Visualizing the Yelp Dataset

Ashutosh Verma, Karthik Balasubramanian, Md Khaled Hassan, Suneil Shrivatsav and Tianji Zhou
Department of Computer Science, Georgia Institute of Technology

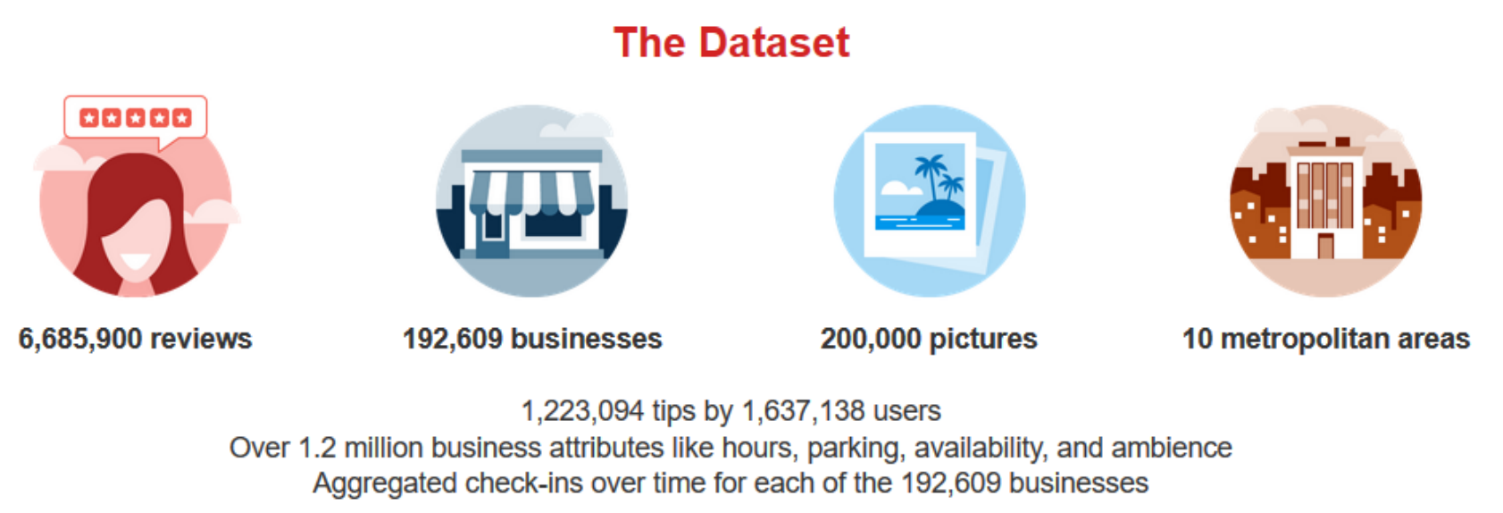
I. Introduction

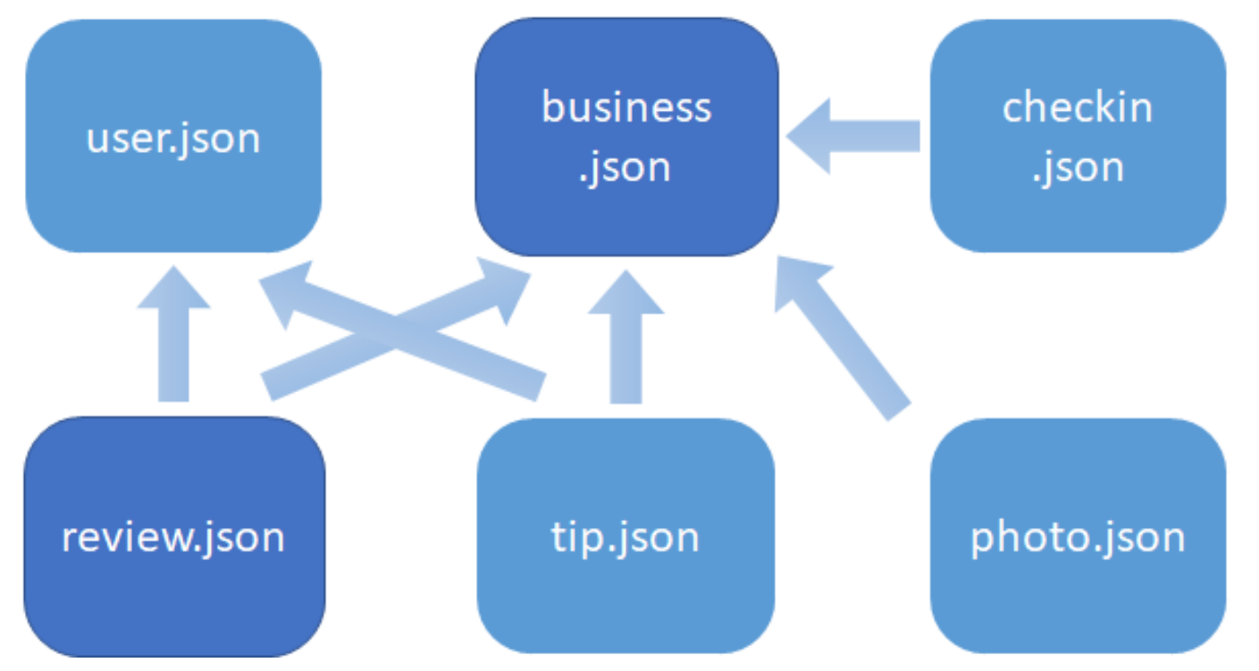
Problem Statement:
The aim of this project is to use very large dataset from Yelp(~8.6GB) and provide meaningful visualizations/dashboard to business analysts, restaurant owners & economic policy makers. The dataset contains a multitude of indicators like user reviews, ratings, co-ordinates, review counts, categories etc. which we intend to correlate and produce such visualizations. We also intend to use the poverty/population dataset per county provided in homework-2 for such visualizations.

Motivations:
Our study will provide detailed insights for restaurant owners, customers, suppliers, economists and business analysts.

