CS-GY 6513 Big Data Project Report

Spatial Profiling for NYC Inspection Data and Airbnb Data

New York University

Prof. Juliana Freire

Fall 2020

Group 5

Annan Zhang, Qiang Cui, Tianying Zhang

https://github.com/TianyingTina/Big-Data-Project

CS-GY 6513 Big Data Project Report

Spatial Profiling for NYC Inspection Data and Airbnb Data

Introduction

Spatial analysis or spatial statistics includes any of the formal techniques which studies entities using their topological, geometric, or geographic properties. In our project, we not only studied the two data sets themselves, making certain analysis. We also visualized our data through the integration of data and spatial graphs. We want to study the relationship between Restaurant Inspection, Airbnb price with geometry.

In our project, we get the datasets from NYC Open Data. We applied pandas and OpenRefine to preprocess the data, using geopandas and matplotlib to visualize the data. After we finished understanding the data and visualizing them. We proposed a hypothesis: The restaurant inspection grade and space are related. The AirBnB price is related to restaurant grade. We use SVM, kNN, k-means clustering, linear regression to build models to see if there exist any relationship. In the k-means clustering method, we can see that geometry and restaurant grade have a relationship. In the logarithmic regression, we can see a strong relationship between restaurant grade and Airbnb price.

Problem Formulation

There are a lot of restaurants and airbnb rooms in New York City. Now we have two datasets corresponding to the two types of data we want. One is New York City Restaurant Inspection Results, the other is 2019 Airbnb Rooms Data. At first glance, it seems that they have no relationship. It is because we never consider the geometry(space) as an attribute.But we want to know is there any relationship between restaurant and space. Is there any relationship between airbnb rooms and space? If both are yes, we can find the relationship between restaurants and airbnb rooms. So our goal is to find the relationship among airbnb rooms, restaurants and space. We propose the hypothesis: restaurant inspection grade, airbnb room price and space are related.

Background & Related Work

In this semester, we learned a lot of knowledge about how to process big data. In week 9, we first touched the spatial analysis. We have some experience in analyzing data without the spatial data. But this is our first time to study a dataset including the geometry data. Not only visualizing the data on a map, but also analyzing the relationship based on space.

We also read and get some inspiration from the notebook in Kaggle. This is a notebook which analyses Airbnb data merely.

https://www.kaggle.com/geowiz34/maps-of-nyc-airbnbs-with-python/notebook#Yankee-Stadium-Maps-&-Analysis

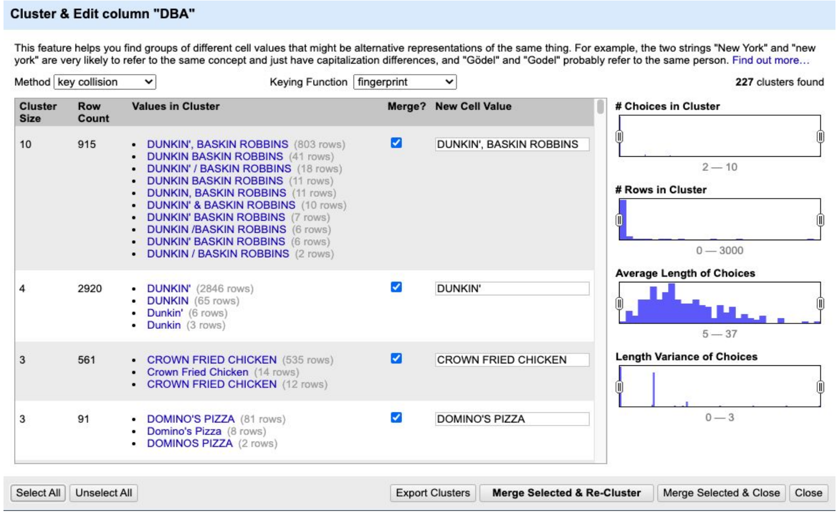
Methodology

1. Preprocessing Data:

Frist, we preprocessed the data. You can find the whole process in our github.

We have 4 datasets: New York Restaurant Inspection Results, Airbnb rooms.

During the preprocessing, we first applied OpenRefine to manipulate the data. The advantage of OpenRefine is that it is really straightforward. For example, in the graph on the left, we can quickly cluster and merge the similar words. This might be because of typo or different abbreviations.



