Cap Stone IBM Project

Compare the top Populated Cities in the US

Introduction

I will view the top populated cities in the USA. I will cluster them into groups and visualize the result on a map. I will adjust the clusters from 5 to 3 and see the effect on the map. One of the questions I want to see if I can answer is in the USA is there a difference or similar in certain areas of the USA?

Example will we see a difference when we compare the West Coast, East Coast, and the Midwest. Also I will compare the very large cities with the smaller cities. Or is the US a "melting pot" and there is not much difference throughout the country when it comes to the types of venues located in the most populated cities in the US.

I would like the compare the cities and determine how similar or dissimilar they are. Through this project I am expecting following people to benefit out of the findings.

- People moving to different cities for work
- Business Companies looking for new locations or to expand
- Restaurants to adjust their menu based on the people's likings and feedbacks

Data

The data I will be using will come from the following web site, https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

I will use beautiful soap to read the tables from this website. I will have to do some data "cleaning" in python pandas to get the data into a correct dataframe, so that the table and data will be ready to be read by the Foursquare API program section on my Python program.

2018 rank	City	State[c]	2018 estimate	2010 Census
1	*New York City*[d]	New York	8,398,748	8,175,133
2	*Los Angeles*	California	3,990,456	3,792,621
3	*Chicago*	Illinois	2,705,994	2,695,598
4	*Houston*[3]	Texas	2,325,502	2,100,263
5	***Phoenix***	Arizona	1,660,272	1,445,632
6	*Philadelphia*[e]	Pennsylvania	1,584,138	1,526,006
7	San Antonio	Texas	1,532,233	1,327,407
8	San Diego	California	1,425,976	1,307,402
9	Dallas	Texas	1,345,047	1,197,816
10	San Jose	California	1,030,119	945,942

Methodology

After loading in the data from the website (https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population)

I had to use Pandas to clean the data to get the following dataframe output with:

- Rank
- City
- Latitude
- Longitude

Completed Cleaning the Data Table

```
[15]: new_data1.head()
```

[15]:

	rank	city	state	estimate	census	area mi	den mi	lat	long
0	1	New York City[d]	New York	8,398,748	8,175,133	301.5 sq mi	28,317/sq mi	40.6635	-73.9387
1	2	Los Angeles	California	3,990,456	3,792,621	468.7 sq mi	8,484/sq mi	34.0194	-118.4108
2	3	Chicago	Illinois	2,705,994	2,695,598	227.3 sq mi	11,900/sq mi	41.8376	-87.6818
3	4	Houston[3]	Texas	2,325,502	2,100,263	637.5 sq mi	3,613/sq mi	29.7866	-95.3909
4	5	Phoenix	Arizona	1,660,272	1,445,632	517.6 sq mi	3,120/sq mi	33.5722	-112.0901

```
[16]: new_data1.tail()
```

[16]:

	rank	city	state	estimate	census	area mi	den mi	lat	long
309	310	Edison[ad]	New Jersey	100,693	99,967	30.1 sq mi	3,389/sq mi	40.5040	-74.3494
310	311	Woodbridge[ad]	New Jersey	100,450	99,585	23.3 sq mi	4,351/sq mi	40.5607	-74.2927
311	312	San Angelo	Texas	100,215	93,200	59.9 sq mi	1,681/sq mi	31.4411	-100.4505
312	313	Kenosha	Wisconsin	100,164	99,218	28.0 sq mi	3,577/sq mi	42.5822	-87.8456
313	314	Vacaville	California	100,154	92,428	29.0 sq mi	3,449/sq mi	38.3539	-121.9728

```
[17]: new_data1.dtypes
```

