Leslie Rosario & Shadman Samad

ECO 32500
Fall 2022
CUNY - CCNY

Your Next Restaurant in NYC Should…

December 20, 2022

Overview

New York City is home to over 27,000 restaurants, disbursed throughout the five boroughs. These diverse businesses in the service industry not only make food, they also add to the melting pot culture that is unique to NYC. With so many options available to New Yorkers to decide where to have their next meal, restaurants have to do more than just serve food to remain in business.

This project intends to guide the decision of opening a restaurant in NYC. Many restaurants have closed due to failure of health inspections that are necessary for operation. All are subject to periodic inspections by The New York City Department of Health and Mental Hygiene (NYC DOHMH), which can alter the reputation and survivability of them. The inspection outcome varies based on the types of violations received and their severity.

We aim to seek factors that influence the failure of an inspection and the survivability of a restaurant in NYC. These factors will then be used to explore the best location to open the restaurant, cuisine type to serve, and when to begin operations.

Our given task was to help a firm open a restaurant in NYC that could be passed down generationally. Essentially, this restaurant must avoid getting closed.

Goals

Our overall goal is to parse existing restaurant and violation data, collected from NYC restaurants, to find patterns that help answer where and what type of restaurant to open that will remain operating for years to come. Therefore we are focused on:

1. **Predicting the best location, time of year, and cuisine type** to open the restaurant under.
2. **Predicting which violations to evade;** those that can result in an immediate failed inspection and eventually the closure of a restaurant.
3. **Choosing a restaurant that can remain running in this competitive market:** all these factors together should provide key insights to opening an operating restaurant.

Data Source

The data used for this project was provided to us in Professor John Droescher's SQL Server. It consisted of a restaurant data table that listed some of the following about 27,483 restaurants in NYC: restaurant name, cuisine description, location (address, district, and coordinates), and camis (unique identifying number for each restaurant permit). The other data table provided contained over 248,000 violations received by these restaurants, recorded between 2015 to August 2022. These records were collected by the NYC DOHMH.

Process of Preparing for Analysis

Identifying the Metrics

To approach the problem of preventing a restaurant to close, the following was considered before diving into the data tables and after meeting with the firm.

1. What violations result in a closure?

   Data needed : violation codes, the inspection types, actions taken by the inspector, violation descriptions, scores.

2. Were there certain areas that had more closures than others?

Data needed: zip code, borough, coordinates.

3. Which types of restaurants have the highest closure rates?

4. Does the time of year impact the likelihood of inspection closures?

Data needed: inspection date with months.

Retrieving the Data

The data was extracted and joined based on the camis key of the restaurant. The columns utilized were: camis, restaurant name ("dba"), camis, borough ("boro"), zipcode, cuisine type ("cuisine_description"), inspection date ("inspection_date"), violation code ("violation_code"), violation description ("violation_description"), score, action, longitude, latitude.

Several queries were created based on only selecting restaurants that have been closed on, all restaurants, all violations, and all violations from restaurants that have been closed on.

Cleaning

- Restaurant data table consisted of "test" camis. This was removed since it does not represent an actual restaurant.
- Violation data table contained:
    - null data entries from January 1, 1900. These rows were removed.
    - restaurants without a zip code or borough. These rows were removed to clearly separate restaurants based on location.

Analysis

Violation Codes

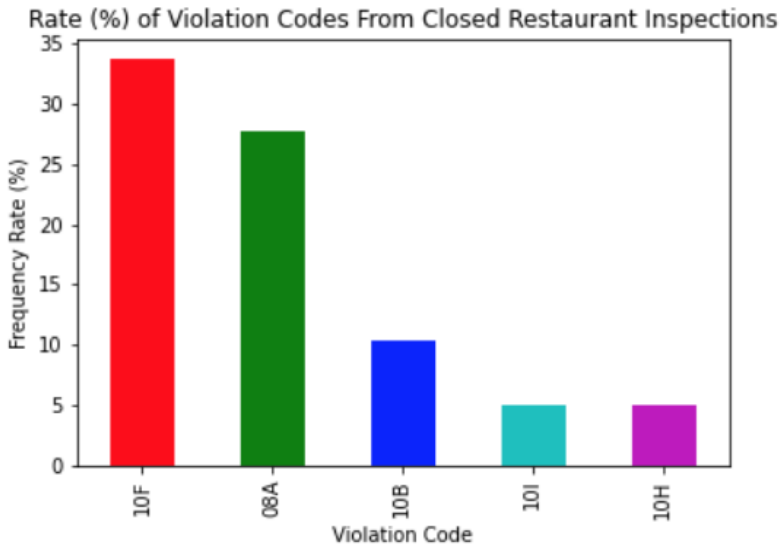
Challenge

Find the most frequent violation codes from inspections that resulted in a restaurant's closure.

