PROCESS DOCUMENTATION

NYC RESTAURANT FOOD SAFETY

Cut Famelia
Shaowei Gong
Shile Zhao
Zhaojun Yang

CS 573 Data Visualization
Final Project
Worcester Polytechnic Institute
Fall 2016

ENTATION

OVERVIEW

Food Safety Awareness

New York City (NYC) is one of the most diverse and cosmopolitan cities in the world, where many people come from many different backgrounds and cultures. Due to this fact, NYC offers many options of basic needs for the people living in the city. One of the needs is restaurants that provide different types of food based on the diversity such as American, Chinese, Indian, Mexican, Mediterranean, Asian, and so forth. All these different choices of restaurant are spread out across the city's five boroughs.

Along with the increasing number of those thematic restaurants and the scaling up population, the food quality of the restaurants has become a serious concern for the government and the people. To address this issue, the government under the Department of Health and Mental Hygiene (DOHMH) has conducted official inspections towards all the restaurants within the city. "In July 2010, the Health Department began requiring restaurants in all five boroughs to post letter grades summarizing their sanitary inspection scores to help achieve three goals: to inform the public about a restaurant's inspection results in a simple, accessible way; to improve sanitary conditions and food safety practices in restaurants; and to reduce illnesses associated with dining out. This report summarizes progress toward these goals." (http://www1.nyc.gov/assets/doh/downloads/pdf/rii/restaurant-grading-18-month-report.pdf)

In fact, since January 2011, when the inspections began, the inspectors have found a big number of food safety violations in many of the restaurants in different ways. According to the NYC Government report, New Yorkers eat food at restaurants almost a billion times every year. While most of them do not get sick, foodborne bacteria, viruses and other contaminants cause millions of cases of illness each year. It is estimated that more than 6,000 New Yorkers are hospitalized and 20,000 visit emergency rooms each year because of foodborne illnesses. Each year, NYC receives approximately 2,700 complaints about restaurant-acquired foodborne illnesses and another 3,000 complaints about restaurant hygiene. Based on a study conducted by Baruch College Survey Research in 2012, 88% of New Yorkers take the grades into account when making dining decisions. Besides, 70%

concern about getting sick from eating out in restaurants, delis and coffee shops, with 38% are very concerned.

Food Safety Grading System

According to the NYC Government report, the inspection program uses letter grades: A, B or C, with dual inspections to help restaurants improve their food safety practices before they are graded. If a restaurant does not get an A on its first inspection (initial inspection), the Health Department does not issue a grade but conducts a second inspection (reinspection) around a month later and issues a grade based on the re-inspection. This series of inspections is called a cycle. Each health code violation results in points; total points are a restaurant's inspection score. The following is the score ranges for each letter grades:

A = 0 – 13 points (excellent food safety practices)
B = 14 – 27 points
C = 28 points or more

Restaurants that earn B or C on their initial inspection begin a new cycle of inspections sooner. Those earning A on their initial inspection start a new cycle in around 12 months, while those getting B or C begin a new cycle in approximately 6 months and 4 months, respectively.

Why Food Safety Inspection Data Visualization Important?

Before letter grades, restaurants were motivated to practice food safety by their own desire to maintain healthful conditions and by the threat of fines for violations found at the time of inspection. From the data the inspectors have collected, in fact, not every restaurant has done improvement based on the inspection findings and recommendations. This fact is indicated by the lower grades given to a number of restaurants after the last inspection. To evaluate the impact of the inspections, it is important for DOHMH to see the trend of the grades given to each restaurant after every inspection was done, whether it is going up or going down, or fluctuating. Thus, DOHMH can easily do some analysis to get some insights for future inspections.

Furthermore, we believe that it is important for New Yorkers, especially those who like to eat outside and care about food safety to learn the inspection data summary of all the restaurants and do analysis accordingly in an easy and interactive way to help them in dining decisions to maintain healthy lifestyle. Also, the visualization will be useful for food safety supervisors and analysts to support their evidence-based works.

