CLUSTERING THE 50 MOST POPULATED CITIES IN THE US

IBM Capstone Final Project

INTRODUCTION

- Foursquare provides location-based data that can be used for a variety of purposes. Two examples of this data is a list of "Top Pick" venues for a given geographical location, and a list of "Trending" venues for a given geographical location and at a specific time (i.e. when the API is called).
- This project focuses on classifying the most populous US cities, based on the type of venues that are listed as "top picks," as well as venues that are "trending" on a Sunday afternoon during the summer season.

INTEREST

- CATEGORIZATION OF CITIES BASED ON VENUES LISTED AS "TOP PICKS":
  - Business owners looking to expand into other large US cities would be interested to see which locations have people with similar interests.
  - Leisure travelers could be interested in exploring similar cities based on their travel preferences.
- CATEGORIZATION OF CITIES BASED ON TRENDING VENUES AT A GIVEN TIME:
  - This type of analysis is useful for businesses to identify peak times and off-peak times to address stock and staffing concerns.
  - This information could also be used by customers to determine personally-ideal times for visiting particular types of venues.
  - Note: This project was run at one specific time – Sunday afternoon during the summer season – but the analysis could be run for all types of times for a more detailed view.

DATA SOURCES

- **Most Populated Cities (Wikipedia)**
  - The list of most populated cities, including names and geographical coordinates, was scraped from Wikipedia using the Beautiful Soup Python library.
- **Top Pick Venues (Foursquare)**
  - The venues listed as "top picks" for each city was extracted using the Foursquare API. I extracted 100 top venues for each city, along with the venues' respective categories.
- **Trending Venues on a summer Sunday afternoon (Foursquare)**
  - Trending venues for each city was extracted using the Foursquare API. I extracted 100 trending venues for each city on a Sunday afternoon during the summer, along with the venues' respective categories.

DATA PREPARATION AND CLEANING

- **Most Populated Cities (Wikipedia)**
  - I scraped the table from Wikipedia as stated above, and removed all columns except for Rank, City, State, Population (2018 Estimate), and Location.
  - I then separated the Latitude and Longitude from the Location column to enable easy parsing from the Foursquare API. A subset of this data is shown below:

2018 Rank	City	State	2018 Estimate	Latitude	Longitude
1	New York	New York	8,398,748	40.6635	-73.9387
2	Los Angeles	California	3,990,456	34.0194	-118.4108
3	Chicago	Illinois	2,705,994	41.8376	-87.6818
4	Houston	Texas	2,325,502	29.7866	-95.3909
5	Phoenix	Arizona	1,660,272	33.5722	-112.0901

DATA PREPARATION AND CLEANING

- **Top Pick Venues (Foursquare)**
  - Using the latitude and longitude for each city, I then called the Foursquare "explore" destination endpoint to get a list of 100 top picked venues for each city.
  - A subset of this data is shown below:

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	New York	40.6635	-73.9387	Mekelburg's	40.687571	-73.962370	Gourmet Shop
1	New York	40.6635	-73.9387	BAM Rose Cinemas	40.686338	-73.977438	Indie Movie Theater
2	New York	40.6635	-73.9387	Los Mariscos	40.742000	-74.005890	Seafood Restaurant
3	New York	40.6635	-73.9387	Carton Brewing	40.411746	-74.038158	Brewery
4	New York	40.6635	-73.9387	Los Tacos No. 1	40.757134	-73.987536	Taco Place

DATA PREPARATION AND CLEANING

- **Trending Venues on a summer Sunday afternoon (Foursquare)**
  - Using the latitude and longitude for each city, I then called the Foursquare "explore" destination endpoint to get a list of 100 trending venues for each city.
  - A subset of this data is shown below:

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	New York	40.6635	-73.9387	Trader Joe's	40.725611	-74.004985	Grocery Store
1	New York	40.6635	-73.9387	Prospect Park (Nethermead)	40.660717	-73.968587	Field
2	New York	40.6635	-73.9387	City Swiggers	40.777515	-73.950820	Beer Store
3	New York	40.6635	-73.9387	Gotham Archery	40.682504	-73.986032	Athletics & Sports
4	New York	40.6635	-73.9387	SoulCycle Brooklyn Heights	40.692253	-73.991042	Cycle Studio

