Abhijith Anil Vamadev

IST 652 - Scripting for Data Analysis

7th November 2022

**NYC Restaurant Inspection Data Analysis**

NYC Restaurant Inspection Data Analysis project, uses the data from an online government website that has a food inspection database (Department of Health and Mental Hygiene (DOHMH)). The data was downloaded as a CSV file and uploaded into and worked on using python. The data contains information from 2018 to 2021. The raw csv file contained, 26 columns with 23,000 data points. Each row represents the information about a restaurant (Restaurant id, name, address, phone, location, latitude, longitude, GRADE, violation description etc). Out of the 26 variables, 9 variables did not have any information regarding the column values, so these values haven’t been taken into consideration for the analysis of this project. The remaining variables that were used intially are: CAMIS (id for restaurants), DBA (Displayd building name), BORO (Boroughs in NY), BUILDING (Building no in NY), STREET (Street No in NY), ZIPCODE (Zip Code in NY), CUISINE DESCRIPTION (Description of food provided by owner), PHONE (Phone no in NY), INSEPCTION DATE (Inspection date for restaurant in NY), ACTION (Action issued for restaurant in NY), VIOLATION CODE (Violation code for restaurant in NY), VIOLATION DESCRIPTION (Description for violation code for restaurant in NY), CRITICAL FLAG (Flag indicater for restaurant in NY), SCORE (Total score for a particular inspection), GRADE (A, B or C grade given to restaurant based on violations) , GRADE DATE (Date grade was issued for restaurant), RECORD DATE (Inspection date for the restaurant) and INSPECTION TYPE (how inspection was conducted for the restaurant).

Restaurant Grade is one of the variables that will be looked at, restaurant grades are divided into A, B and C. GRADE A: The restaurant is clean, up to code, and free of violations. GRADE B: The restaurant has some issues that must be fixed. GRADE C: The restaurant is a public risk and on verge of closure. For the scope of the project, we will be looking to answer 3 questions which are:

1. Identify the 5 most common cuisine types available in NYC
2. Identify the restaurants on a map, by using geographical coordinates for different Boroughs and Grade
3. For different Grades A, B and C analyze the violation description and are there any difference in violation description for the A, B and C grade?

The following libraries were used in this project: Pandas, Seaborn, Matplotlib, Wordcloud Plotly and NLTK.

# **Data Preprocessing**

Once the data was loaded into a data frame using pandas, there were some null values present which were removed using *df.dropna*, then all the columns were converted into their respective data type. Changed the column names by splitting the column names and converting into lower case connecting all the strings with a “\_”. Dropped Duplicates by using the “name” column as a unique identifier.

Then dropped null values present in Grade, Latitude, Longitude and Violation Description. The columns were reduced to 12 with 10,000 data points.

# **Question 1** Identify the 5 most common cuisine types available in NYC? **Unit of analysis:** Cuisine Description **Comparison:** There are many cuisines description types available. Group by the column and then find the size of the column. Comparing the count of each of the restaurant type.

| Cuisine Description | Total Restaurants in NY |
| --- | --- |
| American | 2559 |
| Chinese | 925 |
| Coffee/Tea | 656 |
| Pizza | 590 |
| Italian | 490 |

**Table 1: Top 5 cuisines in New York**

# Chart, bar chart Description automatically generated

**Figure 1: Top 5 Cuisine Types food in New York.**

# **Question 2** Identify restaurants in a map by using geographical coordinates for different Boroughs and by different restaurant grades **Unit of analysis:** Boroughs in NYC, counting the total repetition of it. **Comparison:** Using an online software Open Software Map, downloaded a picture of the required boundaries of NYC latitude and longitude values. Group the data by Borough and latitude, longitude and then plotted the 5 boroughs by using matplotlib and by using the image as a background canvas. The same procedure was used for Grades except the data was grouped by grade instead of borough.

| Boroughs | Total Restaurants in NY |
| --- | --- |
| Manhattan | 1958 |
| Brooklyn | 1242 |
| Queens | 948 |
| Bronx | 364 |
| Staten Island | 244 |

**Table 2: Number of restaurants by Borough in New York**

Map

Description automatically generated  
**Figure 2: Restaurants visualized using longitude and latitude values for every Borough**

| Grades | Total Restaurants in NY |
| --- | --- |
| A | 8207 |
| B | 512 |
| C | 204 |

**Table 3: Number of restaurants by Borough in New York**

Map

Description automatically generated **Figure 3: Restaurants visualized using longitude and latitude values by every GRADE of restaurant**

# **Question 3** For different Grades A, B and C analyze the violation description and are there any difference in violation description for the A, B and C grade? **Unit of analysis:** Violation description column was used – Text/String Column **Comparison:** Tokenize the violation description text seperately for A, B and C grade. Find unique items from the tokenizer, remove any numeric characters from the text and remove stop words which were collected from both the WordCloud library and NLTK library. Find the frequency of each of the unique items, convert the words into a frequency table/data frame, and finally plot each of the analysis data frames using a Tree Map.

Chart, treemap chart

Description automatically generated

**Figure 4: Most common words appearing in GRADE A violation description of restaurants. Contact, Surface, Properly are some of the most common words**

Chart, treemap chart

Description automatically generated

**Figure 5: Most common words appearing in GRADE B violation description of restaurants. Vermin, food, and flies are some of the most common words**

Chart, treemap chart

Description automatically generated

**Figure 6: Most common words appearing in GRADE C violation description of restaurants. Food, flies, and Vermin are some of the most common words**

# **Conclusion**

Through this project, python was used to analyze the three questions, Identify the 5 most common cuisine types available in NYC, Identify the restaurants on a map, by using geographical coordinates for different Boroughs and Grade and for different Grades A, B and C analyze the violation description and are there any difference in violation description for the A, B and C grade?

The most common type of cuisines available in NYC are: American, Chinese, Coffee/Tea, Pizza, Italian Manhattan has the greatest number of restaurants in NYC and there are 8407 GRADE A restaurants, with Staten Island having the lowest number of GRADE C restaurants. We plotted the graphs using an online software for the base canvas and used matplotlib to plot the latitude and longitude coordinates.

There are differences in wording for the different GRADES violation descriptions:

1. Grade A - Contact, Surface, properly. Improperly, equipment, food
2. Grade B - Vermin, Food, Flies, properly, contact, mice
3. Grade C - Food, Flies, Vermin, Contact, properly, improperly, hair etc.

Although the vocabularly dosen’t have a huge change between classes, the words used and their frequency changes with each grade. In grades B and C, words like vermin and flies were much more frequent than in Grade A. Violation description gets worse with each grade from A to B to C, which aligns with research conducted before that stages grade description for restaurant goes worse from A to B to C. With the new information gained, can avoid the Grade C restaurants, pickup good restaurant locations by Borough in NY and hopefully eat better food with my next visit to New York, City.

# **Works Cited**

1. Department of Health and Mental Hygiene (DOHMH). “Dohmh New York City Restaurant Inspection Results: NYC Open Data.” DOHMH New York City Restaurant Inspection Results | NYC Open Data, 6 Dec. 2022, <https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j>.
2. OSM. “Export.” *OpenStreetMap*, https://www.openstreetmap.org/export#map=10/40.7290/-73.7581.