RELATED WORK

One of the visualizations that we are working on has been inspired by an interactive visualization of NYC street trees, which uses a horizontal bar chart with a tree symbol corresponding to each tree type. However, since we have 84 restaurant types in our dataset, which is too many, we think that it is not effective to put all the 84 symbols of restaurant types on our bar chart. http://www.cloudred.com/labprojects/nyctrees/#about

QUESTION

This project aims to address the following questions:

1. How well the inspections have affected the restaurant food safety practices.
2. How the rank of every restaurant type looks like.
3. Which type of restaurants have the highest rate of food safety.
4. Which type of restaurants have the lowest rate of food safety.
5. How every restaurant type's grade performance looks like based on location and period of time.

DATA

We use DOHMH New York City Restaurant Inspection Results dataset retrieved from NYC Open Data website (https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j). This website provides data on restaurant inspections, violations, grades, and adjudication information of 439,301 records.

The structure of data is described below:

- CAMIS: This is a unique identifier for the entity (restaurant)
- DBA:This field represents the name (doing business as) of the entity (restaurant)
- BORO:Borough in which the entity (restaurant) is located.
- BUILDING:This field represents the building number for the entity (restaurant)
- STREET:This field represents the street name at which the entity (restaurant) is located.
- ZIPCODE:Zip code as per the address of the entity (restaurant)
- PHONE :Phone Number
- CUISINE DESCRIPTION: This field describes the entity (restaurant) cuisine.
- INSPECTION DATE: This field represents the date of inspection
- ACTION: This field represents the actions that is associated with each restaurant inspection.
- VIOLATION CODE:This field represents the violation codes that is associated with each restaurant inspection.
- VIOLATION DESCRIPTION:This field is the description that corresponds to the violation codes
- CRITICAL FLAG:This indicates if Violation is critical or not.
- SCORE:Total Score for a particular inspection. If there was adjudication a judge may reduce the total points for the inspection and this field will have the update amount.
- GRADE:• N = Not Yet Graded • A = Grade A • B = Grade B • C = Grade C • Z = Grade Pending • P= Grade Pending issued on re-opening following an initial inspection that resulted in a closure

- GRADE DATE:The date when the current grade was issued to the entity (restaurant)
- RECORD DATE: The date when the extract was run to produce this data set
- INSPECTION TYPE:The type of inspection. A combination of the program and inspection type.

The followings are the details of the data attributes along with each number of the value types:

CAMIS = 25,969

DBA = 20,506

BORO = 6

BUILDING = 7,228

STREET = 3,309

ZIPCODE = 230

CUISINE DESCRIPTION = 84

INSPECTION DATE = 1,335

ACTION = 6

VIOLATION CODE = 97

VIOLATION DESCRIPTION = 95

CRITICAL FLAG = 3

GRADE DATE = 1,256

INSPECTION TYPE = 34

DATA PROCESSING

The dataset takes 160 MB in memory, which is too much to fit into a web page. Besides, the dataset contains a lot of information, so we extract and re-organize the data based on what we need to address each question. Therefore, we do the following data pre-processing work based on the project objectives:

1. Removing "Missing" values from “BORO” column
2. Transforming some parts of the data into JSON data type to make it easier to visualize the restaurant and grade summary based on user selection
3. Building an index JSON for line chart, and help users to navigate among the huge restaurant list
4. Compressing index JSON by using dictionary.
5. Removing the restaurant which has only one record
6. Removing duplicated records

