# DATA DESCRIPTION

### For any “data science project” data is of paramount importance. For this study, we needed data about neighborhoods in each of these metro cities. The data published by the government on postal codes for all India would serve us well for this study. We will specifically download the CSV provided under <https://data.gov.in/resources/all-india-pincode-directory-contact-details-along-latitude-and-longitude>.

For this problem, we will read the csv into a pandas Dataframe and curate it to remove the data related to all other cities, towns, and places which are not in our list. We shall then clean up the unnecessary columns in the CSV, which is not relevant or useful for our current problem. Post office names (*office name*) will be used as the neighborhood names in each of the regions. Duplicate Pincodes will be removed.

***Nominatim*** library will be used to find the longitude and latitude of each of the neighborhoods in both Mumbai ,Pune and Nagpur. This will form the dataset we will use for this study.

# The first few records of the dataset we now have after clean-up and curation appear as below.

# 

### 

### We now have the neighbourhood’s for both the metro cities

### The next step is to enhance the dataset with the required information. We would need the longitude and latitude values for the neighborhoods. We will use the *Nominatim* library from *geocoders.geopy* package to find the longitude and latitude for each of the neighborhoods and would eventually create a dataset having all the necessary columns for our analysis.

### 

### 

### 

### We now have the necessary information to visualize the neighborhoods for both the cities on a *folium* map.

### 

### 

### 