CLUSTERING THE NEIGHBORHOODS OF DATA SCIENCE COMPANIES IN US

Submitted By,

Prabhu Mayilsamy

(Data Scientist)

Introduction- Business Problem:

As I am interested in data science jobs, I would like to choose **clustering the neighborhoods of Data Science companies in US** as my Applied Capstone Project. The basic idea is using the dataset of all the companies in US and location dataset provided by **Four Square API**, we can **cluster the location of different domain companies in US**. We can also cluster the neighborhoods of top venues around all the data science companies in US using Four Square API.

The main problem faced by all the jobseekers is **relocation**. Since we don't have proper dataset of the facilities in the new location, we will always feel uncomfortable and sometimes jobseekers will be frustrated by searching for too many venues around the new place. We should love the environment where we are going to work, so that we will do the work with pleasure.

Our main objective of this project is to find out the locations of all the IT companies in the US and cluster them based on the domain of the company. Then based on the input provided by the jobseeker, we can filter the neighborhood top venues of the selected domain companies in US. In my project, I am going to use Data Science companies as my desired input and I will focus more on clustering the top venues around the selected data science companies. We can also create a **map** to visualize the locations of the different domain companies and the neighborhood top venues of the selected domain companies in US.

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 different domain companies in US and also we can cluster the neighborhoods of selected domain companies. This is how the obtained data can be used to develop our project.

Methodology/Exploratory Data Analysis:

In this section, I will explain about how I extracted the necessary features from the given dataset and I will further explain about how I figured out the relationship between the variables used in the modelling.

Feature Extraction:

In this section, I will explain about how I extracted the necessary features from the given dataset. The company dataset has been downloaded from this link: Link to dataset. For the simplicity I have downloaded the dataset that is in csv format and stored it in my project location. Then I figured out that all the columns in the dataset are not needed for the project purpose. Then I removed few columns.

The open companies 500 dataset contains the following columns/fields

company_name_id
url
city
country
full_time_employees
company_category
business_model
description
data_types
financial_info
company_name
year_founded
state
zip_code
company_type
revenue_source
social_impact
description_short
data_impacts
last_updated

After analyzing the dataset , I came to know that company_id, url, country, full time employees, revenue source, description, data types, data impacts, financial info, last updated fields are no more needed as they don't make any sense when building a model. So I dropped those columns and extracted the remaining columns. Now the dataset in pandas data frame contains only the following columns/fields.

company_name
city
company_type
company_category
social_impact
year_founded
state
zip_code
business_model
last_updated

Replacing the Mismatch Values:

After analyzing the values in the company category columns using **value_counts()** function in pandas package, I have figured out that there are few rows that has the same value but in different way those were represented.For example,

