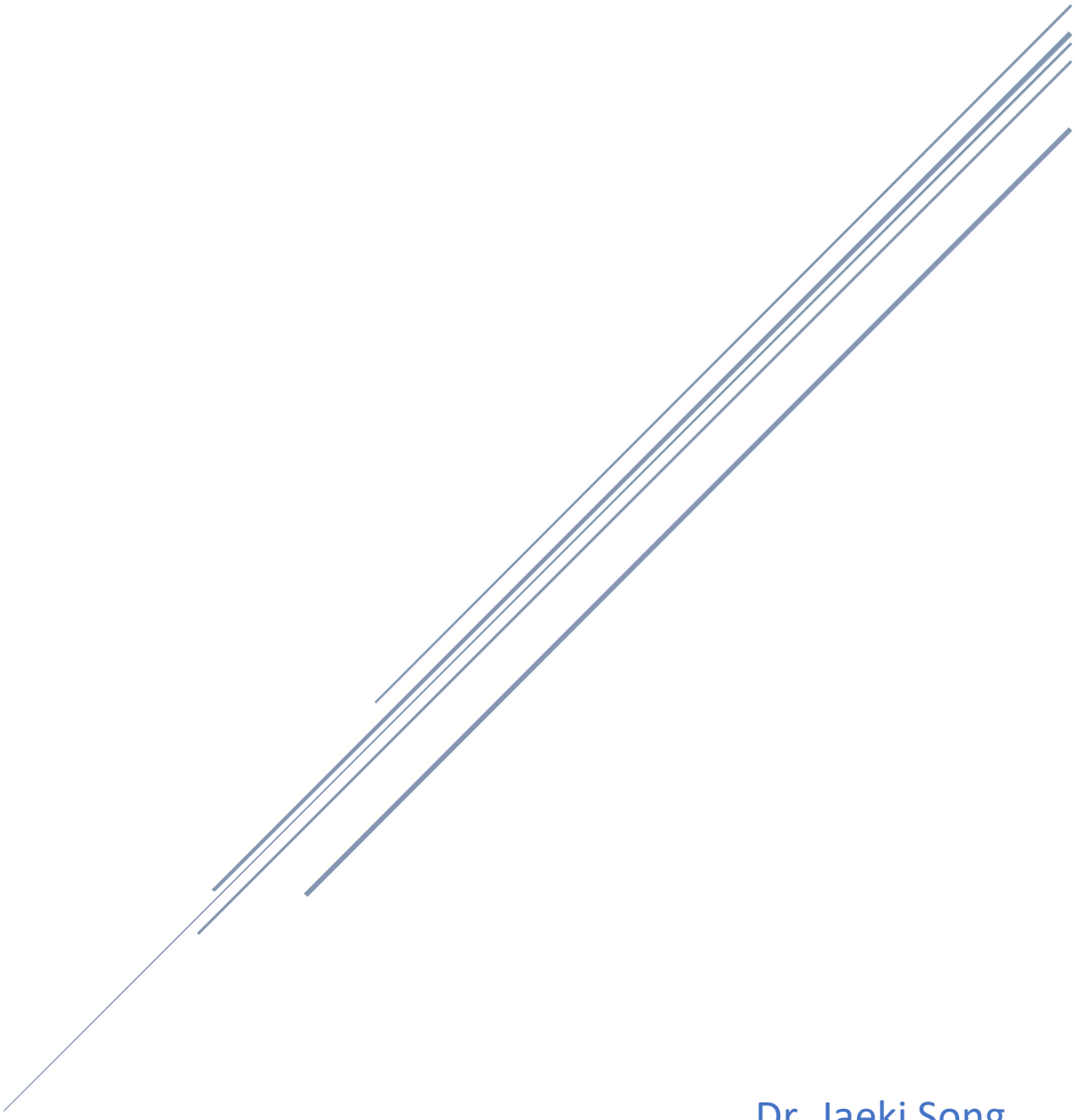
WHERE SHOULD I DINE?
(ANALYSIS OF NYC RESTAURANT INSPECTION DATA)

Siva Manda, Manuj Jha and Krishna Charan Bajanthri Chikalaguriki



Dr. Jaeki Song
ISQS 6337 Scripting Languages

EXECUTIVE SUMMARY

In this study, we are looking at health inspection data for restaurants in the New York City area. This dataset was created by the New York Department of Health and Mental Hygiene (DOHMH) and was downloaded from the NYC Open Data online database. This data set has 17 variables with over 400,000 observations for over 20,000 restaurants in NYC. The variables include things like the name of the restaurant, the address, which borough it is in (Manhattan, Brooklyn, etc.), violations found, and many more attributes. Our objective through the analysis of this data is to help the people of the NYC area narrow down the safest areas/restaurants (food wise) for their type of restaurant/food. Mainly, we seek to provide insights and general trends on how cleanly restaurants in a certain area based on the borough, cuisine type, inspection results (i.e. grade of restaurant) and restaurants with the least violations. We have conducted various analyses like finding the percentage of all restaurants in each area of New York that received a grade of 'A', number of restaurants for different popular cuisine types, restaurants with critical violations in each area of NYC, average score for different cuisine types and many more. A time series analysis is also done between popular pizza chains (i.e. Domino's, Papa John's, etc.). This analyzed data was then put into graphical representations to make it easy to interpret by anyone. We hope that after going through our insights, anyone can make an informed decision on which area to eat at based on his/her cuisine choice and possible violations to look out for.