Process

A query that retrieved all restaurant violations that resulted in a closure was performed and saved to data frame: closed_restaurant_violations. Then, we obtained the rate (%) of each

violation code's frequency; violation code frequency / total number of violations in the data frame, multiplied by a factor of 100.



Interpretation

Violation code 10F (Non-food contact surface improperly constructed) was the most repeated violation, comprising 14.55% of all the violations. Code 08A (Facility not vermin proof) followed up at 10.30% of violations. These five violations were found in inspections that lead to closures and so they must be avoided.

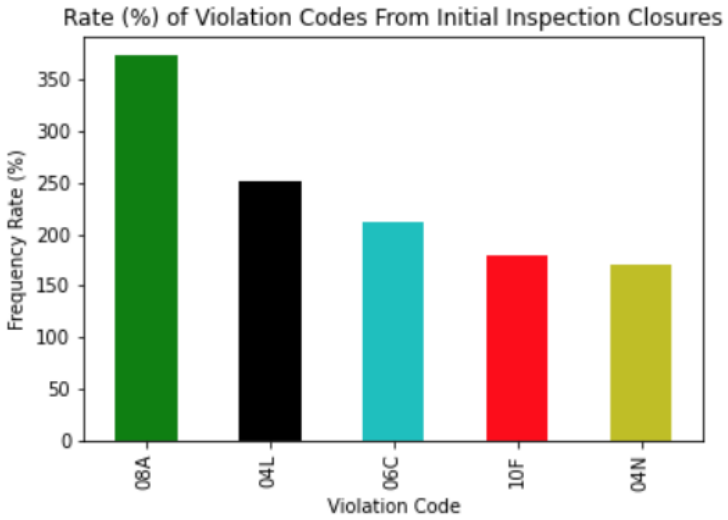
Challenge

Considering some of these restaurants closed on multiple inspections, we want to see what new restaurants have gotten closed on. To do this, the very first inspection a restaurant received that resulted in an immediate closure was extracted.

Process

To extract the first inspections that resulted in a closure, we performed a dense rank over the violation data on inspection date, extracted all rows where dense rank = 1 to get the first inspection date, then filtered for only inspections that resulted in a closure. The same process

was then performed to obtain the rate (%) of each violation code's frequency and graphed the results in a bar chart.



Interpretation

Similarly to the graph with rates of codes from all restaurant closures, code 08A was the most prevalent violation code found in restaurants that closed on their first initial inspection. The next highest was code 04L (evidence of mice or live mice in the establishment's food or non-food area." To avoid the same fate as these restaurants when they were first inspected, these violation codes must be prevented to avoid a closure for the new restaurant.

