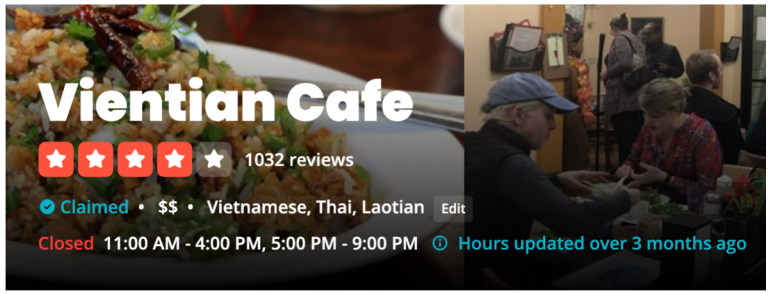
Module 2 Project: Yelp Data Regression

Kenneth Hua

Using Comments to improve Yelp Reviews

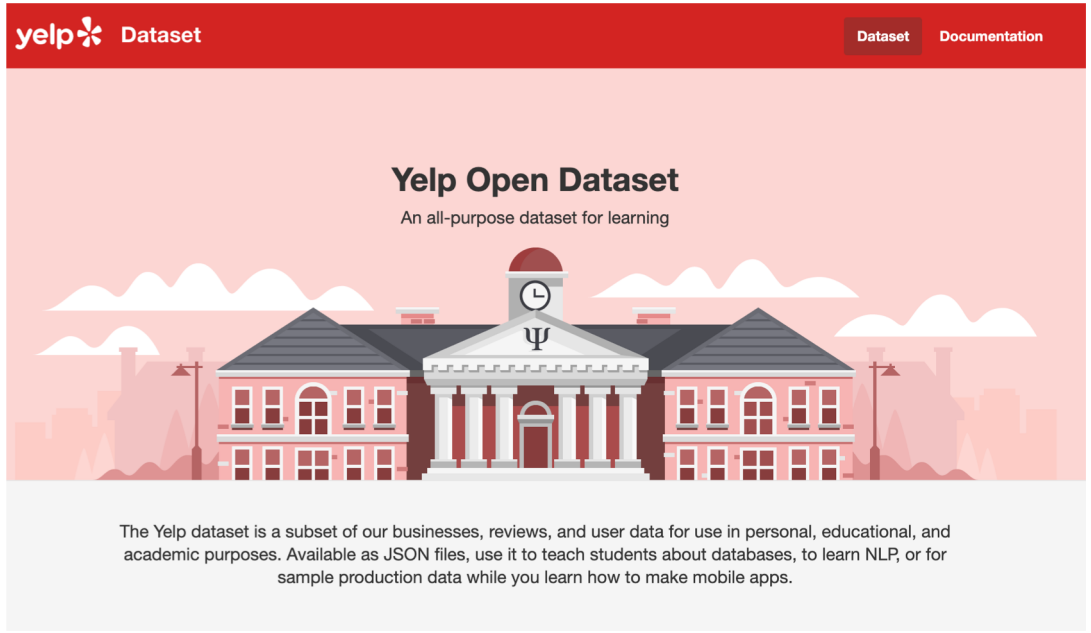
In modern day, people rely on Yelp to choose which places they want to go eat. The Star Rating (1-5) is a key metric users look at to decide if a restaurant is worth visiting or not. **In this study we investigate what factors influence the star rating that users give to restaurant businesses.** Based on this study, a business should know what changes they need to make to their business to improve their rating, and thus attract more customers.