ABOUT THE DATASET

This dataset provides restaurant inspections done, violations, restaurant grades and other information. The dataset has 18 variables for over 20K restaurants and around 400K records.

Data Source: Department of Health and Mental Hygiene (DOHMH) on New York City Restaurant Inspection Results, updated on September 4, 2017.
https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

Assumptions:

- The dataset contained various empty cells and were replaced by NA's using R.
- The Dataset also had observations constituting less than 1% of total observations and were ignored.
- There were multiple inspections record for individual restaurants. So, we filtered the data to consider the inspection for the latest date.

Following are the various fields in the dataset:

Table 1: Variables of the data set

Column Name	Description	Data Type
CAMIS	This is a unique identifier for the entity (restaurant)	Varchar
DBA	This field represents the name (doing business as) of the entity (restaurant)	Varchar
BORO	Borough in which the entity (restaurant) is located	Varchar
BUILDING	This field represents the building number for the entity (restaurant)	Varchar
STREET	This field represents the street name at which the entity (restaurant) is located	Varchar
ZIPCODE	Zip code as per the address of the entity (restaurant)	Varchar
PHONE	Phone number	Varchar
CUISINE DESCRIPTION	This field describes the entity (restaurant) cuisine	Varchar
INSPECTION DATE	This field represents the date of inspection	Datetime

ACTION	This field represents the action that is associated with each restaurant inspection	Varchar
VIOLATION CODE	This field represents each violation associated with a restaurant inspection	Varchar
VIOLATION DESCRIPTION	This field describes the violation codes	Varchar
CRITICAL FLAG	Critical violations are those most likely to contribute to foodborne illness	Varchar
SCORE	Total score for a particular inspection; updated based on adjudication results	Varchar
GRADE	This field represents the grade associated with this inspection	Varchar
GRADE DATE	The date when the grade was issued to the entity (restaurant)	Datetime
RECORD DATE	The date when the webextract was run to produce this data set	Datetime
INSPECTION TYPE	A combination of the inspection program and the type of inspection performed	Varchar

DATA CLEANING AND PREPARATION

Before we could start our analysis, we needed to clean the raw data that was read from the csv file provided by NYC Open Data. We start off with reading the data in R using the read.csv() function. Since the data has many missing values, and the character 'Not Available' in a few fields, we replaced such fields with NA while reading the data itself. This gives us a dataset over which we could perform further analyses. In the next step, we converted the data in columns that contained dates to '%y/%m/%d' format so that R could make sense of the dates. We used the "as.Date()" function for this process. Since the observations where 'No Violations' were found doesn't contain grade or score as well, we ignored such observations as they are not very helpful for our analysis. We performed this operation using the filter() function and stored the filtered data in a new variable for further cleaning. We also got rid of few columns which were not important for our analysis and reduced the size of dataset.

We observed that our data had multiple observations for each restaurant based on different inspection dates by grouping the data by restaurant name, building and name. This repeated occurrence was not a desirable trait in our data as it would lead to multiple grades and scores for each restaurant and the result would not make much sense. So, we selected only the observation pertaining to the latest inspection date for each restaurant. In order to do so, we grouped the data based on restaurant name, street and building using group_by() function, followed by arranging in descending order of inspection date by arrange() function and slicing the first value using the slice() function. This gave us a new dataset containing only one observation for each restaurant. This dataset would be used in most of the future analyses.

ANALYSES

1. Number of Restaurants per Neighborhood in NYC

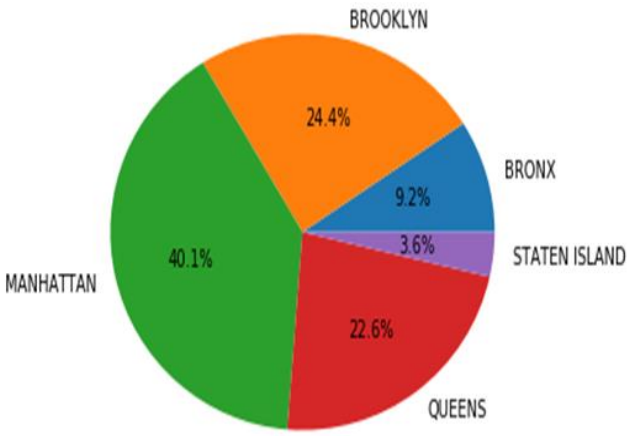


Figure 1: Pie chart for Restaurants Distribution using Python

In this first analysis of the data that was conducted, the list of all the restaurants located in New York City were grouped by their respective borough or neighborhood. In the data set, a small number of these restaurants did not have a specified borough or had missing information for that column, so the ones with boroughs were filtered out after grouping them by their respective neighborhoods. A tally for each borough was then done to get the total number of restaurants for each borough, and then this data was put into a pie chart to make it easier to interpret and compare numbers between each borough. This analysis is important for the project as it serves as the foundation for the rest of the analysis conducted. Through this data, we can see that Manhattan has the most number of restaurants out of all the neighborhoods in New York with around 10,000 restaurants, with Brooklyn and Queens right behind it followed by the Bronx and then Staten Island with the lowest.