EXPLORATORY DATA ANALYSIS

To get the whole picture of what the dataset actually tells us, we firstly looked at the CSV file of the dataset. Here, we found that the dataset has much information that could be useful for analysis. We then tried to figure out some important questions we wanted to answer based on the dataset and we found many interesting questions that can be addressed, which mainly may lead to NYC food safety quality improvement in the future. For instance, the dataset tells the grade of each restaurant so that we can find out which restaurants have the best food safety quality and which restaurants have the worst, also to compare among some restaurants. This will help New Yorkers in making food decision. Besides, the dataset has “VIOLATION TYPE” column that specifies what kind of food safety rules that a restaurant breaks when being inspected. From this data, we can explore what kind of violations mostly done so that the Health Department can analyse why those violations occurred so that they can figure out what kind of actions they need to take towards the restaurants to reduce the violation rate. However, for this final project with a limited amount of time, we would not have enough time to explore and visualize all the important data to answer all the interesting questions. Therefore, we finally decided to focus on the restaurants and the inspection performance, i.e. the grade distribution of the restaurants and the inspection impact on food safety improvement in the restaurants, which we break down into 5 questions as mentioned in “Question” section. However, if we have enough time, we plan to make an additionally supporting visualization to explore the violation types to address the abovementioned question.

After collecting the questions, we deployed Panda to see the data validity and to get some calculation on the data values so that we could see how much workload we would need to do to make an interactive visualization to address the questions. We finally found that it is reasonable to implement our visualization design to address all the five questions in our list with the given length of time and the validity of the data. However, since the number of restaurant is very huge: 20506, which is too many to be put in one visualization, we decided to shorten the number by putting the restaurants into a number of categories based on the food themes/types. We use this number in the restaurant population and grade summary charts only.

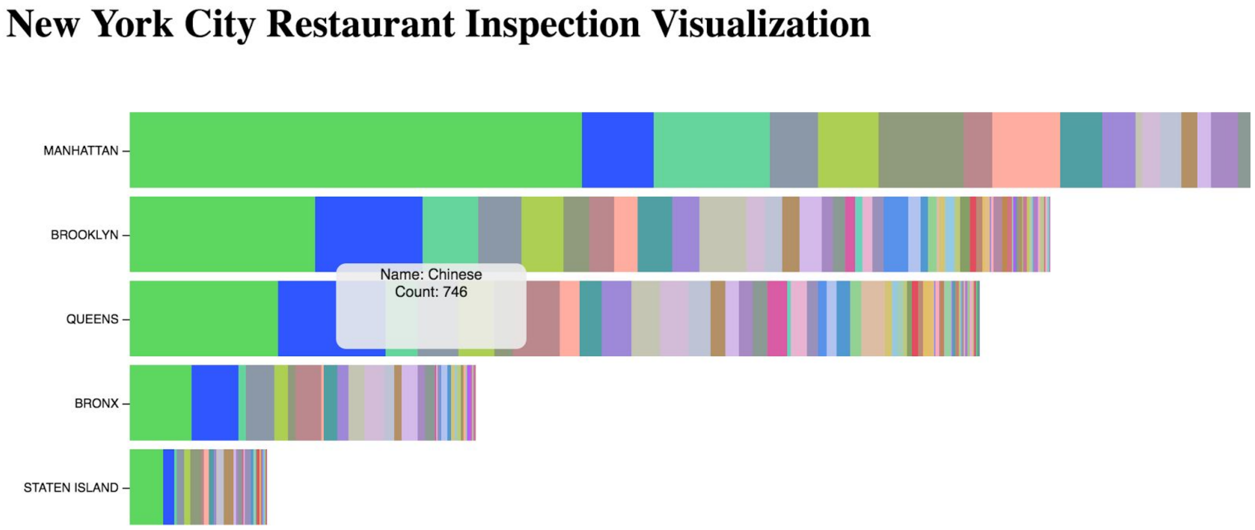
DESIGN EVOLUTION

After doing the exploratory data analysis and finding five questions to be in our list, we decided to create 5 visualizations. Three of them are basically the same as what we described in our proposal, i.e. a horizontal bar chart showing grade A ranking among all restaurants, vertical grouped bar charts illustrating all the grade summary of all restaurants, and the DNA chart describing each restaurant grade performance from time to time. Meanwhile, after considering efficiency factor and some suggestion from Prof. Lane Harrison, we finally changed the tree chart that we mentioned in our proposal into a series of drop-down selection which ends up with line charts of score trends. The last one is a new one: a horizontal bar chart telling the restaurant population. We believe that this visualization is important to give users the whole picture of the restaurant profiles in NYC at the beginning of exploration so that users can always see how the population relates to the rest of the visualizations to get more sense in the analysis.

