Lin, Yuhan
4/28/2018

Glossier Data Challenge

1). Data Preparation and Summary

There are six datasets providing information of New York City's restaurant attributions and violation descriptions, with 'restaurant_attributes' and 'restaurant_violation' being the dominant datasets; the rest of the datasets contain supporting information. All six datasets are joined in R in a way that ensures all the restaurants with violations are captured in the final dataset. One issue occurred during merging is missing values in the supporting data files. For example, there are 65 unique violations being inspected but 'violation_name' file only contains information on 41 unique violations. Therefore, I decided to keep the entries with missing values. In addition, I pulled an additional dataset from an external resource that contains each zip code's affiliated latitude and longitude. A small flaw of the external dataset is that it wasn't able to capture the geolocation of 1.2% of the entries in the final dataset. A brief geospatial analysis will be conducted in the next part of the report.

The final dataset includes 89304 observations and 14 variables, some of the crucial variables are: restaurant_id, zipcode, borough_name, cuisine_description, violation_id, grade, critical_level, and score. It's important to note that this dataset only captures restaurants with violations; the amount of inspections conducted on each restaurant is unknown. There are 10969 distinct restaurant id in the final dataset. Those 10969 restaurants located in 5 different boroughs, covering 83 different types of cuisines.

Three types of violation evaluation are giving during each inspection: Grade includes A,B,C,P,Z and Note Yet Graded; Score ranges from -1 to 82; Critical_flag includes Critical, Not Critical and Not Applicable. As *Figure1* shows, most of the violations inspected in this dataset are restaurants in Manhattan, followed by Brooklyn, Queen, Bronx and Statemen Island. More than 50% of the violations are deemed as critical (*Figure2*). Majority of the restaurants is given grade A (*Figure3*). In addition, Score ranges from -1 to 82, with a median of 11 and a mean of 12.59. Inspections are conducted in the span of 5 years, from 2013 to 2017. Most of the violations in the dataset are inspected from 2014 to 2016.

