Capstone Project - The Battle of Neighborhoods:

Week 5 Assignment → Final Report

Title: Comparing Neighborhoods in LA to SF an NY

1- Problem Description and Background Discussion

A producer of deli meat located in California supplies various types of cured meat to restaurants, supermarkets and sandwich shops located in the Los Angeles metropolitan area. His business is thriving and is considering expanding to another populous city with a similar profile of business distribution and people preferences. Since his business model takes advantage of the appetite for cured meats in Mexican and Italian restaurants and would benefit from expanding to a city with a similar distribution of venues.

To help this producer decide between San Francisco and New York City, we will carry out an analysis of the type of venues and their distribution in each city and determine which of the two shares more similarity with Los Angeles, hence presenting a better chance of success.

We will first collect data on the neighborhoods of all three cities, including name, zip code and geographical coordinates. Then we will extract information on the venues located within each neighborhood. A clustering of the neighborhoods based on the type of venues they contain will follow, together with analysis of the types of venues and plots of the cluster distribution on the map. Finally, we will estimate the similarity between cities to provide a recommendation.

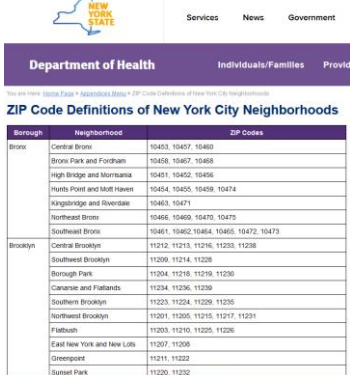## 2- Data Description and Application for Solving the Problem

a) Web data scraping:

To obtain geographical information from each city, one web site was selected in each case that included a table with at least zip-codes and neighborhood names for that city.

The data was scraped using the library **requests** to grab html data and the library **BeautifulSoup** to scrape html data.

The data for each city was wrangled separately according to the characteristics of the URL source and converted to a dataframe with the columns 'Neighborhood' and 'Zip Code'.

Table 1: Data Source for Neighborhood and Zip code information

Los Angeles	San Francisco	New York City
http://www.laalmanac.com/communications/cm02_communities.php	web_link=http://www.healthysf.org/bdi/outcomes/zipmap.htm	web_link=https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm

b) Coordinate extraction:

The coordinates of each neighborhood with distinct zip code was extracted using the function ***Nominatim*** from the library ***geopy.geocoders***, which converts an address into latitude and longitude values. For simplicity and to reduce ambiguity given that some zip codes included more than one neighborhood name, the zip code value was used instead to obtain the latitude and longitude values. Whenever a zip code did not return valid latitude or longitude coordinates, the entire row was deleted. The data was then added to the dataframes as columns 'Latitude' and 'Longitude', respectively.

To understand better the spatial distribution of neighborhoods in each city and their geographical distribution, we use ***Folium*** library and plot the center of each neighborhood (as defined by the latitude and longitude coordinates extracted in Table 2) over the map of each city. The resulting maps, plotted using the same scale, are displayed in table 3 below.

Table 2: Geographical coordinates for each relevant neighborhood/Zip code

Los Angeles	San Francisco	New York City
Number of neighborhoods in LA: 86	Number of neighborhoods in SF: 21	Number of neighborhoods in NY: 42

Los Angeles

	Neighborhood	Zip Code	Latitude	Longitude
0	Arlington Heights (Los Angeles)	90019	34.047371	-118.336046
1	Artesia, Artesia (PO Boxes)	90701	33.868528	-118.077698
2	Athens	90044	33.981914	-118.287489
3	Atwater Village (Los Angeles)	90039	34.118121	-118.264129
4	Avalon (PO Boxes)	90704	33.341730	-118.328136

San Francisco

	Neighborhood	Zip Code	Latitude	Longitude
0	Hayes Valley/Tenderloin/North of Market	94102	37.779491	-122.418224
1	South of Market	94103	37.774425	-122.411091
2	Potrero Hill	94107	37.793634	-122.408295
3	Chinatown	94108	37.791043	-122.406578
4	Polk/Russian Hill (Nob Hill)	94109	37.793815	-122.420597

New York City

	Neighborhood	Zip Code	Latitude	Longitude
0	Central Bronx	10453	40.852348	-73.911965
1	Bronx Park and Fordham	10458	40.861569	-73.888765
2	High Bridge and Morrisania	10451	40.828381	-73.927084
3	Hunts Point and Mott Haven	10454	40.807728	-73.918198
4	Kingsbridge and Riverdale	10463	40.884718	-73.887248

Table 3: Geographical distribution of neighborhoods per city

Los Angeles	San Francisco	New York City
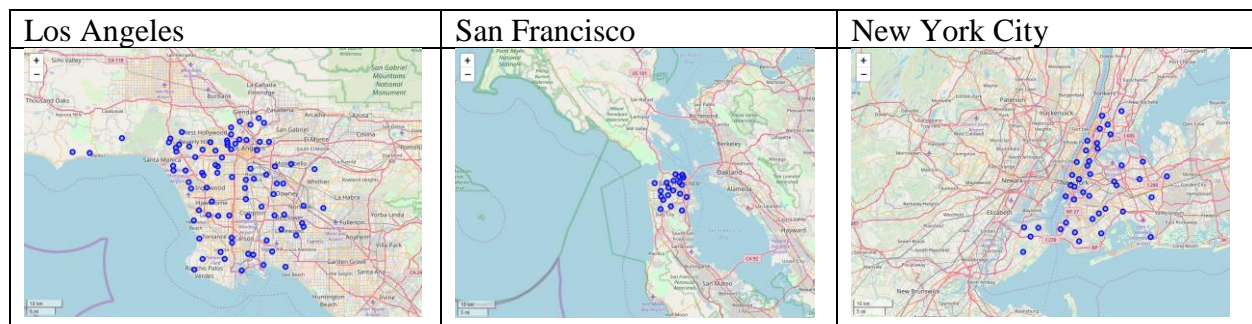
From the results in Table 3, it becomes evident that the geographical extent of the cities is not equal, and the area covered by the neighborhoods in Los Angeles is much larger than that covered by the neighborhoods in San Francisco. Keeping this result in mind, a different radius will be used when sampling each neighborhood for venues in the next section.

