1 Introduction

1.1 Background

Predicting what new venues will be popular and where they will be popular is difficult to do and could benefit from analysis of existing data. Accurate prediction of these trends could yield higher returns on business investments with lower risks. New York City (NYC) is a trendsetting city. Businesses popular there often become popular in other US cities later.

1.2 Problem

What is the best way to determine what business to open and where? Cities that are more similar to NYC are more likely to adopt trends from NYC. So, identifying what cities are similar to NYC is important. Identifying venues that are popular in NYC but aren't yet popular in similar cities is also important.

1.3 Interest

Entrepreneurs interested in the entire US market would benefit from a proper analysis of the similarities between US cities and what venues are and are not currently popular there.

2 Data

2.1 Data acquisition

US city populations as well as latitudes and longitudes were scraped from Wikipedia. Number of types of venues, locations, and popularity (likes) were scraped from Foursquare.

2.2 Data cleaning

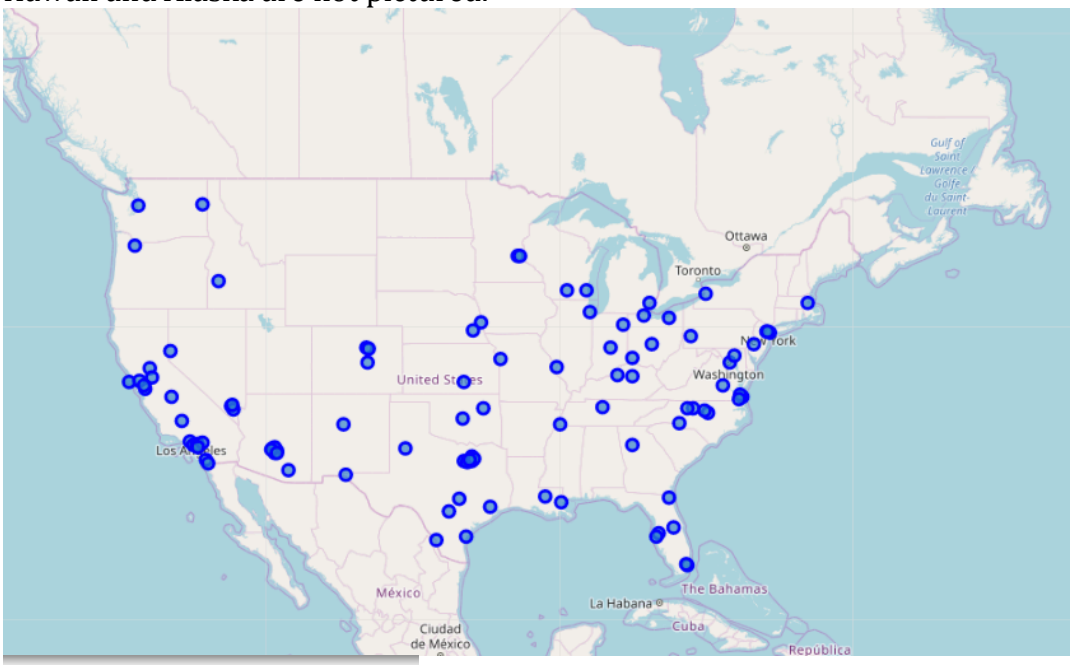
Latitudes and longitudes from the Wikipedia data were in a format not appropriate for folium or Foursquare. It used direction keys (eg W for West) instead of positive or negative values. The direction keys had to be dropped and values need to be properly assigned their positive or negative values accordingly.

One hundred venues were selected from the center of each city within a 50 km radius. This number of venues was chosen because higher numbers ran over the Foursquare API limit available. Obviously, there are more than 100 venues in these areas and a larger limit would allow us to analyze more venues and likely produce more accurate results.

3 Methodology

3.1 Most populated US cities

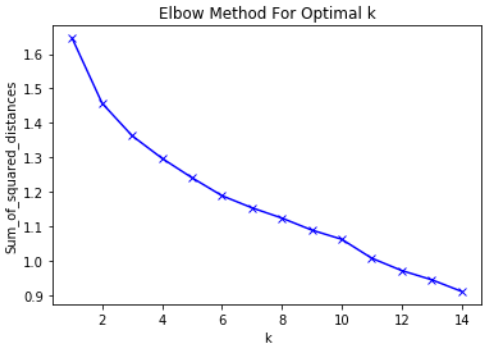
Highly populated cities are geographically diverse and are located across the US. Hawaii and Alaska are not pictured.