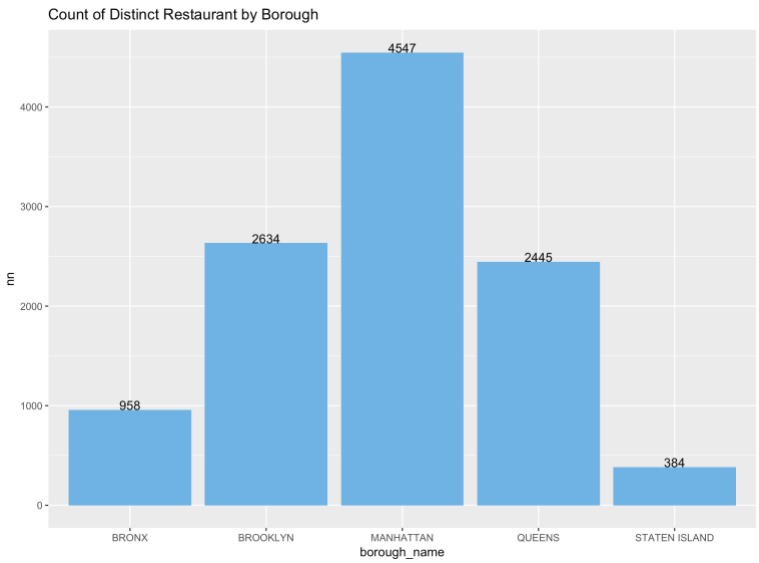


Figure1

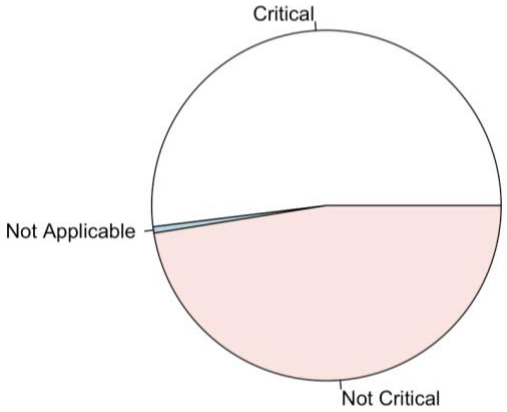


Figure2

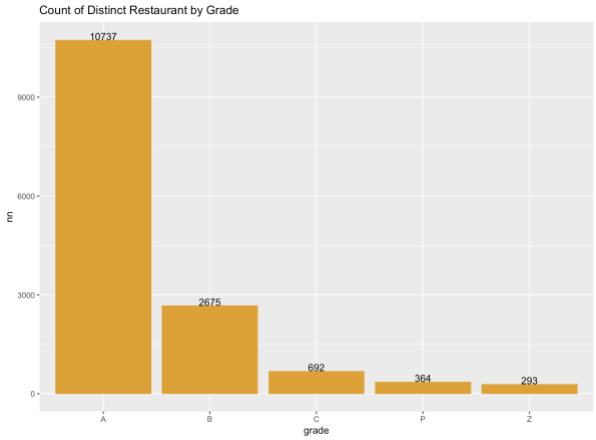


Figure3

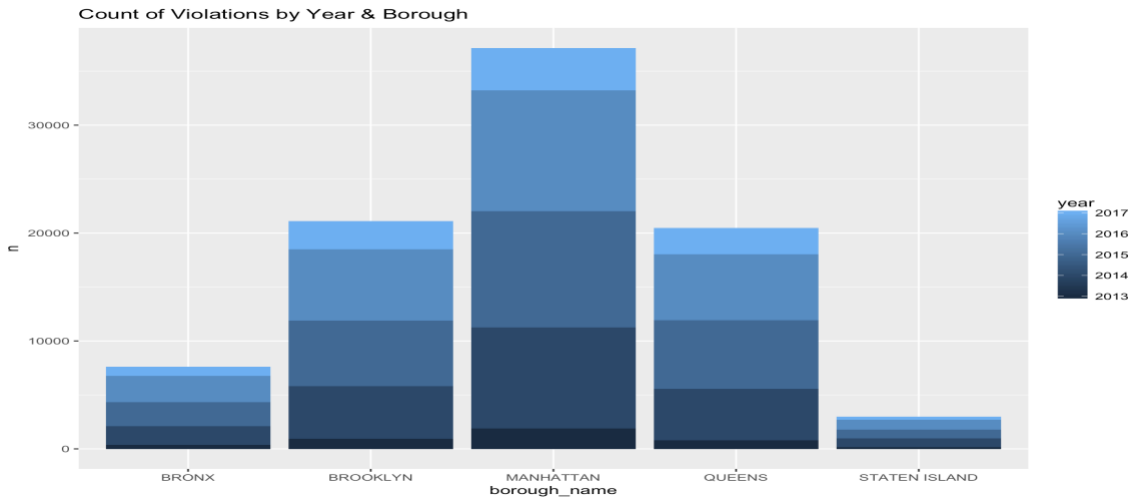


Figure4

2). Process of Analysis

Since violation level is measured in three ways, it's important to investigate further what went into the evaluation and if/how those evaluations are consistent. Assuming Grade follows the standard understanding of course grade, meaning that A stands for the best performance and Z stands for the worst. Figure5 shows that Grade, Score, and Critical_flag are not correlated with each other; higher Score isn't indicative of better Grade; Critical_flag is not indicative of having a lower Score nor worse Grade.

