Data:

To solve the problem, we will need the following data: • List of neighbourhoods in Hyderabad. This defines the scope of this project which is confined to the city of Hyderabad, the capital of Telangana. • Latitude and longitude coordinates of these neighbourhoods. This is required in order to plot the map and also to get the venue data. • Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighbourhoods

Sources of data and methods to extract them :

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India) contains a list of neighbourhoods in Hyderabad , with a total of 200 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful soup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods. After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the shopping category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.