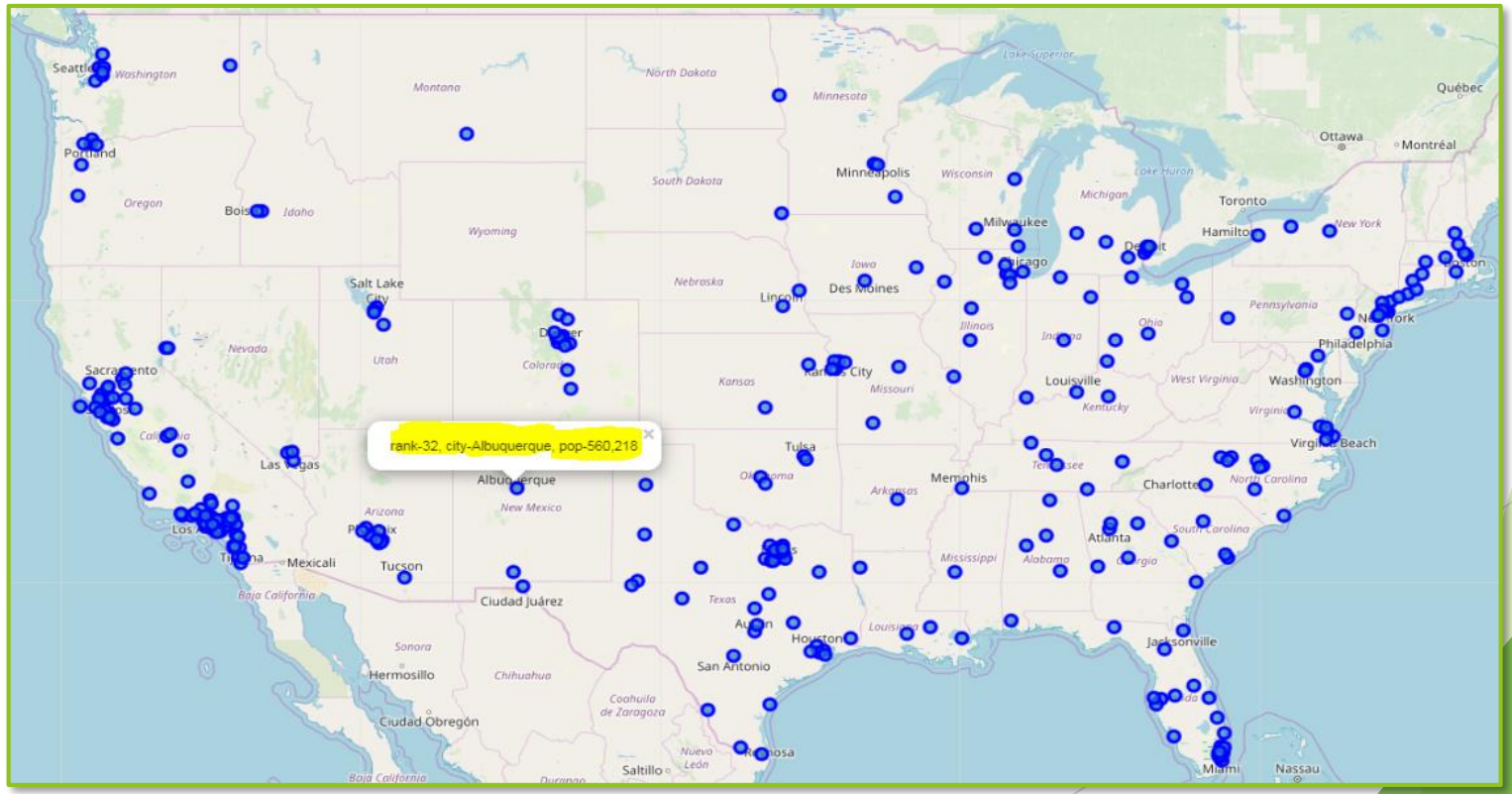
```
[17]: rank         object
      city         object
      state        object
      estimate     object
      census       object
      area mi      object
      den mi       object
      lat         float64
      long        float64
      dtype: object
```

Methodology

Because I am not a master at Pandas yet, it took me several steps to get the data into the right format.

I then used the folium package to visualize the map of the US with displaying the pop-up label with Rank, City and Population.

I did this step so that I knew the data was “clean” and was displaying correctly on the map.



Methodology

I next used the Foursquare section of the program to explore the cities and venues.

I set the limits to 20 venues and with a radius of 1000 meters. This list is 4941 lines deep. So I know I did not get 20 venues for all of the 314 cities in the list. As you can see to the right.

In the picture to the right, you can see New York with 20 venues and then Los Angeles as the next city on the list.

Charleston	15	15	15	15	15	15
Charlotte	20	20	20	20	20	20
Chattanooga	4	4	4	4	4	4
Chicago	20	20	20	20	20	20
Chula Vista	6	6	6	6	6	6
Cincinnati	20	20	20	20	20	20
Clarksville	2	2	2	2	2	2
Clearwater	7	7	7	7	7	7
Cleveland	20	20	20	20	20	20
Clinton[ae]	5	5	5	5	5	5
Clovis	20	20	20	20	20	20