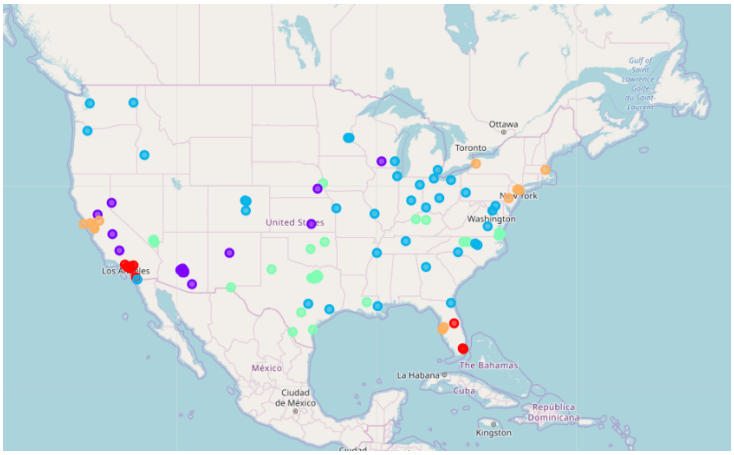


City	Latitude	Longitude
New York[d]	40.6635	-73.9387
Los Angeles	34.0194	-118.4108
Chicago	41.8376	-87.6818
Houston[3]	29.7866	-95.3909
Phoenix	33.5722	-112.0901
Philadelphia[e]	40.0094	-75.1333
San Antonio	29.4724	-98.5251
San Diego	32.8153	-117.1350
Dallas	32.7933	-96.7665
San Jose	37.2967	-121.8189

3.2 Clustering cities based on popular venues

The elbow method for determining the optimal number of clusters to use did not produce an elbow but a gentle downward curve. We chose to use 5 clusters because this will give us clusters with enough cities to draw some conclusions. We were focused on finding cities similar to NYC so our cluster of interest is orange. It contains mostly coastal cities including NYC and San Francisco. Interestingly, the other clusters also appear to share some geographic similarities (eg the light purple cluster is mostly found in the south).

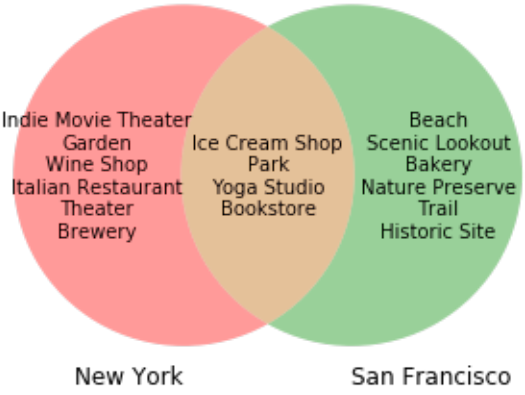




3.3 Cities similar to NYC do not have many of the venues popular in NYC

Based on the lists of popular venues, it seems the NYC cluster mainly shares parks, trails, and beaches as common venues while the other popular venues are different. We chose to further analyze San Francisco (SF) and found that SF and NYC both are popular for parks, yoga studios, ice cream stores, and bookstores. SF was not popular for many venues popular in NYC, including Italian restaurants. We chose Italian restaurants as the new business to open.

