Moving from Indore to Raleigh: Relocating with Data Science

09.07.2020

—

Anmol Lunavat

IBM Applied Data Science Capstone

1. Introduction

Each year there are a number of people who move from one place to another for variety of reasons such as a new job, education, or retirement. Moving to a new city though exciting comes with many challenges. Since we are accustomed to the place and neighborhood we have been staying in, it is easier to make this shift if the new city has a familiar environment. There are multiple factors which constitute in deciding whether two cities are similar such as food, transportation, climate, culture etc. Though, factors such as climate or culture aren't under our control, we can analyze our new city based on food preferences and venues. Therefore, it will be beneficial to predict similar neighborhoods between two cities based on their venues.

Venues surrounding a neighborhood can be analyzed on location and category. Additionally, venue categories can be used to classify similar neighborhoods together. The aim of this project is to predict neighborhoods in Raleigh which are similar to neighborhoods Vijay Nagar and Old Palasia in Indore. The project implements a clustering method which utilizes different categories of venues to group neighborhoods in Raleigh.

The findings of this project will interest people planning to move from Indore to the city of Raleigh.

2. Data Collection and Strategy

To analyze the similarity between neighborhoods of Raleigh with Vijay Nagar and Old Palasia in Indore we gathered a list of neighborhoods and venues in Raleigh.

2.1Neighbourhoods

A list of neighborhoods in Raleigh could be found on the Wikipedia page Raleigh, North Carolina Neighborhoods. Python web scraping techniques with URL handling and beautifulsoup packages was utilized to extract the list and convert it into a pandas dataframe. Following this, Python geocoder package was used to collect latitude and longitude values for all the above collected neighborhoods.

2.2Venues

In the second part of data collection we utilized Foursquare API to gather a list of venues in the neighborhoods of Raleigh. Particularly we relied on Places by Foursquare, a database of more than 105 million places worldwide and API services that enable retrieval of location data. Format of the data received was not appropriate for the purpose of analytics. Thus, we employed data wrangling techniques to transform the data into a more appropriate and valuable format.

3. Methodology

In order to find similar neighborhoods in Raleigh we were required to analyze the venues in each neighborhood and cluster them accordingly. Following this, we were able to use the data of venues in Vijay Nagar and Old Palasia to predict which cluster of neighborhoods they belong to.

3.1Neighborhoods in Raleigh

After collecting the data from Wikipedia, we converted it into a Pandas dataframe. There is a total of 105 neighborhoods in the City of Raleigh.

	Neighborhood
0	Anderson Heights
1	Avent West
2	Belvidere Park
3	Battery Heights
4	Bloomsbury
...	...
100	Southgate
101	Swift Creek
102	Trailwood
103	Walnut Creek
104	Wilder's Grove

105 rows × 1 columns

To utilize Foursqaure API and collect a list of venues we would require longitudes and latitudes of the all the neighborhoods. We utilized Geolocator geocode to fetch the required data for each neighborhood.

```
#Run the above code for all neighborhoods in Raleigh
i = 0
for neigh in raleigh_neigh['Neighborhood']:
    location = geolocator.geocode('{}, Raleigh, North Carolina'.format(neigh))
    if(location):
        raleigh_neigh.at[i,'Latitude'] = location.latitude
        raleigh_neigh.at[i,'Longitude'] = location.longitude
    i = i+1
```

Upon analyzing the accuracy of the Geolocator we found out 15 neighborhoods which did not have a latitude or longitude assigned to them. For the state of accuracy, we filled the missing values by looking them up by ourselves and skipped the ones for which the data was not found.

