# Comparing Neighborhoods: **LA vs SF and NY**

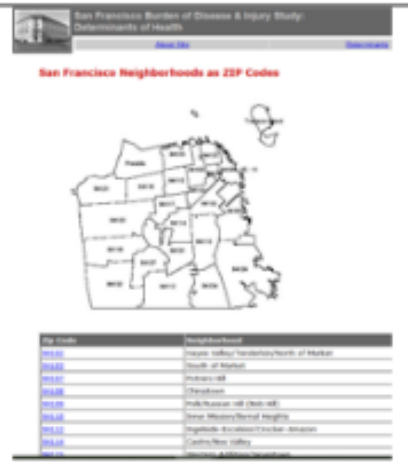
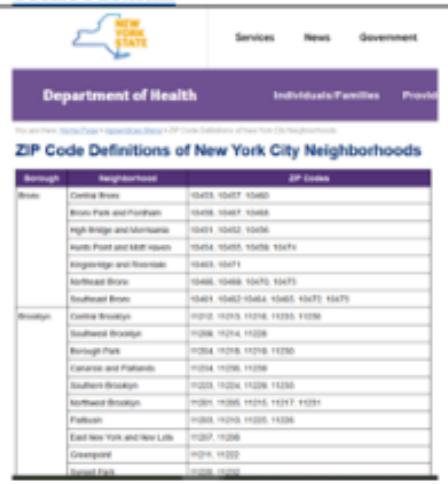
Data Science Capstone Project – **Final Report**

# Background and Motivation

- Deli meat supplier located in Los Angeles wants to expand his business
- Currently considering expanding to another populous city with a similar venue profile
- Candidate cities:
  - San Francisco
  - New York City

# Data Description: Web data scraping

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

| 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 |
|   |   |   |

Number of neighborhoods in LA: 91

| | Neighborhood | Zip Code |
|---|---|---|
| 0 | Arlington Heights (Los Angeles) | 90019 |
| 1 | Artesia, Artesia (PO Boxes) | 90701 |
| 2 | Athens | 90044 |
| 3 | Atwater Village (Los Angeles) | 90039 |
| 4 | Avalon (PO Boxes) | 90704 |

Number of neighborhoods in SF: 21

| | Neighborhood | Zip Code |
|---|---|---|
| 0 | Hayes Valley/Tenderloin/North of Market | 94102 |
| 1 | South of Market | 94103 |
| 2 | Potrero Hill | 94107 |
| 3 | Chinatown | 94108 |
| 4 | Polk/Russian Hill (Nob Hill) | 94109 |

Number of neighborhoods in NY: 42

| | Neighborhood | Zip Code |
|---|---|---|
| 0 | Central Bronx | 10453 |
| 1 | Bronx Park and Fordham | 10458 |
| 2 | High Bridge and Morrisania | 10451 |
| 3 | Hunts Point and Mott Haven | 10454 |
| 4 | Kingsbridge and Riverdale | 10463 |

# Data Description: Coordinate Extraction

- Neighborhood coordinates obtained using the function ***Nominatim*** from the library ***geopy.geocoders***

- Plotting of geographical distribution of neighborhoods using ***Folium*** library

**Los Angeles**

Number of neighborhoods in LA: 86

| | 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**

Number of neighborhoods in SF: 21

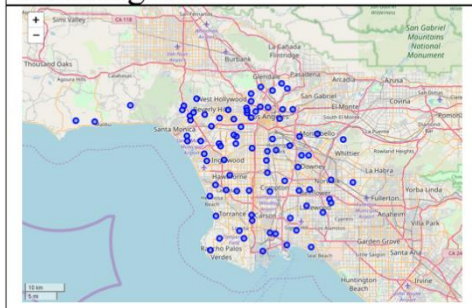
| | Neighborhood | Zip Code | Latitude | Longitude |
|---|---|---|---|---|
| 0 | Hayes Valley/Tenderloin/North of Market | 94102 | 37.779491 | -122.418224 |
| 1 | South of Market | 94103 | 37.774425 | -122.411091 |
| 2 | Potrero Hill | 94107 | 37.793634 | -122.408295 |
| 3 | Chinatown | 94108 | 37.791043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

