Eating “Clean” in New York City

Abstract

What is NYC known for? Good food and big rats. Our project aims to help residents and tourists both figure out exactly where they do not want to eat.

We will do this by reviewing 26,000 restaurants that have received health code violations between 2010 and 2017. We will count how many violations each restaurant has received, count how many rat sightings have occured in the neighborhoods in which they are located, measure how far each restaurant is from the nearest subway, and finally integrate Google user reviews to form our own scorecard for each restaurant and determine the grossest restaurant in New York.

In New York, dining out is one of the best things to do, and we would like to know which restaurants we absolutely should avoid. WE WILL FIND YOU, WE WILL EXPOSE YOU!

Data Sources

- https://www.kaggle.com/new-york-city/nyc-inspections

- https://www.kaggle.com/new-york-city/nyc-rat-sightings

- Google Places API

- Google Distance Matrix API

Steps

1. Data Consolidation of 2 CSV files

2. User Reviews and Lat/Long from Google Places API

3. Distance from Subways using Google Distance Matrix API

4. Make Visualizaitons

Proposed Visualizations

1. Data Breakdown

2. Correlation of Variables

3. Scorecard

Research Questions

1. What is the worst NYC restaurant based on our algorithm? (Research Quesiton)
   1. Cut by Borough
   2. Cut by Cuisine Type
2. What Borough has the most critical violations? (Cut)
3. What Cuisine Type has the most critical violations? (Cut)
4. What Borough has the most rat sightings? (Cut)
5. What Cuisine Type has the most rat sightings? (Cut)
6. Correlation between Number of Violations and Rat Sightings (Viz)
   1. By Borough
7. Violation per Grade Bar Graph (Viz)
8. Rat Sightings per Grade Bar Graph (Viz)

More Cuts

1. Food Type Cleanliness by Borough

2. Cleanest Borough

3. Rats v. Cleanliness by Zip

4. User Reviews v. Cleanliness

5. Distance From Subway Cleanliness

6. Distance From Subway User Reviews

Algorithm

1. Rat Sightings in Neighborhood
   1. Bin based off of ranges
2. Number of Critical Violations
   1. Bin based off of ranges
3. Grade
   1. Look at number of ‘Z’ and ‘P’ values to determine if we drop them
   2. Not all have grades, so we will need to only divide by the number of data points used (2 or 3)
4. Change in letter grade?

Bin each of the components and correspond them to a number (1-5) 🡪 determine bins by range of all data

Add all of the grades and divide by 15