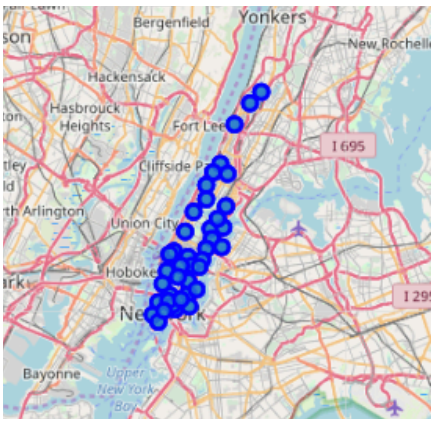
City	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
New York[d]	40.6635	-73.9387	4	Park	Bookstore	Italian Restaurant	Yoga Studio	Theater	Wine Shop	Brewery	Indie Movie Theater	Ice Cream Shop	Garden
Philadelphia[e]	40.0094	-75.1333	4	Park	Brewery	Coffee Shop	Deli / Bodega	Café	Concert Hall	Pizza Place	Trail	Art Museum	Gym
San Jose	37.2967	-121.8189	4	Park	Trail	Beach	State / Provincial Park	Coffee Shop	Bakery	Yoga Studio	Climbing Gym	Wine Shop	Movie Theater
San Francisco[g]	37.7272	-123.0322	4	Park	Beach	Scenic Lookout	Trail	Bookstore	Bakery	Historic Site	Yoga Studio	Ice Cream Shop	Nature Preserve
Boston	42.3320	-71.0202	4	Park	Bakery	Brewery	Seafood Restaurant	Sandwich Place	Indie Movie Theater	Wine Shop	Gym	Mexican Restaurant	Hotel
Oakland	37.7698	-122.2257	4	Park	Yoga Studio	Coffee Shop	Trail	Brewery	Art Museum	Wine Shop	Bakery	Movie Theater	Café
Tampa	27.9701	-82.4797	4	Park	Scenic Lookout	Pizza Place	Convenience Store	American Restaurant	Theme Park Ride / Attraction	Ice Cream Shop	New American Restaurant	Wine Shop	Beer Garden
Stockton	37.9763	-121.3133	4	Park	Trail	Music Venue	Brewery	Coffee Shop	Grocery Store	Burger Joint	Bakery	Scenic Lookout	Café
Newark	40.7242	-74.1726	4	Park	Theater	Bookstore	Bakery	Italian Restaurant	Gym	Cycle Studio	Art Gallery	Ice Cream Shop	Taco Place
Jersey City	40.7114	-74.0648	4	Park	Bookstore	Italian Restaurant	Theater	Cycle Studio	Ice Cream Shop	Gym	Art Gallery	Taco Place	Bakery
St. Petersburg	27.7620	-82.6441	4	Beach	Park	Coffee Shop	Mexican Restaurant	American Restaurant	Scenic Lookout	Pizza Place	Ice Cream Shop	Brewery	Sandwich Place
Buffalo	42.8925	-78.8597	4	Park	Brewery	Café	Italian Restaurant	Scenic Lookout	Waterfall	Bakery	Coffee Shop	Performing Arts Venue	Ice Cream Shop
Fremont	37.4945	-121.9412	4	Park	Trail	Beach	State / Provincial Park	Brewery	Café	Movie Theater	Coffee Shop	Bakery	Climbing Gym



4 Results

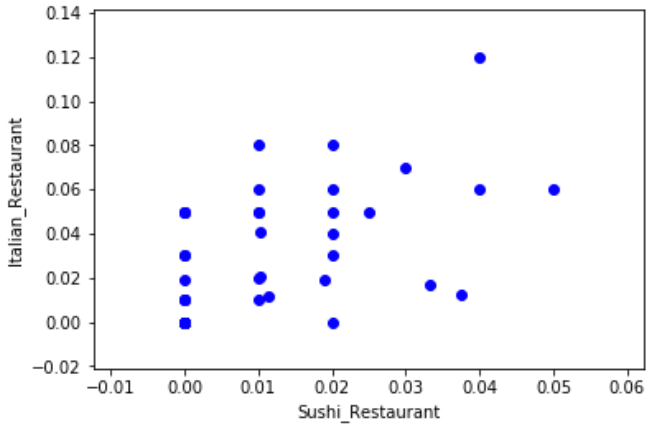
4.1 Italian restaurants are found most often near sushi restaurants relative to other venues in NYC

We identified all Italian restaurants and their locations in NYC. We then generated a list of all the venues within 500 m of those restaurants. Using pearson correlation, we calculated the correlation of different venues in the same vicinity as Italian restaurants. Sushi restaurants were most highly correlated with Italian restaurants in NYC.



