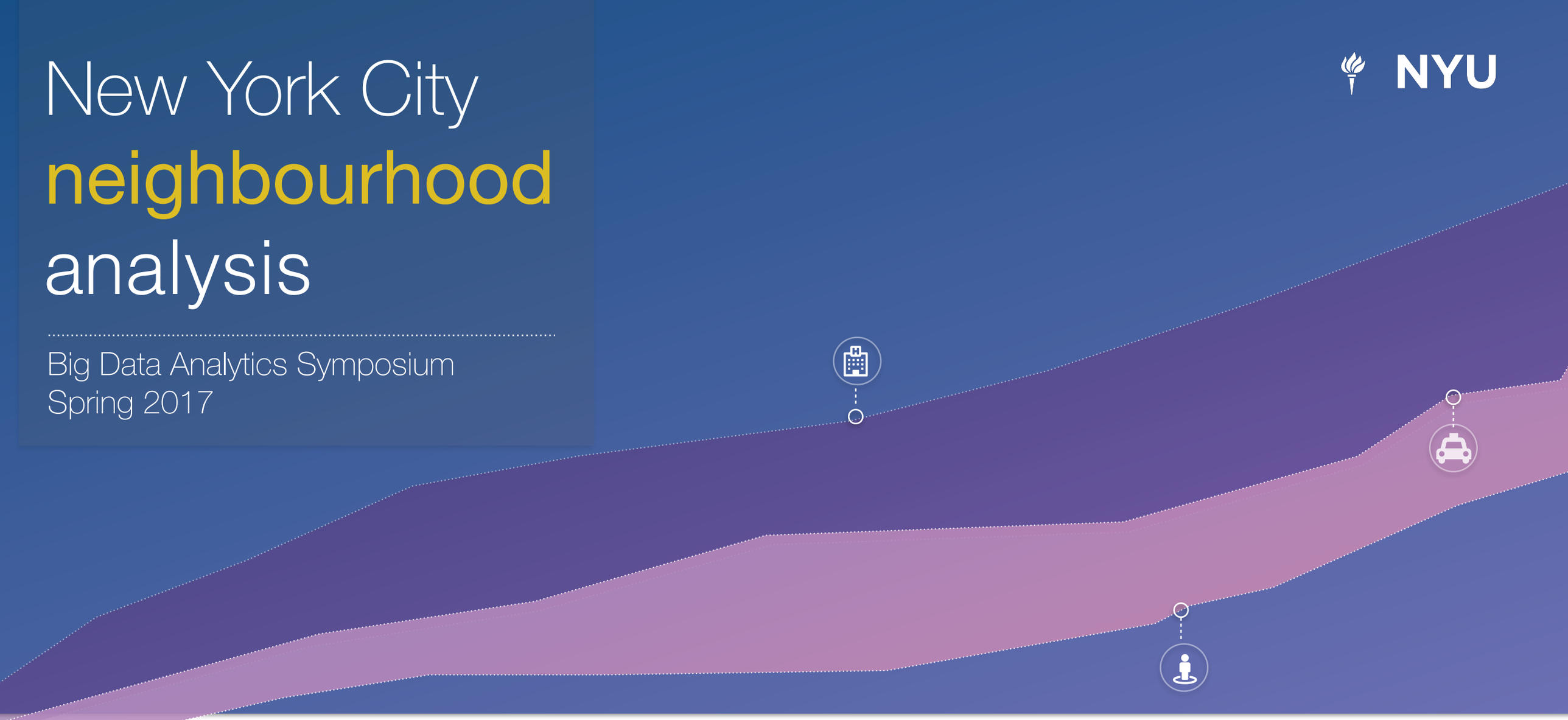
New York City neighbourhood analysis

Big Data Analytics Symposium
Spring 2017



Anish V. Mahesh
anish.mahesh@nyu.edu

Sanchit Mehta
sanchit.mehta@nyu.edu

Rohit R. Muthyala
rrm404@nyu.edu

Naman Kumar
naman.kumar@nyu.edu

Towards the partial requirement for the course **CSCIGA-3033.05**

Abstract

“Our project aims to find the best neighbourhoods in New York City. Parameters that we have taken into consideration are crime data, 311 service requests, availability of cabs and health inspection records of restaurants. We have also created a model which also predicts a restaurant’s health rating in a particular location on the basis of these factors. ”

MOTIVATION

Who are the users of this analytic?

