Relation Between USA Cities by COVID-19 positive cases

Omar Villaseñor

November 4, 2020

1. Introduction

1.1. Background

A coronavirus is a common type of virus that causes an infection in the nose, sinuses, or upper throat. Most coronaviruses are not dangerous. In early 2020, after a December 2019 outbreak in China, the World Health Organization identified SARS-CoV-2 as a new type of coronavirus. The outbreak quickly spread around the world. COVID-19 is a disease caused by SARS-CoV-2 that can trigger what doctors call a respiratory tract infection. It can affect the upper respiratory tract (sinuses, nose and throat) or the lower respiratory tract (trachea and lungs). Currently the world is in a pandemic situation due to this virus.

1.2 Problem

The United States is probably the most important country on the planet, so the pandemic impacts a lot here and the main cities obtained many positive cases, but it is not always known which of these cities have in common to identify which cities can adopt similar techniques to combat . this virus. This project seeks to identify cities by similarity and group them.

1.3 Utility

Now, the advantage of knowing this information is to understand which cities are similar by principal venues and which are capitals of their respective state.

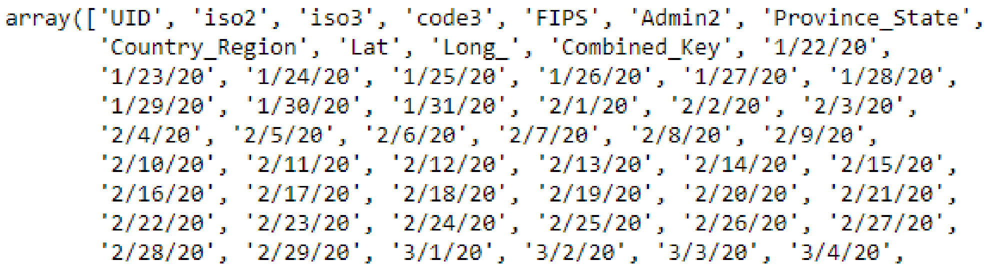
2. Data

2.1. Data Sources

For this project we will obtain the Novel Coronavirus 2019 dataset, using the ***time_series_covid_19_confirmed_US.csv*** file, provided by a user of Kaggle platform (related to this link: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset"). This dataset have the number of positive cases of COVID-19 registered in USA separated by date and cities.

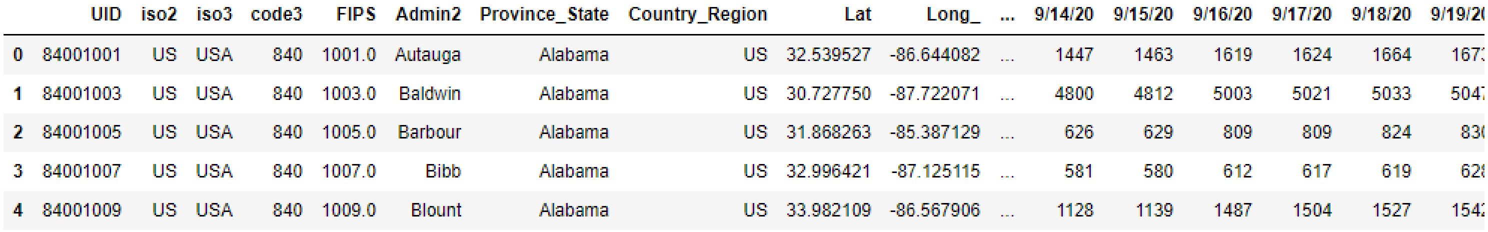
2.2 Data Cleaning

This dataset very organized, but it has some details that we need to change to use in the project. So in the first case we analyze the data and see the names of each column, we got something like this:



```
array(['UID', 'iso2', 'iso3', 'code3', 'FIPS', 'Admin2', 'Province_State',
       'Country_Region', 'Lat', 'Long_', 'Combined_Key', '1/22/20',
       '1/23/20', '1/24/20', '1/25/20', '1/26/20', '1/27/20', '1/28/20',
       '1/29/20', '1/30/20', '1/31/20', '2/1/20', '2/2/20', '2/3/20',
       '2/4/20', '2/5/20', '2/6/20', '2/7/20', '2/8/20', '2/9/20',
       '2/10/20', '2/11/20', '2/12/20', '2/13/20', '2/14/20', '2/15/20',
       '2/16/20', '2/17/20', '2/18/20', '2/19/20', '2/20/20', '2/21/20',
       '2/22/20', '2/23/20', '2/24/20', '2/25/20', '2/26/20', '2/27/20',
       '2/28/20', '2/29/20', '3/1/20', '3/2/20', '3/3/20', '3/4/20',
```

As we can see, we obtained a lot of dates, so we see a sample of rows from this dataset to identify the definition of each column and the content of the same. So, the data looks like this:



	UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	...	9/14/20	9/15/20	9/16/20	9/17/20	9/18/20	9/19/2
0	84001001	US	USA	840	1001.0	Autauga	Alabama	US	32.539527	-86.644082	...	1447	1463	1619	1624	1664	167
1	84001003	US	USA	840	1003.0	Baldwin	Alabama	US	30.727750	-87.722071	...	4800	4812	5003	5021	5033	504
2	84001005	US	USA	840	1005.0	Barbour	Alabama	US	31.868263	-85.387129	...	626	629	809	809	824	83
3	84001007	US	USA	840	1007.0	Bibb	Alabama	US	32.996421	-87.125115	...	581	580	612	617	619	62
4	84001009	US	USA	840	1009.0	Blount	Alabama	US	33.982109	-86.567906	...	1128	1139	1487	1504	1527	154

Thanks to this visualization, we can identify the content of each column and the conclusions are:

- The city name has an incorrect column name (Admin2)
- We need the Province_State column to identify the state of the city
- The coordinates are separated by two columns: ***Lat*** and ***Long_***.
- Each date have an accumulative number of positive cases.

The data has a lot of columns to ignore, as we want to compare the top 200 cities with the most positive cases, so the latest date is perfect for this task.

Now, we separate all the data in a new Dataframe with our specified columns and the top 200 cities, changing the name of the columns like:

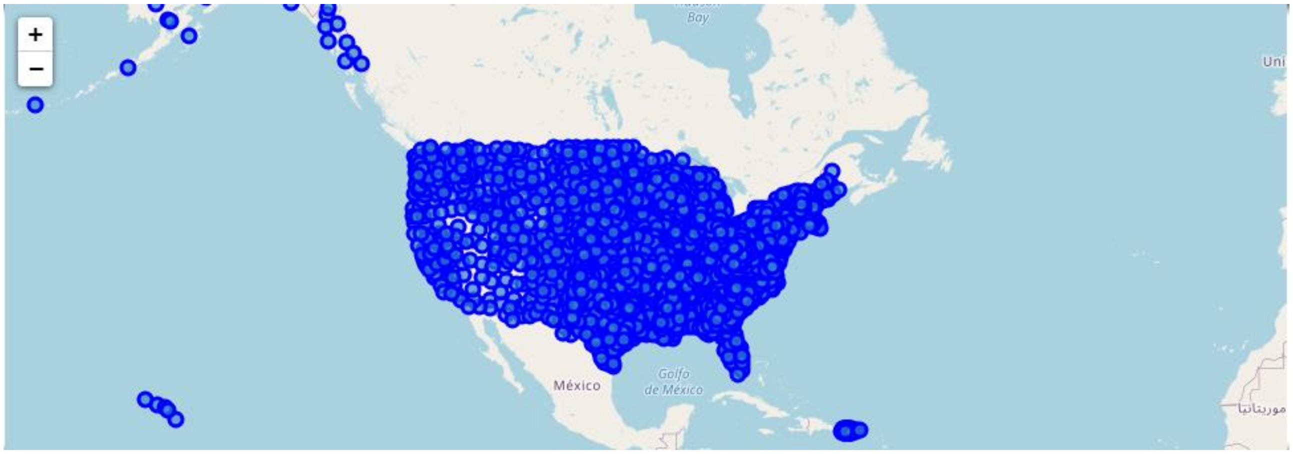
- **Admin2** to ***City***
- **Province_State** to ***State***
- **Lat** to ***Latitude***
- **Long_** to ***Longitude***
- **9/23/20** to ***Total***

3. Methodology

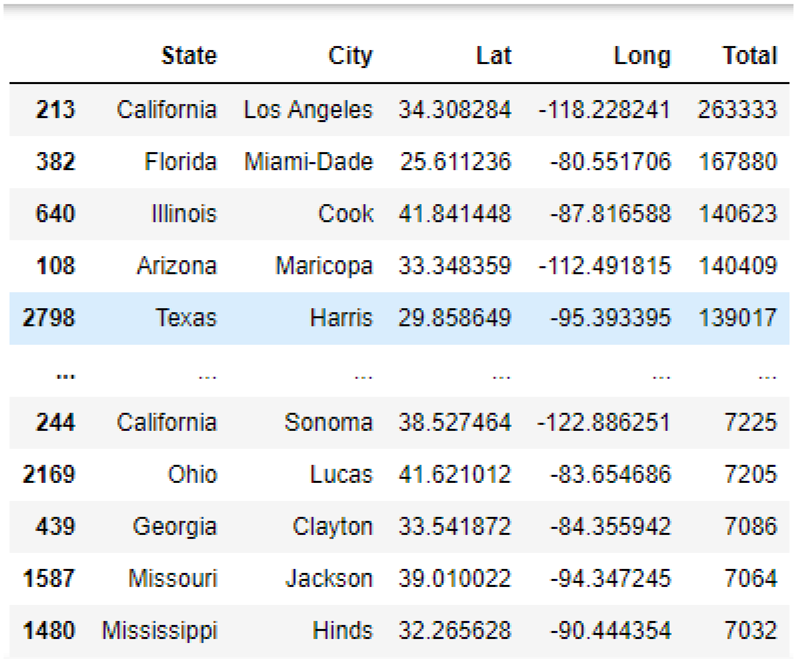
To do this project, the first part is to obtain the data and perform the entire process that we describe in the Data section. Then, we follow the following steps:

As good information, we see all the cities contained in the data set by state.

We plot all the cities in a map to contrast this data, and the result looks like this:



But the US (and our dataset) have many cities, so we need to select a set by applying some criteria. This criterion is: the top 200 cities with the most positive cases of COVID-19. To do this, we sort the dataset by the last date (total accumulated cases) and select the first 200. Later, we take the necessary columns to reach the objective. This looks like:

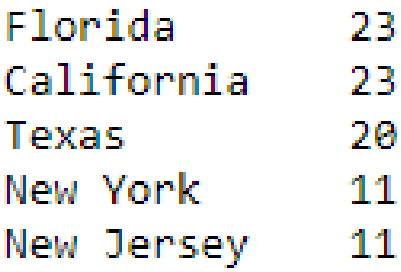


	State	City	Lat	Long	Total
213	California	Los Angeles	34.308284	-118.228241	263333
382	Florida	Miami-Dade	25.611236	-80.551706	167880
640	Illinois	Cook	41.841448	-87.816588	140623
108	Arizona	Maricopa	33.348359	-112.491815	140409
2798	Texas	Harris	29.858649	-95.393395	139017
...	...	...	...	...	...
244	California	Sonoma	38.527464	-122.886251	7225
2169	Ohio	Lucas	41.621012	-83.654686	7205
439	Georgia	Clayton	33.541872	-84.355942	7086
1587	Missouri	Jackson	39.010022	-94.347245	7064
1480	Mississippi	Hinds	32.265628	-90.444354	7032

Now, it’s time to start the process.

The relation between cities

We have 1979 cities by total. But in this project we only select 200. Ok, that´s right, now we want to know: What is the top 5 States with more cities in this 200 set?, well, this is the answer:



```
Florida       23
California    23
Texas         20
New York      11
New Jersey    11
```

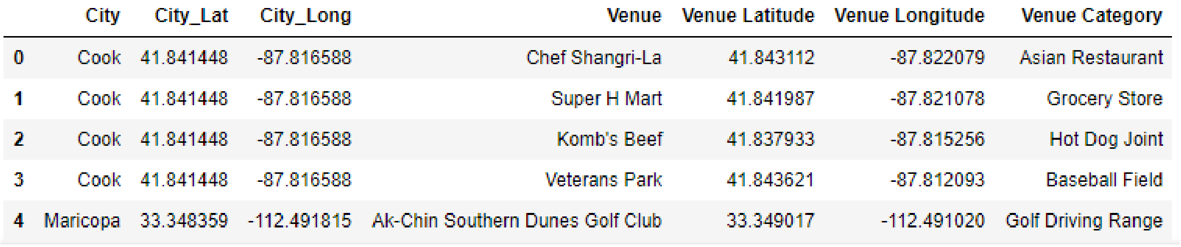
With this info, now we can expect more points in the Florida/California region at the end of the project.

The API

The main objective of the project is to apply Data Science techniques with geographic tools. First, we define the credentials to use the Foursquare API and get 10 places within a 500 meter radius for each city center.

As we search for this information, it will be stored in a new dataset for later analysis.

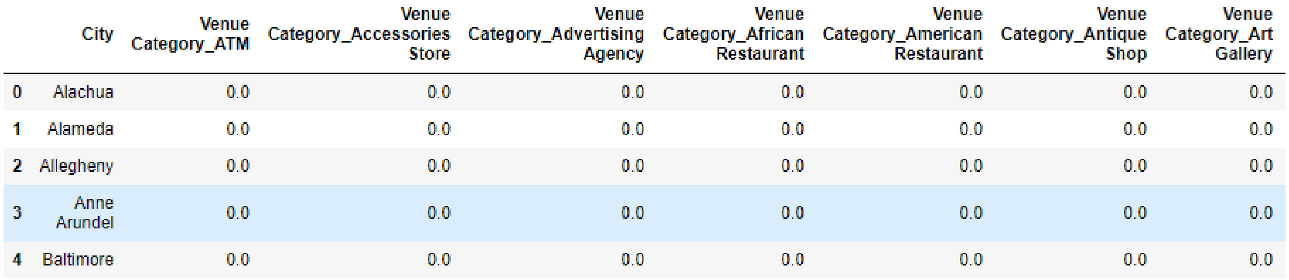
At the end of the search, we got a data set like this:



	City	City_Lat	City_Long	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Cook	41.841448	-87.816588	Chef Shangri-La	41.843112	-87.822079	Asian Restaurant
1	Cook	41.841448	-87.816588	Super H Mart	41.841987	-87.821078	Grocery Store
2	Cook	41.841448	-87.816588	Komb's Beef	41.837933	-87.815256	Hot Dog Joint
3	Cook	41.841448	-87.816588	Veterans Park	41.843621	-87.812093	Baseball Field
4	Maricopa	33.348359	-112.491815	Ak-Chin Southern Dunes Golf Club	33.349017	-112.491020	Golf Driving Range

But of course the dataset is not in a good format to apply a clustering algorithm, so we transform it into a numeric dataset using One-hot encoding technique and getting the mean of each match.

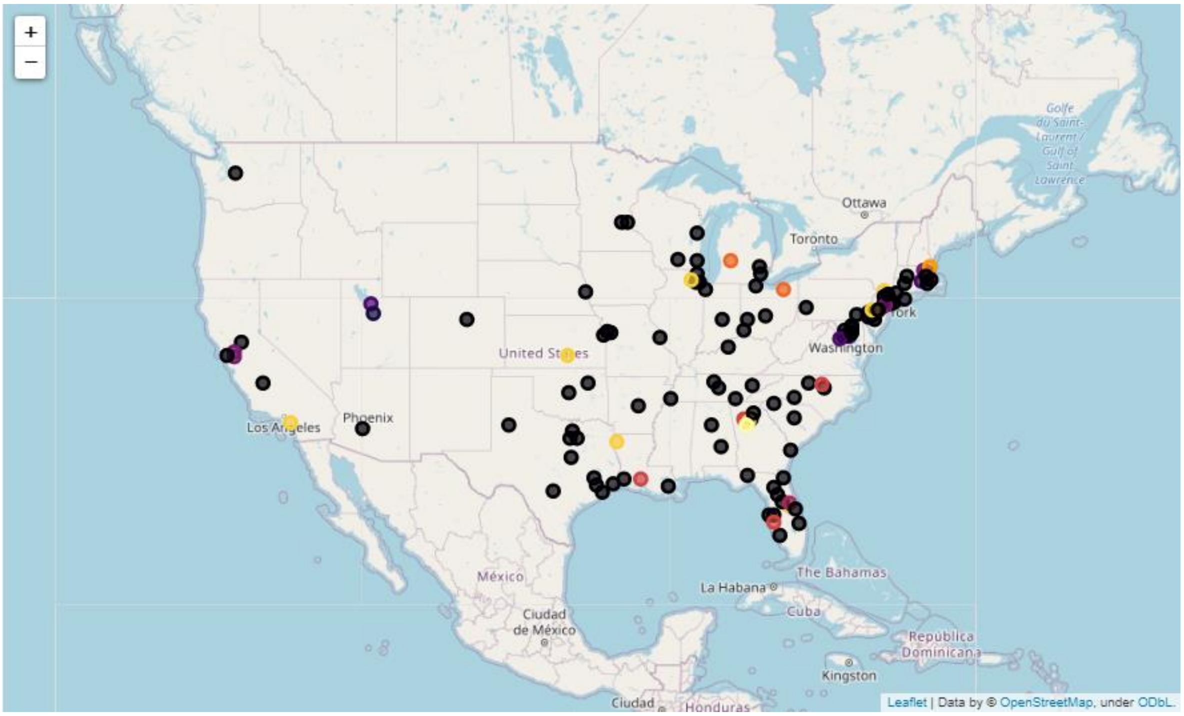
The trainable dataset looks like this:



	City	Venue Category_ATM	Venue Category_Accessories Store	Venue Category_Advertising Agency	Venue Category_African Restaurant	Venue Category_American Restaurant	Venue Category_Antique Shop	Venue Category_Art Gallery
0	Alachua	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Alameda	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Allegheny	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Anne Arundel	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Baltimore	0.0	0.0	0.0	0.0	0.0	0.0	0.0

This dataset is ready to be trained, so let’s apply KMeans, searching 10 clusters.

Once we have the results, we add them to our principal data and plotting in a map like this:



4. Conclusions

The final map shows us some different insights:

Many cities have places in common, which means that we can apply similar rules to these cities to fight the virus. Furthermore, we can see that the distribution of the cities with the most positive cases are located in the western part of the country, so it is necessary to enforce this regulation in this sector.

5. Improvements

This project can be constantly improved with new data on COVID-19, so the scope of this can be greater, because this data has many possibilities.