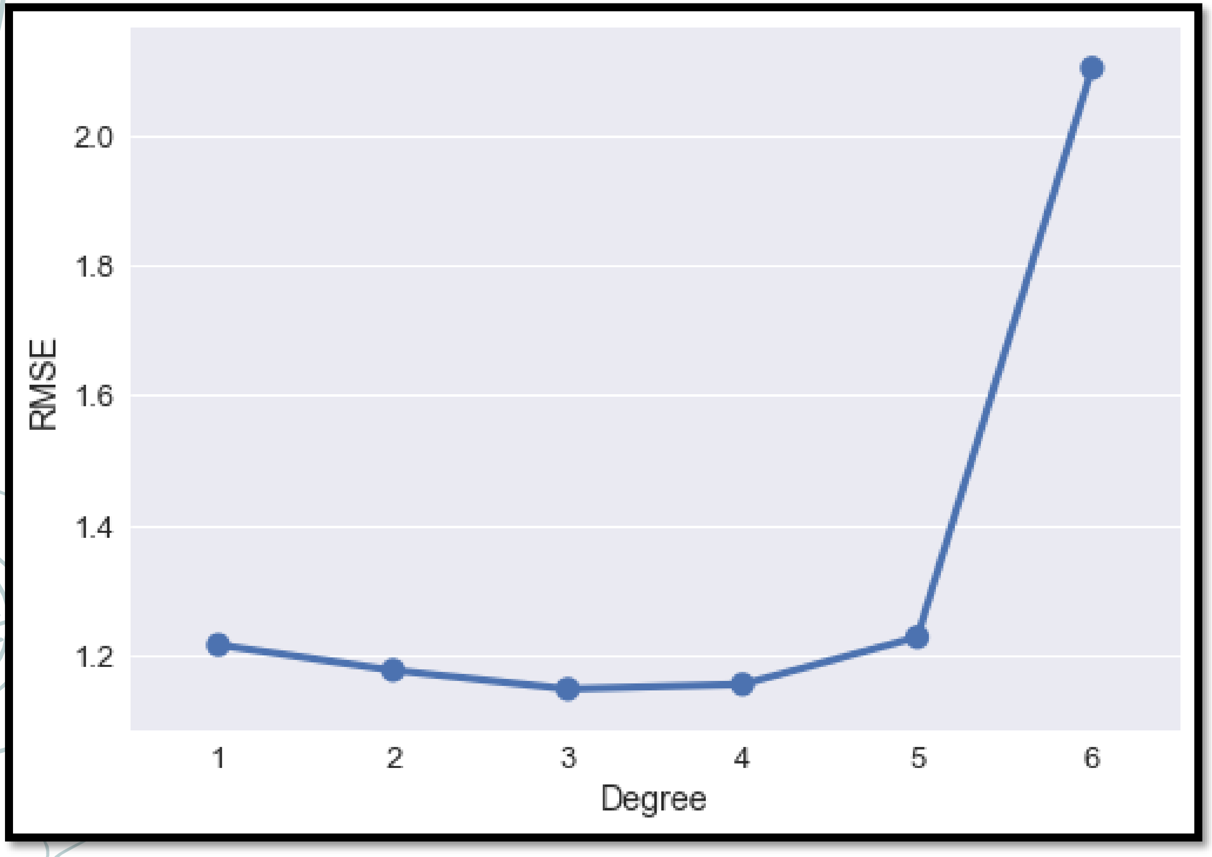
All variables are statistically significant.

Positively Important variables:
‘Polarity’
‘Subjectivity’
‘Review coolness’

Negatively correlated to the ratings:
‘Review funniness’
‘Review usefulness’