Cluster & Edit column "DBA"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more…

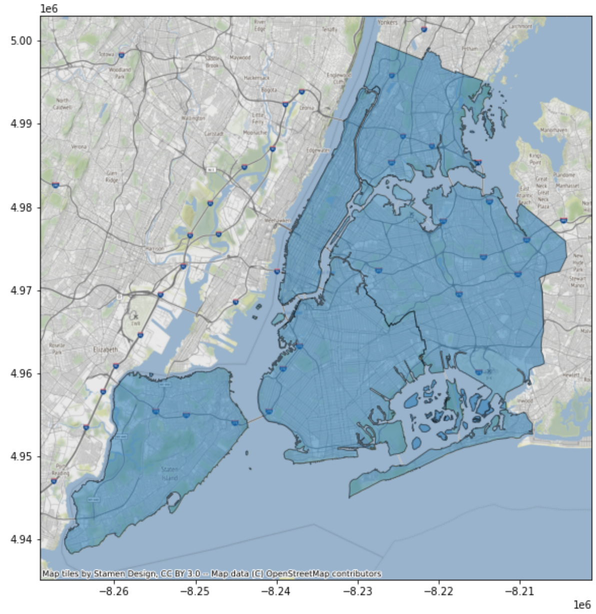
Method: key collision   Keying Function: fingerprint   227 clusters found

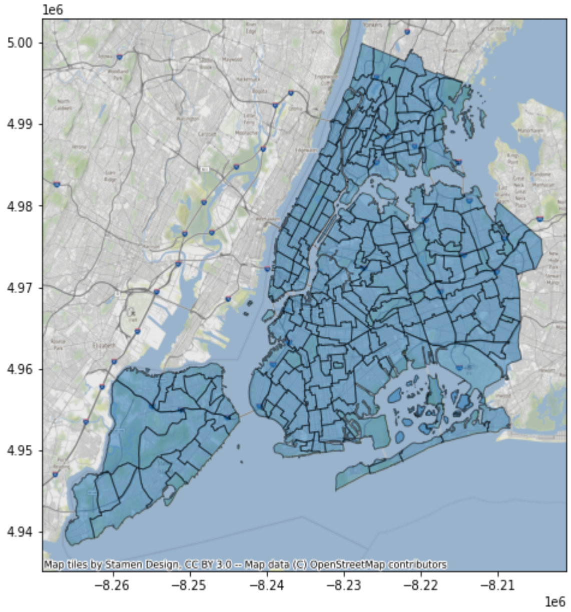
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
10	915	• DUNKIN', BASKIN ROBBINS (803 rows) • DUNKIN BASKIN ROBBINS (41 rows) • DUNKIN' / BASKIN ROBBINS (18 rows) • DUNKIN BASKIN ROBBINS (11 rows) • DUNKIN, BASKIN ROBBINS (11 rows) • DUNKIN' & BASKIN ROBBINS (10 rows) • DUNKIN' BASKIN ROBBINS (7 rows) • DUNKIN /BASKIN ROBBINS (6 rows) • DUNKIN' BASKIN ROBBINS (6 rows) • DUNKIN / BASKIN ROBBINS (2 rows)	☑	DUNKIN', BASKIN ROBBINS
4	2920	• DUNKIN' (2846 rows) • DUNKIN (65 rows) • Dunkin' (6 rows) • Dunkin (3 rows)	☑	DUNKIN'
3	561	• CROWN FRIED CHICKEN (535 rows) • Crown Fried Chicken (14 rows) • CROWN FRIED CHICKEN (12 rows)	☑	CROWN FRIED CHICKEN
3	91	• DOMINO'S PIZZA (81 rows) • Domino's Pizza (8 rows) • DOMINOS PIZZA (2 rows)	☑	DOMINO'S PIZZA

Not only the cluster, we also use facet function to help us to see the distribution of different restaurant grades.
Now we apply the same step on all of the 4 datasets. Fortunately, the other 3 datasets are very clean. The data quality(percentage of usage) of the restaurant dataset is 201190/400046 = 50.29%.

