CITIES CLUSTERING PROJECT

By- Nirali Parekh
Github - nirali25parekh
Date- March 2020

Ideation

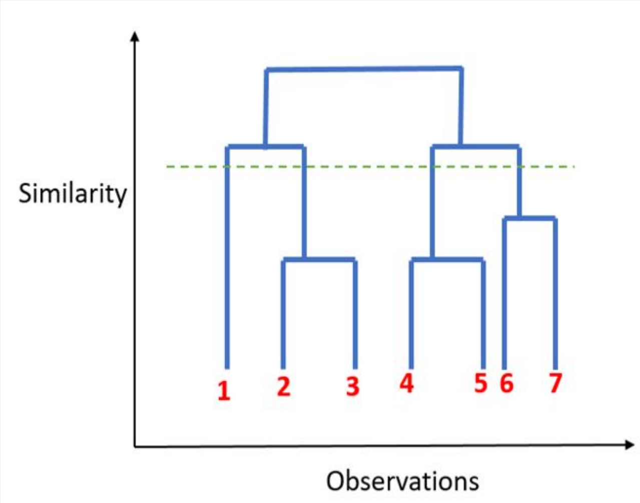
This project focuses on clustering of similar cities types according to venues in the area.
Hence, by the end, we will be able to find and segregate similar cities based on categories of places and venues that are abundant in the cities.

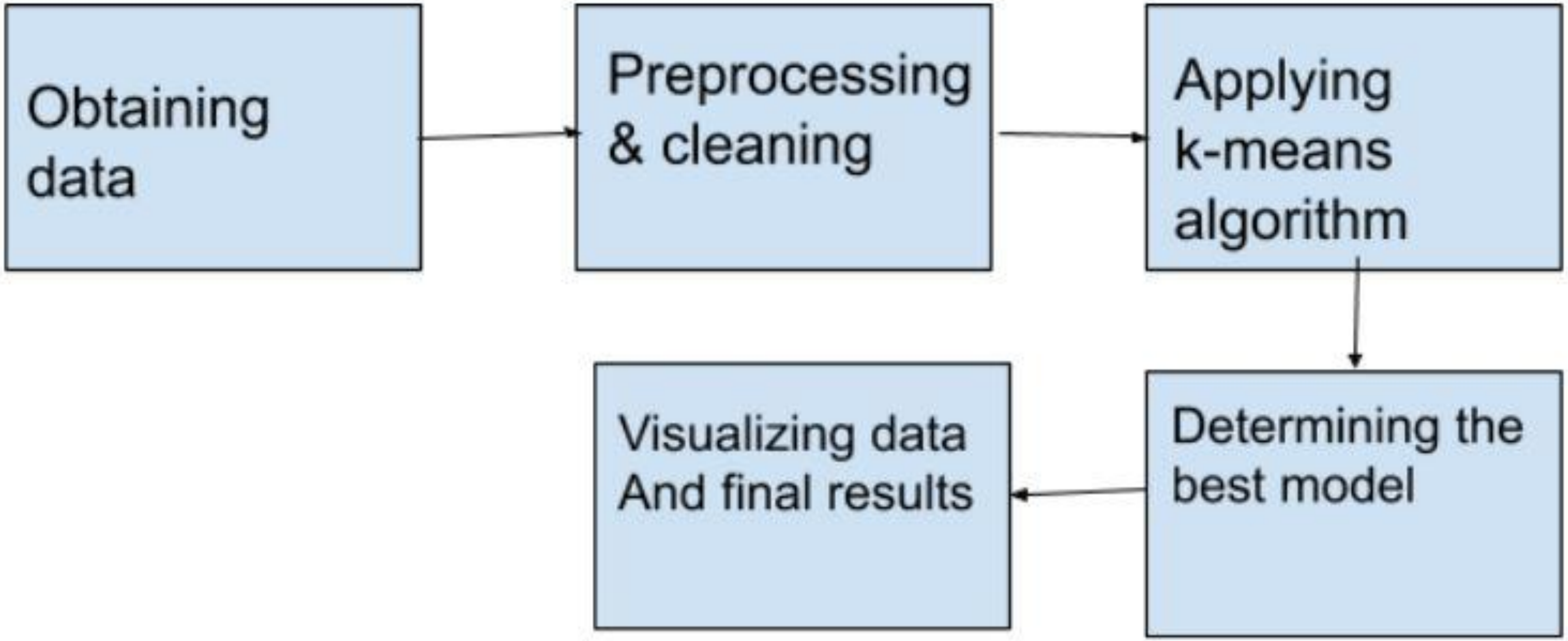
Method

Unsupervised Learning (Clustering) :

“**Clustering**” is the process of grouping similar entities together. The goal of this technique is to find similarities in the data point and group similar data points together.