2. Distributions of restaurant scores in NYC

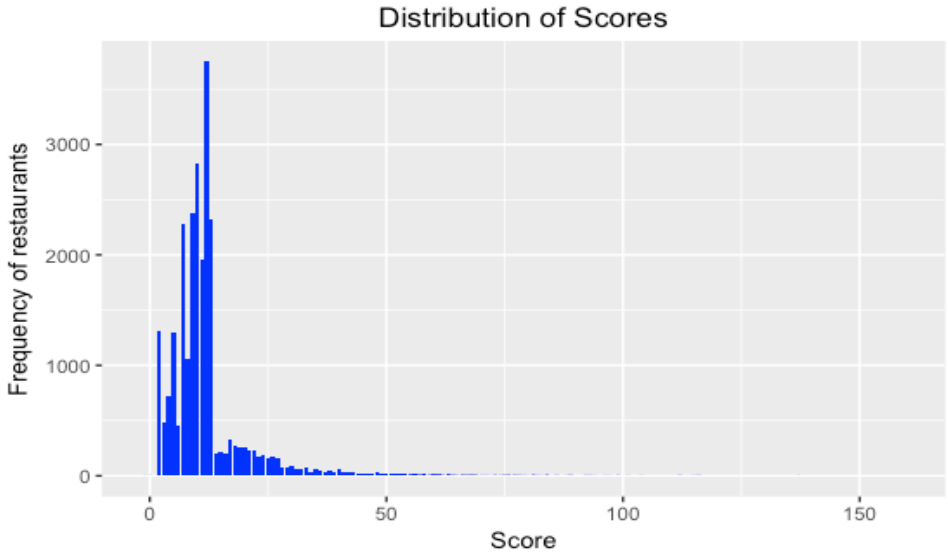


Figure 2: Distribution of Restaurant Scores

In the data set, for each restaurant its respective inspection data (date, Critical flag, violation code, etc.), the score given for the inspection and the respective grade for the score to each restaurant. A restaurant's score depends on how well it follows City and State food safety requirements, specified by NYC health guidelines. Inspectors check for food handling, food temperature, personal hygiene, facility and equipment maintenance and vermin control. The scale for the restaurant grades and scores are:

- Score: 0-13; Grade = A
- Score: 14-27; Grade = B
- Score: 28+; Grade = C

A score in the "A" range of 0-13 is a good score and given to restaurants with no violations or have very small and non-critical violations that do not negatively affect the restaurant. A score in the "B" range means the restaurant either had several small minor violations or a couple major violations. A score in the "C" range of 28 points or more means the restaurant had numerous violations that were critical and posed significant health threats. All the unique observations and latest inspection data was taken for each of the restaurant and the distribution of these scores was plotted. The peak for the distribution is seen at a score of 12, which is the lower bound of the "B" score range. We can see from the data that most of the restaurants have a grade of "A", with a lot being on the border between an A and a B. A small amount of the restaurants do have a grade of "C", with most of the scores being well below 50. This distribution is vital to the whole analysis of the data as it gives us an idea of what scores most of the restaurants are receiving and where these scores range to. This distribution then serves as foundation for further analysis of scores and grades by cuisine type, boro, etc seen in the analysis.

