A Research into NYC Restaurants' Food Safety

By Yuan Yao (yy2884) and Yuming Wang (yw4054)

Abstract: NYC is a city with a huge population and diverse cultures. One result of this is the huge quantity of restaurants in NYC. On some important days (holiday, anniversary, friends gatherings), people would often dine in a restaurant instead of cooking on their own. And when choosing restaurants, an important criteria is food safety. Through this study, we tried to have a look into the food safety level in different NYC restaurants.

I. Data Understanding.

We found our data in NYC Open Data. The Department of Health and Mental Hygiene inspects the restaurants in NYC regularly, and this dataframe contains all their inspection data. It contains 383167 rows and 18 columns. The data in the dataframe can be generally divided into two types. One is the information used to help identify the restaurants. The other is the information used to help identify what kind of violation this restaurant has done.

The identifiers of the restaurants include: streets, names, boroughs, zipcode, cuisines, phone numbers, etc. we think the **boroughs**, the **cuisines**, and the **zipcode** are quite interesting to explore and instead of looking into every single restaurant, we will group our data by boroughs, cuisines, and zipcode and have a look at them as a whole.

To help consumers understand what kind of food safety rule this restaurant has violated, the Department of Health and Mental Hygiene has developed a thorough grading system. In this dataframe, there are columns including the inspection type, the record date, the action that's seen during this inspection, the critical flag, a violation code and the violation description. Among these columns, we think the violation code is the most useful one. Violation code is a combination of a number and a letter, the number gives a general idea of the kind of the violation, and the letter explains it further. Like 02C, 02 tells us that this is a violation that's related to temperature, and when illustrated further by the letter, we can know the violation is "Previously heated and cooled potentially hazardous hot food not reheated to 165°F for 15 seconds within 2 hours". All these specific violations can be found on *https://www1.nyc.gov/assets/doh/downloads/pdf/rii/ri-violation-penalty.pdf.* There are a bunch of violation codes, but we don't want to dig too deep into it. So we are just using **the number in the violation codes** to get a general idea of the violations. Besides, the **score** is the most important identifier of the food safety level of the restaurants. Thus we will certainly explore into it.

II. Data Preparation.

First, when looking at the inspection date, we found some data are quite old. There are even data from the 20th century, which is obviously no longer worth exploring. And other data are not all up to date enough, some are from 2014, some are from 2012. In order to keep our research result up to date, we decided to only keep the data that's after 2015-01-01.

Secondly, the original data in our dataframe is based on every single inspection. Some restaurants are inspected many times, and some are inspected just a few times. So we can't use the score of every single inspection, otherwise some restaurants that are of higher weight will influence the overall result. So we first deleted some irrelevant columns, and then calculated the average score based on CAMIS (an identifier for each restaurant) and made that a new column called "avg_score". Now, we have a new dataframe that will be useful in research of the scores of these restaurants.

	camis	name	boro	building	street	zipcode	type	avg_score
0	40931972	88 PALACE RESTAURANT	MANHATTAN	88	EAST BROADWAY	10002.0	Chinese	10.833333
1	40698807	MAMA MIA 44 SW	MANHATTAN	621	9 AVENUE	10036.0	Italian	13.416667
2	41348161	DELICATESSEN MACBAR	MANHATTAN	54	PRINCE STREET	10012.0	American	9.636364
3	41594717	AUNTIE ANNE'S PRETZELS	BROOKLYN	625	ATLANTIC AVENUE	11217.0	Nuts/Confectionary	6.500000
4	50004809	BAR BACON	MANHATTAN	836	9TH AVE	10019.0	American	11.500000
5	50033575	JACQUES TORRES ICE CREAM	MANHATTAN	89	E 42 ST	10017.0	Ice Cream, Gelato, Yogurt, Ices	14.625000
6	50013014	BE JUICE	MANHATTAN	93	3RD AVE	10003.0	Juice, Smoothies, Fruit Salads	28.166667
7	40401002	JOHNNY MACK'S BAR	BROOKLYN	1114	8 AVENUE	11215.0	American	10.428571
8	50042162	MILE 17	MANHATTAN	1446	1ST AVE	10021.0	American	10.333333
9	50060617	POQUITO PICANT	BROOKLYN	497	ATLANTIC AVE	11217.0	Mexican	22.375000
10	50049945	NEW SUKI SUSHI	BROOKLYN	9208	3RD AVE	11209.0	Japanese	10.750000

Dataframe1: with a new column named 'avg_score'

When calculating the average score, we deleted all the columns that might be different for the same restaurant, so that we can delete the duplicates. But we still want to look at the violation type. So we created new dataframe that contains a column named "violation type" with only the numerical part of the violation codes.

	camis	name	boro	building	street	zipcode	phone	type	inspection date	action	violation description	critical flag	score	grade	grade date	record date	inspection type	violation type
0	40931972	88 PALACE RESTAURANT	MANHATTAN	88	EAST BROADWAY	10002.0	2129418886	Chinese	03/04/2016	Violations were cited in the following area(s).	Pesticide use not in accordance with label or ...	Not Critical	8.0	A	2016-03-04	11/26/2018	Cycle Inspection / Initial Inspection	08
1	40698807	MAMA MIA 44 SW	MANHATTAN	621	9 AVENUE	10036.0	2123154582	Italian	04/25/2018	Violations were cited in the following area(s).	Food contact surface not properly washed, rins...	Critical	10.0	A	2018-04-25	11/26/2018	Cycle Inspection / Initial Inspection	06
2	41348161	DELICATESSEN MACBAR	MANHATTAN	54	PRINCE STREET	10012.0	2122260211	American	09/27/2018	Violations were cited in the following	Non-food contact surface improperly constructe...	Not Critical	9.0	A	2018-09-27	11/26/2018	Cycle Inspection / Re-inspection	10

Dataframe2:with a new column named 'violation type'

III. Behind The Scores.

The most objective and clearest indicator of the food safety of a restaurant is its score. According to the grading system developed by the NYC Health Department, when inspecting these restaurants, every

violation is associated with a certain number of points, and the restaurant's inspection score is the total points at the end of the inspection. The lower the score, the better. According to the NYC Health Department, restaurants with a score between 0 and 13 points earn an A, those with 14 to 27 points receive a B and those with 28 or more a C.

