Restaurants Unwrapped: Deep Dive Analytics on NYC Restaurant Health Inspection Data

Zach Sickles
11 min read · May 9

Introduction

Restaurants are a vital component of the rich cultural fabric of New York City. They serve various cuisines and accommodate different tastes and preferences from around the globe. In 2019, the restaurant industry in New York City was worth $26.9 billion according to the New York State Department of Taxation and Finance. However, not all restaurants are alike. In fact, many restaurants, even under the same name, differ greatly when it comes to the quality of food, service, and ambiance they provide.

I wanted to explore if there were **certain factors that made a restaurant more likely to pass an inspection compared to others.** Anybody in New York City who would like to go to a restaurant is likely to want to consider how they can obtain the best quality food possible. Additionally, from the perspective of investors in the industry, knowing how to avoid businesses that are likely to receive violations or even be closed can be crucial.

The Data

The first dataset honed in on restaurant inspections that occurred in the five New York City boroughs. It was the core of the analysis. The data was collected by the Department of Health and Mental Hygiene (DOHMH), and I used it to address whether certain violations are more prominent in certain neighborhoods, zip codes, cuisine styles, or amongst various restaurant chains. Details will be considered throughout this blog, but most importantly this dataset involves the regular inspections of each restaurant, so the inspections are evenly spread across each restaurant.

The second dataset that I used provided a comprehensive breakdown of the nutritional content of various fast food products from popular fast food chains, including McDonalds, Burger King, and Subway. With information on calories, fat, carbohydrates, protein, and other key nutrients, the dataset provided a valuable resource for me to analyze and compare the nutritional impact of fast food consumption. This also allowed for a basis of comparison to see if food nutrition impacted restaurant inspection results.

My third dataset provided information about the top 50 fast food chains in America in 2021. This dataset, which included columns such as total sales, total stores, and average sales per unit, provided me with a better understanding of which fast food restaurants in particular were successful in comparison to their competitors. This allowed me to also compare what effect, if any, inspection data or health data had on success.

The fourth dataset focused more on the areas that these restaurants were built in. The dataset was originally developed by Golden Oak Research Group using information from the US Census. The statistics included aspects such as the mean and median household income for each zip code.

The fifth and final dataset was from Opportunity Insights and measured values associated with economic mobility for zip codes across the country. This data set was primarily used for its tallying of volunteering rates, which was calculated using Facebook data. Specifically, according to documentation the data represented “the percentage of Facebook users who

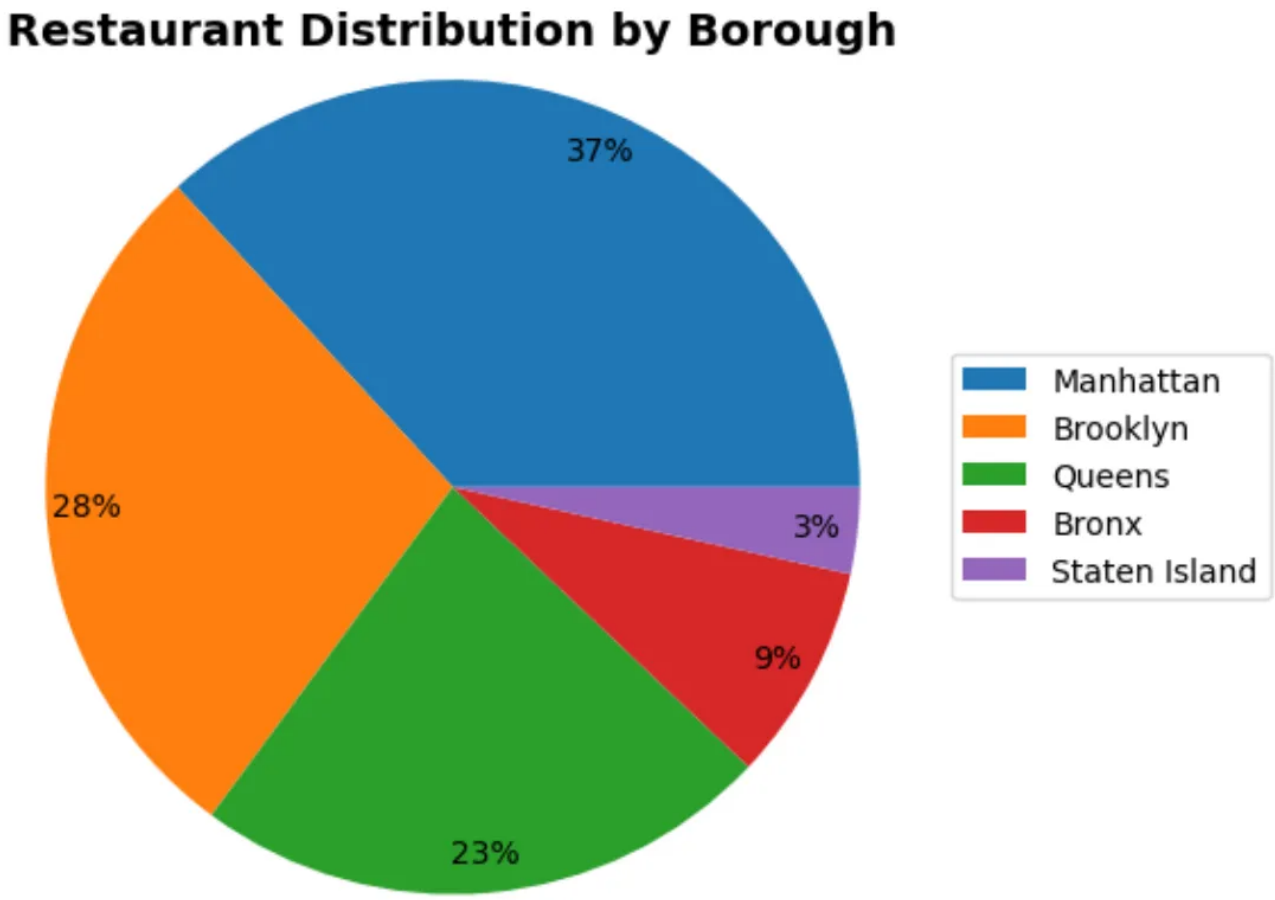
are members of a group which is predicted to be about ‘volunteering’ or ‘activism’ based on group title and other group characteristics” within each zip code.

With these datasets under my belt, I set out to complete a series of insights and analyses that would allow us to better understand the restaurant industry and the factors that contribute to its success or failure.

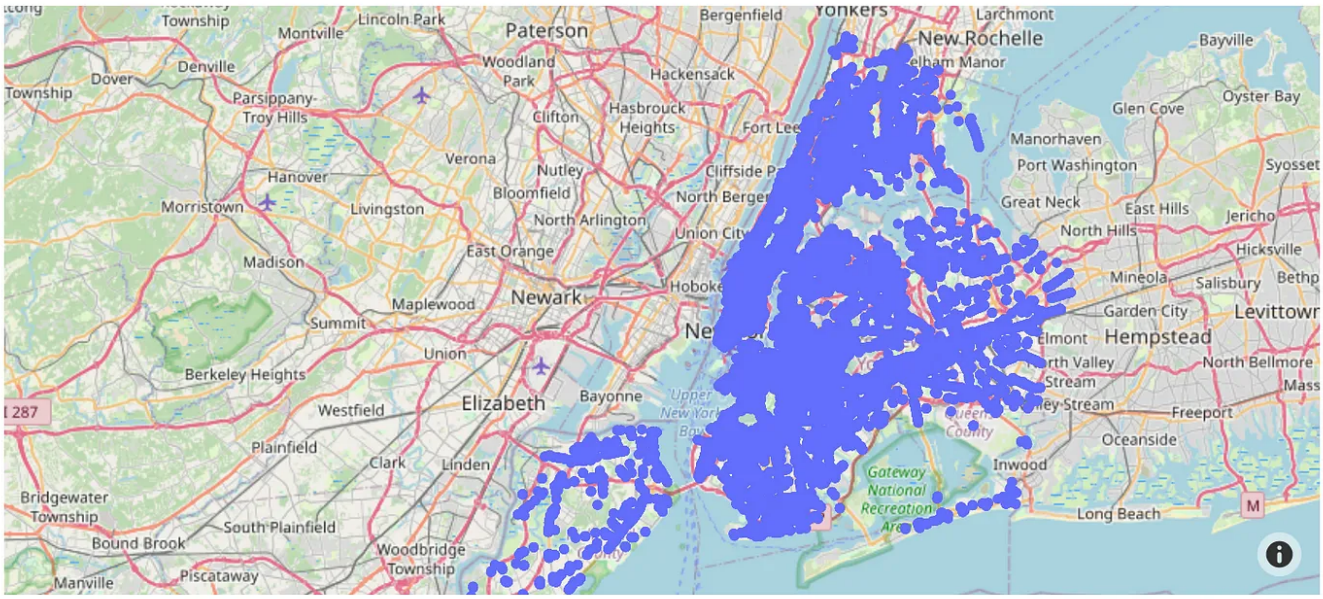
Location Data Analysis

Because I focused the research on New York City, I decided to begin with the most obvious subdivision of the city: its boroughs.

To begin, I created a pie chart that illustrated the distribution of restaurants amongst New York City boroughs. After filtering out any borough with the value of ‘0’ in the borough column (undefined/ no recorded borough), I counted the number of inspections for each remaining borough. Finally, I created the pie chart with a legend indicating the percentage of restaurants in each borough.