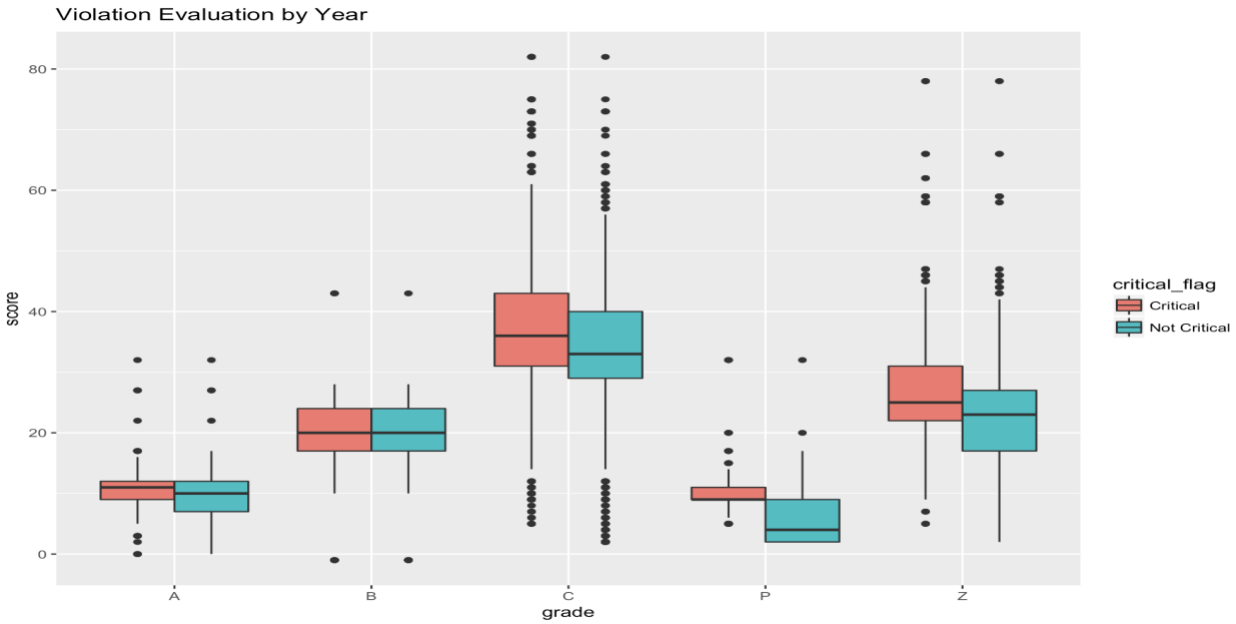


Figure5

Therefore, my next step is to figure out which measurement I should rely on more heavily to evaluate the level of violations, so I looked into the violation descriptions. This part is more subjective than I wanted. I selected two violation descriptions that I think are high-level violations: ID14 and ID16. I used boxplot to capture each violation's Grade, Score, and Critical_flag. My assumption is that ID14 and ID16 are critical violations, therefore should be given a fairly low Grade and Score. However, when the same type of violation occurred across different restaurants, the only consistent measurement is its Critical_level. In addition, regardless of the count and severity of violations, individual restaurant received the same Grade and Score on the same day. My hunch is that there are Critical-flag associated with each type of violation, but Grade and Score are given more subjectively depending on the inspectors.

If the interpretation of Grade and Score follows along with the standard understanding of their meanings, are the inspectors in certain borough more lenient or inconsistent with giving evaluations then? The following graph shows that assuming critical level is the benchmark and inspectors/grading rubric is the same within the same borough, almost all boroughs have pretty inconsistent gradings. When the evaluation is grouped by boroughs, we can see more consistency with Score and Grade (only A, B and C though). However, when a violation is deemed as critical it could be given any type of Grade and Score. Grade Z was only given starting from 2015 and onward. 2017 yielded the lowest level of variability in grading. In addition, there is more variability of Score for restaurants with the same Grade in the Bronx; there are also many

outliers in both Queens and Manhattan. Maybe Grade and Score are not associated entirely with the violation type; rather, it might have to do with the size, design or employee of the restaurant.

