Capstone Project - The Battle of the Neighborhoods

- **Introduction: Business Problem**
  - In this project, I will try to find the patterns of the venues around top universities in America.
  - This report will be targeted to stakeholders interested in investing new facilities (restaurants, gyms, etc.) around top universities.
  - I will try to answer following questions:
    - what kinds of venues are the most popular around those universities,
    - what's the most common patterns of the venues around top universities,
    - which universities' surrounding venues are different from others (for example, have fewer restaurants than the majority).
  - Those outliers (universities with different pattern of venues) might suggest gaps in the local market, which could help those stakeholders to make a wiser business decision.

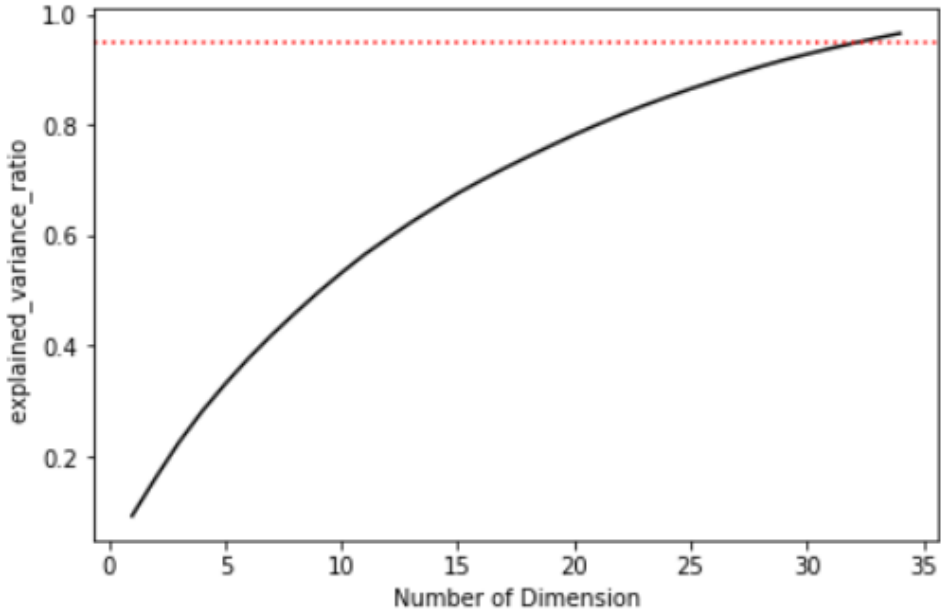
- **Data**
  - Based on the questions, I will need the datasets of:
    - the names of those universities
    - those universities' locations (longitudes and latitudes)
    - the venues around those universities
  - First, I will choose the top 50 universities in American from *The Times Higher Education World University Rankings*, and retrieve their location (longitudes & latitudes) applying ***geocoder*** package.
  - Then, I will extract the venues around them using ***Foursquare API***.

- **Methodology and Results**
  - The universities' locations:

