II. Data acquisition and cleaning

2.1 Data sources

Most details information including longitude, latitude, street name, city, district and zip code can be found in Kaggle datasets here. However, this dataset is very large in size. It contains a lot of duplicate information as well as many other detail information I do not need in this analysis.

2.2 Data cleaning

Before cleaning the data, there were 3,507,168 samples and 11 features in the data. This dataset contains all fourteen Massachusetts counties detail location information. I decided I would only use Middlesex county, Suffolk county and Norfolk county in this analysis because those are the three counties located at the center of the Massachusetts state as well as areas where the majority of people who work in Boston would live in.

The dataset contains many other columns of information I do not need, so I drop five columns that I am not going to use in this project. Moreover, with the same zip code, there are many different locations are recorded in the dataset. Hence, I removed all the rows with the same zip code and only left one row for each zip code. Then, I calculate the missing value rows and found out the missing value rows only occupied a very small portion of the dataset. Thus, I removed all the missing value rows from the dataset.

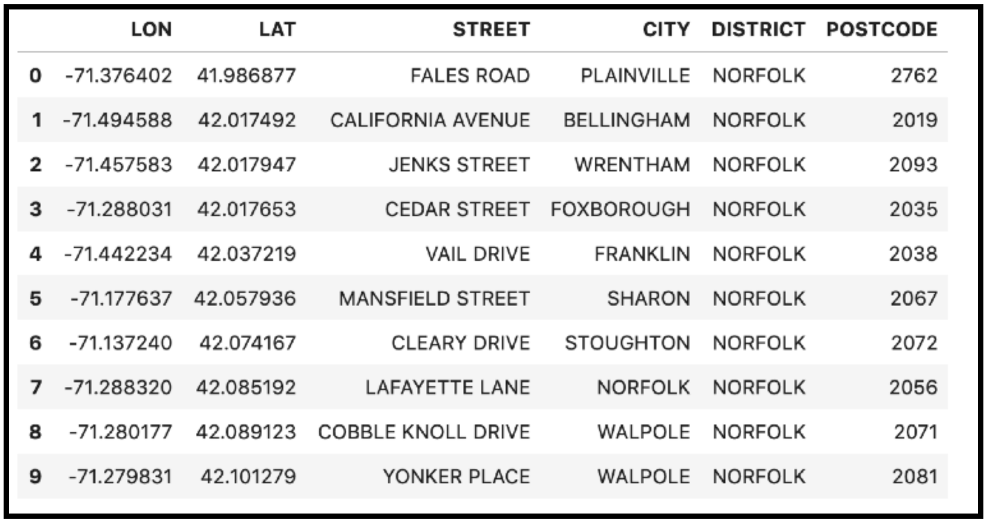
2.3 Feature selection

After data cleaning, there were 188 samples and 6 features in the data (Table 1).

LON - Longitude

LAT - Latitude

Table 1. First ten rows of the dataset



	LON	LAT	STREET	CITY	DISTRICT	POSTCODE
0	-71.376402	41.986877	FALES ROAD	PLAINVILLE	NORFOLK	2762
1	-71.494588	42.017492	CALIFORNIA AVENUE	BELLINGHAM	NORFOLK	2019
2	-71.457583	42.017947	JENKS STREET	WRENTHAM	NORFOLK	2093
3	-71.288031	42.017653	CEDAR STREET	FOXBOROUGH	NORFOLK	2035
4	-71.442234	42.037219	VAIL DRIVE	FRANKLIN	NORFOLK	2038
5	-71.177637	42.057936	MANSFIELD STREET	SHARON	NORFOLK	2067
6	-71.137240	42.074167	CLEARY DRIVE	STOUGHTON	NORFOLK	2072
7	-71.288320	42.085192	LAFAYETTE LANE	NORFOLK	NORFOLK	2056
8	-71.280177	42.089123	COBBLE KNOLL DRIVE	WALPOLE	NORFOLK	2071
9	-71.279831	42.101279	YONKER PLACE	WALPOLE	NORFOLK	2081

Upon examining the meaning of each feature, it was clear that some of the features were not needed in the report as well as map visualization. The features are dropped including “ID”, “Number”, “Unit”, “Hash” and “Region”. Features I kept including “LON"(longitude), “LAT"(latitude), “Street”, “City”, “District” and “Post code”. The reason I kept “City” and “Street” is that these information is needed when using Foursquare API to explore the neighborhood. Moreover, it would be easier for the audience to read.

Table 2. Simple features selection during data cleaning

Kept features	Dropped features	Reason for dropping features
LON, LAT, Street, City, District, Postcode	ID, number, unit, hash, region	Dropped features are not needed for this report and will not be used in map visualization