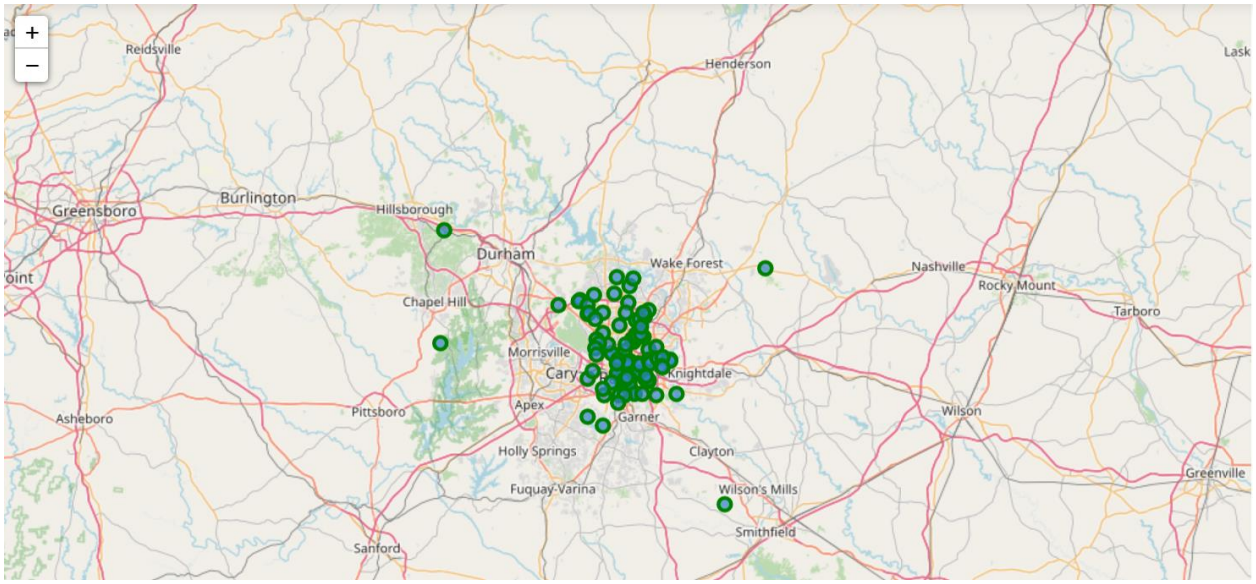
```
#All the null data points have either been filled or removed
raleigh_data.isnull().any()
```

```
Neighborhood    False
Latitude        False
Longitude       False
dtype: bool
```

```
raleigh_data
```

	Neighborhood	Latitude	Longitude
0	Anderson Heights	35.808897	-78.648599
1	Avent West	35.774159	-78.652102
2	Belvidere Park	35.785779	-78.655470
3	Battery Heights	35.790361	-78.660413
4	Bloomsbury	35.809100	-78.649100
...	...	...	...
96	Southgate	35.800649	-78.564618
97	Swift Creek	35.710600	-78.724900
98	Trailwood	35.759300	-78.691600
99	Walnut Creek	35.749400	-78.576300
100	Wilder's Grove	35.798800	-78.564400

101 rows × 3 columns



Upon completion of the above steps our final dataset had a total of 101 neighborhoods from the city of Raleigh.

3.2Venues in the Neighborhoods of Raleigh

To fetch a list of venues in the neighborhoods of Raleigh we utilized Foursqare API. From the result set we collected Venue Name, Category, Latitude and Longitude of the venue.

	Neighborhood Name	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Anderson Heights	35.808897	-78.648599	NOFO @ the Pig	35.805081	-78.646896	Café
1	Anderson Heights	35.808897	-78.648599	Lilly's Pizza	35.805108	-78.646553	Pizza Place
2	Anderson Heights	35.808897	-78.648599	The Third Place Coffeehouse	35.805076	-78.646515	Coffee Shop
3	Anderson Heights	35.808897	-78.648599	Martian Creations @ MFM	35.806227	-78.649612	Arts & Crafts Store
4	Avent West	35.774159	-78.652102	Boulted Bread	35.772535	-78.648729	Bakery

Venue category was an important aspect of the analysis. We used this particular feature to cluster the neighborhoods and find similarities between Raleigh and Indore. While analyzing the result set, we found out that we have a total of 210 unique venue categories in Raleigh.

	Neighborhood Name	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
ATM	1	1	1	1	1	1
Accessories Store	5	5	5	5	5	5
American Restaurant	46	46	46	46	46	46
Antique Shop	6	6	6	6	6	6
Arcade	2	2	2	2	2	2
...	...	...	...	...	...	...
Whisky Bar	1	1	1	1	1	1
Wine Bar	5	5	5	5	5	5
Wine Shop	3	3	3	3	3	3
Women's Store	6	6	6	6	6	6
Yoga Studio	7	7	7	7	7	7

210 rows × 6 columns

Following this, we converted the data into one hot encoded set and grouped them on the basis of neighborhood. While implementing grouping, we took the sum of each category in a particular neighborhood respectively.