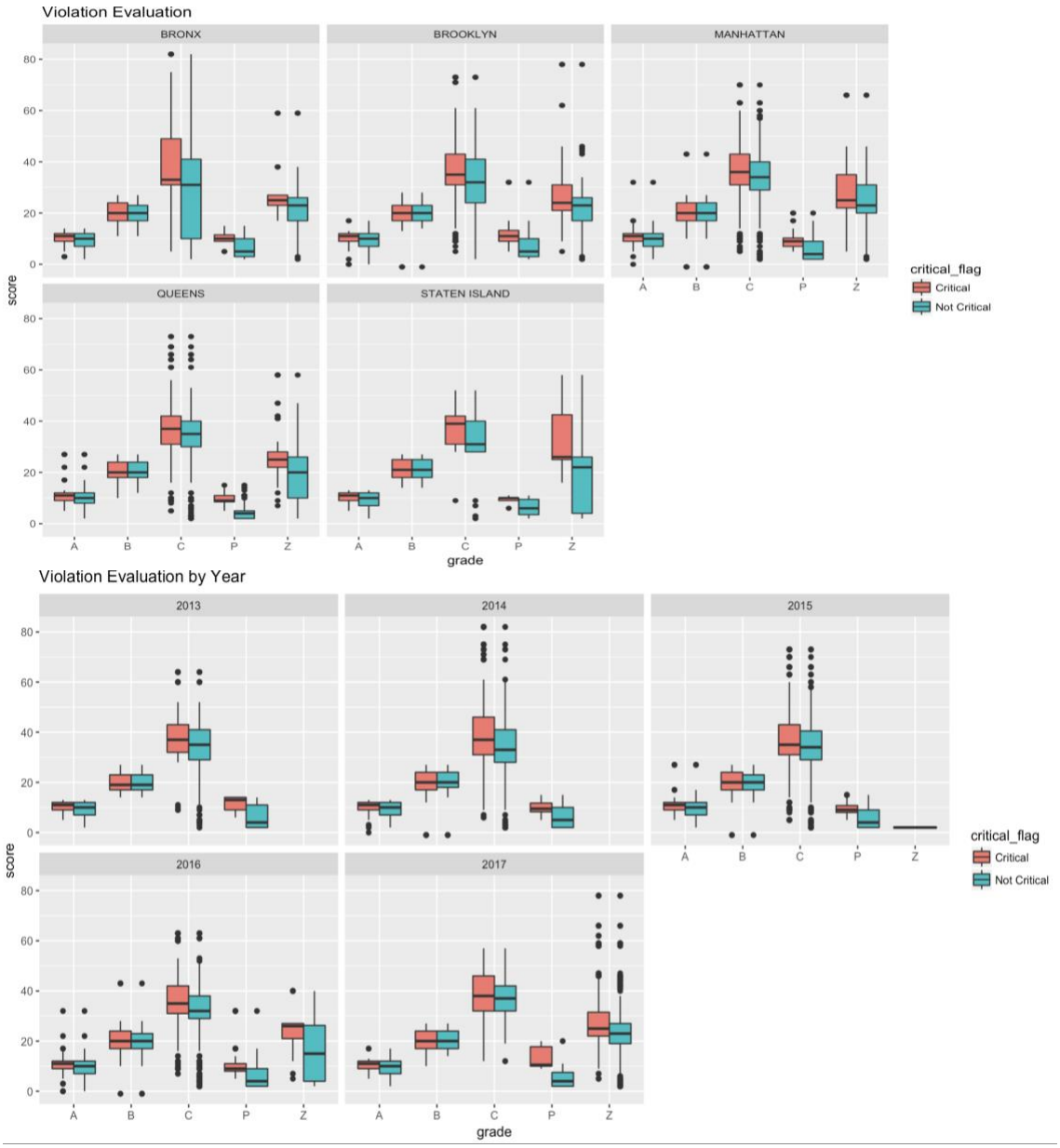


Figure6

Once I have determined which benchmark to use, I can then discover the spread of the critical violations in each borough. Figure7 shows that most of the restaurant receive less than 5 counts of violations, they can be marked as either critical or not critical. However, when the count of violations a restaurant receives is more than 10, almost all of them are marked as critical.