2. Boro and Neighbourhood Dataset visualization

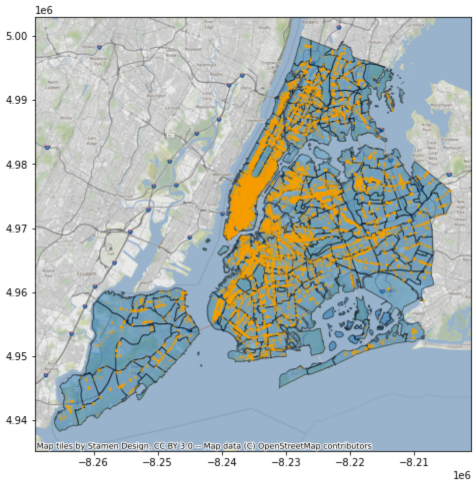
We use geopanda to visualize these two datasets. We first choose Borough dataset because this is an attribute in Restaurant dataset. But this border is too large and may not reflect what we want exactly, so we add a neighborhood map to help us see the data more clearly.

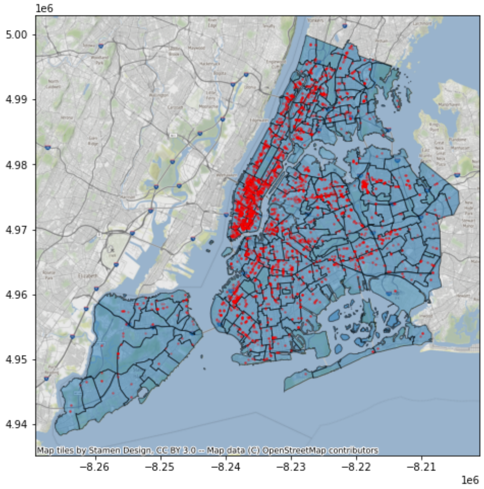

1e6
5.00
4.99
4.98
4.97
4.96
4.95
4.94
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
−8.26
−8.25
−8.24
−8.23
−8.22
−8.21
1e6


1e6
5.00
4.99
4.98
4.97
4.96
4.95
4.94
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
−8.26
−8.25
−8.24
−8.23
−8.22
−8.21
1e6

Now we have two maps with neighborhoods and boroughs.

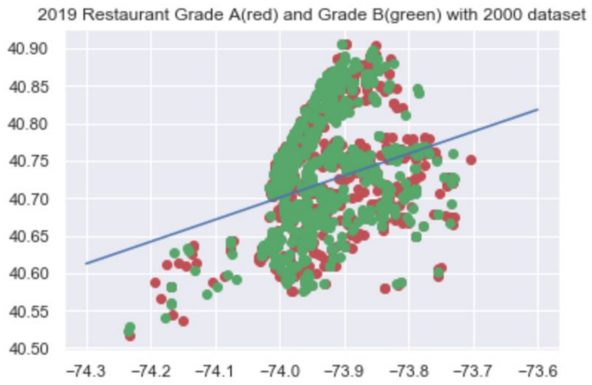
3. Spatial Analysis

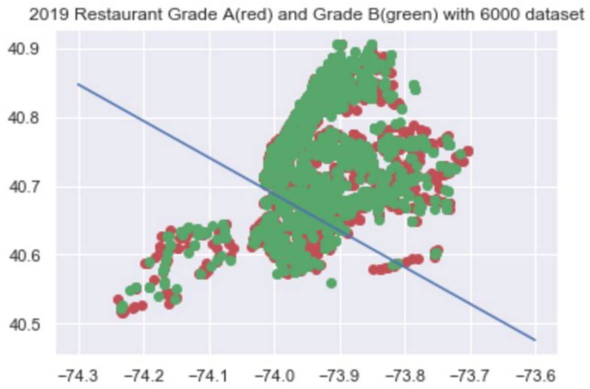

1e6
5.00
4.99
4.98
4.97
4.96
4.95
4.94
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
−8.26
−8.25
−8.24
−8.23
−8.22
−8.21
1e6


1e6
5.00
4.99
4.98
4.97
4.96
4.95
4.94
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
−8.26
−8.25
−8.24
−8.23
−8.22
−8.21
1e6

In this part, we first visualized the restaurants with grade A and B in 2019, we want to get some relationship between these two graphs. According to the graph, It seems that the distribution of both grading restaurants are very similar.
We still need some solid models to prove our intuition based on the two graphs.

