Jongwook Choe:
Professor Tao He
Math 448 Tuesday-Thursday
Due 3/12/2024

MATH 448 Project Report: Understanding Public Data and Predict Probability of Success of Newly Opening Business.

Introduction:

Haven't you worried about stepping on poop while walking down the road? Well, I did. “San Francisco's public poo problem hits record level. They call it the Bay Area brownout. San Francisco authorities have recorded more than 125,000 cases of human feces found on public streets since 2020.” (Washington Times) According to the article, just the cases reported in 2020 are more than those reported from 2011 to 2019. I believe it started after the covid hit. The city is starting to get better but we want clean shoes. I thought if we opened a shoe cleaning business in a neighborhood where there is a higher chance people shoes it could be a big hit! Throughout the analysis, the goal is to predict where will be the optimized opening place for the new business. The data will be trained by merging public data and Yelp API. I will create a new equation to estimate the success rate based on review counts and counts of surrounding human waste.

Data Description:

This dataset includes cases generally associated with a place or thing (for example parks, streets, or buildings) and created July 1, 2008 or later. Cases generally logged by a user regarding their own needs (for example, property or business tax questions, parking permit requests) are not included.

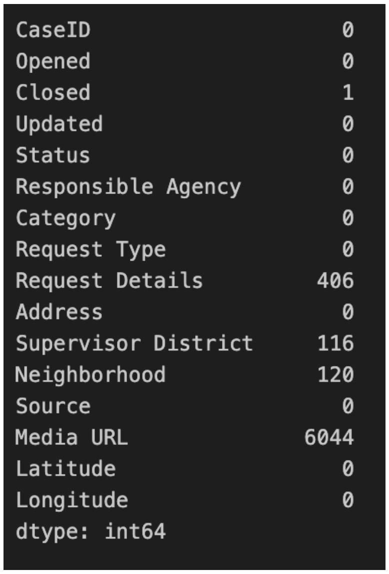
The first data set consists of 8,183 rows and 15 columns. The following table briefly describes the variables/columns of the data set.

Column Name	Description	Type
CaseID	The unique ID of the service request was created.	Number
Opened	The date and time when the service request was made.	Date & Time
Closed	The date and time when the service request was closed.	Date & Time
Updated	The date and time when the service request was last modified.	Date & Time
Status	A single-word indicator of the	Plain Text

	current state of the service request. (Note: GeoReprot V2 only permits ‘open’ and ‘closed’)	
Responsible Agency	The agency responsible for fulfilling or otherwise addressing the service request.	Plain Text
Category	The human-readable name of the service request type.	Plain Text
Request Type	The human-readable name of the service request subtype.	Plain Text
Request Details	The human-readable name of the service request details.	Plain Text
Address	Human readable address or description of location.	Plain Text
Supervisors District	San Francisco Supervisor Distrity as defined in ‘Supervisor Districts as of 2022’	Number
Neighborhood	San Francisco Neighborhood as defined in ‘SF Find Neighborhoods’	Plain Text
Point	Combination of Latitude and Longitude for Sorcrata native maps.	Location
Source	Mechanism or path by which the service request was created typically ‘Phone’, ‘Text/SMS’, ‘Website’, ‘Moblie App’, ‘Twitter’, etc but terms may vary by system.	Plain Text
Media URL	A URL to media associated with the request, e.g. an image.	Website URL

This dataset provides comprehensive information about various service requests in San Francisco, including their status, categories, locations, and timestamps. Analyzing this data can provide insights into service request patterns, response times, and resource allocation for different agencies and neighborhoods in the city.