METHODOLOGY: ONE HOT ENCODING

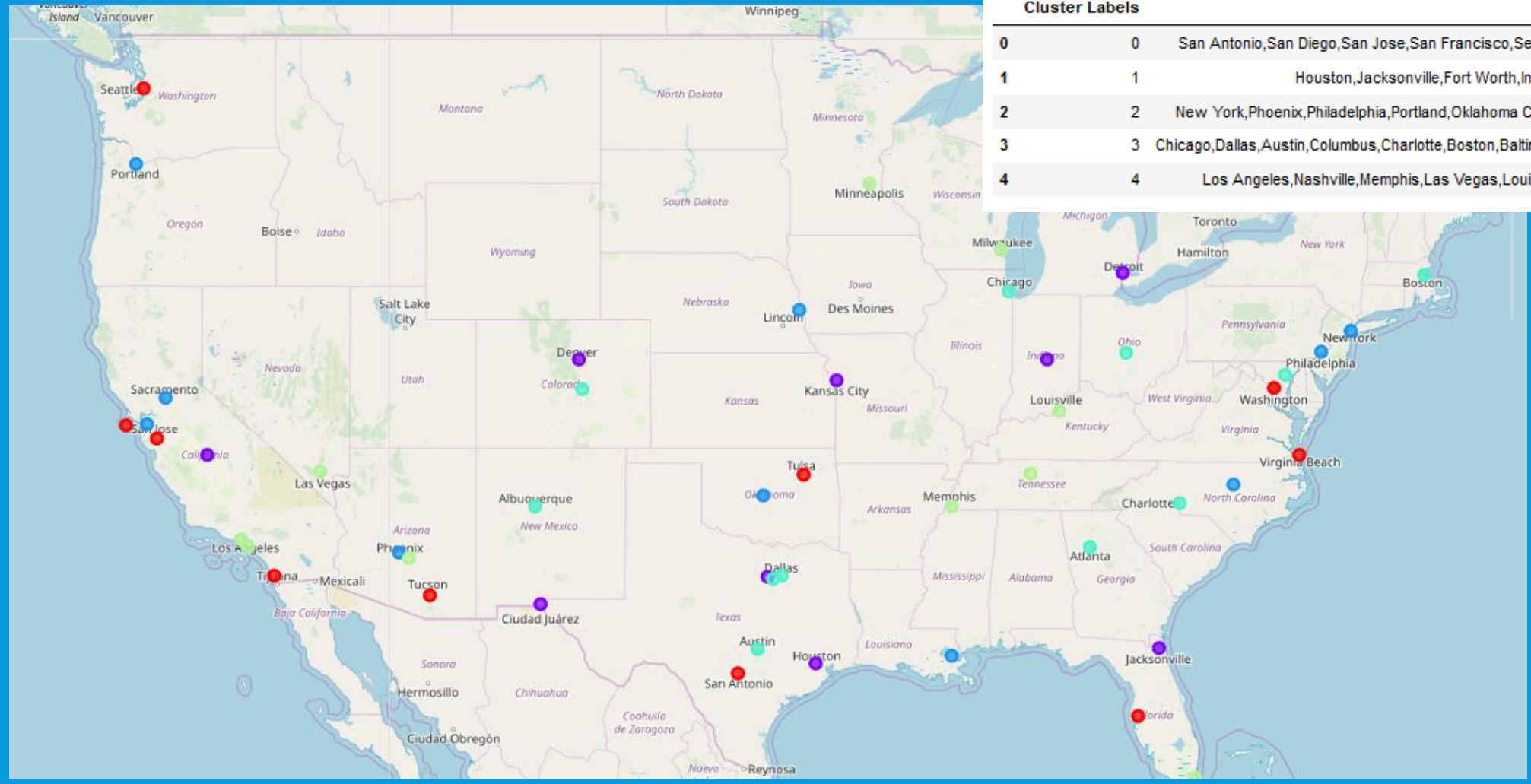
- Using One-Hot Encoding, I sorted the categories of most common top pick venues and most common trending venues for each city. A subset of this data is shown below:

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Albuquerque	Brewery	Pizza Place	Mexican Restaurant	Movie Theater	Science Museum
1	Arlington	Brewery	Coffee Shop	American Restaurant	Seafood Restaurant	Trail
2	Atlanta	Trail	Park	Sandwich Place	American Restaurant	Coffee Shop
3	Austin	Coffee Shop	Taco Place	Ice Cream Shop	Movie Theater	Pizza Place
4	Baltimore	Seafood Restaurant	Park	BBQ Joint	Ice Cream Shop	Pizza Place

