Group N (NYC Rats): Process Book

Group members: Brendan Mapes, Pratishta Yerakala, Prajwal Seth

Part 1: Project Proposal and Brainstorming

For context, here we have included our original brainstorming and project proposal. This was the first iteration of our project, but we incorporated many of these ideas into the final version.

Proposal

Title: NYC Rats

Abstract: We're interested in rat data. Since NYC is known for being so populated and dense, it is inevitable that there will be vermin sightings and nuisances. We would like to take a look at the
geographic nature of these sightings with relation to other maps with geographic features (i.e. sewage systems, neighborhoods with many restaurants, etc.).

- What is the distribution of rat sightings across New York City and its boroughs?
- How do rat sightings compare when overlaid with sanitation pick up, neighborhoods with densely packed restaurants, neighborhoods with high or low income, etc.
- Current status of rat inspections, where they fail or pass or are in process the most
- Compare data from 311 rat sighting calls and rodent inspections to see if there is a correlation between the complaint and the response from the city
- Rat sightings and proximity to parks
- Rat sightings in relation to dog ownership density (because dogs have to be licensed, and this data is available)

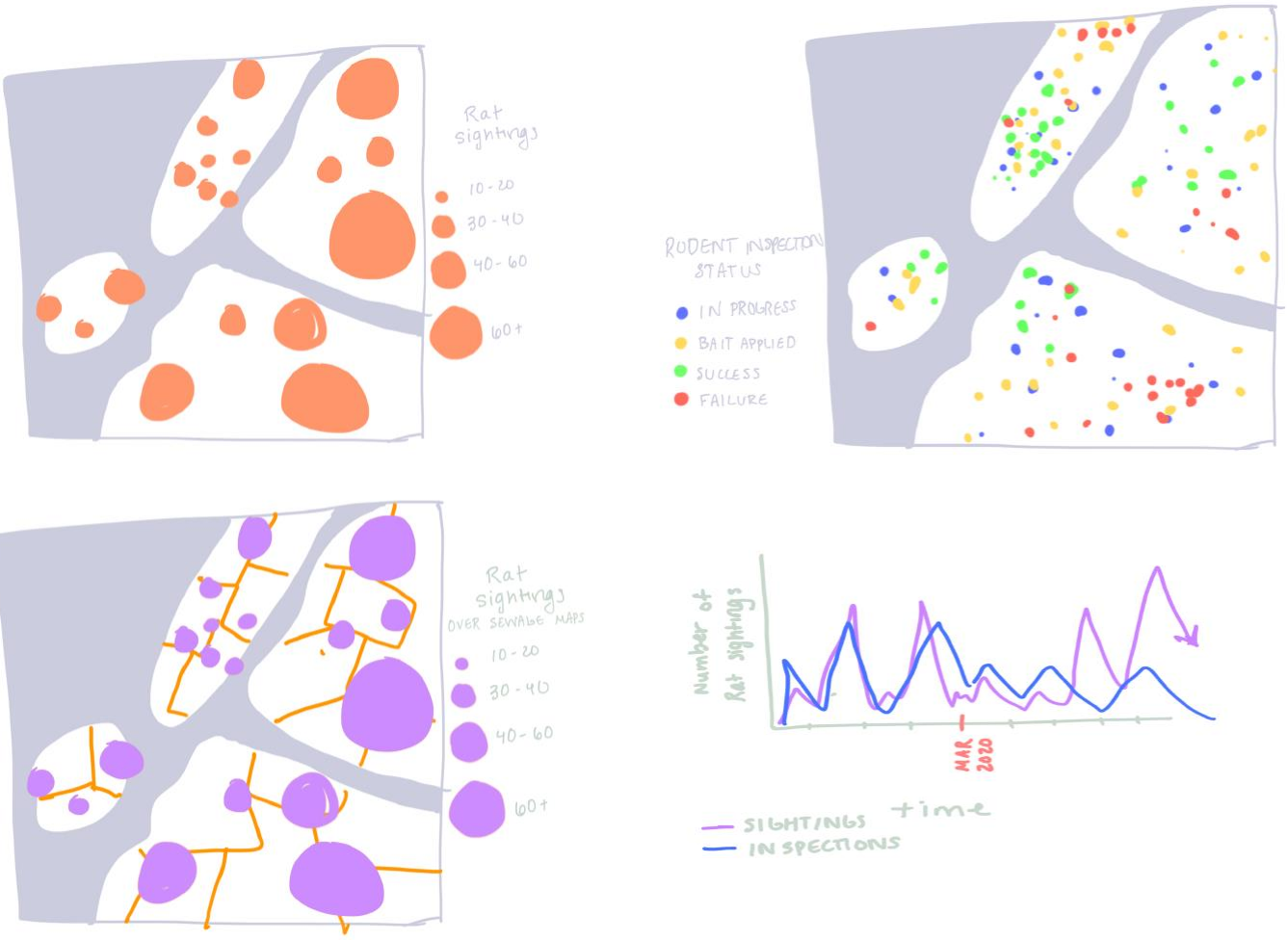
Data:

- Rodent Inspection - NYC OpenData
- OPEN SEWER ATLAS NYC (tumblr.com)
- DOB Stalled Construction Sites | NYC Open Data (cityofnewyork.us)
- NYC Dog Licensing Dataset | NYC Open Data (cityofnewyork.us)
- NYC OpenData - 311 calls filtered by rat sightings
- Housing Maintenance Code Violations | NYC Open Data (cityofnewyork.us)
- Potential others to include:
    - City parks location / shapefile data
    - Restaurant data
    - Sanitation data
    - [some demographic / census data]

Visualization brainstorming:

- Primary visualization style: **maps**
    - Rat sightings data
        - Centroid data, size of each centroid represents frequency of rat sightings
        - Heatmap with rat sightings data
        - Heatmap / choropleth with zip code data to see if there is a significantly rat dense neighborhood

Could look something like this:



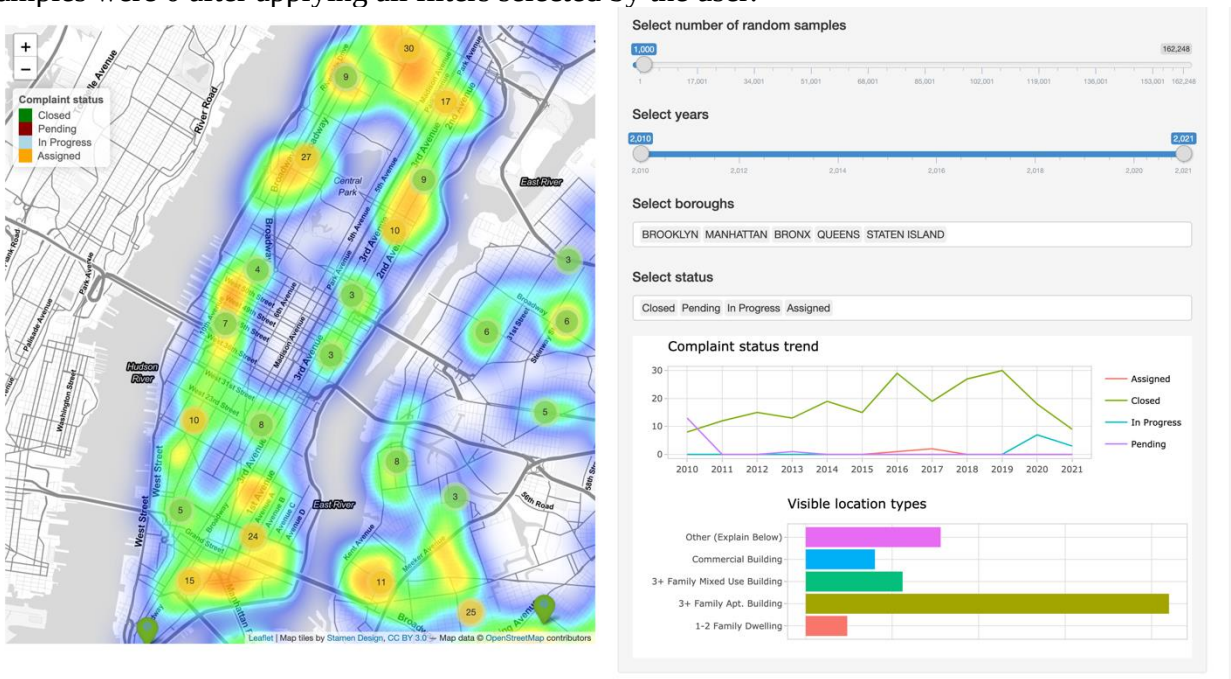
Part 2: Description of Final Output

Prajwal Seth:

I initially started off with a simple map which plotted 10 points from the NYC rats dataset. I also included a “New points” button which would refresh the map with new points when clicked. The point of starting in this manner was to take baby steps into using Shiny (as I was new to it). I referred to the leaflet + Shiny documentation at this page (https://rstudio.github.io/leaflet/shiny.html) in order to create this map. I also went through the source code for the SuperZip explorer visualization (https://shiny.rstudio.com/gallery/superzip-example.html) which is listed as a sample project on Shiny’s website. After going through the code for the SuperZip explorer app, I began adding more functionality such as tooltips (for when the markers are clicked), clusters (for when the map is zoomed out), a heat layer (using addHeatMap) and filters for the year, borough, and case status.

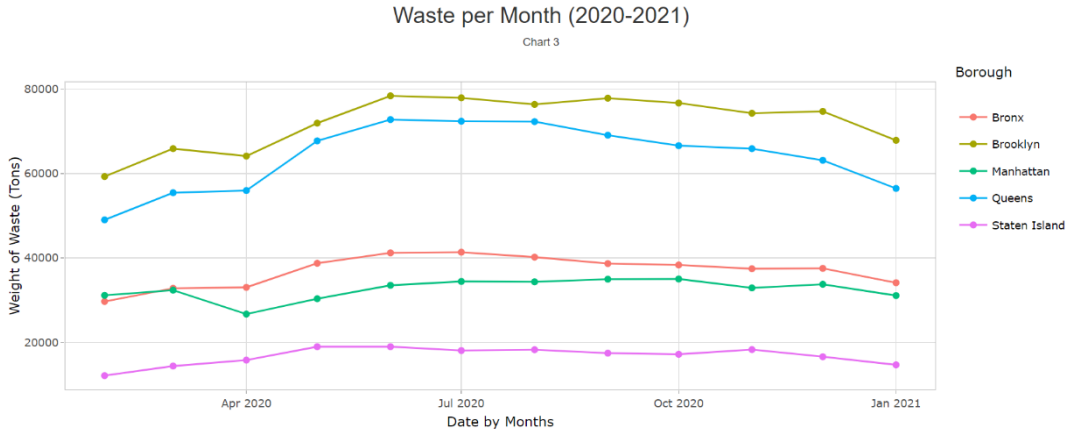
After following the steps listed before, I was able to create a map which looked like this- it had filters for years, boroughs, and case status, had a multi-colored tooltip which was colored according to the case status, had a layer of heat added to it, and a simple plotly graph showing the rat sightings from 2010 onwards in the visible area. I used random sampling to generate the data for the markers (the value for the random samples was set between 100 and 10,000 while I was debugging my app). I later on added a slider so that the user could select the number of random samples to draw from the original NYC 311 dataset which had 162K rows. I also added another plotly visualization showing the visible location types (such as school, commercial, 3bhk home, etc.). Finally, I wrote my code in such a way that it would not crash if the number of

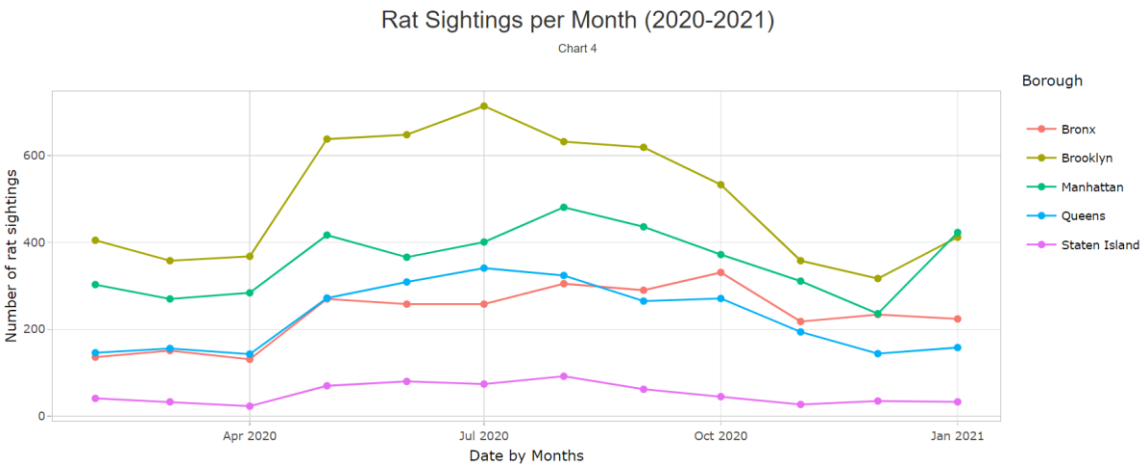
samples were 0 after applying all filters selected by the user.