3.1 SVM Model


2019 Restaurant Grade A(red) and Grade B(green) with 2000 dataset

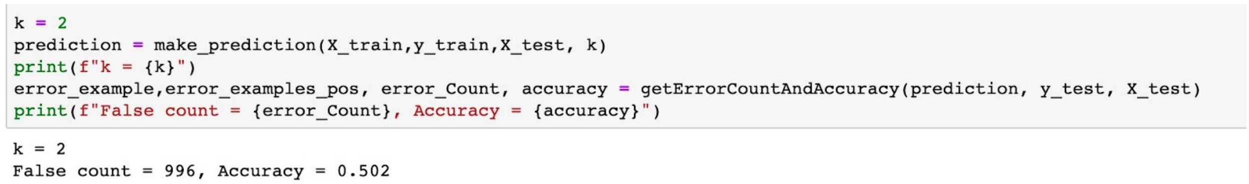

2019 Restaurant Grade A(red) and Grade B(green) with 6000 dataset

SVM is better used to classify several groups of data. By applying the SVM model on the dataset to grade A and B restaurants, we can plot a decision boundary for classification.
For the graphs above, we first plot 2019 Grade A and Grade B restaurants with 1000 data for respectively. The decision boundary makes sense because it roughly separates the Manhattan area and the Brooklyn area. However, when we increased the data size from 2000 to 6000, the line changed significantly, even the slope

changed from positive to negative. What we conclude is that the SVM model may not be a good choice to classify Grade A and Grade B restaurants.

3.2 kNN

Secondly, we decided to use the kNN model to train our dataset. The main advantages of kNN for classification are: very simple implementation; robust with regard to the search space; for instance, classes don't have to be linearly separable. We first set aside enough train dataset and test dataset. Then, we implemented a kNN algorithm to classify our restaurant with Grade A and Grade B. The following is our result:



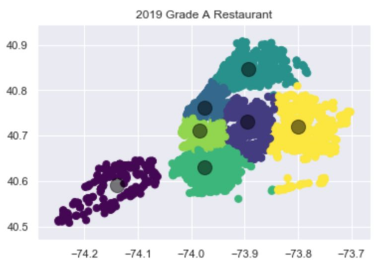
```
k = 2
prediction = make_prediction(X_train,y_train,X_test, k)
print(f"k = {k}")
error_example,error_examples_pos, error_Count, accuracy = getErrorCountAndAccuracy(prediction, y_test, X_test)
print(f"False count = {error_Count}, Accuracy = {accuracy}")
```

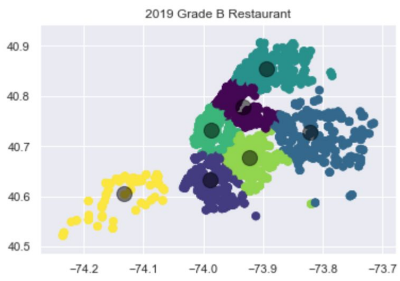
```
k = 2
False count = 996, Accuracy = 0.502
```

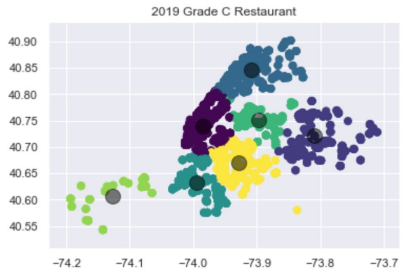
We set k equals to 2 and tested on 2000 datasets. The false count is 996 and the accuracy is 50.2%, which really makes sense because it is indeed hard to find the relationship between Grade A and Grade B restaurants according to their Longitude and Latitude. The result is similar when we change k to 1 or 3 or 4. So, the conclusion is that we can hardly identify whether a restaurant is Grade A or Grade B according to its location.

3.3 K-Means Clustering

Since we cannot make any conclusion on the relationship between restaurant location and Grade. How about finding Location distribution of restaurants at the same grade? We apply k-Means Clustering to our dataset and find that the result is amazingly beautiful.