3. Number of Restaurants by the type of Cuisine

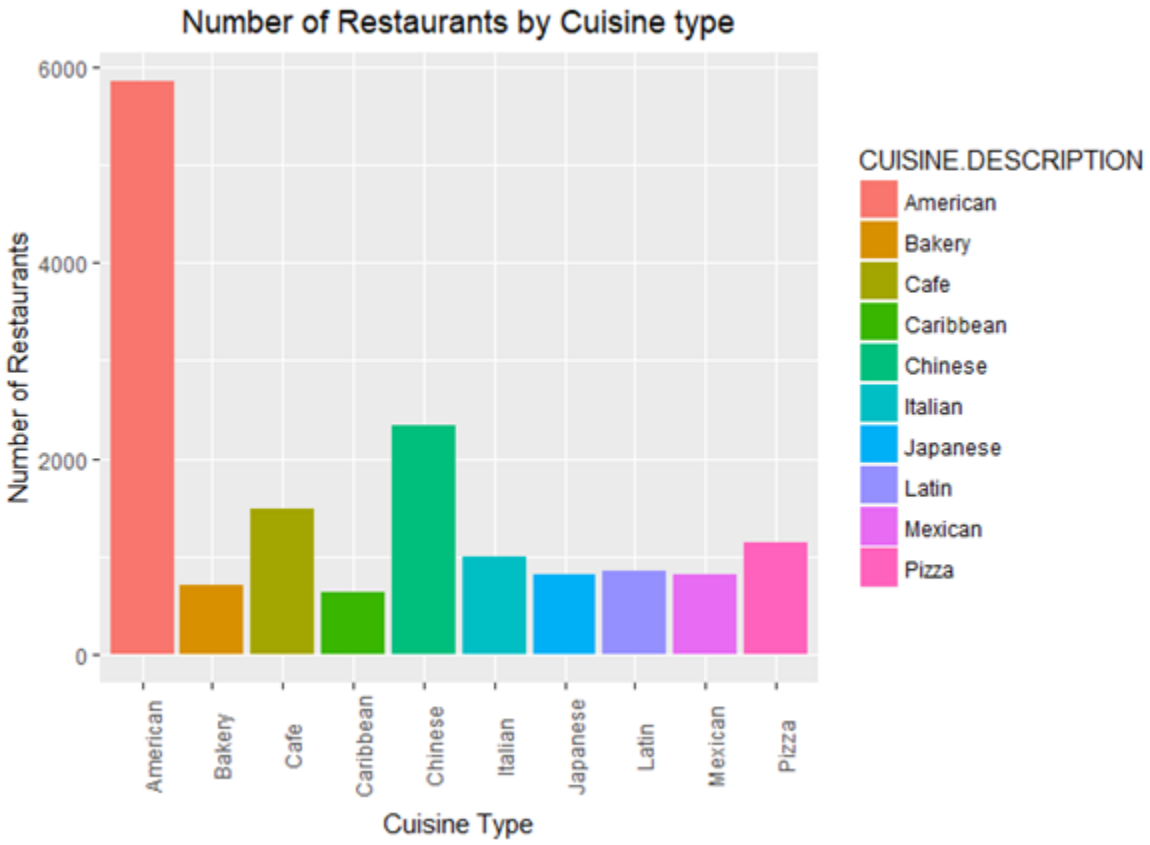


Figure 3: Distribution of Number of Restaurants for each Cuisine type

While analyzing the distribution of restaurants by the type of cuisine they served, we started with grouping the dataset by the type of cuisine. There were more than 80 types of cuisines served in the New York out of which we selected the top ten for easy visualization in limited space. After grouping, counting the number of restaurants using tally function, arranging and selecting the top ten types of cuisines by number of restaurants, we used ggplot2 in R to plot the data. From the bar chart above, we can infer that American cuisine is the most prevalent among restaurants in New York followed by Chinese and Cafe. The following analysis can be very helpful while selecting any kind of cuisine we would want to eat as it would describe the likelihood of getting the desired kind of restaurant in a neighborhood.

4. Distribution of Grades of Restaurants in each Borough:

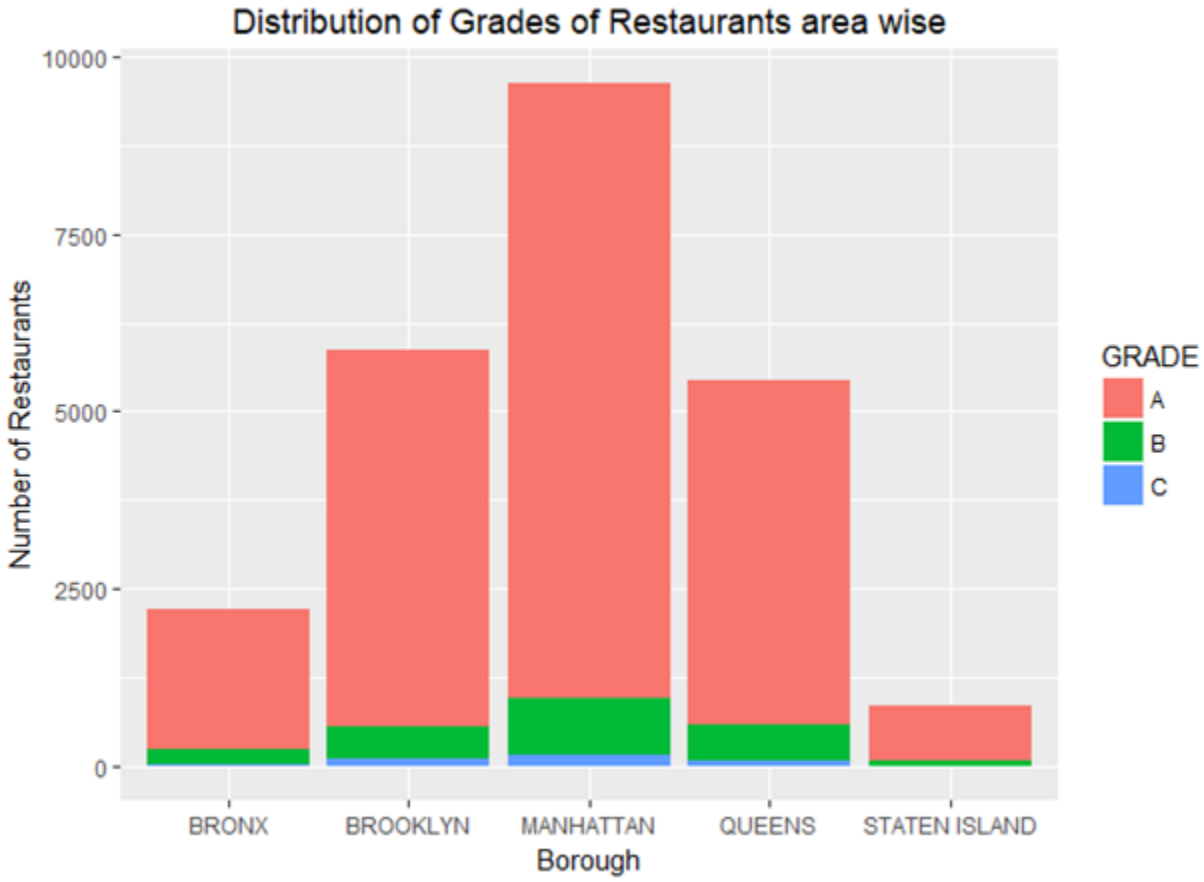


Figure 4: Distribution of Grades for Restaurants in each Borough

All of us want to eat at a restaurant which has been graded best by the Department of Health and mental hygiene. Distribution of restaurant grades in each neighborhood gives us the insight on the distribution of restaurants in each borough so that insight about the likelihood of finding a high graded restaurant could be generated. For the above analysis, we started with grouping the data with latest inspection dates by the borough they were situated in and the grades they received in inspections. After getting the count of number of observations in each such group, we plotted the stacked bar chart to show the distribution of restaurants in each neighborhood by grade in an effective manner. From the chart, Manhattan has the highest number of restaurants graded 'A', at the same time it also had the highest number of restaurants with grade 'B' and 'C'. Bronx and Staten Island have most of its restaurants graded 'A', and very less number of 'C' graded restaurants.

