Comparing the Rising Tech City Hubs: Comparing and Investing Similarity to San Francisco

IBM Data Science Specialization

Applied Data Science Course: Capstone Project

Pedro Junior Vicente Valdez

December 31, 2019

Introduction

People have different preferences when looking for a place/state/city to live in. Many factors will play into the final decision including location, weather, social environment, population, etc. In the STEM field, many cities are known as hubs for research and development. These Tech Hubs are usually overpopulated and facing issues like excessive increases to cost of living (rent, food, etc.). A good example is San Francisco. San Francisco faces a unique housing challenge. Due to the lack of regulation of business needs by the city, housing needs cannot be match. Due to the increasing employment, new commercial or office buildings are required which comes at the cost of residential housing. There are other theories but the facts are the same.

San Francisco has been, and still is, one of the biggest tech cities in the United States but recently its growth has caused many issues in terms of cost and infrastructure. The growth in population is causing traffic jams throughout the city and nearby bridges, apartment renting prices are increasing to unsustainable levels, and the cost of living is getting too high for the average worker. Because of these and additional factors, many of the tech industry workers are applying for jobs elsewhere. Many of these persons looking for similar environments to San Francisco. In this project, I investigate which of the seven major rising tech hub cities (Seattle, New York, Colorado, Austin, Los Angeles, Chicago, and Boston) is most similar to San Francisco.

These tech hubs were selected based on a recent articles ranking the top tech cities in the United States (https://builtin.com/tech-hubs). This project is mostly targeted to STEM professionals looking to move out of San Francisco due to the current housing crisis but wants to maintain a similar lifestyle/neighborhood.

Data Sources

To answer this question, we will need information regarding the most popular recommended venues which may include coffee places, restaurants, malls, and parks. For this we will use the new Foursquare API endpoint, venues/explore (https://developer.foursquare.com/docs/api/venues/explore). The idea is that the top recommended venues in San Francisco will be used to build a profile using several characteristics like the venue category. Similarly, profiles will be created for the previously mentioned cities. All this information can be gathered using regular calls from the Foursquare API. In order to increase the profile details, premium calls will be made to get the venues ratings. The ratings will provide a "weight" that gives adequate importance to the gathered venues. It serves as a measure of the quality of the venue but more specifically of it's category.

For the purpose of this project, only 100 venues for each city will be called. This is due to the constraints (limits) of premium calls. Once the data is gathered, it will be processed to create a purely numerical dataset which will be used as input to a clustering (the K-means algorithm from Scikit-Learn).

Methodology

section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

For this project a variety of python packages were used:

- *pandas* is an "open source library providing high-performance, easy-to-use data structures and data analysis tools".
- *NumPy* is "the fundamental package for scientific computing with Python. It contains among other things: a powerful N-dimensional array object, sophisticated (broadcasting) functions, tools for integrating C/C++ and Fortran code, useful linear algebra, Fourier transform, and random number capabilities. NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of database."
- *Matplotlib* is a "Python 2D plotting library which produces publication quality figures".
- *Scikit-Learn* is an open source "simple and efficient tools for predictive data analysis built on NumPy, SciPy, and matplotlib
- *Folium* makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map.
- *Geopy* makes it easy for Python developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources.

At first, Geopy was used to automatically get the longitude and latitude information for each city. An example of this call can be seen in Figure 1. Later on, a function to collect the location information for an arbitrary number of cities (Table A).

```
# Specifying the address
address = 'Downtown, San Francisco'

# Gathering the location coordinates
location_sf = geolocator.geocode(address)

# Placing coordinates to variables
latitude_sf = location_sf.latitude
longitude_sf = location_sf.longitude
print('The geograpical coordinate of Downtown, San Francisco are {}, {}.'.format(latitude_sf, longitude_sf))
```

```
The geograpical coordinate of Downtown, San Francisco are 37.7875138, -122.407159.
```

	Cities	Latitude	Longitude
0	San Francisco, California	37.787514	-122.407159
0	Denver, Colorado	39.751770	-105.013873
0	Seattle, Washington	47.604872	-122.333458
0	Los Angeles, California	34.498713	-118.584307
0	New York City, NY	40.599756	-73.946390
0	Chicago, Illinois	41.893648	-87.621960
0	Austin, Texas	30.268054	-97.744764
0	Boston, Massachusetts	42.362918	-71.068737

The response from the Foursquare Places API gives a specified number of venue recommendations given the CLIENT_ID, CLIENT_SECRET, and Foursquare Version. Details about the recommended venues include the address, its location (latitude and longitudes), and the venue's category which can range from gourmet shop, coffee place, all the way to schools and restaurants. The response is of JSON type which can be processed using the *json_*normalize call in the pandas module. This allows to convert responses into data frames. An example of a response for one venue can be seen in the foursquare's developers website. The JSON response contains more information that is not useful for the purposes of this project. Limiting the response to only the needed fields results in the following data frame.