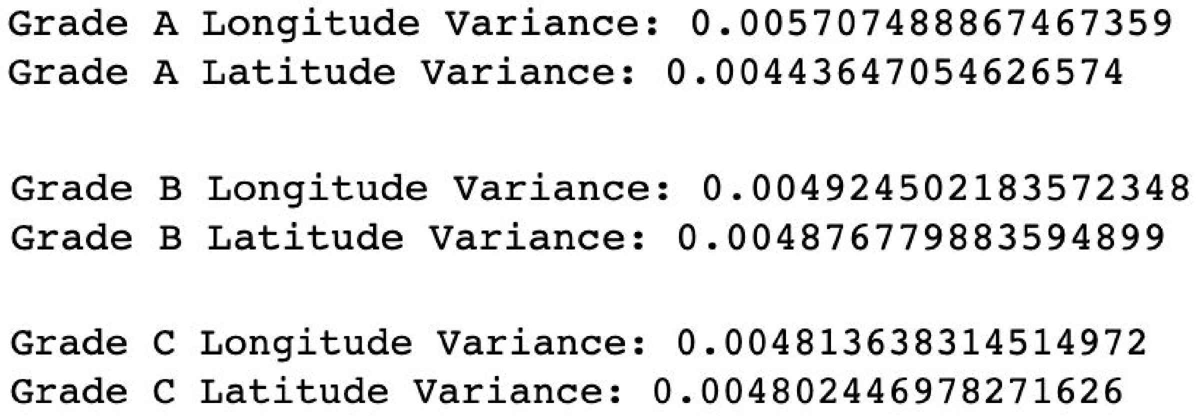




We define the clustering number as 7. Since K-means clustering has randomness, the results are based on several times of algorithm running to make sure we have the correct clustering central. We can find that the location distribution of restaurants with different grades is quite different. By analyzing the relationship between different Grades, we find that Grade A pattern is quite similar to Grade B pattern but different to Grade C pattern. We can conclude that Grade A and Grade B restaurants have similar location distribution. Then, we try to analyze the relationship between clustering center and boros. The clustering center nearly follows with the boro center.

3.4 Variance

Based on previous data exploration, we find that the distribution of different grades in different boroughs are nearly the same. All of them have 80% of restaurant graded A. So we cannot say A grade is more frequent in some boroughs. Each data contains the longitude and latitude, because the total mean of the dataset is the same, so we can calculate the variance separately.



```
Grade A Longitude Variance: 0.005707488867467359
Grade A Latitude Variance: 0.00443647054626574


Grade B Longitude Variance: 0.004924502183572348
Grade B Latitude Variance: 0.004876779883594899


Grade C Longitude Variance: 0.004813638314514972
Grade C Latitude Variance: 0.004802446978271626
```

Longitude (A - B) / A 0.13718584513725976

Latitude (A - B) / A 0.09924766382136264

Longitude (A - C) / A 0.15661012639855132

Latitude (A - C) / A 0.08249269958840043

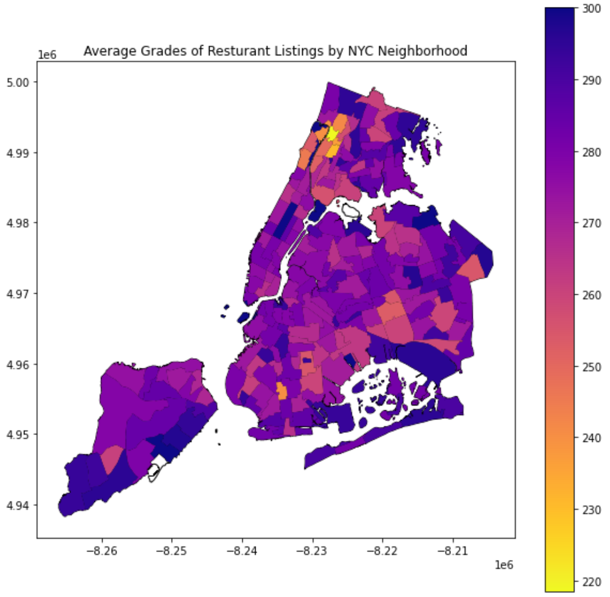
Longitude (B - C) / B 0.022512705838004785

