BIG DATA ANALYTICS GROUP PROJECT

NEW YORK CITY HYGIENE DATA ANALYSIS

December 19, 2016

Munaf Arshad Qazi
Yingnan Li
Rishabh Jain
Pratik Shirish Kamath

Contents

1 OBJECTIVE

As we all know, good sanitary condition is very important for restaurants, and everyday people search information of NYC restaurants from yelp, which does not contain significant hygiene information. So the objective of this project is to map New York eateries and bars to their expected hygiene rating by the New York City Department of Health. Next we intend to determine patterns in eatery properties which lead to certain ratings and subsequently be able to train a recommendation engine to show the distribution of high rated eateries based in New York City in next two years.

2 APPROACH

In a bird's eye view, to achieve this goal, we got the NYC restaurants inspection results dataset from Department of Health and mental Hygiene to indicate the sanitary condition about each restaurant. Next, we got the NYC restaurants detail information from yelp API. We imported all the dataset to mongodb and did data cleaning, aggregation and map reduce by mongo scripts or pymongo. This condensed our datasets into one which comprised of detailed information of each restaurant and the significant hygiene score. This was used to visualize data along with using machine learning algorithms to predict future ratings and hygiene scores.

We used mongodb for storage because of following reasons

- Scheme less, mongodb is a document database in which one collection holds different documents. Number of fields, content and size of the document can differ from one document to another.
- Index on any attribute, mongodb gives every document a index which makes query faster.
- Rich functions, mongodb has a lot of functions to do aggregation, CRUD operations, text searcing or map reduce. It makes data processing easier to do.

There are also many reasons to choose geoplotlib and R to do visualizations. They are open source, and have plenty of useful packages and libraries.

3 SYSTEMS ARCHITECTURE

In this project, we used mongodb as central database to store datasets. Also to do data cleaning and aggregation with pymongo or javascript. We also use mongodb and javascript for map reduce. The machine learning part, was written in python because of its flexibility. Finally, we used R and goeplotlib to do information virtualization. See figure 1.

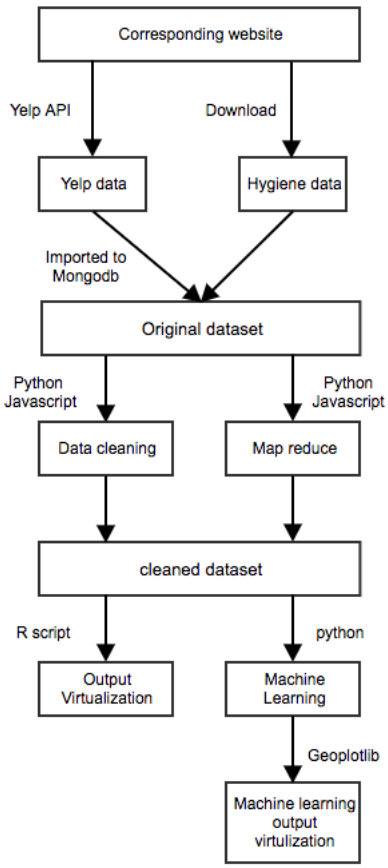

Figure 1: Designed System Architecture

4 DATA ARCHITECTURE

In this project, we used two datasets, the hygiene inspection dataset in New York city and yelp's restaurants information dataset in New York City. The details of the two datasets are shown in Table 1 and 2.

Table 1: Hygiene Data Architecture Table.

Dataset	Field Name	Comment
Hygiene data	CAMIS	Unique identifier for each restaurant
	DBA	Name of each restaurant
	BORO	Borough in which the restaurant is located
	BUILDING	The building number for the entity
	STREET	The street name at which the entity
	ZIPCODE	Zip code as per the address of the entity
	PHONE	Phone Number
	CUISINE DECRIPTION	This field describes the entity cuisine.
	INSPECTION DATE	The date of inspection
	ACTION	The actions associated with each inspection.
	VIOLATION CODE	Violation codes associated with each inspection
	VIOLATION DESCRIPTION	Violation description with each inspection
	CRITICAL FLAG	This indicates if Violation is critical or not
	SCORE	The score with each inspection
	GRADE	The grade with each inspection
	GRADE DATE	The grade date with each inspection
	RECORD DATE	The date when the current grade was issued to the entity
	INSPECTION TYPE	The type of inspection

Table 2: Hygiene Data Architecture Table.

Dataset	Field Name	Comment
Yelp data	Is_claimed	Indicate entity is claimed or not
	Is_closed	Indicate entity is closed or not
	Name	Name of each restaurant
	City	The city each restaurant located
	State	The state for the entity
	Address	The address for each entity
	Zip code	Zip code as per the address of the entity
	Phone	Phone Number

Continued on next page

Table 2 – continued from previous page

Dataset	Field Name	Comment
Yelp data	Display_phone	Phone number in display format
	Latitude	Latitude of restaurant location
	Longitude	Longitude of restaurant location
	Category_alias	Alias of category
	Category_name	Name of category
	Review_count	The review count of each restaurant
	Rating	The rating score for each restaurant

5 INTEGRATION ARCHITECTURE

In this part we will introduce the architecture of the result collection. This collection contains each inspection score, average score and date for this inspection and the business information for this restaurant. We get the result collection by joining the yelp dataset and hygiene dataset. Then exported it from mongodb and used it to do result virtualization and machine learning. See detail in Table3.