c) Venue sampling and data extraction:

With the coordinates of each neighborhood of each city collected into each corresponding dataframe as shown in Table 2, the **FourSquare API** that we set up in week 1 of the course was employed to sample each city for venues and extract coordinates for common venues within a pre-defined radius around the center of each neighborhood. As discussed in the previous section, given that the center of the neighborhoods in Los Angeles are farther apart and they can extend larger areas than those in San Francisco or New York, the search radius was adjusted slightly for each city so that each neighborhood included approximately the same amount of potentially meat-serving venues. In particular, a radius of 850 meters was selected for Los Angeles while 600 meters was selected for San Francisco and 550 meters for New York.

The current settings of the FourSquare API only allow to extract a maximum of 100 venues per neighborhood, which combined with the limited spatial extent of the search means that the results provide a representative sample of the venue distribution in each city and our conclusions will need to be extrapolated to the entire venue population.

Re-using code from previous weeks assignments, only relevant information about each venue returned was collected into each of the three separate dataframes, one per each city. These dataframes are used in subsequent steps to analyze the neighborhood information and extract insights. A representative example of the dataframes can be seen in Table 4.

Table 4: Dataframes with venue samples for each city

Los Angeles

	Venue	Venue Latitude	Venue Longitude	Venue Category
0	PizzaRev	34.048585	-118.336439	Pizza Place
1	Smart & Final Extra!	34.047692	-118.335932	Grocery Store
2	Planet Fitness	34.047774	-118.338605	Gym / Fitness Center
3	Jersey Mike's Subs	34.048449	-118.337419	Sandwich Place
4	PetSmart	34.048184	-118.335489	Pet Store
5	La Fayette Square	34.043205	-118.333813	Neighborhood
6	Midtown Crossing	34.048047	-118.337077	Shopping Mall
7	Mateo's Ice Cream & Fruit Bars	34.047588	-118.327972	Ice Cream Shop
8	El Compita	34.048592	-118.332846	Mexican Restaurant
9	Panda Express Mid-City	34.048654	-118.337556	Chinese Restaurant

San Francisco

	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Asian Art Museum	37.780178	-122.416505	Art Museum
1	Louise M. Davies Symphony Hall	37.777976	-122.420157	Concert Hall
2	Herbst Theater	37.779548	-122.420953	Concert Hall
3	Philz Coffee	37.781433	-122.417073	Coffee Shop
4	War Memorial Opera House	37.778601	-122.420816	Opera House
5	San Francisco Ballet	37.778580	-122.420798	Dance Studio
6	Ananda Fuara	37.777693	-122.416353	Vegetarian / Vegan Restaurant
7	Siam Orchid Traditional Thai Massage	37.777111	-122.417967	Massage Studio
8	August 1 Five	37.780537	-122.420188	Indian Restaurant
9	War Memorial Court	37.779042	-122.420971	Park

New York City

	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Liberato	40.853744	-73.907966	Latin American Restaurant
1	Accra Resturant	40.853871	-73.908421	African Restaurant
2	Wingstop	40.854093	-73.907899	Wings Joint
3	Bravo Supermarkets	40.853936	-73.914144	Grocery Store
4	Papa John's Pizza	40.852429	-73.908976	Pizza Place
5	Dunkin Donuts	40.853817	-73.908724	Donut Shop
6	Chase Bank	40.854087	-73.907631	Bank
7	Food Dynasty	40.853772	-73.909267	Supermarket
8	Subway	40.853887	-73.907285	Sandwich Place
9	Chase Bank	40.850381	-73.916217	Bank

3- Data Analysis Methodology

We want to compare the venue profile between LA and either NY or SF in order to produce a recommendation for expanding the business of our meat producer client. The analysis will have the following components:

a) Data understanding and preparation
   In this section we try to gain insight into the representative set of venues returned for each city by finding the most common venues, computing percentages and plotting the distribution of the most common venue types.
   In a second step, only venues that potentially serve or sell meat are analyzed in a similar fashion. Finally the geographical distribution of the neighborhoods with the most potentially meat-serving venues is analyzed graphically.

b) Modeling
   In this section we group the meat-serving venues into a few clusters by applying the k-means algorithm on the set. We then analyze the top 5 meat-serving venues for each cluster to evaluate similarities.