1. Is NYC a safe place to eat?

We calculated the overall average score for all the restaurants in NYC, and we calculated the number of restaurants with a grade of A, B, C. The results are quite pleasant: the overall average score for NYC restaurant is 11.26, that is an A. there are 20445 restaurants with an A, 4360 restaurants with a B, only 464 restaurants get an C. So the overall food safety level in NYC is rather high, generally when dining in a restaurant the residents don't need to worry too much about the food safety problem.

2. Which borough to eat in?

There are five boroughs in the New York City: Manhattan, Brooklyn, Queens, Bronx, and Staten Island.

We want to explore the distribution of restaurant inspection scores across boroughs, hoping to find some relationship between the food safety of a NYC restaurant and the borough in which it is located. From our perspective, the average score of the restaurant within a borough can best reflect the overall food safety level of that specific borough. Therefore, we grouped our data by borough, and took the mean of the scores. To make the results more intuitive, we plotted the average scores, and we ended up with a chart like this.

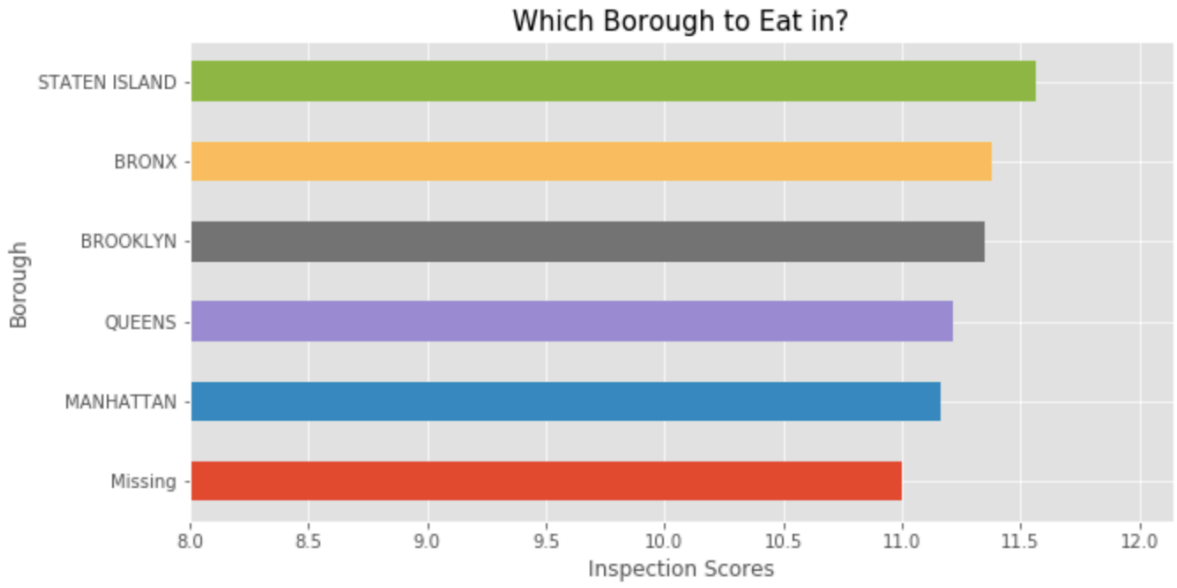


Chart1: which borough to eat in?