Table 3: Yelp Hygiene final dataset.

Data	Field Name	Comment
Yelp Hygiene Final	Is_claimed	Indicate entity is claimed or not
	City	The city each restaurant located
	Review_count	The review count of each restaurant
	Name	Name of each restaurant
	State	The state for the entity
	Address	The address for each entity
	Zip code	Zip code as per the address of the entity
	Phone	Phone Number
	Latitude	Latitude of restaurant location
	Longitude	Longitude of restaurant location
	Category_alias	Alias of category

Continued on next page

Table 3 – continued from previous page

Data	Field Name	Comment
	hygiene_average_score	Average hygiene score for this entity
	Category_name	Name of category
	Category_alias	Alias of category
	Time	The date of the inspection
	Score	The hygiene score for each restaurant
	Rating	The rating score for each restaurant

6 DESIGN

Data Cleaning

This part will introduce how our project do data cleaning. At first we download the hygiene dataset from NYC open data website and imported it into mongodb. Next, we changed the phone number format of hygiene dataset, so that we can use them to get yelp data. The original phone number type is just integer, so ewe converted it to string and add '+1' before every phone numbers. Then we removed documents which score are empty. After that we used yelp API and distinct phone number from hygiene data to grab yelp restaurant data and save them into a json file. After that, we import the json file to mongodb. Next, we did map reduce on hygiene collection that calculate the average hygiene score of each restaurant. The average hygiene score will be uses in analysis and machine learning part. After map reduce, we got the new hygiene data with average hygiene score, then we join this collection with yelp dataset to get the final result collection, which contains every inspection and restaurant information and remove duplicated records. Finally we exported this collection into a csv file and do virtualization and machine learning on that. See Table 3 to get the detail of result collection.

Machine Learning

For the purpose of Machine Learning we utilize the fact that we have mostly quantities which are either of float or int type. We first come across the problem of having character/string based quantities which needs to be used in the regression algorithms to be able to use the incredible details that we are expecting the data set to contain.

To solve this problem we use a method called **Sparse encoding**. This is a Label encoding pre-processing mechanism in which the categorical data is encoded in a way where each of the categories of a feature is used as a Boolean feature leading to a sparse matrix of each categorical data feature. Using Label encoding we achieve a fully quantifiable feature vector for the data-set which can be used for regression analysis.

For the purpose of this study we perform 4 kinds of regression analysis:

- **Linear Regression** Linear Regression is the most used statistical method for predictions. This algorithm is used for linear additive analysis between features and their contribution to the outcome. The feature predictor in Linear regression looks like

  $$\hat{Y} = b_0 + b_1 X_{1t} + b_2 X_{2t} + b_3 X_{3t} + ... + b_k X_{kt}$$

  where $\hat{Y}$ is the predicted field, k is the number of features, X are the values for each of the features, and b is the coefficient of the features.

  The machine learning algorithm is designed to learn the values of the coefficients such as to most closely approximate the predicted field.

  Before the beginning of the learning phase, the coefficients are initialized to random values. These are updated according by the following principle (also known as **Gradient Descent Algorithm**):

  $$b = a - \gamma \nabla F(a)$$

  where b is the updated coefficient vector, a is the old coefficient vector, and $F(a)$ is the error, which in our case remains constantly to be the L-2 loss.

- **Ridge Regression** Ridge Regression is a variation of the linear regression algorithm with a an added regularization term which regulates the effect of large updates on the coefficient vector. The update term of the coefficient vector is given by

  $$b = a - \gamma \nabla F(a) - \tau a.a$$

  where the notation is the same as given in the linear regression equation and τ is the regularization factor.

  This helps relieve the algorithm of surges of gradient explosion on occurrence of noise or one off data points which tend to go towards local minimas.

- **Bayesian Ridge Regression** Bayesian Ridge Regression adds a probabilistic aspect to the ridge regression regularization term by adding an identity matrix with a probabilistic occurrence of terms. The probabilistic term is given by

$$p(w|\lambda) = \aleph(w|0, \lambda^{-}1 I_p)$$

  where p is the probabilistic term, λ is the Gaussian distribution, I_p is the identity matrix and w is the coefficient matrix.

- **Lasso Regression** Least Absolute Shrinkage and Selection Operator, also known as Lasso regression is a regression technique with variable selection as well as regularization. This is given by a performance parameter tuning operator

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

  and the loss function is given by the following equation

$$PRSS(\beta)_{l_1} = \sum_{i=1}^{n} (y_i - z_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Data Visualization

Taking results from our predictions of the 4 different ML algorithms, we constructed visualizations in R and using the python geoplotlib library. Geoplot library is a python toolbox for visualizing geographical data and making maps and can be used to plot results according to latitude and longitude. It was used to create interactive visualizations. The created plot can be moved around in zoomed in zoomed out just like for a normal map. Snapshots for the generated plots are present in the outputs folder. R visualizations were created for static analysis.