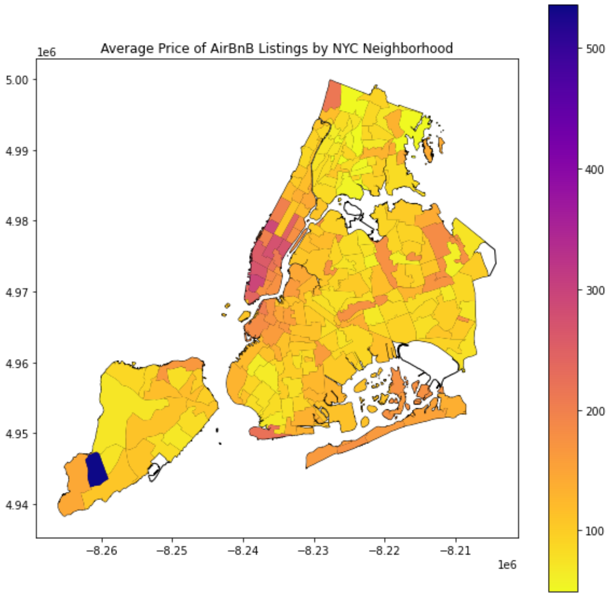
Latitude (B - C) / B 0.015242210453935556

By looking at the Longitude and Latitude variance between A, B, and C. We find that A and B have similar variance while A and C have quite different variance, which justify our conclusion on 3.3. To our surprise, we find that B and C are quite similar which we do not notice on 3.3.

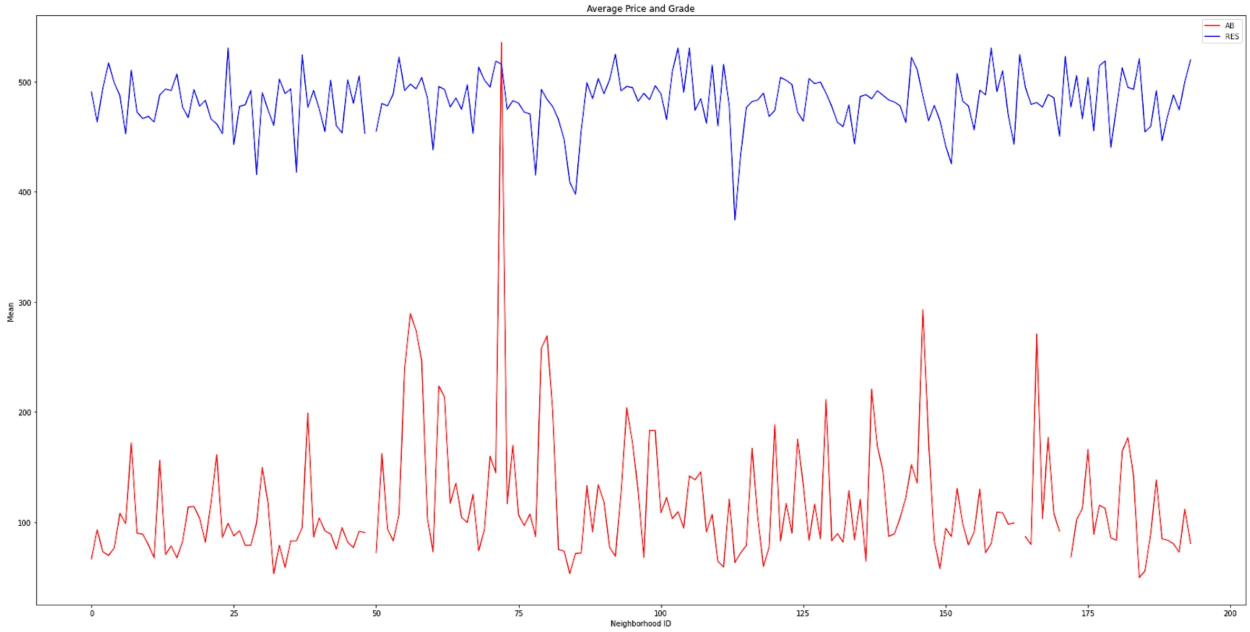
4. Relationship Between Restaurants and AirBnB

We plotted two geo-location based graphs as shown below. Since the restaurants were graded by ‘ABC’ which was not easy to show in graph, we assigned a grade of 300 to A, a grade of 200 to B and a grade of 100 to C. We took neighborhoods as units and calculated the average value of the AirBnB price and that of the grades from each restaurant. The darker the color, the higher the value. After we plotted the two graphs, we found that some of the higher value areas of the two graphs collided. We want to find out if they are correlated.