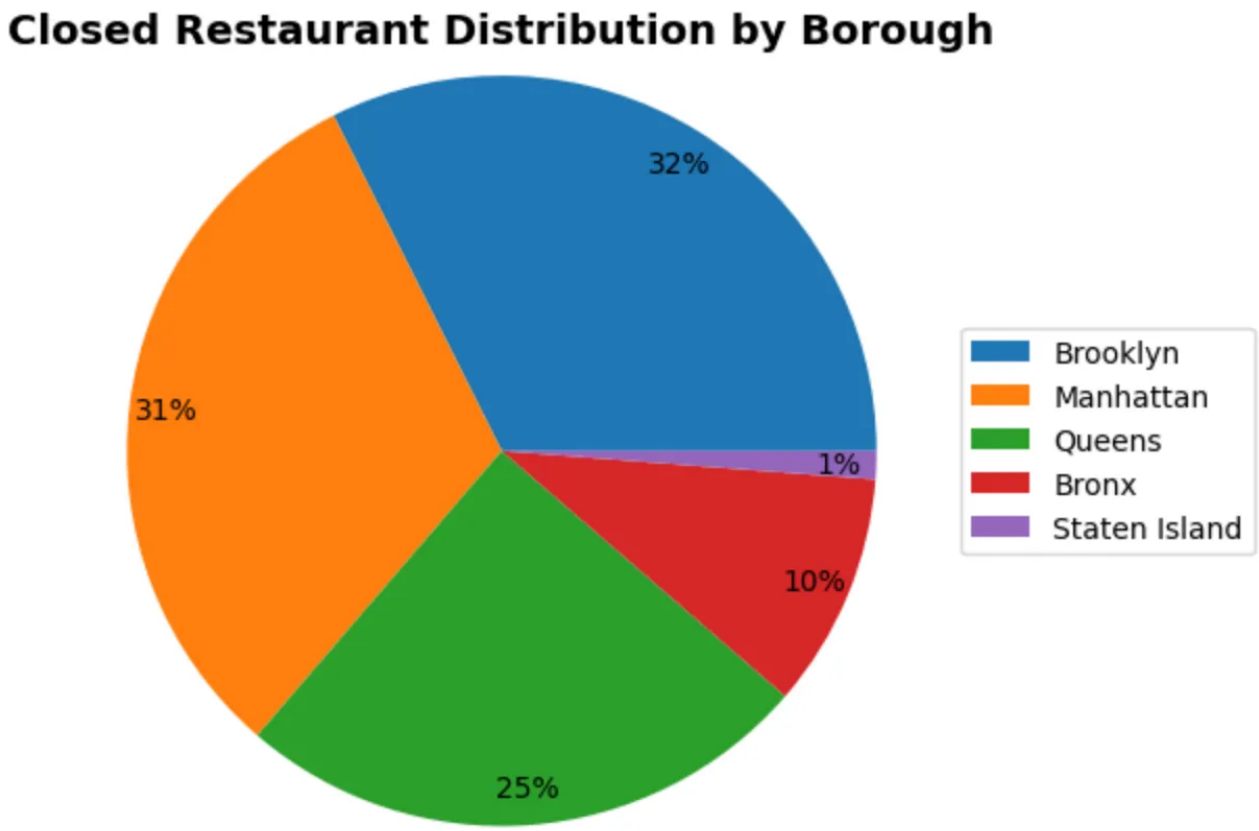


My analysis tracked generally with the population size of each borough. This makes sense — restaurants are largely proportional to population. To get a closer look at how this restaurant spread occurs across the boroughs, I also visualized each restaurant on a map.



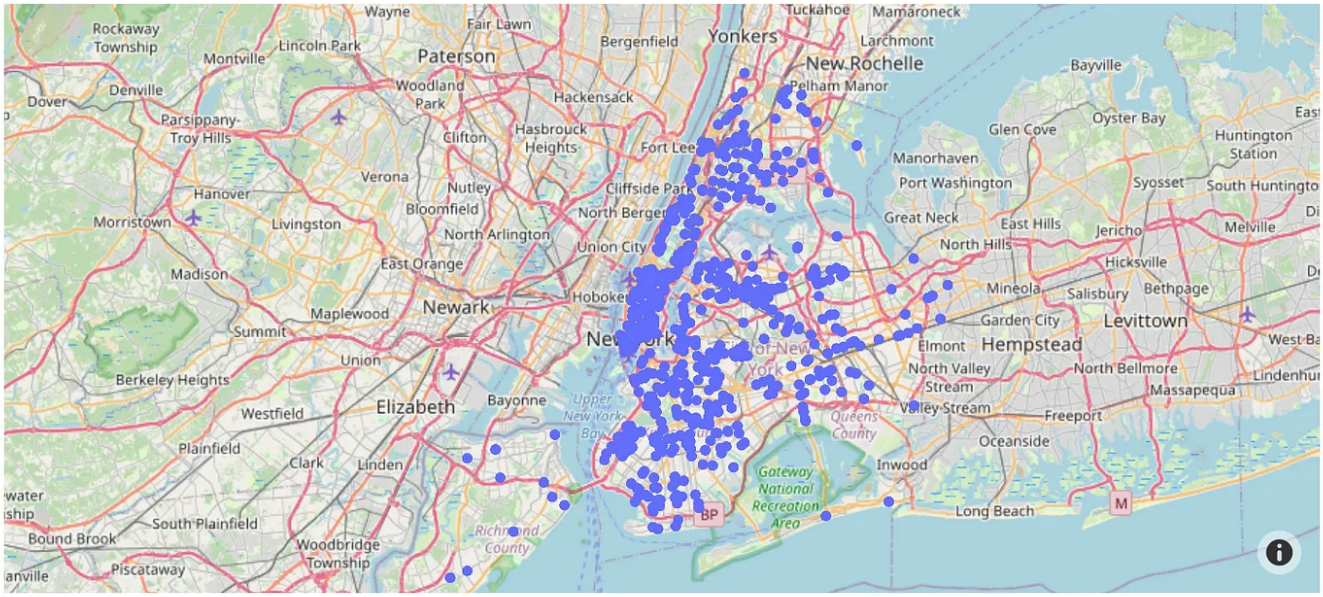
Restaurant Distribution By Location

We can observe here even more clearly that most restaurants are close to downtown Manhattan and nearby in Brooklyn. I then wanted to see if there was a certain borough that had more restaurant closures than the other four. To do this, I first generated another pie chart, this time looking at the distribution of closed restaurants specifically. I counted the number of closed restaurants in each borough. The resulting data is used to generate a pie chart with each borough's percentage of closed restaurants labeled.



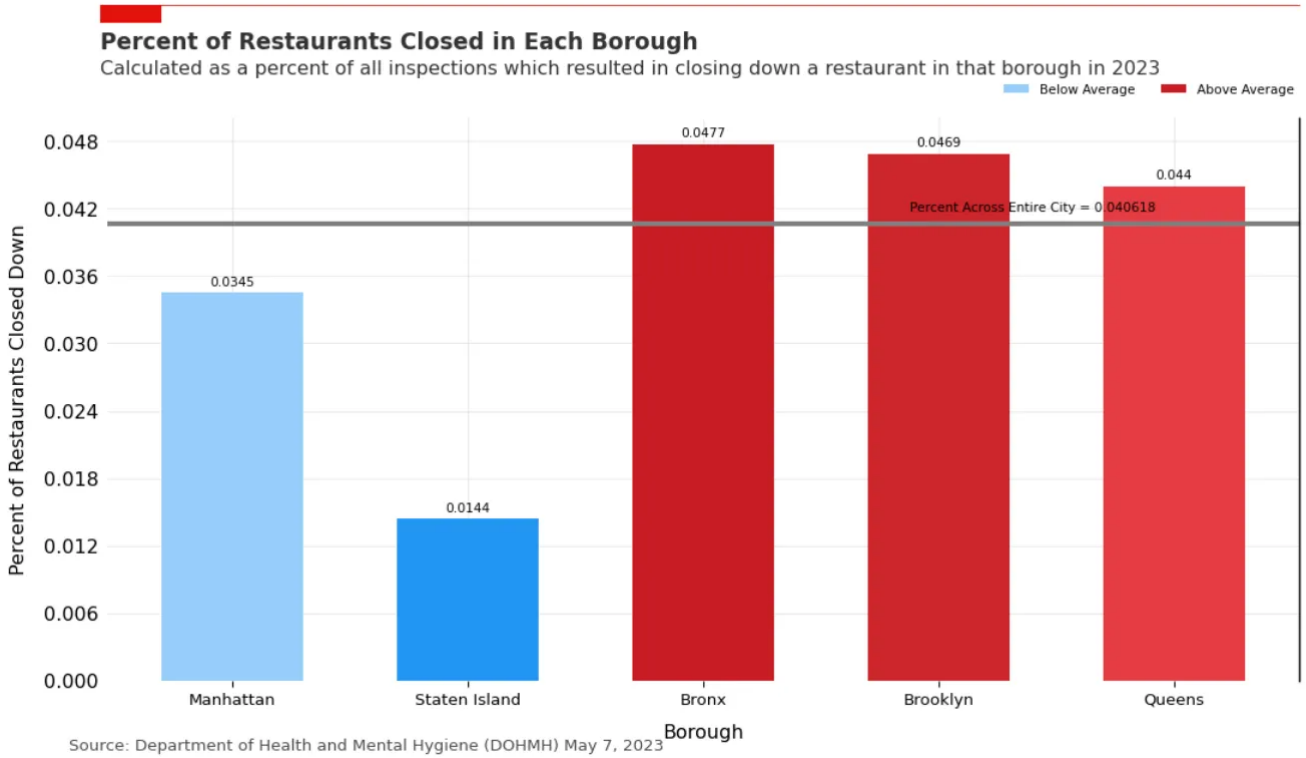
If you look at the two pie charts, there is a clear visual correlation between the number of restaurants in each borough and the number of closed restaurants in each borough. This may seem obvious, considering that areas with more restaurants would be subject to more closures.

Given the information I observed from the pie chart, I continued visualizing the data by plotting the restaurant location data out on the map below.