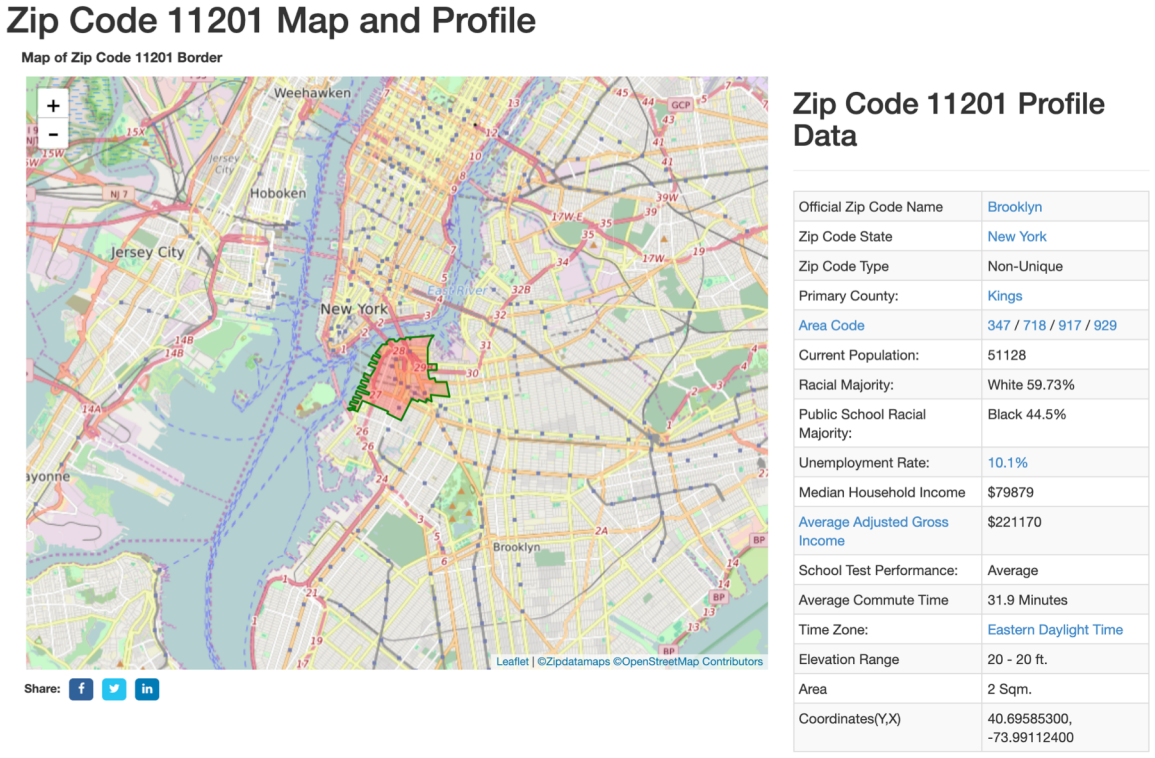


Data Source + Process

Steps:

- Data Sourcing (Yelp API, Zip code Web Scraping - Beautiful Soup, Selenium)
- Data Cleaning/Sorting(Pandas)
- Statsmodels (Regression Exploratory Analysis)
- Scikitlearn (Model/Feature Engineering)
- Nktl, Vader (NLP)
- Visualization (Seaborn)
- Analysis + Conclusions





Initial Regression Datatable Structure

X Feature:
Features related to restaurant listing

X Feature:
Features specific to zipcode of restaurant

X Feature:
Sentiment Analysis of sentence containing relevant phrase:
Sanitation: Sentiment of sentence containing: 'clean','dirty','sanitary', 'unclean','spotless'

Y Target:
Star Rating (1-5)