	venue.name	venue.id	venue.location.formattedAddress	venue.categories	venue.location.lat	venue.location.lng
0	Maison Margiela	551cfcaf498e23f2c0115449	[134 Maiden Ln, San Francisco, CA 94108, Unite...	[{'id': '4bf58dd8d48988d104951735', 'name': 'B...	37.788261	-122.405765
1	Saint Laurent	528d4fe211d2543b7663f4fd	[108 Geary St, San Francisco, CA 94108, United...	[{'id': '4bf58dd8d48988d104951735', 'name': 'B...	37.787774	-122.405412
2	Williams-Sonoma	4aa45625f964a5207b4620e3	[340 Post St (btwn Powell & Stockton), San Fra...	[{'id': '58daa1558bbb0b01f18ec1b4', 'name': 'K...	37.788377	-122.407446
3	Tiffany & Co.	4a791992f964a520efe61fe3	[350 Post St (btwn Powell & Stockton), San Fra...	[{'id': '4bf58dd8d48988d111951735', 'name': 'J...	37.788598	-122.407708
4	UNIQLO	50043438e4b0f448ea4f447f	[111 Powell St, San Francisco, CA 94102, Unite...	[{'id': '4bf58dd8d48988d103951735', 'name': 'C...	37.785850	-122.408041

The *venue id* was then used as input to subsequently make premium calls to obtain the venue's rating. A total of 100 venue recommendation per city were gathered due to the limits and constraints of Foursquare's premium calls.

```
venue_id = "551cfcaf498e23f2c0115449"

url = "https://api.foursquare.com/v2/venues/{}".format(venue_id)

params = dict(
    client_id=CLIENT_ID,
    client_secret=CLIENT_SECRET,
    v='20191129')

resp2 = requests.get(url=url, params=params)
data2 = json.loads(resp2.text)
```

```
venue_details = json_normalize(data2)
venue_details[["response.venue.name", "response.venue.rating"]]
```

	response.venue.name	response.venue.rating
0	Maison Margiela	9.2

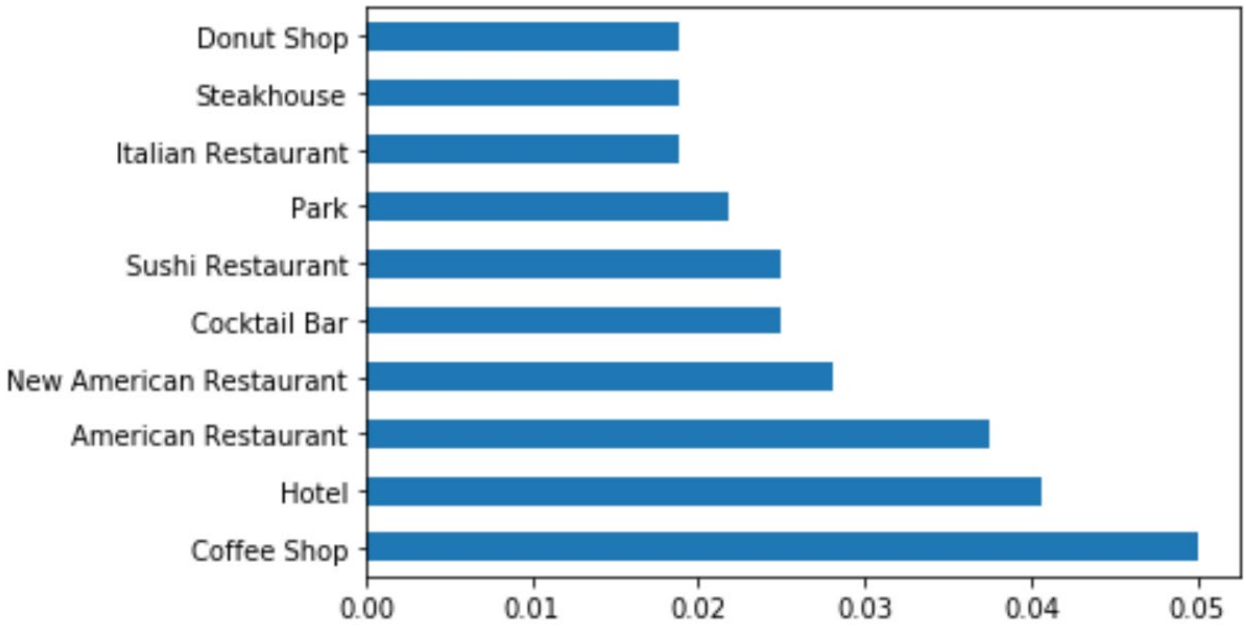
After calling the API, the dataset was cleaned of unnecessary features. A snippet of the head of the data frame can be seen in the next figure. A similar data frame for each city was created.

