NYU SCHOOL OF PROFESSIONAL STUDIES

New York City Restaurant Citations Analysis

Team #3

Members:

1. Jen Arriaza (ja3477, N13997347)
2. Chris Ni (zn377, N14680097)
3. Shirad Ryan Rahim (srr483, N16543099)

Course Name: Applied Data Analytics I

Course Number: ADAV1-UC 1000

Title: New York City Restaurant Citations Analysis

Section I: Research Statement

For the final project, we have chosen the data set, *New York City Restaurant Inspection Results*, which is released by the New York City Department of Health and Mental Hygiene. The inspection data ranges from the past three (3) years to present and is updated on a daily basis. The paper examines the differences in health scores and correlating grades that were received by restaurants throughout the five (5) boroughs of New York City; Brooklyn, Bronx, Manhattan, Staten Island and Queens. The paper studies the trends in the different types of cuisines and what factors and violations may have influenced their received health scores, grades, the grades of restaurants by location and borough as well as patterns of restaurants that have been closed by the New York City Department of Health and Mental Hygiene. The paper also studies what types of restaurants are more popular in which areas and which restaurants seem to be the least apparent.

The calculated statistics and information after studying the different trends can be important and useful to prospective restaurateurs who are considering opening up their businesses in New York City and can help them determine their restaurant type and location. They can use this data to see which areas may have been closed due to violations of roaches and rodents and determine what areas may be more susceptible to and may have high vermin and rodent infestations due either to geographic location and/ or its surroundings. Restaurateurs can also use this information to see which types of cuisines are popular in certain boroughs and which areas are lacking a specific type of cuisine. Moreover, the general public of New York City can benefit from this data since they would have full access to the data and the trends to make their own judgement if they choose to eat at a certain restaurant or not. It will allow them to have a more informed opinion of what was violated and what may have influenced the establishment's grade.

Section II: Data Set Description

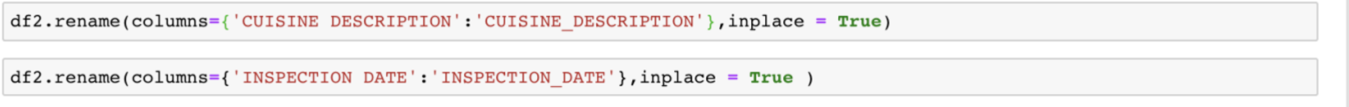
The data set, *New York City Restaurant Inspection Results,* is a considerably large dataset at approximately 395,909 rows and 26 columns in its full unaggregated form. The data is owned by NYC Open Data and provided by the Department of Health and Mental Hygiene. The dataset is available directly at the following URL: https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j/data; the dataset was made public in 2015 and is now updated daily through data provided by the DOH agency in conjunction with NYC OpenData.

The columns in the dataset mainly store nominal data, such as classifications of the cuisine type, type of inspection being performed, or whether the entity is considered critical. The citations result in numerical scores in the form of interval data, with the differences in score. There is also ordinal data in the form of letter grades — those that the consumers see displayed at restaurants in order to determine the current rank received by the Department of Health. There are two columns that record the interval data for relevant dates, such as the date the restaurant was inspected and the date their letter grade was made official. The more interesting or important information within the dataset — such as reasons for violations (i.e. evidence of live mice in food area violations) will be in the form of textual categorical data. The dataset is made easier to explore to the public by categorization by borough, zip code, etc. The remaining ancillary columns consist of nominal data that serve as identifying information such as longitude/latitude, affiliated council districts, etc. The extraction of information of this dataset can be useful to consumers in NYC restaurants and constituents utilizing NYC OpenData.