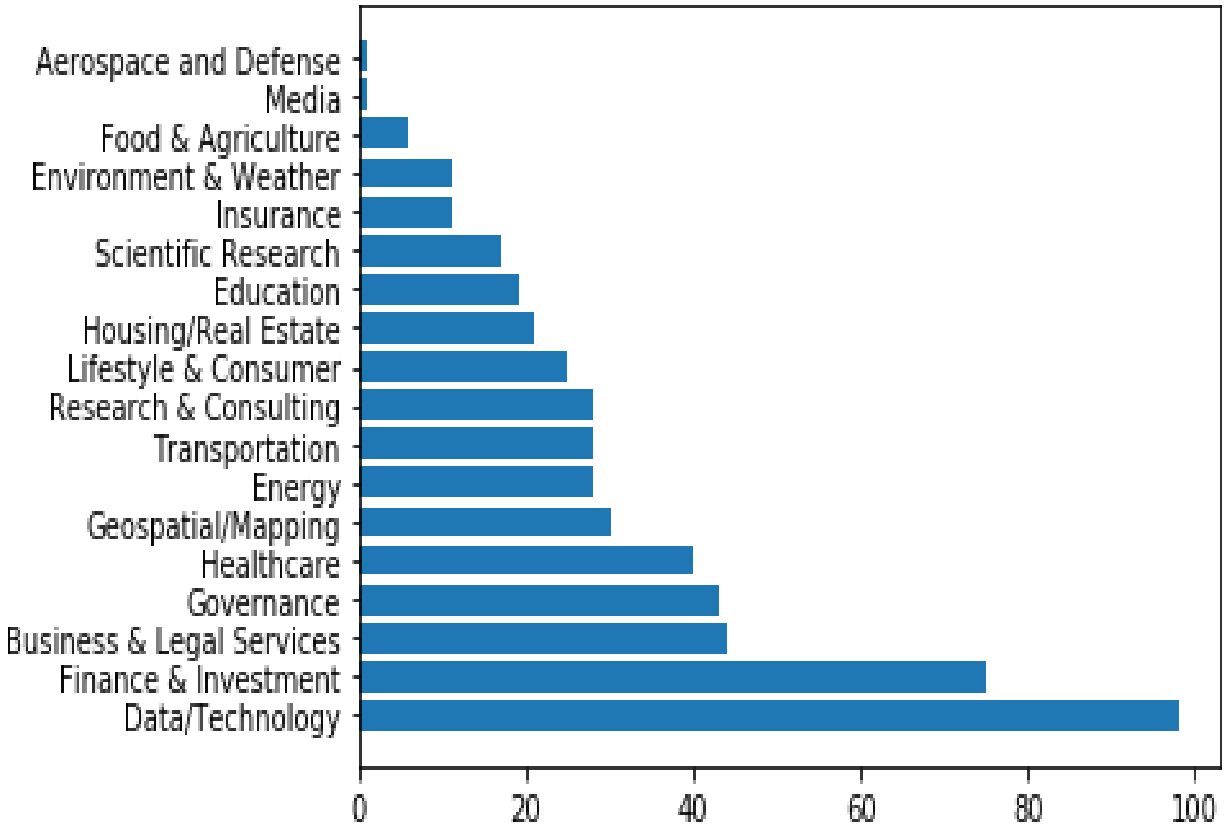
Data/Technology as Data/Technology,

Housing/Real Estate as Housing/Real Estate,

I have changed the values in these rows and made it same by renaming the values in the data frame in python pandas

Visualizing Each Type of Companies Count in US :

Using matplotlib and python pandas function, I have count the number of different category companies and their count using value_counts() function and converted that into data frame and visualized it .



From the above bar chart, we can clearly say, data/technology companies are the most featured companies in US. From this, we come to know that data science is the growing technology in US. As expected finance and investment took runner up place as both are almost closely related.

Handing Missing Data:

From the analysis, we come to know that company category plays a vital role in clustering the different domains and city and state columns are also very important. We have to handle the missing values in these columns. Since only 3 values are missing in company category field, we can remove these 3 rows as they wont affect the result much. In city and state columns also , only few rows are containing NaN values. So we can drop those rows also.

Categorical Values to Numerical Values Mapping:

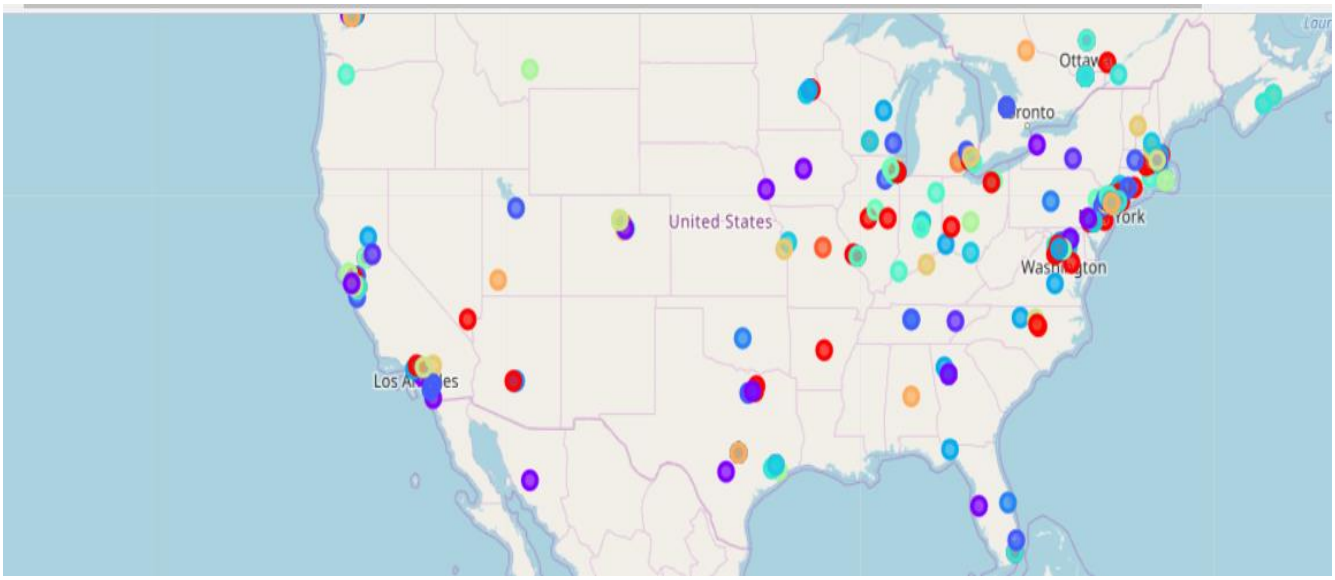
Company category column contains around 18 unique entries. Machine learning algorithms are good to handle numerical data in best way, we can convert those columns into numerical mapped values and store it in another column named category Label.