name	review_count	is_open	Population	Median Household Income	Avg Commute Time	Managemer..	Atmosphere	Sanitation	Speed	Service	Food	High End	review_stars
Baja Betty's Burritos	102	0	20679	102000	29.4	0.4429	0.404817	0.28732	0.104061	0.194045	0.346164	0.210442	3.16505
Bajio Mexican Grill	167	1	26010	101979	25	0.273802	0.509624	0.560557	0.24211	0.27378	0.428167	0.387133	3.63793
Bakery on the Common	154	0	32786	94395	29.8	-0.178198	0.426611	0.009975	0.075299	-0.156812	0.288215	0.121913	2.25749
Bales Cedar Mill Market Place	65	0	58217	96211	23.1	0.219034	0.419525	0.47089	0.140644	0.0581738	0.432144	0.414326	3.40299
Bali Hai	66	0	12108	108609	27.7	0.349257	0.248325	0.29865	0.317652	0.318301	0.466729	0.38825	3.18841
Bamboo	244	1	13221	122221	25.6	0.0597694	0.343127	0.527598	0.173907	0.300233	0.303027	0.338153	3.28854
Bamboo Asian Grille	68	0	26814	101003	23.9	0.115887	0.651983	0.589889	0.224314	0.431782	0.302108	0.2351	3.21429
Bamboo Sushi	88	1	18905	105969	22.9	0.23751	0.496036	-0.0772667	0.182911	0.345148	0.455709	0.290329	3.61538
Bangkok Blue Thai Restaurant	117	0	20628	92954	24	0.228177	0.412668	0.542778	0.236212	0.349829	0.304475	0.273921	3.03419
Bangkok Spice Thai Restaurant	70	1	23956	103983	29.5	0.16449	0.449	-0.02168	0.266713	0.393231	0.382598	0.344846	3.94286

- Which features have the greatest effect on our target variable?
- What is the best model to build with these features to predict restaurant review rating?

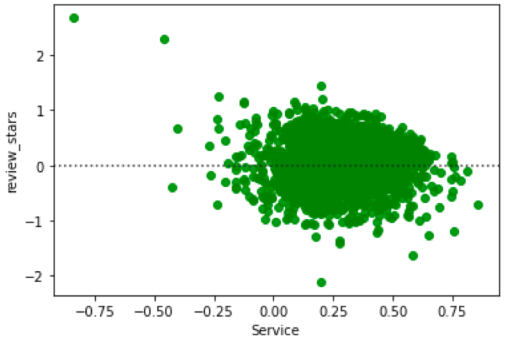
High Income Zip Codes

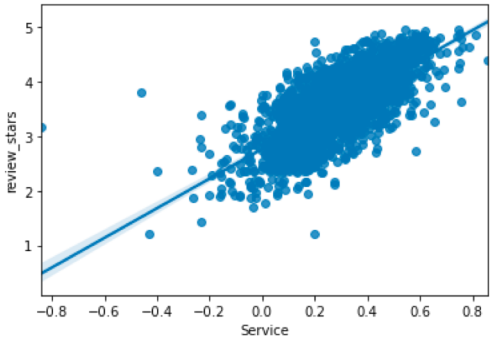
OLS Regression Results

Dep. Variable:	review_stars	R-squared:	0.705
Model:	OLS	Adj. R-squared:	0.704
Method:	Least Squares	F-statistic:	1000.
Date:	Sun, 12 Sep 2021	Prob (F-statistic):	0.00
Time:	18:48:09	Log-Likelihood:	-681.02
No. Observations:	2941	AIC:	1378.
Df Residuals:	2933	BIC:	1426.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.0123	0.024	82.818	0.000	1.965	2.060
is_open	0.1473	0.012	12.181	0.000	0.124	0.171
Management	0.3223	0.030	10.842	0.000	0.264	0.381
Sanitation	0.1477	0.020	7.230	0.000	0.108	0.188
Speed	0.7282	0.072	10.158	0.000	0.588	0.869
Service	1.3757	0.052	26.672	0.000	1.275	1.477
Food	1.2267	0.052	23.694	0.000	1.125	1.328
High End	0.7604	0.052	14.688	0.000	0.659	0.862

Omnibus:	253.160	Durbin-Watson:	1.970
Prob(Omnibus):	0.000	Jarque-Bera (JB):	881.695
Skew:	0.395	Prob(JB):	3.49e-192
Kurtosis:	5.564	Cond. No.	19.4





- After exploring all the features, the most impactful features were being open (is_open) and sentiment from the categories: Management, Sanitation, Speed, Service, Food and High End
- From the sentiment analysis features, Service and Food were most important.
- Residual check of 'Service' looks okay, the correlation between this and the review passes the visual test
- **In order for a restaurant in a high income zip code to improve its ratings, it needs to have positive performance in Service and Food.**