Section III: Applied Data Analytics

Part 1. Data Visualization and Exploration

New York is not merely the capital of finance but also the capital of myriad gourmet food. The first step is to standardize the dataset and prepare it to be analyzed. In other words, we have to disambiguate the values or features on which we would do analytics in the later part. The first difficulty we encountered in the very beginning is how to deal with the missing data including key elements like Grade or Score. Instead of importing random data, we decided to drop any row that has only one single value missing. After that the number of items in the dataset is trimmed from 395909 to 196904 – still an impressive amount of data to be representative. To further refine our dataset, we dropped some of the columns that may not be relevant with our project, such as PHONE, CENSUS TRACT, BBL. At the same time, a new column NEIGHBORHOOD is added by Python data munging, linking a NYC neighborhood to usually multiple zip-codes provided in the original column. Since zip-codes are not specific enough if we would like to know which neighborhood in the city has the cleanest restaurant environment. Also, we add the column YEAR whose values are extracted from column INSPECTION DATE so we can narrow our focus to the inspection taken with the recent three years. Additionally, latitude and longitude can be useful for creating a geographical heat map. As a result, we settle to 13 columns in the dataset. Finally we modified the format of the two column names: "CUISINE DESCRIPTION" and "INSPECTION DATE". The space between the words can become tricky when writing the code.

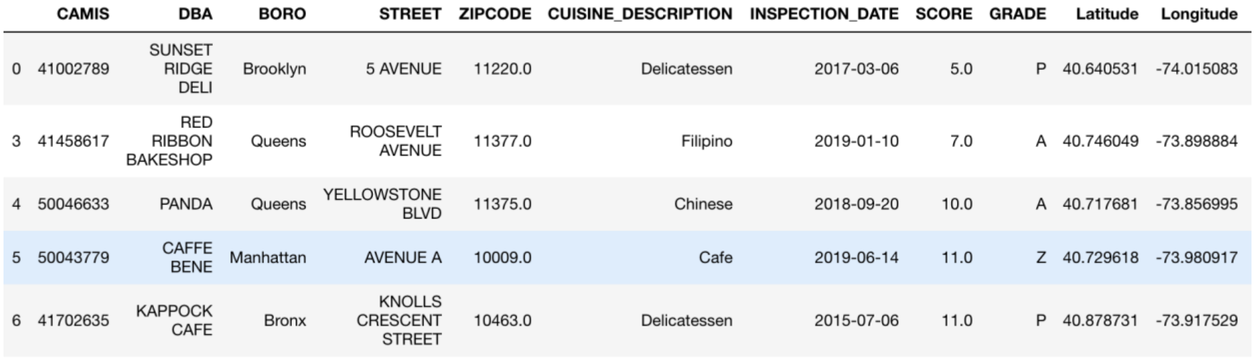


```
df2.rename(columns={'CUISINE DESCRIPTION':'CUISINE_DESCRIPTION'},inplace = True)
```

```
df2.rename(columns={'INSPECTION DATE':'INSPECTION_DATE'},inplace = True )
```

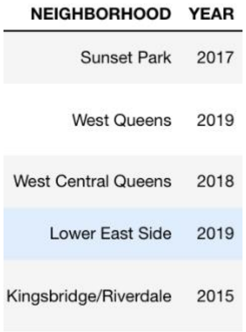
There are two column names we need to clarify. The restaurants' name is assigned to column DBA (doing business as (name)). CAMIS is the column reserved for the unique identification assigned to each restaurant in view of the fact that the same restaurant can be inspected multiple times within years, resulting in several records regarding the restaurant in the dataset.

Here is the first five rows of the clean dataset:



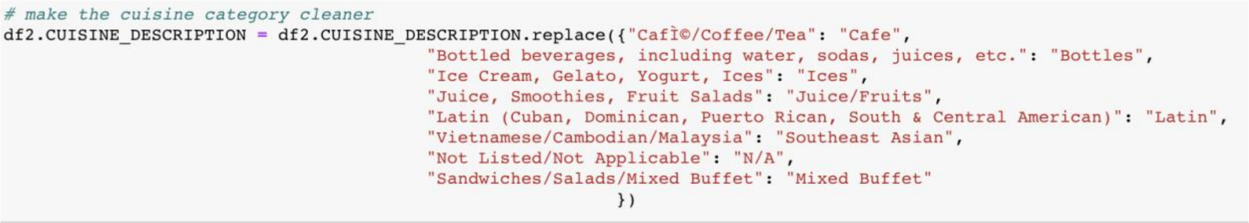
	CAMIS	DBA	BORO	STREET	ZIPCODE	CUISINE_DESCRIPTION	INSPECTION_DATE	SCORE	GRADE	Latitude	Longitude
0	41002789	SUNSET RIDGE DELI	Brooklyn	5 AVENUE	11220.0	Delicatessen	2017-03-06	5.0	P	40.640531	-74.015083
3	41458617	RED RIBBON BAKESHOP	Queens	ROOSEVELT AVENUE	11377.0	Filipino	2019-01-10	7.0	A	40.746049	-73.898884
4	50046633	PANDA	Queens	YELLOWSTONE BLVD	11375.0	Chinese	2018-09-20	10.0	A	40.717681	-73.856995
5	50043779	CAFFE BENE	Manhattan	AVENUE A	10009.0	Cafe	2019-06-14	11.0	Z	40.729618	-73.980917
6	41702635	KAPPOCK CAFE	Bronx	KNOLLS CRESCENT STREET	10463.0	Delicatessen	2015-07-06	11.0	P	40.878731	-73.917529

Here is the same clean rows with added columns.



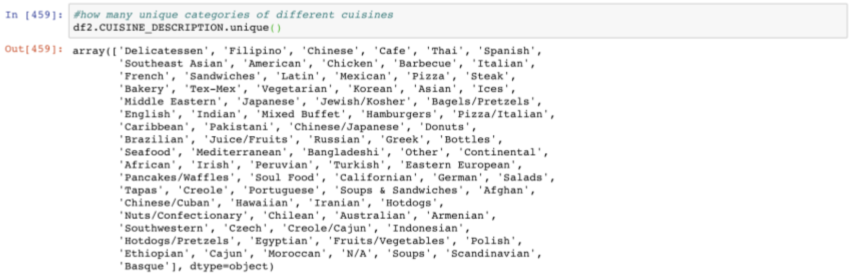
NEIGHBORHOOD	YEAR
Sunset Park	2017
West Queens	2019
West Central Queens	2018
Lower East Side	2019
Kingsbridge/Riverdale	2015

Our first dabble in the dataset is to know how many kinds of cuisines available in the city. When doing the Pandas inquiries, we notice that some names of the cuisine categories are too long and will cause blurs in the visualization. So we try to make it short and simple.



```
# make the cuisine category cleaner
df2.CUISINE_DESCRIPTION = df2.CUISINE_DESCRIPTION.replace({"Cafì©/Coffee/Tea": "Cafe",
                                                           "Bottled beverages, including water, sodas, juices, etc.": "Bottles",
                                                           "Ice Cream, Gelato, Yogurt, Ices": "Ices",
                                                           "Juice, Smoothies, Fruit Salads": "Juice/Fruits",
                                                           "Latin (Cuban, Dominican, Puerto Rican, South & Central American)": "Latin",
                                                           "Vietnamese/Cambodian/Malaysia": "Southeast Asian",
                                                           "Not Listed/Not Applicable": "N/A",
                                                           "Sandwiches/Salads/Mixed Buffet": "Mixed Buffet"
                                                                                   })
```

With the help of Python and Pandas, we can find out that there are 83 kinds of unique cuisines available in the city.