	Neighborhood Name	ATM	Accessories Store	American Restaurant	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Warehouse Store
0	Anderson Heights	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0
1	Asbury	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
2	Avent West	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0
3	Battery Heights	0	0	2	1	0	0	0	0	1	...	0	0	0	2	0
4	Belvidere Park	0	0	2	0	0	0	0	0	0	...	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
80	Wayland Heights	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
81	Westlake	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
82	Westover	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
83	Wilder's Grove	0	0	1	0	0	0	0	0	0	...	0	0	0	2	0
84	Woodcrest	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0

85 rows × 211 columns

3.3 Common Venue Categories in Raleigh and Indore

Following a similar process, we fetched the venues in Vijay Nagar and Old Palasia. To create a common clustering mechanism, we analyzed the common venue categories in both data sets. This was done to ensure that the algorithm trained on Raleigh is also able to predict cluster labels for Vijay Nagar and Old Palasia.

```
#Taking out the common venues from both indore and raleigh for analysis
a = set(indore_grouped.columns).intersection(set(raleigh_grouped.columns))
a.add('Neighborhood Name')
a
```

```
{'Café',
 'Chinese Restaurant',
 'Clothing Store',
 'Coffee Shop',
 'Dessert Shop',
 'Fast Food Restaurant',
 'Greek Restaurant',
 'Hot Dog Joint',
 'Hotel',
 'Indian Restaurant',
 'Neighborhood Name',
 'Pizza Place',
 'Plaza',
 'Restaurant',
 'Sandwich Place',
 'Shopping Mall',
 'Snack Place'}
```

Since we wanted to predict cluster labels for Indore, it was important to ensure we were considering only the venue categories which are common between both. Otherwise, an algorithm trained on actual 210 venue categories of Raleigh would have not been able to predict cluster labels for Indore as it had lesser number of categories. For our model to work accurately we require both our data sets to have the same number of features. Therefore, the next step was to remove uncommon venue categories from the data set of Raleigh before we could run our clustering algorithm on this data.

```
raleigh_grouped_cluster = raleigh_grouped[a]

neighname = raleigh_grouped_cluster.pop('Neighborhood Name')
raleigh_grouped_cluster.insert(0, 'Neighborhood Name', neighname)

raleigh_grouped_cluster
```

	Neighborhood Name	Chinese Restaurant	Indian Restaurant	Hotel	Fast Food Restaurant	Dessert Shop	Clothing Store	Coffee Shop	Pizza Place	Shopping Mall	Hot Dog Joint	Plaza	Sandwich Place	Restaurant	Greek Restaurant
0	Anderson Heights	0	0	0	0	0	0	1	1	0	0	0	0	0	0
1	Asbury	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Avent West	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Battery Heights	0	0	1	1	0	1	2	0	0	0	0	2	0	0
4	Belvidere Park	0	0	1	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
80	Wayland Heights	0	0	0	0	0	0	0	0	0	0	0	0	0	0
81	Westlake	0	0	0	0	0	0	0	0	0	0	0	0	0	0
82	Westover	0	0	0	0	0	0	0	0	0	0	0	0	0	0
83	Wilder's Grove	0	0	1	1	0	1	1	0	0	0	0	1	0	0
84	Woodcrest	0	1	0	0	0	0	1	2	0	0	0	1	0	0

```
indore_grouped.head()
```