	Linear Model	**Ridge Model**	Lasso Model
Validation R^2	0.703774	**0.704329**	0.675897
Alpha (Hyperparameter)	NA	**0.74016**	0.01

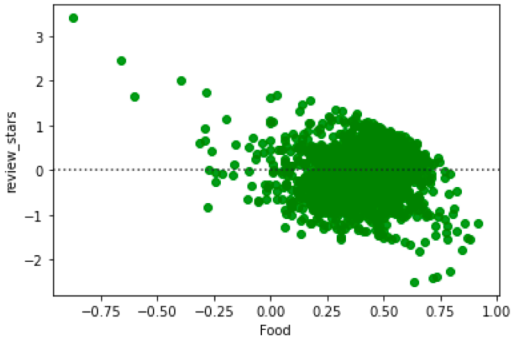
Low Income Zip Codes

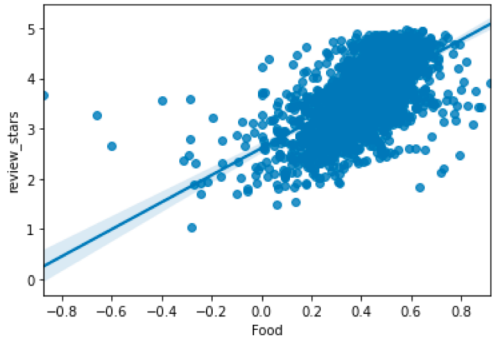
OLS Regression Results

Dep. Variable:	review_stars	R-squared:	0.696
Model:	OLS	Adj. R-squared:	0.696
Method:	Least Squares	F-statistic:	1023.
Date:	Sun, 12 Sep 2021	Prob (F-statistic):	0.00
Time:	20:40:27	Log-Likelihood:	-993.46
No. Observations:	3578	AIC:	2005.
Df Residuals:	3569	BIC:	2061.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.0127	0.025	81.823	0.000	1.965	2.061
is_open	0.0976	0.011	8.686	0.000	0.076	0.120
Management	0.3662	0.029	12.597	0.000	0.309	0.423
Atmosphere	0.0731	0.027	2.664	0.008	0.019	0.127
Sanitation	0.1394	0.020	7.010	0.000	0.100	0.178
Speed	0.9422	0.068	13.846	0.000	0.809	1.076
Service	1.3023	0.047	27.573	0.000	1.210	1.395
Food	1.1385	0.048	23.666	0.000	1.044	1.233
High End	0.9555	0.050	19.079	0.000	0.857	1.054

Omnibus:	179.694	Durbin-Watson:	1.948
Prob(Omnibus):	0.000	Jarque-Bera (JB):	658.081
Skew:	0.053	Prob(JB):	1.26e-143
Kurtosis:	5.098	Cond. No.	20.3





- The same features from High Income were present in Low Income, with the addition of ‘Atmosphere’ sentiment
- From the sentiment analysis features, Service and Food were also most important. But Speed and High End were also stronger predictors in low income neighborhoods.
- Residual check of ‘Service’ looks a little non random but mostly okay, the correlation between this and the review passes the visual test
- **In order for a restaurant in a high income zip code to improve its ratings, it needs to have positive performance in Service and Food, but also keep in mind Speed of service and a sense of high end/luxury for its customers**

	Linear Model	Ridge Model	Lasso Model
Validation R^2	**0.684503**	0.684455	0.629273
Alpha (Hyperparameter)	**NA**	0.4659	0.01

Summary

- Sentiment analysis of text, and opening time are key features that impact the review rating of a restaurant (1-5 stars)
- In order to improve the rating, the two most important factors of a restaurant to work on are Service and Food
- Low Income neighborhoods need to also think a bit more about speed of service, and how to tailor a nice high-end experience to customers, as well as how to generate a good atmosphere.

For more info: https://github.com/kenhua15/Metis-Projects