**New York City**

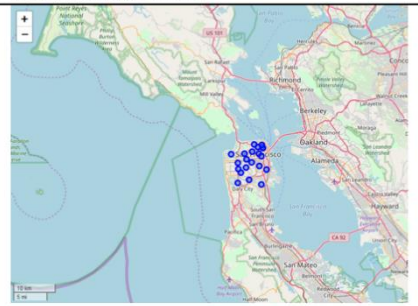
Number of neighborhoods in NY: 42

| | 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 |

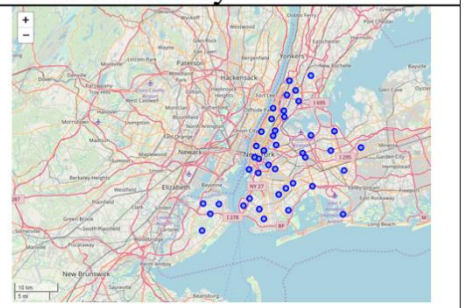
Los Angeles



San Francisco



New York City



# Data Description: Venue Sampling

- ***FourSquare API*** used to extract coordinates and information about venues within a radius of each neighborhood center.
  - LA Radius = 850 meters
  - SF Radius = 500 meters
  - NY Radius = 550 meters
  - Maximum number of venues per neighborhood = 100
- Search radius adjusted to include same number of meat-serving venues per neighborhood.

# Data Analysis Methodology

**Data understanding and preparation**

- Venue distributions and percentages
- Meat-serving venues filtering
- Visualization

**Modeling**

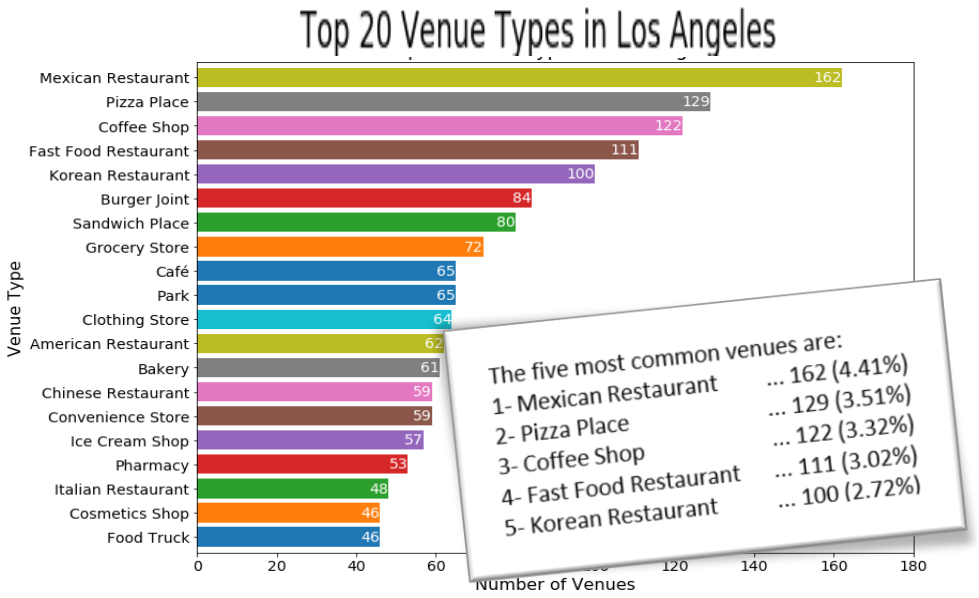
- Clustering using k-means algorithm
- Cluster characteristics and distribution

