Data Acquisition and Cleaning

1. Data Acquisition

First, I had to find the data for all the localities in Bangalore. For that I had to search through google and I reached this site – click here to visit the site. The source is from a well reputed news channel in India so I thought to continue with it.

I scrapped the data from their site into a csv file for its further use. At first look itself I could understand that it had some inconsistencies like null data, similar data also sub location of the locality listed as a separate locality. So, these were the first things in my mind to sort.

Before starting data cleaning process I even added Latitude and Longitude columns in the dataframe with the values added into them using geopy package of Python.

	Office	Taluk	District	State	Pincode	Latitude	Longitude
0	A F Station Yelahanka	Bangalore North	Bangalore	KARNATAKA	560063	NaN	NaN
1	Agram	Bangalore South	Bangalore	KARNATAKA	560007	NaN	NaN
2	Air Force Hospital	Bangalore North	Bangalore	KARNATAKA	560007	12.964027	77.627500
3	Amruthahalli	Bangalore North	Bangalore	KARNATAKA	560092	13.066513	77.596624
4	Anandnagar Bangalore	Bangalore North	Bangalore	KARNATAKA	560024	13.033377	77.589523

Look of the data after 1st stage of Data Acquisition

2. Data Cleaning

I started with searching for duplicate data and null values. First, I removed all the null valued areas as they were mostly the inner neighbourhood of the localities already present in the dataset.

Secondly, I combined the localities which had their same latitude and longitude into one row.

Third, I removed the unnecessary columns i.e ‘Taluk’ and ‘State’ and renamed the column ‘Office’ as ‘Locality’ which is a better suited name for the column.

After this I visualized the data using folium package on the map and found may outliers i.e the locations which are far away from the Bangalore.

To get rid of those locations as well as outskirt areas of Bangalore, I removed all those locations which were more than 35kms apart from the center of Bangalore using ‘Haversine

distance’ method so that our analysis gets concentrated in urban areas of Bangalore which offers much better scope of business.

After this a total of 142 localities were left which are now ready to be used for further work.

At a later stage after using FourSquare API , I even found that some localities had no venue within specified radius as a result I had to remove them too as they couldn’t be clustered in any group.

3. FourSquare API

Now, after the data cleaning phase we were left with 142 localities. I used the FourSquare API and ran through all those localities to find the venues present inside the 500 metres radius from their location and to store and maintain a different dataframe for it. It had total of 1032 venues in it.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1027	Kannur	13.103799	77.601617	Udupi Grand, Kogilu Cross	13.103455	77.600627	Fast Food Restaurant
1028	Kannur	13.103799	77.601617	Yelahanka lake	13.103729	77.602205	Lake
1029	Kannur	13.103799	77.601617	Gobi Adda	13.103732	77.602401	Food Truck
1030	Madanayakanahalli	13.059247	77.461511	Khimaj Caterers	13.060353	77.460115	Restaurant
1031	Neriga	12.911766	77.776525	Shristhi Village	12.913261	77.773636	Resort

Results returned from using FourSquare API on the localities