Comparing Manhattan, NY and San Francisco, CA

1. Introduction **2**

1.1 Description of the problem and a discussion of the background 2

1.2 Description of the data and how it will be used to solve the problem 2

2. Methodology **2**

2.0. Importing all necessary libraries to work with the data. 2

2.1 First, we need the data about neighborhoods in both cities, including the geo coordinates for each neighborhood. 3

2.2. Downloading information about venues in both cities using Foursquare API 6

2.3. Next I performed clustering of the neighborhoods. 7

2.3.2. Mapping the clusters with folium 9

2.4. Venues diversity comparison in both cities. 10

2.5 Create a dataframe with the most common venues in Manhattan and SF and visualize it with Word Cloud: 11

2.6. Finally, we will explore which cities are unique to Manhattan and San Francisco. 13

2.6.1. First, we create a combined file for comparison: 13

2.6.2. Export a list of venues that are present in San Francisco but missing in Manhattan: 14

2.6.3. Export a list of venues that are present in Manhattan but missing in San Francisco: 14

3. Results **15**

4. Discussion **15**

5. Conclusion **16**

1. Introduction

1.1 Description of the problem and a discussion of the background

I live in New York City and know it well. With the possibility of remote work, I want to research other cities where I can live that will be similar to New York. New York is often compared to San Francisco. Although the two cities are located on opposite sides of the USA, it's a popular opinion that both cities have diverse cultures, lots of creative spaces and various venues.
I want to explore the venues in both cities and compare the neighborhoods to test if they are really so similar to each other. In this study, I will explore the diversity of the neighborhoods in both cities.
The study can be extended to any other pairs of cities. In the current situation when many people are working remotely, many are traveling away from their home cities, using it as an opportunity to explore other places within the country or even abroad.
The results might also be interesting for business owners. For instance, if some venues are popular in New York but not in San Francisco or vise versa, businesses can further research the possibility of expanding their business in another city.

1.2 Description of the data and how it will be used to solve the problem

- I will use the New York City neighborhoods data provided in the previous Lab.
- In addition, I will use Foursquare API to download the list of venues for NY and SF.
- San Francisco neighborhoods data will be downloaded from this website:
- http://www.healthysf.org/bdi/outcomes/zipmap.htm
- I will also use some libraries (geopy, geocoder, geopandas) to extract geo coordinates for the neighborhoods.

2. Methodology

2.0. Importing all necessary libraries to work with the data.

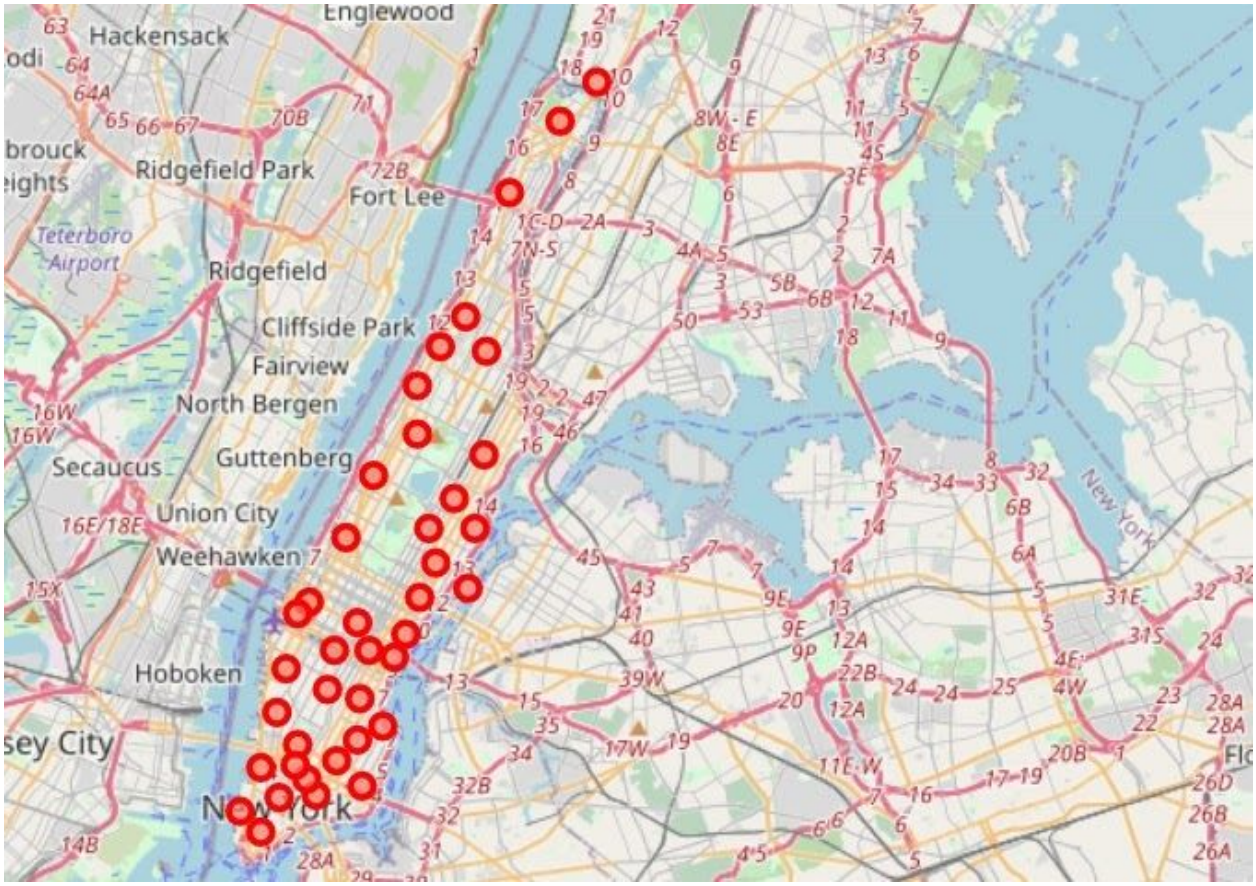
The libraries include: numpy, pandas, json, requests, matplotlib, sklearn - Kmeans, folium, seaborn, geocoder, geopandas, beautiful soup, lxml, and wordcloud.
I also created a file with my Foursquare credentials and imported it. That allows me to use my credentials without revealing them in the notebook.

