Data Section:

It took me some hours to find out the proper dataset for this project. First, I tried to do it for Companies for India. Since here in India datasets are not publicly available, I came up with an idea of doing it for US. Then I extracted the csv file format of US Open 500 Companies dataset. Along with dataset, we can explore the nearby venues of these Companies using Four Square API.

The **Open 500 Companies dataset** contains the following fields or columns.

- Company ID - The ID of the Company
- Company Name - The Name of the Company
- URL - URL of the company
- Year Founded - The year when the company is founded
- City - City of the company
- State - State of the company
- Full time Employees - Number of Full time Employees
- Company Type - Either Full Time or Part Time
- Company Category - Domain of the company
- Description - Single line description about the company
- Data Type - In which field the company is focusing the most

Some columns are ignored as they don't make sense when building a model for clustering.

The **Venues Dataset** – obtained through **Four Square API**

- Counts of Venues closed to the Neighborhood
- The frequency of each Venues Category such as Office, Bus Stop, Pizza Place, Coffee, Chinese Restaurant, Italian Restaurant, etc.

Using the above datasets, we can cluster the top revenue generated private sector companies in US and also we can cluster the neighborhoods of selected private sector companies. This is how the obtained data can be used to develop our project.