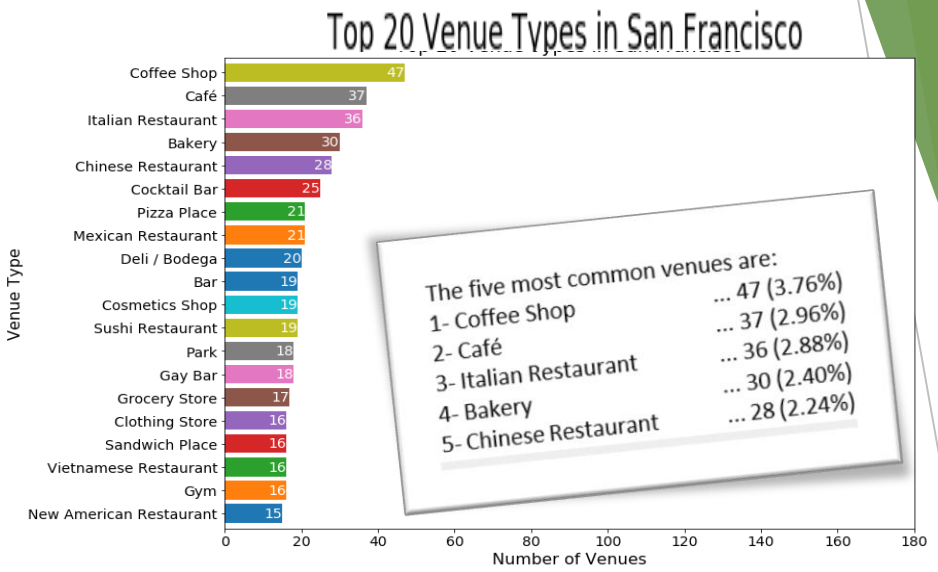
**City Comparison**

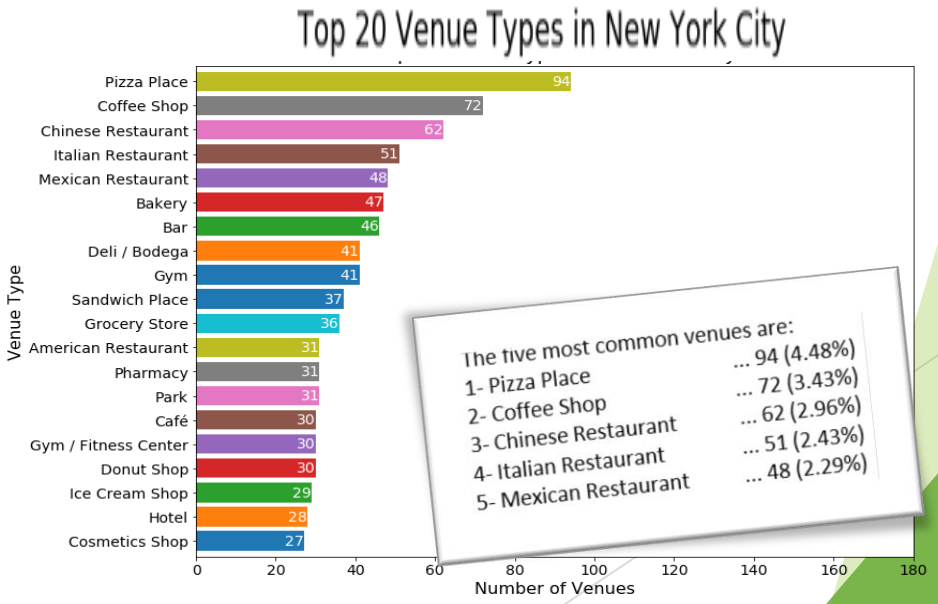
- Visual comparison
- Euclidean distance