II. Dataset

The Yelp Open Dataset (https://www.yelp.com/dataset) is a subset of Yelp's business, reviews and users data. It includes 192609 business in 10 metropolitan areas. The dataset contains a 8GB core data set and a supplementary data set for user uploaded images.


The Dataset
6,685,900 reviews
192,609 businesses
200,000 pictures
10 metropolitan areas
1,223,094 tips by 1,637,138 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 192,609 businesses



The core data set is a combination of six different json files: business, review, user, tip, check-in and photo. Each item in the business, user and review data files is given an ID, by which these data files can be connected.

To investigate the relation between Yelp dataset business with county economy and population, we will also explore two other data file from the United States Department of Agriculture (https://www.ers.usda.gov/), county_poverty.csv and county_detail.csv.

III. Approaches

The proposed methodology broadly consists of three parts:
(1). **Data cleaning and refinement**: Data present in Business.json was not directly usable in Tableau as it required some Data cleanup and Data refining. We use Python and Google OpenRefine to cleanup and filter information.

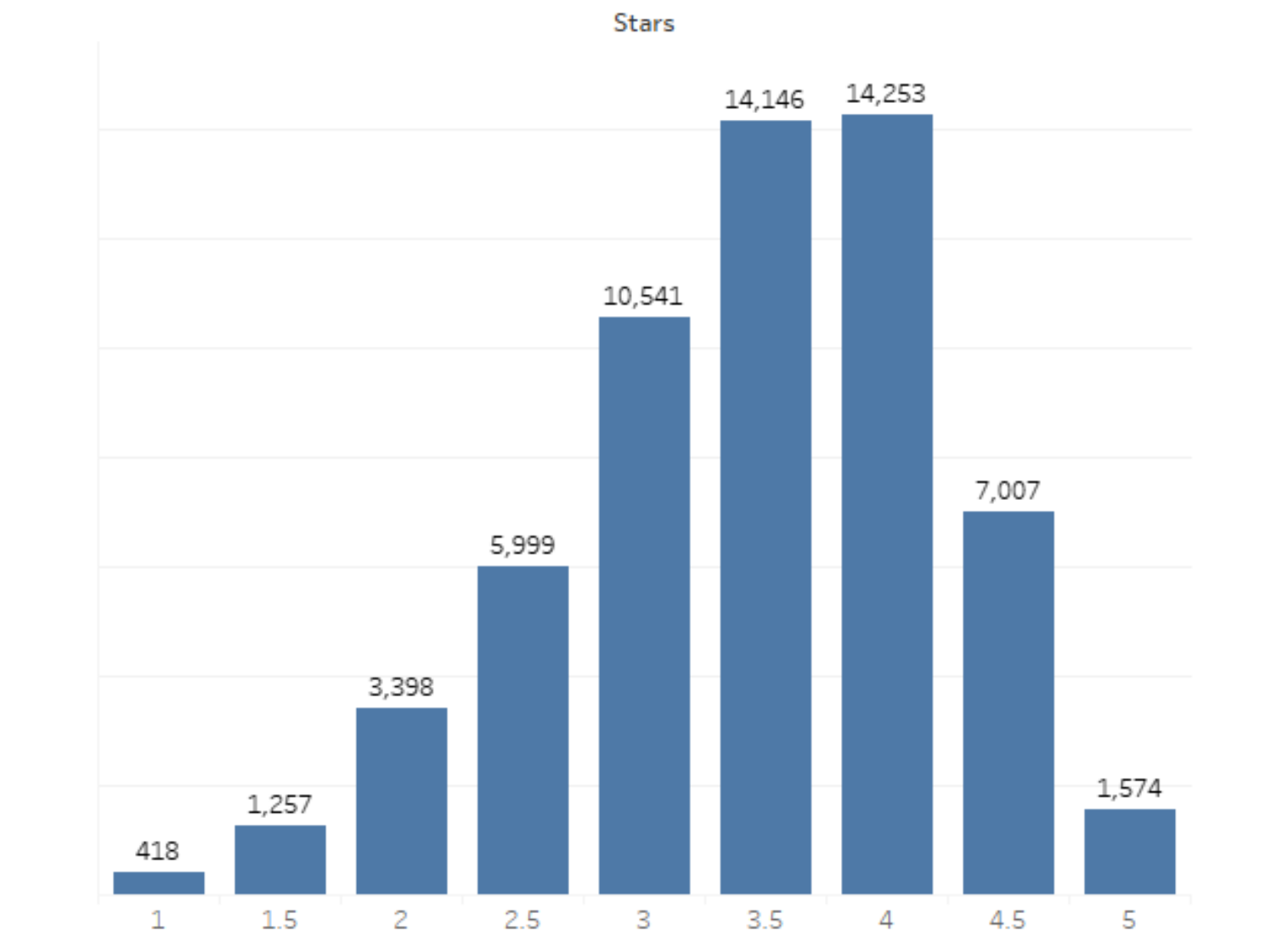
(2). **Data visualization and analysis**: The current data visualization is looking at the business data table. We use Tableau too visualize these on a map to allow more intuitive understanding of food diversity data and its intersection with poverty.

(3). **Machine learning and text mining**: The machine learning and text mining analysis are done using Python and scikit-learn package, for the purpose of predicting ratings or sentiment analysis based on the given attributes.