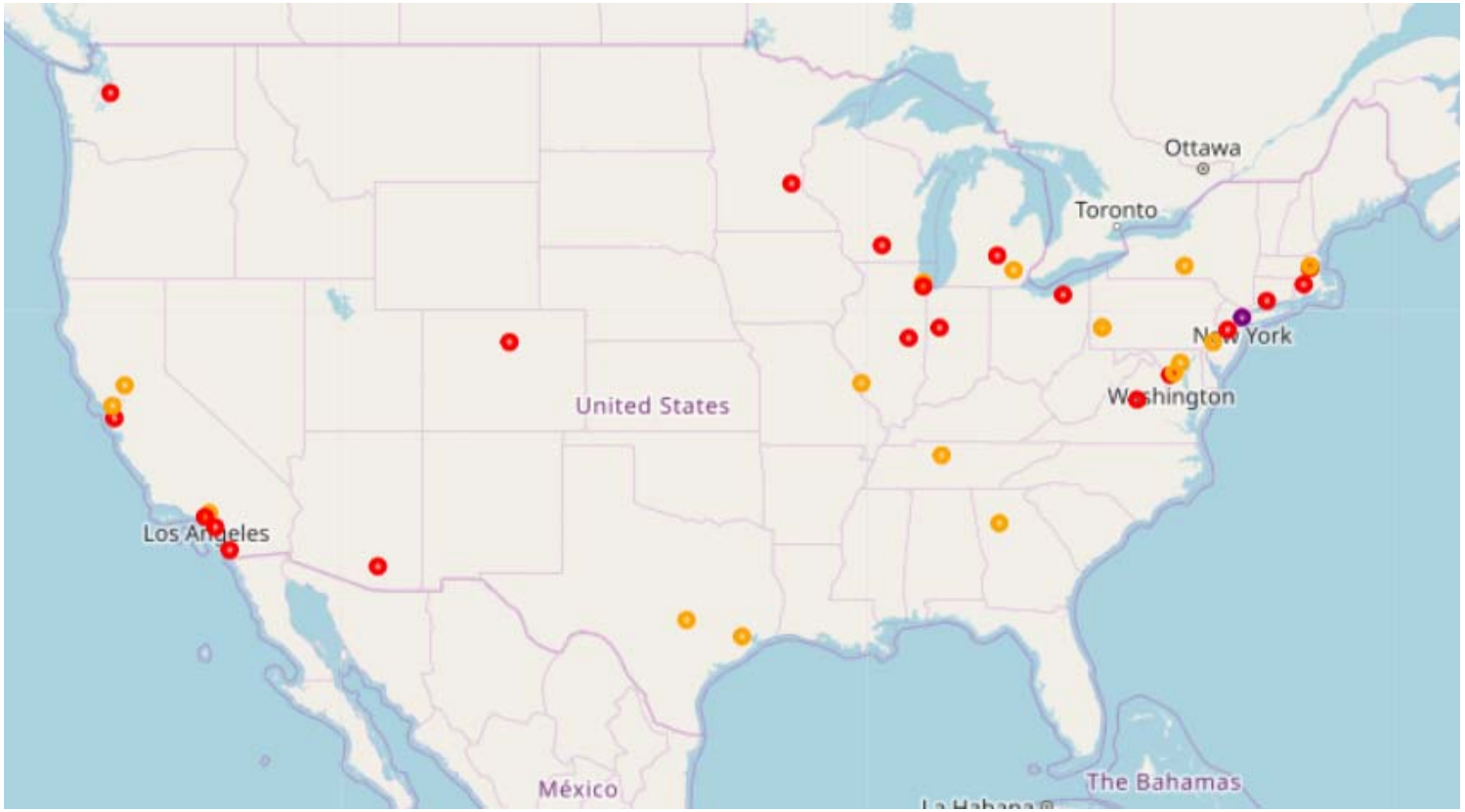
- **Methodology and Results**
  - Some of the universities' surrounding venues (one-hot encoding):

	Name	Accessories Store	African Restaurant	American Restaurant	Amphitheater	Aquarium	Arcade	Arepa Restaurant	Art Gallery	Art Museum	...
0	Boston University	0	0	4	0	0	0	0	0	0	...
1	Brown University	0	0	1	0	0	0	0	1	0	...
2	California Institute of Technology	0	0	3	0	0	0	0	0	0	...
3	Carnegie Mellon University	0	0	1	0	0	0	0	0	1	...
4	Case Western Reserve University	0	0	1	0	0	0	0	1	2	...
5	Columbia University	0	0	3	0	0	0	0	0	0	...

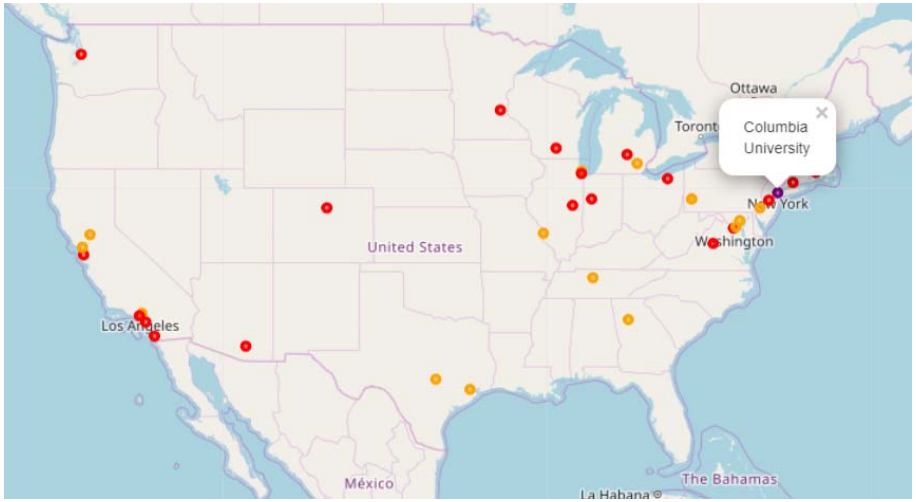
- **Methodology and Results**
  - There are altogether 267 features (different venues' categories).
  - Since many of the categories only occur once or twice, I decide to apply Principal Component Analysis (PCA) to reduce the dimension.
  - By observing the curve, we could reduce the dimension to 30, and still retain 95% variance of the data.

