2. Data

2.1 Data Source

The data we will use in this project will be gathered from a variety of online sources using web scraping techniques such as Beautiful Soup.

The arrondissements information was found on this Wikipedia page. Its corresponding latitude and longitude were compiled manually using Nominatim. The top must-see Parisian landmarks were found on the this website as well as from Google Search, compiled with its corresponding latitude and longitude were found using Nominatim.

We will utilize a Paris arrondissements geojson file from Carto to map out the boundaries of each arrondissements and create visuals such as choropleth maps for population.

The venues data was found by using the Foursquare API.

2.2 Data Cleaning

The Arrondissement data that was scraped from Wikipedia site was fairly clean except for the "Arrondissement (R for Right Bank, L for Left Bank)" column, so we added another column to the data set named "Arrondissement" that included just the numbered arrondissement (i.e. 1st, 2nd, 3rd, 4th, etc.). We changed the longitude and latitude columns from an object to a float type. We also added another two columns for each arrondissement's average longitude and latitude.

The famous landmarks data was compiled manually into a list and appended with its associated longitude and latitude.

Utilizing Foursquare data, we downloaded the top 100 venues in each arrondissement within a 2000-meter distance from the arrondissement coordinate. The resulting table contained the arrondissement number, longitude, latitude, venue, venue longitude and latitude, and venue category. Since each arrondissement's area is varied between 0.3 square miles to 3.3 square miles, there may be duplicates in our search results. We therefore dropped duplicate venues that appeared in our search.