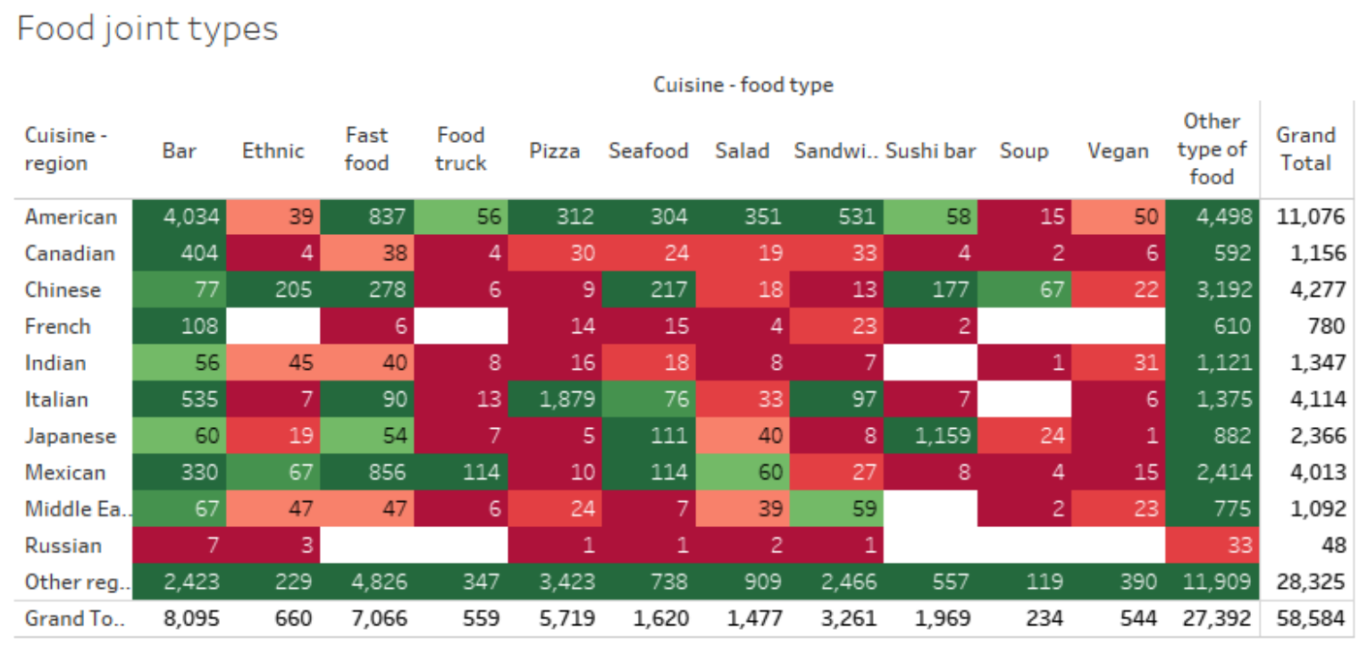
List of Innovations:
(1). Exploring Yelp's data through a lens of food diversity and poverty.
(2). Geographical data blend.

IV. Business Data Analysis

We first look at the distribution of rating: how many Business ids receiving a particular star rating. The plot shows that most of the ratings are centered around 3-4.5 stars on a scale of 5. However, the number of ratings of 1-2 stars and 5 stars are low indicating a skewed dataset. The dataset shows an average rating of 3.44.



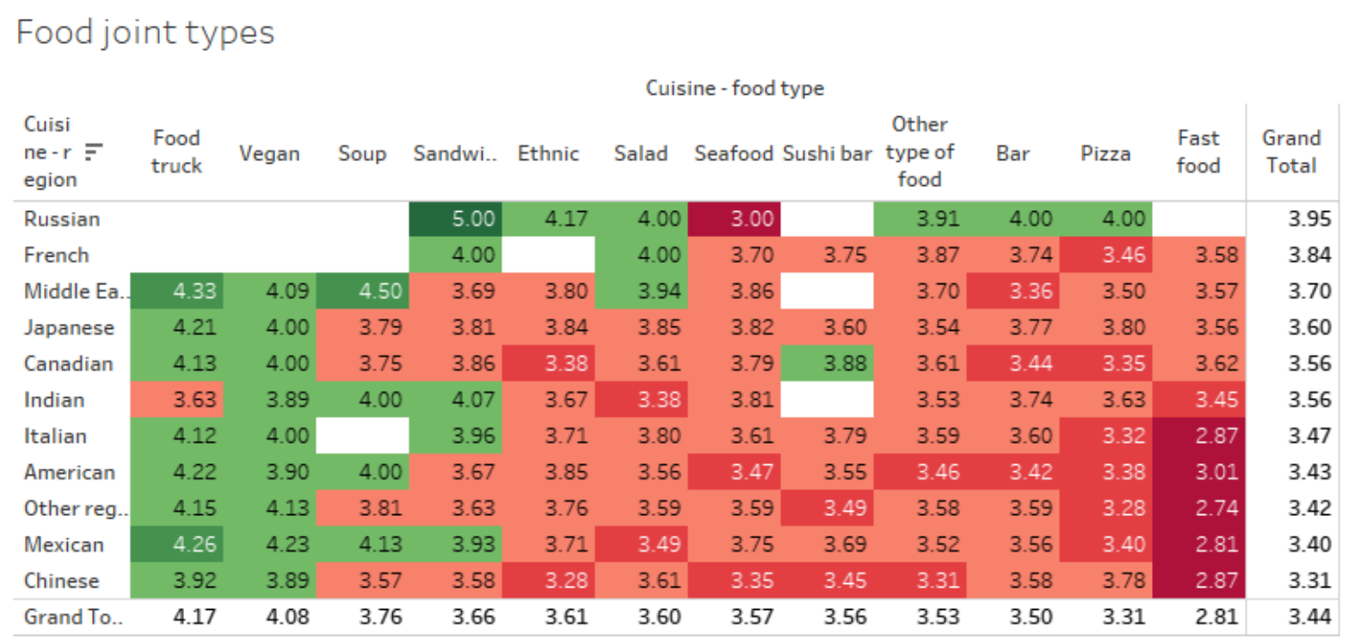
We then look at the restaurant types, especially the cuisine type and region, and plot them vs. the restaurant type counts and star ratings in heat maps.