METHODOLOGY: K MEANS CLUSTERING

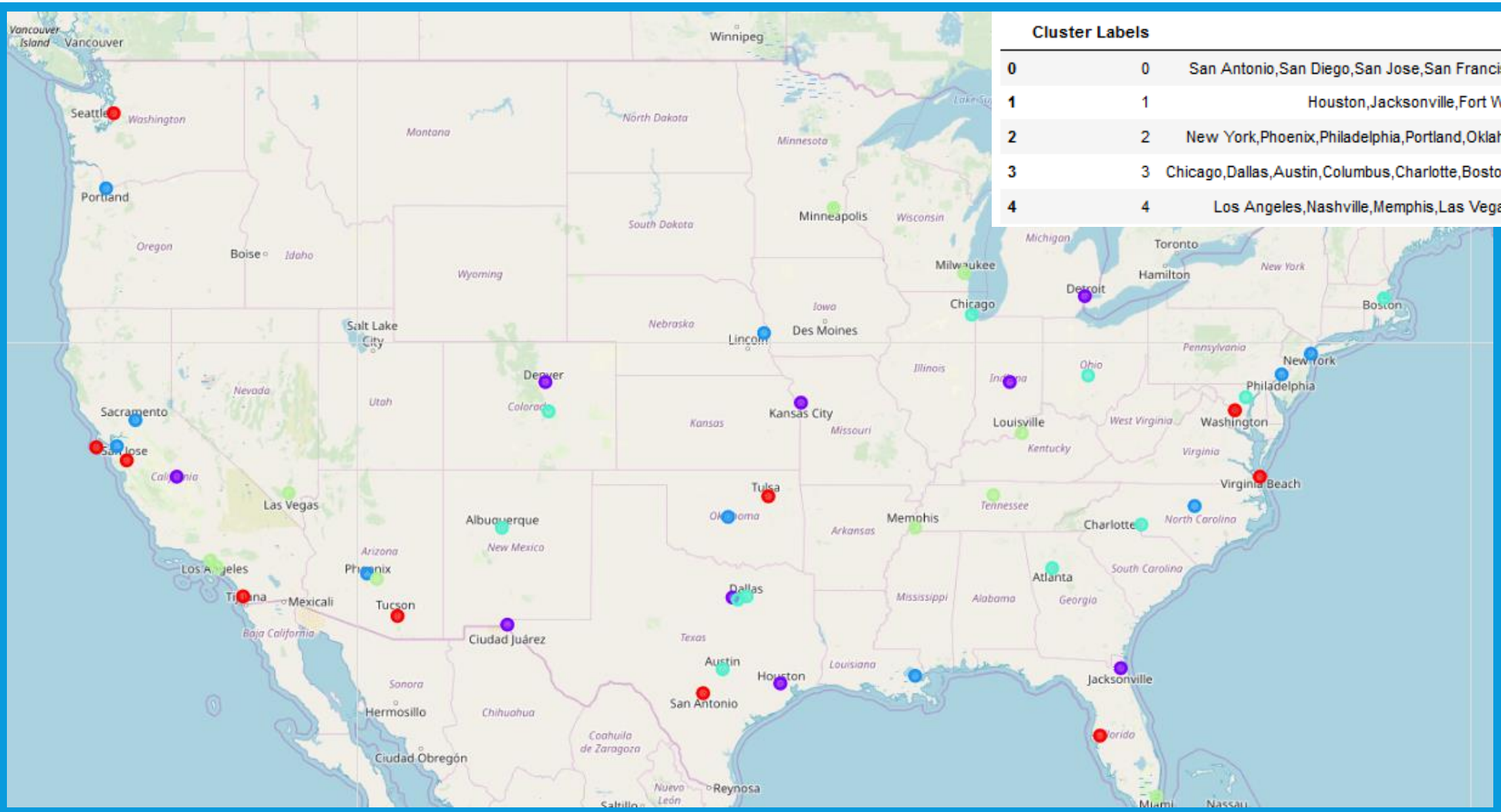
- I then used K Means Clustering to cluster cities based on the most common top pick venues and most common trending venues, and generated two different sets of clusters.