5. Percentage of ‘A’ graded restaurants in each neighborhood

The previous analysis gave us about the distribution of number of restaurants in each neighborhood by grade. We saw that Manhattan had the highest number of top graded restaurants, but it gives us no idea to if we select a restaurant how likely is that a top graded venue because it also has the highest number of restaurants in total. To fix that issue, we need to look at the percentage of ‘A’ graded restaurants in each neighborhood.

BORO	GRADE	n	percentage
BRONX	A	1972	88.98917
BROOKLYN	A	5297	90.43879
MANHATTAN	A	8655	89.87539
QUEENS	A	4848	89.31466
STATEN ISLAND	A	770	89.43089

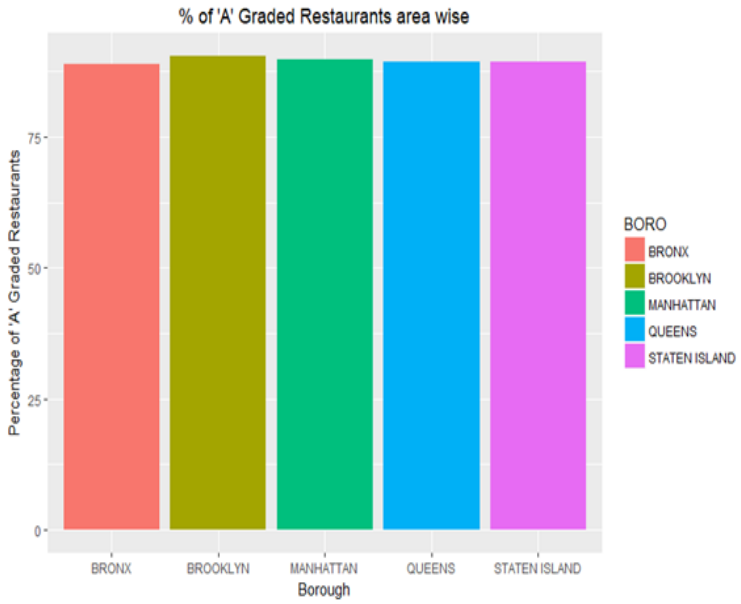


Figure 5: Percentage of "A" Graded Restaurants in respective Boroughs

For this analysis, we grouped the dataset with latest inspection dates for each restaurant by neighborhood and grade. Then we find the number of ‘A’ graded restaurants using filter(), tally() and mutate() functions and find the percentage of ‘A’ graded restaurants in each borough. The analysis gives us insights on how likely a person is to find a top graded restaurant in the neighborhood. From the chart and table above, Brooklyn has the highest percent of ‘A’ graded restaurants whereas Bronx has the least, but not by a big margin.

6. Percentage of Restaurants with Critical Violations

Whenever someone goes out to dine, it is always a good idea to understand how often the restaurants in the neighborhood have critical violations. Critical violations include presence of mice, roaches or fungi in the kitchen or the food not being stored at appropriate temperature and many more. With the aim of educating diners about the critical violations, we conducted the analysis involving number of critical and non- critical violations in each neighborhood. Below is the plot for the distribution of critical/non-critical violations per borough and similarly for each cuisine type.

BORO	CRITICAL.FLAG	n	percentage
BRONX	Not Critical	1100	48.67257
BRONX	Critical	1160	51.32743
BROOKLYN	Not Critical	2879	48.10359
BROOKLYN	Critical	3106	51.89641
MANHATTAN	Not Critical	4631	47.21656
MANHATTAN	Critical	5177	52.78344
QUEENS	Not Critical	2575	46.45499
QUEENS	Critical	2968	53.54501
STATEN ISLAND	Not Critical	427	48.13980
STATEN ISLAND	Critical	460	51.86020

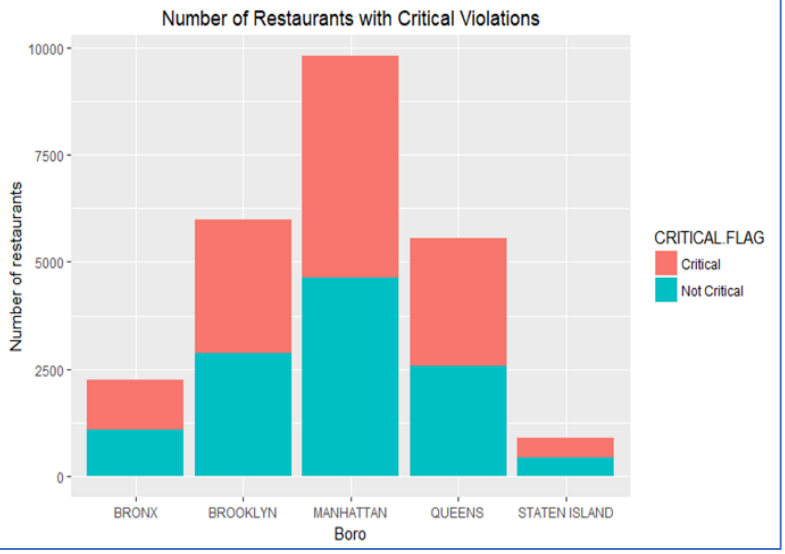


Figure 6 : Distribution of Restaurants with Critical Violations in each Borough

CUISINE.DESCRIPTION	CRITICAL.FLAG	n	percentage
American	Critical	3037	51.85248
American	Not Critical	2820	48.14752
Caribbean	Critical	330	51.32193
Caribbean	Not Critical	313	48.67807
Chinese	Critical	1280	54.81799
Chinese	Not Critical	1055	45.18201
Japanese	Critical	480	58.39416
Japanese	Not Critical	342	41.60584