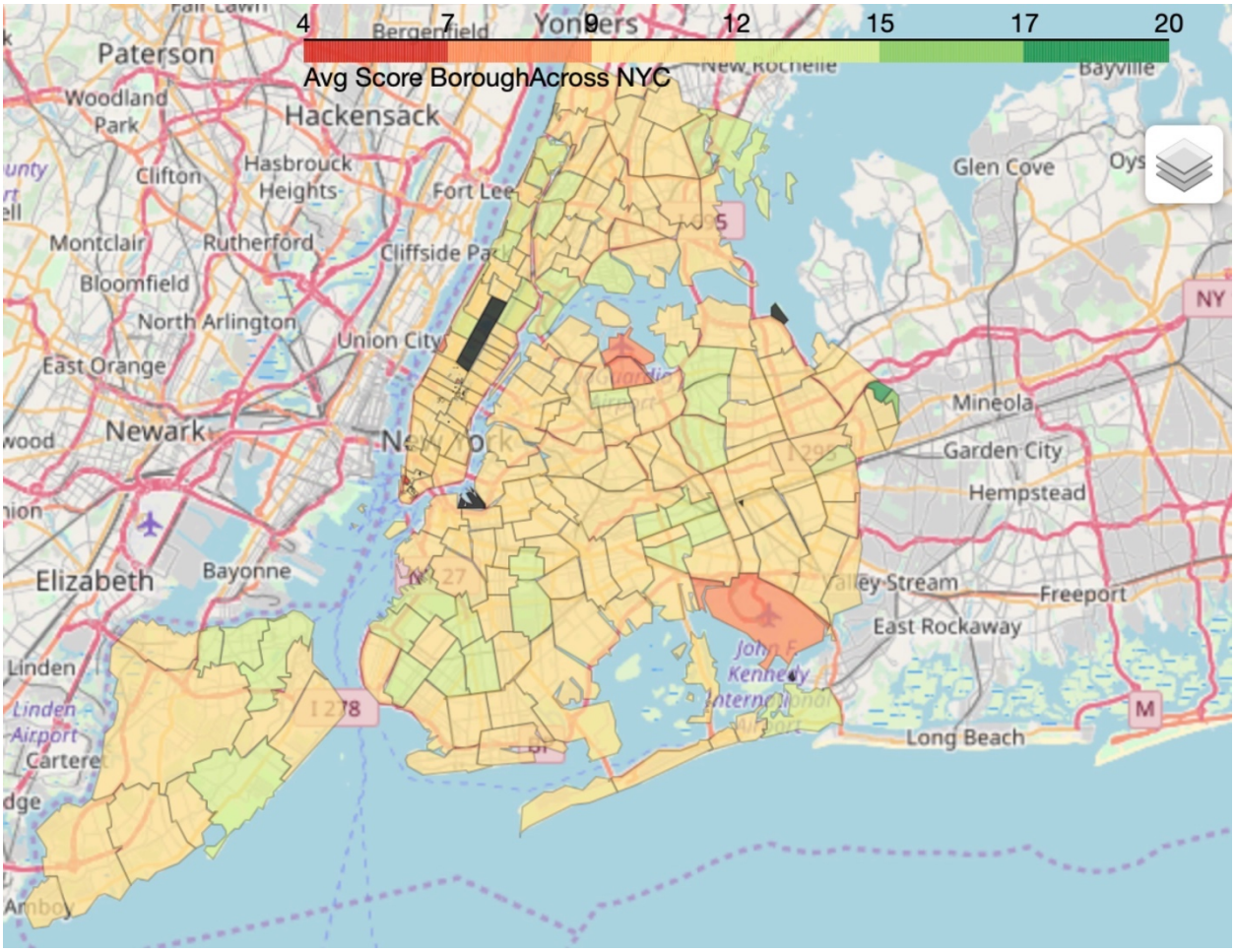
As is shown in the bar chart, overall food safety of the five boroughs are basically at the same level, with Staten Island having the highest score (which indicates more violations), and Manhattan having the lowest. However, the differences are just marginal.

Well, it seems hard to find solid evidence supporting differences between boroughs with regard to food safety. We presume that this is because restaurants in these five boroughs of the New York City are basically competing in the same market and supervised by the same authorities. These common factors contribute to there being no significant differences concerning food safety among New York restaurants.

3. Zipcode and Scores

Now, since there are no significant differences between different boroughs, we might as well dig deeper into different boroughs and try to see in each borough, which parts have the best(worst) food quality.

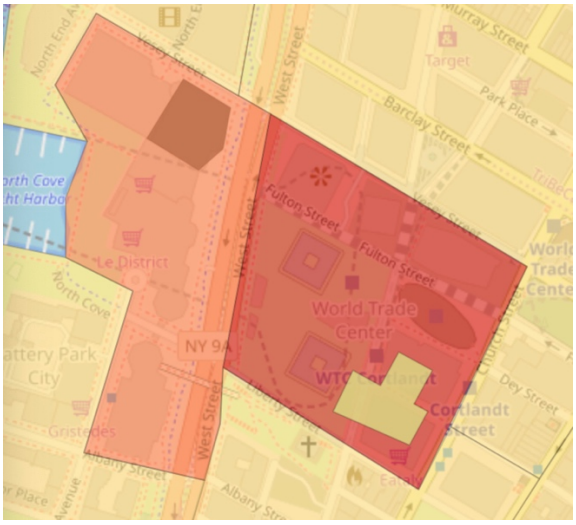
We grouped these restaurants by their zipcode and we calculated the average score within different zipcode zones. Then we plotted it out in the map. The greener the worse, the redder the better.



Graph1: restaurant score distribution in NYC (dynamic map in our jupyter notebook)

After analyzing this map, we have the following discoveries:

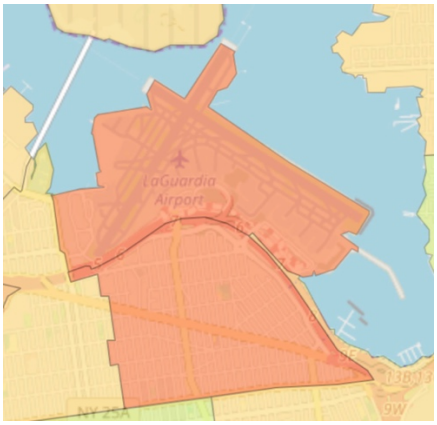
(1) Downtown Manhattan, with no zipcode zone plotted in green, has higher general food safety level than midtown and uptown. The best performing zipcode zones are the two zones around World Trade Center. They are the only two red zones in downtown Manhattan. World Trade Center did significantly better than the other zones.



Graph2: best performing zones around WTC

(2) In Manhattan, Bronx and Staten Island, the underperforming restaurants are all near the water.