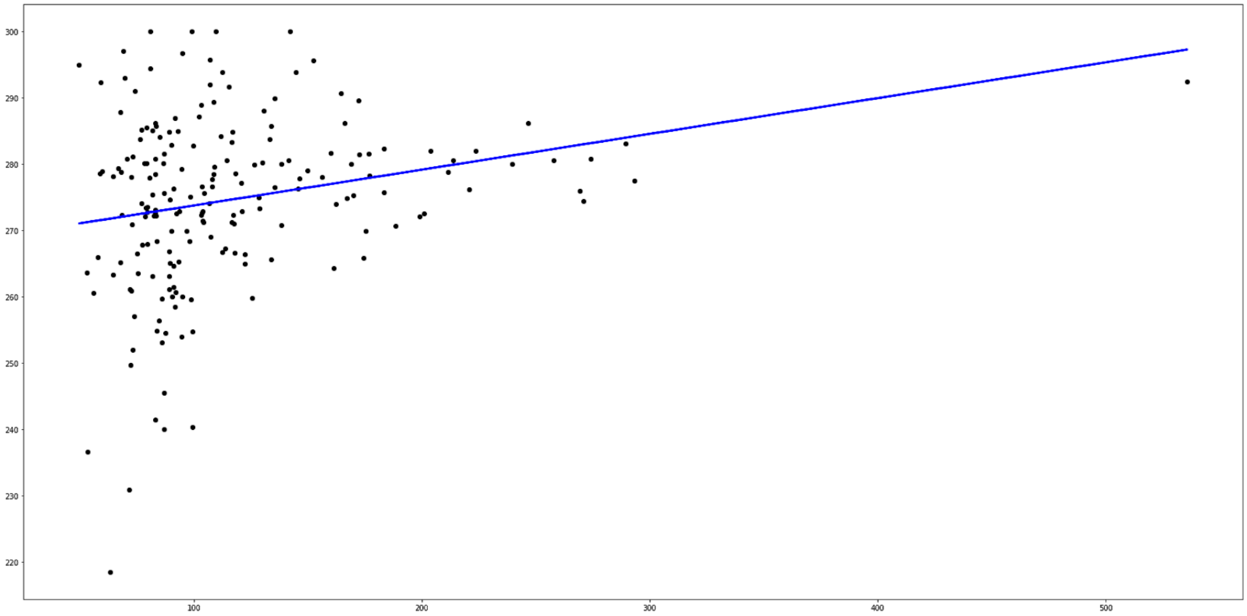




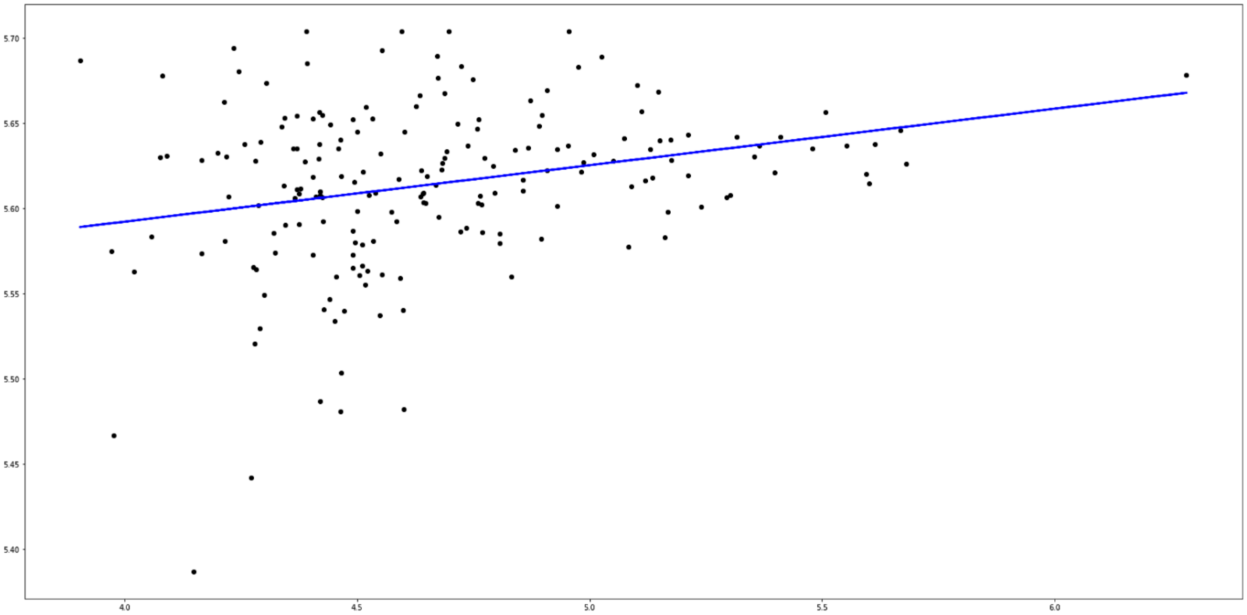
To see it more clearly, we plotted a line graph as shown below. We found that the peaks are very similar in both lines. Therefore, we made a hypothesis that AirBnB price and restaurant grade are linearly related.