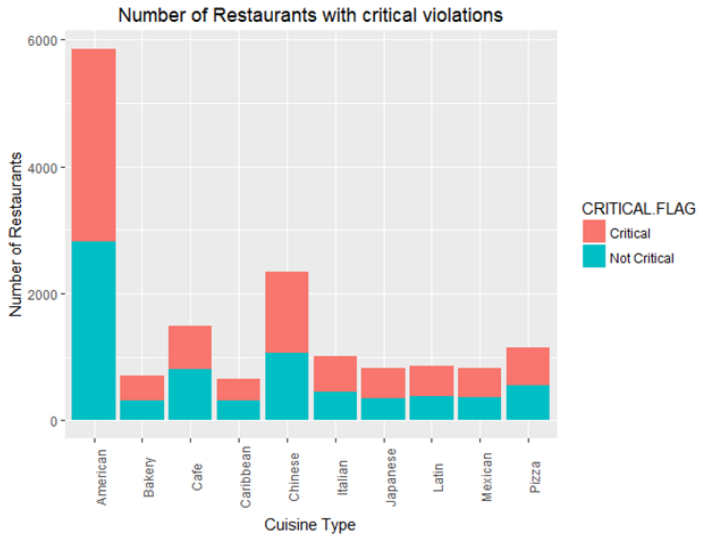


Figure 7: Distribution of Restaurants with Critical Violations for each Cuisine type

For this, we started with grouping the dataset with latest inspection dates by borough and the critical flag. Then using tally() function, we found out the number of restaurants in each group. Following which, we used summarise() function to find percent of critical and non-critical restaurants in each neighborhood as well.

From the chart and table above, Manhattan has the highest number of restaurants with reported critical violations while Staten Island has the least. But when we look at the percentage, Queens has the highest percentage of restaurants with critical violations whereas Bronx has the least. Similarly, Japanese cuisine type has the highest percentage of critical violations whereas Caribbean cuisine type has the least.

7. Choosing the best neighborhood for given choice of cuisine type

Considering a person's choice of cuisine being Caribbean/ Chinese/ Indian/ Latin, we here attempt to recommend the best neighborhood to visit based on his food type. We assume that the person is looking for the best place in NYC, without taking into consideration the distance to be travelled.

Using group_by(), filter() and summarise() functions we compute the average scores of each borough for each cuisine type. We plot below the average scores of each borough for the different cuisine types available.

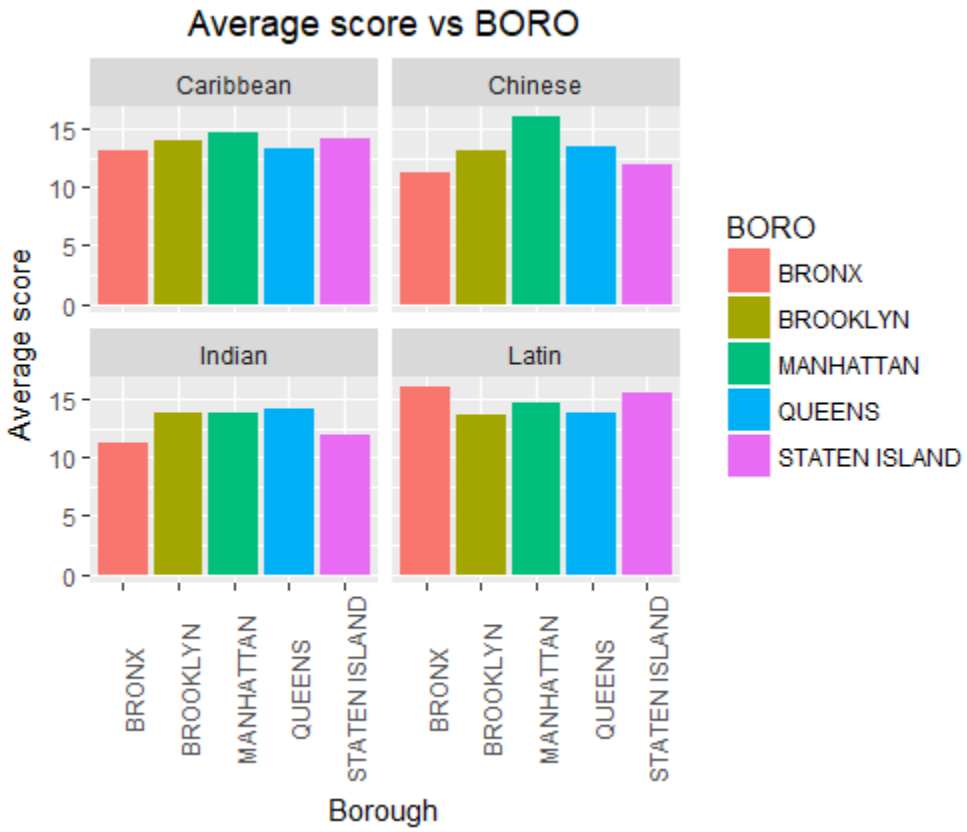


Figure 8: Distribution of Average Scores in each Borough for 4 Cuisine types

Considering for food choice of "Chinese" type, we can clearly observe that Bronx has lower mean score and hence the best available borough. Similarly, for "Indian" cuisine type, we have Bronx as the best borough to head.

