**Data Preparation**

Since we need to compare the neighbourhoods of the IT SEZ’s in and around Chennai, first we need to have the list of IT Park’s located in and around Chennai, for which there cannot be a better source than Wikipedia which looks like the below.

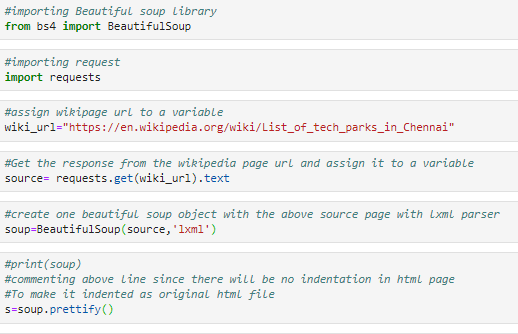


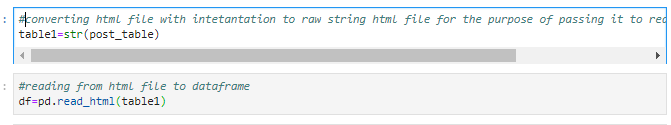
The Above table consists of SEZ name, Area (in Sq. ft),Address, Investment(in crores),Start Date Current Status and Notable Clients who rented out the SEZ space.

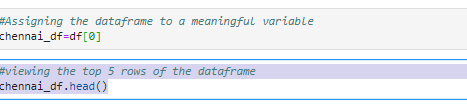
**2.1 Web Scraping:**

Since the above table is the part of the Wikipedia Page, we need to scrap the table alone from the page. For this, Requests library are used to get the XML and HTML source page from the Web and BeautifulSoup4 library object with lxml parser are used to scrap the table we need from the Source page.

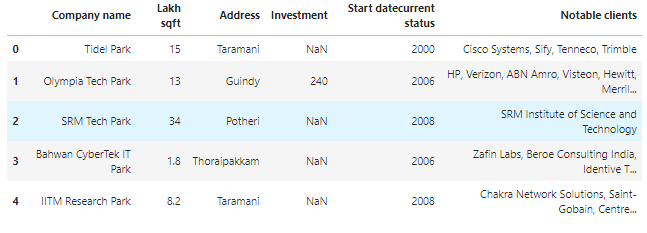
Below is the rough code snippet for it.







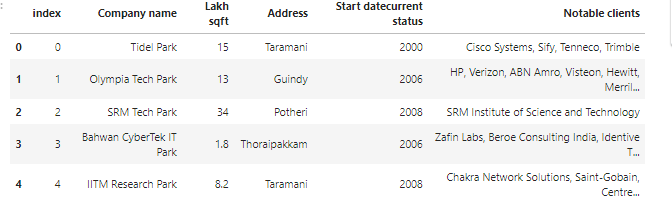
Below is the first version of the dataframe we get after the above code



**2.2 Feature Selection and Engineering:**

Firstly, I drop the Investment column, since many rows does not have it.Then I chose only the rows which are either already completed or just now completed. For this I have replaced the Under Construction and Planning rows with Null values and then dropping the corresponding rows.

After this I have resetted the row index, since I have dropped some of the rows, we get the below dataframe.

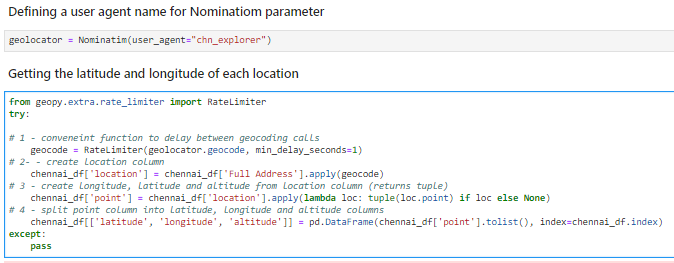


* 1. **Getting the co-ordinates of the IT Parks.**

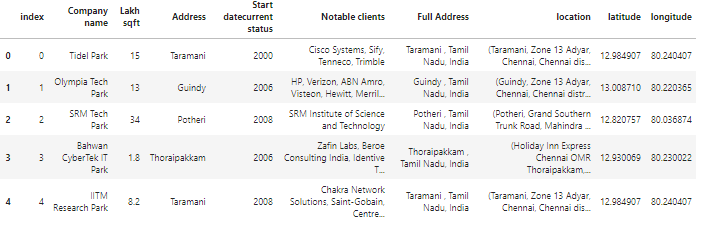
Then we need to get the co-ordinates of each of the IT SEZ in our dataframe. For this, we use Geopy library.

Then we add ‘Tamil Nadu, India’ to each values of the Address column to get the latitude and longitude of each places and store it in a separate column

Below is the rough code snippet.



After some manipulation, we get the below dataframe.



* 1. **Using FourSquare Location data**

Foursquare data is very comprehensive and it powers location data for Apple, Uber etc. For this business problem I have used, as a part of the assignment, the Foursquare API to retrieve information about the popular spots around these IT parks within the radius of 500m.

The popular spots returned depends on the highest foot traffic and thus it depends on the time when the call is made. So, we may get different popular venues depending upon different time of the day. The call returns a JSON file and we need to turn that into a data-frame which looks like the below.