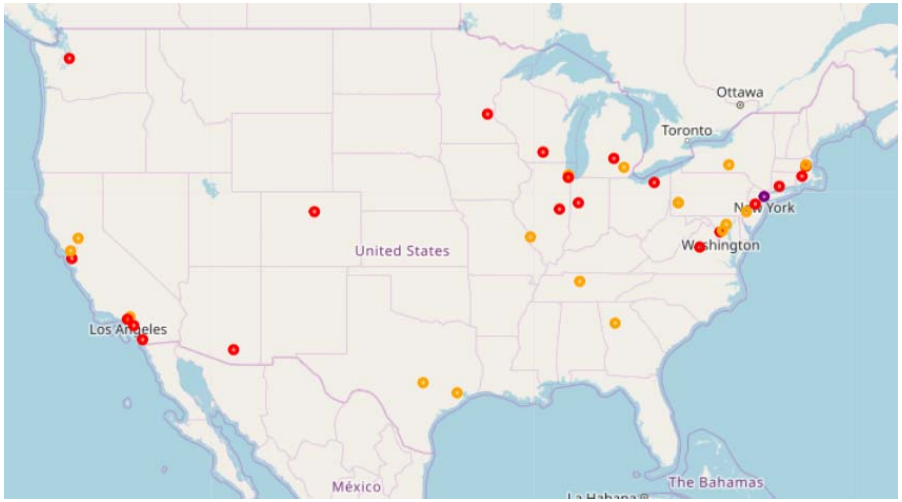


- **Methodology and Results**
  - Now, let's apply K-Means to classify those universities into several groups, by their surrounding venues' categories.



- **Methodology and Results**
  - 20 universities were assigned to group 1, and 19 universities were assigned to group 2, which were very close in value.
  - While the remaining 1 university (*Columbia University*) might have different patterns of venues around it, and seem to be an outlier.
  - By observing the map, we can infer that geographical locations have little to do with the venues' patterns, as different groups are mixed in the map.





- **Methodology and Results**
  - The most popular venues for the universities in group 1 (left) and group 2(right), and the average number:
  - We can see that in both groups, most common venues are about catering, like coffee shop and different kinds of restaurants.

	avg_number
Coffee Shop	4.25
Sandwich Place	2.35
Pizza Place	1.80
American Restaurant	1.50
Café	1.20
Ice Cream Shop	1.05
Bar	0.95
Mexican Restaurant	0.85
Bakery	0.85
Park	0.80
Italian Restaurant	0.75
Burger Joint	0.70
Chinese Restaurant	0.70
Art Museum	0.65
Bagel Shop	0.65

	avg_number
Coffee Shop	2.105263
American Restaurant	1.947368
Pizza Place	1.736842
Café	1.631579
Hotel	1.578947
Indian Restaurant	1.052632
Italian Restaurant	1.000000
Sandwich Place	0.947368
Japanese Restaurant	0.894737
Mexican Restaurant	0.842105
Zoo Exhibit	0.842105
Burger Joint	0.789474
Sushi Restaurant	0.789474
Ice Cream Shop	0.736842
Food Truck	0.736842

- **Methodology and Results**
  - Compare the two main groups:
    - It is mainly the structure of the local catering industry that differ those two main groups.
      - For example, group 1 has more Coffee Shops and Sandwich Places around it, while group 2 has more Japanese Restaurants and Indian Restaurants around it.
    - Apart from the differences in catering industry, group 2 also has more Hotels, Zoo Exhibits and Performing Arts Venues around it.