2.1 First, we need the data about neighborhoods in both cities, including the geo coordinates for each neighborhood.

New York Neighborhoods dataset

We will be analysing Manhattan neighborhoods in New York using the datafile provided in the Lab session.

Below is the folium map of neighborhoods in Manhattan, New York:



San Francisco Neighborhoods dataset.

2.1.1. Download the table using Beautiful Soup library:
http://www.healthysf.org/bdi/outcomes/zipmap.htm

```
url = "http://www.healthysf.org/bdi/outcomes/zipmap.htm"

response = requests.get(url)
soup = BeautifulSoup(response.text, "lxml")
table = soup.find_all("table")
sf_neighborhoods = pd.read_html(str(table))
sf_neighborhoods = pd.DataFrame(sf_neighborhoods[4])
```

2.1.2. Clean the data: remove unnecessary columns, create column names, reset indexes. Final file:

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

2.1.3. Add geo coordinates. I used geocode library to add coordinates to the neighborhoods:

```
latitude = []
longitude = []
for n in range(0, 21):
    zip_code = sf_neighborhoods['Zip Code'][n]
    location = locator.geocode(zip_code)
    latitude.append(location.latitude)
    longitude.append(location.longitude)
```

```
sf_neighborhoods['latitude'] = pd.Series(latitude)
sf_neighborhoods['longitude'] = pd.Series(longitude)
```

2.1.4. Some of the coordinates looked a little off. The next step is to check correctness of these coordinates:

```
geocode = RateLimiter(locator.geocode, min_delay_seconds=1)
```

```
sf_neighborhoods['location'] = sf_neighborhoods['Zip Code'].apply(geocode)
```

