3. METHODOLOGY

3.1. DATA EXTRACTION AND CLUSTERING

In this section, first collect the zonal information data from the mentioned website in the data section. Beautiful soup is used to extract data from HTML, which is useful for web scraping. First, inspect the webpage to identify the class attribute for the table of interest. Raise a get request to fetch the raw HTML content and parse the content. Read the table of interest with the identified class attribute and extract all table information.

After obtaining the zonal information, latitude and longitude location data for each zone is obtained using Geopy, Nominatim package which is commonly used for converting address to location information. The location data is visually rendered on a map using Folium.

Each neighborhood location is explored using foursquare API requests and the JSON response from the API request is processed to obtain only relevant information. The Obtained venue data is explored for total number of venues and unique categories.

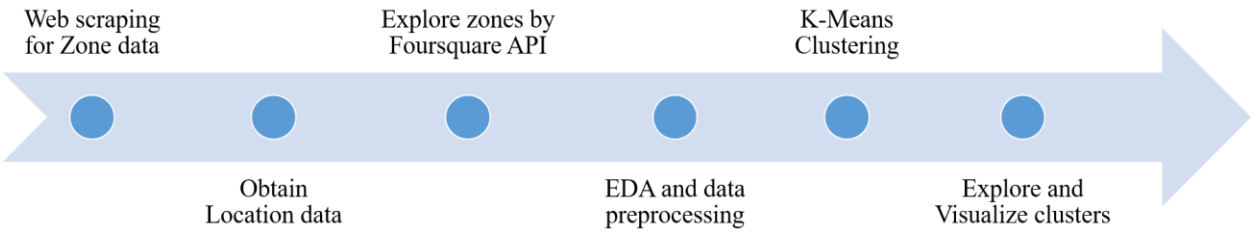

Figure 1 Data Extraction and clustering process

One hot coding is done to convert the unique categories to venues to numerical data and the average occurrence of a particular category of venue is obtained. Most commonly occurring venues in a given neighborhood are identified and a data frame consisting of the neighborhood and its corresponding 10 most commonly occurring venues is prepared.

K-Means clustering algorithms is used to group the venues based on the similarities of the occurrence of venues. In this project, a cluster value of five is chosen to group the neighborhoods into five different categories. The Clustered information is then

visually rendered using folium. Individual cluster groups are explored for their similarities and their descriptions. The complete data extraction and clustering process is shown in Figure 1.

3.2. IDENTIFYING SIMILAR NEIGHBORHOODS

The above extraction and clustering process is repeated for two main locations i.e. Chennai or hometown location and Bangalore or job location. Neighborhoods of both job and home location are explored first for similarities. Home location neighborhoods are filtered for regions of interest for comparison with job location neighborhoods. Neighborhoods in job location are filtered based on distance from office area which is chosen to be less than 10 Km.

The radial distance between two latitude and longitude location information is calculated using Geopy geodesic distance calculator module.

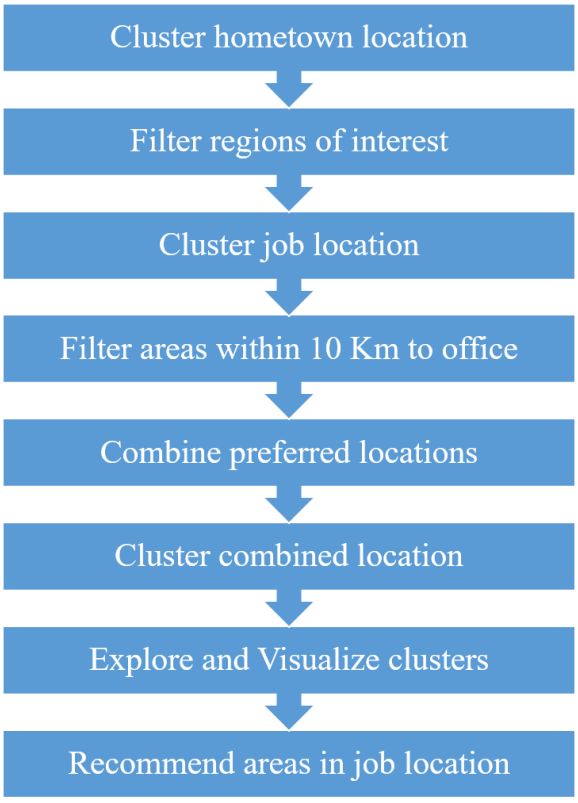

Figure 2 Identifying similar neighborhood

One important fact to observe is that the distance obtained is radial or perpendicular distance between two latitude and longitude location data and not road wise distance that is noticed commonly on any navigator like google maps.

The filtered location information from hometown and job location are combined in single data frame. The data is then visualized using folium renderer and explore of sufficient information. Already preprocessed data for previous clustering is used which includes the average occurrence of a particular venue category for each neighborhood.

The combined filtered data is then clustered using K-Means into five groups. The groups are then visualized and explored individually for similarities and descriptions.

The area in job location, which falls in the same group of the hometown, would become the first choice for selection. In case of multiple areas and groups identified, the most similar hometown location is chosen and the job area in the same group, which is closest to office, are recommended. The process for identifying similar neighborhood is shown in Figure 2.