RESULTS: CITY CLUSTERS BASED ON TOP PICK VENUES



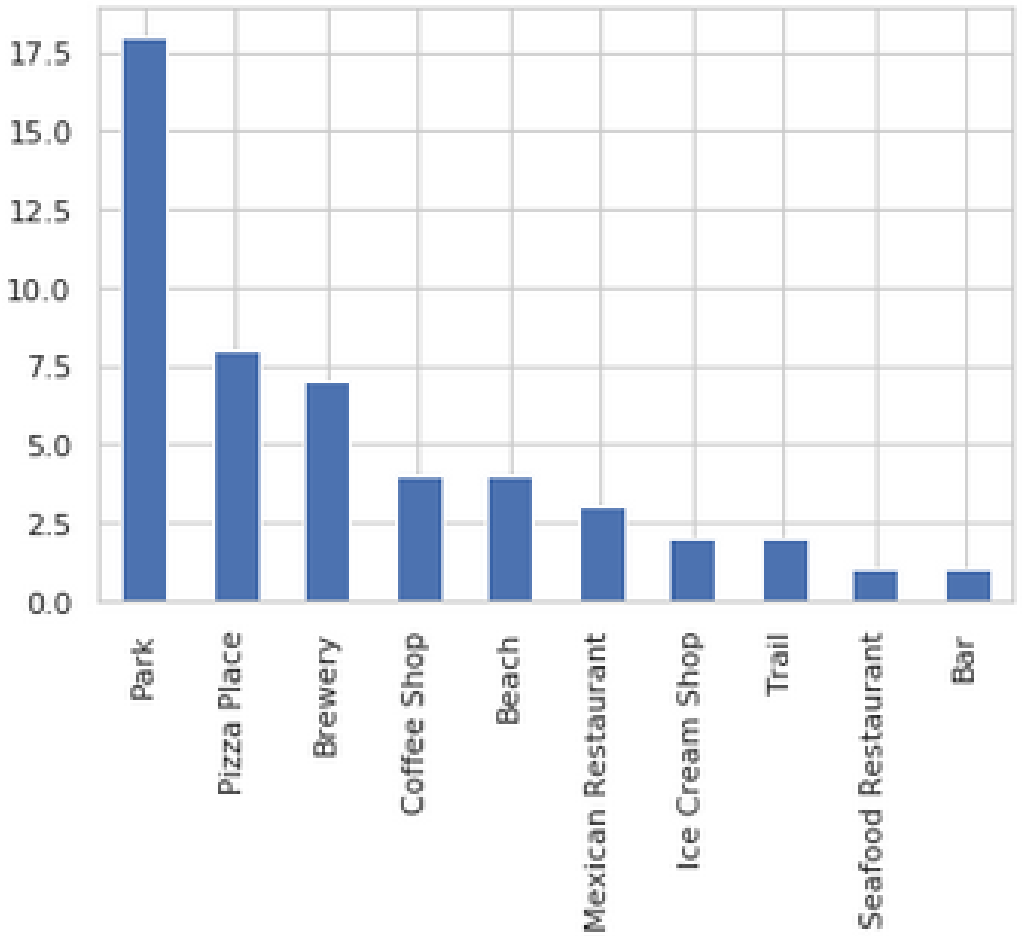
	Cluster Labels	City
0	0	San Antonio,San Diego,San Jose,San Francisco,Seattle,Washington,Tucson,Virginia Beach,Tulsa,Tampa
1	1	Houston,Jacksonville,Fort Worth,Indianapolis,Denver,El Paso,Detroit,Fresno,Kansas City
2	2	New York,Phoenix,Philadelphia,Portland,Oklahoma City,Sacramento,Raleigh,Omaha,Oakland,New Orleans
3	3	Chicago,Dallas,Austin,Columbus,Charlotte,Boston,Baltimore,Albuquerque,Atlanta,Colorado Springs,Arlington
4	4	Los Angeles,Nashville,Memphis,Las Vegas,Louisville,Milwaukee,Mesa,Miami,Long Beach,Minneapolis