Map of Italian restaurants in NYC

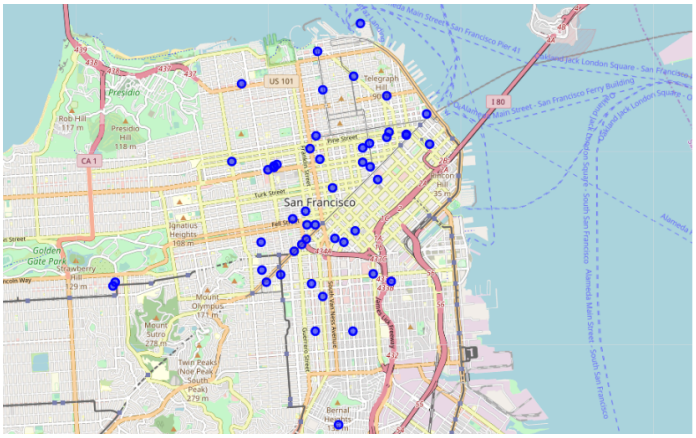
	Pearson Correlation with Italian Restaurant
Italian Restaurant	1.000000
Sushi Restaurant	0.524317
Lebanese Restaurant	0.498653
Gaming Cafe	0.498653
Eastern European Restaurant	0.498653
Udon Restaurant	0.471621
Gourmet Shop	0.453924
Creperie	0.451432
Electronics Store	0.375364
French Restaurant	0.357963



Frequency of sushi restaurants relative to Italian restaurants in NYC

4.2 SF neighborhoods with the most popular sushi restaurants

We found the locations of all the sushi restaurants in SF and categorized them based on their neighborhoods (postal code). We then obtained the number of "likes" for each restaurant. All the sushi restaurant likes were added together for each neighborhood. We found that postal code 94102 had the most "liked" sushi restaurants.

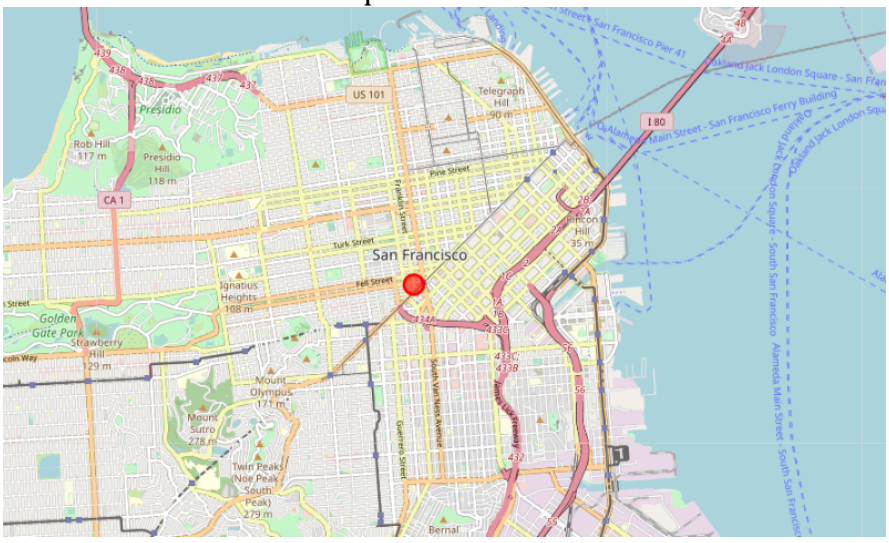


postal_code	likes
94102	945
94103	577
94104	3
94105	14
94108	499
94109	482
94110	148
94111	7
94114	336
94115	126
94117	37
94122	70
94123	78
94128	25
94133	97

Locations of sushi restaurants in SF and the number of likes for sushi restaurants by postal code

5 Discussion

Using our approach we determined that the SF neighborhood with postal code 94102 is the best area to open a new Italian restaurant.



This approach found that US cities that are geographically similar will also cluster together when grouped by the most popular venues in those cities. NYC and SF are in a cluster with mostly coastal cities and rather large cities. This approach generated many cities that cluster with NYC and also many venues that are popular in NYC that are not yet popular in other cities in the same cluster. We chose Italian restaurants and SF but many other possibilities are available.

6 Conclusions

Our analysis would benefit from a higher limit on Foursquare inquiries. We believe we would be able to more optimally choose the number clusters for a k-means analysis. Additionally, with the proper computational resources this analysis could be performed on every city worldwide with sufficient Foursquare data. Another dimension that could be added to this approach is to add analysis

of trending venues to determine the time of day certain venues are most visited. For example, if our approach found a coffee shop was a good business to start, one might want to correlate where coffee shops are found most often with venues that popular in the morning.