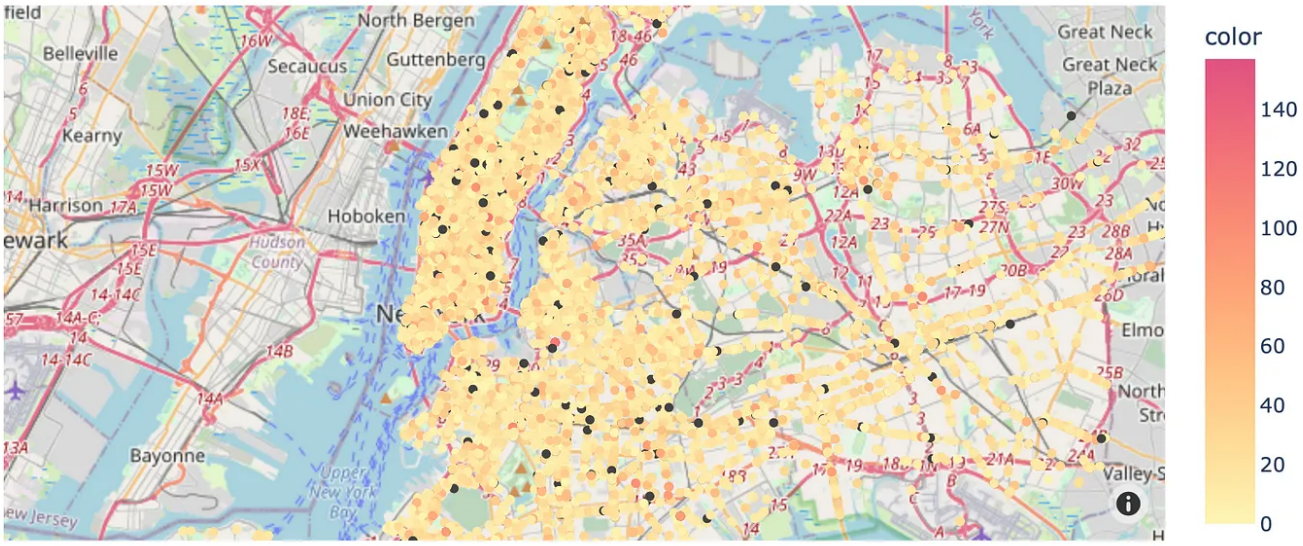
Closed Restaurant Distribution By Location

In order to get a closer look at the percentage of restaurants closed in each borough, I normalized the data and created a bar graph to get a better visualization of the closures across the five boroughs. The chart is divided into two parts, the boroughs where the percentage is above average and the boroughs where the percentage is below average. The average percentage is also plotted as a horizontal line. The chart uses a color map to distinguish between the two groups of boroughs, with red indicating above average and blue indicating below average. This strategy allows us to control for the different amount of restaurants in each borough.



In my analysis I found that Staten Island was significantly less likely to have restaurant closures than other boroughs. Manhattan was also less likely, with the other boroughs maintaining a similarly increased risk.

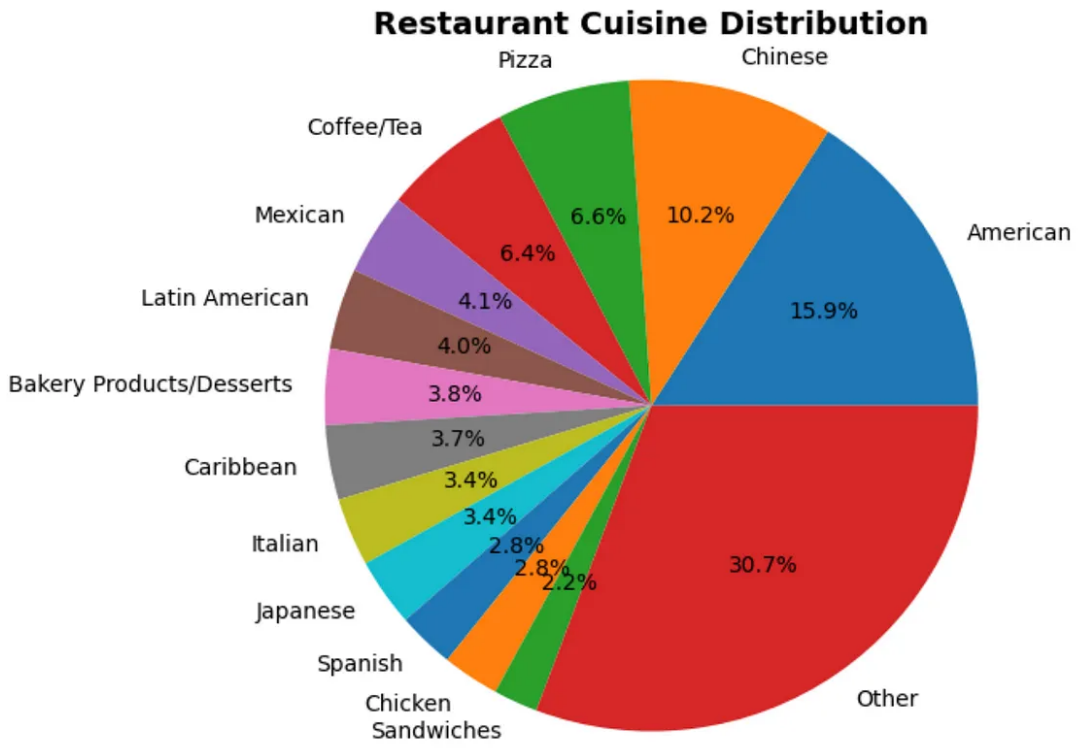
Finally, as another visualization of how location impacts quality, I looked at the inspection score. The higher an inspection score the lower the quality of the establishment. An inspection score of under 13 is an A, and above a 27 is a C or greater, with a 14–26 rated a B. Below I mapped each location inspection with a color code regarding their score. My analysis found that it largely mapped to a similar distribution as the closure data.



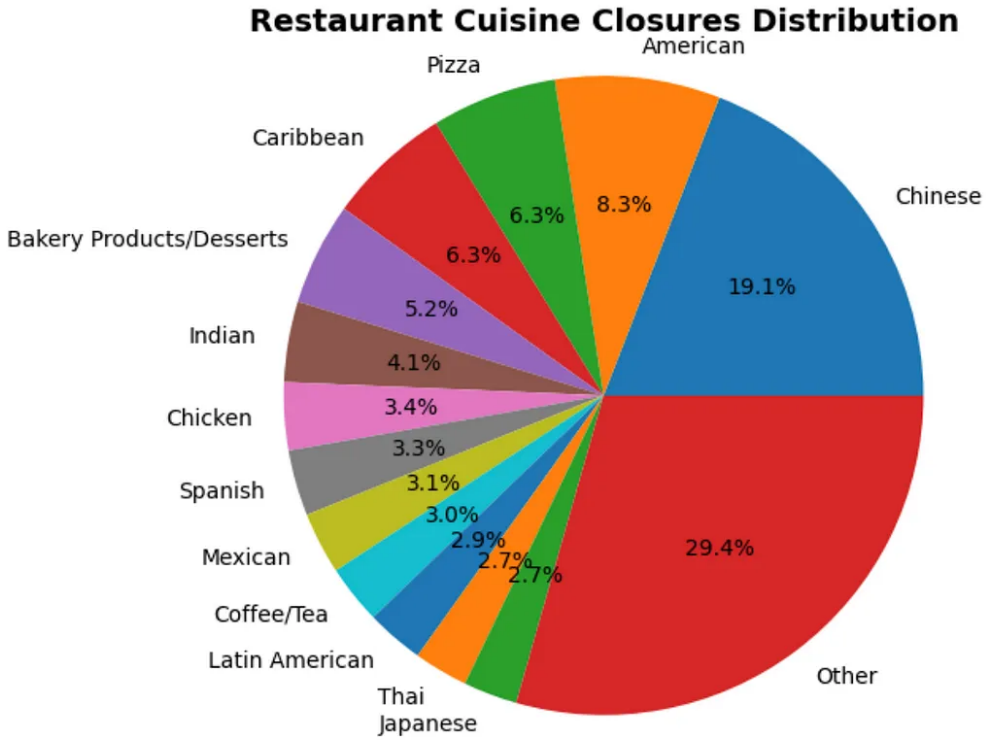
Restaurant Inspection Scores by Location

Cuisine Analysis

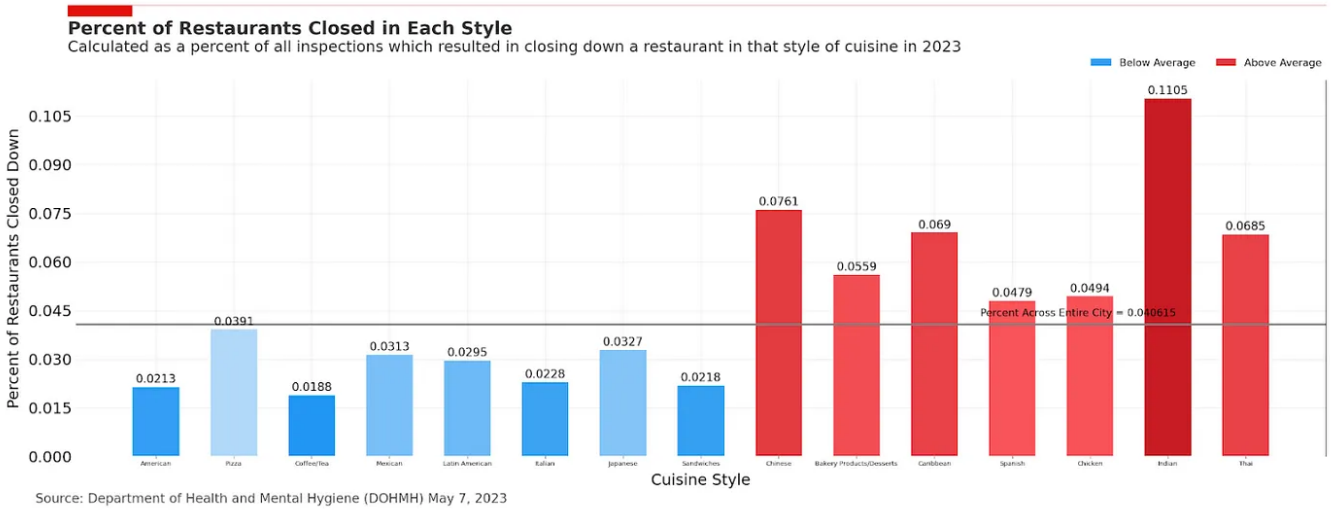
I then looked into the distribution of the type of restaurant cuisine. I observed that American, Chinese, Coffee/Tea, and Pizza are the top four types of cuisines that are common in NYC.