To effectively manage and analyze the service request dataset from San Francisco, it's imperative to perform data cleaning and reduce the number of columns to enhance efficiency and accuracy in subsequent analyses.



```
CaseID                    0
Opened                    0
Closed                    1
Updated                   0
Status                    0
Responsible Agency        0
Category                  0
Request Type              0
Request Details         406
Address                   0
Supervisor District     116
Neighborhood            120
Source                    0
Media URL              6044
Latitude                  0
Longitude                 0
dtype: int64
```

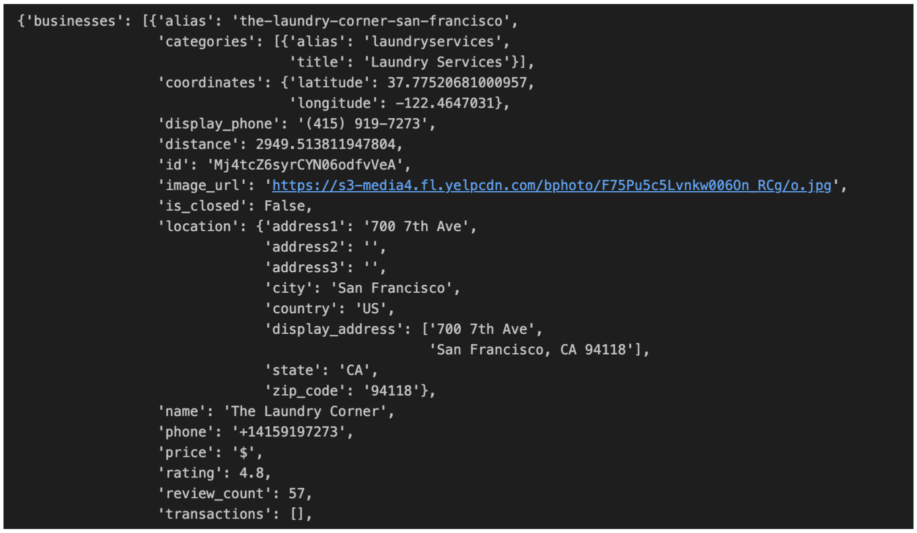
The above chart shows the number of missing values in each column. Request Details, Supervisor District, Neighborhood, and Media URL are the ones.

Furthermore, reducing the number of columns can streamline the analysis process and improve computational efficiency. We can simplify the dataset structure without sacrificing critical information by selecting relevant columns that contribute significantly to the analysis objectives while discarding redundant or unnecessary ones. This targeted approach facilitates faster processing and enhances the interpretability of the analysis results.

Therefore, I deleted the few columns and separated points into Latitude and Longitude.

The next data set is from the Yelp API returns up to 50 businesses with some basic information based on the provided search criteria in this case it will be laundry.

The following is an example of JSON output:



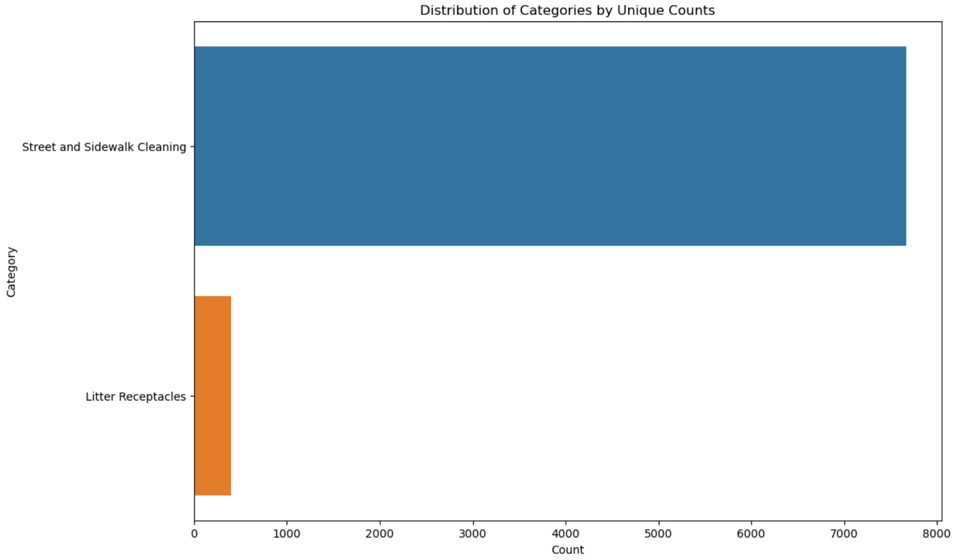
```
{'businesses': [{'alias': 'the-laundry-corner-san-francisco',
                 'categories': [{'alias': 'laundryservices',
                                 'title': 'Laundry Services'}],
                 'coordinates': {'latitude': 37.77520681000957,
                                 'longitude': -122.4647031},
                 'display_phone': '(415) 919-7273',
                 'distance': 2949.513811947804,
                 'id': 'Mj4tcZ6syrCYN06odfvVeA',
                 'image_url': 'https://s3-media4.fl.yelpcdn.com/bphoto/F75Pu5c5Lvnkw006On_RCg/o.jpg',
                 'is_closed': False,
                 'location': {'address1': '700 7th Ave',
                              'address2': '',
                              'address3': '',
                              'city': 'San Francisco',
                              'country': 'US',
                              'display_address': ['700 7th Ave',
                                                  'San Francisco, CA 94118'],
                              'state': 'CA',
                              'zip_code': '94118'},
                 'name': 'The Laundry Corner',
                 'phone': '+14159197273',
                 'price': '$',
                 'rating': 4.8,
                 'review_count': 57,
                 'transactions': [],
```