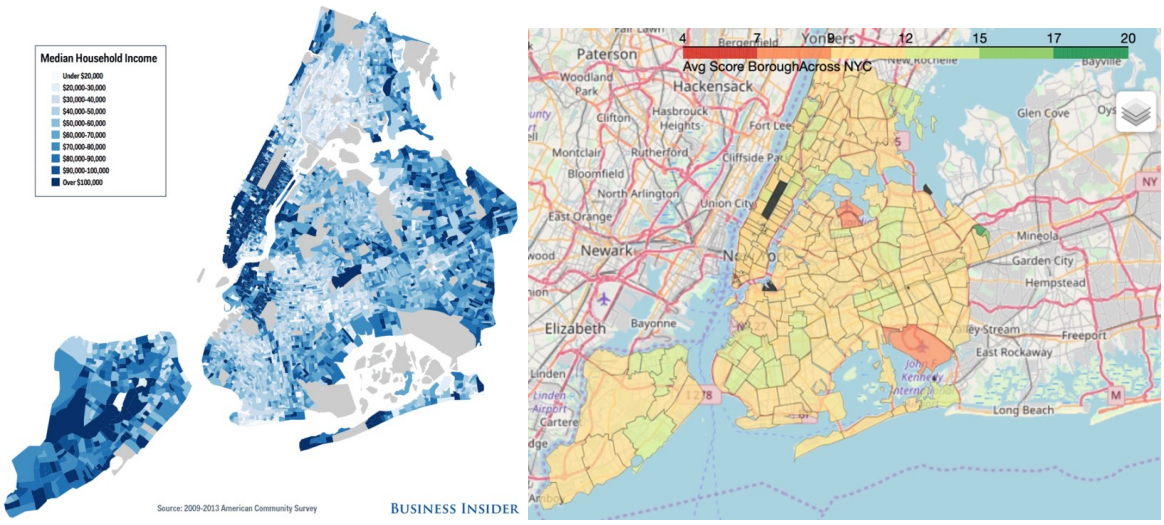
(3) In the areas surrounding the two airports in NYC (LGA, JFK), the food quality is higher than the other areas.



Graph3: LGA

Graphs4: JFK

(4) comparing our map with the medium household income distribution within NYC, provided by Business insider, we found that districts with a higher income level tend to have restaurants of higher food safety level.



Graph5: comparison between medium income and food quality

4. Cuisine and the Scores

What about the type of cuisines? Would the kind of food served by the restaurants have anything to do with food safety? That sounds an interesting aspect to probe into. Before we started, we had a look at how many cuisine types are there and how many restaurants each type has, we found that there are a total of 85 different types of cuisines in our dataset, and that would be too much for us to cope with.

(1) Best performing and worst performing cuisine types.

Some of our cuisines only have a few restaurants around NYC, they wouldn't be representative. So we decided to delete those cuisines with less than 50 restaurants, and then we sorted them according to the average score, from small to big (best performing to worst performing). Now we see, the best performing type of restaurants is the Donuts, with an average score of only 8.64, despite their huge inspection numbers. And the worst performing one is Peruvian, with a score of 14.65. Let's have a closer look at those two types.

First we looked at the donut type. After counting the names, we found that a large part of the donut restaurants is Dunkin' Donuts. So we decided to separate the Dunkin' Donuts restaurants and the non-Dunkin' Donuts restaurants and have a closer look at it.

count	505.000000
mean	8.607010
std	2.990381
min	0.000000
25%	6.857143
50%	8.666667
75%	10.142857
max	24.500000

Dunkin' Donuts

count	25.000000
mean	9.249434
std	3.392404
min	2.500000
25%	7.222222
50%	9.875000
75%	11.285714
max	15.952381

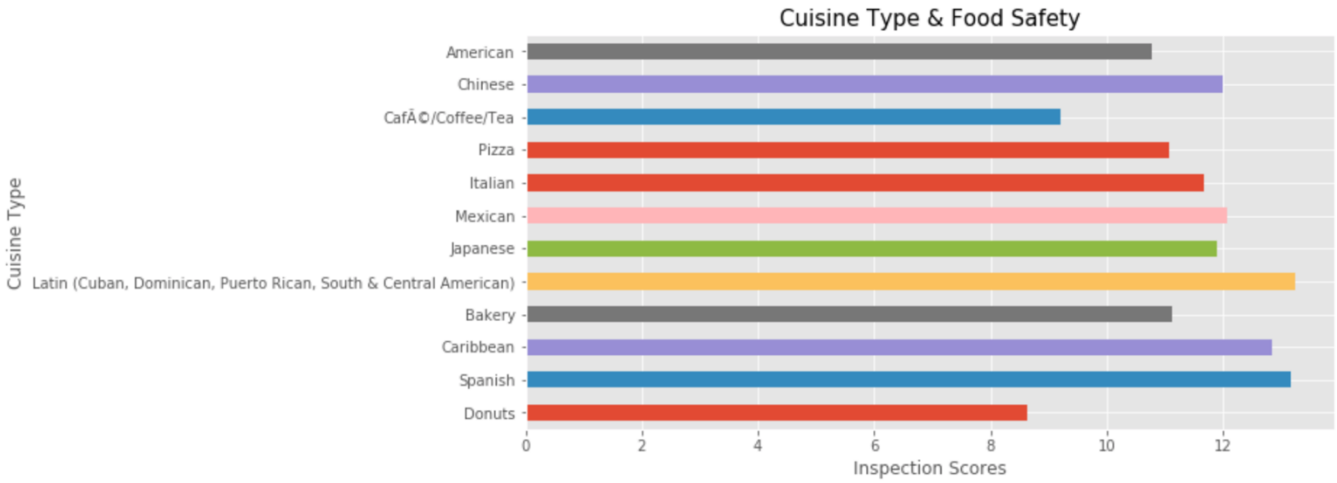
non-Dunkin' Donuts

Among the 530 donuts restaurants, more than 95% are Dunkin' Donuts. And their average score (8.61) is significantly lower than the average score of other donut restaurants (9.24). It is quite impressive that a chained store with so many branches can perform so well. So, when you need to grab something to eat, choose Dunkin' Donuts, it wouldn't let you down.