	City	City Latitude	City Longitude	Venue	id	Venue Latitude	Venue Longitude	Venue Category	Venue Rating
0	San Francisco	37.787514	-122.407159	Maison Margiela	551cfcaf498e23f2c0115449	37.788261	-122.405765	Boutique	9.2
1	San Francisco	37.787514	-122.407159	Saint Laurent	528d4fe211d2543b7663f4fd	37.787774	-122.405412	Boutique	9.2
2	San Francisco	37.787514	-122.407159	Williams-Sonoma	4aa45625f964a5207b4620e3	37.788377	-122.407446	Kitchen Supply Store	8.9
3	San Francisco	37.787514	-122.407159	Tiffany & Co.	4a791992f964a520efe61fe3	37.788598	-122.407708	Jewelry Store	8.9
4	San Francisco	37.787514	-122.407159	The Archive	4b4bd8caf964a5207ba926e3	37.789494	-122.405766	Men's Store	9.3

The features/columns include:

- City
- City Latitude and Longitude:
- Venue Name
- Venue ID
- Venue Latitude and Longitude
- Venue Category
- Venue Rating

The most recommended venues will depend on popularity among the citizens of each city. Seattle maybe more of a coffee place than San Francisco while the latter may have higher density of boutiques or shopping malls. The venue category will have an important role on the final result along with the venue rating (see next Figure).



We can see that the most popular category in the entire data frame are coffee shops followed by hotels, restaurants, bars going all the way to parks, steakhouses and donut shops. A list of the top 10 most common categories can be seen in the following table.

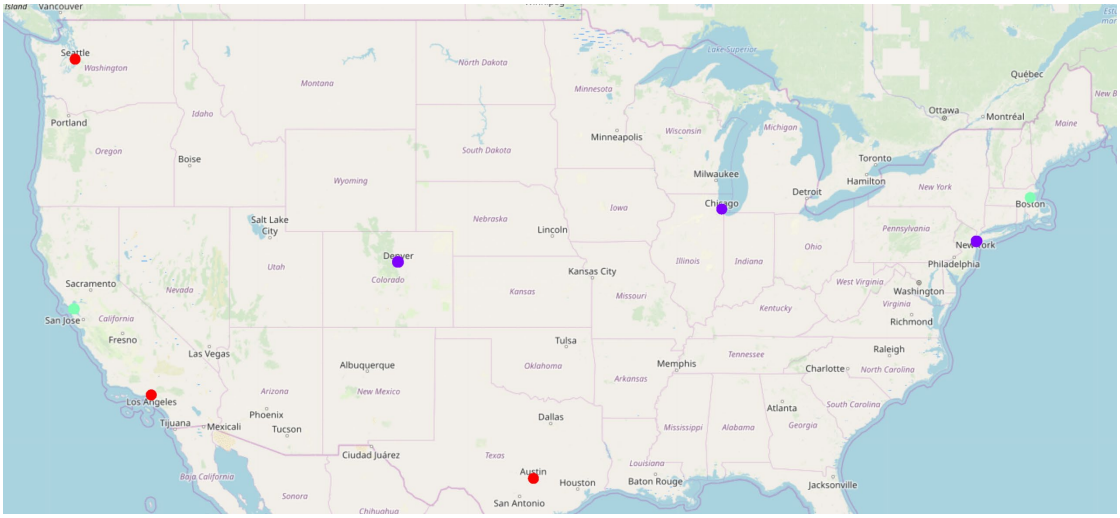
	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Austin	Hotel	Cocktail Bar	Lounge	Steakhouse	Coffee Shop	Burger Joint	Park	Cajun / Creole Restaurant	Chinese Restaurant	Salad Place
1	Boston	Coffee Shop	New American Restaurant	Gym / Fitness Center	Steakhouse	Sandwich Place	Salad Place	Gastropub	Hotel	Falafel Restaurant	Restaurant
2	Chicago	American Restaurant	New American Restaurant	Donut Shop	Grocery Store	Restaurant	Cosmetics Shop	Yoga Studio	Café	Salon / Barbershop	Resort
3	Colorado	Theme Park Ride / Attraction	Coffee Shop	Park	Yoga Studio	Ice Cream Shop	Café	Sushi Restaurant	Brewery	Pizza Place	Seafood Restaurant
4	Los Angeles	Plaza	Ice Cream Shop	Speakeasy	Coffee Shop	Theater	Jazz Club	Park	School	Candy Store	Historic Site
5	New York	Pizza Place	Sushi Restaurant	Bakery	Italian Restaurant	Food & Drink Shop	Bagel Shop	Deli / Bodega	Farmers Market	Mexican Restaurant	Bubble Tea Shop
6	San Francisco	Boutique	Hotel	Men's Store	Bubble Tea Shop	Clothing Store	Plaza	Gym / Fitness Center	Shoe Store	Music Venue	Food Truck
7	Seattle	Hotel	Coffee Shop	Cocktail Bar	Concert Hall	Café	Donut Shop	Deli / Bodega	Seafood Restaurant	Scenic Lookout	Sandwich Place

Next, we proceeded to normalize all features since algorithms like K-means are deeply influenced by the magnitude of all numbers. For this both a MinMaxScaler and the StandardScaler was implemented to see the influences in the final results.