Geoplot Library The 2 sets of graphs generated by the Geoplot library display results from the ridge, linear, lasso and bayesian ridge regression for both ratings and hygiene score. These are predictions for the next two years of how these measures vary. They are discussed in further detail in the Analysis phase.

Geoplot creates a dataframe from the read csv with one latitude, longitude and value mapped as one point on the plot. A part of the dataframe looks like Table 4.

The colorbrewer selects the color based on the value of the rating or hygiene score. For Weisheng, we have distributed the rgb matrix to evaluate to three colors: Yellow represents a

values	latitude	longitude
4	40.8242006	-73.9455163
3.6	40.83073	-73.81807
0.9	40.6681595	-73.7957993
2.5	40.75057	-73.98536
2.7	40.75057	-73.98536

Table 4: Data Frame for GeoPlotLib. Values represent either the hygiene score or rating.

low score/rating, orange means the it belongs to the middle third while a red indicates a high score/rating. It should be kept in mind a high hygiene score means a bigger violation while a high rating score means a better rated restaurant. The batch painter and drawer plot the values on the map depending upon the zoom level.

R Visualizations For the current ratings and hygiene scores, we found a perfect shapefile for NYC and imported them in R. Then we plotted the restaurants on the shapefile. After that we went forward with the actual Ratings and Hygiene Score analysis. We split the Scores and Ratings into bins of 10 and then we made a heatmap for them to show which places have better hygiene and better ratings. The higher the hygiene score, more the violations a restaurant would have faced. The rating distributions are directly proportional to as good the restaurants are. In this way, we gathered conclusive results regarding Hygiene scores and rating distribution of restaurants across NYC using data from Yelp API and NYDOH Hygiene data in R.

7 PERFORMANCE OBJECTIVES

The performance objective of this analysis is a two part problem. Since we are trying to predict the expected hygiene score as well as yelp rating of all the bars and restaurants in the New York City area, we do not have a traditional testing set to compare our results with. Hence, the objectives for performance would be

- To identify the best algorithm among the machine learning methods we run on the data set between themselves.
- To test the performance of the algorithms individually on the data set and find the absolute performance metrics.

The performance benchmarks would be a comaprative cross validation score between the different algorithms. The quantifyable quantity achieved by this method should justify the

results and reliably judge the model produced by the four algorithms which we use to generate the models and predict the score/rating. However, we should note that these models however accurate would still generate some error on some data points attributing to the following reasons

- Noise in the training data
- Fitting accuracy
- Unexperienced change in pattern

Therefore, these results should not be taken as an absolute vector but rather as an soft indicator of expectation.

8 VALIDATION

Validation in the current dataset is done via cross-validation. We, for the purpose define the following terms.

- ***k*-fold validation** In the k-fold cross validation technique, the training dataset is divided into k number of equal sized disjoint subsets. After the division, one of these subsets are used for testing and the remaining $k-1$ subsets are used for training.
- **Random Sampling** Random sampling refers to selecting rows in a dataset randomly. In the implementation, we used a seeded *pseudo* random generation to select random samples.

For validation on this dataset, we use 10-fold cross validation with the division made by a pseudo random generator. We shall present the results in a later section after discussing the fitting performance as well as the generated visualizations.

9 EXECUTION

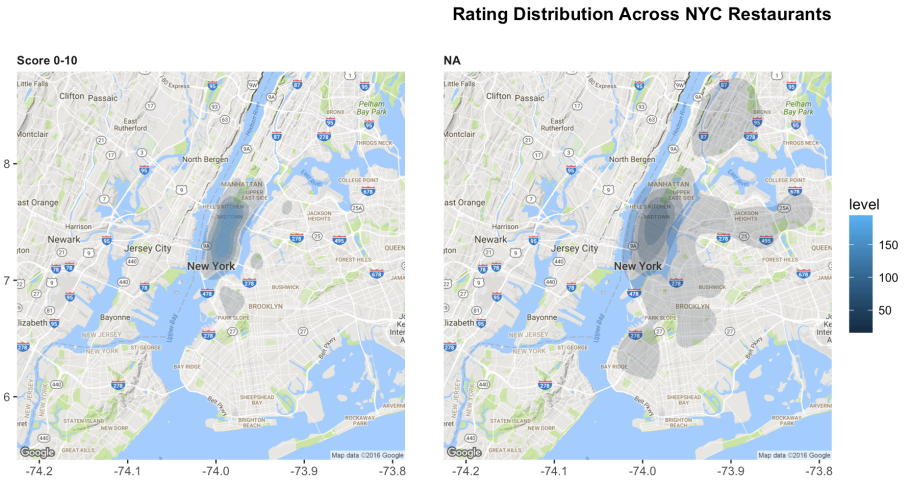

Figure 2: A plot of the distribution of Yelp rating across NYC Restaurants and Bars as obtained by the **YELP Fusion API**

We used python for the final evaluation as well as for the purpose of data cleaning and joining. We performed sparse dictionary construction and constructed pickle files for the serializing and storing the state of the intermediate forms of the data. You should find the python scripts in the scripts directory and these specific tasks are in *provider.py*. A snippet from the code which is self explanatory,

```
    # Construct sparse vector
        vector_maker = DictVectorizer()
        train_vector = vector_maker.fit_transform(train).toarray()

        # For reference print the labels for vector
        vector_maker.get_feature_names()

        # Save Data into files
        with open('train.pkl', 'wb') as output:
            pickle.dump(train_vector, output, pickle.HIGHEST_PROTOCOL)
        print "\nTraining data saved"
```