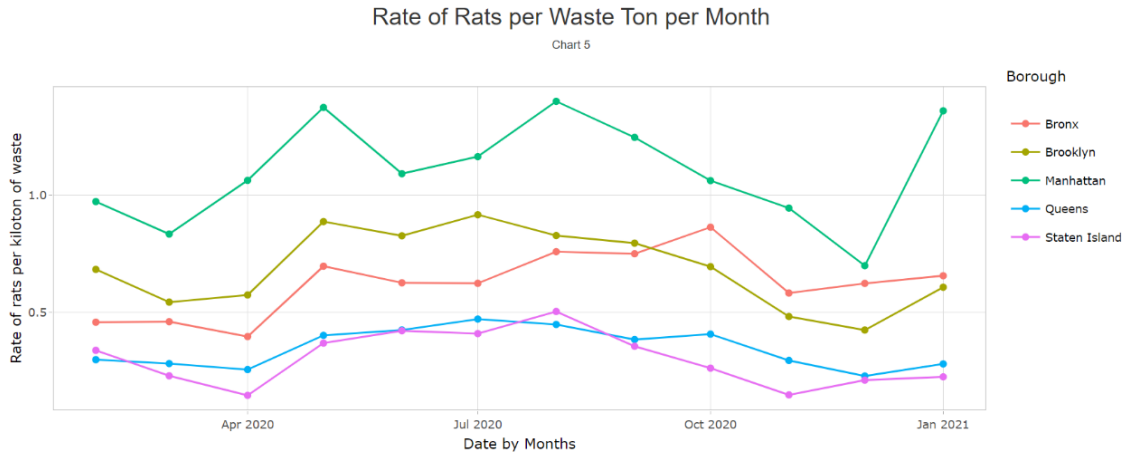


This is the final result for the interactive map of rodent complaints. In this version, I have added trend lines broken down by complaint status (as opposed to simply showing the total complaints like in the previous version). I also added the visualization for visible location types, the slider for the user to select the number of random samples to draw from and changed the legend on the map to color each complaint status differently.

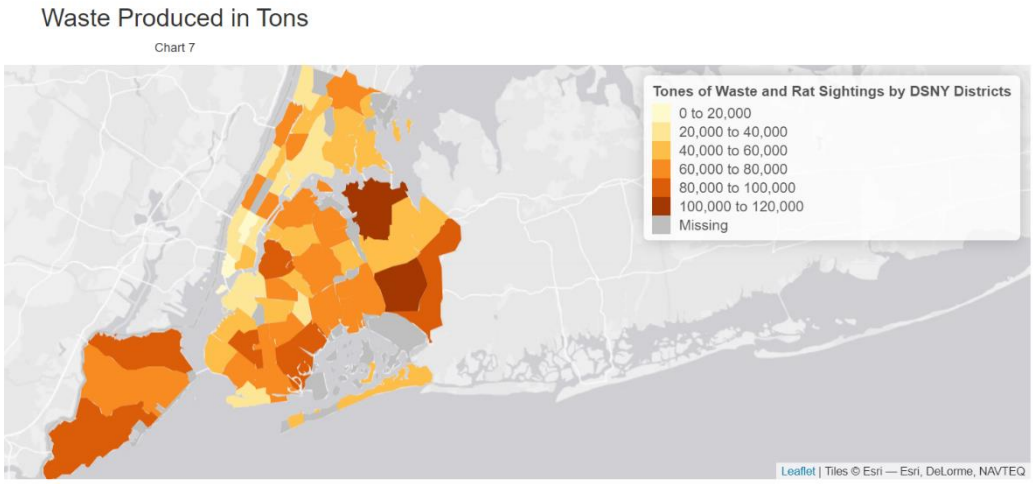
Pratishta Yerakala:





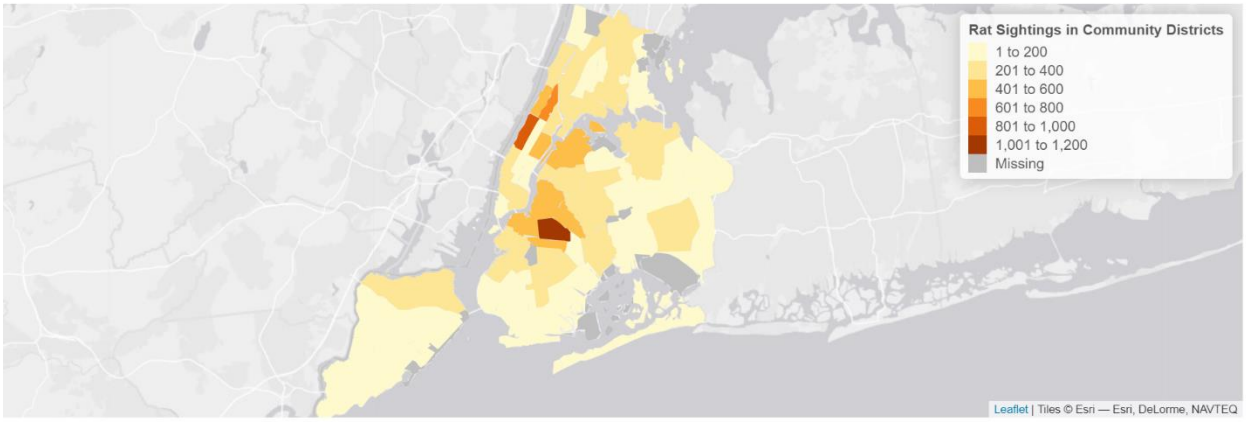


The most difficult aspect of data wrangling for the time series data was in the DSNY Monthly Tonnage Data and converting the months format of %Y / %m to a DateTime data type. Since that data type requires a date value, I input dummy values of 01. Then I converted the Rat Sightings Data dates from the created column and converting it into %Y-%m-%d format. Then after tallying (rat data) and summing (DSNY tonnage data) I was able to plot the time series data grouping by borough.