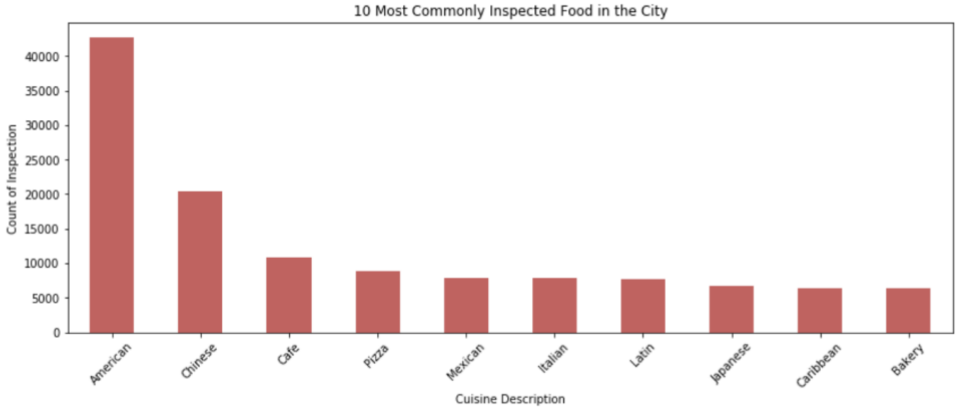


```
In [459]: #how many unique categories of different cuisines
          df2.CUISINE_DESCRIPTION.unique()
```

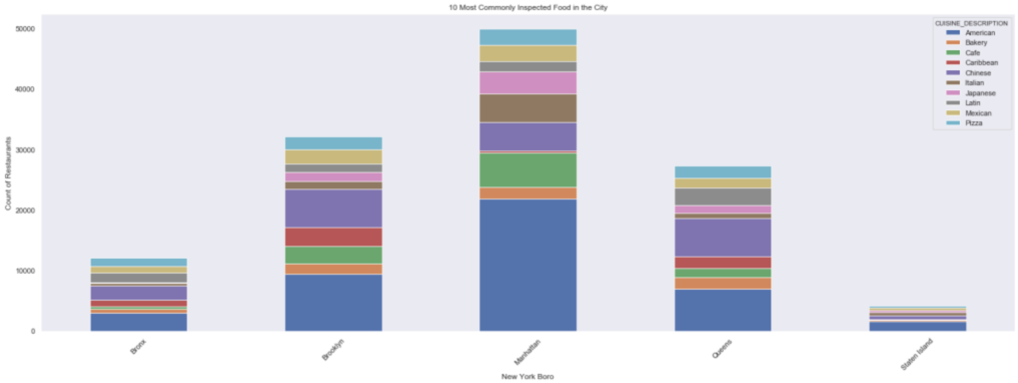
```
Out[459]: array(['Delicatessen', 'Filipino', 'Chinese', 'Cafe', 'Thai', 'Spanish',
                 'Southeast Asian', 'American', 'Chicken', 'Barbecue', 'Italian',
                 'French', 'Sandwiches', 'Latin', 'Mexican', 'Pizza', 'Steak',
                 'Bakery', 'Tex-Mex', 'Vegetarian', 'Korean', 'Asian', 'Ices',
                 'Middle Eastern', 'Japanese', 'Jewish/Kosher', 'Bagels/Pretzels',
                 'English', 'Indian', 'Mixed Buffet', 'Hamburgers', 'Pizza/Italian',
                 'Caribbean', 'Pakistani', 'Chinese/Japanese', 'Donuts',
                 'Brazilian', 'Juice/Fruits', 'Russian', 'Greek', 'Bottles',
                 'Seafood', 'Mediterranean', 'Bangladeshi', 'Other', 'Continental',
                 'African', 'Irish', 'Peruvian', 'Turkish', 'Eastern European',
                 'Pancakes/Waffles', 'Soul Food', 'Californian', 'German', 'Salads',
                 'Tapas', 'Creole', 'Portuguese', 'Soups & Sandwiches', 'Afghan',
                 'Chinese/Cuban', 'Hawaiian', 'Iranian', 'Hotdogs',
                 'Nuts/Confectionary', 'Chilean', 'Australian', 'Armenian',
                 'Southwestern', 'Czech', 'Creole/Cajun', 'Indonesian',
                 'Hotdogs/Pretzels', 'Egyptian', 'Fruits/Vegetables', 'Polish',
                 'Ethiopian', 'Cajun', 'Moroccan', 'N/A', 'Soups', 'Scandinavian',
                 'Basque'], dtype=object)
```