Listing 1: Construction of data for Regression

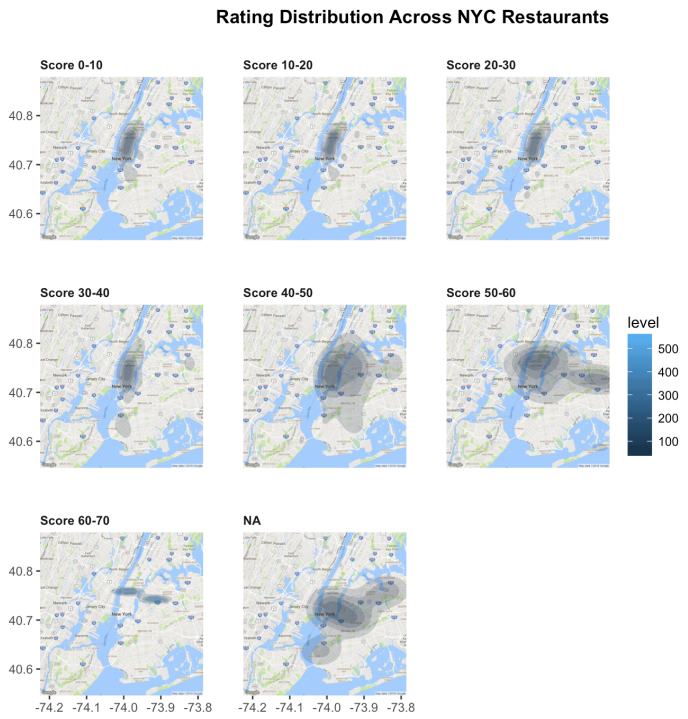

Figure 3: A plot of distribution of hygiene offenses based on the rating provided by the NYC Department of Health

Once we construct the appropriate tensors we use regression to construct the models after several epochs of training. Our machine learning pipeline looks like the following in python, which can be found in *learning.py*.

```
        # Ridge Regression
        reg = linear_model.Ridge(alpha=0.25)
        reg.fit(train, labels)

        results = []
        for index in range(0, len(train)):
            result = reg.predict(train[index])[0]
            latitude = train[index][531]
            longitude = train[index][532]
            results.append({'value': result, 'lat': latitude, 'lon': longitude})
        keys = ["value", "lat", "lon"]
        with open('ridge_regression_' + type + '.csv', 'wb') as output:
            dict_writer = csv.DictWriter(output, keys)
            dict_writer.writerows(results)
```

Listing 2: Snippet of one of the methods we used for learning the data

You will find that the parameters of the algorithms are already defined. We have had many iterations of runs for the entire dataset of more than half a million restaurants which led to different results and tuned these parameters accordingly. For some algorithms like Bayesian Ridge Regression however, we do not have to define so many arguments as the probabilistic model takes care of these parameters.

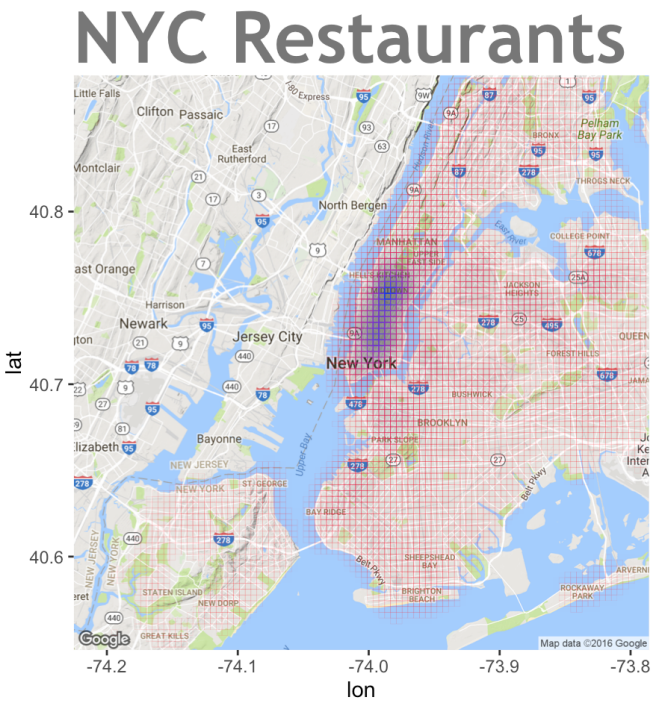

Figure 4: A plot of the distribution of location of each retrieved restaurant, bar or eatery in NYC

All the figures used in the report are also generated in a python wrapper library called geoplotlib, which results form self explanatory code which can be found in *R_viz.R*, *PlotRatingsfromLearntCSVs.py* and *PlotHygienefromLearntCSVs.py*.

With the help of the libraries and some hard coding we were able to achieve the visualizations presented in the next section of the report. For running the code and environment setup, please refer to *Instructions.txt* and *requirements.txt*.

