Final Project Proposal

Team Members: Humaira Ahmed, Juna Kawai-Yue, Katherine Lin

Title: Restaurant Health Code Violations in New York City

Abstract: Our team is interested in investigating patterns in restaurant health code violations across New York City. We want to look at a few key aspects of these violations in tandem with reviews of these restaurants to see whether or not health code violations impact customer experiences. Using analysis of NYC restaurant health code violation and government data and natural language processing of restaurant review data, we will be able to determine the relationships between restaurant violations, customer experience, and geographical region over time.

- *Regional differences:* How do health code violations differ from borough to borough? Which are the most common in each borough, and which borough has the most severe health code violations? What zip code has the least or most violations?
- *Cuisine:* Which cuisines are most common in which neighborhoods? Which cuisine tends to receive the most health code violations?
- *Restaurant reviews and ratings:* How did restaurant reviews change for these restaurants from before and after their health code violations were issued? What topics were most prominent when it came to restaurant reviews, more specifically bad ones.
- *Health code violations vs customer experience:* Does receiving health code violations impact the customer experience with the restaurant?
- *Time differences:* How do health code violations change over time? What is issued most over different years? How does it change regionally over time?
- *Socioeconomic factors:* How do health code violations correlate with socioeconomic factors, like income, of different regional areas (like zip codes)?

Techniques:
More broadly we plan on doing: general analysis, spatial mapping, and natural language processing techniques

- ggplot2, ggmap, tidytext, web scraping, Lemmatization, TF-IDF

Data description:

- NYC restaurant health code violation dataset: https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j/about_data
    - This dataset is a compilation by NYC's Department of Health & Mental Hygiene of NYC restaurants and any health code violations that they may have received. This was made public in 2015 and continues to update daily.
    - The restaurants that are included in this dataset are all active.

    - The dataset contains information such as the borough of the restaurant, zip code, cuisine description, inspection date, violation code/description, a critical flag, score, and restaurant grade.

- Income by area/neighborhood dataset:
    - https://simplemaps.com/data/us-zips
        - https://simplemaps.com/city/new-york/zips/income-household-median?color=d6a4a4
    - This dataset is a compilation of various social and economic variables per zip code as designated by the US Postal Service (USPS) system. From here, we can specifically pull out median household income by zip code within NYC.
    - However, this data is only kept up to date to the 2023 census (5-year estimate) for certain variables like median income. This would limit out analysis with this dataset to just 2023, but could still yield interesting results and visuals.

- Google Reviews Data: https://developers.google.com/maps/documentation/places/web-service/overview
    - Through research on APIs to gather restaurant reviews and ratings, it seems we can use Google Places APIs to pull this data.
    - This API is not entirely free. It comes with a 90 day free trial or $300 credit free trial, but this is possible to work within due to the deadline of the project and the fact that we have 3 team members that could help pull data. We will most likely narrow down the restaurants we will pull reviews from due to these restrictions as well. Another limitation of this data set is that we may only be able to pull a certain number of reviews, as other APIs like Yelp API only allow you to pull 3 reviews per location.
    - However, if we are unable to make this work, we may look into alternatives that do not incorporate API use/web scraping, such as looking at other data that may correlate with restaurant experience or health code violations.

Visualizations:

- *Interactive maps:* We can do mapping to visualize the number of violations and cuisines per region of New York as of this year (e.g. density of violations per borough)
    - We can also do this for different years to visualize how it changes as the years change – this could be an interactive element where the map would change for each year selected
    - Within these maps, we could have points where viewers can click on a restaurant, borough, or zip code, and see the number of violations it has

- *Static/interactive bar charts:* We can use these types of visualizations for a variety of different purposes.
    - Ranking different regions (by zip code or borough) by number of health violations
    - Top 10 most common words in reviews over time
    - Most common violations over time

- *Static/interactive line plots:* We could use a line plot to track:

    - Restaurant rankings over time (by year) to see if there was any significant change as restaurants received health code violations. To prevent overcrowding, we could visualize a handful of restaurants (~5) that fall on the higher end of the number of health code violations.
    - Relationships between median income per zip code and number of health code violations

- *Static word clouds:* We can do an introductory graph on some key words for different types of Google Reviews (negative and positive) to identify what is noticed most about restaurants when they have a poor experience. We will select a handful of these to highlight, which will most likely be those who have the most severe and highest number of health violations.