Location - Borough

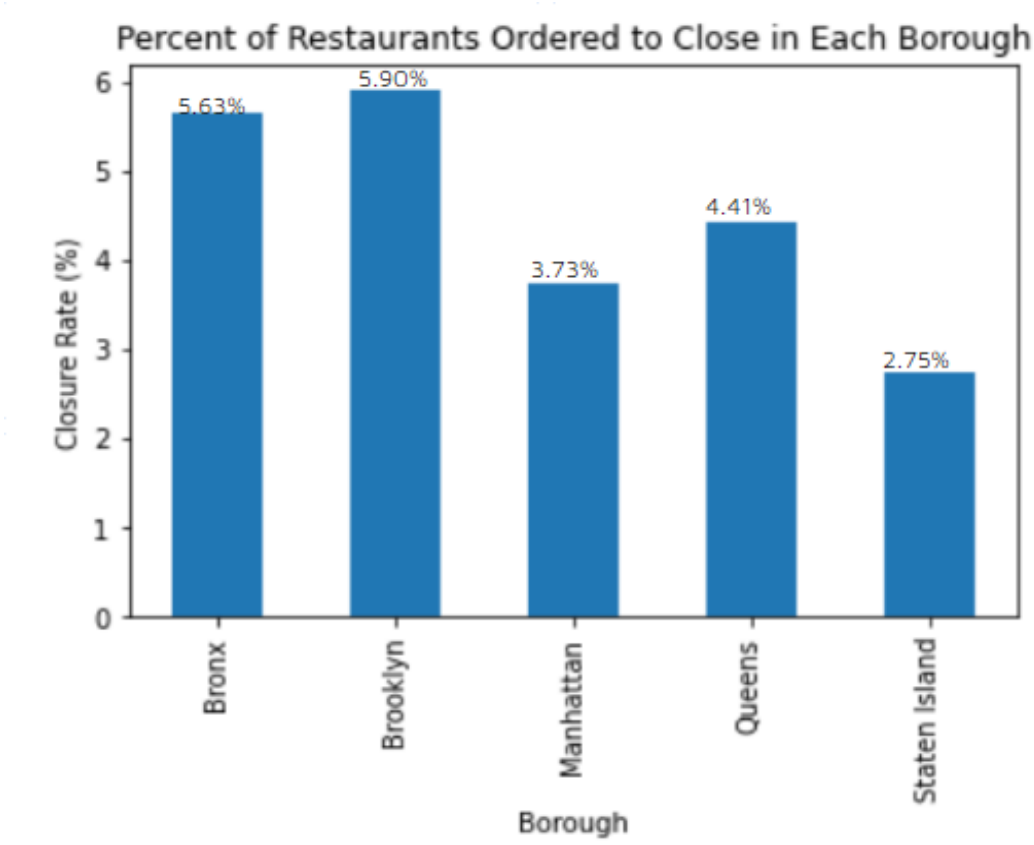
Challenge

Find the most frequent area, in terms of borough, that resulted in a restaurant's closure.

Process

First we created a dataframe which held the information regarding the distinct counts of closed restaurants for each borough. Then we created a separate dataframe in python which held the distinct count of restaurants per borough. The reason that we got distinct counts was to ensure that there would be no restaurant repeated multiple times in the dataframes. Afterwards we got the ratio of closed restaurants by borough by dividing the dataframe which held the information

of the distinct count of closed restaurants by borough by the dataframe which held the information of the distinct count of all restaurants for each borough. This returned the rate of closures for restaurants per borough. Afterwards we graphed the data in python using the matplotlib.pyplot package.



Interpretation

Staten Island had the lowest percentage of restaurant closures, making up only 2.75% out of the total number of restaurants in the borough. On the other hand, Brooklyn had double the rate of restaurants that have closed in their borough. Opening a restaurant in Staten Island or Manhattan would decrease the chances of receiving a closure.