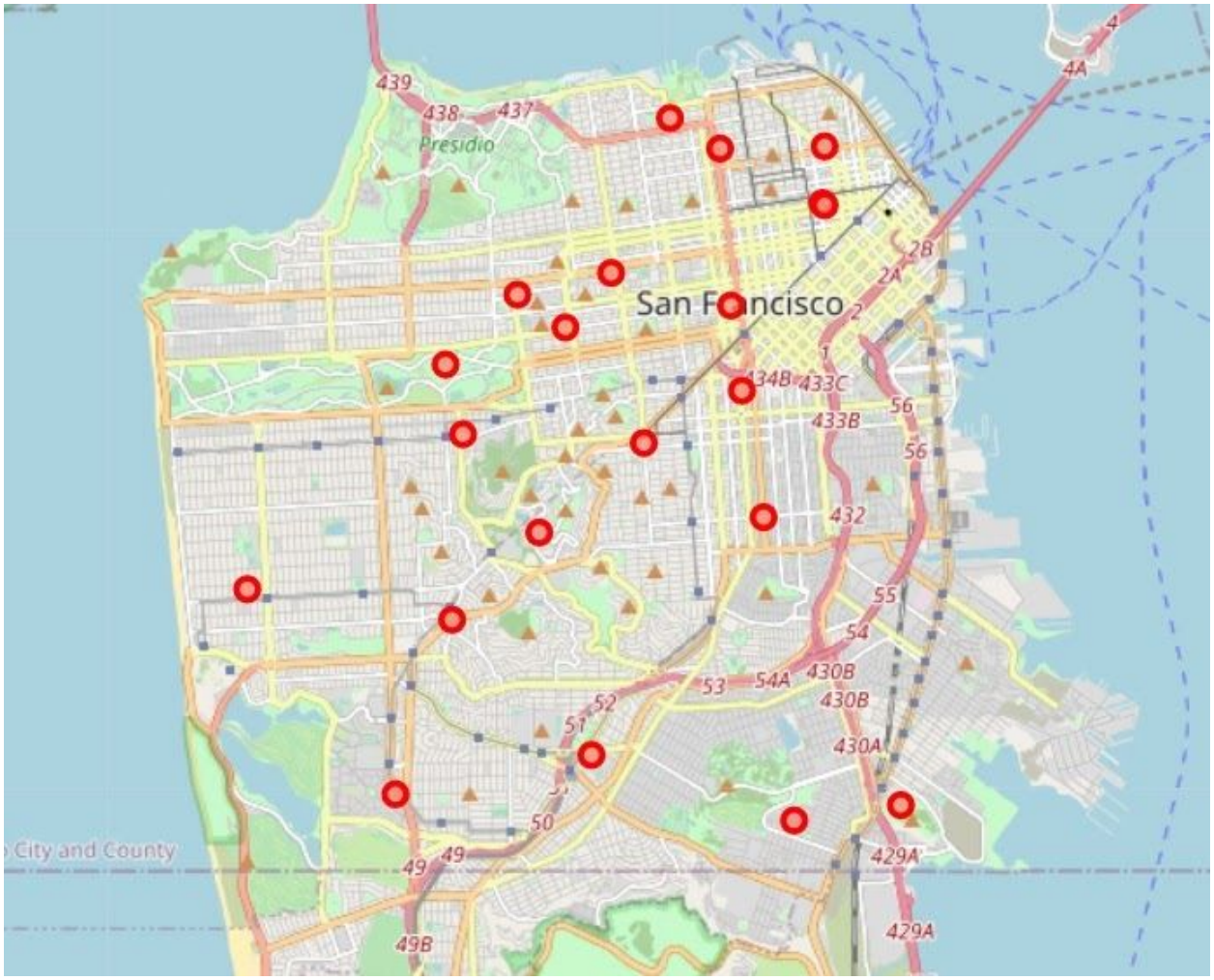
Resulting table:

```
sf_neighborhoods
```

	Zip Code	Neighborhood	latitude	longitude	location
0	94102	Hayes Valley/Tenderloin/North of Market	48.511538	38.620331	(Тополь, Брянка, Брянківська міська рада, Луга...
1	94103	South of Market	37.768063	-122.419700	(San Francisco, San Francisco City and County,...
2	94107	Potrero Hill	37.791436	-122.406974	(San Francisco, San Francisco City and County,...
3	94108	Chinatown	37.791074	-122.406559	(San Francisco, San Francisco City and County,...
4	94109	Polk/Russian Hill (Nob Hill)	37.798012	-122.422964	(San Francisco, San Francisco City and County,...
5	94110	Inner Mission/Bernal Heights	37.752172	-122.416104	(San Francisco, San Francisco City and County,...
6	94112	Ingelside-Excelsior/Crocker-Amazon	37.722630	-122.443304	(San Francisco, San Francisco City and County,...
7	94114	Castro/Noe Valley	37.761403	-122.435242	(San Francisco, San Francisco City and County,...
8	94115	Western Addition/Japantown	37.782757	-122.440178	(San Francisco, San Francisco City and County,...
9	94116	Parkside/Forest Hill	48.688171	13.481367	(Lenzingerberg, Hutthurm, Landkreis Passau, Ba...
10	94117	Haight-Ashbury	37.775865	-122.447288	(San Francisco, San Francisco City and County,...
11	94118	Inner Richmond	37.779911	-122.454970	(San Francisco, San Francisco City and County,...
12	94121	Outer Richmond	37.771355	-122.466327	(San Francisco, San Francisco City and County,...
13	94122	Sunset	37.762451	-122.463423	(San Francisco, San Francisco City and County,...
14	94123	Marina	37.801901	-122.430807	(San Francisco, San Francisco City and County,...
15	94124	Bayview-Hunters Point	37.716300	-122.394562	(San Francisco, San Francisco City and County,...

Unfortunately, two of the zip codes were incorrectly identified. I manually corrected the table with correct latitude and longitude values.
After that I dropped the column with location as we won't need it for further analysis.

2.1.5. Based on the final file, I created the neighborhoods map for San Francisco neighborhoods:



2.2. Downloading information about venues in both cities using Foursquare API

```
CLIENT_ID = cr.CLIENT_ID
CLIENT_SECRET = cr.CLIENT_SECRET
VERSION = cr.VERSION
```

```
neighborhood_latitude = manhattan_data.loc[0, 'Latitude']
neighborhood_longitude = manhattan_data.loc[0, 'Longitude']
neighborhood_name = manhattan_data.loc[0, 'Neighborhood']
```

```
LIMIT = 100
radius = 500
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    neighborhood_latitude,
    neighborhood_longitude,
    radius,
    LIMIT)
```

```
results = requests.get(url).json()
```

Following the process from the Lab session, I used one-hot encoding and then created a table with 10 most common venues for each neighborhood in both cities.