c) Comparison between cities
   In this final section, the geographical distribution of clusters is displayed on the map by coloring the neighborhoods according to the cluster they belong. In addition, Euclidean distances are computed between the venues in each of the cities as well as those in the two most populous clusters.
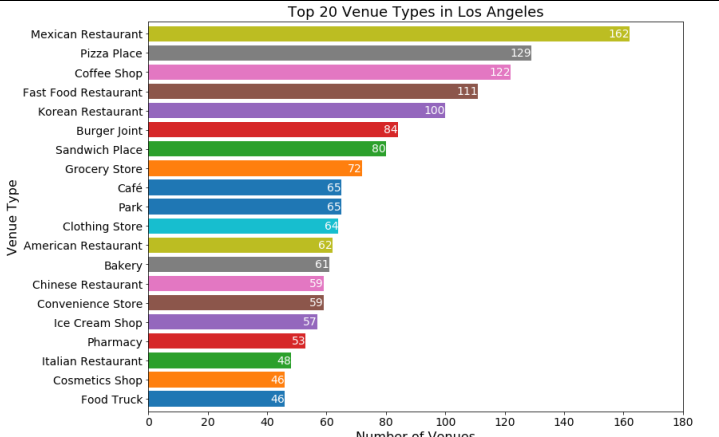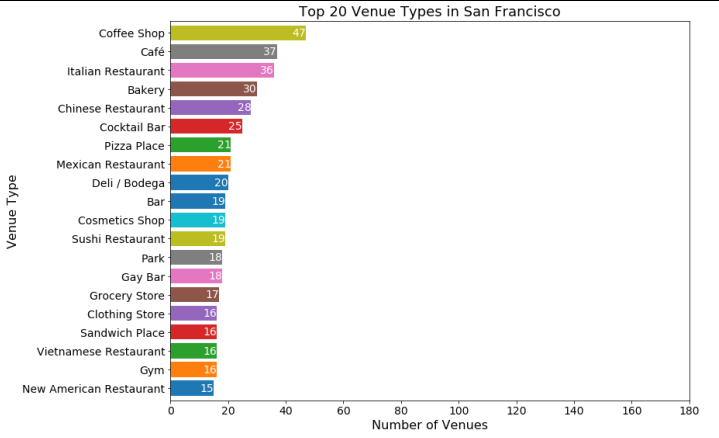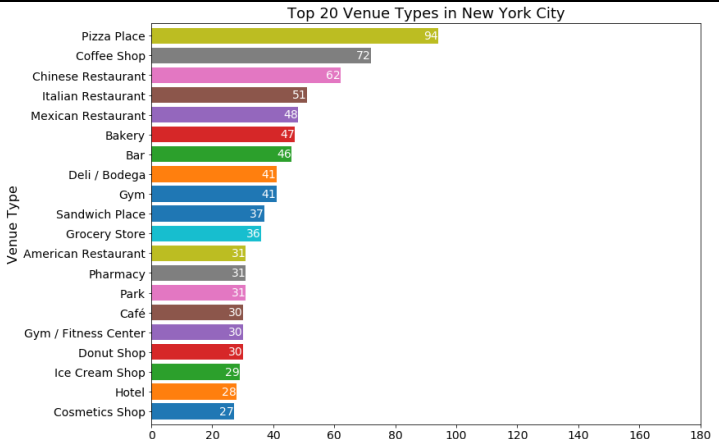
4- Results

a) Data Understanding and Preparation

Venues:

After collecting a representative sample of venues for all neighborhoods in each city, we start analyzing each of the sets by first grouping them by venue type, understanding their distribution and what are the top most common venue types in absolute and percentage terms. Table 5 below displays the results for all three cities, combining some general statistics of the venue types and horizontal bar plots of their distributions.

Table 5: Analysis of venue type distribution per city

Los Angeles Venues	
Top 20 Venue Types in Los Angeles chart showing: Mexican Restaurant 162, Pizza Place 129, Coffee Shop 122, Fast Food Restaurant 111, Korean Restaurant 100, Burger Joint 84, Sandwich Place 80, Grocery Store 72, Café 65, Park 65, Clothing Store 64, American Restaurant 62, Bakery 61, Chinese Restaurant 59, Convenience Store 59, Ice Cream Shop 57, Pharmacy 53, Italian Restaurant 48, Cosmetics Shop 46, Food Truck 46	Number of venues in LA: 3677 Avg. number of venues per neighborhood: 42.76 Number of venue types in LA: 321 The five most common venues are: 1- Mexican Restaurant ... 162 (4.41%) 2- Pizza Place ... 129 (3.51%) 3- Coffee Shop ... 122 (3.32%) 4- Fast Food Restaurant ... 111 (3.02%) 5- Korean Restaurant ... 100 (2.72%)

San Francisco Venues	
Top 20 Venue Types in San Francisco chart showing: Coffee Shop 47, Café 37, Italian Restaurant 36, Bakery 30, Chinese Restaurant 28, Cocktail Bar 25, Pizza Place 21, Mexican Restaurant 21, Deli / Bodega 20, Bar 19, Cosmetics Shop 19, Sushi Restaurant 19, Park 18, Gay Bar 18, Grocery Store 17, Clothing Store 16, Sandwich Place 16, Vietnamese Restaurant 16, Gym 16, New American Restaurant 15	Number of venues in SF: 1250 Avg. number of venues per neighborhood: 59.52 Number of venue types in SF: 246 The five most common venues are: 1- Coffee Shop ... 47 (3.76%) 2- Café ... 37 (2.96%) 3- Italian Restaurant ... 36 (2.88%) 4- Bakery ... 30 (2.40%) 5- Chinese Restaurant ... 28 (2.24%)

New York City Venues	
Top 20 Venue Types in New York City chart showing: Pizza Place 94, Coffee Shop 72, Chinese Restaurant 62, Italian Restaurant 51, Mexican Restaurant 48, Bakery 47, Bar 46, Deli / Bodega 41, Gym 41, Sandwich Place 37, Grocery Store 36, American Restaurant 31, Pharmacy 31, Park 31, Café 30, Gym / Fitness Center 30, Donut Shop 30, Ice Cream Shop 29, Hotel 28, Cosmetics Shop 27	Number of venues in NY: 2097 Avg. number of venues per neighborhood: 99.85 Number of venue types in NY: 272 The five most common venues are: 1- Pizza Place ... 94 (4.48%) 2- Coffee Shop ... 72 (3.43%) 3- Chinese Restaurant ... 62 (2.96%) 4- Italian Restaurant ... 51 (2.43%) 5- Mexican Restaurant ... 48 (2.29%)

At first glance it can be noted that both Los Angeles and New York City contain 4 eateries among their top 5 most common venue types while San Francisco only contains 2 restaurants. Instead, coffee shops, cafes and bakeries seem to be some of the most common venue type in San Francisco.