| | 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 |

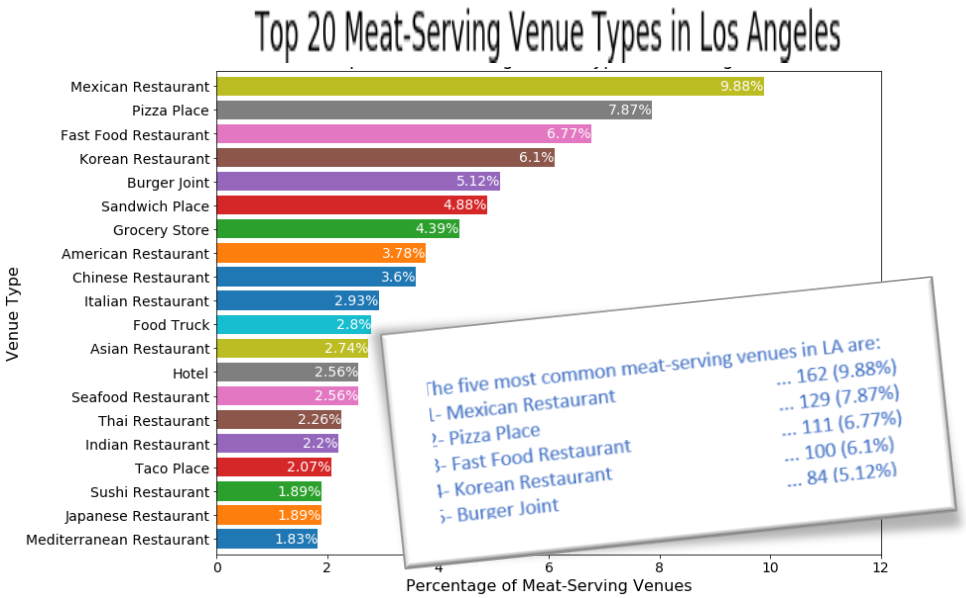
# Results: Venues

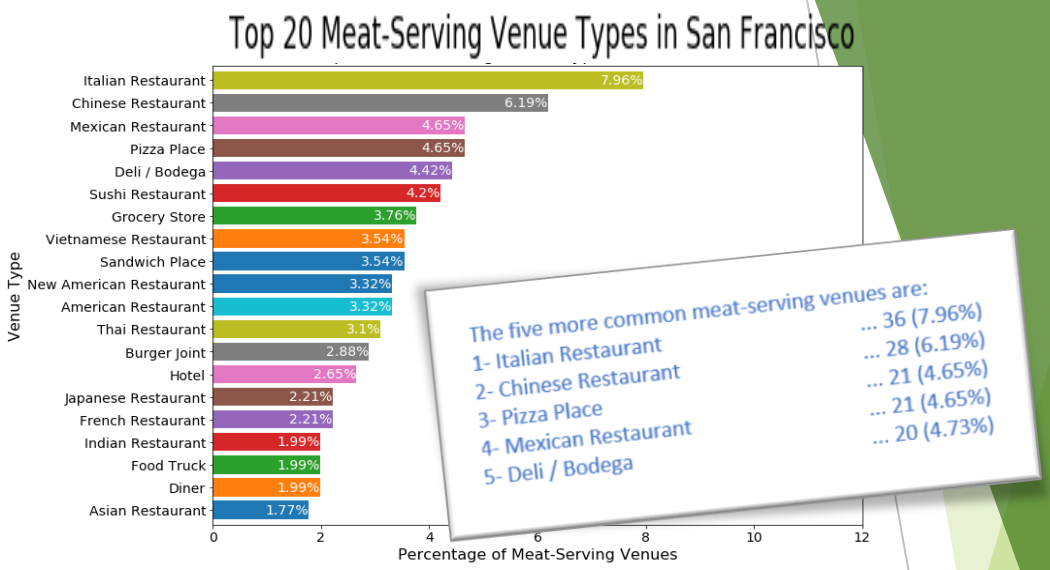


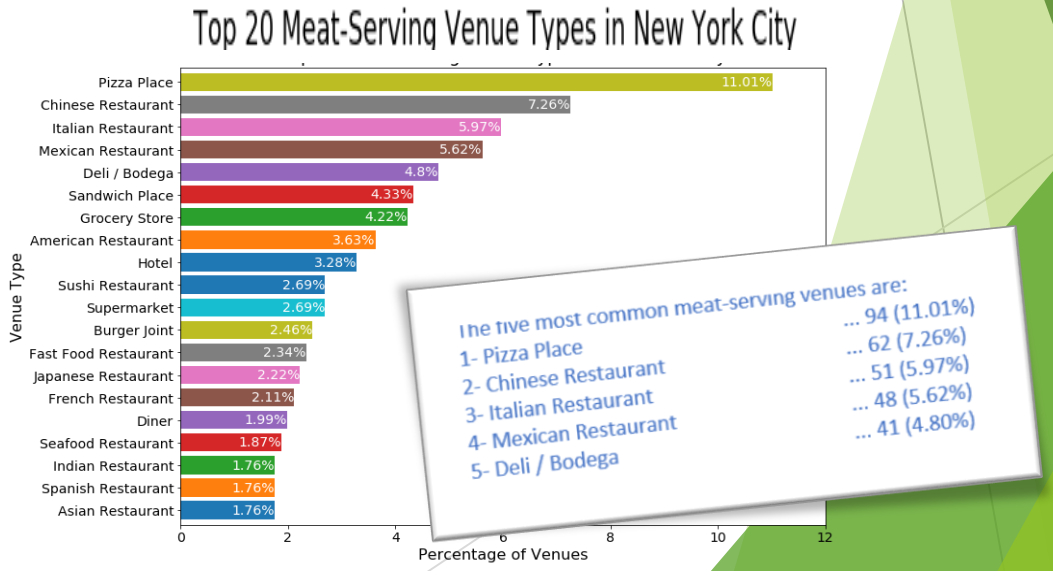



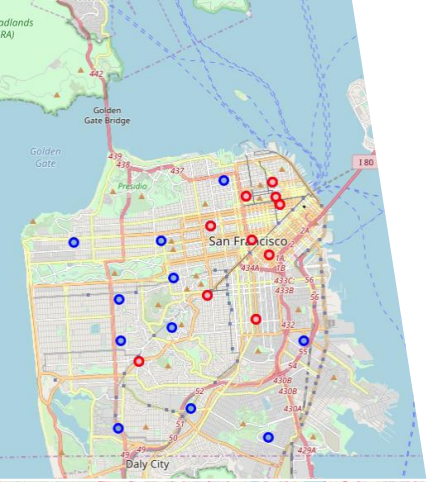



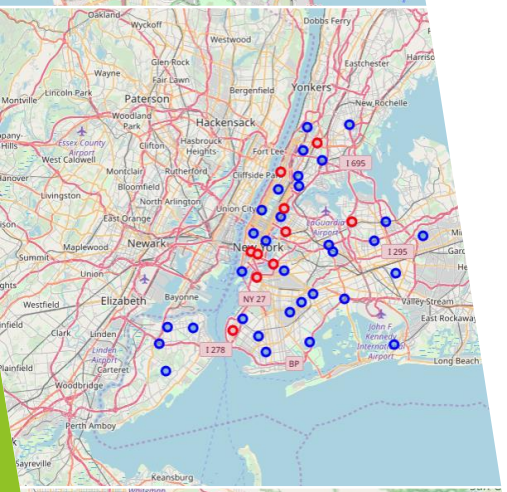
# Results:
# Meat-Serving Venues

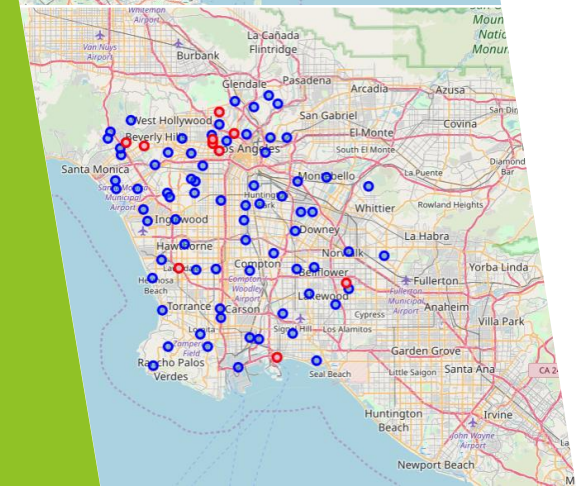





# Results:
# Distribution of Meat-Serving Venues

- SF → Higher concentration on north-eastern part of the city
- NY → Higher concentration on Manhattan
- LA → Higher concentration on northern part of the city

# Modeling:
# K-means Clustering

- Top 2 clusters concentrate most of the neighborhoods in all cities but specially in LA.
- Most common in LA:
  - Pizza/Mexican/Fast food
- Most common in SF:
  - Sushi/Deli/Italian/Chinese
- Most common in NY:
  - Pizza/Deli/Italian/American