Manhattan:

	Neighborhood	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	Battery Park City	Park	Hotel	Gym	Coffee Shop	Memorial Site	Plaza	Playground	Boat or Ferry	Food Court	Burger Joint
1	Carnegie Hill	Coffee Shop	Café	Yoga Studio	Wine Shop	Italian Restaurant	Gym / Fitness Center	Gym	French Restaurant	Pizza Place	Bookstore
2	Central Harlem	African Restaurant	Chinese Restaurant	Bar	Seafood Restaurant	American Restaurant	French Restaurant	Cosmetics Shop	Caribbean Restaurant	Fried Chicken Joint	Café
3	Chelsea	Coffee Shop	American Restaurant	Art Gallery	Bakery	Ice Cream Shop	Italian Restaurant	French Restaurant	Market	Bookstore	Japanese Restaurant
4	Chinatown	Chinese Restaurant	Bakery	Cocktail Bar	American Restaurant	Dessert Shop	Hotpot Restaurant	Ice Cream Shop	Optical Shop	Noodle House	Salon / Barbershop

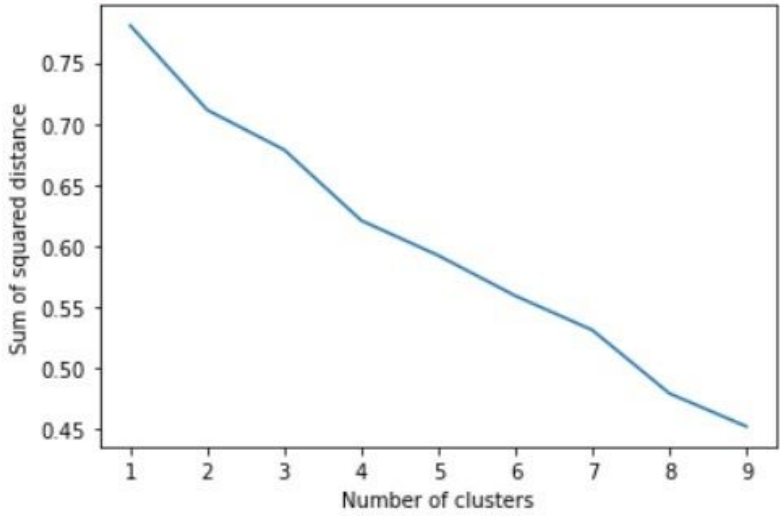
San Francisco:

	Neighborhood	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	Bayview-Hunters Point	Breakfast Spot	Park	Mountain	Scenic Lookout	Light Rail Station	Exhibit	Eye Doctor	Falafel Restaurant	Farmers Market	Fast Food Restaurant
1	Castro/Noe Valley	Gay Bar	Coffee Shop	Thai Restaurant	New American Restaurant	Yoga Studio	Seafood Restaurant	Gym / Fitness Center	Indian Restaurant	Japanese Restaurant	Mediterranean Restaurant
2	Chinatown	Hotel	Coffee Shop	Clothing Store	Boutique	Jewelry Store	Sushi Restaurant	French Restaurant	Lounge	Men's Store	Electronics Store
3	Haight-Ashbury	Café	Coffee Shop	Bank	Salon / Barbershop	Supermarket	Massage Studio	Mexican Restaurant	Middle Eastern Restaurant	Deli / Bodega	Yoga Studio
4	Hayes Valley/Tenderloin/North of Market	Sushi Restaurant	Wine Bar	Coffee Shop	Boutique	French Restaurant	Cocktail Bar	Clothing Store	Optical Shop	Pizza Place	Dessert Shop

2.3. Next I performed clustering of the neighborhoods.

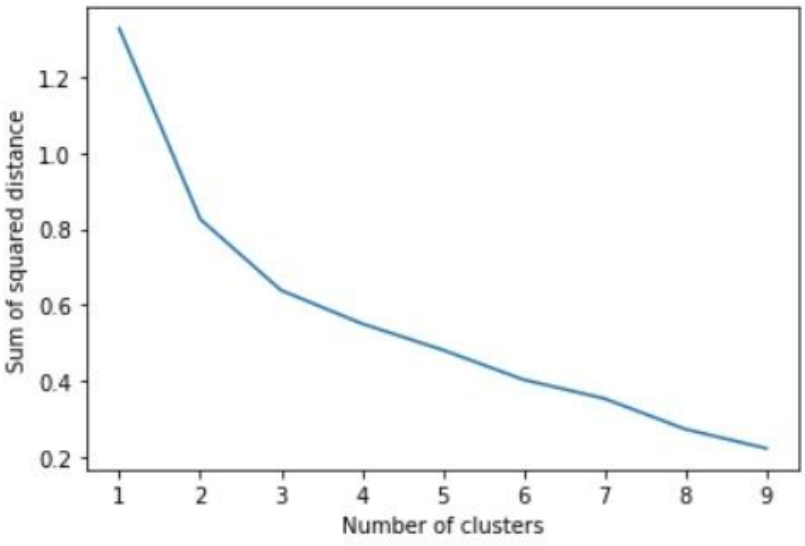
2.3.1. First, determine the optimal number of clusters for the cities using the elbow method:

Manhattan:



Based on the graph, 7 clusters are optimal for Manhattan.
Please, not, that if you run the above algorithm on different days, the results can change because of the Foursquare request limits. When I was working on the dataset, 7 clusters provided the reasonable separation of neighborhoods. When you increase the number of clusters, the algorithm separates single neighborhoods into individual clusters, and it doesn't add useful information.