Out of those restaurants, I looked into the cuisine of the restaurants that were closed in 2023. I observed that Chinese, American, Caribbean, and Pizza are the top four types of cuisines that are most common for closing due to violations in terms of overall numbers.

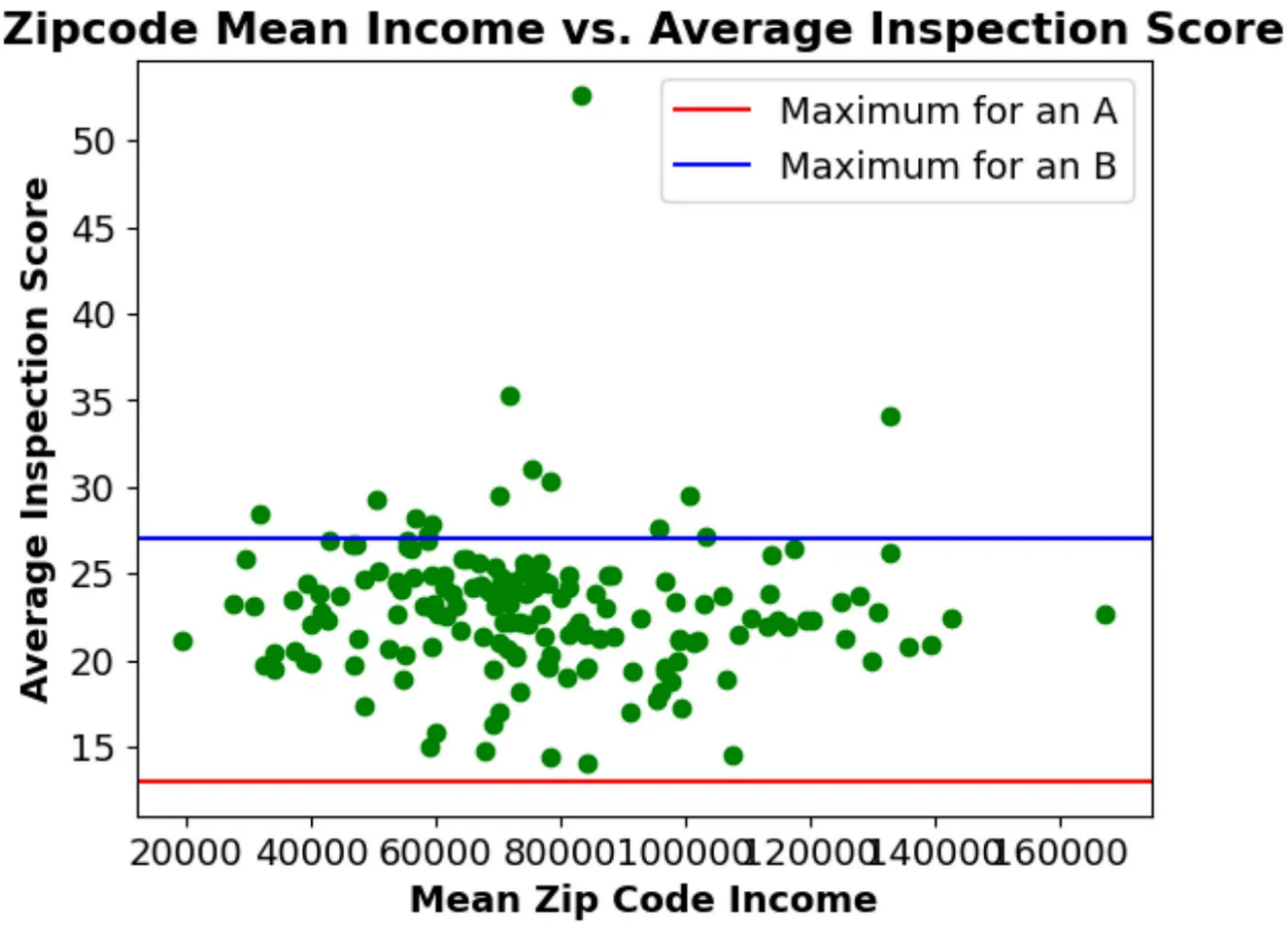


To bring my analysis further, I converted these values into percentages to find the cuisine styles that are the most likely to be closed. That is, what proportion of restaurant inspections of establishments in that cuisine style end up closing it down. I found that Indian restaurants are by far the most likely, with inspections of Indian restaurants in New York City ending in a closure 11% of the time, almost triple the overall average. Chinese, Thai, and Caribbean restaurants were also significantly more likely to be closed. On the other hand, Italian and American restaurants were about half as likely to be closed as the average across all cuisine styles.



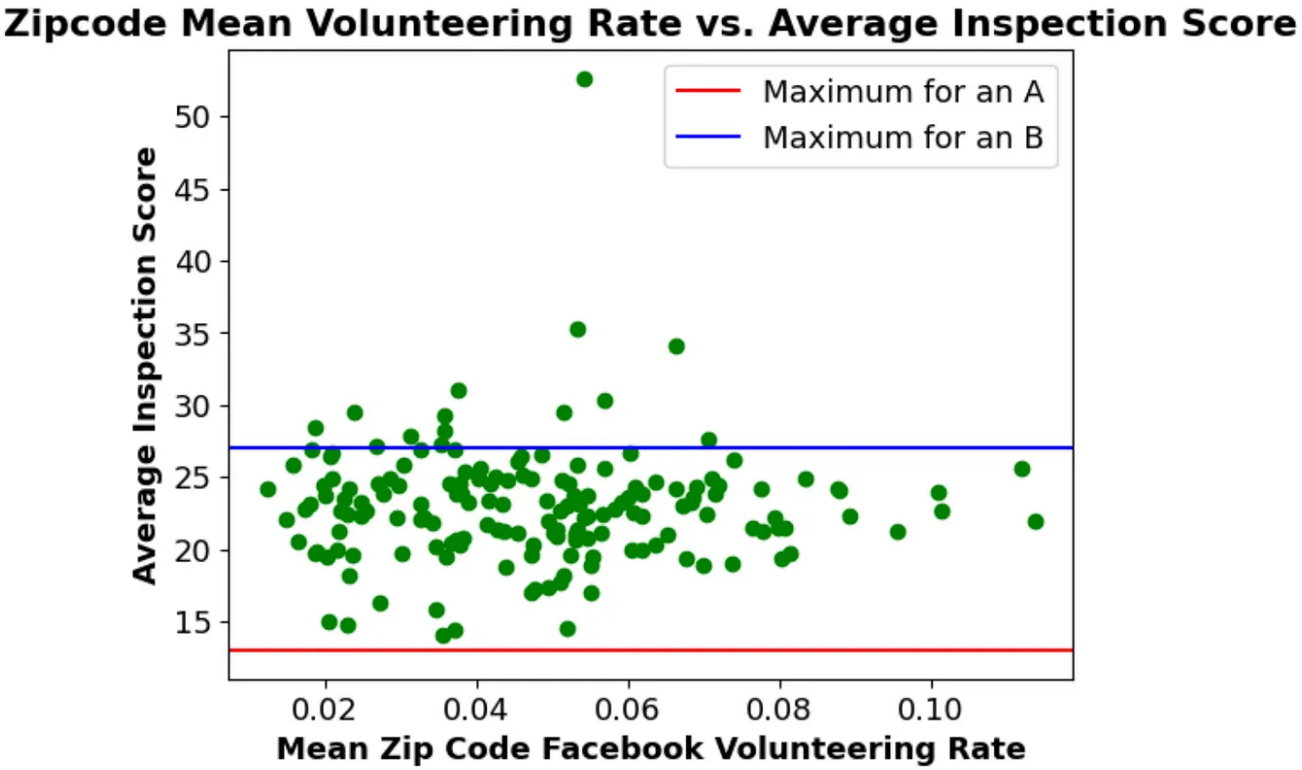
Income and Volunteer Zip Code Analysis

I then wanted to understand more about the areas in which these restaurants are situated. To begin my investigation, I created a scatterplot that examined the mean income of a zip code compared to the average inspection score that restaurants in that zip code received.



While I expected to see a correlation between the income and the inspection score, that was not the case. The income of a zip code did not appear to have an influence on the average inspection score of restaurants within the zip code. This was surprising, as my original expectation was that higher income neighborhoods would have fewer violations due to a clientele which could afford higher resourced establishments.