Get Location Details:

Our dataset contains only city and state details. But we need longitude and latitude details to use folium maps. So we are using **open cage package** to get location details using city and state. After successfully getting the location details in a list, I have added those details in new columns of the company data frame.

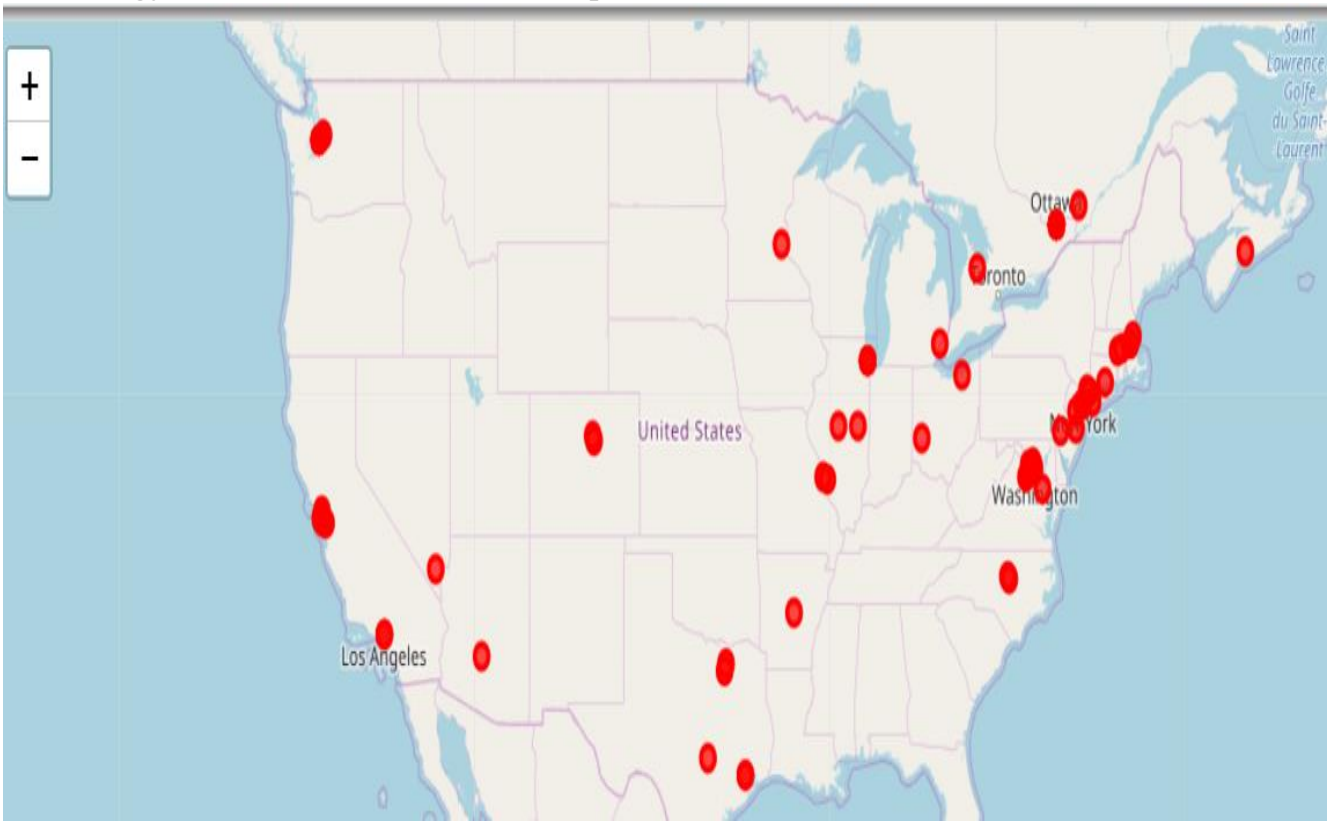
Visualizing Different Domain Companies using Folium Map:

After successfully added the location details in the data frame, I have formed clusters of different domain companies in the data frame and then visualized the different clusters using the folium map. Folium is a great visualization library that is used to visualize the map and help us to zoom in and zoom out in the created map. The map with clusters of different domain companies are as follow.



Visualizing Selected Domain Companies using Folium Map:

After Visualizing all the domain companies in the map, we can visualize only the given domain companies in the map using the prompt like thing. For simplicity, I am going to visualize the data /technology domain companies alone in the map and then we can explore the neighborhood top venues. The map of data / technology domain companies are as follows.



Connecting to Four Square API to Get Top Venues:

Once we visualized all the selected domain companies (here in our notebook I have selected Data/Technology domain companies) , I started exploring the near by top venues in the selected areas. We always want to live in the places where we will get all the facilities. So its important for us to explore all the top venues in nearby neighborhood places. Hence I have used Four Square API to explore all these details. Four Square is a famous API that is used by many companies those want to use location details and explore the nearby top venues and all. To connect to Four Square API , we need the following details. Client ID ,Client Secret and Version details and all. After connected to Four Square API, we can start exploring the near by top venues and all.

Clustering the Top Venues of Data/Technology Companies in US:

After we successfully connected to Four Square API, we can explore top venues using Four Square API and then cluster them based on the top venues. For this I used K-Means Clustering algorithm.

K-Means Algorithm:

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the *K*-means clustering algorithm are:

1. The centroids of the *K* clusters, which can be used to label new data

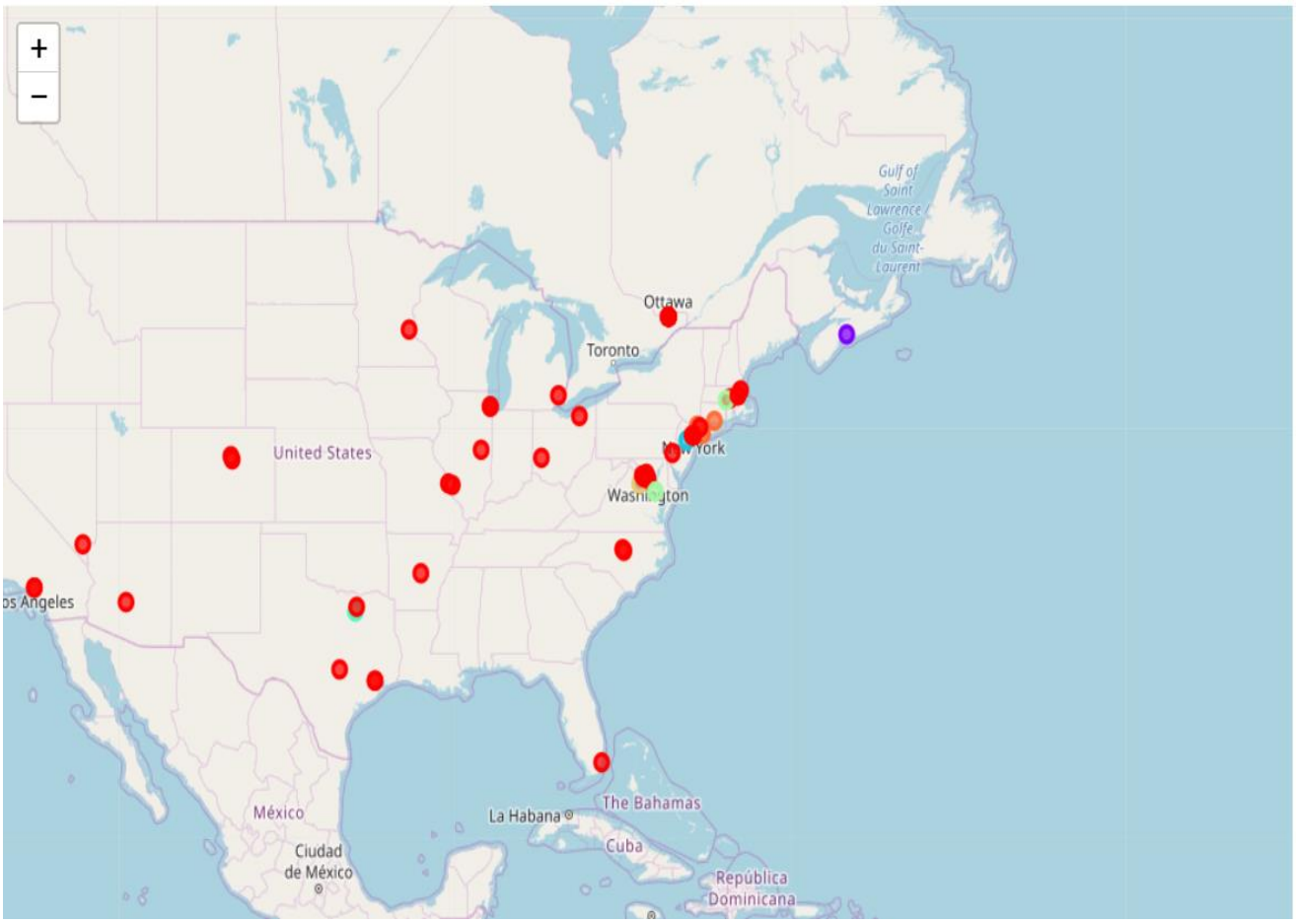
2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.

Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

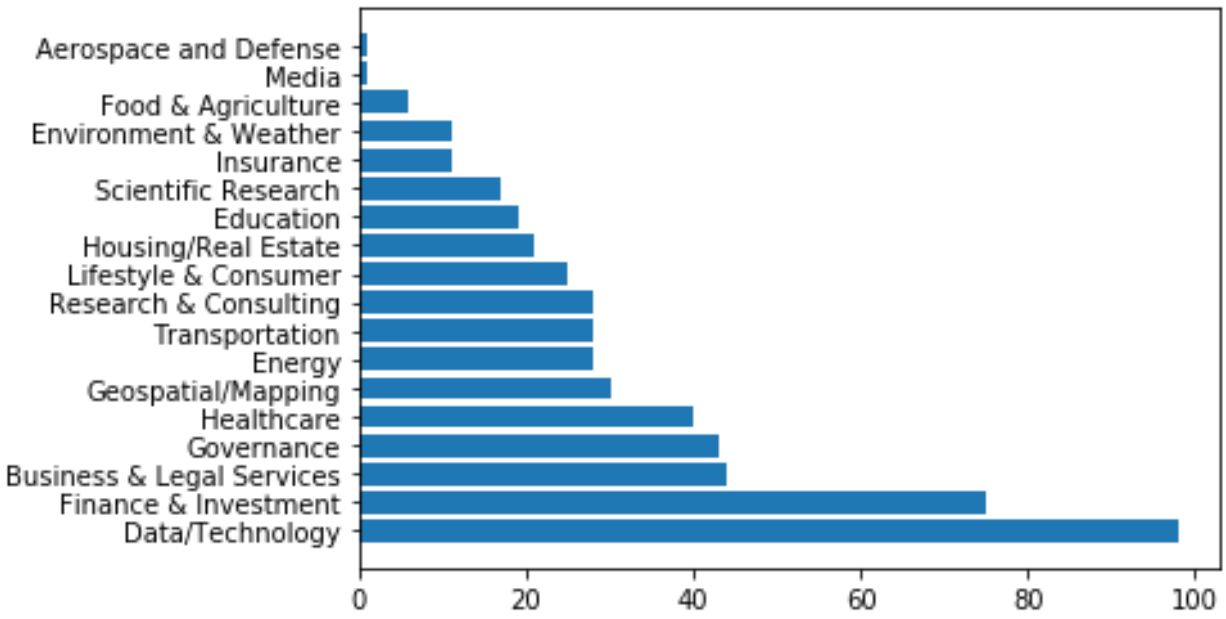
Examining Clusters:

Using K-Means Clustering Algorithm, I have clustered the top venues of the data science companies in US into 8 different clusters. We can cluster into how much clusters we want to do. These 8 clusters can be put into Folium Map and visualized using 8 different colors. This is how I have implemented clustering the near by top venues of the data science companies in US.