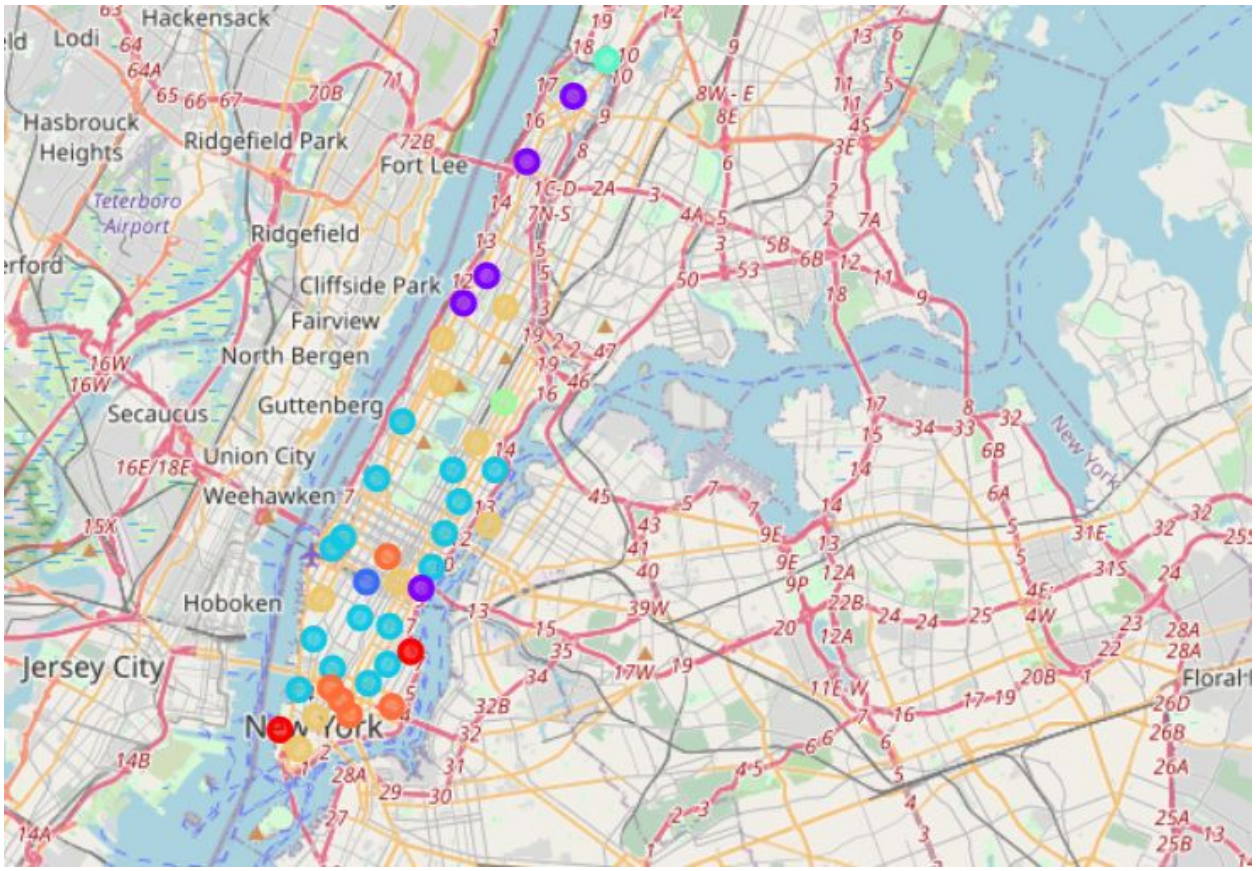
San Francisco



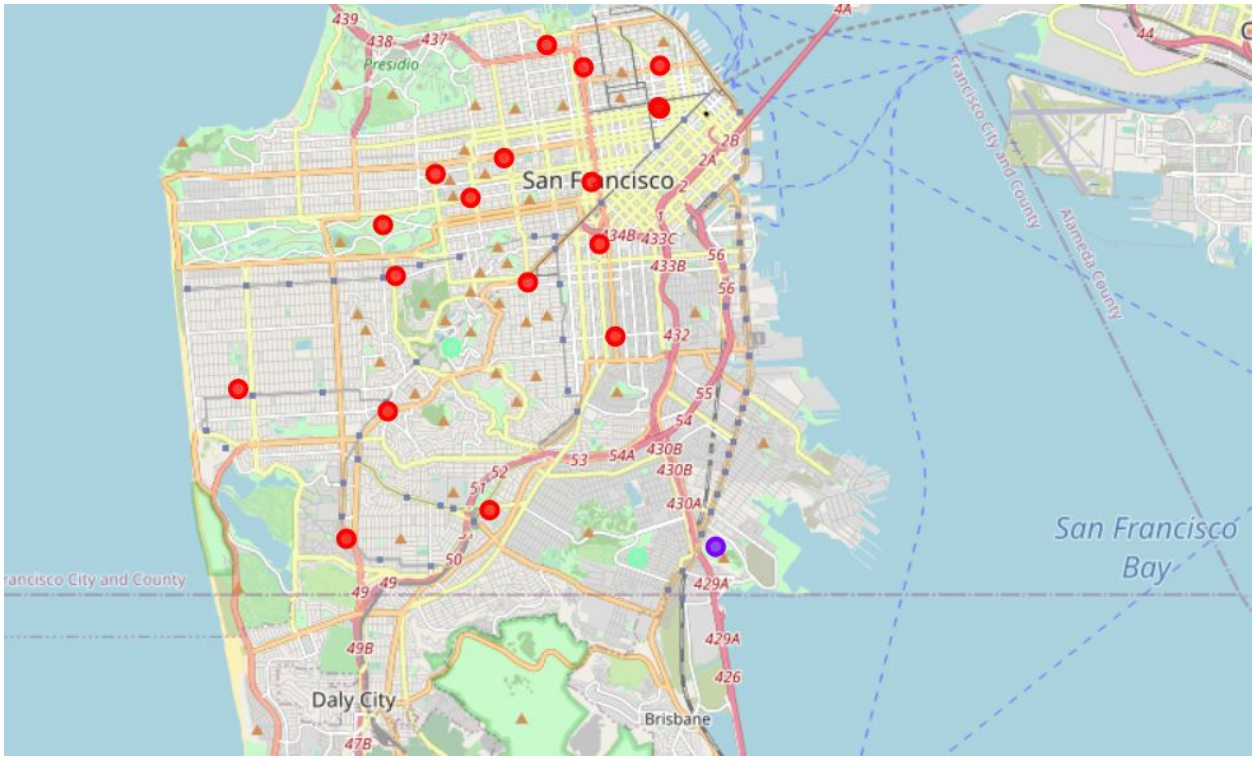
Based on the elbow method, 3 is an optimal number of clusters for San Francisco.

2.3.2. Mapping the clusters with folium

Map of Clusters in Manhattan:



Map of Clusters in SF:



By examining these maps, we can already see that Manhattan in New York is more diverse than San Francisco. First, we have more clusters of neighborhoods in Manhattan. Second, In San Francisco most neighborhoods are very similar and belong to one large cluster and only 3 neighborhoods for the remaining 2 clusters, while clusters in Manhattan are represented with several neighborhoods.

Let's further explore venue diversity in the cities.

2.4. Venues diversity comparison in both cities.

Manhattan:

```
manhattan_venues.shape[0]
```

```
3219
```

```
manhattan_venues['Venue Category'].nunique()
```

```
325
```

San Francisco:

```
sf_venues.shape[0]
```

```
1053
```

```
sf_venues['Venue Category'].nunique()
```

```
247
```

As we see, there are a lot less venues(**1053 vs 3219** in Manhattan only) and also less unique venue categories in San Francisco (**247 vs 325** in Manhattan). This confirms our findings from the clustering analysis that that there is less diversity in San Francisco compared to New York.

2.5 Create a dataframe with the most common venues in Manhattan and SF and visualize it with Word Cloud:

I created the table with all venue categories in Manhattan with the count of these venues. However, we have to remember that there are limits on Foursquare requests and in reality there will be more venues. The information we have is still enough to analyse the diversity of the cities.
After that, I created a Wordcloud to quickly visualize the most popular places in both cities. Below is the code executed for Manhattan, a similar approach was for SF venues:

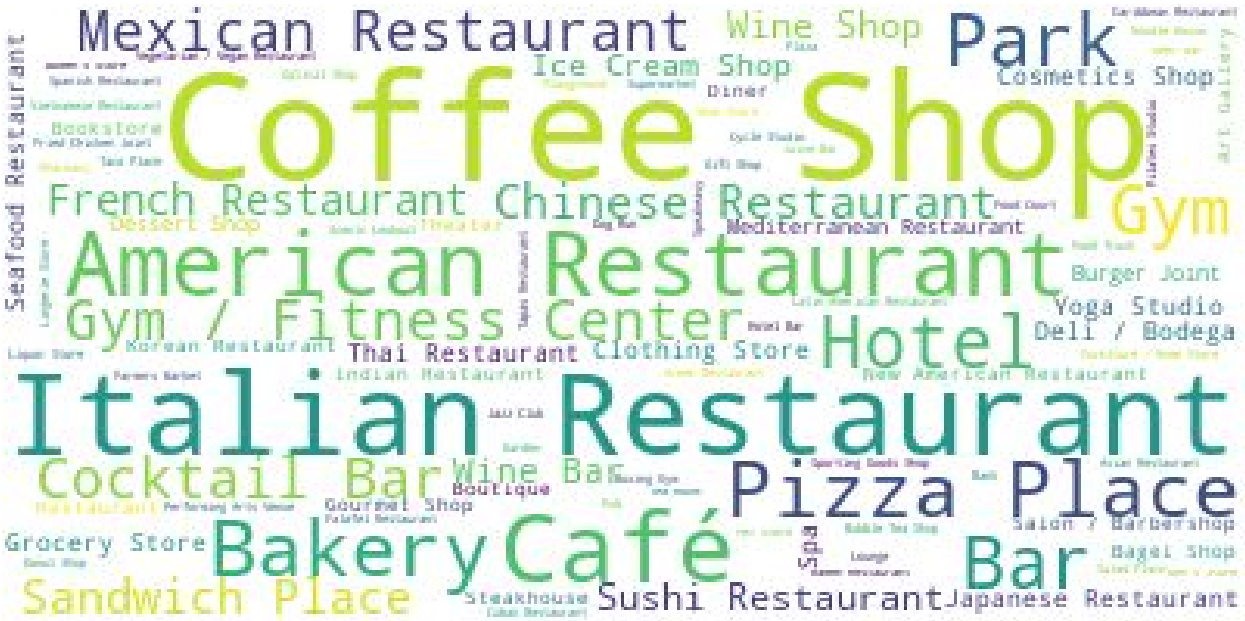
```
manhattan_venue_count = manhattan_venues \
    .groupby('Venue Category', as_index = False) \
    .agg({'Venue': 'count'}) \
    .rename(columns = {'Venue': 'Venue Count'}) \
    .sort_values('Venue Count', ascending = False) \
    .rename(columns = {'Venue Count' : 'Manhattan Venue Count'}) \
    .reset_index(drop = True)
```

WordCloud code:

```
manhattan = {}
for category, count in manhattan_venue_count.values:
    manhattan[category] = count
```

```
wordcloud = WordCloud(max_font_size=50, max_words=100, background_color="white")
wordcloud.generate_from_frequencies(frequencies=manhattan)
plt.figure()
plt.figure(figsize=[16,8])
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
```

Manhattan WordCloud:



San Francisco WordCloud:

Interesting finding, coffee shops are very popular in both cities. However, other venues are different in levels of popularity (for example, Mexican places and Sushi restaurants are a lot more popular in SF than in Manhattan).

2.6. Finally, we will explore which cities are unique to Manhattan and San Francisco.

2.6.1. First, we create a combined file for comparison:

```
venues_comparison = manhattan_venue_count.merge(sf_venue_count, how = 'outer', on='Venue Category')
```

```
venues_comparison.head()
```

	Venue Category	Manhattan Venue Count	SF Venue Count
0	Coffee Shop	139.0	52.0
1	Italian Restaurant	129.0	21.0
2	Café	86.0	23.0
3	American Restaurant	77.0	10.0
4	Pizza Place	76.0	27.0

2.6.2. Export a list of venues that are present in San Francisco but missing in Manhattan:

Output:

'Science Museum', 'Bus Stop', 'Church', 'Light Rail Station', 'Field', 'Brewery', 'Motel', 'Road', 'Neighborhood', 'College Gym', 'Lake', 'Comic Shop', 'Track', 'Storage Facility', 'Planetarium', 'Street Food Gathering', 'Sicilian Restaurant', 'Skating Rink', 'Tanning Salon', 'Tuscan Restaurant', 'Souvlaki Shop', 'ATM', 'Physical Therapist', 'Cultural Center', 'Car Wash', 'Business Service', 'Bus Line', 'Burmese Restaurant', 'Botanical Garden', 'Aquarium', 'Gluten-free Restaurant', 'Parking', 'Pakistani Restaurant', 'Nabe Restaurant', 'Music Store', 'Mountain', 'Motorcycle Shop', 'Metro Station', 'Marijuana Dispensary', 'Hill', 'Herbs & Spices Store'

Some venue categories are missing because we analyzed Manhattan only and not the entire New York. In addition, some venues missing in New York are probably due to the limits in Foursquare API calls (as there are ATMs, Churches, Bus stops, Road, lakes, Physical Therapists, parkings and car washes in Manhattan). However, there are a few distinct categories that are missing in Manhattan, such as Nabe Restaurant, Planetarium, Skating Rink(that will change is we run this research in winter), Burmese Restaurant and a few others).