IMPLEMENTATION

The followings are the intent and functionality of all the five visualizations:

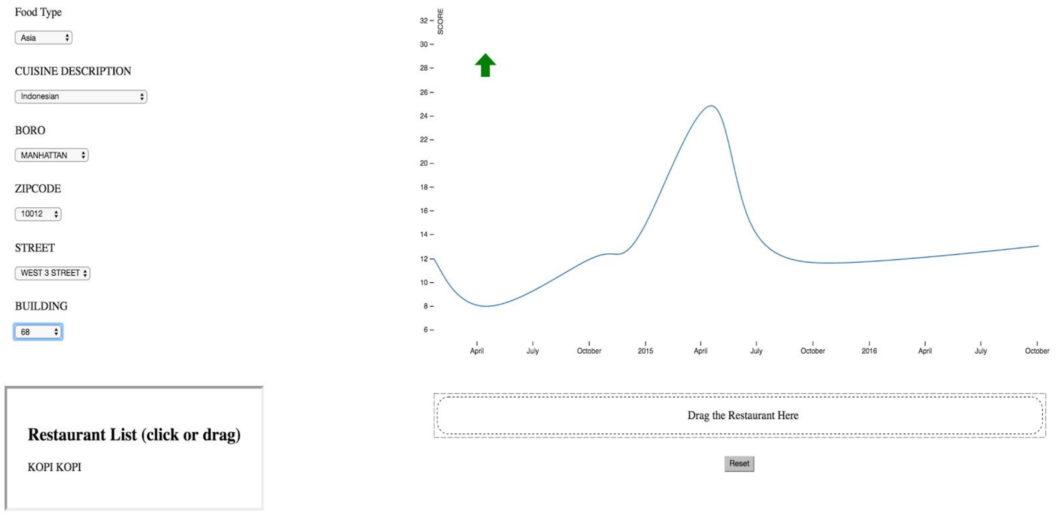
1. **Independent Horizontal Bar Chart (Restaurant Population)**



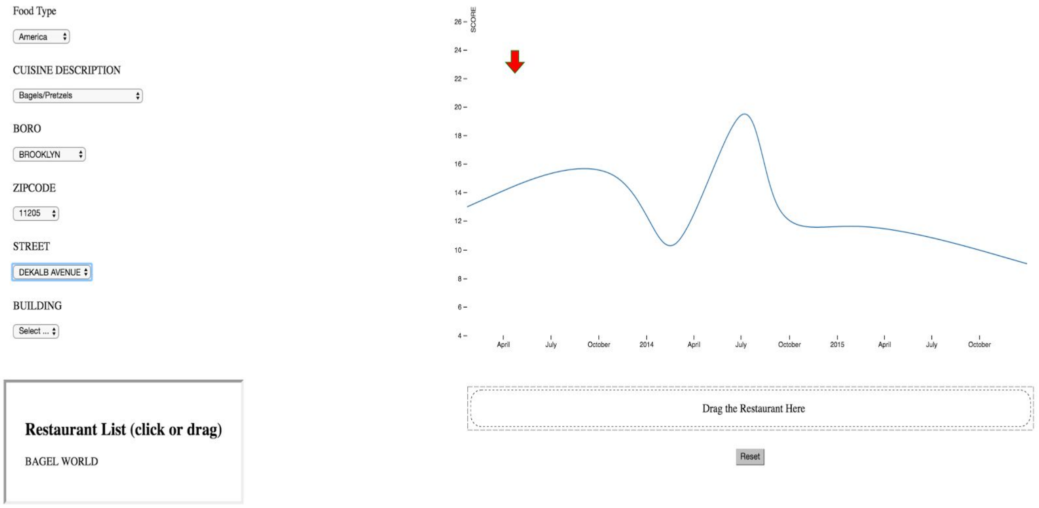
New York City Restaurant Inspection Visualization

The bar is designed to expose the whole picture of the restaurants (names and counts) in each borough. Since the number of restaurant is too many (20506), we categorize the restaurants' names based on food type retrieved from CUISINE DESCRIPTION column. There are 84 food types based on the dataset and all of them are shown with 84 different colors on the bars. Once a user puts a cursor on one of the color bar, a tooltip showing the name and count of the corresponding restaurant will turn up. This chart can help users figure out the population of the restaurants to get the context of the whole visualization before they start the exploration on the rest of the charts based on their needs.