	city	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	New York City[d]	40.6635	-73.9387	Izzy's Brooklyn Smokehouse	40.664869	-73.937023	BBQ Joint
1	New York City[d]	40.6635	-73.9387	Ali's Roti Shop	40.666436	-73.931346	Caribbean Restaurant
2	New York City[d]	40.6635	-73.9387	Sweet Expressions	40.668382	-73.942337	Candy Store
3	New York City[d]	40.6635	-73.9387	Bodega	40.668388	-73.932366	Deli / Bodega
4	New York City[d]	40.6635	-73.9387	The Market Place	40.662153	-73.943010	Grocery Store
5	New York City[d]	40.6635	-73.9387	Mama Louisa's Hero Shoppe	40.659496	-73.947519	Sandwich Place
6	New York City[d]	40.6635	-73.9387	Conrad's Famous Bakery, III, Inc.	40.667942	-73.931033	Bakery
7	New York City[d]	40.6635	-73.9387	Calabria	40.670420	-73.942230	Pizza Place
8	New York City[d]	40.6635	-73.9387	Crunch Fitness - Crown Heights	40.663335	-73.932808	Gym / Fitness Center
9	New York City[d]	40.6635	-73.9387	Jewish Children's Museum	40.669017	-73.942086	Museum
10	New York City[d]	40.6635	-73.9387	Sushi Spot	40.664320	-73.942705	Japanese Restaurant
11	New York City[d]	40.6635	-73.9387	Blink Fitness Crown Heights	40.669828	-73.931352	Gym
12	New York City[d]	40.6635	-73.9387	Fish N Chips (Soul of the Sea)	40.661951	-73.940055	Fish & Chips Shop
13	New York City[d]	40.6635	-73.9387	Kings County Nurseries	40.659171	-73.947301	Garden Center
14	New York City[d]	40.6635	-73.9387	Old Boys High Field	40.658667	-73.938427	Field
15	New York City[d]	40.6635	-73.9387	Three Star Juice Lounge	40.661465	-73.931672	Juice Bar
16	New York City[d]	40.6635	-73.9387	Family Dollar	40.662678	-73.933558	Discount Store
17	New York City[d]	40.6635	-73.9387	White Castle	40.663357	-73.932238	Fast Food Restaurant
18	New York City[d]	40.6635	-73.9387	Rite Aid	40.661066	-73.931816	Pharmacy
19	New York City[d]	40.6635	-73.9387	Bakerie	40.672013	-73.939183	Café
20	Los Angeles	34.0194	-118.4108	Oldfield's Liquor Room	34.016286	-118.411881	Cocktail Bar
21	Los Angeles	34.0194	-118.4108	Yogurtland	34.017710	-118.407116	Frozen Yogurt Shop
22	Los Angeles	34.0194	-118.4108	Pampas Grill Culver City	34.016929	-118.406503	Brazilian Restaurant
23	Los Angeles	34.0194	-118.4108	Bella Vista Brazilian Gourmet Pizza	34.016824	-118.409644	Pizza Place
24	Los Angeles	34.0194	-118.4108	Robeks Fresh Juices & Smoothies	34.017295	-118.406101	Smoothie Shop
25	Los Angeles	34.0194	-118.4108	Kogi Taqueria	34.024653	-118.411534	Taco Place

Methodology

When I looked at all of the Venue Categories, I had a total of 379 venues.

Next I took only the top 10 venue categories for each city, see picture on the right, where some cities did not have a percentage for all 10 venue categories.

Next I created a table showing the 10 ten most common venues and merged that data with the original list of 314 cities.