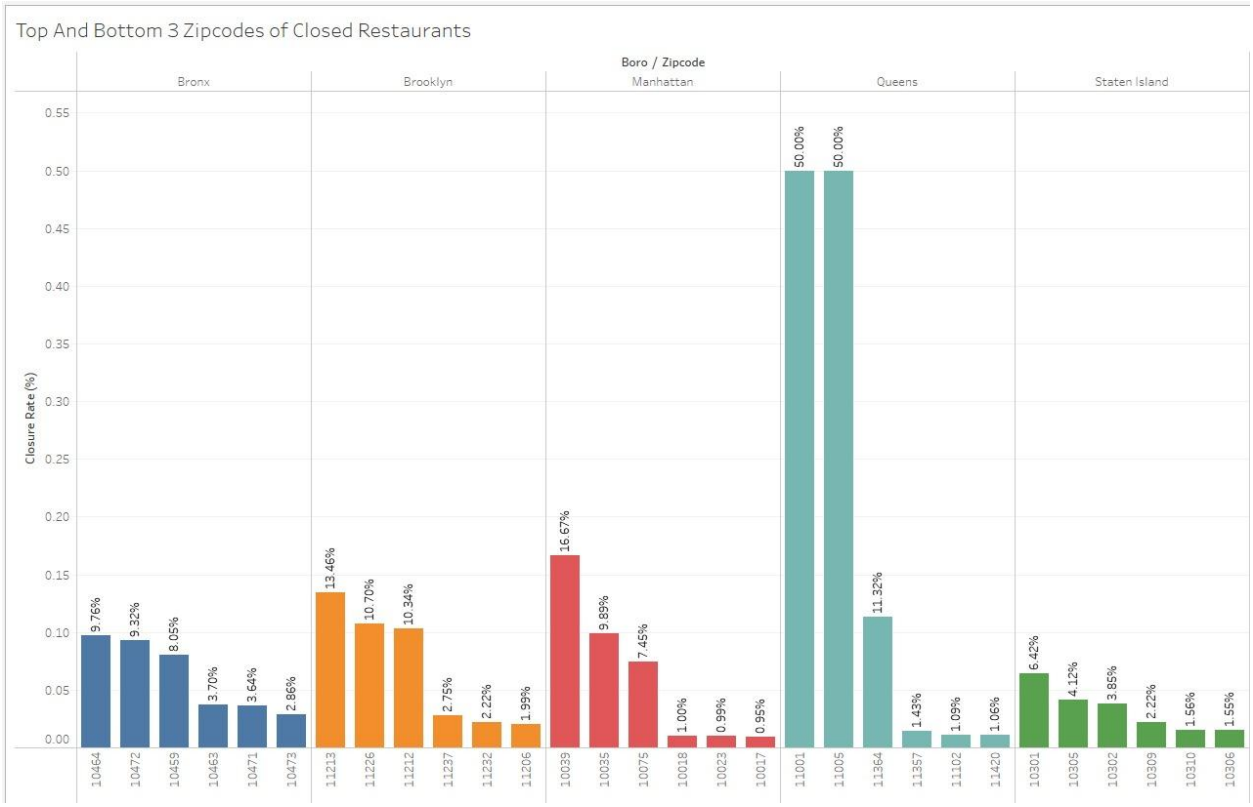
Location - Zip Code And Borough Closures

Challenges

Find the closure rate by zip codes for each borough.

Process

Once we got the closure rates of each borough the next step was to get the closure rates for each zip code within the boroughs. So using the dataframes that we created earlier we added a column to each dataframe to incorporate zip codes. Then we divided the closed restaurants in each zip code by the total number of restaurants in each zip code. After we did this calculation we transferred the results into its own dataframe. The next step was to transfer this data over to tableau which was then used to graph and visualize this data. We then successfully graphed the data in tableau. After graphing this data we then only showed the rate of closures for the top 3 and bottom 3 zip codes for each borough.



Interpretation

A quick glance at borough Queens shows zip codes 11001 and 11005 had half of the restaurants in those areas close at some point of operation. Focusing on the boroughs with the overall lowest closure rates, Staten Island and Manhattan, zip codes like 10017 and 10306 have less than 2% of restaurants that have closed in the areas.