It can also be noted that the overall number of venues extracted in Los Angeles is larger than in the other two cities, which can be explained by the larger number of neighborhoods and the larger radius sent to the FourSquare API to extract venues. On the other hand, while roughly the same radius was used on both New York and San Francisco, the number of venues collected in New York is ~68% higher than in San Francisco, pointing to a much denser distribution of venues in New York. This is corroborated by the average number of venues per neighborhood result, where New York shows over 130% higher density than Los Angeles and a similar ~68% more than in San Francisco.

Analysis with larger radius sizes for all cities (e.g. 1000 meters in LA and 800 meters on both NY and SF) changed the results only slightly, switching the position of some of the top 10 venues, but the overall percentage remained very similar.

Meat-Serving Venues:

To focus our analysis on venues relevant to our meat-supplier client, in this section we filter all venues that are not related to meat and keep only those that may be serving or selling meat. In particular, we keep only the venue types containing the following strings in their '*Venue Category*' column:

- Restaurant
- market
- Market
- Food
- Pizza
- Hotel
- Burger
- Diner
- Hot Dog
- Grocery
- Sandwich
- Snack
- Salad
- Taco
- BBQ
- Bodega
- Burrito
- Meat

We then repeat the analysis surrounding Table 5 above, but apply it now to the new data frames containing only meat related venues. The results for all three cities are displayed in Table 6 below, combining as before some general statistics of the venue types and horizontal bar plots of their distributions.

Table 6: Analysis of meat-serving venue type distribution per city
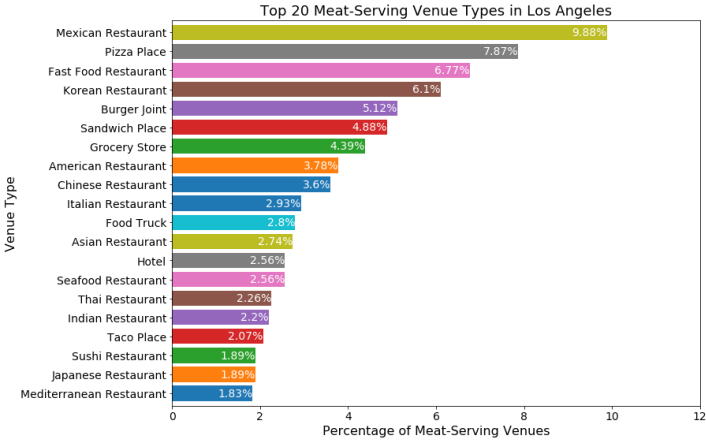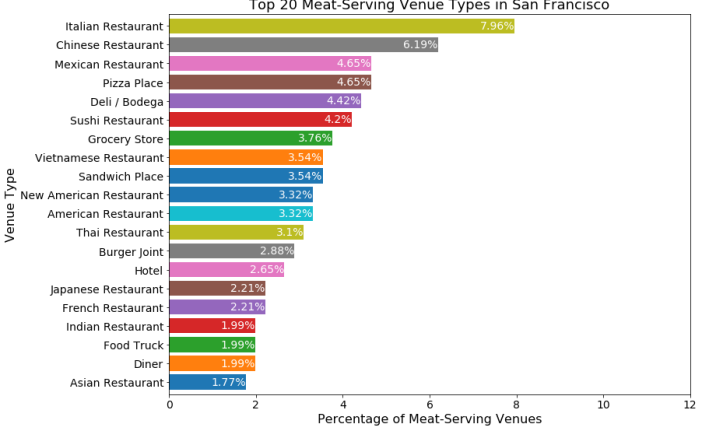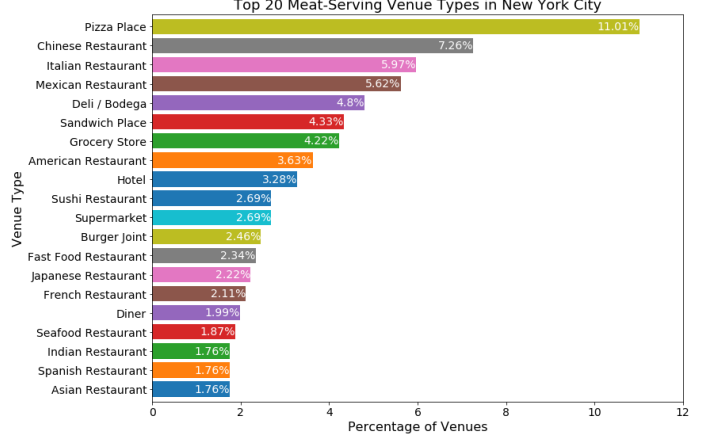
Los Angeles Venues

Top 20 Meat-Serving Venue Types in Los Angeles

Number of meat-serving venues in LA: 1640
Avg. number of meat-serving venues per neighborhood: 19.1
Number of meat-serving venue types: 79

The five most common meat-serving venues in LA are:
1- Mexican Restaurant ... 162 (9.88%)
2- Pizza Place ... 129 (7.87%)
3- Fast Food Restaurant ... 111 (6.77%)
4- Korean Restaurant ... 100 (6.1%)
5- Burger Joint ... 84 (5.12%)

Number of neighborhoods in LA: 86
Number of neighborhoods with at least one venue: 85
Number of nbhds with at least one meat-serving venue: 80
Max. number of meat-serving venues per neighborhood: 63

San Francisco Venues

Top 20 Meat-Serving Venue Types in San Francisco

Number of meat-serving venues in SF: 452
Avg. number of meat-serving venues per neighborhood: 21.5
Number of meat-serving venue types: 77