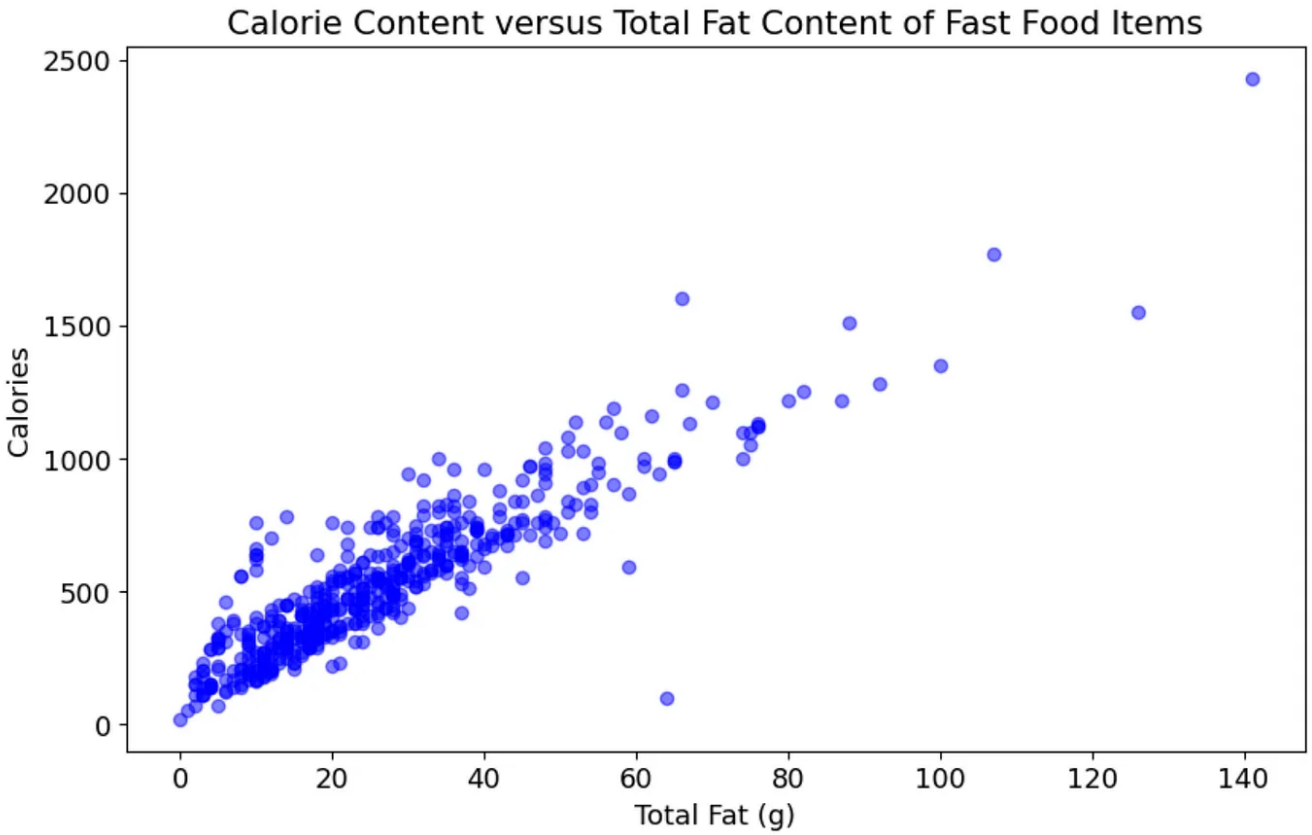
Then I utilized the data regarding the percentage of Facebook users in each zip code which are in groups predicted to be about volunteerism or activism. I created a scatterplot that examined the volunteerism rate of a zip code compared to the average inspection score that restaurants in that zip code received.



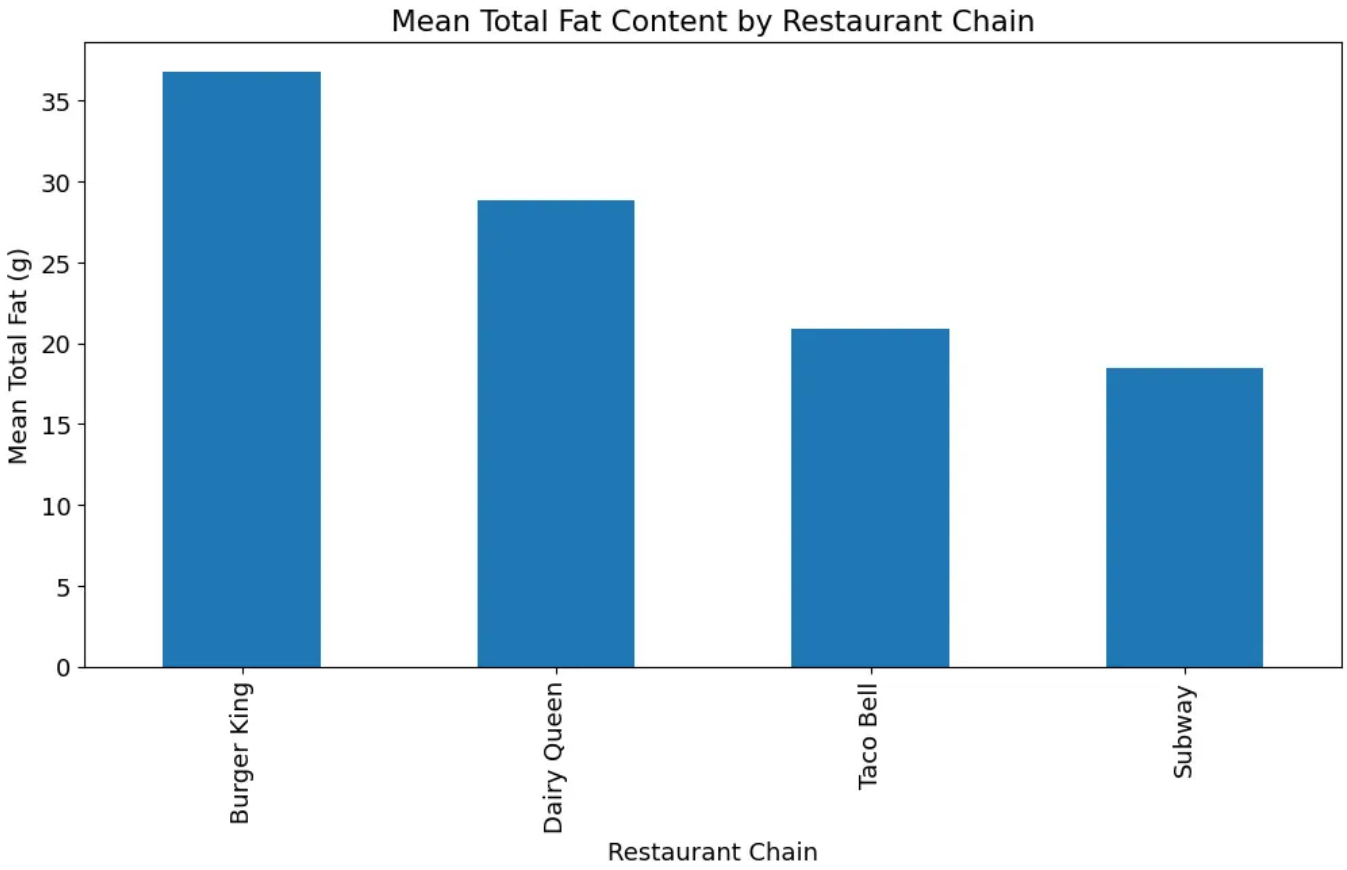
I originally expected to see that passionate and involved citizens would have a positive impact on quality of the restaurants around them. However, there was once again not a meaningful correlation between the rate of volunteerism in an area and the average inspection score of the restaurants in an area.

Nutrition and Sales Analysis

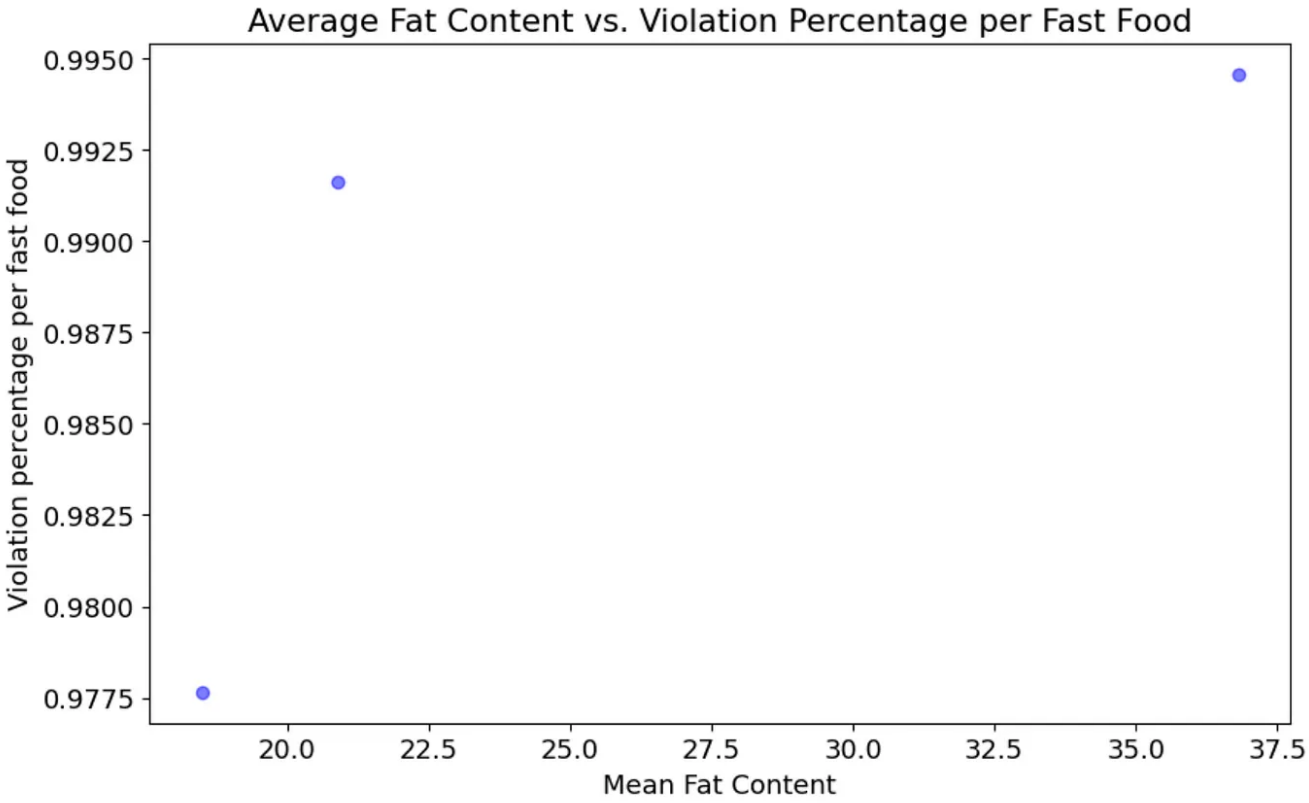
I created a few graphs to explore what exactly these restaurants were serving up. The scatter plot shown shows the caloric content of the food items on the menu of the top 50 fast food chains vs. the total fat contained in the items. The positive correlation shows that the more fat in the item, typically that leads to more calories in the item. This is not surprising.