Food joint types

Cuisine - region	Bar	Ethnic	Fast food	Food truck	Pizza	Seafood	Salad	Sandwi..	Sushi bar	Soup	Vegan	Other type of food	Grand Total
American	4,034	39	837	56	312	304	351	531	58	15	50	4,498	11,076
Canadian	404	4	38	4	30	24	19	33	4	2	6	592	1,156
Chinese	77	205	278	6	9	217	18	13	177	67	22	3,192	4,277
French	108		6		14	15	4	23	2			610	780
Indian	56	45	40	8	16	18	8	7		1	31	1,121	1,347
Italian	535	7	90	13	1,879	76	33	97	7		6	1,375	4,114
Japanese	60	19	54	7	5	111	40	8	1,159	24	1	882	2,366
Mexican	330	67	856	114	10	114	60	27	8	4	15	2,414	4,013
Middle Ea..	67	47	47	6	24	7	39	59		2	23	775	1,092
Russian	7	3			1	1	2	1				33	48
Other reg..	2,423	229	4,826	347	3,423	738	909	2,466	557	119	390	11,909	28,325
Grand To..	8,095	660	7,066	559	5,719	1,620	1,477	3,261	1,969	234	544	27,392	58,584

On the right we plot the cuisine region and type vs counts. The most common cuisines are American Bar, Italian pizza and Japanese Sushi.

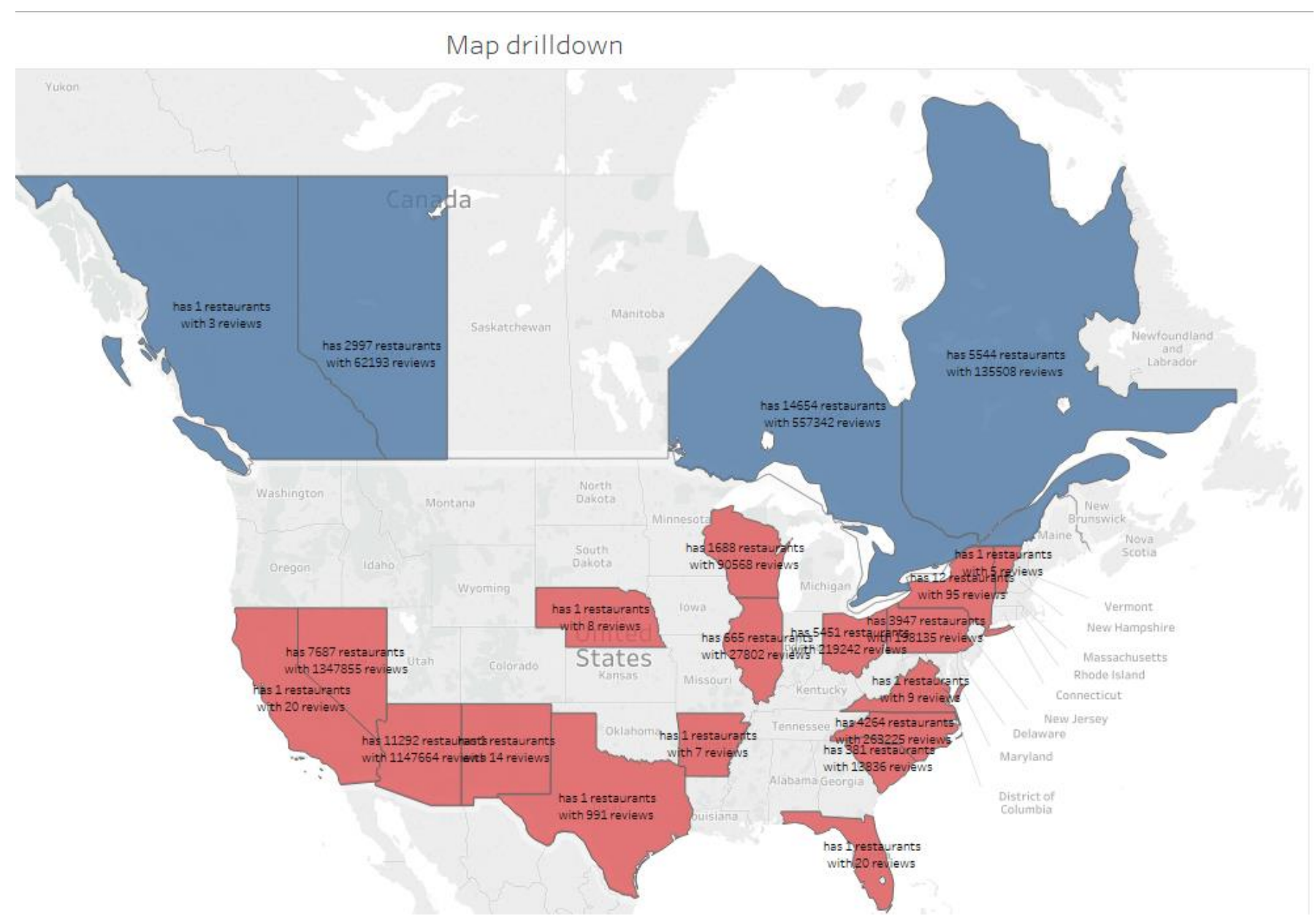


Food joint types

Cuisine - region	Food truck	Vegan	Soup	Sandwi..	Ethnic	Salad	Seafood	Sushi bar	Other type of food	Bar	Pizza	Fast food	Grand Total
Russian				5.00	4.17	4.00	3.00		3.91	4.00	4.00		3.95
French				4.00		4.00	3.70	3.75	3.87	3.74	3.46	3.58	3.84
Middle Ea..	4.33	4.09	4.50	3.69	3.80	3.94	3.86		3.70	3.36	3.50	3.57	3.70
Japanese	4.21	4.00	3.79	3.81	3.84	3.85	3.82	3.60	3.54	3.77	3.80	3.56	3.60
Canadian	4.13	4.00	3.75	3.86	3.38	3.61	3.79	3.88	3.61	3.44	3.35	3.62	3.56
Indian	3.63	3.89	4.00	4.07	3.67	3.38	3.81		3.53	3.74	3.63	3.45	3.56
Italian	4.12	4.00		3.96	3.71	3.80	3.61	3.79	3.59	3.60	3.32	2.87	3.47
American	4.22	3.90	4.00	3.67	3.85	3.56	3.47	3.55	3.46	3.42	3.38	3.01	3.43
Other reg..	4.15	4.13	3.81	3.63	3.76	3.59	3.59	3.49	3.58	3.59	3.28	2.74	3.42
Mexican	4.26	4.23	4.13	3.93	3.71	3.49	3.75	3.69	3.52	3.56	3.40	2.81	3.40
Chinese	3.92	3.89	3.57	3.58	3.28	3.61	3.35	3.45	3.31	3.58	3.78	2.87	3.31
Grand To..	4.17	4.08	3.76	3.66	3.61	3.60	3.57	3.56	3.53	3.50	3.31	2.81	3.44