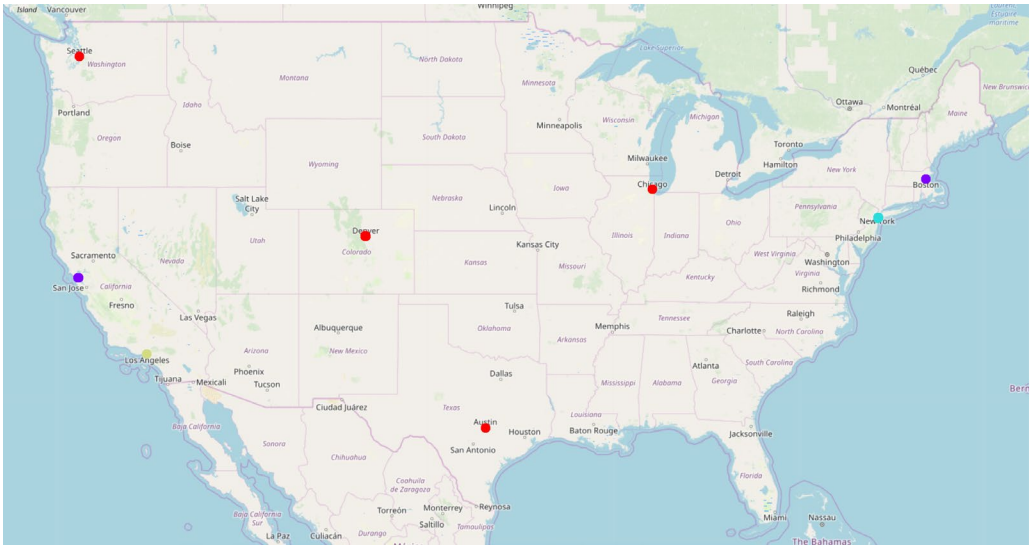
For the algorithm, the Scikit-Learn K-Means packages was implemented. 3 clusters where arbitrarily specified to have a good distribution of classes. The clustering algorithm was implemented in dataframe 1 (normalize using Standard Scaler) and dataframe 2 (normalize using Min Max). After fitting the algorithm the results where added back to the original dataframe and visualize using a Folium map.

Result and Discussion

Based purely on location closeness, it is expected that Los Angeles or Seattle be amongst the most similar cities to San Francisco. It was found that, solely based on the venues, its categories and the rating, the most common city to San Francisco is Boston, both of which belong to cluster 1. Cluster 2 includes Seattle, Los Angeles and Austin while Cluster 3 includes Denver, Chicago, and New York. The similarity of Boston to San Francisco can be explained by it's highly rated hotels, gyms/fitness centers and coffee shops. While the number of venues/datapoints used to fit the model is not enough, it is still a valid answer for the purposes of this class. A more accurate portrayal of each city would involve getting a higher number of venue recommendations. To go one step further, a picture for each venue could be gathered followed by fitting a convolutional neural networks to add the visual dimensions to the problem. This way, the model could be aided by comparing which places look most alike.



To corroborate the results, the number of clusters was increased to 4. The final results is still the same. San Francisco and Boston belong to Cluster 1. With 4 clusters, Seattle and Los Angeles were placed in different clusters.



Conclusion

Our hypothesis purely based on distance was wrong (expectedly). We observe that Boston is the most similar city to San Francisco based on the information gathered. It is important to know the limitations of the current results. This result is purely the result of the clustering algorithm fitted to a limited number of recommended venues (limited by the number of premium calls that allowed us to get the venue's ratings). This is not representative of an entire city but for the purposes of this project it will suffice.

From the qualitative analysis we can see that both, Boston and San Francisco, have a high venue density of Boutiques, Stores and Hotels followed by Tea Places and Gyms. Other clusters include Parks as one of the most common categories. Parks are not even in the top 10 most common venues of San Francisco cluster #1. This is in contrast to other cities like Denver, Chicago, and New York which has Parks as the 3rd most common venue. As mentioned previously, a larger dataset is needed to accurately build a profile for each city. Architecture has a big impact on similarity and therefore the venues picture can be used by a convolutional neural network to classify which city is most similar to one another based on the cities design.