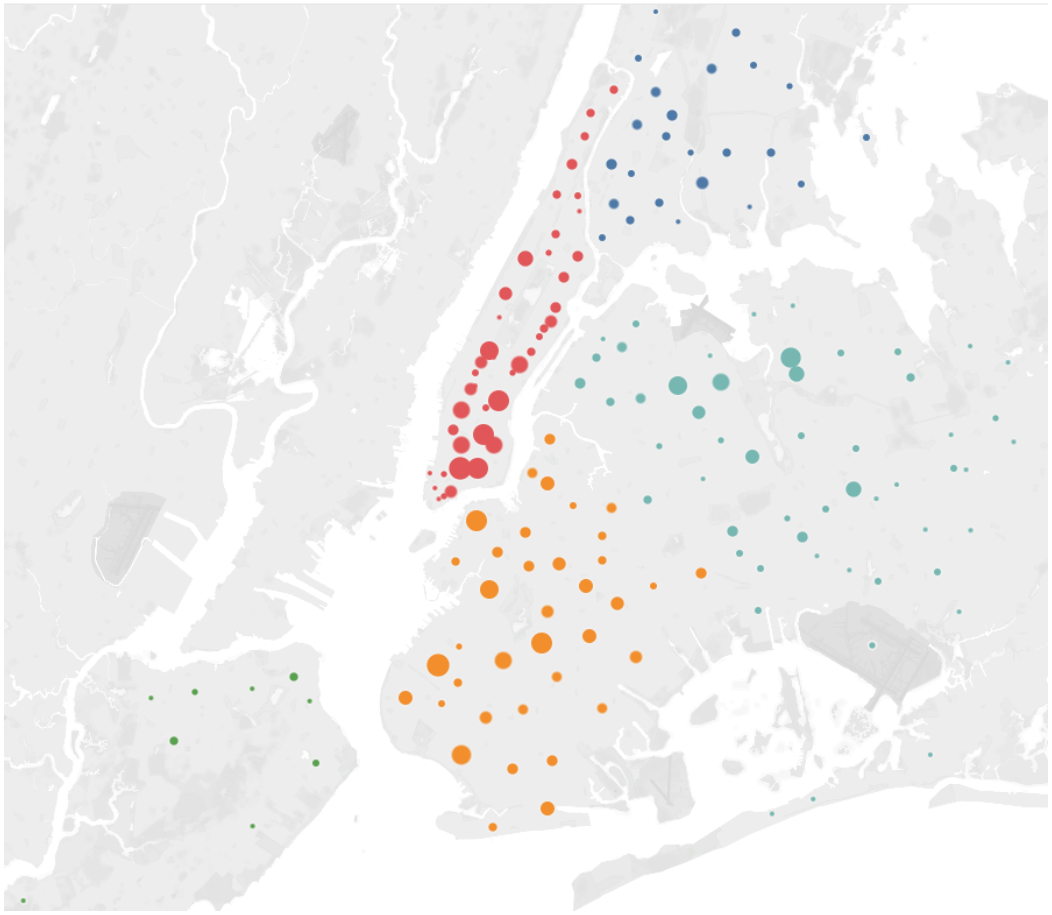
Location | Map of Restaurant Closures Visual

Challenges

During this part of the project we were trying to find a way to visualize the closures by the zipcodes on a map.

Process

To do this, a new dataframe was created from the closed_restaurant_violations data frame to display the distinct count of closures for each zip code. This meant instead of looking at 27,000 restaurants in the data, we parsed just the 1,259 restaurants that have had a closure at some point. This allowed plotting of the points on the map, and size them accordingly based on the count of closed restaurants for each zip code. To graph this map, the longitude and latitude was used from the closed restaurants dataframe to plot all their zip codes and colored based on borough. This assisted in distinguishing the boroughs. The size of each marked zip code was the count of closures in that area.



Cuisine Type

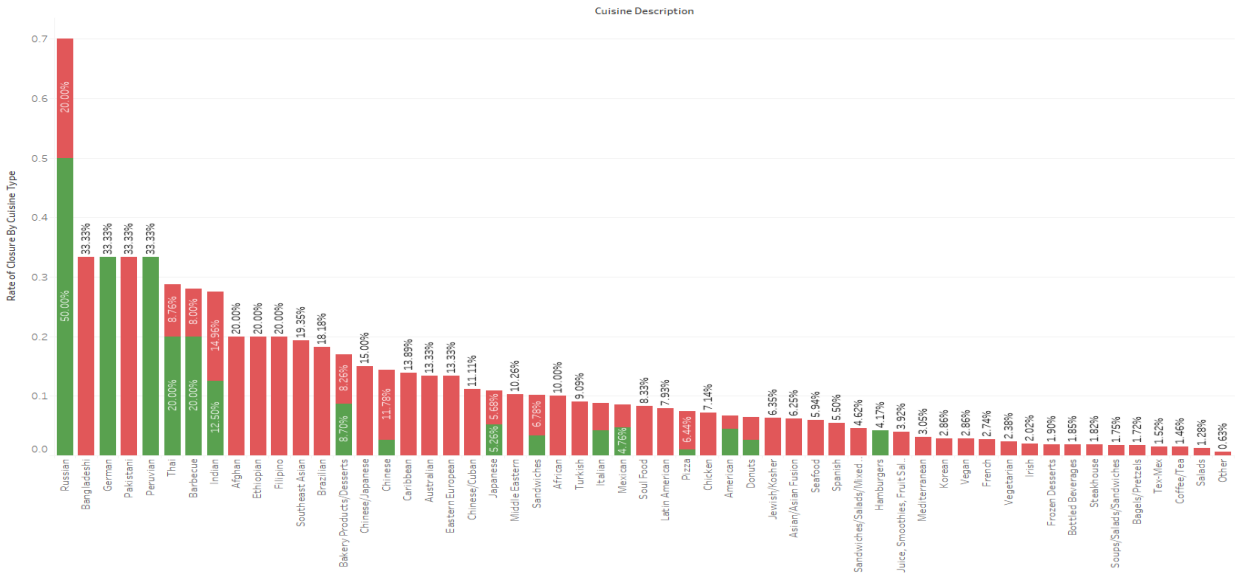
Challenge

Discover relationships with cuisine types and closures.