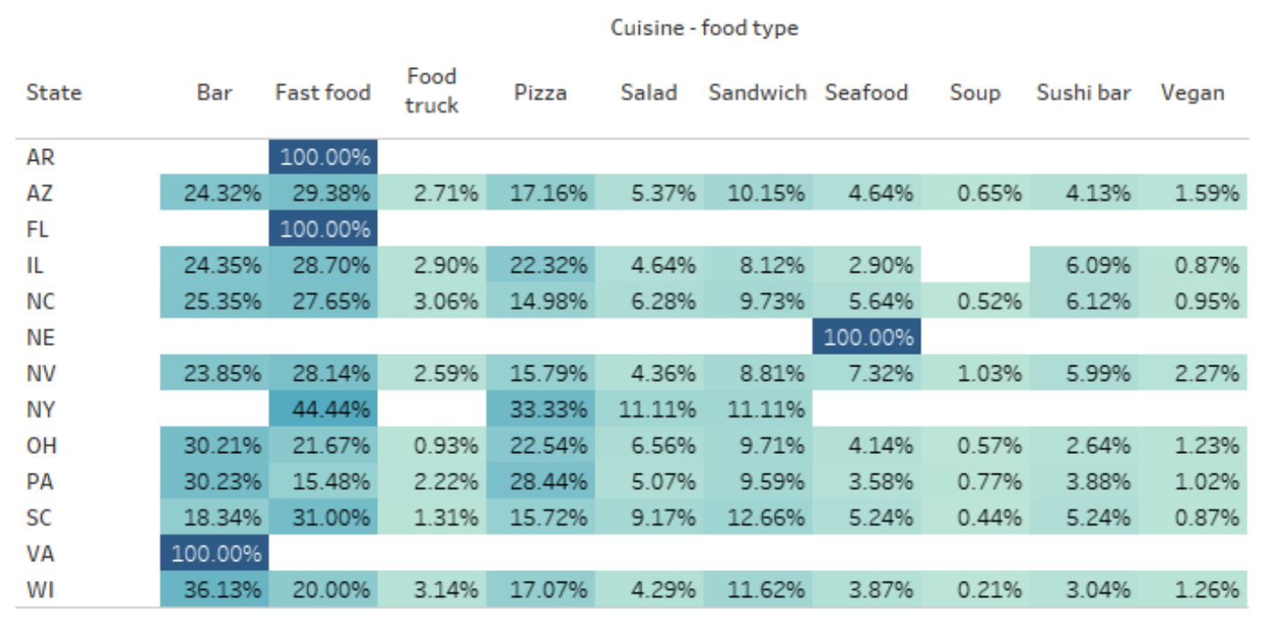
On the left we plot the cuisine region and type vs. star ratings. Food trucks and vegan food are highest rated while fast food and pizza are the worst rated.

V. Geographical Data and County Economy

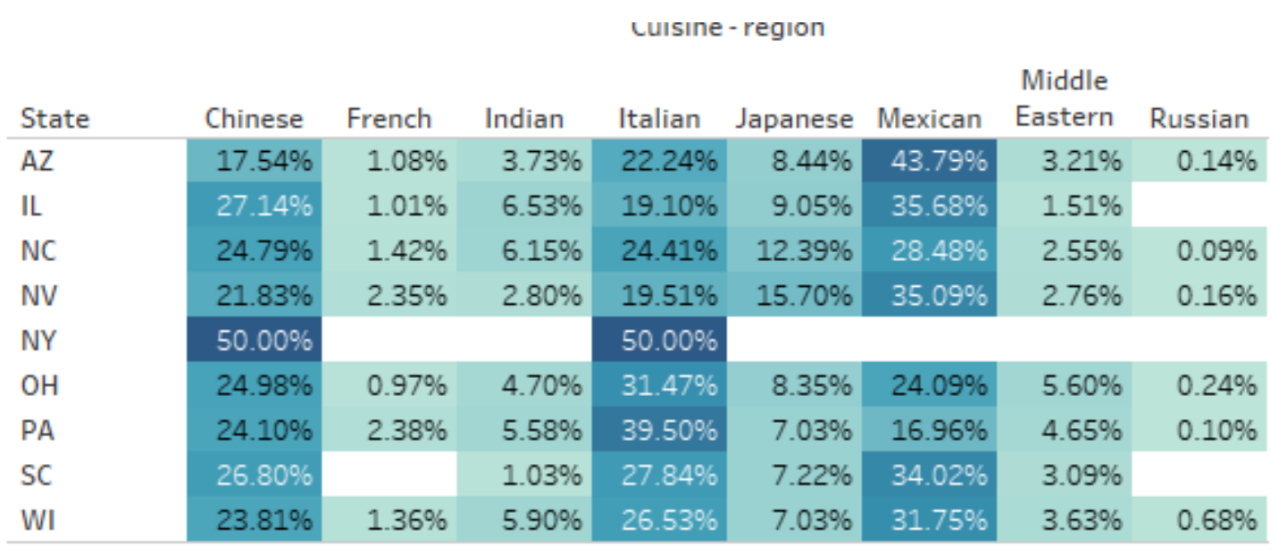


On the left is a choropleth map showing the number of restaurants in the different states which gives us an idea of the income levels which can be correlated to number of restaurants across the states.

The following two tables show type of restaurants in each state where data is available. Arizona, Illinois, Ohio and Pennsylvania all have a robust combination of food by region and by cuisine type. On the other hand, New York is the state with the least food diversity in terms of region of origin of the cuisine.