1. 10 Common Inspected Cuisines

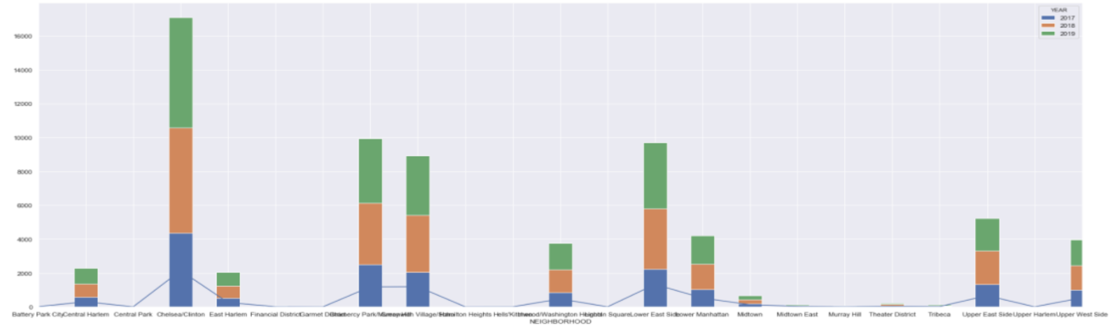
Here is a bar chart for the 10 most commonly inspected cuisines in the city. We can see that American cuisines are much more prone to be inspected than other types of cuisines.



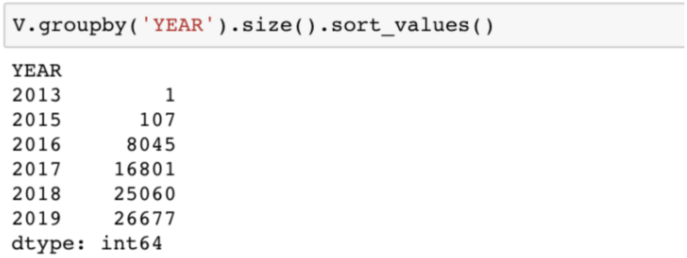
Another visualization is categorized by Boros, based on the 10 types of cuisines above. We can see Manhattan occupies a relatively large portion compared to other Boros.



Under the same criteria, we use a stacked chart to see which year(2017,2018,2019) has more inspections. Although there are other years available such as 2013, 2015 and 2016, they contain relatively fewer inspections. If we count them in our visualization, it might bring about unwelcomed outliers.



For instance, there is only one inspection in year 2013.

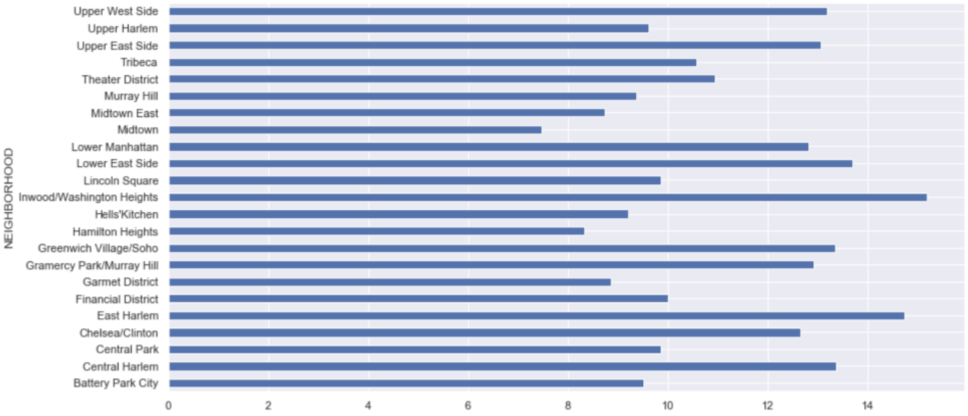


```
V.groupby('YEAR').size().sort_values()
```