In our case, we are obtaining data from various sources and using it, we are clustering or ‘grouping’ cities together.





Obtaining data

Preprocessing & cleaning

Applying k-means algorithm

Determining the best model

Visualizing data And final results

Fig: process of the cities clustering flow

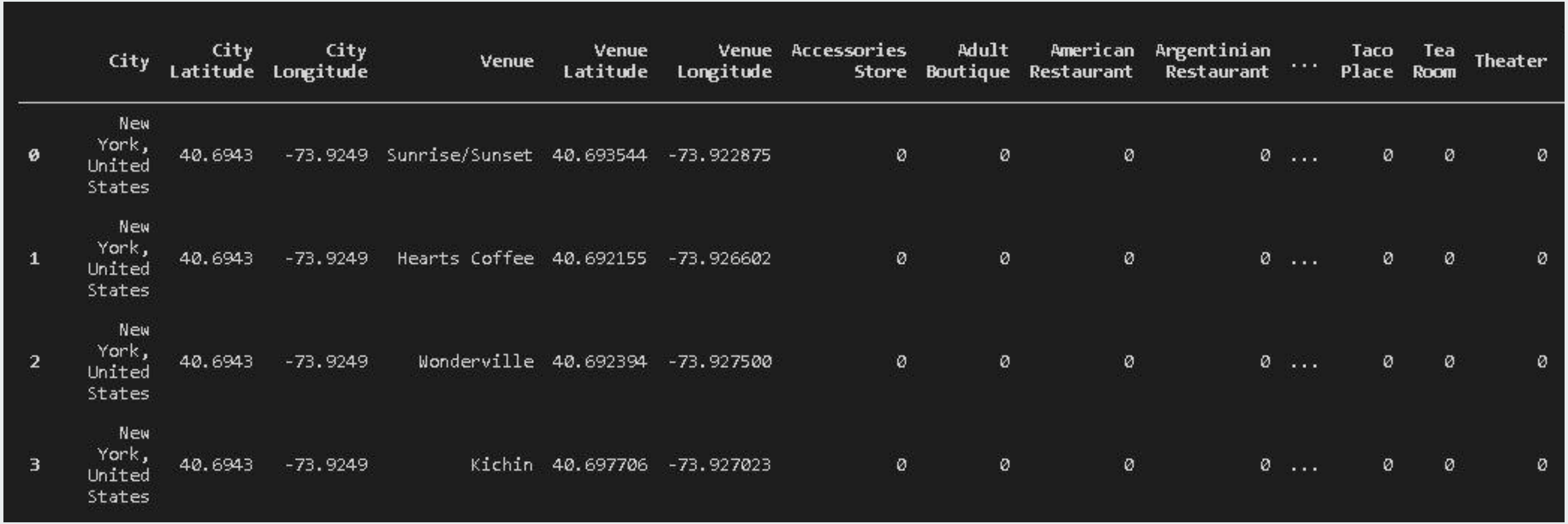
1. Obtaining Data

- The geographical data was obtained by the free ‘Single Maps’ Database.
- The information from the database (considered as features) were:
  - City
  - Latitude
  - Longitude
  - Country
  - Population
  - Capital typel (if any)
- Here, FourSquare API was used to collect data of venues for each city.
  https://developer.foursquare.com/docs/

Step 2: Preprocessing and Cleaning:

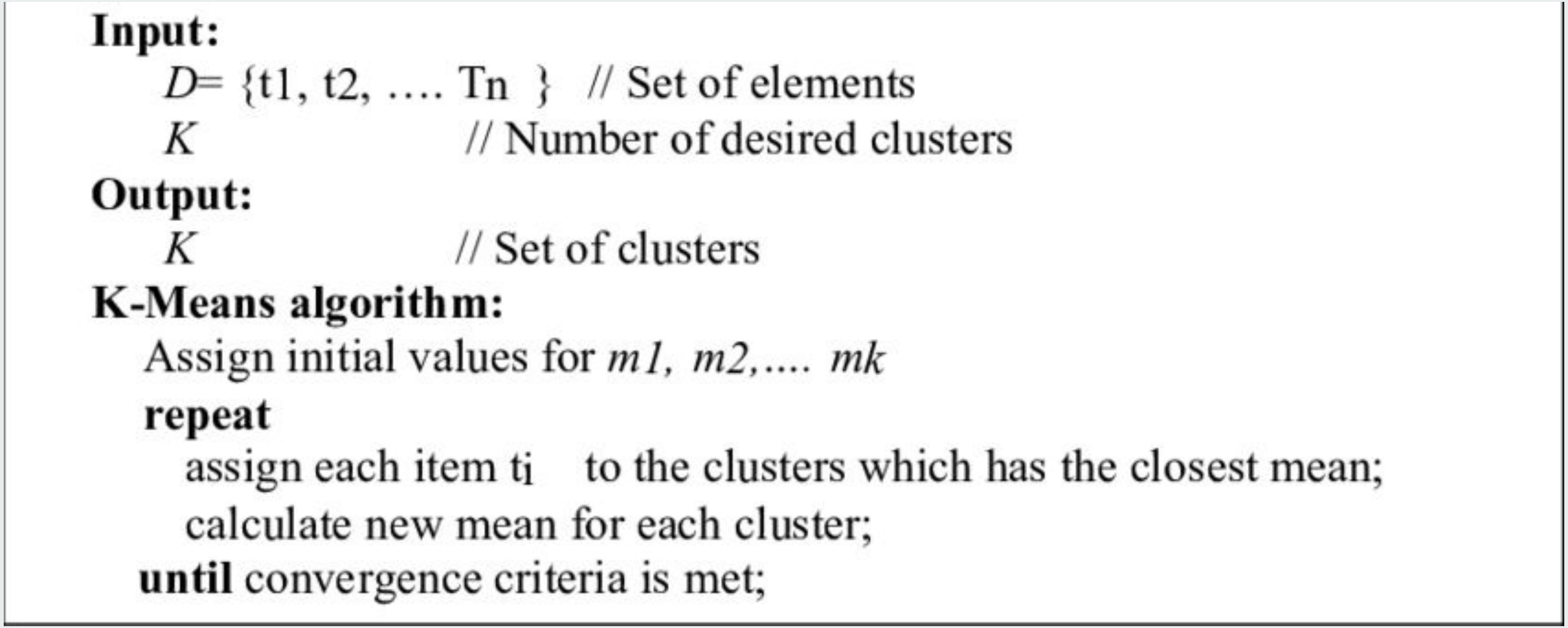
- This consists of getting rid of unwanted data (null values) and tweaking data according to our needs.
- Here we, used One-Hot-Encoding to convert our categorical values to digits.

One Hot Encoding



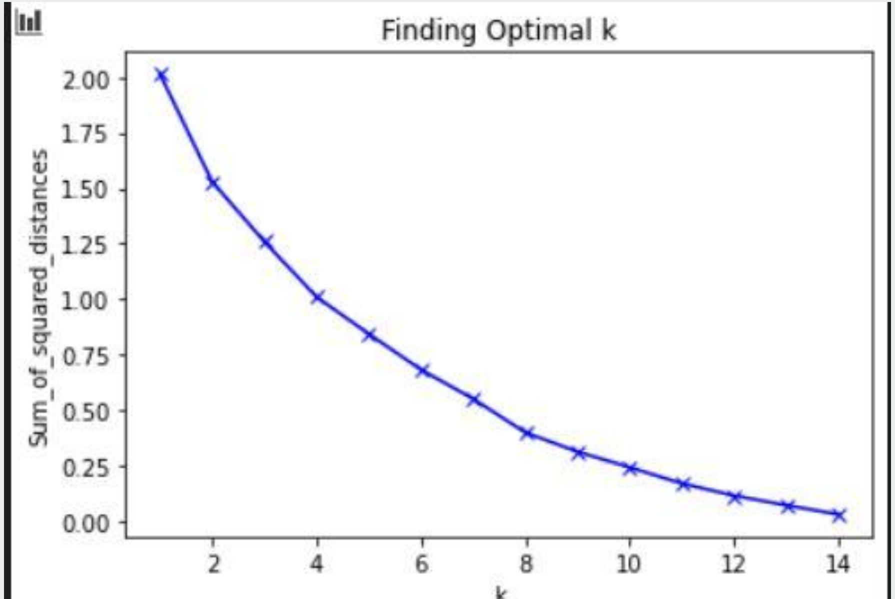
	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Accessories Store	Adult Boutique	American Restaurant	Argentinian Restaurant	...	Taco Place	Tea Room	Theater
0	New York, United States	40.6943	-73.9249	Sunrise/Sunset	40.693544	-73.922875	0	0	0	0	...	0	0	0
1	New York, United States	40.6943	-73.9249	Hearts Coffee	40.692155	-73.926602	0	0	0	0	...	0	0	0
2	New York, United States	40.6943	-73.9249	Wonderville	40.692394	-73.927500	0	0	0	0	...	0	0	0
3	New York, United States	40.6943	-73.9249	Kichin	40.697706	-73.927023	0	0	0	0	...	0	0	0

Step 3: K Means Algorithm:



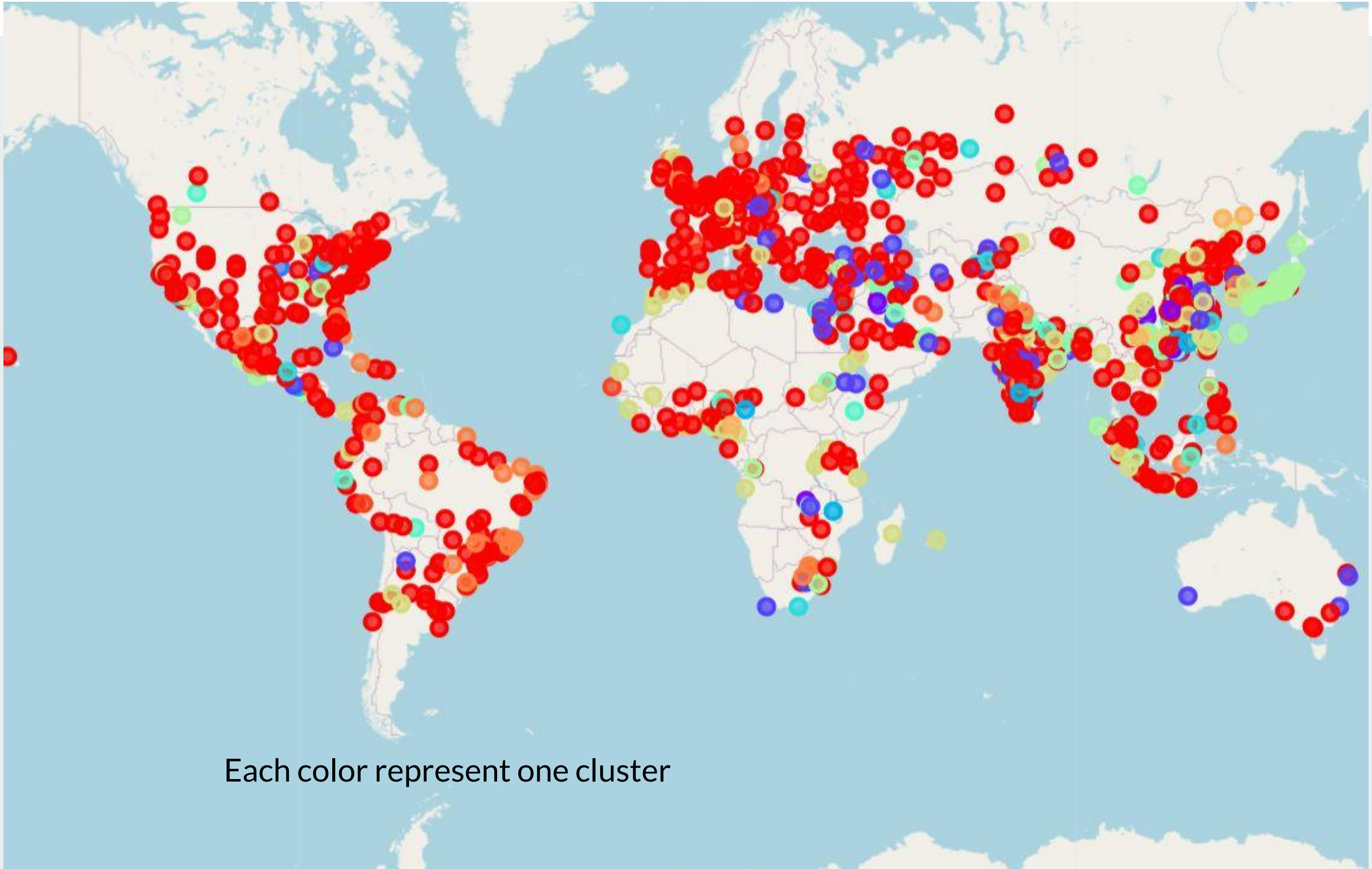
```
Input:
    D= {t1, t2, …. Tn }   // Set of elements
    K                     // Number of desired clusters
Output:
    K                     // Set of clusters
K-Means algorithm:
  Assign initial values for m1, m2,…. mk
  repeat
    assign each item ti  to the clusters which has the closest mean;
    calculate new mean for each cluster;
  until convergence criteria is met;
```

Step 4: Choosing the best model:



Step 5: Visualizing Data and Final Results:

- **Data visualization** is the graphical representation of information and **data**. By using visual elements like charts, graphs, and maps, **data visualization** tools provide an accessible way to see and understand trends, outliers, and patterns in **data**.
- Visualizing final data is an important step in data science.
- It helps the people to understand what to make sense of the data.
- Tools used: folium - for maps
- Matplotlib - for graphs



Each color represent one cluster

Thank you!