With such a high correlation between fat content and calories, very often the following fat content data is mirrored by calorie data and is therefor not included in my analysis. Due to their prevalence in each data set, I decided to take a further look at Burger King, Taco Bell, Dairy Queen, and Subway specifically.

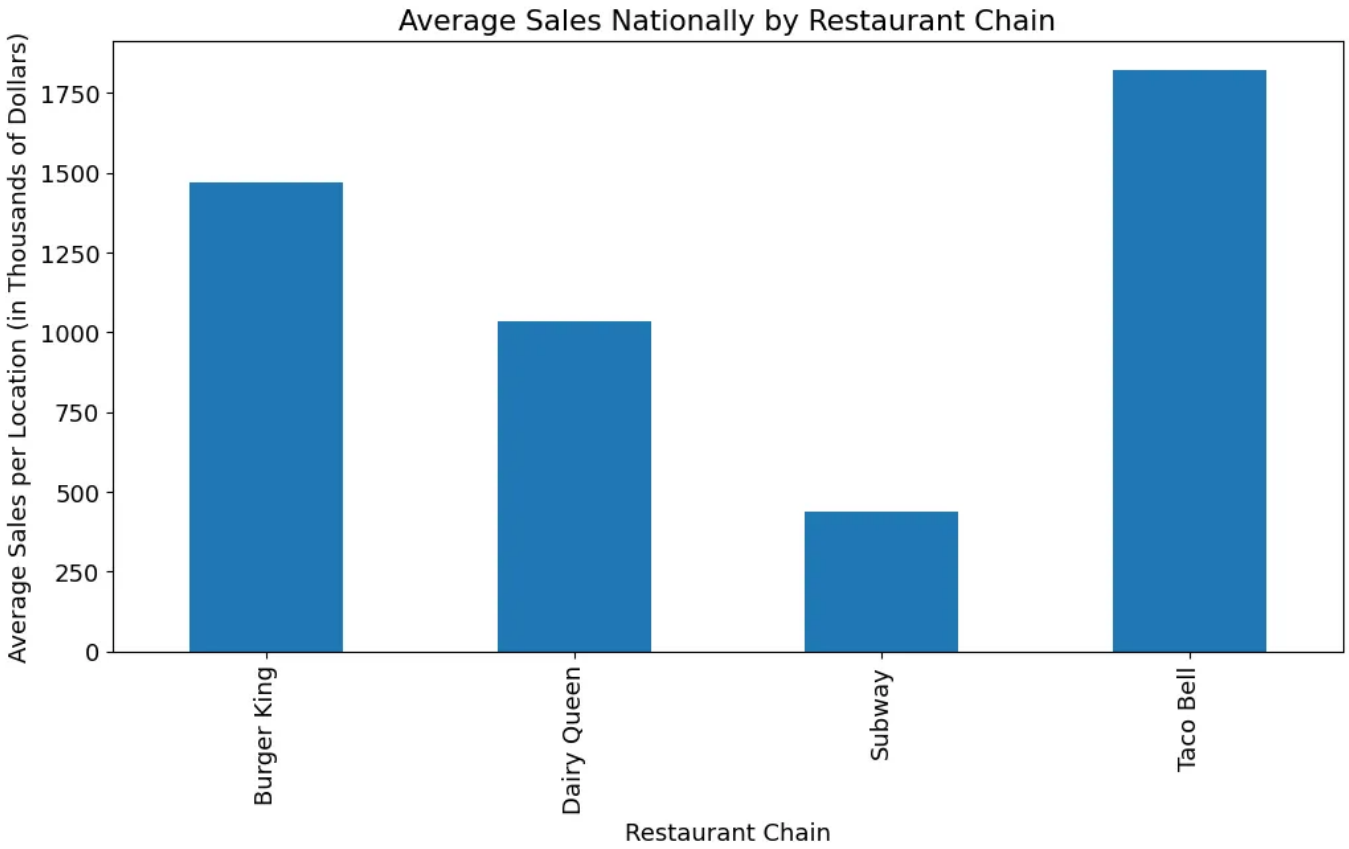


First, I found that Burger King had the highest average fat content, compared to Subway which had the least. Applying these values to my inspection data which did not include Dairy Queen, I got the following results:

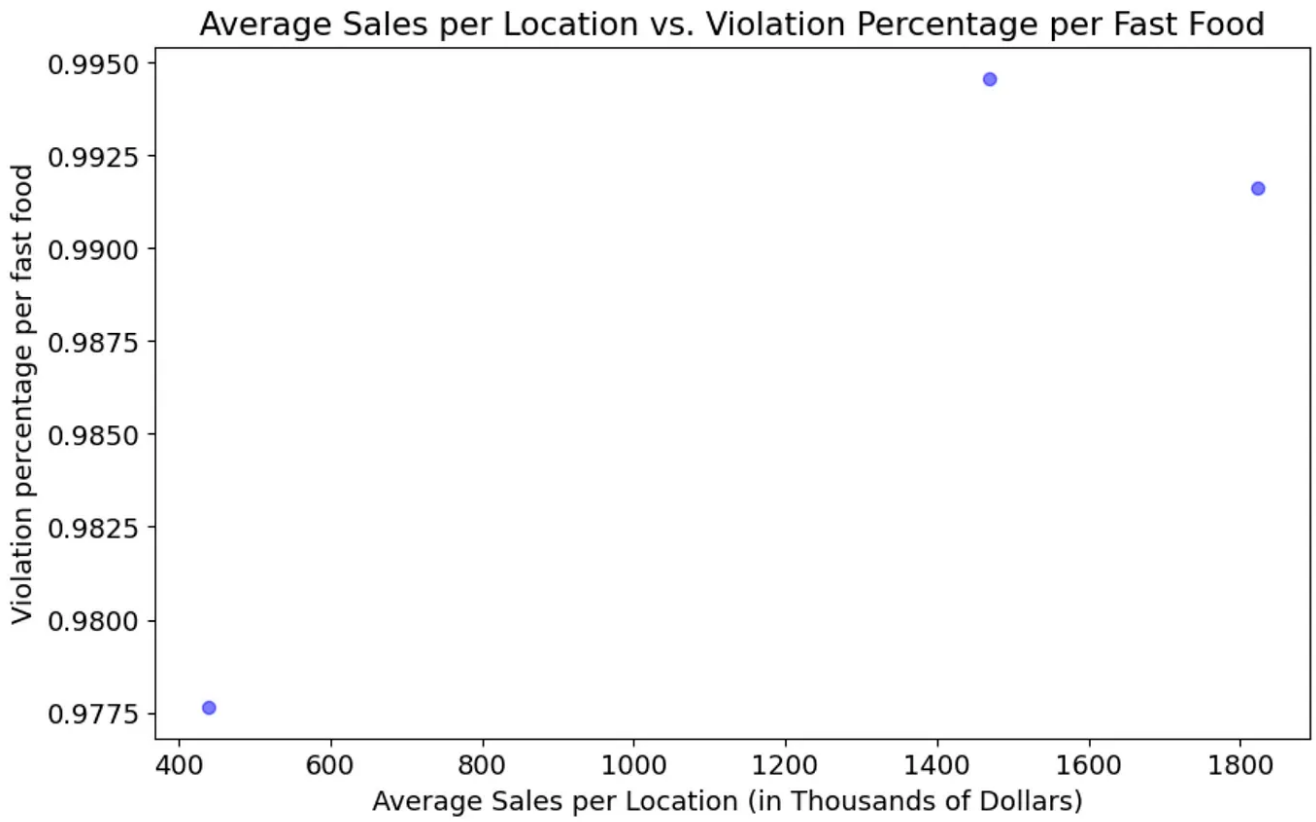


In this scatterplot that explores the relationship between the average fat content in menu items and the concentration of inspections violations, there are only a few dots on the graph so it is hard to come to significant conclusions. However, there is a correlation between higher fat content and more violations. If this trend was to hold true with more data points, it may be because of the related health rules in regards to more meat consumption.