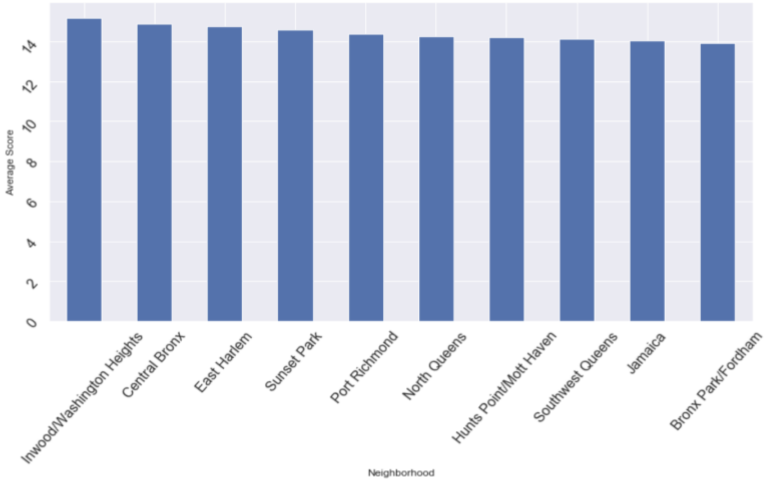
```
YEAR
2013        1
2015      107
2016     8045
2017    16801
2018    25060
2019    26677
dtype: int64
```

2. Manhattan Neighborhood: the Cleanest and Dirtiest

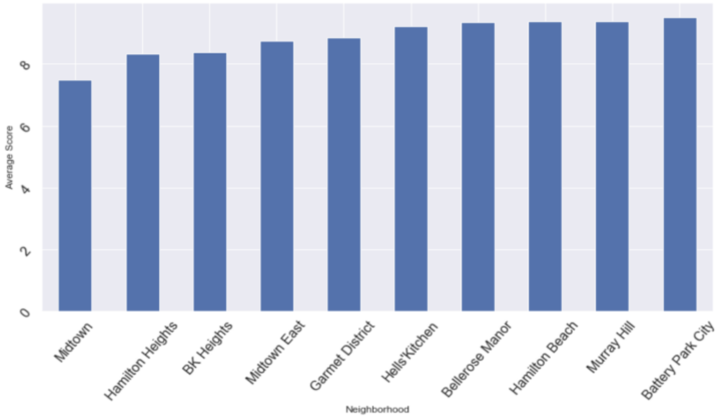
Before we discuss the scores in each Manhattan neighborhood, we can ask ourselves, if there is a neighborhood where restaurants are much cleaner than the other? Based on our visualization, here is the average score for all the neighborhood in Manhattan.



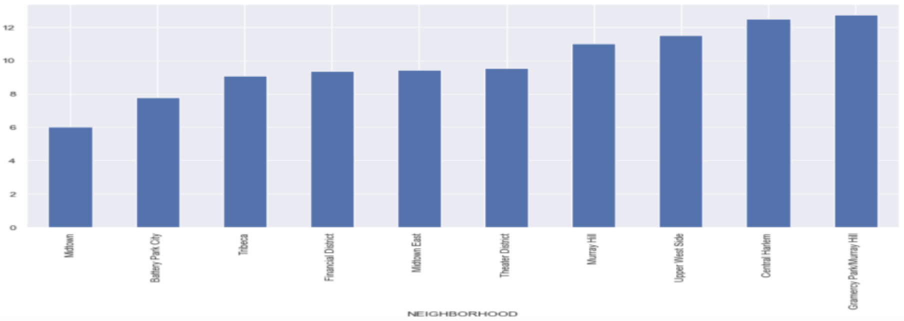
First, here is a bar chart for the 10 neighborhoods that have the dirtiest restaurant environments. The top three neighborhoods are all in Upper Manhattan.



