**Classification of Delhi Metro stations**



**Introduction**

Delhi Metro is a rapid transit system serving Delhi and its satellite cities in the National Capital Region of India. As of now, there are a total of 229 metro stations including the Airport Express stations. The first section of the Delhi Metro opened on 25 December 2002 with the Red Line, and has since been expanded to around 347.66 km (216.03 miles) of route length as of 4 October 2019. The network has nine operational lines and is built and operated by the Delhi Metro Rail Corporation Limited (DMRC). The Delhi Metro Rail Corporation makes 2,700 trips per day carrying 1.5 million passengers, who on an average travel a distance of 17 kilometres each.

For this project, we will try to look at the places surrounding these metro stations and classify them according to the similarity of nearby venues. Almost every one use metro transit to migrate from one place to another for reasons which can be personal of professional. If there are more professional places like companies, offices surrounding a station then it will mostly be used by working professionals. Then there are some stations with many universities or colleges nearby and is used by Students mostly. Stations which have places like amusement parks, malls, monuments are used by people for recreation.

We can classify stations by primary usage analysing the data that contains the number of nearby venues according to their category. This can help plan further extension of the network and find places for new development.

**Data**

In this section we will describe our base data which we will analyse to reach the goal we want.

We will be requiring some Python libraries and modules like Pandas, Numpy, JSON, requests, geopy, bs4, sklearn, etc.to move further in the project.

The main source of stations data that will contain the list of all the metro stations under DMRC in Delhi. Is the following Wikipedia page.

<https://en.wikipedia.org/wiki/List_of_Delhi_Metro_stations>

We have to scrape the relevant table data from this url like Station name, Line.

To get the latitude and longitude values we will use geopy.geocoder module and will geocode each station.

**Assumption :-**

* There are some stations with more than one line that pass through it. So we have assumed that only the line that is written first on the above url will be the data of our choice to nullify the ambiguity that we may face while plotting on the graph

To get the nearby venues of all stations we will be using Foursquare API for which we need to create an account on the following url.

<https://developer.foursquare.com/>.

After this we will be required to pre-process data and create a data frame. After the pre-processing our data will look something like this.