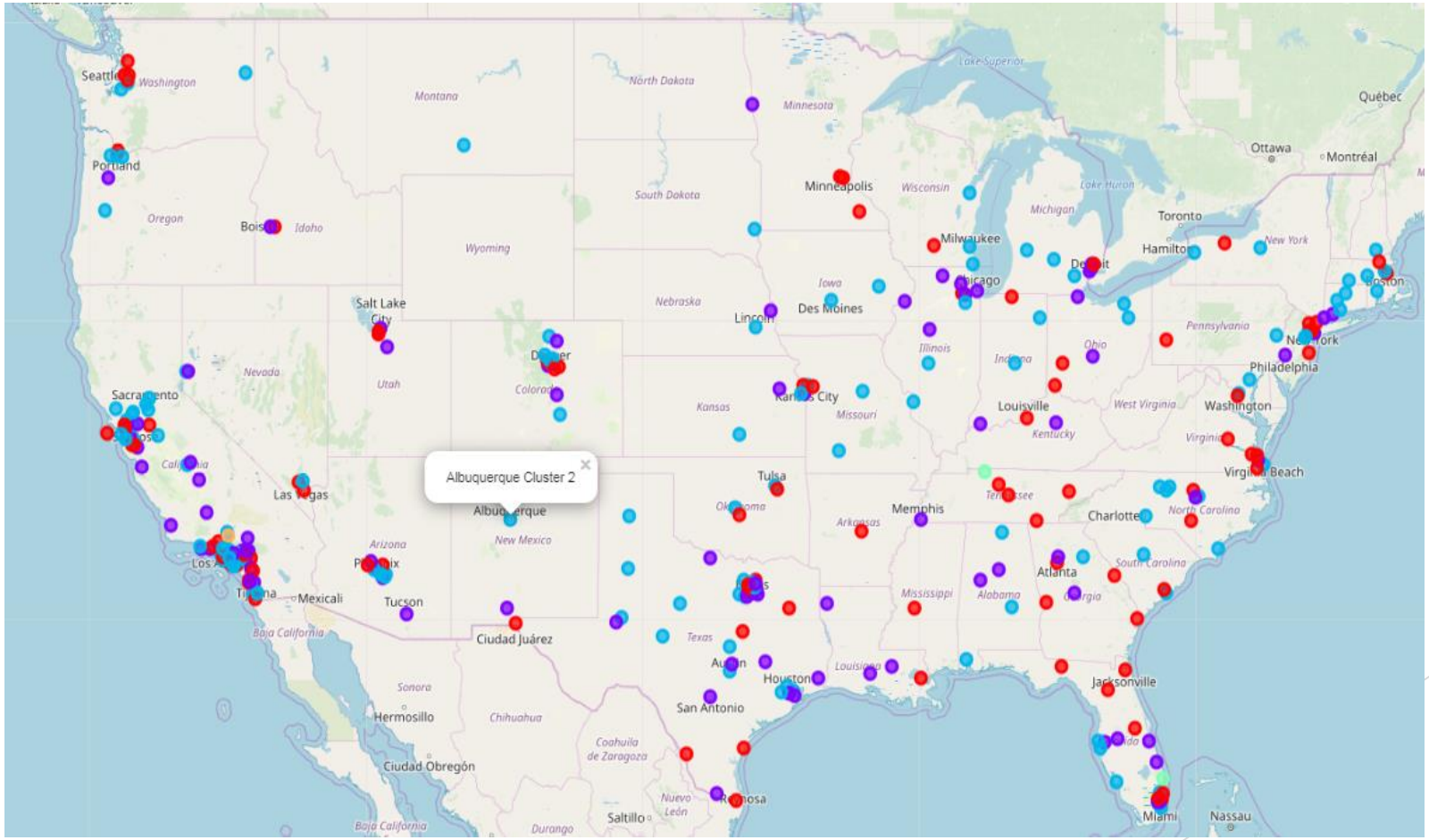
```
----Oxnard----
                venue  freq
0    Mexican Restaurant  0.15
1      Airport Terminal  0.15
2              Pharmacy  0.15
3          Optical Shop  0.08
4       Doctor's Office  0.08
5  Fast Food Restaurant  0.08
6     Fish & Chips Shop  0.08
7     Convenience Store  0.08
8               Airport  0.08
9   Rental Car Location  0.08


----Palm Bay----
                         venue  freq
0                 Home Service   0.2
1                Ice Cream Shop   0.2
2            Chinese Restaurant   0.2
3                   Pizza Place   0.2
4                   Golf Course   0.2
5                  Music School   0.0
6                        Office   0.0
7                          Park   0.0
8  Paper / Office Supplies Store   0.0
9              Paella Restaurant   0.0
```

	rank	city	state	estimate	census	area mi	den mi	lat	long	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	1	New York City[d]	New York	8,398,748	8,175,133	301.5 sq mi	28,317/sq mi	40.6635	-73.9387	1.0	BBQ Joint	Grocery Store	Field	Fast Food Restaurant	Museum
1	2	Los Angeles	California	3,990,456	3,792,621	468.7 sq mi	8,484/sq mi	34.0194	-118.4108	2.0	Pizza Place	Indian Restaurant	Café	Brazilian Restaurant	Mediterranean Restaurant
2	3	Chicago	Illinois	2,705,994	2,695,598	227.3 sq mi	11,900/sq mi	41.8376	-87.6818	1.0	Fast Food Restaurant	Diner	Coffee Shop	Video Game Store	Intersection
3	4	Houston[3]	Texas	2,325,502	2,100,263	637.5 sq mi	3,613/sq mi	29.7866	-95.3909	2.0	Café	Cajun / Creole Restaurant	Italian Restaurant	Beer Store	Coffee Shop
4	5	Phoenix	Arizona	1,660,272	1,445,632	517.6 sq mi	3,120/sq mi	33.5722	-112.0901	1.0	Fast Food Restaurant	Storage Facility	Video Store	Pharmacy	Sandwich Place

Results

I set the Clusters to 5 to display the map below, as you can see the clusters do not show a significant determination of territory difference. It appears as thought the clusters are evenly distributed thought the US.