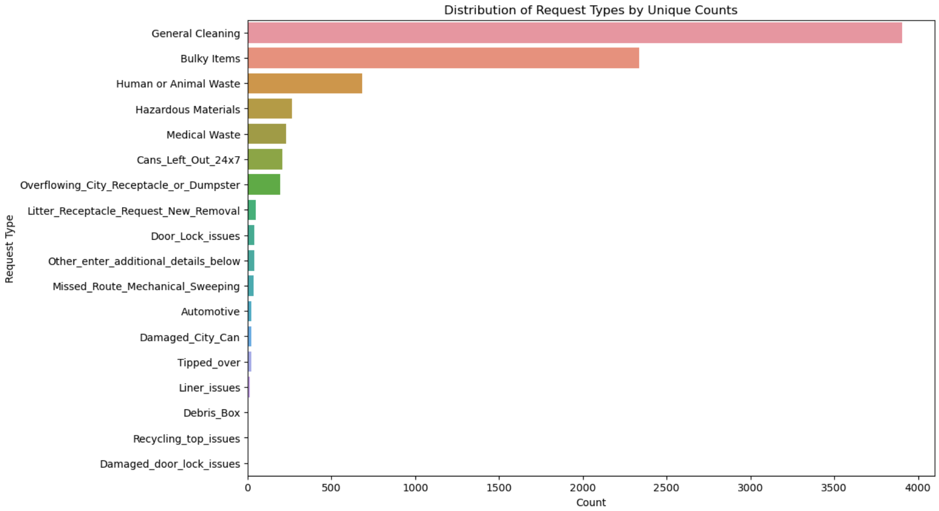
The decision to utilize the datasets separately was made after initially considering their merger.

The decision to utilize the datasets separately stemmed from the realization that the public dataset contains a substantial volume of data, while the Yelp API only provides 50 results per call. Consequently, the Yelp API will serve as a focal point for analysis, leveraging insights from the extensive public dataset to enrich and augment its training process. This approach maximizes the utility of both datasets, harnessing the depth of the public dataset while integrating the real-time and specific insights provided by the Yelp API.