2. Interactive Line Chart (Score Distribution Trend of Each Restaurant)


Food Type
Asia
CUISINE DESCRIPTION
Indonesian
BORO
MANHATTAN
ZIPCODE
10012
STREET
WEST 3 STREET
BUILDING
68

Restaurant List (click or drag)
KOPI KOPI


Food Type
America
CUISINE DESCRIPTION
Bagels/Pretzels
BORO
BROOKLYN
ZIPCODE
11205
STREET
DEKALB AVENUE
BUILDING
Select ...

Restaurant List (click or drag)
BAGEL WORLD

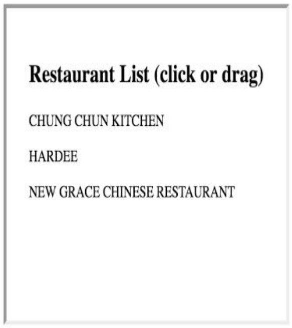
The line chart is made to show the score, not the letter grade, distribution trend of each restaurant from 2014 to 2016 by month. To see the trend of a restaurant, a user can select an attribute item from each of the drop-down selection bars, but it is not necessary to pick a value from each bar. So a user can pick a value based on his need. If he wants to know the score trend of every Indonesian restaurant in Manhattan no matter where they are located based on zip codes, streets, and the buildings, then he can just pick a value on each of Food Type, Cuisine Description, and Boro bars, and let the rest of the bars without any value.

Once a user picks all the values that he needs, the name(s) of the corresponding restaurant(s) will pop up in the box under the “building” selection bar. If there are more than one restaurant shown up in the “Restaurant List” box, then he needs to click on each restaurant name, one by one, to see the trends. The chart will only show the trend of the restaurant on the top by default. If the box only shows one restaurant, then the trend will pop up automatically without a mouse click. From the trends, users can see how well the inspections have affected the food safety quality of the restaurants.

If the score trend of a restaurant is going down, it means the restaurant has made better improvement in food safety practices, and a green arrow with up direction will turn up. Meanwhile, if a trend is going up, then it indicates that the food safety practices in the restaurant becomes worse, signed by a red arrow with down direction.

However, we made a mistake by accident. We just realised that we made the arrow works in a reverse way, so we will make it correct it soon later.

3. Interactive DNA Chart (Grade Performance and Comparison among Restaurants)

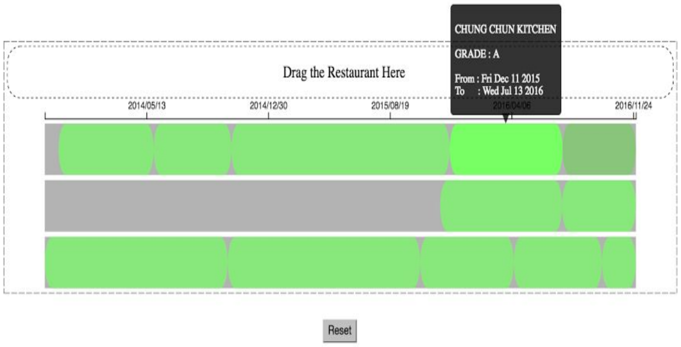


Restaurant List (click or drag)