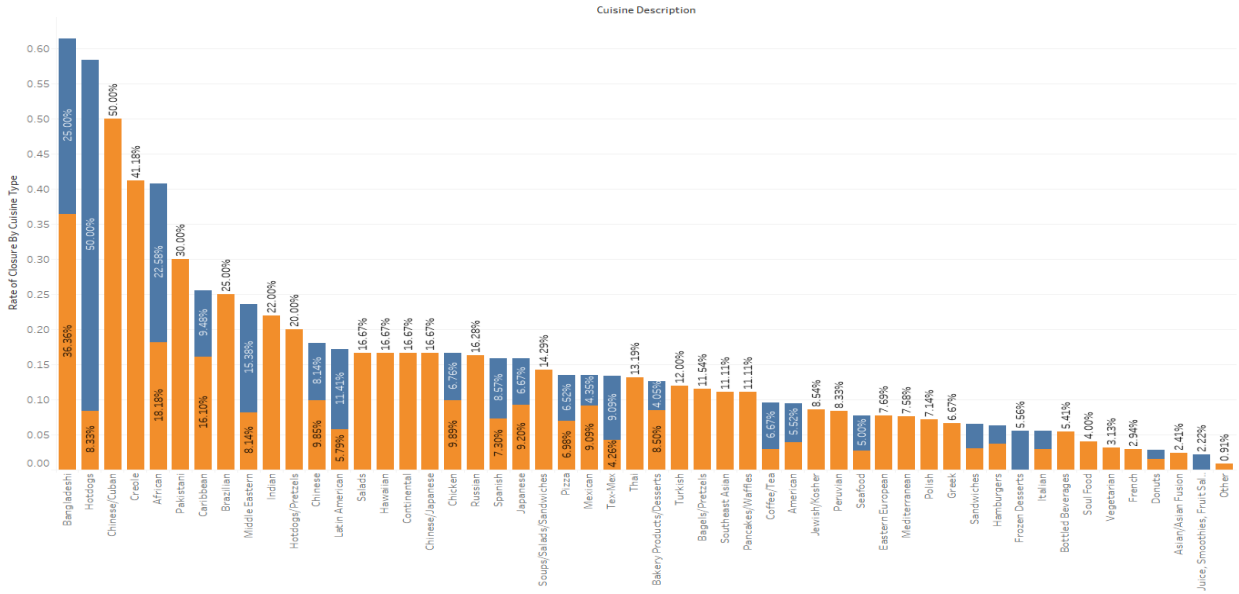
Process

To explore this relationship, the frequency of cuisine type from restaurants that have had a closure were obtained from each borough. To accurately report this, the rate of this frequency was extracted to compare each cuisine type. Data frames were created from the individual closed restaurant data frame, sliced by borough. The data frame was then utilized in tableau to create a stacked bar chart, comparing the cuisine type closures against two boroughs.

Most Common Cuisine of Closed Restaurants by Borough (Manhattan And Staten Island)



Most Common Cuisine of Closed Restaurants by Borough (Bronx And Brooklyn)



Interpretation

Highlighting Staten Island and Manhattan, one should avoid opening a restaurant based on Russian, Bangladeshi, German, Pakistani, and Peruvian. These cuisine types represented the

majority of restaurant closures. Instead, restaurants at the right hand side of the chart, such as Donut shops, would be less likely to close in these boroughs.

When to Open the Restaurant

Challenge

Discover when an inspection resulting in a closure would most likely occur.

Process

For a restaurant to get closed, they'd have to fail 3 consecutive inspections (an initial and two reinspections). Another way to get closed on is to receive a score of over 28 and/or contain a Public Health Hazard. Since the restaurants were scattered throughout these conditions, we decided to extract the month the closure occurred in a new column for the closed restaurant violations data frame, and graph a line chart on the recurrence of a closure by month.