- Government Agencies
- People seeking information about NYC neighbourhoods
- Restaurant Owners

Who will benefit from this analytic?

- Realtors, Restaurant owners, people looking to move in to NYC, New York City Council.

Why is this analytic important ?

- Municipal governments around the globe are employing big data and Internet-of-Things applications to improve many aspects of daily life. This analytic will help govts in making these decisions.

- With our model, we have proven the hypothesis that even a health rating of a restaurant which is thought to be primarily dependent on the cleanliness of a restaurant - also depends on the above external factors.

GOODNESS

CROSS VALIDATION

We split out input data in the ratio 80%-20% into train and test and cross validated the test data values by plugging them into our model.

WEB INDEX LIVEABILITY

We compared our results with the popular online surveys (eg Bloomberg) about the areas in NYC with highest Crime, best areas to live etc

MULTIPLE MODELS

We tried different models including OLS, Regression Trees, Random Forests, K Means ensuring the our root mean square error was low and predictions were more accurate.

DESIGN DIAGRAM

Datasets

RestaurantNYC

NYC 311 Service Requests

NYC Crime

NYC Taxi

Iterative

Process

- Data Cleansing
- Join Datasets
- Remove Duplicates
- Prepare Data

hadoop PIG

Store

Save

- HDFS

Data Analysis

Analyse

- Analyse Data
- Build Models
- Iteratively Test Models
- Leverage MLLib in Scala

APACHE Spark

Visualise

- Visualise the analytics
- Iteratively improve the model based on these visualisations

Action

- Predict Health Grade for new Businesses

We ran all our analytics on NYU HPC Cluster: **Dumbo**

DATA SOURCES

311

Size: 1.2 GB

Description:

This dataset includes the all 311 Service Requests from 2010 to present. Records include type of calls, NYC department that handled the call, lat long of the incident and zip of the incident:

Features used:

1. Latitude and Longitude of the incident
2. Incident Type
3. Department that handled the incident
4. Incident Zip
5. Status of the complaint
6. Date of the complaint

Total Entries: 46,087,462 entries

Link to the data: https://data.cityofnewyork.us/dataset/311-Service-Requests-From-2011/fpz8-jqf4

Crime

Size: 1.4 GB

Description:

This dataset includes New York City Police Department's records reported crime and offense data based upon the New York State Penal Law and other New York State laws.

Features used:

1. Type of Crime Violation
2. Cime Latitude and Longitude
3. Felony Misdemeanour Count
4. Attempted Crime Count
5. Completed Crime Count

Total Entries: 1,312,321 entries

Link to the data: https://data.cityofnewyork.us/Public-Safety/Historical-New-York-City-Crime-Data/hqhv-9zeg

Cabs

Size: 10 GB

Description:

This dataset includes trip records from all trips completed in yellow taxis from in NYC. Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

Features used:

1. Pickup Latitude and Longitude
2. Dropoff Latitude and Longitude
3. Number Of passengers in the cab
4. Total Fare amount
5. Tip Given
6. Distance Traveled.

Total Entries: 1,598,223 entries

Link to the data: https://data.cityofnewyork.us/view/ba8s-jw6u

Health Records

Size: 153 MB

Description:

This dataset includes the unannounced inspections of restaurants for compliance in food handling, food temperature, personal hygiene and vermin control. Each violation of a regulation gets a certain number of points. Each restaurant also has an inspection score—the lower the score, the better the Grade.