The five more common meat-serving venues are:
1- Italian Restaurant ... 36 (7.96%)
2- Chinese Restaurant ... 28 (6.19%)
3- Pizza Place ... 21 (4.65%)
4- Mexican Restaurant ... 21 (4.65%)
5- Deli / Bodega ... 20 (4.73%)

Number of neighborhoods in SF: 21
Number of neighborhoods with at least one venue: 21
Number of nbhds with at least one meat-serving venue: 18
Max. number of meat-serving venues per neighborhood: 50

New York City Venues

Top 20 Meat-Serving Venue Types in New York City

Number of meat-serving venues: 854
Avg. number of meat-serving venues per neighborhood: 20.3
Number of meat-serving venue types: 83

The five most common meat-serving venues are:
1- Pizza Place ... 94 (11.01%)
2- Chinese Restaurant ... 62 (7.26%)
3- Italian Restaurant ... 51 (5.97%)
4- Mexican Restaurant ... 48 (5.62%)
5- Deli / Bodega ... 41 (4.80%)

Number of neighborhoods in NY: 42
Number of neighborhoods with at least one venue: 42
Number of nbhds with at least one meat-serving venue: 41
Max. number of meat-serving venues per neighborhood: 50

Table 7: Neighborhood Distribution of Meat-Serving Venues

Los Angeles	
	⬤ Neighborhoods ⬤ Neighborhoods with most number of meat-serving venues

San Francisco	
	⬤ Neighborhoods ⬤ Neighborhoods with most number of meat-serving venues

New York City	
	⬤ Neighborhoods ⬤ Neighborhoods with most number of meat-serving venues
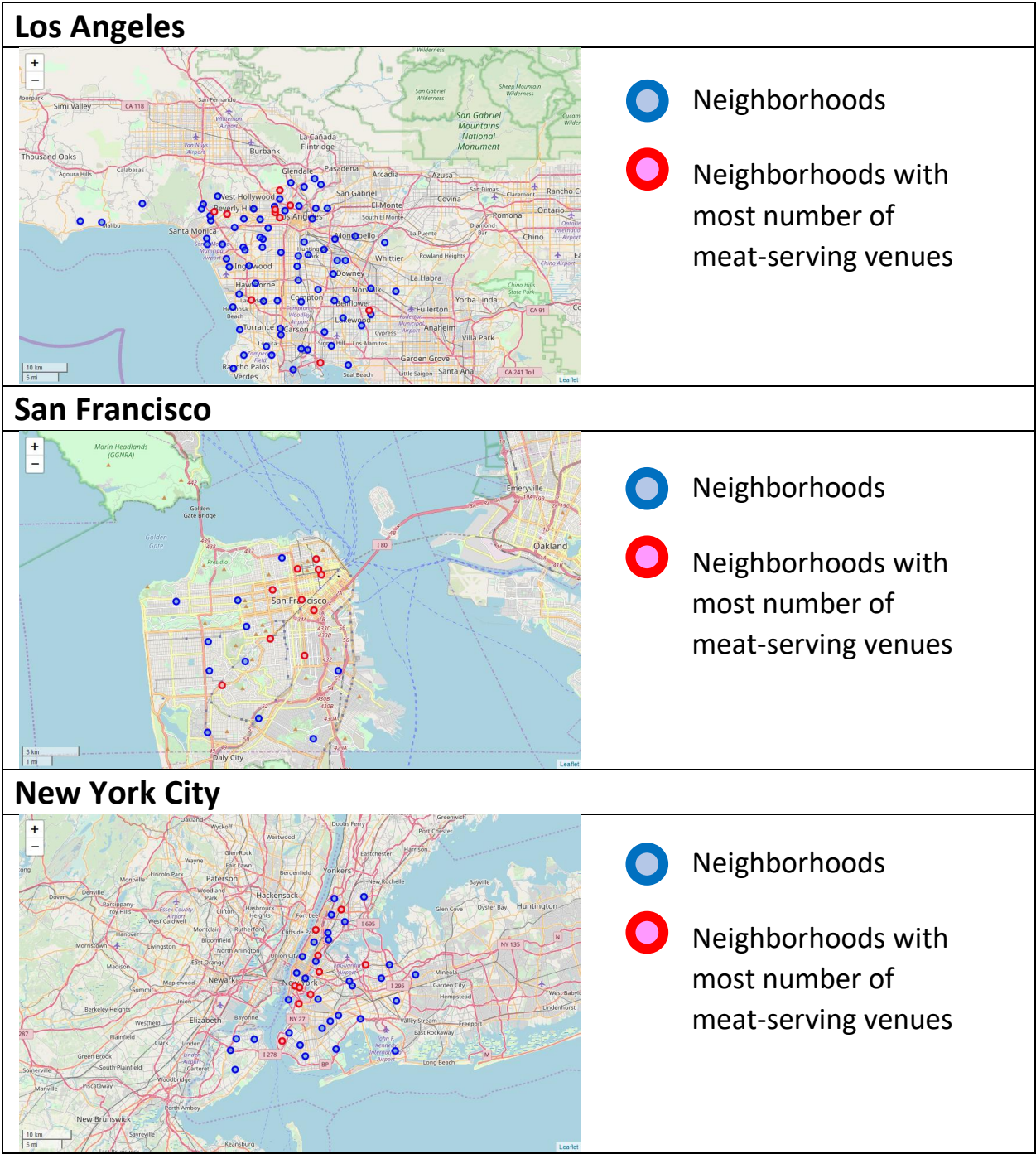
Regarding the results for potential meat-serving venues in Table 6, they show a similar trend in venue volume between cities, that is, with Los Angeles showing a larger absolute number of venues. When focusing on the average number of meat-serving venues per neighborhood, however, roughly the same number around 20 venues is observed in all cities, with San Francisco

actually showing the largest number. When looking at the distribution of potential meat-serving venues, however, we observe that Mexican restaurants and pizza places are the most popular venues in LA while pizza places seem to dominate in New York and Italian and Chinese restaurants share the first two positions in San Francisco.

