Coursera
IBM Applied Data Science Capstone

The Battle of Neighborhoods - final report

1. Introduction

The report will outline how I have attempted to solve the business problem of comparing neighborhood similarity in different cities - in this case New York and Los Angeles. In addition to serving as a reference when comparing data, a “comparability-index” would also allow individuals interested in moving to new countries to quantify how similar or dissimilar other neighborhoods or cities are to their current place of living. Some may be more interested in moving to something familiar (i.e. high “comparability-index”) while others are looking for a change (i.e. low “comparability-index”). My hope is that this report will show the usefulness of such a comparison. The comparability-index can be used on several factors. In this case, I will use it to compare food venues in New York and Los Angeles.

2. Data

The data used for this report include location data extracted using the Foursquare API for the two selected cities - New York and Los Angeles. This data includes information on the venues within a given radius, including the type of category. By comparing the frequency of each category, thereby showing us which venue categories are more frequent in each city.

3. Methodology

The data was gathered using the Foursquare API and loaded into two separate dataframes, see figure 1 below. In order to only include data on venues in the food category, a specific categoryId was fed to the search query.

```
dataframe_ny.head()
```

[7]:

	id	name	categories	referralId	hasPerk	location.lat	location.lng	location.labeledLatLngs	location.distance	location.postalCode	...	locati
0	5095b31be4b0a7ac715f2a6d	L&B Spumoni Gardens at the Barclays Center	[{'id': '4bf58dd8d48988d1ca941735', 'name': 'P...	v-1615276027	False	40.683093	-73.975888	[{'label': 'display', 'lat': 40.68309316821465...	4165	11217	...	
1	579c12cd498e6e815b3acf50	Eataly Downtown	[{'id': '4bf58dd8d48988d1f5941735', 'name': 'G...	v-1615276027	False	40.710075	-74.011976	[{'label': 'display', 'lat': 40.710075, 'lng':...	583	10007	...	10
2	4fa2c66ae4b0dc1d01359ad5	WNYC Cafe	[{'id': '4bf58dd8d48988d128941735', 'name': 'C...	v-1615276027	False	40.726449	-74.004947	[{'label': 'display', 'lat': 40.72644854260348...	1530	10013	...	16 N
3	5f5e1b49af65ef27d803a4f0	The Tyger	[{'id': '4bf58dd8d48988d142941735', 'name': 'A...	v-1615276027	False	40.718835	-73.999480	[{'label': 'display', 'lat': 40.718835, 'lng':...	875	10013	...	1
4	4d669e1e13a6236a71789b68	Sing Wah Sing	[{'id': '4bf58dd8d48988d145941735', 'name': 'C...	v-1615276027	False	40.677846	-73.910819	[{'label': 'display', 'lat': 40.67784554923943...	8923	11233	...	14

Figure 1: Dataframe of New York location data

The two dataframes were then combined into a single one, see figure 2, and desired columns selected. In this case we were primarily interested in the venue category (categories) in order to enable the future data analysis, the location data in terms of latitude and longitude, as well as the state. Additional information such as the venue

name, its address and the city was included as this was deemed relevant.

	name	categories	address	lat	lng	city	state
0	L&B Spumoni Gardens at the Barclays Center	Pizza Place	NaN	40.683093	-73.975888	Brooklyn	NY
1	Eataly Downtown	Gourmet Shop	101 Liberty St	40.710075	-74.011976	New York	NY
2	WNYC Cafe	Cafeteria	160 Varick St, New York, NY 10013	40.726449	-74.004947	New York	NY
3	The Tyger	Asian Restaurant	1 Howard St	40.718835	-73.999480	New York	NY
4	Sing Wah Sing	Chinese Restaurant	143 Rockaway Ave	40.677846	-73.910819	Brooklyn	NY
...	...	...	...	...	...	...	...
45	Starbucks	Coffee Shop	1742 S. La Cienega	34.046120	-118.376188	Los Angeles	CA
46	Soho Warehouse	Restaurant	1000 S Santa Fe Ave	34.031513	-118.229770	Los Angeles	CA
47	Chick-fil-A	Fast Food Restaurant	6750 W Sunset Blvd	34.097757	-118.338254	Los Angeles	CA
48	By Chloe	Vegetarian / Vegan Restaurant	2520 Glendale Blvd	34.102877	-118.258451	Los Angeles	CA
49	Starbucks	Coffee Shop	1258 S La Brea Ave	34.050983	-118.344030	Los Angeles	CA

Figure 2: Combined dataframe with filtered columns

The data was plotted on a map, see figure 3, to verify that there weren't any issues (e.g. wrong city).