In order to testify our hypothesis, we used a linear regression model from Sklearn to train our datasets. We used AirBnB price as X and we wanted to see if we can use a linear regression model to predict the restaurant grade in the same neighborhood. The result is as follows. There is no clear sign showing that they are linearly related.



The nonlinear result may be caused by the fact that the data points are too scattered. Therefore, we took a logarithm of the data points and checked if they are linearly related. The graph we plotted is shown below.

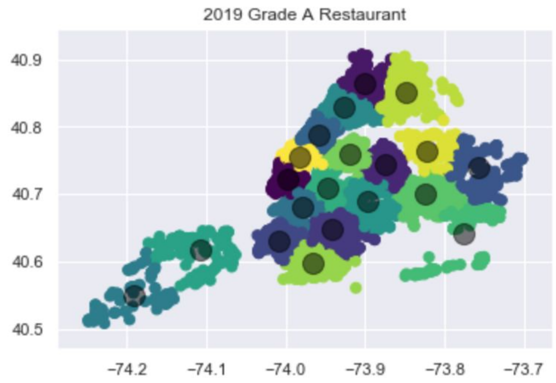


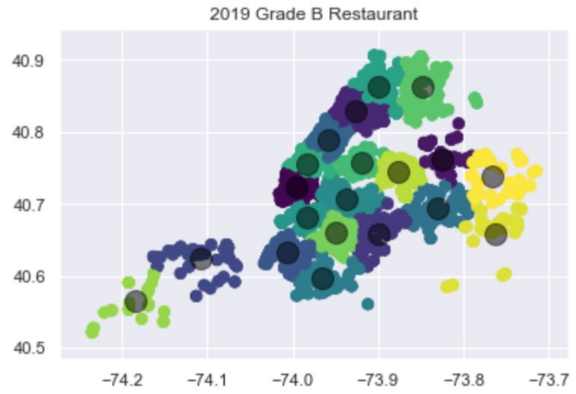
As we can see from the graph above, we can make a conclusion that logX and logY are linearly related to some extent.

Conclusion

In our intuition, we guess that the grade of the restaurant may have a strong relationship with the geometry. Based on our analysis, the geometry is different but not so relevant. Each area has almost the same distribution of each grade of restaurants.

They have different clustering centrals. In graph 3.3, we can see the details of the different centrals, and adding more centrals can help us to see the differences more clearly. For example, set n = 20.





We can see the different centrals. If we plot grade C, it would be more different. There is no clear sign showing that AirBnB price and restaurant grade are linearly related. However, after we took a logarithm of them, we could see more clearly that logX and logY are linearly related, which indicates that it's highly possible that restaurants with high grades usually appear around AirBnB with high prices. If we want a more precise model to determine the relationship between these two variables. A non-parametric regression model would be a good choice. We can also draw a graph to both indicate the Airbnb price and restaurant grade to show.
And for many other attributes we have, for example the violation codes and cuisine types, we can use the same way to find whether there is a relationship between them and the space(geometry). For the Restaurant Inspection Result dataset itself. There are still a lot of interesting relationships we can find.

Citation

NYC OpenData. (2020). *DOHMH New York City Restaurant Inspection Results, 2015-2020* https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

Kaggle. (2019). *AB_NYC_2019*, 2019
https://www.kaggle.com/chadra/ab-nyc-2019