Features used:

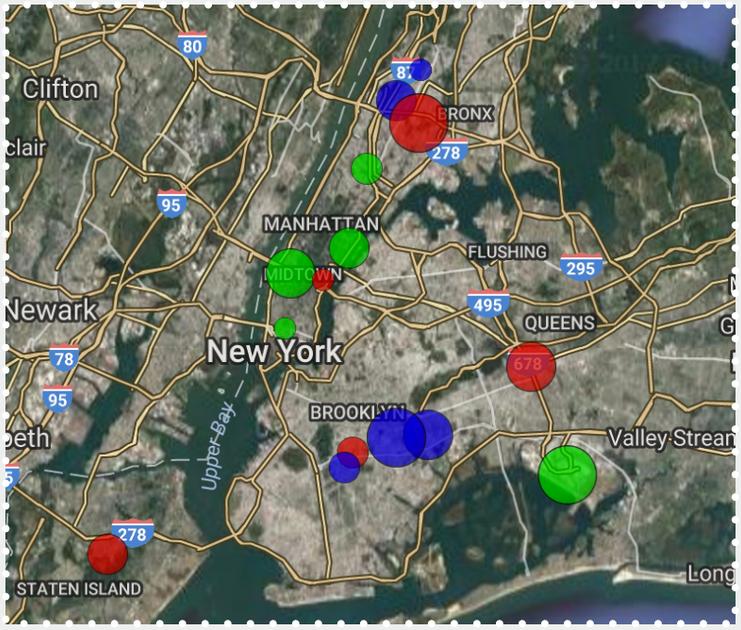
1. Restaurant Zip
2. Cuisines
3. Facilities
4. Hygiene
5. Violations
6. Restaurant Name.

Total Entries: 24,000 entries

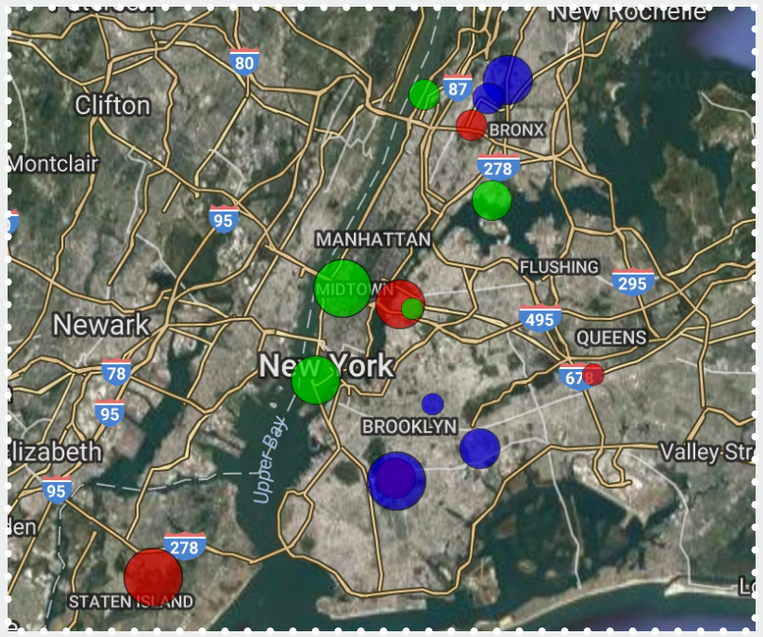
Link to the data: http://www1.nyc.gov/site/doh/services/restaurant-grades.page