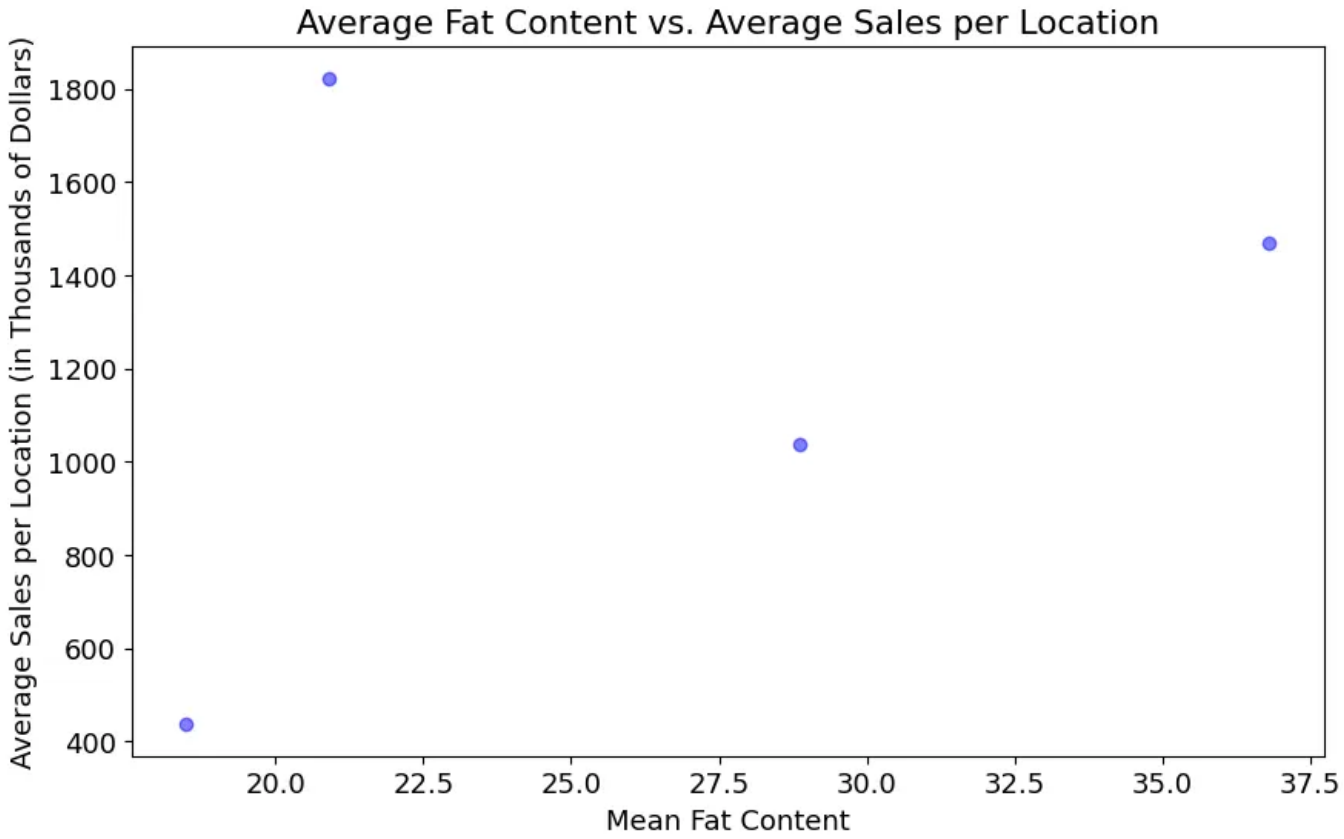
Now, I did a similar analysis of these restaurants, except using data relating to the average sales in thousands of dollars per unit across the United States. I am using this as a proxy for the success of each establishment, to see if there is a correlation between success and these other factors. First I looked at the overall sale totals for each chain.



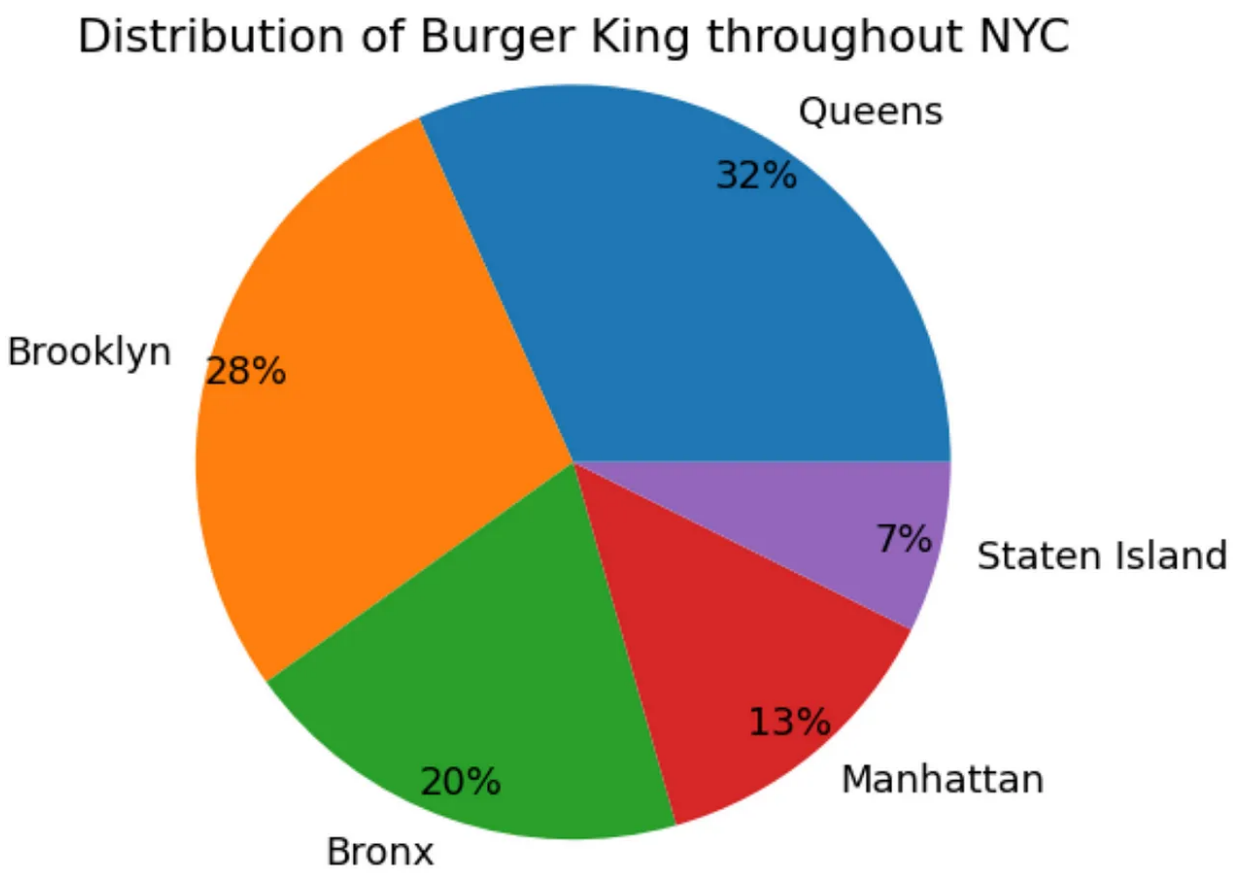
Next, I looked at the relationship between these values and the percent of inspections that acquire violations.