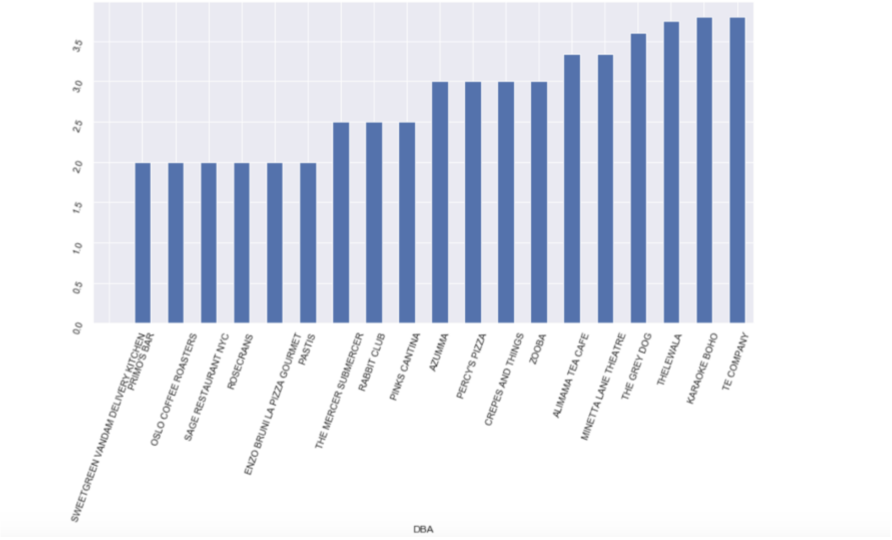
Next, another bar chart for neighborhoods that have the cleanest environments.



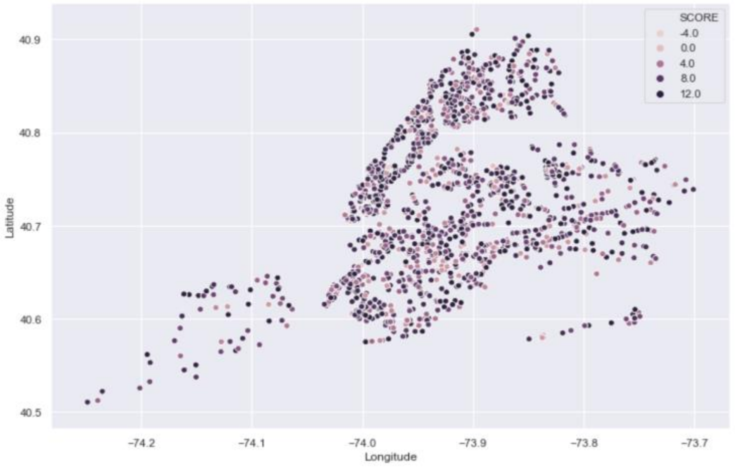
This bar chart examines the neighborhoods that have the cleanest American cuisine Restaurants. This chart is peculiarly interesting since neighborhoods Midtown and Midtown East, Downtown (Battery Park City, Financial District and Tribeca) are all business or corporation-based. We can draw a conclusion that there is a positive relation between cleanliness and neighborhood business prosperity.



Since New York University is in the Greenwich Village/Soho Neighborhood, we examined the 20 cleanest restaurants in the neighborhood.