10 RESULTS

The results presented here onwards depict the predictions that we make by using the 4 machine learning algorithms that were described above. The results are presented in two folds,

- The predicted Yelp review score for individual restaurants and bars in 2017 which are plotted over the map of NYC to get an accurate idea of the expected distribution of the positions of concentration of higher or lower rated restaurants.
- The expected health code violations in the city per restaurant or bar which is plotted across the map where the number score is directly proportional to the severity of the heal code violation.

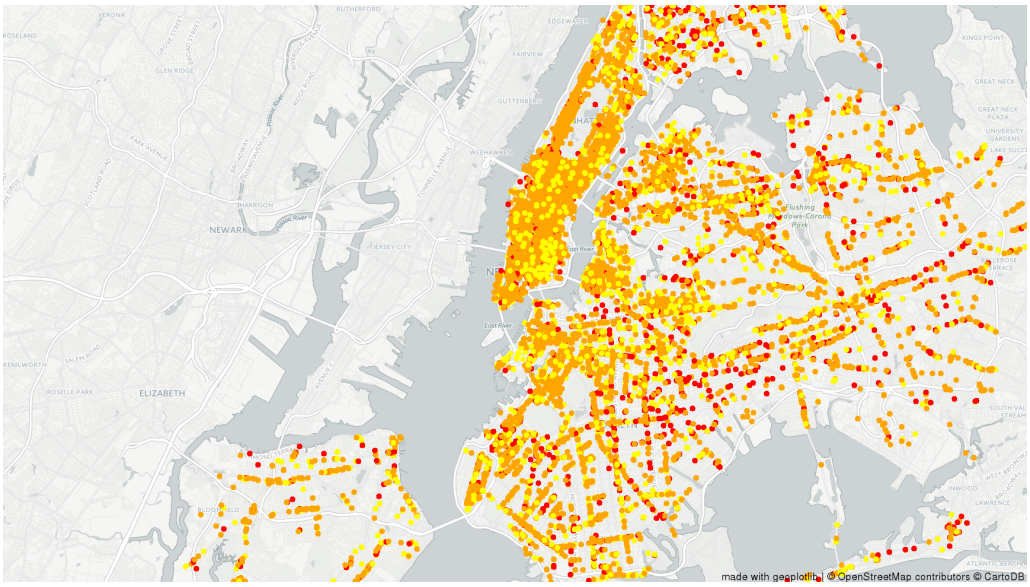

Figure 5: Plot of the expected rating on YELP of NYC restaurants, Bars and Eateries in 2017 by Linear Regression Algorithm

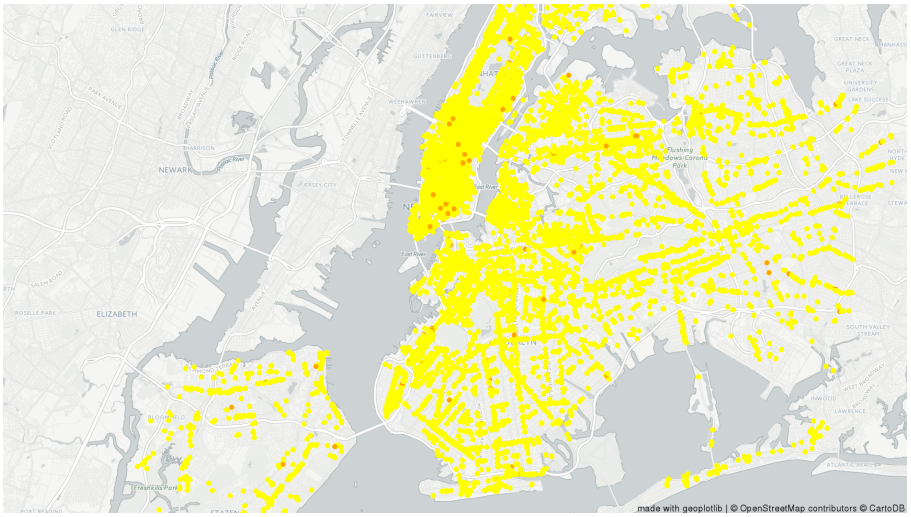

Figure 6: Plot of the predicted health code violations of NYC restaurants, Bars and Eateries in 2017 by Linear Regression Algorithm

By an visual inspection we can easily determine the clusters of belts where there is an expectation of a higher dining experience leading to a better yelp rating, which also coincides with the traditionally known places for dining like hell's kitchen and midtown. We can also visibly see that the violations are a direct coincidence with the economic demography of the city depicting that the known centers of rules violations also follow in food based commercial hubs.