As for the Peruvian restaurants, after some basic analysis, we think it's not as interesting to dig into as the Dunkin' Donuts. So we just left it there.

(2) Major cuisine types

Next, we are going to have a look at the main cuisine types in NYC. We selected the cuisine types with more than 500 restaurants, that leaves American, Chinese, Café, Pizza, Italian, Mexican, Japanese, Latin, Bakery, Caribbean, Spanish and Donuts. As we did before, we plotted it out in a bar chart:



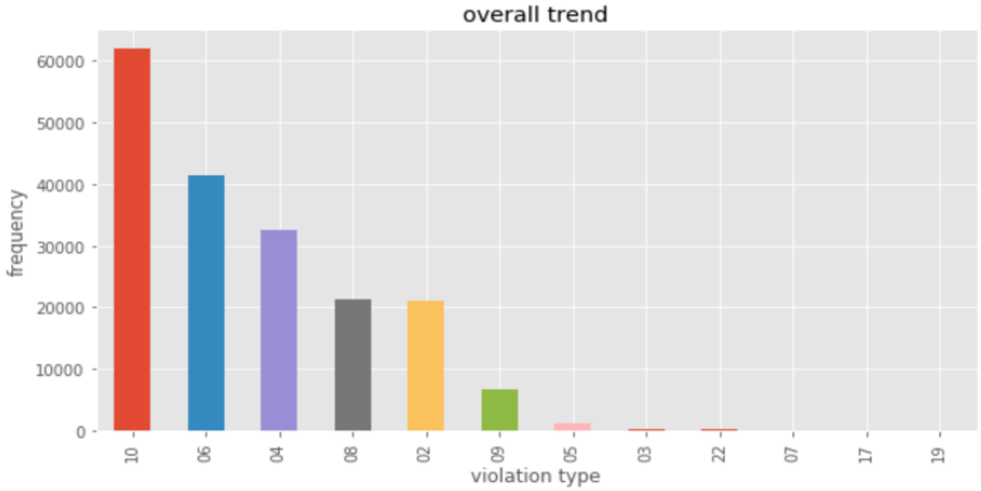
One thing we can tell immediately is that the number of restaurants has no significant relationship with their scores. Among those major cuisines, the best performing one is still donuts, while the worst performing one is the Latin. Only two types of major cuisines (Latin & Spanish) have their average scores over 13 (Grade B), Café and Donuts have significantly lower average scores than other cuisines. And based on those trends and numbers we had an assumption: the cuisine that takes a long procedure to make is more likely to have a higher score, a lower grade, while foods that can be made fairly quickly wouldn't have many chances to violate the food safety law. Thus their scores are usually lower.

IV. Behind the Violation Types

During the inspection of those restaurants, NYC Health Department have given different restaurants their violation codes to help consumers see what kind of inappropriate action those restaurants have made. In total, there are 99 different violation codes. But as mentioned before, we didn't want to dig too

deep into this. So we decided to use just the numerical part as indicators of what kind of violation it is. Violation types could also tell us something interesting. First, let's have a look at the overall trend.