Figure 3: Map with data points

As Foursquare appears to use inconsistent naming for the state (i.e. both NY and New York), we do some further data cleaning. We finally create a pivot table where we have filtered all the categories where the difference in number of venues per category is greater than one, see figure 4.

state	CA	NY
categories		
Bakery	0.0	3.0
Café	5.0	1.0
Coffee Shop	14.0	8.0
Fast Food Restaurant	6.0	0.0
Ice Cream Shop	0.0	2.0
Shopping Mall	3.0	0.0

Figure 4: Cleaned pivot table

4. Results

Visualizing the results, see figure 5, we see that there are six categories where there is a big difference in the number of venues. New York appears to have more of a sweet tooth as bakeries and ice cream shops are more frequent, while Los Angeles has a greater number of café's, fast food restaurants and coffee shops.

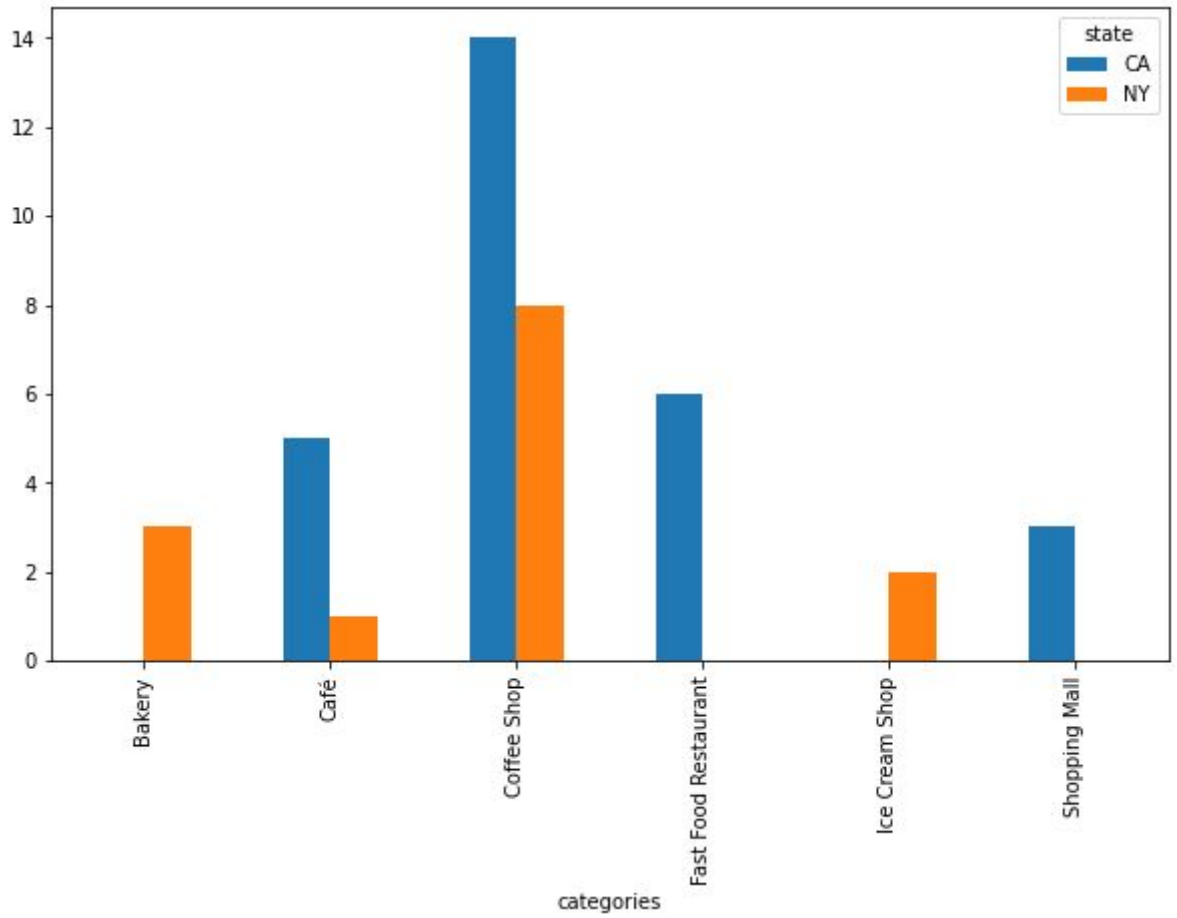

Figure 5: Visualization of results

6. Discussion

During this project, I found that creating this comparability index is more challenging than initially thought. There are a multitude of different dimensions to consider when comparing cities and one is probably best suited to selecting one which one cares most about, rather than trying to do everything. For example, if you care about food, a similar exercise to the one conducted above could be useful, while if you care about e.g. architecture, other categories are of interest.

Furthermore, there is likely a need to further investigate Foursquare's categorization. For example, some categories are quite similar (e.g. japanese food and sushi) but still appear as separate categories. Ideally, one would go through all categories and lump them together to avoid mis-categorization. However, this might also be quite subjective and thus not ideal.

7. Conclusion

This project aimed at creating a comparability-index between New York and Los Angeles in terms of their food venues. Although I was far away from actually achieving this, I believe I made relatively good progress in at least identifying which categories of food are most different between the two cities. Further studies can likely build and improve on this work-