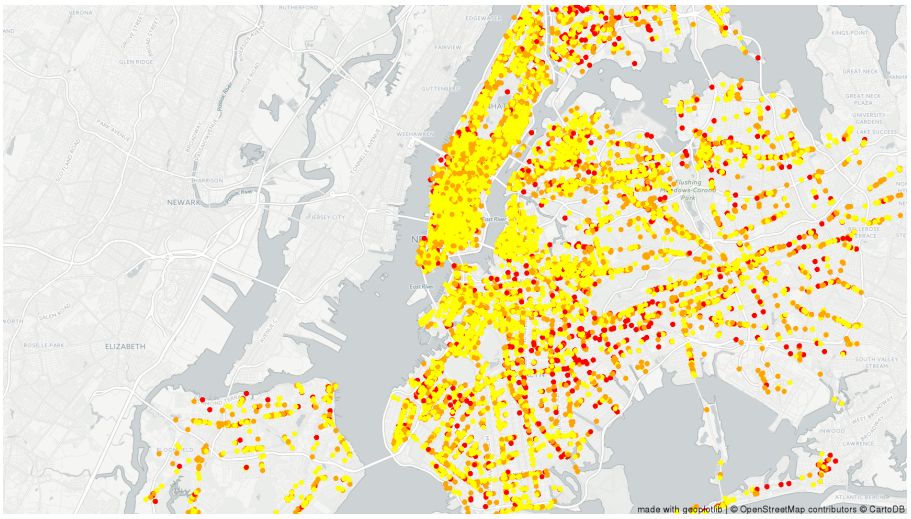

Figure 7: Plot of the expected rating on YELP of NYC restaurants, Bars and Eateries in 2017 by Ridge Regression Algorithm

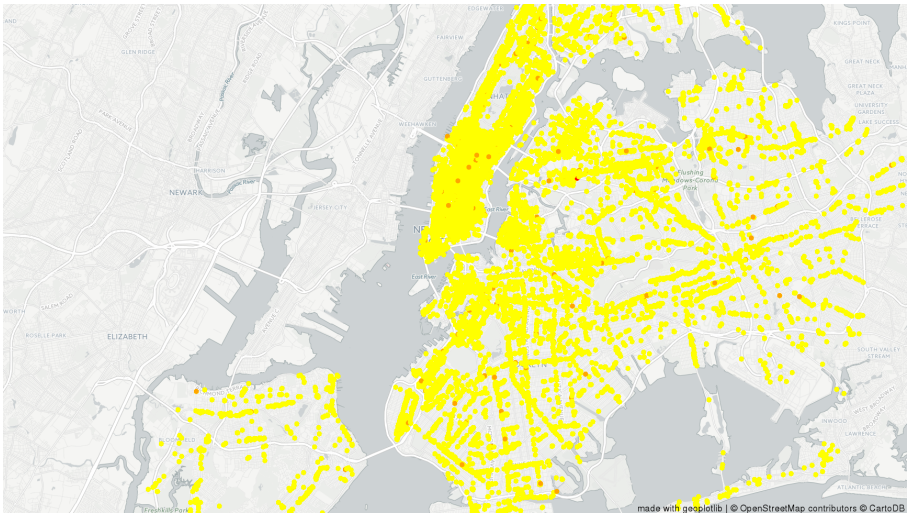

Figure 8: Plot of the predicted health code violations of NYC restaurants, Bars and Eateries in 2017 by Ridge Regression Algorithm

We also find that the different machine learning algorithms also make some marginal but noticeably different results for predictions, perhaps owing to the threshold decisions, despite having very similar low error.

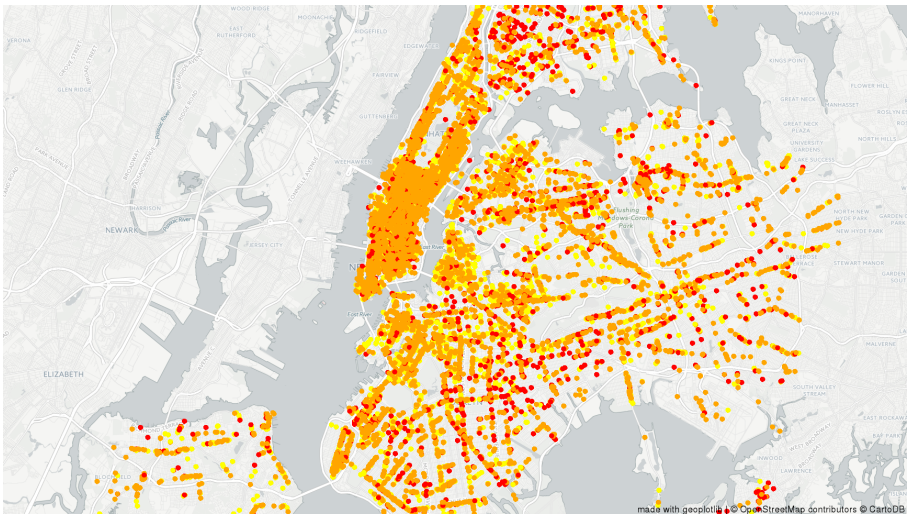

Figure 9: Plot of the expected rating on YELP of NYC restaurants, Bars and Eateries in 2017 by Bayesian Ridge Regression Algorithm