CHUNG CHUN KITCHEN

HARDEE

NEW GRACE CHINESE RESTAURANT



CHUNG CHUN KITCHEN

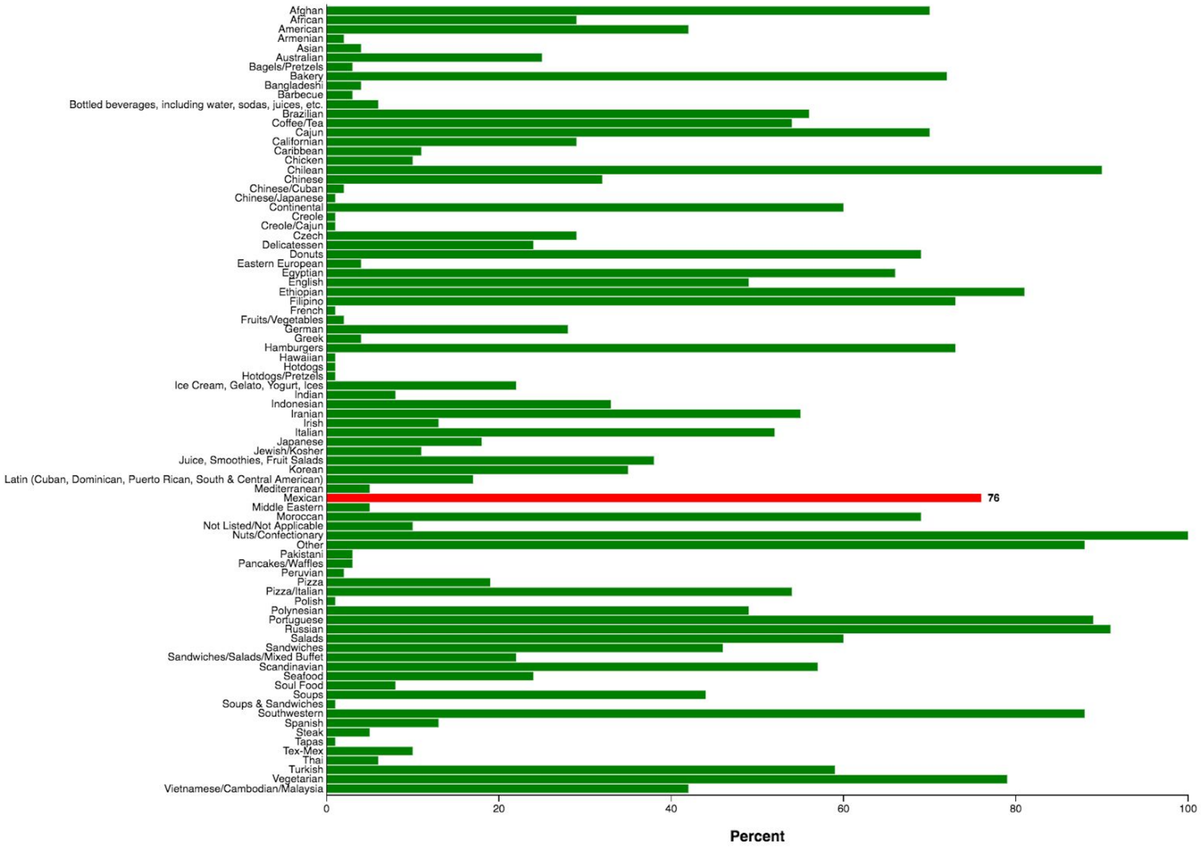
GRADE : A

From : Fri Dec 11 2015
To : Wed Jul 13 2016

The purpose of creating the DNA chart is to show users how every restaurant’s grades change from time to time since the first inspection was conducted. Once a user explores visualization #2, he can drag any of the restaurant(s) in the “Restaurant List” box to the “Drag the Restaurant Here” bar so that he can see the grade evolution by date. If he has more than one restaurants in the box, then he can drag more than one restaurant to the dragging bar so that he can see the grade evolution comparison among those restaurant.

If he moves a cursor on any of the DNA bar, a tooltip showing the restaurant's name, grade, and grade date range will turn up. At this point, users can learn how every restaurant's grade performance looks like based on location and period of time.

