This week, I focused on getting the dataset ready for clustering. I explored its structure, checked for missing or inconsistent values, and worked on cleaning the data. The dataset contained several inconsistent or misspelled borough names, which could affect analysis and clustering accuracy. To fix this, I normalized the text by converting all entries to lowercase, trimming extra spaces, and then created a mapping dictionary to correct the variations.

For example, names like “manhhattan,” “quen” or “staten isalnd” were standardized to their proper boroughs (“Manhattan,” “Queens,” and “Staten Island”). After applying this mapping, I filtered the data to include only the five official boroughs — Manhattan, Brooklyn, Queens, Bronx, and Staten Island.

To strengthen my understanding, I watched a lecture on Clustering and Facility Location Problems at Microsoft Research and read sections from Sara Ahmadian’s PhD thesis. Explained how clustering can be reframed as a facility-location optimization problem finding a small number of centers (facilities) that efficiently serve demand points (clients).

Converted the dataset into a GeoDataFrame using GeoPandas, giving each incident an actual spatial geometry. I re-projected coordinates from latitude–longitude (EPSG:4326) to the New York State Plane coordinate system (EPSG:2263) for accurate distance computation in feet, and extracted x and y values for numerical clustering. Created a heat map visualization with Folium, giving an intuitive spatial picture of where incidents are concentrated across New York City.

Once the data was cleaned, I ran initial visualizations to understand patterns and relationships between features.