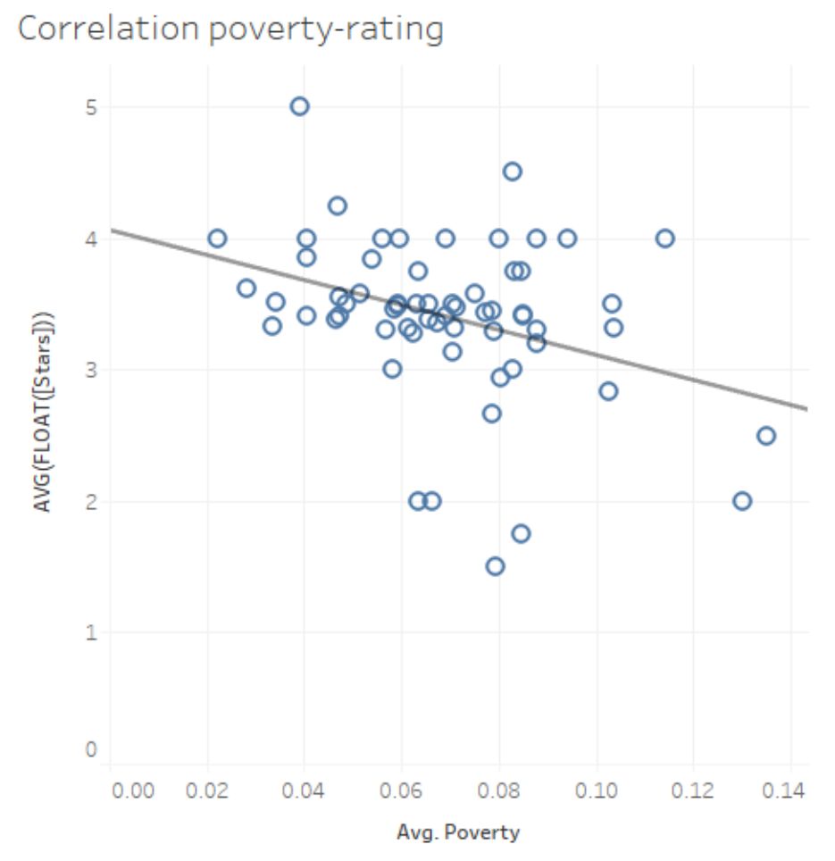


State	Bar	Fast food	Food truck	Pizza	Salad	Sandwich	Seafood	Soup	Sushi bar	Vegan
AR		100.00%								
AZ	24.32%	29.38%	2.71%	17.16%	5.37%	10.15%	4.64%	0.65%	4.13%	1.59%
FL		100.00%								
IL	24.35%	28.70%	2.90%	22.32%	4.64%	8.12%	2.90%		6.09%	0.87%
NC	25.35%	27.65%	3.06%	14.98%	6.28%	9.73%	5.64%	0.52%	6.12%	0.95%
NE							100.00%			
NV	23.85%	28.14%	2.59%	15.79%	4.36%	8.81%	7.32%	1.03%	5.99%	2.27%
NY		44.44%		33.33%	11.11%	11.11%				
OH	30.21%	21.67%	0.93%	22.54%	6.56%	9.71%	4.14%	0.57%	2.64%	1.23%
PA	30.23%	15.48%	2.22%	28.44%	5.07%	9.59%	3.58%	0.77%	3.88%	1.02%
SC	18.34%	31.00%	1.31%	15.72%	9.17%	12.66%	5.24%	0.44%	5.24%	0.87%
VA	100.00%									
WI	36.13%	20.00%	3.14%	17.07%	4.29%	11.62%	3.87%	0.21%	3.04%	1.26%



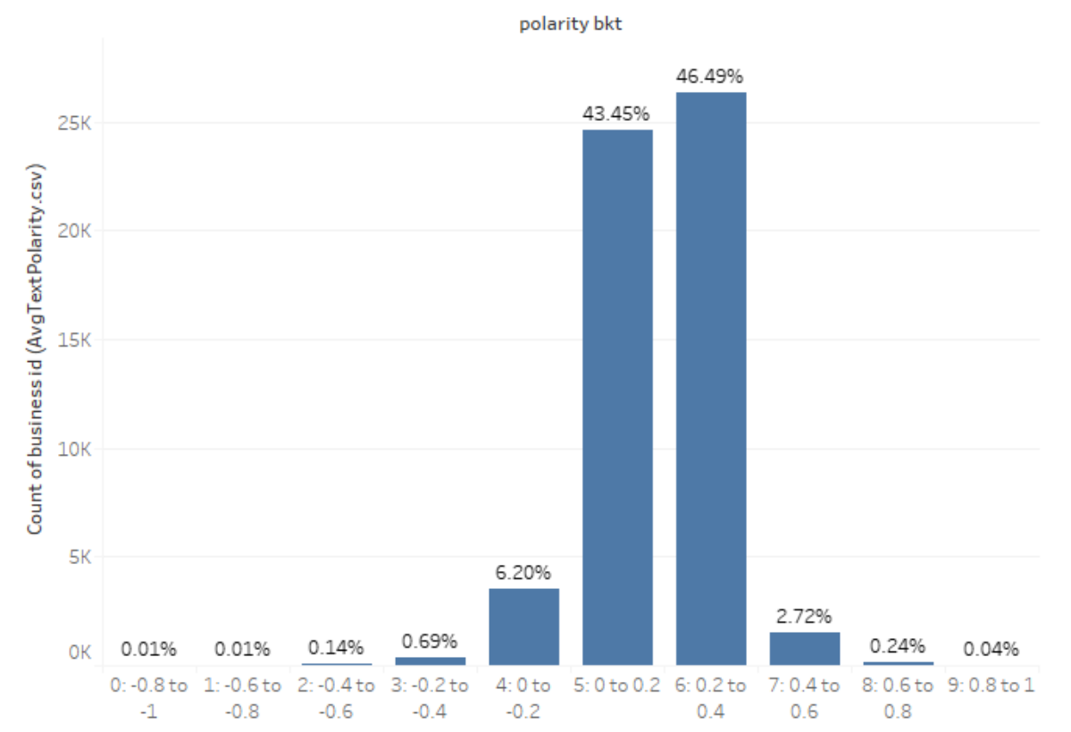
State	Chinese	French	Indian	Italian	Japanese	Mexican	Middle Eastern	Russian
AZ	17.54%	1.08%	3.73%	22.24%	8.44%	43.79%	3.21%	0.14%
IL	27.14%	1.01%	6.53%	19.10%	9.05%	35.68%	1.51%	
NC	24.79%	1.42%	6.15%	24.41%	12.39%	28.48%	2.55%	0.09%
NV	21.83%	2.35%	2.80%	19.51%	15.70%	35.09%	2.76%	0.16%
NY	50.00%			50.00%				
OH	24.98%	0.97%	4.70%	31.47%	8.35%	24.09%	5.60%	0.24%
PA	24.10%	2.38%	5.58%	39.50%	7.03%	16.96%	4.65%	0.10%
SC	26.80%		1.03%	27.84%	7.22%	34.02%	3.09%	
WI	23.81%	1.36%	5.90%	26.53%	7.03%	31.75%	3.63%	0.68%