Finally, Table 7 collects plots of the city neighborhoods where the top 10 neighborhoods with the the highest number of potentially meat-serving venues highlighted in red. It is observed that while in Los Angeles this group of top 10 neighborhoods seem concentrated in the northern part, the distribution is more uniform in the other two cities.

Only potentially meat-serving venues are used from this point on.

b) Modeling and Evaluation

Get_Dummy Method:

To transform venue categories into vectors in order to perform operations such as calculate distances and apply clustering algorithms, we apply the pandas function get_dummies to the venue dataframes. This function "converts categorical variable into dummy/indicator variables".

Moreover, since we want to compare the venues across 3 cities, we will create one big dataframe containing the venues of all cities and then apply the function get_dummies to it. Afterwards we separate the rows on this big dataframe, grouped by neighborhood, into three dataframes, one for each city as displayed in Table 8.

Table 8: Venue dataframes after applying get_dummy function from pandas library

Los Angeles One_Hot

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	BBQ Joint	...	Taiwanese Restaurant	Tapas Restaurant	T Res
0	Arlington Heights (Los Angeles)	0.0	0.0	0.035714	0.0	0.0	0.000000	0.0	0.0	0.035714	...	0.000000	0.0	
1	Artesia, Artesia (PO Boxes)	0.0	0.0	0.013889	0.0	0.0	0.041667	0.0	0.0	0.013889	...	0.013889	0.0	
2	Athens	0.0	0.0	0.041667	0.0	0.0	0.000000	0.0	0.0	0.000000	...	0.000000	0.0	
3	Atwater Village (Los Angeles)	0.0	0.0	0.060606	0.0	0.0	0.000000	0.0	0.0	0.000000	...	0.000000	0.0	
4	Avalon (PO Boxes)	0.0	0.0	0.076923	0.0	0.0	0.000000	0.0	0.0	0.000000	...	0.000000	0.0	

5 rows × 119 columns

San Francisco One_Hot

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	BBQ Joint	...	Taiwanese Restaurant	Tapas Restaurant	F
0	Bayview-Hunters Point	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.111111	...	0.0	0.00	
1	Castro/Noe Valley	0.0	0.0	0.060606	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.00	
2	Chinatown	0.0	0.0	0.088235	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.00	
3	Haight-Ashbury	0.0	0.0	0.050000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.05	
4	Hayes Valley/Tenderloin /North of Market	0.0	0.0	0.027778	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.00	

5 rows × 119 columns

New York City One_Hot

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	BBQ Joint	...	Taiwanese Restaurant	Tapas Restaurant	Tex-I Restaur
0	Borough Park	0.0	0.000000	0.000000	0.0	0.0	0.064516	0.0	0.0	0.0	...	0.000000	0.0	
1	Bronx Park and Fordham	0.0	0.000000	0.023256	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.000000	0.0	
2	Bushwick and Williamsburg	0.0	0.000000	0.000000	0.0	0.0	0.037037	0.0	0.0	0.0	...	0.037037	0.0	
3	Canarsie and Flatlands	0.0	0.000000	0.058824	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.000000	0.0	
4	Central Bronx	0.0	0.058824	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.000000	0.0	

5 rows × 119 columns

k-means clustering:

In this next step, the neighborhoods are clustered based on the type of meat-serving venues that they contain using **k-means** algorithm from **sklearn.cluster** library. The number of clusters was set to 5, that is, k=5, which helps simplify the comparison between cities.

To understand the clusters better, we create a new dataframe adding the cluster labels to all meat-serving venues and then group them by cluster number. We then display the top 5 venues for each cluster and the cluster size. The results can be seen in Table 9.

We observe that, in general, the top two clusters tend to group most of the neighborhoods in the city. This is particularly clear in the case of Los Angeles, but also seen in New York City and to lesser extend in San Francisco.

Pizza, Mexican and Italian restaurants seem to dominate the top two clusters in both LA and NY, while Chinese and Sushi restaurants appear more often in SF.

Table 10 shows also the 5 clusters along with the top 5 most common meat-serving venues, the size of the cluster and the frequency of the venue per city.

Table 9: Dataframes with the top 5 most common meat-serving venues per cluster and the cluster size.

Los Angeles Clusters

	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Cluster Size
4	4	Pizza Place	Mexican Restaurant	Fast Food Restaurant	Burger Joint	Grocery Store	56
0	0	Mexican Restaurant	Fast Food Restaurant	Food	Sandwich Place	Food Truck	11
1	1	Hotel	Italian Restaurant	Restaurant	Sandwich Place	American Restaurant	5
2	2	Burger Joint	Grocery Store	Food & Drink Shop	American Restaurant	Fast Food Restaurant	4
3	3	Korean Restaurant	Asian Restaurant	Mexican Restaurant	Grocery Store	Japanese Restaurant	4

San Francisco Clusters

	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Cluster Size
0	0	Sushi Restaurant	Deli / Bodega	Italian Restaurant	American Restaurant	Sandwich Place	8
2	2	Italian Restaurant	Chinese Restaurant	Pizza Place	Mexican Restaurant	Burger Joint	6
1	1	Chinese Restaurant	Hotpot Restaurant	Sushi Restaurant	Grocery Store	Italian Restaurant	2
3	3	Asian Restaurant	Sandwich Place	BBQ Joint	Vietnamese Restaurant	French Restaurant	1
4	4	Pizza Place	Burger Joint	Mexican Restaurant	Food Truck	French Restaurant	1

New York City Clusters

	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Cluster Size
3	3	Pizza Place	Deli / Bodega	Mexican Restaurant	Sandwich Place	Supermarket	15
0	0	Italian Restaurant	American Restaurant	Mexican Restaurant	Pizza Place	Grocery Store	13
2	2	Chinese Restaurant	Pizza Place	Grocery Store	Italian Restaurant	Fast Food Restaurant	11
1	1	Moroccan Restaurant	Vietnamese Restaurant	Ethiopian Restaurant	Farmers Market	Fast Food Restaurant	1
4	4	Italian Restaurant	Vietnamese Restaurant	French Restaurant	Farmers Market	Fast Food Restaurant	1

Table 10: Top 5 most populous clusters with their top 5 meat-serving venues.