2.6.3. Export a list of venues that are present in Manhattan but missing in San Francisco:

Output:

'Speakeasy', 'Caribbean Restaurant', 'Cuban Restaurant', 'Fried Chicken Joint', 'Food Court', 'Snack Place', 'Turkish Restaurant', 'Hotpot Restaurant', 'Boat or Ferry', 'Rock Club', 'Smoke Shop', 'Filipino Restaurant', 'School', 'Memorial Site', 'Residential Building (Apartment / Condo)', 'Hawaiian Restaurant', 'Australian Restaurant', 'Tattoo Parlor', 'Supplement Shop', 'Malay Restaurant', 'Office', 'Pet Café', 'Pet Service', 'Video Game Store', 'Tailor Shop', 'Bike Rental / Bike Share', 'Flea Market', 'Bistro', 'Bridal Shop', 'Bridge', 'Discount Store', 'Design Studio', 'Club House', 'Israeli Restaurant', 'Kebab Restaurant', 'Cooking School', 'Shopping Mall', 'Bike Trail', 'Skate Park', 'Soup Place', 'Soccer Field', 'Taiwanese Restaurant', 'Creperie', 'College Theater', 'Record Shop', 'Paella Restaurant', 'Climbing Gym', 'Sports Club', 'Udon Restaurant', 'Lebanese Restaurant', 'Library', 'Karaoke Bar', 'Wings Joint', 'Hobby Shop', 'Heliport', 'Hardware Store', 'Used Bookstore', 'Athletics & Sports', 'Basketball Court', 'German Restaurant', 'Train Station', 'Tennis Stadium', 'Theme Restaurant', 'South Indian Restaurant', 'Circus', 'Check Cashing Service', 'Tech Startup', 'Waterfront', 'Tourist Information Center', 'Volleyball Court', 'Cajun / Creole Restaurant', 'Cafeteria', 'Strip Club', 'Board Shop', 'Veterinarian', 'Venezuelan Restaurant', 'Baby Store', 'Auditorium', 'Social Club', 'Kosher Restaurant', 'Soba Restaurant', 'Moroccan Restaurant', 'Outdoors & Recreation', 'North Indian Restaurant', 'General

Entertainment', 'Non-Profit', 'Golf Course', 'Gym Pool', 'Gymnastics Gym', 'Harbor / Marina', 'Moving Target', 'Molecular Gastronomy Restaurant', 'Food Stand', 'High School', 'Hookah Bar', 'Medical Center', 'Leather Goods Store', 'Gaming Cafe', 'Financial or Legal Service', 'College Academic Building', 'Laundry Service', 'College Arts Building', 'College Bookstore', 'College Cafeteria', 'Comfort Food Restaurant', 'Community Center', 'Scandinavian Restaurant', 'Coworking Space', 'Czech Restaurant', 'River', 'Rest Area', 'Resort', 'Drugstore', 'Duty-free Shop', 'Empanada Restaurant', 'Pier', 'Pie Shop', 'Piano Bar', 'Daycare'

Looking at the missing venues in San Francisco we see a lot of various cuisines that we can enjoy in Manhattan, New York, but cannot find in San Francisco.

In addition, some other interesting venue categories are missing, such as: Climbing Gym, Golf course, Piano bar, Gaming cafe, Flea market, Heliport and some others.

If some of these venue categories are popular in New York, business owners might consider opening these types of venues in San Francisco as well to offer more diversity in the city and expand options for people living and traveling there.

3. Results

We analysed neighborhood venues in Manhattan, New York and San Francisco, California. The goal of the study was to determine whether both cities have equally diverse venues.

By performing clustering analysis on neighborhoods in both cities, we learned that Manhattan is more diverse than San Francisco. There are more neighborhood clusters that are different from each other. In addition, clusters are formed with several neighborhoods, while in San Francisco almost all neighborhoods formed one large cluster and the remaining 3 neighborhoods formed the remaining 2 clusters.

Furthermore, there are more venues that are unique to Manhattan and are not present in San Francisco. It also proves that Manhattan has more diverse venues.

4. Discussion

Our analysis can be further improved if we collect information about all venues in both cities. As some neighborhoods reached the 100 API calls limit, we are certainly missing some data.

To further improve our analysis, we can expand it to all boroughs in New York and include the nearby areas in San Francisco as a lot of people live in suburbs and not only in San Francisco.

Even the analysis performed on limited data demonstrated a large difference in the number of venue categories in two cities.

For a person who is looking to move to a different city, this analysis is sufficient to estimate the gain or loss in diversity. However, if the business owners are looking to expand their venue category presence in another city, it is important to collect full data about the city.

5. Conclusion

Based on our analysis, New York proved to be a city with diverse culture. We compared it to San Francisco, and learned that Manhattan, only one borough in New York, offers a lot more different venues than San Francisco.
It is an important point for people looking for a new place to live.
Businesses can also use these findings to explore what venues are missing in San Francisco.