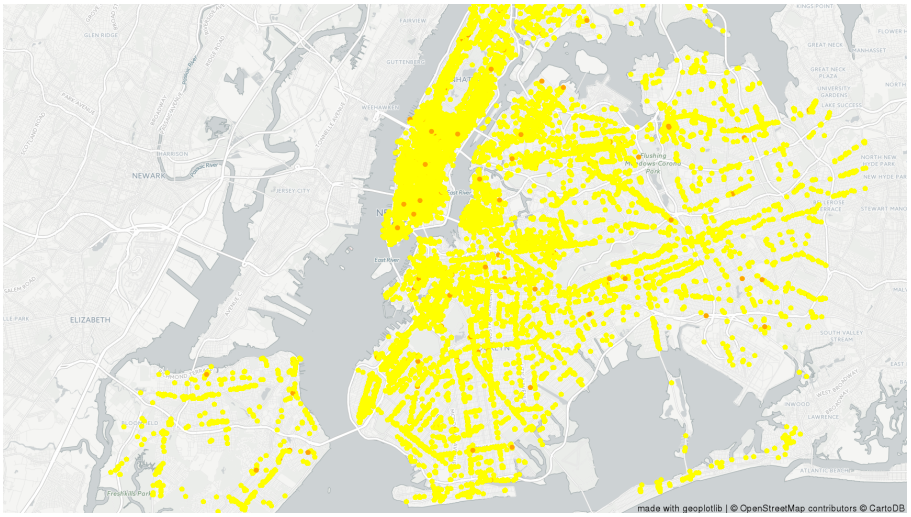

Figure 10: Plot of the predicted health code violations of NYC restaurants, Bars and Eateries in 2017 by Bayesian Ridge Regression Algorithm

One of things to note in these algorithms is that these are not lazy algorithms which mean that they are disjoint from the context of one-shot learning, i.e. the model has to be trained proactively in order to be used for the purpose of predictions and cannot passively generate any decisions, like some algorithms like Generative Adverserial Networks.

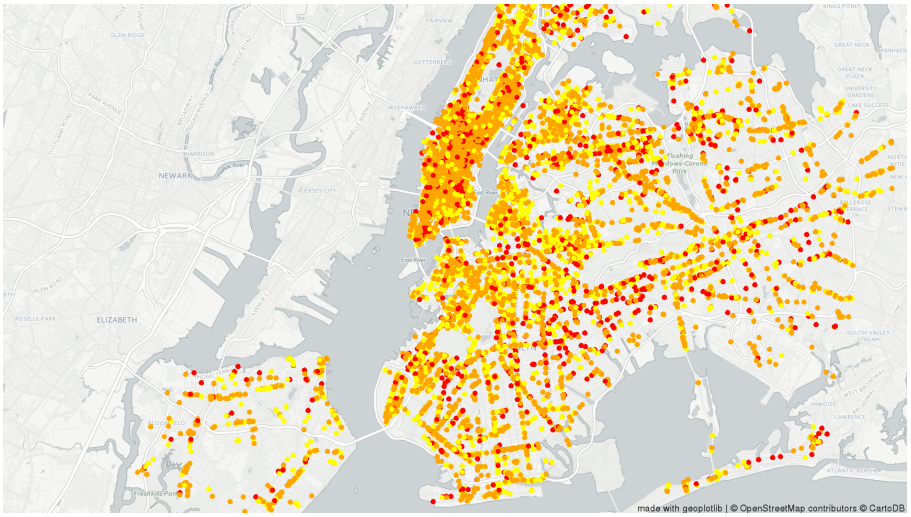

Figure 11: Plot of the expected rating on YELP of NYC restaurants, Bars and Eateries in 2017 by Lasso Regression Algorithm

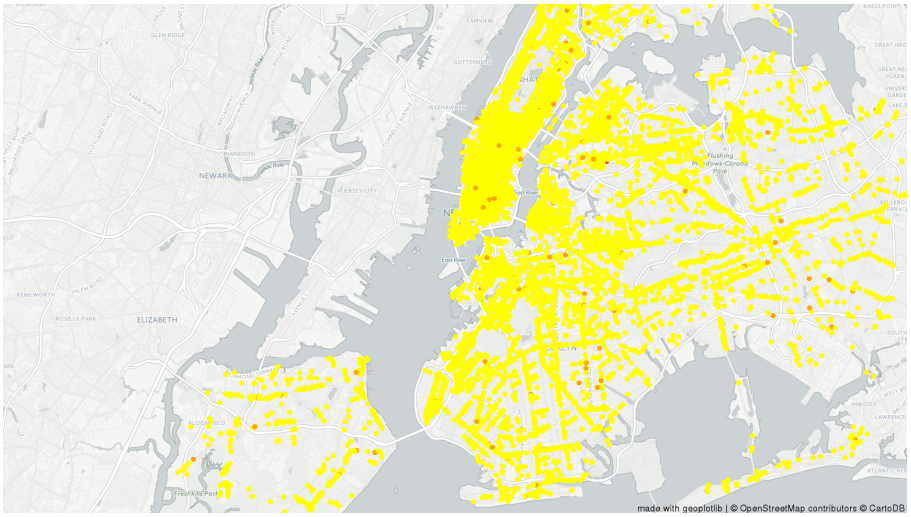

Figure 12: Plot of the predicted health code violations of NYC restaurants, Bars and Eateries in 2017 by Lasso Regression Algorithm

11 CONCLUSION

With the extensive visualization techniques and the analysis we performed on the dataset we can easily determine the changes in the patterns of data hygiene of the city as well as individually predict the hygiene score as well as the expected yelp rating for restaurants and bars in NYC. This analysis can be used in for recommendations in engines such that of yelp, zomato, foursquare, etc. as well as can be used by the city authorities for an optimized evaluation route. we can see that

REFERENCES

[1] Nabiha Asghar. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*, 2016.

