Coursera Capstone Project
Battle of the Neighborhoods

Predictive Modelling & Clustering of New York City Restaurant Inspection Results for the New York City Department of Health and Mental Hygiene

April 13, 2020

1. Introduction (1 of 2)

- The *New York City Department of Health and Mental Hygiene* has very limited inspection capacity (while restaurants in New York are plentiful). They would like to understand if there is a predictive relationship between borough and critical food violations/inspection grades, and if data could be meaningfully clustered.

- This understanding would allow them to focus their inspections on those restaurants where the likelihood for food violations is highest.

1. Introduction (2 of 2)

The following business problems will be answered.

1. *To what extent does borough reliable predict critical food violations?*

2. *To what extent does borough reliable predict inspection grade?*

3. *Can the restaurant data be meaningfully clustered by critical food violations?*

2. Data

- The "DOHMH New York City Restaurant Inspection Results" dataset (as obtained at https://opendata.cityofnewyork.us/) was used for this project.

- This dataset will be used to pull restaurant zipcode, borough, food violations and inspections grades.

3. Methodology (1 of 3)

3.1. Exploratory Data Analysis:

- Firstly, the data types were obtained to ensure the right methods could be called when scripting the analysis in Python (table 1).
- Secondly, an informative overview was generated for the count (and relative amount) of cuisines in the dataset to provide insights in the relative cuisine composition of restaurants in New York (table 2).
- Lastly, averages were calculated for critical food violations (where yes =1, and no =2) in each borough to show whether one particular borough had a substantially higher average of critical food violations (when compared to the other boroughs or overall, table 3).

3. Methodology (2 of 3)

3.2. Inferential Statistical Methods

Logistical Regressions are fit for categorial dependent variables and were conducted to analyze if borough (BORO) can reliable predict critical food violations (CRITICAL) and inspection grade (GRADE).

The entire dataset was used for analyses, but was split up in a train (75%, n=300,666) and test set (25%, n=100,222) for both regressions. A Jaccard Similarity Score was calculated to see how accurate the predicted values were with the actual outcomes in the test set.

3. Methodology (3 of 3)

3.3. Machine Learning

- DBSCAN was used (for example, over K-Nearest Neighbors) as it can identify arbitrarily shaped clusters (ideal for geographical clustering) to meaningfully group restaurants with critical food violations.
- Due to the size of the dataset (n_pop= 400,888), it was decided to reduce the amount of observation in the analysis to a randomized sample of 10,000 observations (n = 10,000).
- Three steps were completed in the Machine Learning analysis:
  - Firstly, the selected data points were plotted on the map to show the geographic distribution of restaurants.
  - Second, the selected data points were clustered according to their location.
  - Thirdly, the selected data points were clustered according to i) critical food violations results and ii) their location.

4. Results (1 of 6)

4.1. Exploratory Data Analysis

- The results of the exploratory data analysis show that the data types in the dataset are integers, objects and floating numbers (see table 1).

- The exploratory data analysis further shows that Queens ($\bar{x}$= 1.437535) and Staten Island ($\bar{x}$= 1.429506) have a lower average of critical food violations than i) the overall dataset ($\bar{x}$= 1.441575) and ii) the boroughs Bronx ($\bar{x}$= 1.442299), Brooklyn ($\bar{x}$= 1.445071) and Manhattan ($\bar{x}$= 1.442545, see table 3.

4. Results (2 of 6)

4.1. Exploratory Data Analysis

Table 1. Datatype in the dataset

Column	Datatype
CAMIS	int64
DBA	object
BORO	object
BUILDING	object
STREET	object
ZIPCODE	float64
PHONE	object
CUISINE	object
INSPECTION DATE	object
ACTION	object
VIOLATION CODE	object
VIOLATION DESCRIPTION	object
CRITICAL	object
SCORE	float64
GRADE	object
GRADE DATE	object
RECORD DATE	object
INSPECTION TYPE	object
Latitude	float64
Longitude	float64
Community Board	float64
Council District	float64
Census Tract	float64
BIN	float64
BBL	float64
NTA	object
Council District	float64
Census Tract	float64
BIN	float64
BBL	float64
NTA	object