## Los Angeles Clusters

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

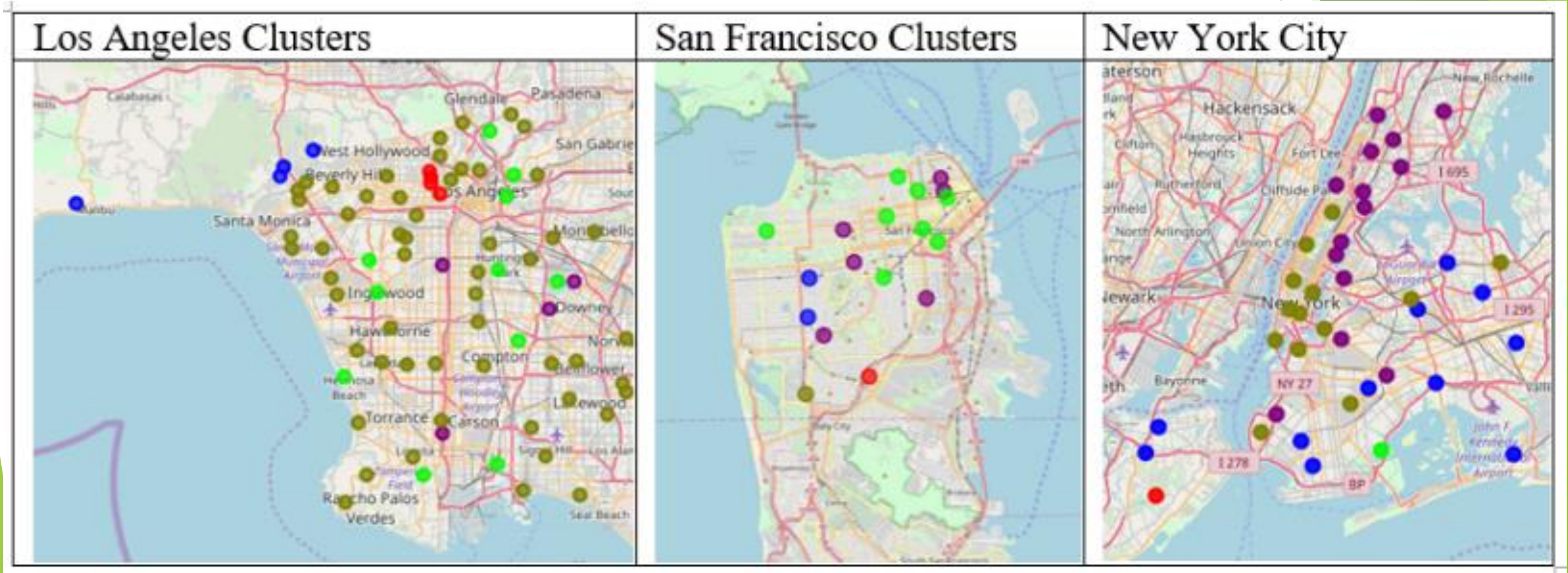
## San Francisco Clusters

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

## New York City Clusters

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

# Modeling: **Cluster Distribution**



# Results: Euclidean Distance

- Between LA and SF
- Between LA and NY

- For all venues
- For venues in top 2 clusters
- Between pairs of clusters

- NY results in the smallest distance in most cases

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

**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 |

# Discussion: Takeaways

- Smaller number of venues in in SF.
- Venue types in SF dominated by coffee shops, cafes and bakeries
- Venue types in LA and NY show more potentially meat-serving eateries or stores.
- Most neighborhoods concentrated on top 2 clusters in all cities
- Euclidean distance between pair of top 2 clusters shows smaller distance between LA and NY than between LA and SF.

# Recommendation

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.

# Conclusions

- In this project we collected information about neighborhoods and venues for three cities
- Filtered the venues to focus on venues with potential to serve or sell meat.
- We explored the data for each city by grouping the neighborhoods into clusters based on their venue profile.
- Computed Euclidean distances between venues in each city and between pairs of clusters in each city.
- Provided recommendation to our client base on the study.

# Links:

- Code repository

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

- Report

https://github.com/jaionet/CapstoneProjectNotebook/blob/master/Data%20Science%20Capstone%20Final%20Project%20-%20Week5.pdf

- Presentation

https://github.com/jaionet/CapstoneProjectNotebook/blob/master/DS%20Capstone.%20Final%20Project%20Presentation.pdf

- Watson Studio Notebook

https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/ddff3368-8fa4-4678-8b77-b7111235fcc3/view?access_token=c54775c3744c4c168c3838021f81b732e46cc2b67c37fa4d03ab45a40470e9bd