This analysis can be further extended to finding the best restaurant for your choice of cuisine in your nearby location by taking location parameters into consideration along with the modelling techniques.

8. Time series analysis of grades and mean scores for pizza chains

We usually refer to the reviews or history of restaurant while choosing the best place, which is very important to make sure you are at the best place. Considering we should choose the best place to have a pizza between Papa John's and Domino's, we focus on the performance of these restaurants on a time scale.

Let us understand the performance of the major pizza chains, Papa John's and Domino's across the entire NYC region. Initially this comparison was made for McDonald's and Domino's which are different types of restaurants, but with feedback during presentation we found it more fitting to compare two similar restaurant types (i.e. Domino's and Papa John's). Observing the percentage of grades, they received each year over the period 2001 to 2012, we can understand the past history of their performances, giving us the review of restaurant chain. This analysis helps us in predicting their grades for the coming years, if we are extending it for further prediction and modelling techniques.

We have used select(), group_by(), filter() and tally() functions to manipulate the data and used facet wrapping technique to visualize the data as follows:

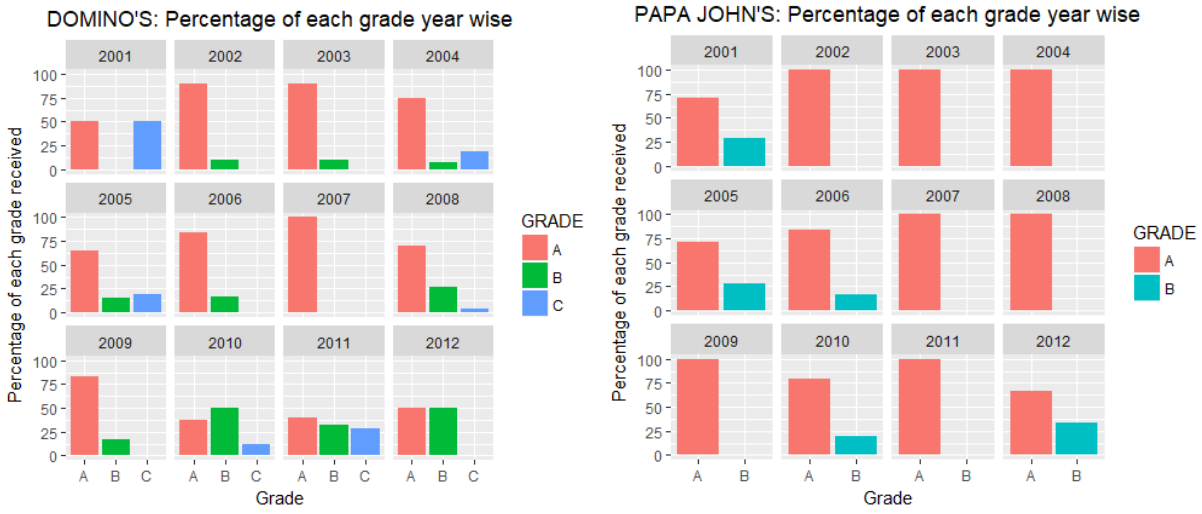


Figure 9: Percentage of Grades received by Domino's and Papa John's for each year

From the above two distributions of grades for Domino's and Papa John's over the period 2001 to 2012, we can observe that Domino's has not been consistently performing good and more over we can observe that it does not show any improvements in the last 3 years. Whereas, observing the distribution of grades for Papa John's, we can notice that it never received a grade of "C" in the last 12 years and have been performing decent compared to Domino's. Even though it does not have an ideal grade distribution, we can suggest or recommend Papa John's over DOMINO's as a better restaurant for pizza.

The scores and the grades are correlated with each other. For an instance let us consider the scoring of the 4 major pizza chains in the NYC region, i.e. Papa John's, Domino's, Pizza Hut and Little Caesars over the time scale of 2001 to 2012.

Initially this comparison was made for popular fast food restaurants which are different types of restaurants, but with feedback during presentation we found it more fitting to draw a comparison of similar restaurant types.