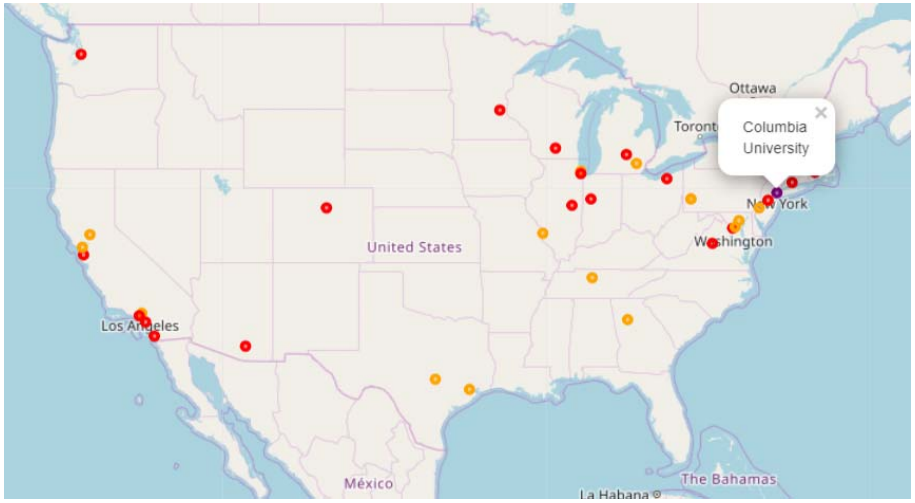
	delta_1_2	abs(delta_1_2)
Coffee Shop	2.144737	2.144737
Sandwich Place	1.402632	1.402632
Hotel	-0.978947	0.978947
Japanese Restaurant	-0.844737	0.844737
Zoo Exhibit	-0.842105	0.842105
Indian Restaurant	-0.752632	0.752632
Sushi Restaurant	-0.639474	0.639474
Food Truck	-0.486842	0.486842
American Restaurant	-0.447368	0.447368
Asian Restaurant	0.442105	0.442105
Café	-0.431579	0.431579
Bar	0.423684	0.423684
Juice Bar	0.392105	0.392105
Bagel Shop	0.386842	0.386842
Performing Arts Venue	-0.378947	0.378947

- **Methodology and Results**
  - Compare the outlier with the two main groups:
    - Group 0 (outlier, Columbia University) has more Parks, Bookstores, Italian Restaurants etc.
    - Besides, compared with group 1, group 0 has fewer Pizza Places and Coffee Shops; compared with group 2, group 0 has fewer Pizza Places, Hotels and India Restaurants.

	delta_0_1
Park	4.20
Bookstore	2.40
Italian Restaurant	2.25
Playground	2.00
Farmers Market	1.80
Pizza Place	-1.80
Food Truck	1.75
Seafood Restaurant	1.70
Grocery Store	1.65
American Restaurant	1.50
Burger Joint	1.30
Coffee Shop	-1.25
Whisky Bar	1.00
Dog Run	1.00
Historic Site	1.00

	delta_0_2
Park	4.526316
Bookstore	2.263158
Italian Restaurant	2.000000
Farmers Market	1.894737
Playground	1.842105
Seafood Restaurant	1.842105
Pizza Place	-1.736842
Grocery Store	1.631579
Hotel	-1.578947
Food Truck	1.263158
Burger Joint	1.210526
Sandwich Place	1.052632
American Restaurant	1.052632
Indian Restaurant	-1.052632
Dog Run	1.000000

- **Methodology and Results**
  - So, considering the potential gaps in the market, I suggest the stakeholders invest new Pizza Places around Columbia University .



	delta_0_1
Park	4.20
Bookstore	2.40
Italian Restaurant	2.25
Playground	2.00
Farmers Market	1.80
Pizza Place	-1.80
Food Truck	1.75
Seafood Restaurant	1.70
Grocery Store	1.65
American Restaurant	1.50
Burger Joint	1.30
Coffee Shop	-1.25
Whisky Bar	1.00
Dog Run	1.00
Historic Site	1.00

	delta_0_2
Park	4.526316
Bookstore	2.263158
Italian Restaurant	2.000000
Farmers Market	1.894737
Playground	1.842105
Seafood Restaurant	1.842105
Pizza Place	-1.736842
Grocery Store	1.631579
Hotel	-1.578947
Food Truck	1.263158
Burger Joint	1.210526
Sandwich Place	1.052632
American Restaurant	1.052632
Indian Restaurant	-1.052632
Dog Run	1.000000

- **Discussion**
  - Based on the analysis and results, I suggest the stakeholders invest new **Pizza Places** around *Columbia University*. However, there are several potential limitations in the analysis:
    - I did not consider the population density around those universities. Higher density implies that there should be more venues to fulfill their need.
    - I did not consider the local people' nationalities. For example, if there are many Chinese students in a certain university, to invest a Chinese restaurant nearby could be a wise choice.
    - I couldn't find a way to create proper weight factors to different venues. Since many of the venues are in the catering industry, other venues' importance could be shadowed. If you have any suggestions or comments, please feel free to leave me a note.

- **Conclusion**
  - There are generally two different patterns of venues around top universities in America.
  - The two patterns are mainly differed by the structure of the local catering industry.
  - Columbia University has different venues' pattern. To invest new Pizza Places around them might fulfill the potential gaps in local market, thus helping the stakeholders to make profits and develop connection with excellent university students and faculties.