Table 2. Count of cuisine type in the dataset

Cuisine Type	Count	Perc.
American	83424	20,81%
Chinese	42348	10,56%
Café/Coffee/Tea	19843	4,95%
Latin	17643	4,40%
Pizza	17448	4,35%
Mexican	16693	4,16%
Italian	16228	4,05%
Caribbean	14493	3,62%
Japanese	14457	3,61%
Bakery	12493	3,12%
Spanish	12476	3,11%
Pizza/Italian	8296	2,07%
Chicken	7603	1,90%
Indian	7015	1,75%
Asian	6607	1,65%
Delicatessen	6364	1,59%
Jewish/Kosher	5792	1,44%
Thai	5627	1,40%
Korean	5505	1,37%
Donuts	5284	1,32%
Juice, Fruit, Salads	4757	1,19%
French	4656	1,16%
Hamburgers	4397	1,10%
Mediterranean	4185	1,04%
Sandwiches	4039	1,01%
Irish	3424	0,85%
Seafood	3170	0,79%
Ice Cream	3106	0,77%
Middle Eastern	2846	0,71%
Sandwiches/Salads/Mixed	2836	0,71%
Bagels/Pretzels	2611	0,65%
Other	2511	0,63%
Greek	2192	0,55%

Table 2. Continued

Vietnamese/Cambodian/Malaysia	1996	0,50%
Vegetarian	1929	0,48%
African	1846	0,46%
Peruvian	1796	0,45%
Tex-Mex	1760	0,44%
Eastern European	1443	0,36%
Turkish	1301	0,32%
Steak	1275	0,32%
Bangladeshi	1260	0,31%
Bottled Beverages	1248	0,31%
Russian	1128	0,28%
Soul Food	1027	0,26%
Salads	1007	0,25%
Chinese/Japanese	853	0,21%
Barbecue	818	0,20%
Filipino	704	0,18%
Soups & Sandwiches	635	0,16%
Pakistani	634	0,16%
Hawaiian	590	0,15%
Continental	557	0,14%
Tapas	555	0,14%
Creole	520	0,13%
Brazilian	481	0,12%
Polish	481	0,12%
German	474	0,12%
Australian	432	0,11%
Chinese/Cuban	415	0,10%
Armenian	307	0,08%
Hotdogs/Pretzels	280	0,07%
English	249	0,06%
Pancakes/Waffles	229	0,06%
Ethiopian	227	0,06%
Afghan	219	0,05%
Egyptian	187	0,05%

Table 2. Continued

Hotdogs	183	0,05%
Moroccan	177	0,04%
Creole/Cajun	159	0,04%
Portuguese	150	0,04%
Soups	132	0,03%
Indonesian	128	0,03%
Cajun	102	0,03%
Californian	94	0,02%
Scandinavian	88	0,02%
Not Listed	85	0,02%
Southwestern	83	0,02%
Fruits/Vegetables	79	0,02%
Iranian	69	0,02%
Czech	43	0,01%
Nuts/Confectionary	43	0,01%
Chilean	32	0,01%
Basque	9	0,00%

Table 3. Averages of food violations in New York boroughs

Borough	Food Violation (average, 1 = yes, 2= no)
Bronx	1.442299
Brooklyn	1.445071
Manhattan	1.442545
Queens	1.437535
Staten Island	1.429506
Overall	1.441575

4. Results (3 of 6)

4.2. Inferential Statistical Methods

The results of the logistical regression show that the predictive relationship between i) borough and critical food inspections (Jaccard Similarity Score = 0.55) and ii) borough and inspection grade (Jaccard Similarity Score= 0.40) is low (see table 4).

Table 4. Results of the regression 1 (BORO, CRITICAL) and 2 (BORO, GRADE)

Regression	***Jaccard Similarity Score***
BORO -> CRITICAL	*0.5485222805372074*
BORO -> GRADE	*0.39872483087545646*

4. Results (4 of 6)

4.3. Machine Learning

The results of the machine learning tactics (i.e. DBSCAN) to cluster restaurants based on critical food violations are more promising.

- Firstly, the results of plotting the sample of the dataset (n=10,000) geographically (see figure 1) shows that the sample shows the distribution between the different boroughs in New York city, with no apparent biases for either borough (although no thorough comparison was made between the sample and the population of the relative amount of restaurants per borough).
- Secondly, the results for clustering the sample of the dataset (n=10,000) based on location shows the densest areas in New York for restaurants.
- Thirdly, the results for clustering the sample of the dataset (n=10,000) based on i) location and ii) critical food violation shows the 15 areas (incl. zipcodes) where restaurants with critical food violations are most dense (eps = 0.1 and min. sample = 30, see figure 3 and table 5).