We use ggplot() and geom_line() to visualize the means scores data on a time scale as follows:

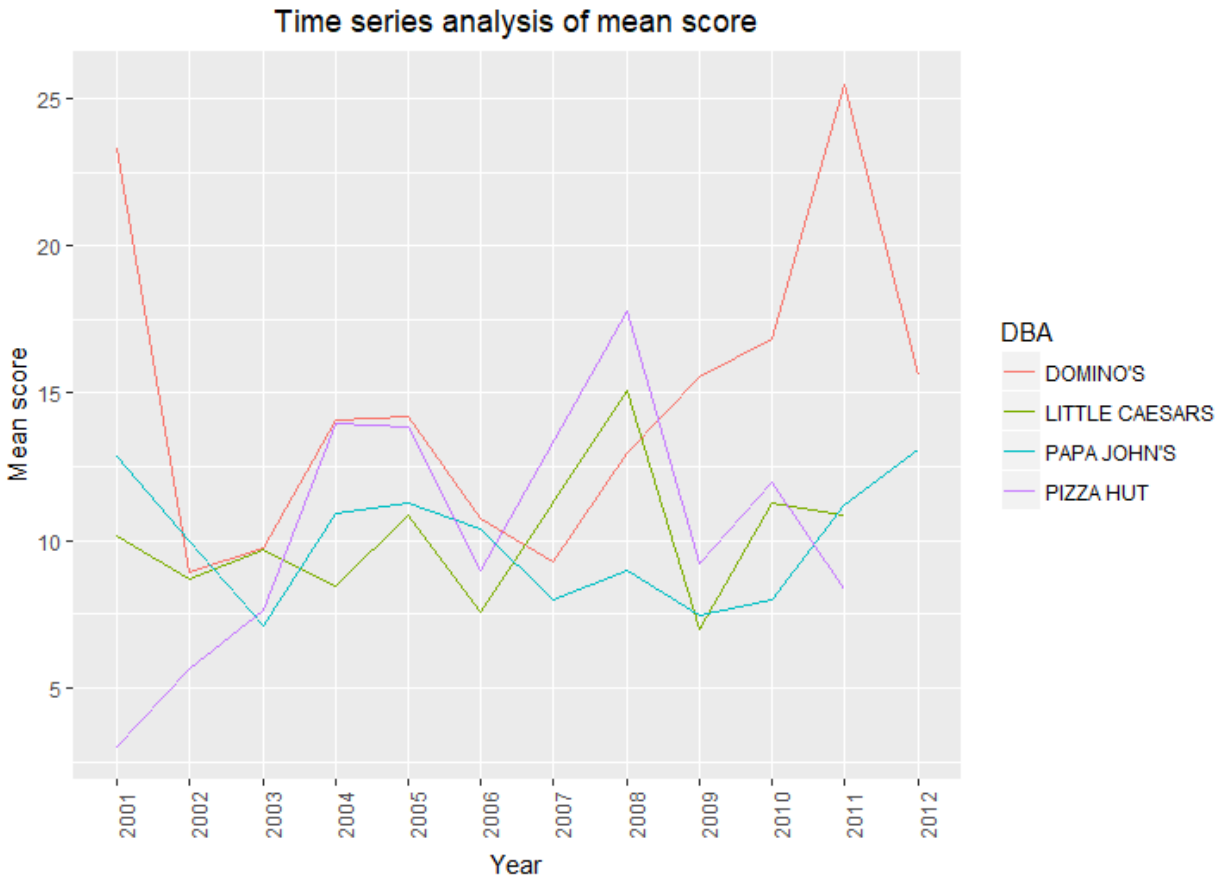


Figure 10: Time scale Plot of Mean Scores

From the above time series plot, we can observe the performance of the 4 restaurant chains, the Little Caesars and papa john's have considerably lower mean scores compared to the rest. Domino's and Pizza hut have extremely varying scores and hence we would not recommend them when having an option of Little Caesars or Papa John's. Following table provides the mean scores for the restaurants over the said period.

We can observe that Little Caesars and Papa John's have lower mean scores and have less variability around it. Hence, among the choice of above 4 restaurants we recommend Little Caesars and Papa John's.

	DBA	MeanScore
1	DOMINO'S	14.727852
2	LITTLE CAESARS	9.963907
3	PAPA JOHN'S	9.945397
4	PIZZA HUT	10.051573

CONCLUSION

After performing the different types of analyses on the data set of NYC restaurant inspection data, we gained some insights and statistics from the data that would be useful to people in the NYC area. First off, we looked at the number of restaurants in each borough to see which borough had the most. From the plot, we found that Manhattan had the most number of restaurants, followed by Brooklyn and Queens next. This make sense as Manhattan is the most populated area of NYC. Then we looked at what kind of health inspection scores these restaurants in these boroughs had, which could be an A, B or C. After looking at the distribution of scores, we found that most of the restaurants had a score of an A with a significantly low amount of Bs and Cs. Looking at the information on how these restaurants are graded by the department of health, restaurants not given an A on its initial inspection will have chances to improve their grade over reinspection in the following months. A time series analysis of popular pizza chains (i.e. Domino's, Pizza Hut, etc.) was also done to observe how they were rated by the department of health over time. After looking at past trends and scores by year from 2001-2012, we concluded that Little Caesar's and Papa John's are better choices in terms of food safety.

In the future, we look to further extend our analysis to focus more on individual restaurants and apply some more advanced concepts such as prediction and modeling. For example, we would pick out a pizza place (i.e. Graziella's in Brooklyn), see how it is performed in the past and try to predict its future inspection scores and grades. In the future, we could also build a model that would recommend the best restaurant based on food type and location choices. We could also perform further statistical analysis of the data using sample statistics and dividing the data into test, validation and verification data sets. We can also expand our model to help not only customers, but to also help any restaurant improve its inspection scores and food safety practices based on past scores and target goal they seek to achieve in the future.

REFERENCES

1. https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j
2. https://www1.nyc.gov/assets/doh/downloads/pdf/rii/how-we-score-grade.pdf
3. https://www.r-project.org/other-docs.html
4. https://docs.python.org/3/tutorial/index.html