Mean error = 1.3722157328468485

Polynomial Regression

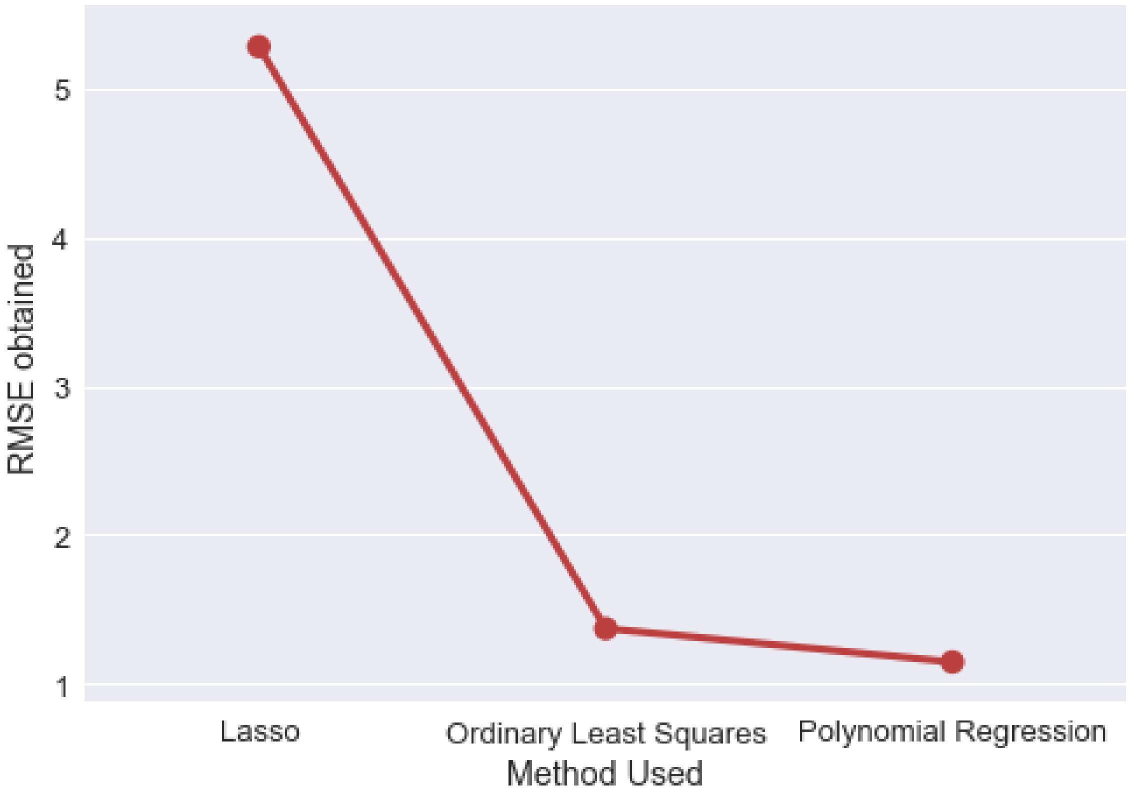


- RMSE decreases with increasing degree initially, but then increases due to overfitting of the data.

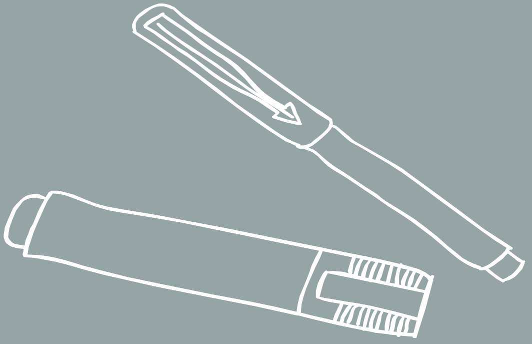
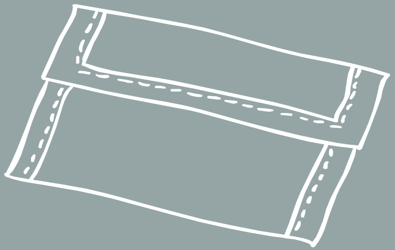
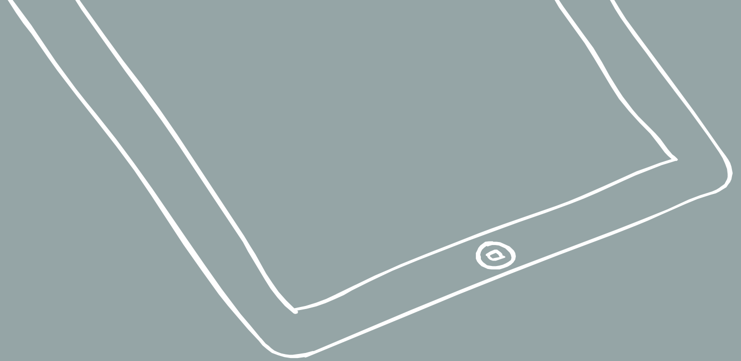
- The best RMSE for degree = 3
- Mean error = 1.14

Conclusion

Method Used	RMSE obtained
Lasso	5.28035644422
Ordinary Least Squares	1.37221573285
Polynomial Regression	1.14738827266



Polynomial Regression is the best one with degree = 3!

Questions?