4. Results

Figure 1. Sample restaurants plotted on them map of New York (n=10,000)

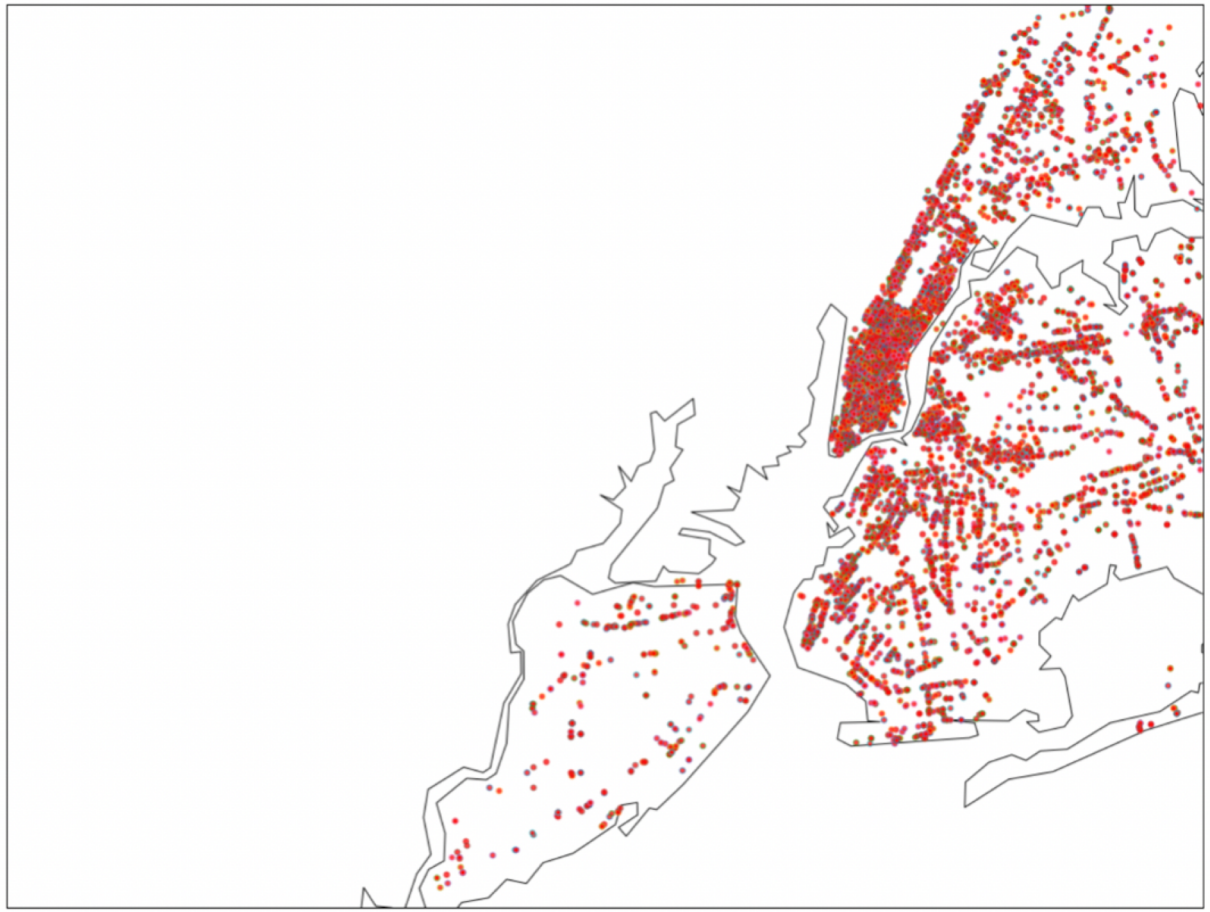
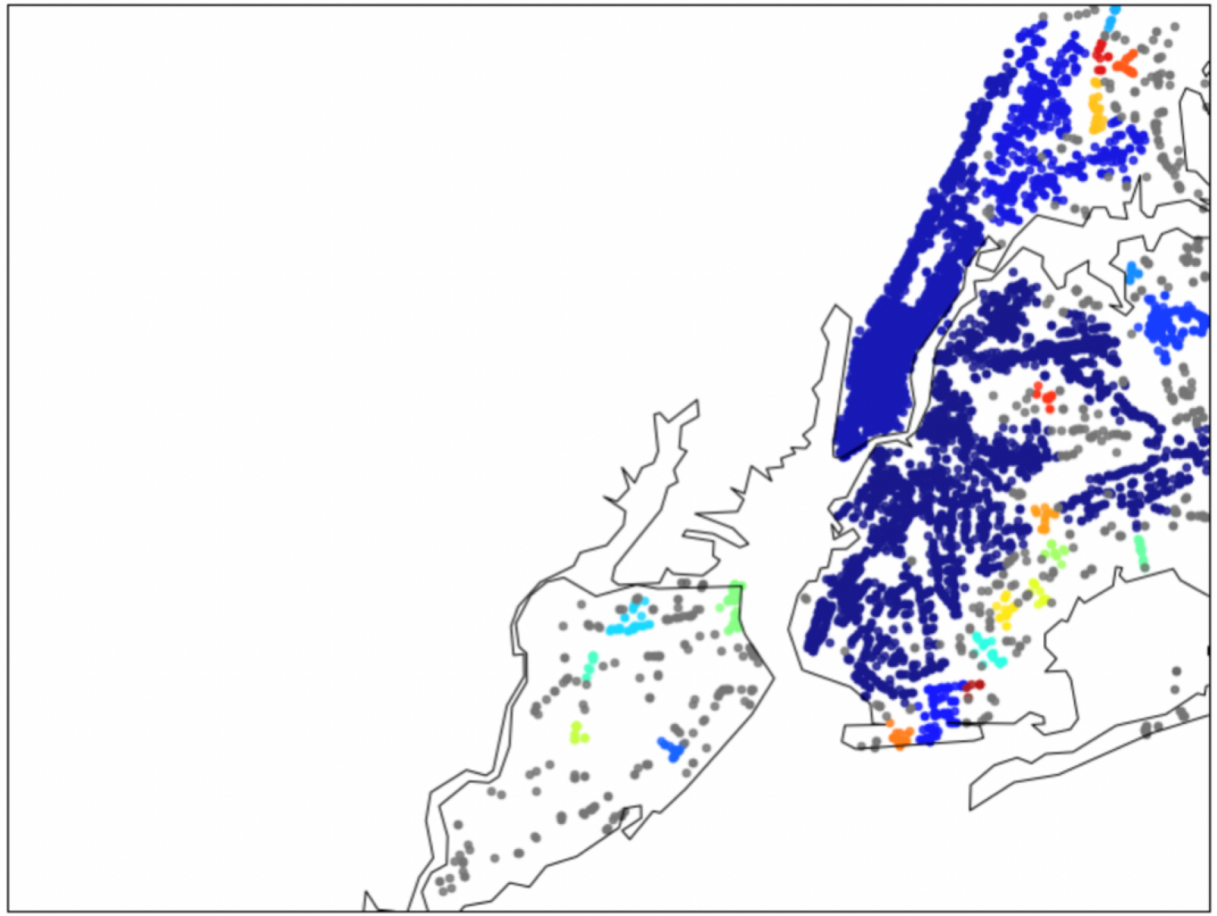