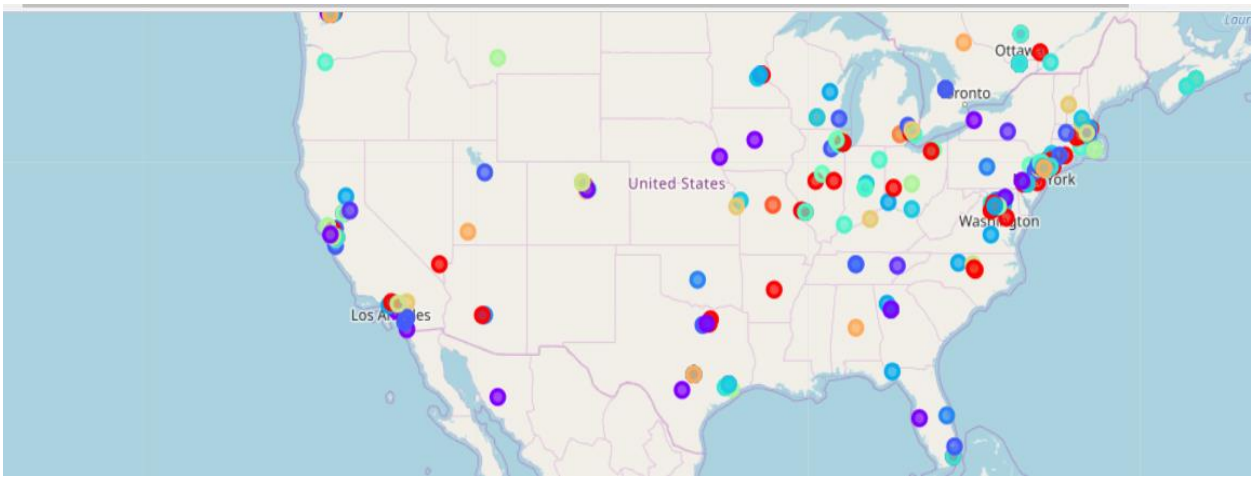


Results Section:

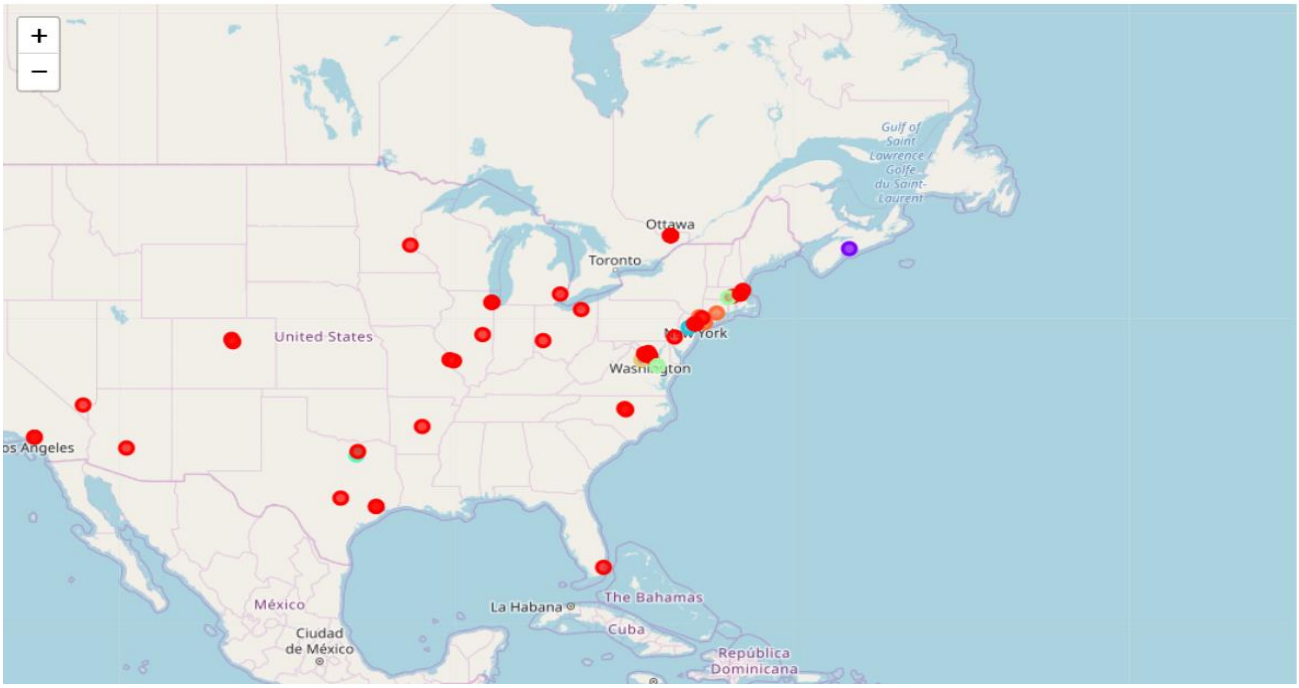
1) From the project ,we came to know that in US, there are many data science(i.e Data/Technology) companies are founded. So we can say that data science is the best domain to work with.