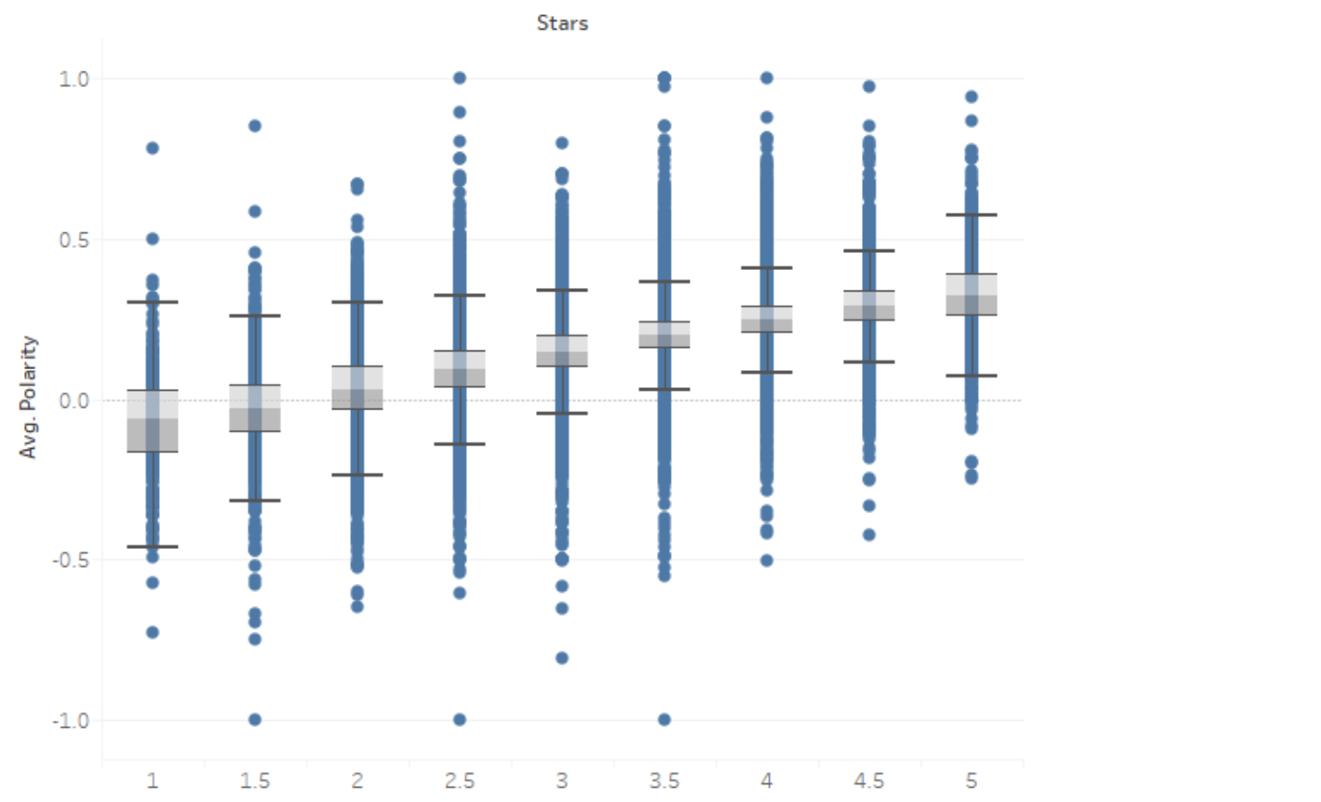
The plot on the right shows a statistical analysis of the dataset fitting the poverty level and average rating of the restaurants. The plot indicates that with increasing poverty, the quality of the restaurants goes down.



VI. Text Mining: Review Text Sentiment Polarity

We carried out a sentiment analysis on more than 1.8 million reviews using TextBlob library of python. We extracted the polarity of each of the reviews in a range of [-1.0 1.0], where, -1 being an extremely negative and +1 being an extremely positive review.