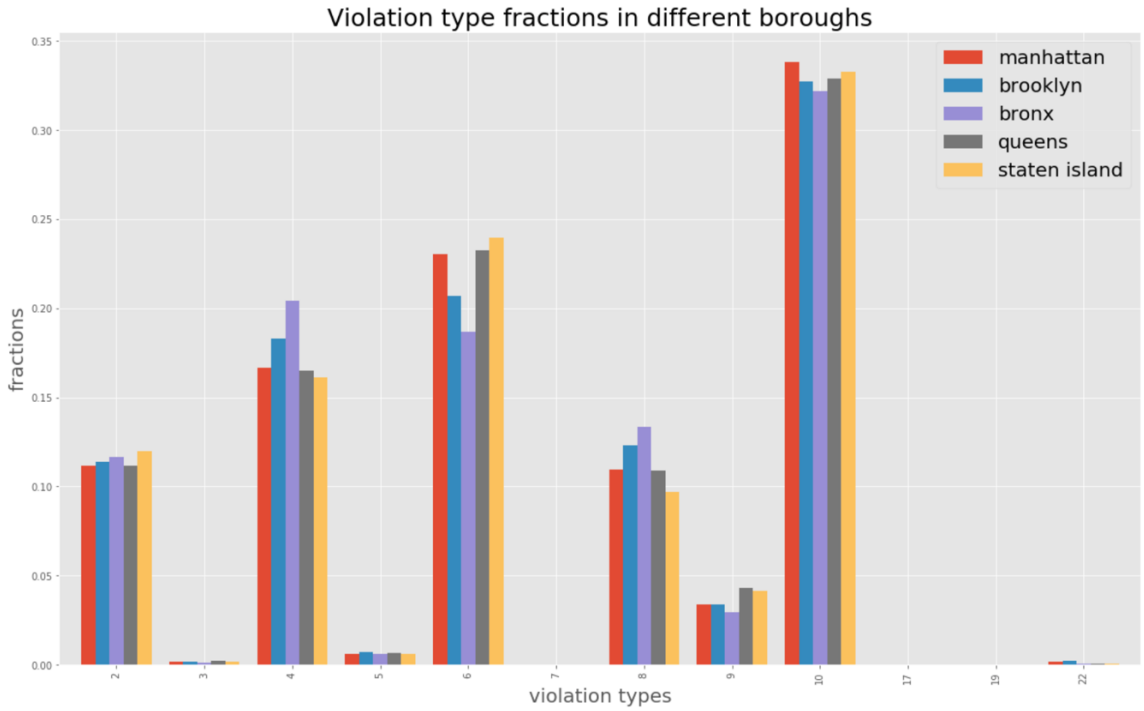
1. overall trend



Overall, the violation type appeared most is 10 (dining environment), that is over 21000 violations more than the second most: 06 (workers), almost the size of the fourth and fifth violation code: 08 (pest) and 02 (temperature). 04 (contamination of food) is worth noticing as well, more than 30000 times of violation. Apart from these, other violations occur rather infrequently, with less than 10000 violations.

2. Within the boroughs.

In order to wipe out the influence of different inspection times, we calculated the fraction of violation types within different boroughs. And then we plotted them out in the same chart



The major violation type is 10 for all 5 boroughs. There are no distinguishable difference between the 5 boroughs can be found in type 2, 9, 10. For the other violation types, Bronx has either the highest fraction or the lowest. Bronx suffers more from food contamination and pests.

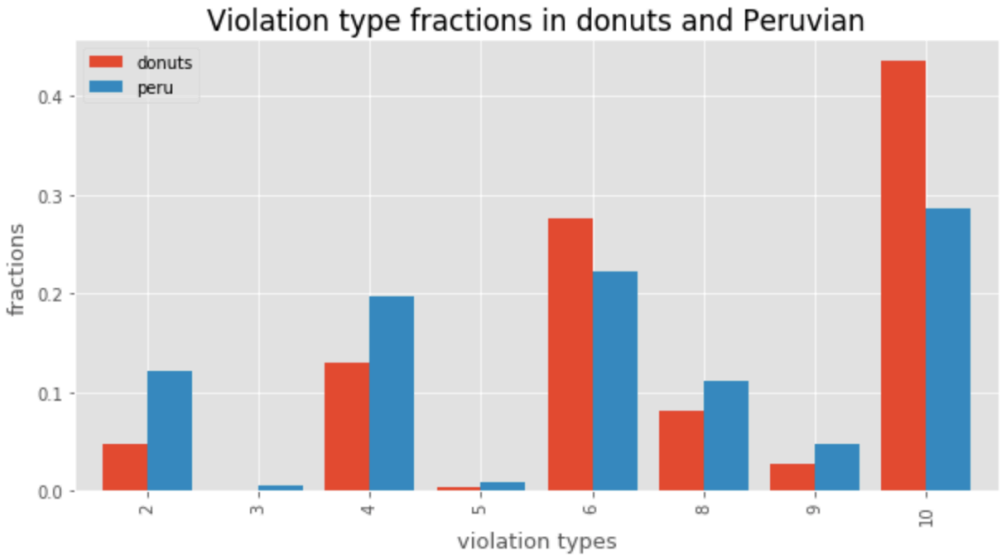
3. Cuisine types and violation types.

Will violation types have anything to do with the cuisine types?

Hypothetically, we think they are related. For example, for Japanese restaurants, an important part of their food is sushi, thus it's less likely that they are going to violate rules with regards to temperature. Or, some Indian dishes might require a lot of manual work, while American restaurants and Cafe may just have a short cooking procedure. Thus Indian restaurants might have more violations towards workers.

(1) Best performing and worst performing ones

To justify those assumptions, we need to look at different cuisines individually. Of course we are not going to look at every single cuisine type. First let's have a look at the best performing one: donut, and the worst performing one: Peruvian.



As we can see in the bar chart above, donuts, our best performing cuisine type, has its violations concentrated in 10, restaurant environment. That is 16% larger than the second most violated: 6, workers. 06 is 15% higher than the third one, 04, contamination of foods. So we can see the violation types for donuts are quite concentrated. On the contrary, Peruvian restaurants' violation types are relatively evenly distributed.

(2) 4 major cuisine types.

After looking at the best performing and worst performing ones let's look at some of the largest cuisines. We decided to look at the 4 cuisines with the largest number of restaurants: American (5837), Chinese (2375), Café (1677), and Pizza (1176).

The conclusions are as follows:

(a) 10 is still the highest within the 4 major cuisines.

(b) Surprisingly, café, where most people go for their comfortable environment, ended up with almost half of their violations related to environment.

(c) Pizza has its violation types evenly distributed.

(d) This conclusion comes from all the analysis we have done before: food contamination (04) and pests (08) tend to come together. The ones with highest fraction in pest tend to suffer more from food contamination.