Los Angeles	San Francisco	New York City

```
---- Cluster Number = 4 ----
---- Cluster Size = 56 neighborhoods ----
                     venue  freq
0              Pizza Place  0.09
1        Mexican Restaurant  0.08
2       Fast Food Restaurant  0.07
3             Grocery Store  0.06
4              Burger Joint  0.06


---- Cluster Number = 0 ----
---- Cluster Size = 11 neighborhoods ----
                     venue  freq
0        Mexican Restaurant  0.29
1       Fast Food Restaurant  0.20
2                     Food  0.06
3            Sandwich Place  0.05
4         Seafood Restaurant  0.04


---- Cluster Number = 1 ----
---- Cluster Size = 5 neighborhoods ----
                    venue  freq
0                    Hotel  0.42
1         Italian Restaurant  0.13
2            Sandwich Place  0.11
3                Restaurant  0.11
4        American Restaurant  0.06


---- Cluster Number = 2 ----
---- Cluster Size = 4 neighborhoods ----
                    venue  freq
0              Burger Joint  0.62
1             Grocery Store  0.15
2          Food & Drink Shop  0.12
3            Sandwich Place  0.02
4         Chinese Restaurant  0.02


---- Cluster Number = 3 ----
---- Cluster Size = 4 neighborhoods ----
                    venue  freq
0          Korean Restaurant  0.40
1             Grocery Store  0.05
2           Asian Restaurant  0.05
3         Mexican Restaurant  0.05
4        Japanese Restaurant  0.04
```

```
---- Cluster Number = 0 ----
---- Cluster Size = 8 neighborhoods ----
                     venue  freq
0       American Restaurant  0.06
1         Italian Restaurant  0.06
2           Sushi Restaurant  0.06
3             Deli / Bodega  0.06
4      Vietnamese Restaurant  0.05


---- Cluster Number = 2 ----
---- Cluster Size = 6 neighborhoods ----
                    venue  freq
0              Pizza Place  0.09
1         Italian Restaurant  0.09
2         Chinese Restaurant  0.09
3         Mexican Restaurant  0.08
4              Burger Joint  0.05


---- Cluster Number = 1 ----
---- Cluster Size = 2 neighborhoods ----
                    venue  freq
0         Chinese Restaurant  0.29
1           Sushi Restaurant  0.10
2           Hotpot Restaurant  0.10
3             Grocery Store  0.10
4         Italian Restaurant  0.06


---- Cluster Number = 3 ----
---- Cluster Size = 1 neighborhoods ----
                    venue  freq
0            Sandwich Place  0.33
1           Asian Restaurant  0.33
2                 BBQ Joint  0.33
3         African Restaurant  0.00
4         Mexican Restaurant  0.00


---- Cluster Number = 4 ----
---- Cluster Size = 1 neighborhoods ----
                    venue  freq
0              Pizza Place  0.4
1         Mexican Restaurant  0.2
2              Burger Joint  0.2
3                Food Truck  0.2
4         African Restaurant  0.0
```

```
---- Cluster Number = 3 ----
---- Cluster Size = 15 neighborhoods ----
                     venue  freq
0              Pizza Place  0.17
1             Deli / Bodega  0.09
2        Mexican Restaurant  0.08
3            Sandwich Place  0.07
4               Supermarket  0.06


---- Cluster Number = 0 ----
---- Cluster Size = 13 neighborhoods ----
                     venue  freq
0         Italian Restaurant  0.08
1        American Restaurant  0.08
2         Mexican Restaurant  0.06
3              Pizza Place  0.06
4            Sandwich Place  0.04


---- Cluster Number = 2 ----
---- Cluster Size = 11 neighborhoods ----
                     venue  freq
0         Chinese Restaurant  0.25
1              Pizza Place  0.17
2             Grocery Store  0.09
3         Italian Restaurant  0.06
4       Fast Food Restaurant  0.05


---- Cluster Number = 1 ----
---- Cluster Size = 1 neighborhoods ----
                     venue  freq
0        Moroccan Restaurant  1.0
1         African Restaurant  0.0
2        Romanian Restaurant  0.0
3           Ramen Restaurant  0.0
4              Pizza Place  0.0


---- Cluster Number = 4 ----
---- Cluster Size = 1 neighborhoods ----
                     venue  freq
0         Italian Restaurant  1.0
1         African Restaurant  0.0
2         Mexican Restaurant  0.0
3           Ramen Restaurant  0.0
4              Pizza Place  0.0
```

c) Comparison between Cities

Table 11: Geographical distribution of neighborhood clusters based on meat-serving venues