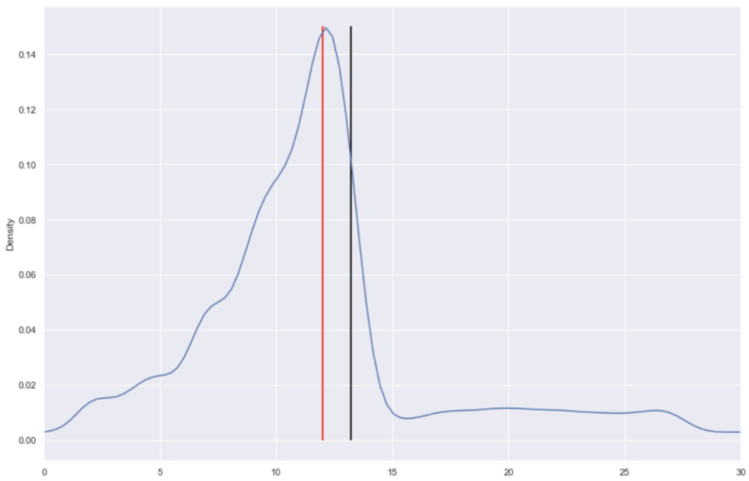


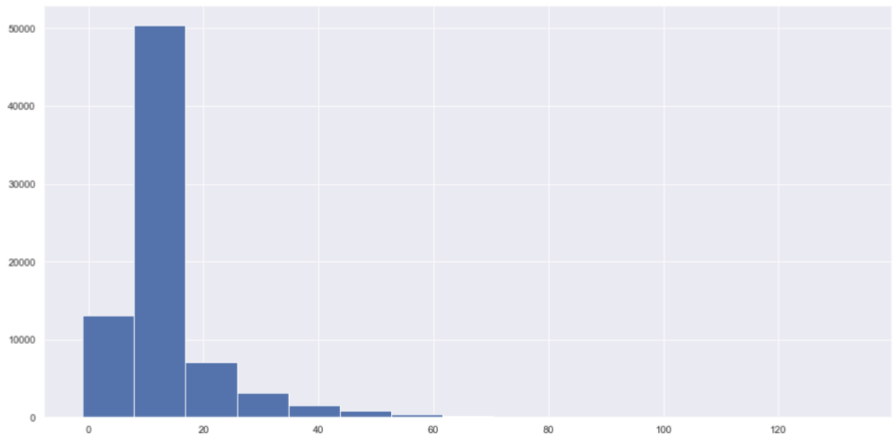
3. Heat Maps of Chinese Restaurants in New York with a score less than 10:

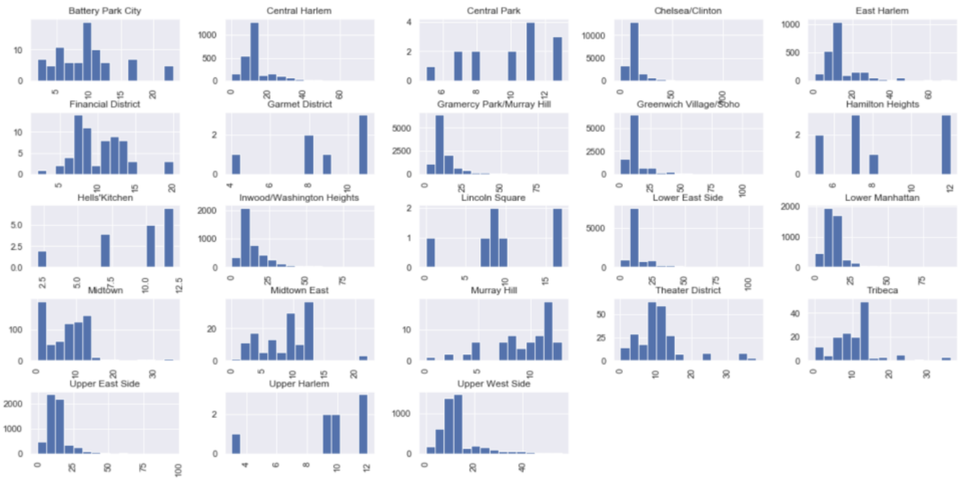


Part 2. Descriptive Statistical Measures

We can see that the median is on the left side of the mean in a skewed distribution.


Density


50000
40000
30000
20000
10000
0
0
20
40
60
80
100
120


Battery Park City
Central Harlem
Central Park
Chelsea/Clinton
East Harlem
Financial District
Garmet District
Gramercy Park/Murray Hill
Greenwich Village/Soho
Hamilton Heights
Hells'Kitchen
Inwood/Washington Heights
Lincoln Square
Lower East Side
Lower Manhattan
Midtown
Midtown East
Murray Hill
Theater District
Tribeca
Upper East Side
Upper Harlem
Upper West Side