2) We have found out where and all different domain companies are located in US. From the map plotted below, we can say that most of the data science companies are located in Washington, New York and Toronto. So these places are best to relocate to US.



3) From the below mentioned map, we can say that the most common places near Data/Technology companies in US are restaurants, coffee shops and Bakery shops



4) We can view each and every clusters. For example, the first cluster containing mostly the restaurant venues.

Cluster 1

```
In [69]: comp_merged.loc[comp_merged['Cluster Labels'] == 0, comp_merged.columns[[1] + list(range(5, comp_merged.shape[1]))]]
```

Out[69]:

	City	CategoryLabel	lat	lon	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Washington	0	38.894893	-77.036553	0.0	Park	Hotel	Government Building	Garden	Gift Shop	History Museum	Sandwich Place
1	Fairfax	0	38.846224	-77.306373	0.0	Pizza Place	Yoga Studio	Pharmacy	Brewery	Smoothie Shop	Cafeteria	Café
2	Little Rock	0	34.746481	-92.289595	0.0	Government Building	Liquor Store	Train Station	Southern / Soul Food Restaurant	Café	Sandwich Place	Gym
3	San Jose	0	37.336191	-121.890583	0.0	Mexican Restaurant	Sandwich Place	Cocktail Bar	Sushi Restaurant	Pub	Coffee Shop	Restaurant
4	Beverly	0	42.558428	-70.880049	0.0	Pizza Place	Pharmacy	Donut Shop	Breakfast Spot	Sandwich Place	BBQ Joint	Automotive Shop
5	Seattle	0	47.603832	-122.330062	0.0	Coffee Shop	Hotel	Café	Cocktail Bar	Japanese Restaurant	Sandwich Place	Salad Place

Discussion Section:

Our problem statement is like when job seekers want to search for jobs in particular location, they will find it difficult to google it. If we provide them with list of different domain companies in a map and let them explore the top venues in the nearby places makes them easy to come to know about the feasibility level of migration so that they can easily decide and don't need to spend too much time in web surfing for the details. This project was implemented only for US companies. **When we collect all the country company details, we can process the huge data (Big Data) using Hadoop and make it work for all the countries**. That's what my recommendation for this project. That may be future work of this project.

Conclusion:

First I identified the problem statement I took . The business requirement is like, when we want to migrate to some other company and in some other location, then we will find it difficult to migrate since we don't know about the location. In this project, I have addressed about how to visualize the different domain companies in US country. Then I clustered the companies based on their domain and then I visualized the data using Folium map so that it will be easy to understand them. The another important thing what job seekers will expect is like, when they migrate, will they avail all the facilities in the new location. To address this problem, I have clustered the top venues around the Data/Technology domain companies and visualized it using Folium Map. We can collect all the country company details and make this project work for all the countries and all domain companies. From the project, I came to know that Data Science is the most growing field in the World especially in US.