Figure 2. Sample restaurants clustered by location (DBSCAN, n =10,000)



4. Results (6 of 6)

Figure 3. Sample restaurants clustered by location and critical inspection results (DBSCAN, n=10,000

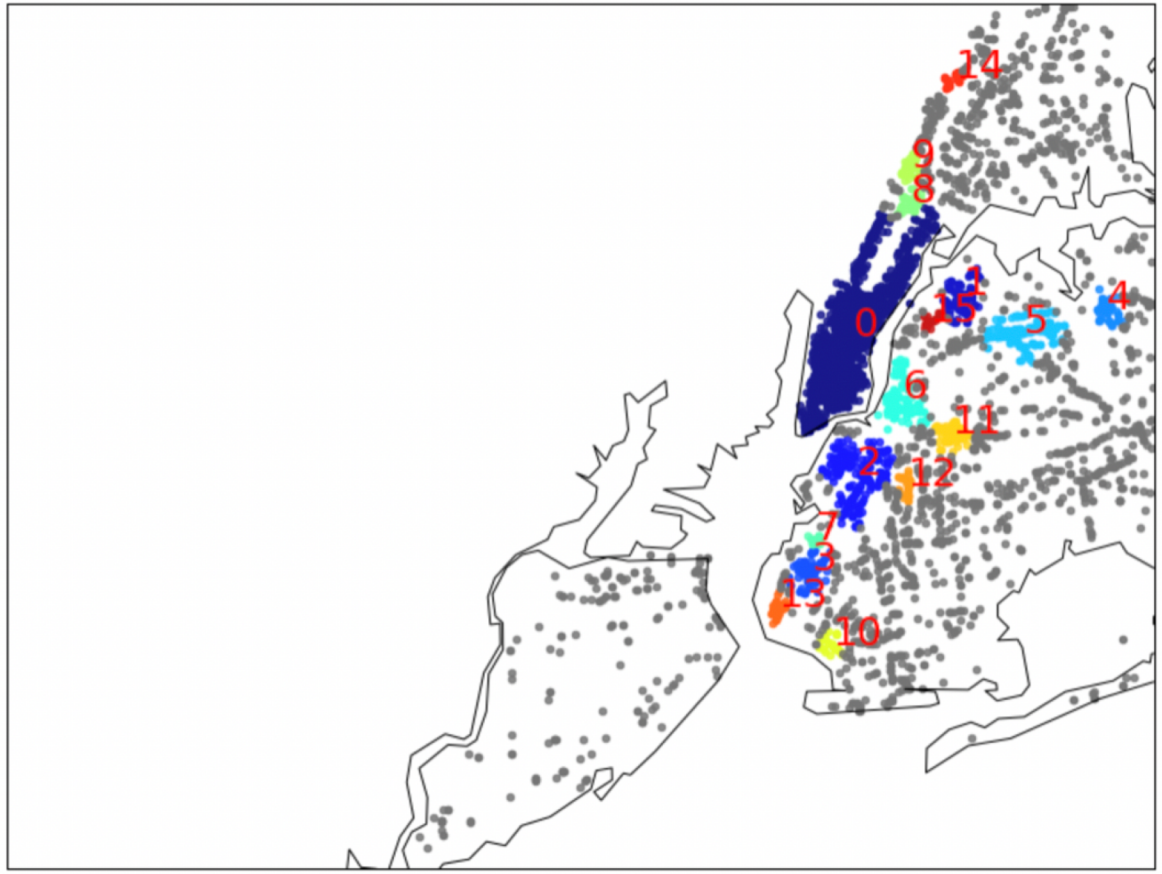


Table 5. List of zipcodes where restaurants are clustered by critical inspection results (eps = 0.1, min. sample = 30)

Cluster 0	11368, 11377, 11372, 11373, 11369
Cluster 1	11106, 11103, 11101, 11105, 11102, 11377
Cluster 2	10010, 10009, 10011, 10014, 10013, 10002, 10028, 10029, 10036, 10017, 10003, 10012, 10007, 10016, 10001, 10019, 10103, 10022, 10025, 10021, 10128, 10024, 10065, 10018, 10118, 10020, 10023, 10005, 10004, 10075, 10038, 10026, 10035, 10006, 10121, 10041, 10169, 10112, 10280, 10027, 10178, 10281, 10106, 10000, 10165, 10167, 10153, 10170, 10271, 10282
Cluster 3	11354, 11355
Cluster 4	11211, 11206, 11222, 11249
Cluster 5	11209, 11220
Cluster 6	11215, 11217, 11238, 11231
Cluster 7	11220, 11232, 11219, 11228
Cluster 8	11201, 11231
Cluster 9	11237, 11221, 11206
Cluster 10	11377, 11104
Cluster 11	10032, 10033
Cluster 12	11238, 11205
Cluster 13	10458, 10457
Cluster 11	10032, 10033
Cluster 12	11238, 11205
Cluster 13	10458, 10457
Cluster 14	10032, 10033, 10040
Cluster 15	10454, 10451, 10455

5. Discussion

Inferential Statistical Methods

- A predictive relationship between i) borough and food violations (BORO -> CRITICAL) and ii) borough and inspection grade (BORO -> GRADE) exist with very low predictability (i.e. Jaccard Similarity Score of 0.55 and 0.40 respectively).
- Both predictive models do not provide the DOHMH with practical predictive capabilities. Picking an independent variable that covers smaller geographical area (e.g. zipcodes or council districts) or other restaurant characteristics (i.e. cuisine) might produce different results. ***In short, borough do not reliable predict critical food violations or inspection grades for restaurant in New York.***

Machine Learning

- Interestingly, the machine learning approach did show that clusters could be meaningfully created. In total 15 different clusters were identified, which could be further detailed by their zipcodes (see figure 3 and table 5). These results provide meaningful insights for the DOHMH. With limited resources at their disposal and many restaurants in New York state the inspections should be focused on those fifteen areas where critical food violations are most common. Interestingly, these outcomes seem to reflect the exploratory insights of average critical food violations per borough (see table 3) as Queens and Staten Island show a relatively few number of clusters. ***In short, the restaurant data can be meaningfully clustered by critical food violations and location for restaurants in New York.***

6. Conclusion

- The conducted analysis shows that critical food violations and inspection grades cannot be reliable predicted by the borough a restaurant is located in. Clustering the data by critical food violation and location does provide a meaningful cluster of restaurants with critical food violations.
- The identified clusters could assist the DOHMH in identifying clusters for future inspections, where critical food violations are more rampant.
- Future research might produce more reliable predictive models by picking dependent variables spanning smaller geographical areas (i.e. zipcodes or council district) or other restaurant characters (i.e. cuisine).
- Furthermore, there might be other underlying conditions that could predict where these clusters form (i.e. rodent populations, income levels or even the abundance one-time customers (like tourists) that does not incentivize keeping a clean shop).