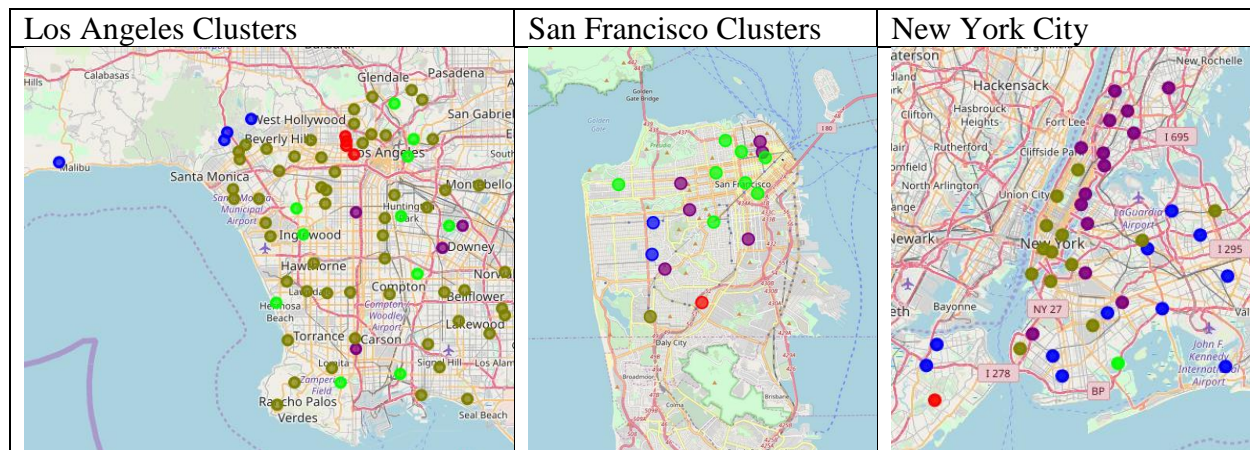| Los Angeles Clusters	San Francisco Clusters	New York City

Table 11 collects the maps with the geographical distribution of the neighborhood clusters, grouped and colored based on clusters of similar potential meat-serving venues. Los Angeles neighborhoods seem to belong mostly to one cluster while the neighborhoods in New York and San Francisco appear more distributed into 2 or 3 clusters, although when focusing only on Manhattan in New York City, only two clusters appear to dominate the scene.

Next we compute a couple of metrics to decide which city has a closer venue profile to LA. In particular we use the *scipy.spatial* library to compute Euclidean distances between vectors of venues.

First we compute the Euclidean distance between all venues returned for each city and then for the venues contained within the top 2 clusters. Table 12 collects all the results between cities, highlighting the city with the minimum distance in each case. It can be seen that when considering the entire sample of representative meat-serving venues per city, the Euclidean distance between cities slightly smaller in the case of San Francisco than for New York City. However, when the computation is narrowed down to the top 2 clusters per city, which represents the most number of neighborhoods, the Euclidean distance computed for the city of New York results in a smaller value than that of San Francisco.

Table 12: Euclidean distance between venues in each city

LA vs	All venues	Top 2 Clusters
	Distance	Distance
SF	0.142	0.226
NY	0.151	0.197

Next we compute distance between venues in each individual cluster and collect the results in Table 13. It is seen that when computing the Euclidean distance between pairs of clusters in Los Angeles vs either San Francisco or New York, focusing on the top 2 per each city, the distance between LA and NY always results in a smaller value than that between LA and SF. As a consequence, the results in the last row in Table 13 shown New York as the city with the closer meat-serving venue profile to Los Angeles.

Euclidean Distances			
SF		LA First Cluster	LA Second Cluster
	SF First Cluster	0.171866	0.375680
	SF Second Cluster	0.143351	0.341144
NY		LA First Cluster	LA Second Cluster
	NY First Cluster	0.165150	0.337169
	NY Second Cluster	0.123031	0.336007
Preferences		LA First Cluster	LA Second Cluster
	SF or NY First Cluster	NY	NY
	SF or NY Second Cluster	NY	NY

5- Discussion and Recommendations

Considering the results obtained in the previous section, we can collect here a few points to highlight in the topic of city similarity based on venue profile. In particular:

- ➢ The number of neighborhoods and their geographical extent in SF is the smallest of all three cities. As a consequence, the number of venues in absolute terms is smaller in SF.

- ➢ The list of venue types in SF is dominated by coffee shops, cafes and bakeries while the venue types in LA and NY seem dominated by pizza places and other potentially meat-serving restaurants.

- ➢ The distribution of the top 10 neighborhoods with the most meat-serving venues in LA is rather concentrated in a couple of areas while the neighborhoods with the most meat-serving venues seem more informingly distributed in the other two cities.

- ➢ Neighborhood clusters based on venues is much more concentrated in LA were the most neighborhoods are collected in one cluster, than the other two cities.

- ➢ However, when focusing on Manhattan in NY, it seems dominated by the top 2 clusters.

- ➢ Euclidean distance between cities is almost always smaller for NY than for SF except when considering all venues, in which case SF shows smaller distance. When looking at the top 2 most populous clusters, the results all point to NY as the city with the most similar profile to Los Angeles.

Based on this discussion, the recommendation provided to our meat-supplier client is to look to New York City to expand his business as the city with the closer profile to his home base of Los Angeles.

6- Conclusions

In this project we collected information about neighborhoods and also about venues contained in those neighborhoods for three cities, focusing on venues with potential to serve or sell meat. Then we explored the data for each city by grouping the neighborhoods into clusters based on their venue profile and computed Euclidean distances between venues in each city and between pairs of clusters in each city. The objective of the study is to find the city between San Francisco and New York City that more closely resembles the meat-serving venue profile found in the city of Los Angeles. The result will be used to provide a recommendation to our client who owns a meat supplying business currently operating in Los Angeles and who is looking to expand his business in one of those cities. Based on the results of the study the recommendation was to expand in the city of New York.

7- Link to Github Notebook Repository

https://github.com/jaionet/CapstoneProjectNotebook/blob/master/DSCapstoneProject-Week5.ipynb

8- Link to Github Report Repository

9- Link to Github Presentation Repository