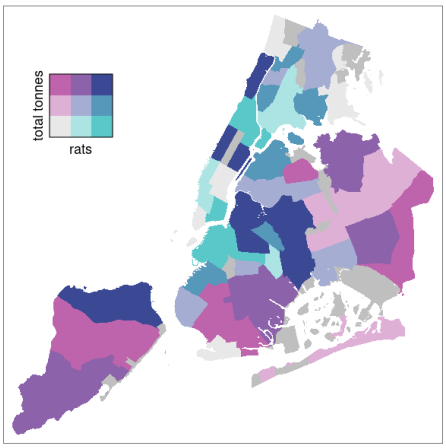
Number of Rat Sightings per Month

Chart 6



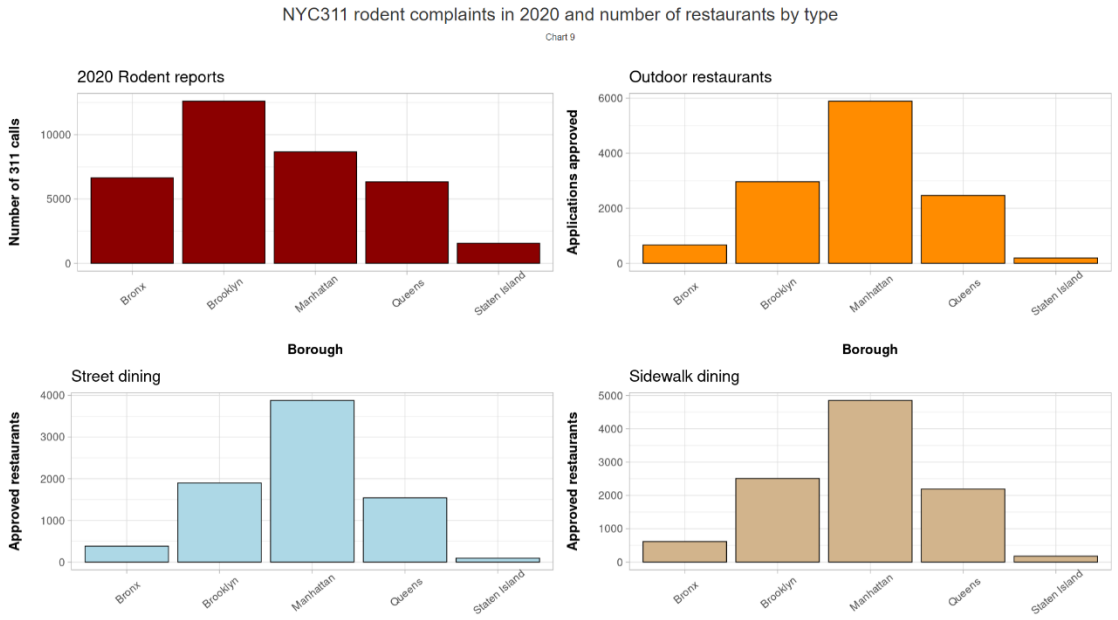
Rat Sightings and Waste Produced By Community District

Chart 8

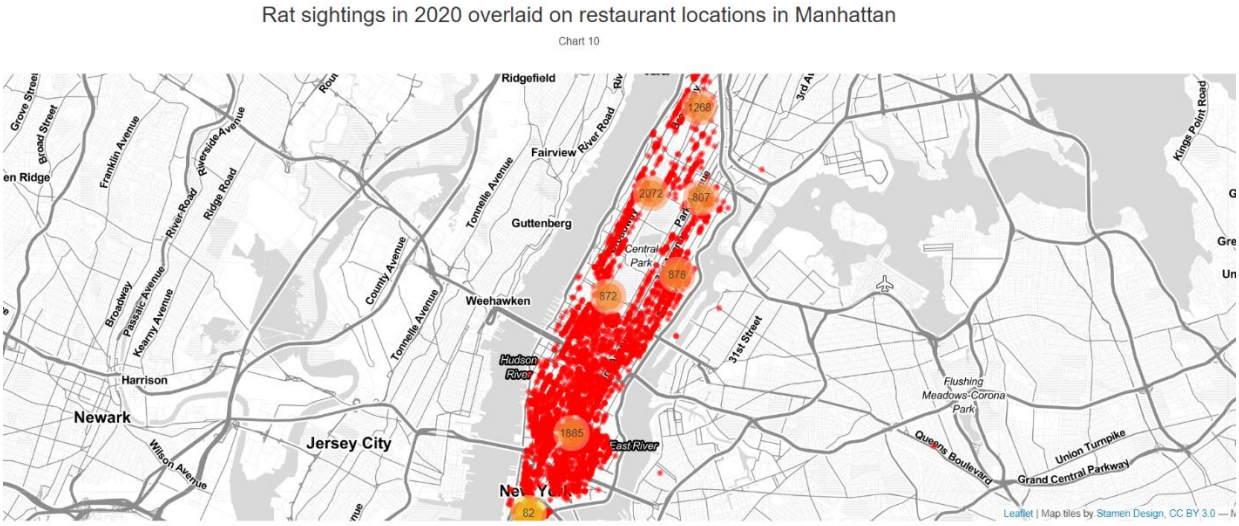


The biggest issue was to decide by which geographic feature that we could map the rat sighting *and* the DSNY data. I choose community districts because that was the only geographic feature that was readily available in both datasets along with a convenient shape file available. However, the biggest challenge here was converting the community district codes for each dataset along with the shape file data. For the rat data, the community districts were split into two columns. And the shape file had them in the form of numbers. I created two functions (the two convertToCommunityDistrict functions I forget the exact name). I used those functions when cleaning and setting up the data in the mutate function to create a new clean column with community districts that match the format of all the datasets. This required quite some string manipulation but thankfully not as granular as needing to use regular expressions.