[2] Marco Antonio Andrade Barrera. Predicting stars: application of three machine learning algorithms. 2015.

[3] Jaime G Carbonell, Ryszard S Michalski, and Tom M Mitchell. An overview of machine learning. In *Machine learning*, pages 3–23. Springer, 1983.

[4] Patricio Córdova. Analysis of real time stream processing systems considering latency. *University of Toronto patricio@ cs. toronto. edu*, 2015.

[5] Andrea Cuttone, Sune Lehmann, and Jakob Eg Larsen. Geoplotlib: a python toolbox for visualizing geographical data. *arXiv preprint arXiv:1608.01933*, 2016.

[6] Tri Doan and Jugal Kalita. Sentiment analysis of restaurant reviews on yelp with incremental learning.

[7] Ludwig Fahrmeir, Thomas Kneib, et al. Bayesian smoothing and regression for longitudinal, spatial and event history data. *OUP Catalogue*, 2011.

[8] Wei Fan and Albert Bifet. Mining big data: current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter*, 14(2):1–5, 2013.

[9] Yifei Feng and Zhengli Sun. Yelp user rating prediction.

[10] Tadayoshi Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21(2):137–146, 2011.

[11] Huiji Gao and Huan Liu. Data analysis on location-based social networks. In *Mobile social networking*, pages 165–194. Springer, 2014.

[12] Nitish Gupta and Sameer Singh. Collective factorization for relational data: An evaluation on the yelp datasets. Technical report, Citeseer, 2015.

[13] Louis Alberto Gutierrez, Ron Eglash, and Mukkai Krishnamoorthy. Modeling good data in dynamically changing datasets.

[14] Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

[15] Hilary M Hearnshaw, David J Unwin, et al. *Visualization in geographical information systems.* John Wiley & Sons Ltd, 1994.

[16] Arthur E Hoerl, Robert W Kannard, and Kent F Baldwin. Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, 4(2):105–123, 1975.

[17] Bo Hu and Martin Ester. Social topic modeling for point-of-interest recommendation in location-based social networks. In *2014 IEEE International Conference on Data Mining*, pages 845–850. IEEE, 2014.

[18] James Huang, Stephanie Rogers, and Eunkwang Joo. Improving restaurants by extracting subtopics from yelp reviews. *iConference 2014 (Social Media Expo)*, 2014.

[19] Jun Seok Kang, Polina Kuznetsova, Michael Luca, and Yejin Choi. Where not to eat? improving public policy by predicting hygiene inspections using online reviews. In *EMNLP*, pages 1443–1448. Citeseer, 2013.

[20] Aditya Kankanala. *Yelp data into insights.* PhD thesis, California State University, Sacramento, 2016.

[21] Daniel A Keim. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[22] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995.

[23] Krzysztof Koperski, Junas Adhikary, and Jiawei Han. Spatial data mining: progress and challenges survey paper. In *Proc. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada*, pages 1–10. Citeseer, 1996.

[24] Jia Le Xu and Yingran Xu. Recommendation system using yelp data.

[25] Lon-Mu Liu, Siddhartha Bhattacharyya, Stanley L Sclove, Rong Chen, and William J Lattyak. Data mining on time series: an illustration using fast-food restaurant franchise data. *Computational Statistics & Data Analysis*, 37(4):455–476, 2001.

[26] Frederick Mosteller and John Wilder Tukey. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

[27] Thomas Palomares, Alexis Weill, and Arnaud Guille. Cs229 project report: Yelp personalized reviews. 2014.

[28] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[29] Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. In *Encyclopedia of database systems*, pages 532–538. Springer, 2009.

[30] Sujit K Sahu, Dipak K Dey, and Márcia D Branco. A new class of multivariate skew distributions with applications to bayesian regression models. *Canadian Journal of Statistics*, 31(2):129–150, 2003.

[31] Sumedh Sawant and Gina Pai. Yelp food recommendation system.

[32] André Skupin and Sara Irina Fabrikant. Spatialization methods: a cartographic research agenda for non-geographic information visualization. *Cartography and Geographic Information Science*, 30(2):99–119, 2003.

[33] Robert Spence. *Information visualization*, volume 1. Springer, 2001.

[34] Masahiro Takatsuka. An application of the self-organizing map and interactive 3-d visualization to geospatial data. In *Proceedings of the 6th International Conference on GeoComputation*, pages 24–26, 2001.

[35] Wei Tan, M Brian Blake, Iman Saleh, and Schahram Dustdar. Social-network-sourced big data analytics. *IEEE Internet Computing*, 17(5):62–69, 2013.

[36] Shashank Uppoor and Shreyas Pathre Balakrishna. Predicting restaurant health inspection penalty score from yelp reviews.

[37] Colin Ware. *Information visualization: perception for design*. Elsevier, 2012.

[38] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[39] CS240A Guang Yang and Proposal Mingjia Han. Topic: Parallelize: Efficient yelp data mining.

[40] Yinshi Zhang. Semantic feature analysis and mining for yelp rating prediction.

LIST OF FIGURES

LISTINGS