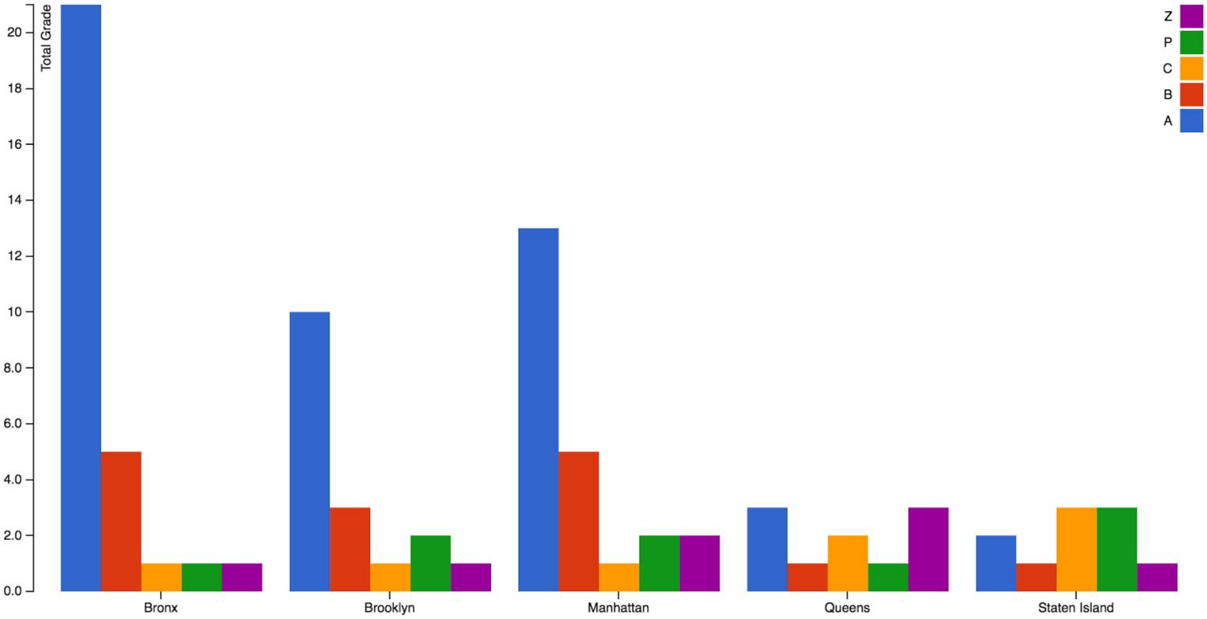
4. Interactive Horizontal Bar Chart (Restaurant Ranking Based on Grade A)



This cart is intended to show users the comparison among the restaurants across boroughs based on the percent of grade A accumulation that each restaurant type has got from the latest inspection. So, there are 84 bars, each of which represents each restaurant type. We use tooltip to show the associated percentage. Once a user click on a bar, it will be highlighted and it will be connected to visualization #5 if being clicked.

However, since the data that we are using for the bar chart has to go through many filtering and calculation to get the mapping, we manipulated and converted the select data from CSV to JSON using Panda. This data pre-processing in fact has been taking longer time, so we are still working on this visualization. The figure above is a trial of the chart using some parts of the data, to show how the chart will look like.

5. Interactive Vertical Grouped Bar Chart (Restaurant Grade Summary)



If a user goes over visualization #4, when he mouseovers any of the bar, it will be highlighted and if he clicks on it, five separate vertical grouped bar charts representing five boroughs will pop up. This chart will illustrate the summary of each letter grade (A, B, C, P, Z) associated to the restaurant type clicked by in each borough. At this point, users can learn in which borough a restaurant type has the most A or the most C.

Since the carts are connected to visualization #4, we are still working on this visualization. The figure above is a trial of the chart using some parts of the data, to show how the chart will look like.

REFERENCE

https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

http://www1.nyc.gov/assets/doh/downloads/pdf/rii/restaurant-grading-18-month-report.pdf

http://www.cloudred.com/labprojects/nyctrees/#about