Exploratory Data Analysis:

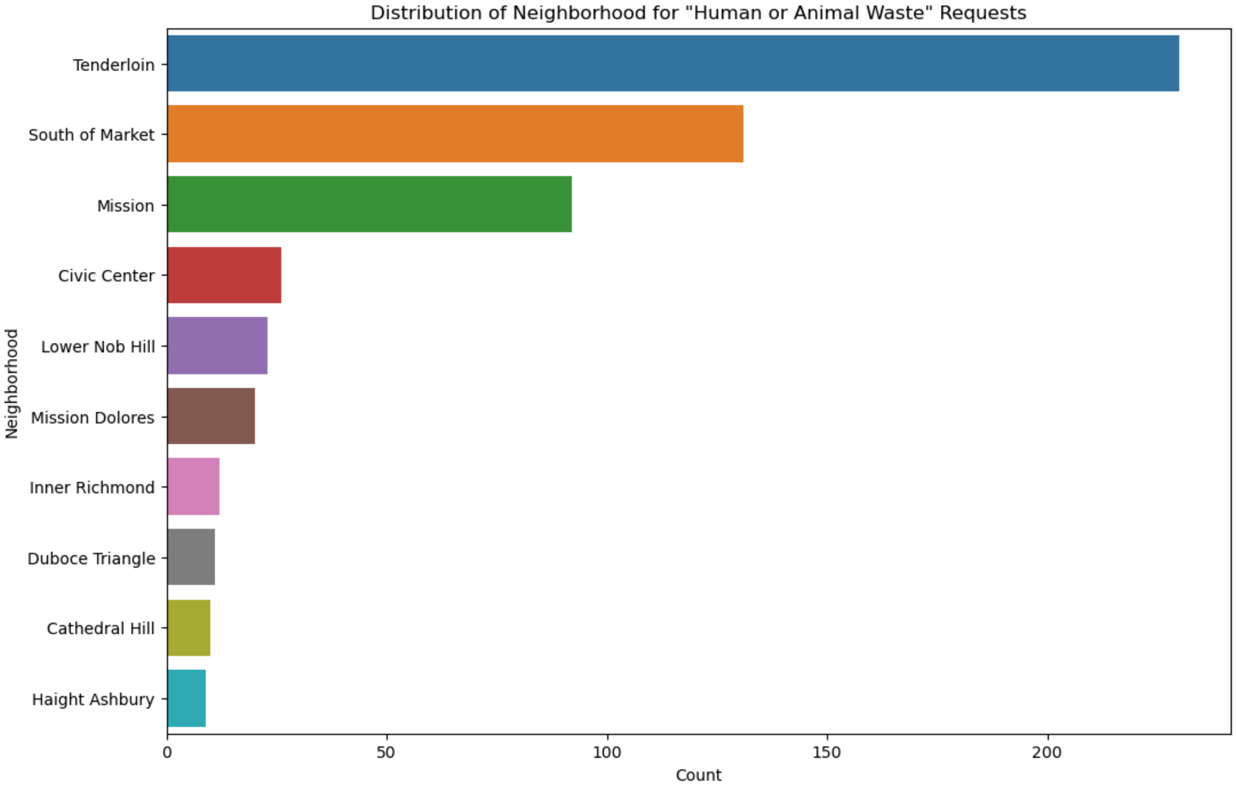
Let's see the distribution of the “Category” and “Request Type” columns by their unique counts. The first one shows the distribution of categories by their unique counts and the other shows the distribution of request types by their unique counts.





The dataset predominantly comprises entries related to street and sidewalk cleaning, with the top five request types being general cleaning, bulky items, human or animal waste, hazardous materials, and abandoned trash.

Now to visualize the distribution of the "Neighborhood" column when the "Request Type" column is "Human or Animal Waste", you can filter the DataFrame first and then plot the distribution of neighborhoods.



It can be observed that Tenderloin is the most frequently reported neighborhood for cases involving Human or Animal Waste, with SOMA and Mission closely following suit in the frequency of reports.

Furthermore, investigation is needed to make a prediction model of “Count of Human and Animal Waste Case” while using Yelp data as the reference. The target variable is the count of reported human and animal waste cases at each point. In this case, each data point in your dataset would represent a specific location (identified by latitude and longitude), and the target variable would be the number of reported cases of human and animal waste at that location.