Interpretation

The graph shows closures occurred during the summer months most, followed by the fall. The lowest months of closures occurred were April, May, November, and February.

When to Open a Restaurant (Cont.)

Challenge

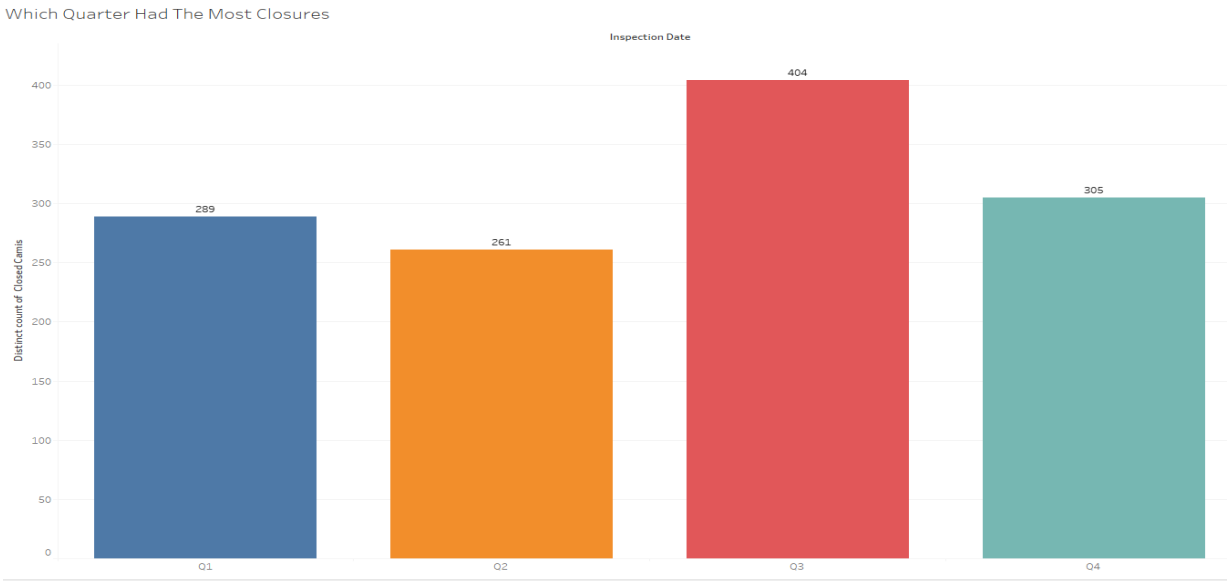
Since there have been new restaurants that have closed on their first received inspection, we wanted to learn what they failed on so as to not make the same mistake. The challenge here was to find out when restaurants, who closed on their original inspection, failed and get closed on.

Process

To extract restaurants that failed their original inspection that resulted in them getting closed on, we created a new query. The query used a CTE to first, dense rank all the inspection dates by restaurant camis. We then called only the restaurants whose ranking equated to 1, since this represented their original inspection. An additional condition was added after this to call only the restaurants that were closed on during their first inspection. This query was then parsed on python to extract the month number and plot the count of these restaurants.

Interpretation

These newly opened restaurants tended to get closed in the late summer to mid Fall. To avoid an early closure, a new restaurant should open just after Mid Fall.



Conclusion

The data revealed important insights for opening a new restaurant in NYC. This restaurant should open in November or December where the lowest amount of closures occurred across the city. It should also be a donut based restaurant opened in boroughs Staten Island or Manhattan. The data recommends avoiding opening a Russian, German, or Peruvian restaurant in Staten Island. Also, avoid Bangladeshi, Pakistani based restaurants in Manhattan. Finally, take extra precautions in avoiding violation codes: 08A, 04L, 06C, 10F, 04N, 06C, 10H, 10I to prevent a closure.

Future Recommendations

To improve on this project, we recommend to utilize regression modeling to deeper explore the relationships found. Feedback from our instructor suggests exploring what caused inspection failures and when did they happen for restaurants that closed on their second reinspection. Finally, the project would benefit from exploring the violations that restaurants who never received a failing inspection receive, so we'd know which violations wouldn't pose a grand threat to the restaurant's success.

Acknowledgments

Professor John Droescher

Stackoverflow

pandas.pydata.org