RESULTS: CITY CLUSTERS BASED ON TRENDING VENUES ON SUNDAY AFTERNOON IN SUMMER



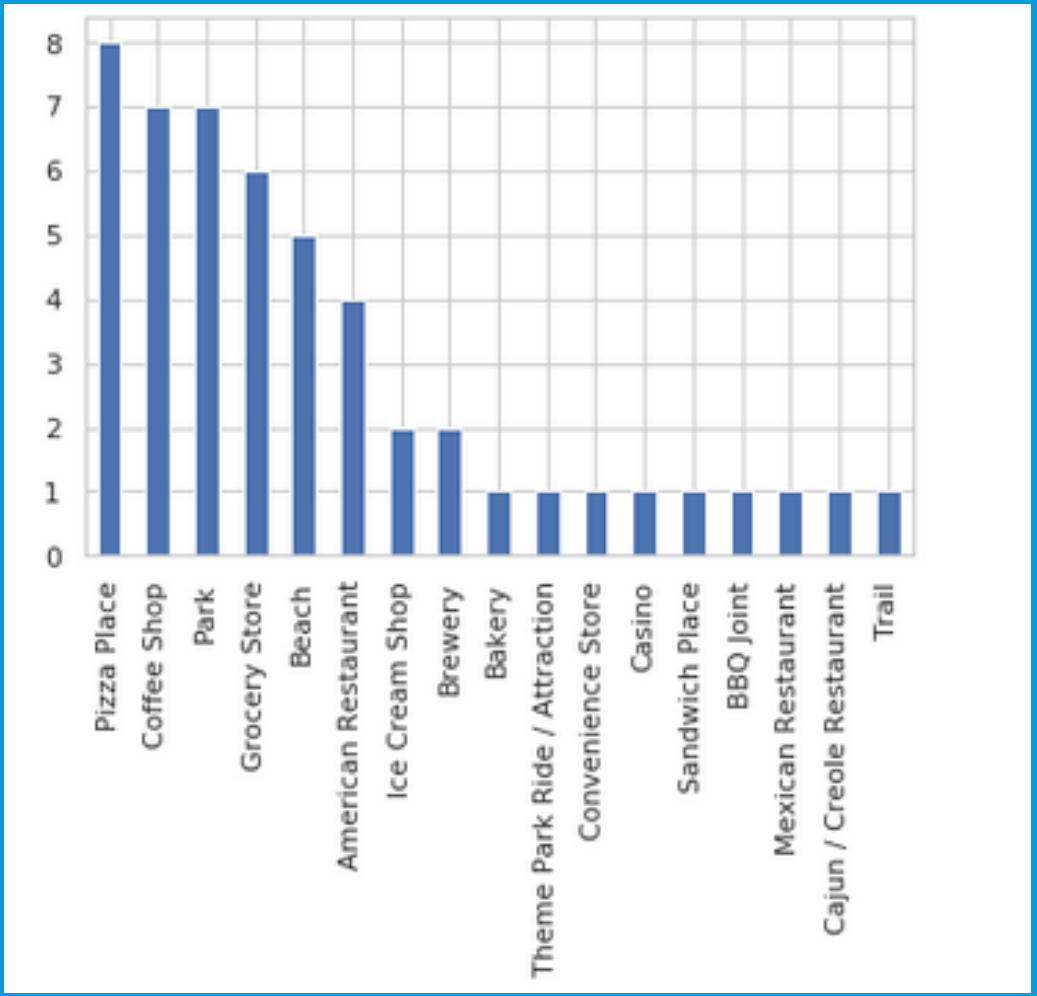
	Cluster Labels	City
0	0	San Antonio,San Diego,San Jose,San Francisco,Seattle,Washington,Tucson,Virginia Beach,Tulsa,Tampa
1	1	Houston,Jacksonville,Fort Worth,Indianapolis,Denver,El Paso,Detroit,Fresno,Kansas City
2	2	New York,Phoenix,Philadelphia,Portland,Oklahoma City,Sacramento,Raleigh,Omaha,Oakland,New Orleans
3	3	Chicago,Dallas,Austin,Columbus,Charlotte,Boston,Baltimore,Albuquerque,Atlanta,Colorado Springs,Arlington
4	4	Los Angeles,Nashville,Memphis,Las Vegas,Louisville,Milwaukee,Mesa,Miami,Long Beach,Minneapolis

RESULTS: MOST COMMONLY OCCURRING TOP PICK VENUE CATEGORY



- People overwhelmingly love parks contained in big cities, as shown here.
- In 18 out of 50 cities, the top picked venue was a park.
- Pizza places and breweries followed next, while coffee shops and beaches were also popular.

RESULTS: MOST COMMONLY OCCURRING TRENDING VENUE CATEGORY



- During the summer, people spend their Sunday afternoons primarily at pizza places, or getting coffee, or outdoors at the park and the beach.
- For six cities, grocery stores were popular on Sunday afternoons, reinforcing that some people run errands during this time.

DISCUSSION

- Several recommendations can be drawn from the data here:
  - Parks are overwhelming favorite spots for people in many cities, which means that businesses would do well to locate themselves near parks or sell their wares at parks if city regulations allow.
  - Pizza places tend to be very popular in the US, especially on Sunday afternoons.
  - The clustering of cities also shows interesting information. As a leisure traveler, for example, if I enjoy visiting the venues in San Antonio, then I will likely also enjoy traveling to San Francisco or Tucson, but perhaps not Los Angeles. Likewise, if I enjoy Los Angeles, then I may also enjoy traveling to Miami, but perhaps not Denver.
- Another interesting result was that even though the types of venues trending on Sunday afternoons was slightly different from the overall top picks for each city, the clustering of the cities based on both sets of data was ***exactly the same.***