Discussion

When looking at the maps of the US, I did not see a major difference in the type of venues when comparing East Coast, West Cost and the Mid-West. We do see a slight difference in the largest cities comparing to the smaller cities, but not enough in my opinion to confidently state there is a difference.

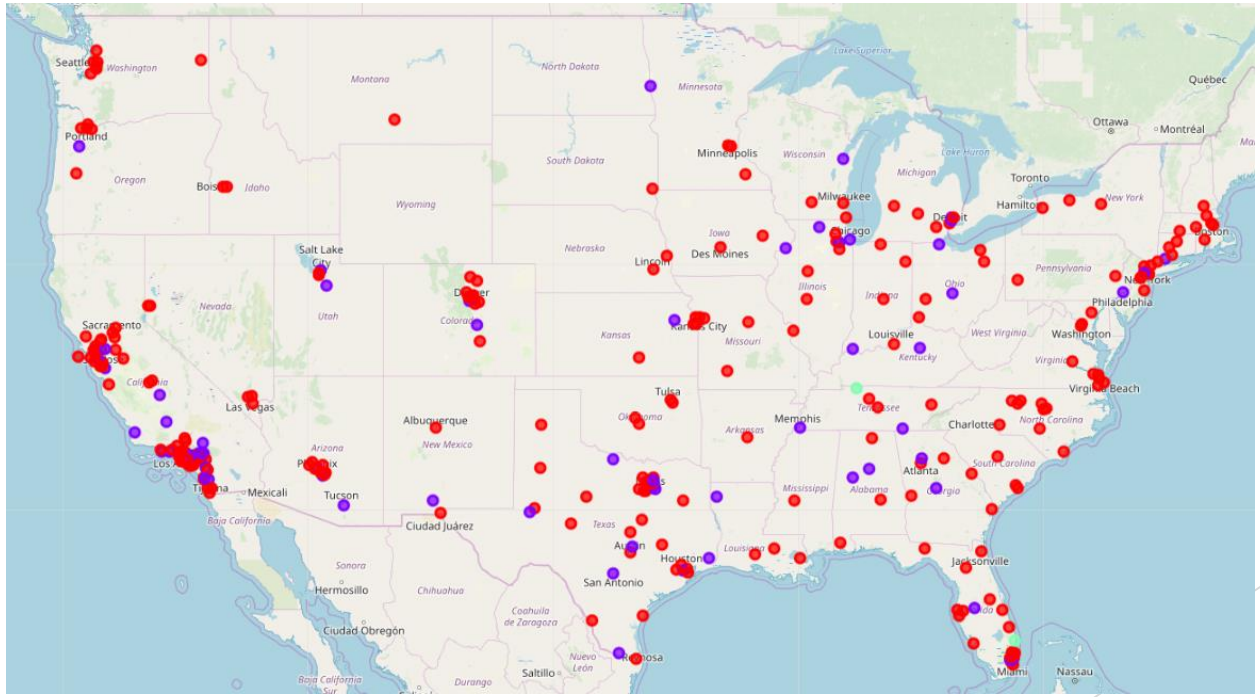
When I had the clusters set to 5, there was a great drop of for clusters 4 and 5.

- Cluster 4 has 2 cities
- Cluster 5 has 1 city

When I had the clusters set to 3, there was a great drop of for clusters 3.

- Cluster 3 has 2 cities

Cluster set to 3



I think trying to compare over 300 cities was not a great project for a 1st Capstone project.
I think comparing under 10 cities would have been more meaningful and easier to analyze.

Conclusion

In this study, I analyzed the largest US cities to see if there is a difference in the venue categories in the country. When I cluster the groups from 5 clusters to 3 clusters. I did not see a large difference in the visual map. It seems the venues in the US are very consistence no matter where you go .
During my travels throughout the US, this conclusion seems correct. When we travel from city to city or state to state in the very large cities in the US, it does seem very similar.

This is my 1st project using Python Pandas, Foursquare and Mapping visualization and felt a learned a great deal about the tools, Next I need to further this project study with more indebt detailed analysis.

As I showed this report to my colleges, some suggestions were made for future studies.
1 -select the top 50 cities around the world
2 – select only the top 10 cities in the US

As I move forward and increase my knowledge with Data Science, I hope to do more projects in this space.