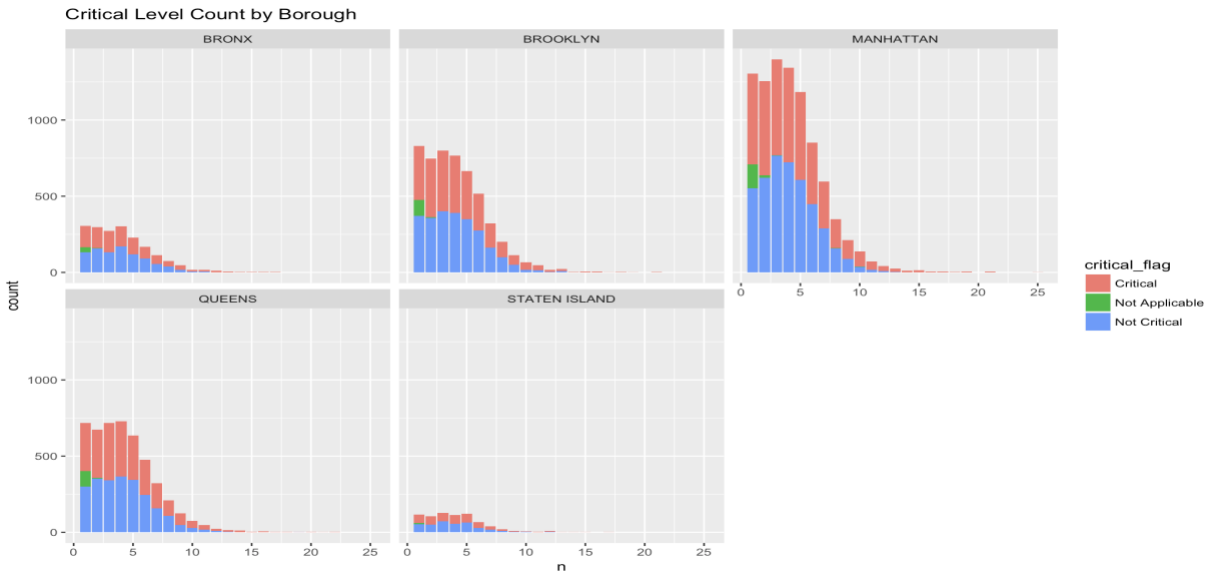
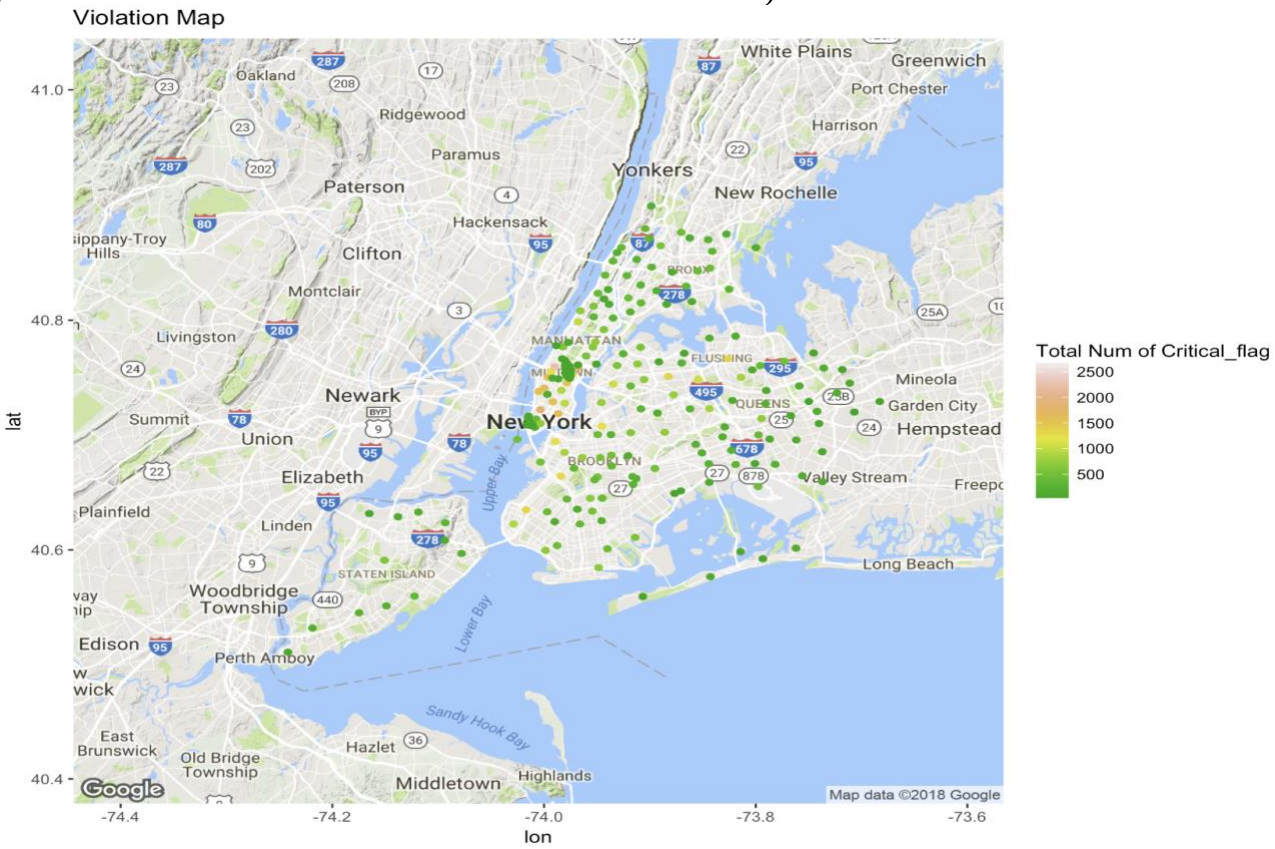