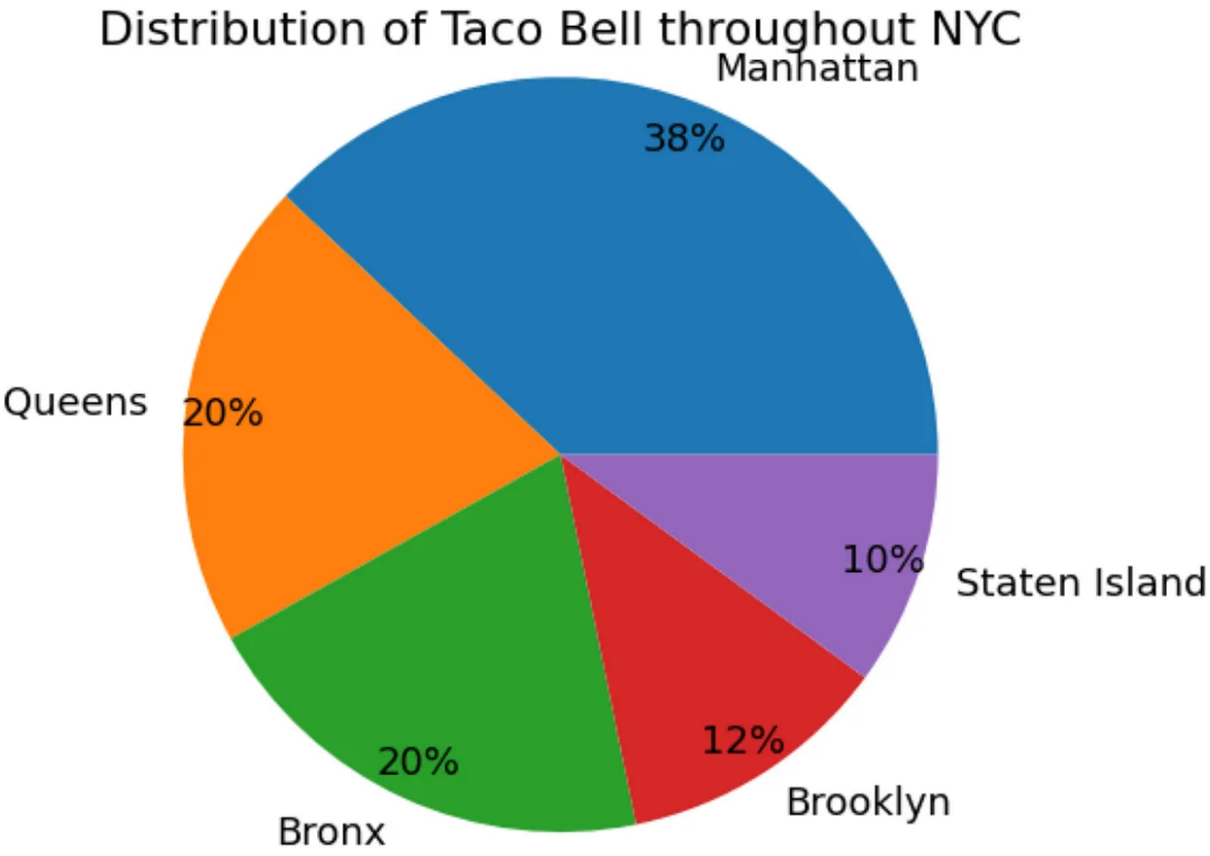


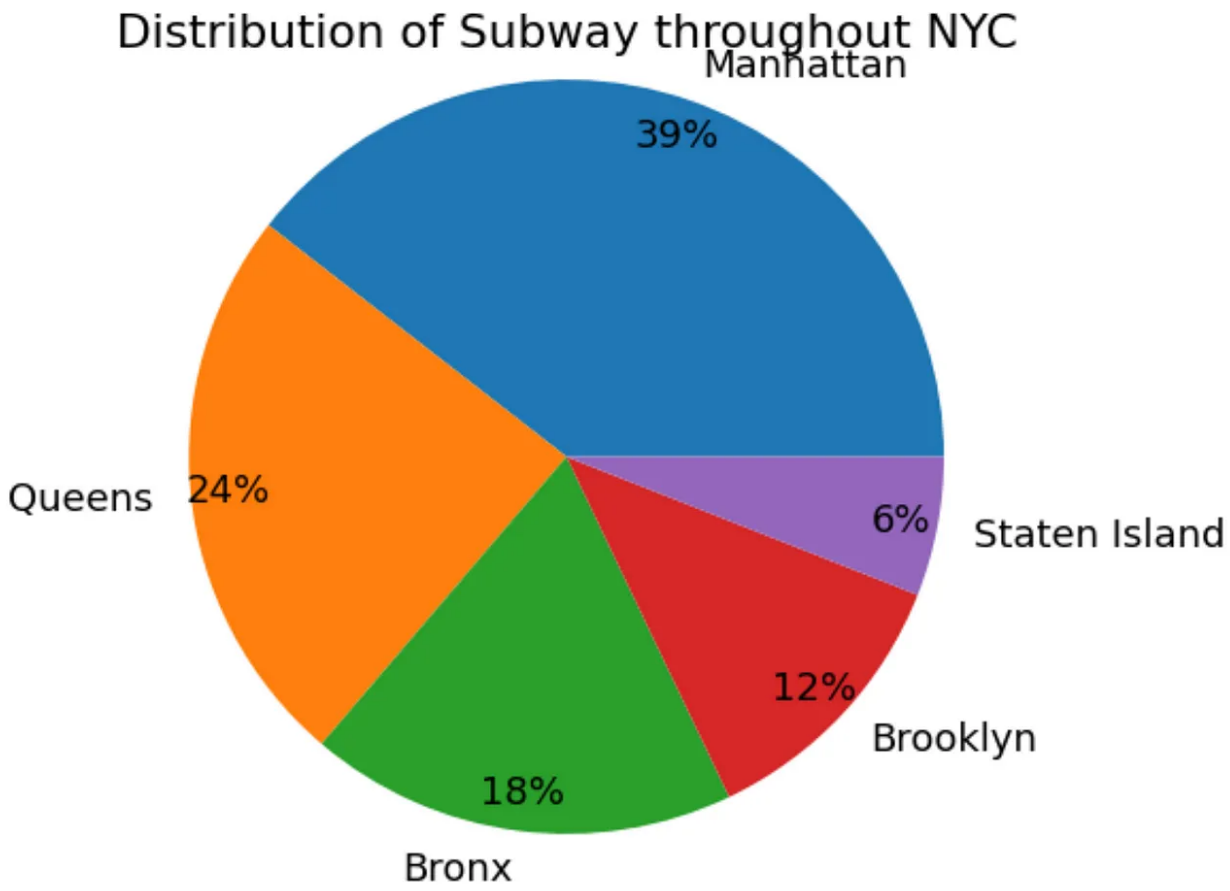
It is impossible to come to any clear conclusion from this data. With Subway being so different from the other two values, it is difficult without more data points to determine if Subway is an outlier, or if more sales equate to more violations. Finally, I looked to see if sales and fat content correlate with each other. I found no correlation between the two variables in the four restaurant chains.



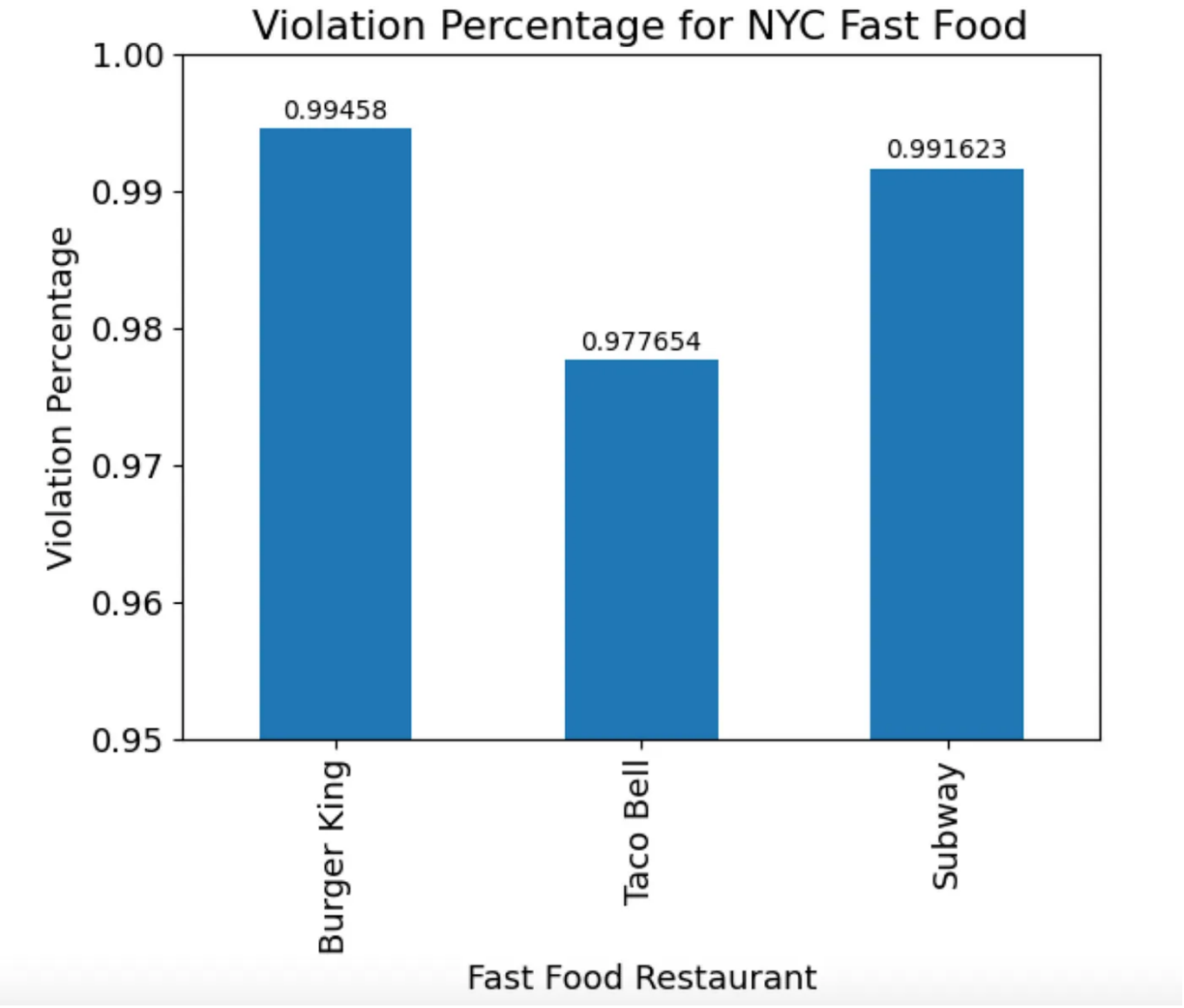
Next, I first plotted the distribution of Burger King, Taco Bell and Subway throughout the boroughs. Graphs are shown below. I was surprised that Queens has the highest number of Burger Kings in the city, as well as Brooklyn's significantly greater proportion of Burger Kings compared to Taco Bell and Subway establishments.


Distribution of Burger King throughout NYC
Queens
32%
Brooklyn
28%
7%
Staten Island
13%
Manhattan
20%
Bronx


Distribution of Taco Bell throughout NYC
Manhattan
38%
Queens
20%
10%
Staten Island
12%
Brooklyn
20%
Bronx



To complete this portion of the assessment, I wanted to visualize the violation percentage of these restaurants overall across the city. This means the percent of inspections that create some sort of violation. Consider that there are many different types of inspections, both serious as well as mundane. Burger King reports the highest percentage of violations followed by Subway.



Restaurant Prediction Model

I decided to also create a simple model to predict if a restaurant will be closed based on the violations that it has. To process the data, I created a feature matrix based on the violations each restaurant has. Each column represents a violation and each row represents a particular restaurant. The target array indicates whether a restaurant is closed or not.

I then designed a simple neural network using keras. The model design is shown below.

```
#create and train model
model = Sequential()
model.add(Dense(60, input_shape=(size,), activation='relu'))
model.add(Dense(128, activation='relu'))
model.add(Dense(60, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
model.fit(xt_matrix, yt, epochs=10, batch_size=10)
```

After training for 10 epochs I got the testing results:

```
785/785 [==============================] - 1s 2ms/step - loss: 0.0349 - accuracy: 0.99
43
Accuracy: 99.43
```

Which showed that you could predict whether a restaurant would be closed to a very high degree of accuracy, 99.43%, using violation data. This model can be used by inspectors trying to be less biased with their decision to close, managers determining any corruption with their monitoring staff, and by restaurant owners deciding which violations they need to consider the most seriously when managing their establishment.

Conclusion

In conclusion, I loved having the chance to look into this data. Analysis of the New York City restaurant industry provided valuable insights into the distribution of restaurants, their inspection scores, and the factors that may contribute to closures. By utilizing multiple datasets, I was able to compare nutritional information of fast food chains, explore the success of fast food restaurants in America, and dive into detailed information about restaurant inspections in the five boroughs of New York City.