V. Trends in Restaurant Scores

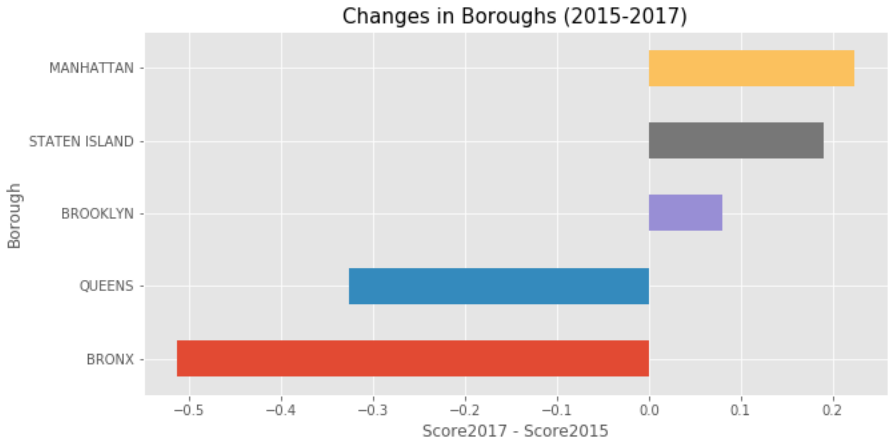
Food safety of the restaurants is changing over time, and so do the restaurant scores. In addition to studying the restaurant scores in a static way, we also want to employ a more dynamic perspective: we are curious about the trends in food safety. Have the restaurants become safer over the years? If so, which borough has experienced the greatest improvement? Which type of food has seen the most significant change? These are the topics that we are going to discuss in the following coding experiments.

In order to do so, we focus our attention on the data collected in 2015 and 2017. By taking the difference between these two years, we can find out the changes among all the restaurants that existed in both years.

1. Overall trend.

Over the 3 years, the overall restaurant score has decreased by 0.015, from 10.584 in 2015 to 10.569 in 2017. The difference is not worth noticing. We can say that the overall food safety level of NYC remains stable.

2. Within boroughs.



The bar chart tells us that over the two years, food safety improved in Bronx and Queens, while deteriorated in Manhattan, Staten Island and Brooklyn. The large leap in the Queens and Bronx offset the deterioration in the other 3 boroughs, making NYC as a whole enjoy a slight improvement in food safety.

3. Trends within zipcode zones.

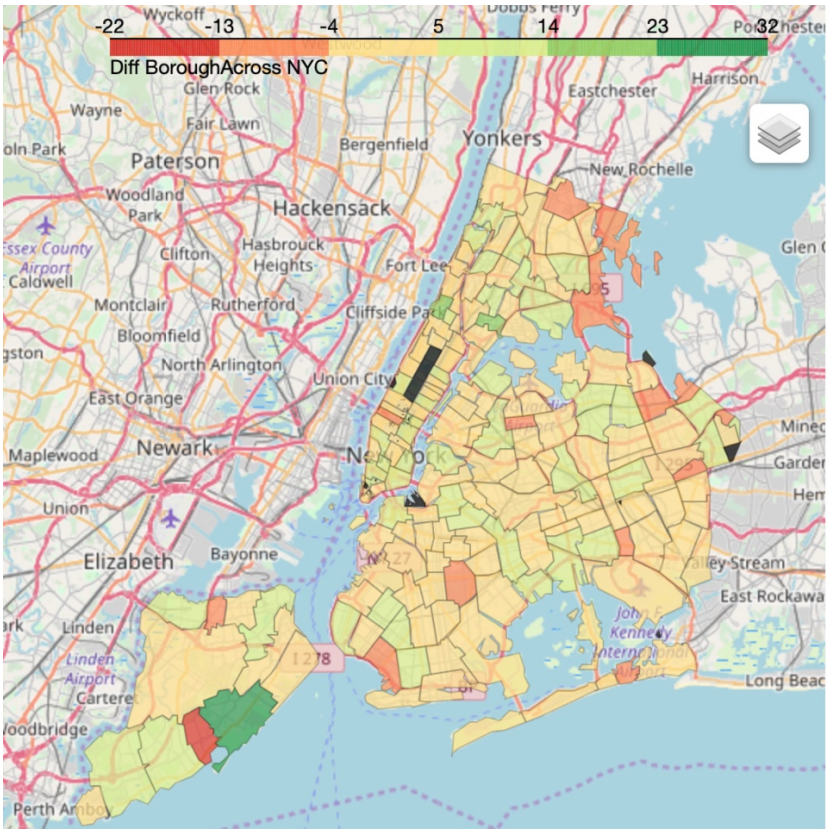


Chart6: overtime trend and zipcode zones.(dynamic map in our jupyter notebook)

There are several interesting findings that we can get from the map:

(1) The majority of the restaurants have no difference in the scores across the years.

(2) The east part of Bronx saw a huge improvement in food safety.