Figure7

I then used the information provided by the external file and plotted those restaurants on a map based on their longitude and latitude for visualization purpose. Since there are many restaurants in the same zip code with the same geo-location, I grouped the data by zip code to avoid overlapping of the dots that could be created otherwise. This following map would be most useful for customers when it can be made interactive, which I have attempted to achieve using Plotly. (when you hover your mouse over the zip code, the Plotly version of the map should give you the number of restaurants received violations in the area).



3). Recommendations

I focused heavily on the analysis of the three different evaluation index and found inconsistency. In addition, gradings are most inconsistent from 2014 to 2016. Assuming Grade and Score are given based on the violation type, then I would recommend a more thorough rubric to associate violation type with Grade and Score. However, as I mentioned before, due to the lack of information on the context of the grading, Grade and Score can also be influenced by the size of the restaurant, or other factors.

The recommendation I would give to customers is that although Manhattan has the most violations in this dataset, half of them are deemed as not critical, just like every other borough with less amount of violations; in addition, the amount of inspection on each restaurant isn't given, so I wouldn't simply avoid Manhattan when it comes to dining choices. Throughout my analysis, I also found that restaurants serve American food have the highest amount of both critical and non-critical violations across five boroughs, especially American chain restaurants, such as Starbucks and Subway.

Something else I would consider doing if I had more time is to bring in Yelp data and check whether there is a correlation between a restaurant's rating and its violation level; I think it would be very interesting! Another type of analysis I would do is to delve deeper into the duration of inspection on individual restaurants and see whether the grading has improved in the later times.

Appendix

borough_name	critical_flag	Max	Mean	Min
BRONX	Critical	17	4.345475	1
BROOKLYN	Critical	21	4.307236	1
MANHATTAN	Critical	25	4.491447	1
QUEENS	Critical	22	4.482612	1
STATEN ISLAND	Critical	17	4.180108	1

critical_flag	Max	Mean	Min
Not Critical	12	3.932755	1
Not Critical	13	4.045309	1
Not Critical	21	4.051157	1
Not Critical	19	4.200688	1
Not Critical	12	3.892265	1

borough_name	cuisine_description	Max
MANHATTAN	American	6224
BROOKLYN	American	2156
QUEENS	American	1887
BRONX	American	626
STATEN ISLAND	American	443