Brendan Mapes:

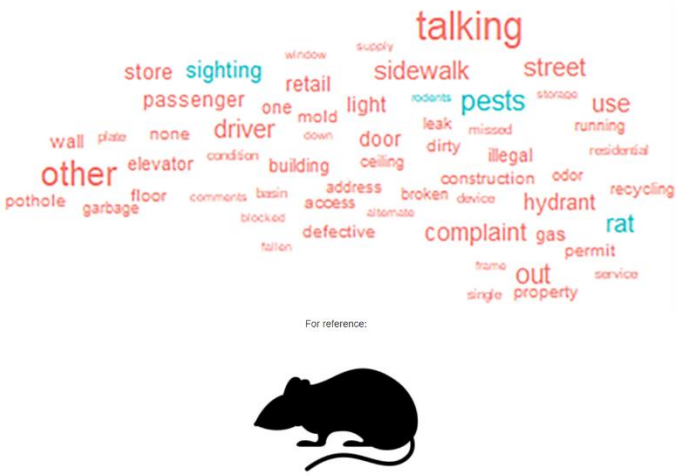


Data for these visualizations came from NYC OpenData's "311 Service Calls from 2010 to Present" and "Open Restaurant Applications" datasets. The 311 calls data was filtered to the year 2020, and for the complaint type rodent. These plots were created simply using ggplot2's geom_bar. From simple bar plots like these, viewers can quickly compare the boroughs of New York in rat sightings and outdoor dining, which is the goal of these first four plots. Manhattan shows to have by far the most restaurants with outdoor dining, while Brooklyn has by far the most rat sightings. Because this section of the of webpage is aimed at exploring outdoor restaurants as they relate to rat sightings during the COVID-19 pandemic, the next visualization focuses specifically on Manhattan, where, if restaurants are affecting the rat issue anywhere in New York, would be the place to look.



Here I have used again the 311 reports and open restaurant applications data from NYC Open Data to create interactive map visualization of rats as and open restaurants. To better connect the restaurant data and the rat sightings data, visualization that included geospatial information was necessary. The most interesting piece of information in exploring restaurants' practices as they affect rat sightings was the geographic information that this visualization includes.

This map was created using the leaflet package. The popup argument allowed for additional information on the restaurant to be included upon the clicking of that restaurant's dot. ClusterOptions allowed for the rats to be clustered when zoomed out, and disableClusteringAtZoom allowed for those clusters to be undone upon zooming into the map. These clusters of rats may be of interest to those looking to identify the most densely populated regions of rats, but not needing to know precise locations of the rat sightings. The icon for each rat sighting location was made to be a small rat image, using the makeIcon function of the leaflet package. Using these icons not only makes the map more visually appealing but also makes it easier to understand which dataset each point is representing. Rather than having two different colored dots or generic symbols on the same map, one for restaurants and one for rats, viewers know right away what data the rat icons represent.



Word clouds were created to fulfill the requirement of text analysis on this assignment, using the descriptor variable of the 311 service calls data set, again filtered to 2020, but not filtered by complaint type as was done for previous visualizations. This was not an ideal dataset to use for text analysis, with little variation and often standardized responses. Nevertheless, a word cloud could be useful as an alternative way to see which topics are most common in the dataset, based off text included in the descriptor column. Such a strategy is not as accurate as simply using counts of complaint types. It was necessary to filter out complaints related to noise, illegal parking, and other non-emergency police calls, which were the most common types of 311 reports ahead of rodent sightings. The descriptor variable was formatted as tidy text using the dplyr and tidytext packages, and then counted by word. An additional column was added to the word frequency data frame, for text color, to highlight rodent related words in the data frame. Stop words were removed and words used under 2000 times were filtered out.

A second word cloud was made, providing the same information to the viewer, but in the shape of a rat. This was accomplished with ggwordcloud as well, using the makeIcon function with the PNG file provided below the word cloud for reference. The only advantage to this version of the word cloud is its visual appeal, utilizing the shape of the rat to remind viewers the intention of these visualizations is to explore rat sightings data and rat related words.

This concludes the process book. As you can see, many, but not all the visualizations that we brainstormed in the early stages of our work made their way into the final version of our project. Our goal with these visualizations was to paint a more detailed picture of NYC's rat problem, to better understand questions like when, where, and why.