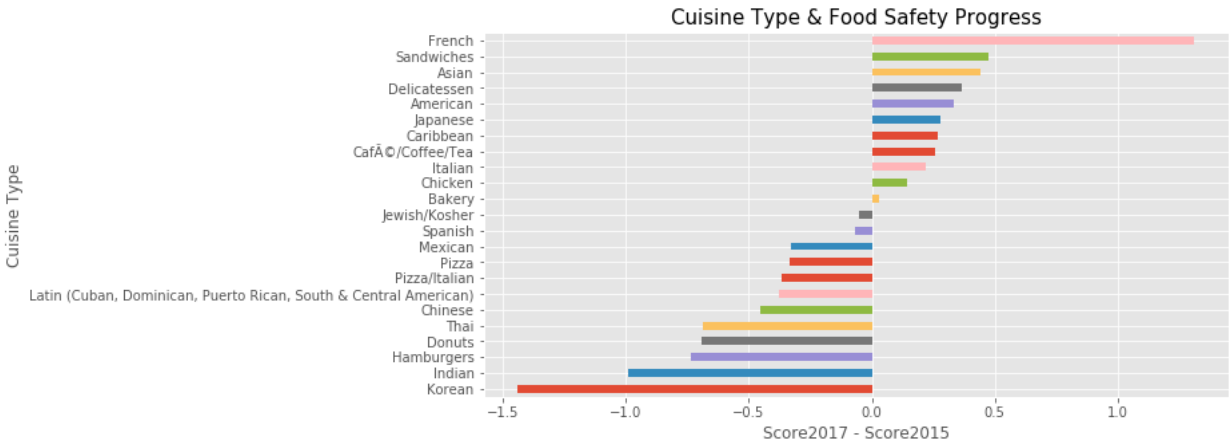
(3) Two adjacent zipcode zones on the southern coast of Staten Island displayed the greatest improvement and the largest deterioration in restaurant scores.

(4) The two airports areas both remained the same level.

4. Which cuisine experienced the greatest progress over the years?

Just as what we have done in the previous parts of the project, we are eager to see the changes of food safety among different kinds of cuisines. After all, the production procedure, key ingredients, as well as the equipment involved can vary across different food types, and these may contribute to different changes in food safety.

To start with, we created a dataframe with the differences of restaurant scores between 2015 and 2017, and sorted the values for further analysis. Then we selected the cuisines with more than 100 restaurants and plotted it out.



Wow! Beautiful!

The differences are significant here. Around half of the main cuisine types experienced food safety improvements, while the other half suffered different levels of deterioration.

It is French and Korean that attracted most of our attention: there was a 1.3-point-increase in the overall restaurant scores of French type, and a 1.4 points decrease in that of Korean. What happened to them?

We now extract French and Korean from our dataframe to take a closer look at them:

count	116.000000
mean	9.722024
std	5.285862
min	2.000000
25%	7.000000
50%	9.000000
75%	12.000000
max	38.000000

French restaurants

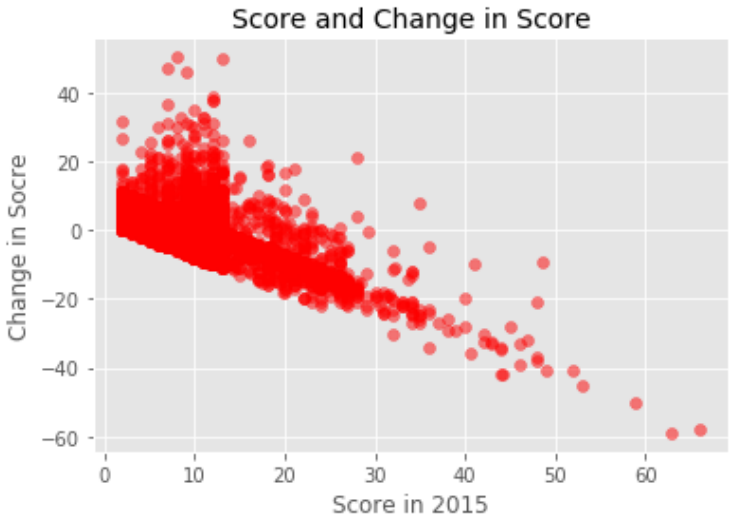
count	99.000000
mean	12.787142
std	7.561093
min	2.000000
25%	8.750000
50%	11.000000
75%	13.000000
max	48.000000

Korean restaurants

OK, it seems to us that the restaurants are following an "approaching the average" trend. French restaurants, which started with a relatively lower score (9.72) in 2015, had its score increase to 11.03, while Korean restaurants, started with a relatively higher score (12.79) in 2015, had its score decrease to 11.35, making these two types of restaurants end up approximately the same level in 2017.

5. Trend for all restaurants

We want to see if this trend applies to all of the restaurants in our dataframe, so we decided to plot it out in a scatter plot. And we ended up with a scatter plot like this:



There is indeed a relationship between the score. We ran a regression to quantify the relationship.

OLS Regression Results

Dep. Variable:	diff	R-squared:	0.447
Model:	OLS	Adj. R-squared:	0.447
Method:	Least Squares	F-statistic:	5491.
Date:	Wed, 12 Dec 2018	Prob (F-statistic):	0.00
Time:	10:50:39	Log-Likelihood:	-20743.
No. Observations:	6790	AIC:	4.149e+04
Df Residuals:	6788	BIC:	4.150e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.9223	0.136	65.722	0.000	8.656	9.188
score2015	-0.8444	0.011	-74.104	0.000	-0.867	-0.822

Omnibus:	3613.012	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	45624.717
Skew:	2.267	Prob(JB):	0.00
Kurtosis:	14.862	Cond. No.	26.1

We can conclude from the regression that the change in score and the score in 2015 followed a negative relationship: every 1 point increase in the 2015 restaurant score was associated with a 0.8444 decrease in the score changes.

VI. Conclusion

We all need to eat and food is important. After our research, we can conclude that food safely is related to both the cuisine type and the affluence of the area. So it might seem wise to eat donuts, ice cream or sandwiches in a rich area. But as the regression has shown, the restaurants' food quality in NYC tend to

converge to the average. And the differences are not significant. So maybe, we don't need to care too much about food safety when choosing the restaurant. Just try new things out and enjoy your life!

Github link for our codes: https://github.com/YvonneYao233/Data_Bootcamp_final_project