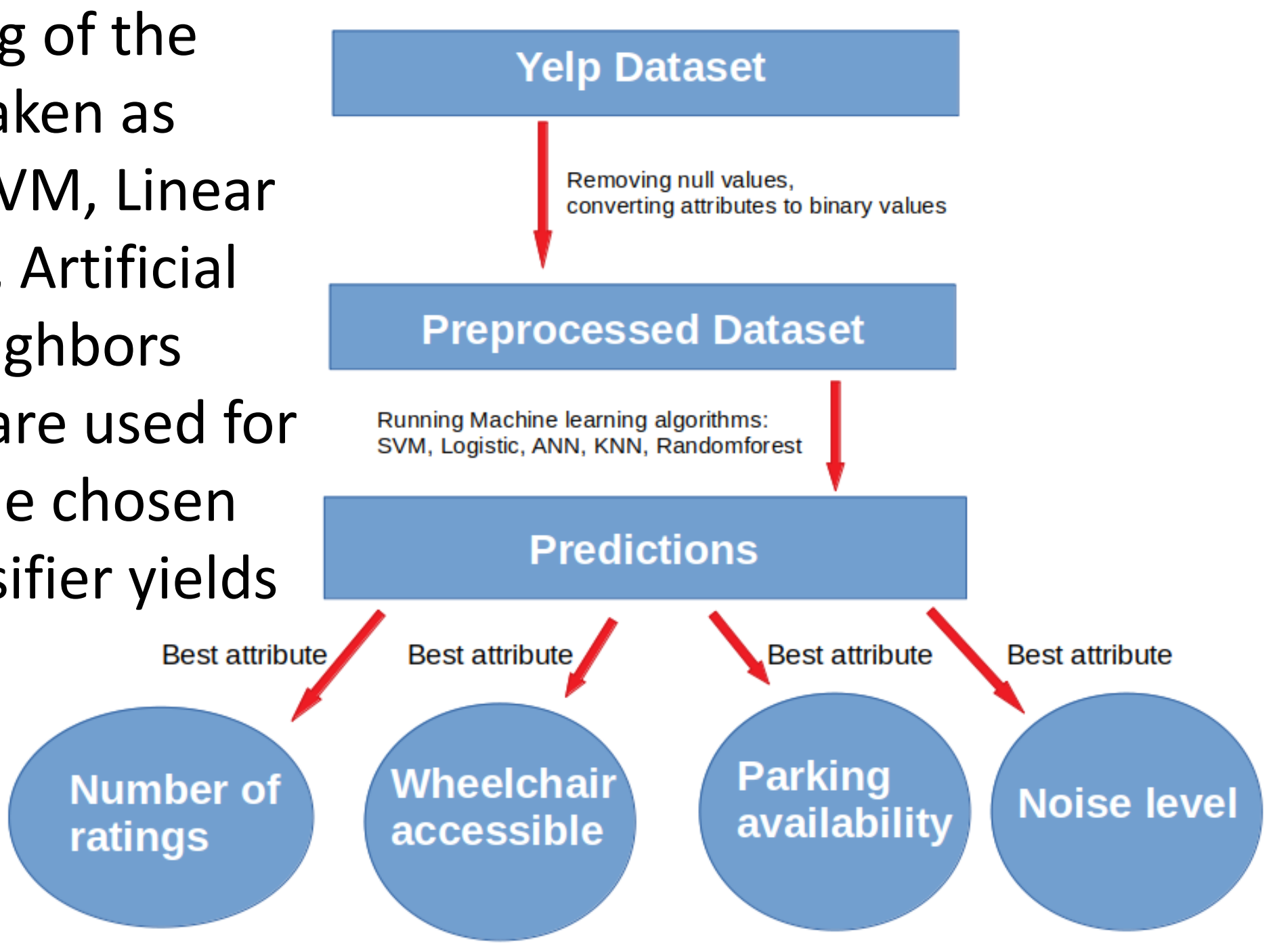




On the left is the distribution of the polarity of the reviews given to restaurants. Most reviews are in the range of [0,0.4]. On the right we explore the correlation between polarity and star ratings. It appears the higher star ratings have slightly higher polarity.

VII. Machine Learning: Best Features Extraction

The business dataset consisting of the entries for the restaurants is taken as the dataset for explorations. SVM, Linear regression, Logistic regression, Artificial neural networks, k-nearest neighbors and RandomForest classifiers are used for predicting the ratings. From the chosen classifiers, RandomForest classifier yields the best results.



The best features which contribute most to the ratings include "Number of ratings", "Wheelchair accessible", "Availability of parking" and "Noise level".

VIII. Conclusions

The key points that we could analyze from these visualizations are:
(1). High poverty areas have lower rating restaurant.
(2). A few businesses are predominant in a specific area.
(3). Food Trucks and Vegan Food are the highest rated ones while Fast Food and Pizza places are the worst rated.
(4). The sentiment polarity is well correlated with average review rating of the restaurant.
Based on all the above data points a business analyst or an investor could make a decision on which sector he should invest in and where for a successful business .

References

[1]. Reviews, Reputation, and Revenue: The Case of Yelp.com, L. Michael, Harvard Business School NOM Unit Working Paper No. 12-016 (2016).
[2] Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud, Luca, Michael & Zervas, Georgios. (2013). SSRN Electronic Journal. 10.2139/ssrn.2293164.
[3]. Improving Restaurants by Extracting Subtopics from Yelp Reviews, Huang, J.; Rogers, S.; Joo, E. Improving Restaurants by Extracting Subtopics from Yelp Reviews. In iConference Social Media Expo (2014).
[4]. Restaurant Setup Business Analysis Using Yelp Dataset , Sindhu B Hegde, Supriya Satyappanavar, Shankar Setty, Advances in Computing Communications and Informatics (ICACCI) 2018 International Conference on, pp. 1455-1462, (2018).
[5]. The Geography of Taste: Using Yelp to Study Urban Culture, Rahimi, S.; Mottahedi, S.; Liu, X. ISPRS Int. J. Geo-Inf. 7, 376(2018).
[6]. Spatial analysis of users-generated ratings of yelp venues, Sun, Y. & Paule, J.D.G. Open geospatial data, softw. stand. https://doi.org/10.1186/s40965-017-0020-9 (2017).
[7]. Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity E. L. Glaeser, H. Kim, M. Luca, National Bureau of Economic Research,Working Paper No. 24010 (2017).
[8]. Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews, Yu B., Zhou J., Zhang Y., Cao Y., arXiv e-prints, arXiv:1709.08698 (2017).
[9-15]. Learning from data. Chapters 1-9 Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. New York, NY, USA:: AMLBook, 2012.
[16]. Restaurants Review Star Prediction for Yelp Dataset, M. Yu, M. Xue, W. Ouyang, pdfs.semanticscholar.org (2015).