ANALYTIC 1

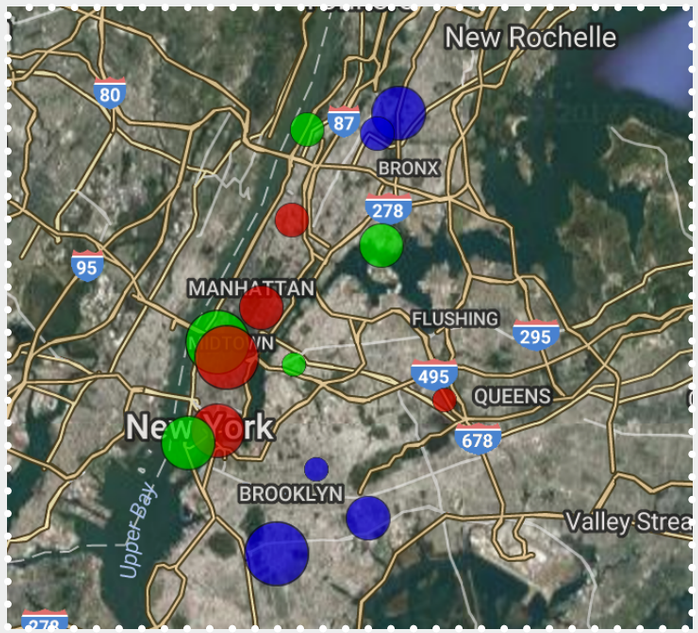
Clustering of Areas



Crime



311



Cabs

Total K-Means Ratio

ANALYTIC 2

Contribution of factors to the Liveability Index

Elastic Net

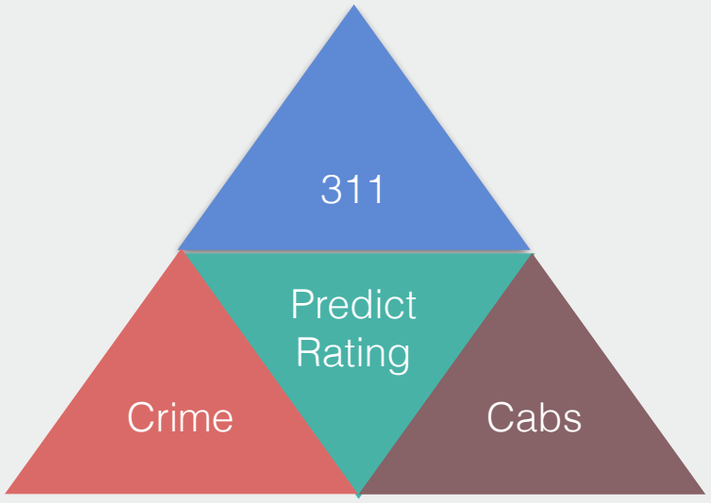
Feature	Weight
grade_C_count	-1.57155790709
grade_A_count	0.717074159885
311_sev1	0.596956381908
grade_B_count	0.395171959285
311_sev4	-0.167361289566
avg_cost	-0.144629209067
311_sev3	-0.0338164394846
311_sev2	0.032743313675
violation	-0.0080300600252
crime_sev2	-0.00603298325994

Random Forest

Feature	Weight
311_sev1	0.114508055309
311_sev2	0.0994473511203
avg_cost	0.0982458848024
avg_tip	0.0787866200138
crime_attempted	0.0669060607174
crime_sev1	0.06650557626
grade_A_count	0.0560312216581
311_sev4	0.0458198618283
crime_sev2	0.0454442041732
311_sev3	0.0426292275173

ANALYTIC 3

Important external features that affect the Restaurant Health Grade



Key

311_sev - 311 Complaint Severity Level
crime_sev - Crime Report Severity

GRADE I FEATURES FROM MODELS

A

Model	Elastic Net	Model	Random Forest
Features Selected	avg_tip, avg_cost, crime_attempted, 311_sev1, 311_sev4, 311_sev3, felony, crime_sev1	Features Selected	311_sev2, 311_sev4, crime_sev2, 311_sev3, misdemeanour, 311_sev1, avg_cost

B

Model	Elastic Net	Model	Random Forest
Features Selected	avg_tip, avg_cost, crime_attempted, 311_sev1, 311_sev4, 311_sev3	Features Selected	311_sev4, 311_sev2, 311_sev3, avg_tip, 311_sev1, crime_sev2

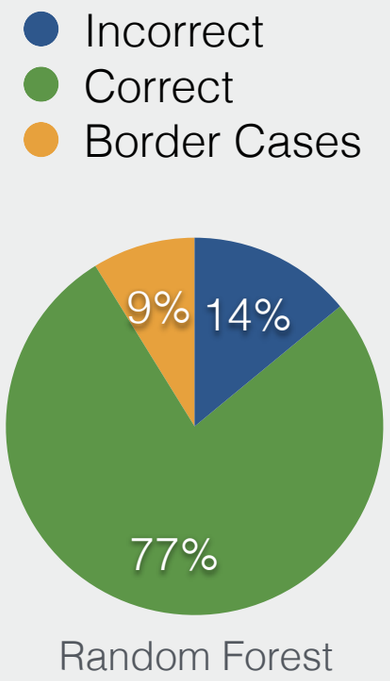
C

Model	Elastic Net	Model	Random Forest
Features Selected	311_sev3, violation	Features Selected	311_sev4, crime_sev2

ANALYTIC 4

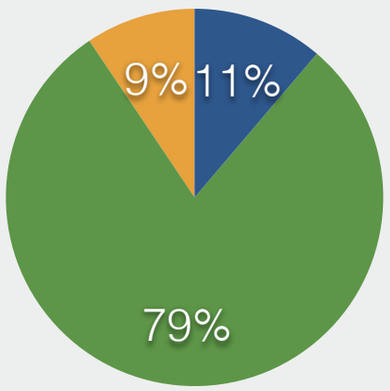
Restaurant Health Grade

Prediction Results



Random Forest

Model	Random Forest
RMSq Error	5.90238



Linear Regression

Model	Linear Regession
RMSq Error	5.258439
Average Rating	15.41648590021692 (15.792275774951598)

- Estimated Restaurants ratings using external features like cab data, 311 data & crime data
- Each of these datasets were processed for feature selection - giving each of them a severity score
- From the above graphs we can see than the Linear Regression model performed better.

OBSTACLES

- We scraped data from Yelp but records were not enough to build an accurate model. We found similar features in a new dataset from NYC Health Inspection Records.

- We did not have a deep understanding of Scala, Spark and ML-Lib, and hence took time to understand the syntax and their working in order to perform necessary computations for the analytic.

- There were a lot corrupt rows in 311, crime, cab datasets.

- Outages in HPC cluster

SUMMARY.

"We performed a zip code level analysis of NYC. Our first analytics showed us the variation in the data and that zip code are good classification points.We then used this classification to find feature importance in determining the liveability index. To add on to this we proved our hypothesis that external factors do affect restaurant's health inspection grade. In order to utilise the information and analytics that we gathered, developed models that can predict a restaurant's health grade. Thus bridging the gap between information and action."

ACKNOWLEDGEMENTS

NYU HPC Team : Mr. Shenglong Wang

REFERENCES

[1] Nitish Gupta and Sameer Singh, *Collectively Embedding Multi-Relational Data for Predicting User Preferences.* Published in arXiv:1504.06165v1 [cs.LG] 23 Apr 2015

[2] Huiyu Sun and Suzanne McIntosh, *Big Data Mobile Services for New York City Taxi Riders and Drivers*. Published in 2016 IEEE International Conference on Mobile Services (MS)

[3] H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin and M. Chau, *Crime data mining: a general framework and some examples.* Published in: Computer (Volume: 37, Issue: 4, April 2004)

[4] Scott Minkoff, *NYC 311: A Tract-Level Analysis of Citizen–Government Contacting in New York City*, Published in sagepub.com/journalsPermissions.nav DOI: 10.1177/1078087415577796

[5] Harvard Data Science, *"Big Data Analysis of NYC Restaurant Health Inspections", Big Data Analysis of NYC Restaurant Health Inspections, 2017.* [Online]. Available: https://harvarddatasciencerestaurantinspections.wordpress.com/2014/12/12/big-data-analysis-of-nyc-restaurant-inspections/. [Accessed: 04- May- 2017].

[6] "Apache Hadoop".

[7] "Apache Hive".

[8] "Apache Pig".

[9] "Apache Spark".

Thank You ☺