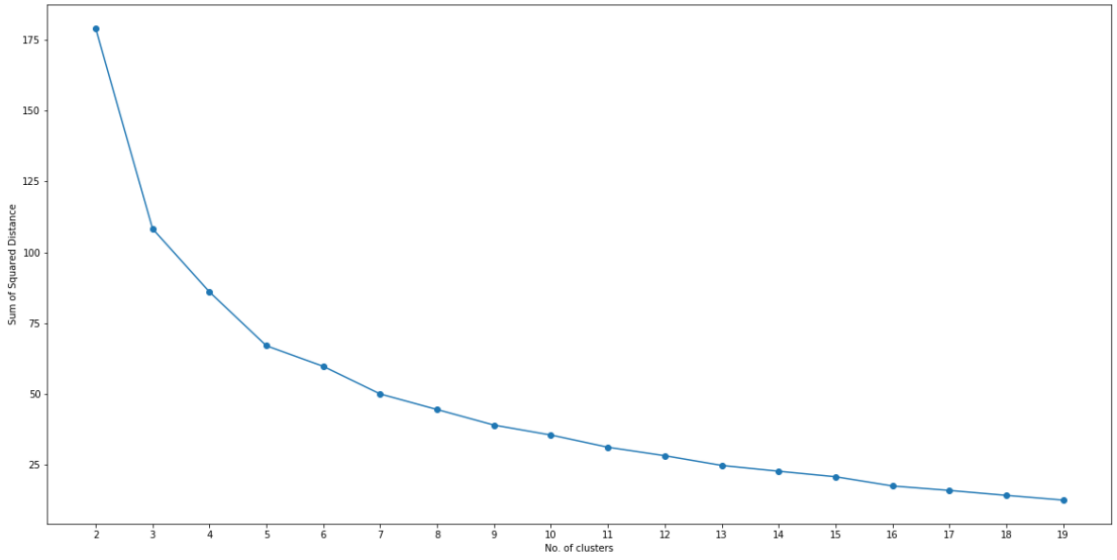
	Neighborhood Name	Café	Chinese Restaurant	Clothing Store	Coffee Shop	Dessert Shop	Fast Food Restaurant	Greek Restaurant	Hot Dog Joint	Hotel	Indian Restaurant	Pizza Place	Plaza	Restaurant	Sandwich Place	Shopp Mall
0	Old Palasia	4	0	1	1	1	1	1	1	1	3	0	2	1	1	1
1	Vijay Nagar	0	1	0	0	0	1	0	0	1	2	1	0	1	0	0

4. Modelling

K-means Clustering algorithm was utilized to cluster the neighborhoods of Raleigh. Clustering the data identified subgroups in the data set such that neighborhoods in the same cluster were very similar while neighborhoods in different clusters were very different. Clustering requires a feature which could be used to find subgroups amongst the data. In our use case, the feature venue categories was used to find best subgroups of neighborhoods. Once we found the common value for the feature venue categories between Indore and Raleigh, we were able to fit the algorithm and cluster neighborhoods of Raleigh.

4.1Optimal number of K

To find the optimal number of clusters for our data we calculated Sum of Squared Distance. We ran for K in range of 2 to 20 and drew an elbow curve to find the best value of K.



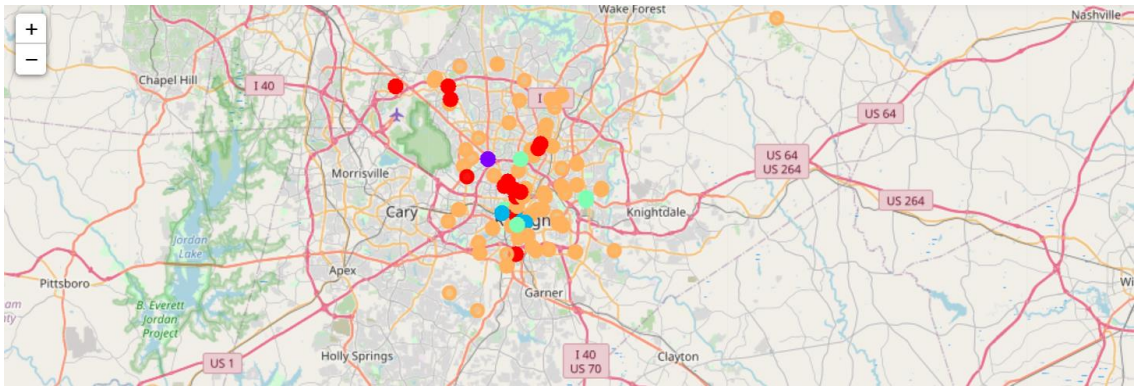

```
#Finding out the Knee point for optimal K
from kneed import KneeLocator
kn = KneeLocator(np.arange(2, max_range), sse, curve='convex', direction='decreasing')
optimalK = kn.knee
print(optimalK)
```

```
5
```

As visualized the knee of the curve occurs when the number of clusters is equal to 5. Thus, we used 5 as the value of K for our algorithm.

4.2 K-means Clustering

We ran K-means on our data and clustered the neighborhoods in Raleigh into 5 different clusters. Each cluster contained neighborhoods which were similar to each other based on the categories of venue it had.



```
#Cluster 0
cluster_0 = raleigh_merged['Neighborhood Name'].loc[raleigh_merged['Cluster Labels'] == 0].unique()
print(cluster_0)
```

```
['Asbury' 'Avent West' 'Belvidere Park' 'Biltmore Hills' 'Bloomsbury'
 'Boylan Heights' 'Brier Creek' 'Brookhaven' 'Capitol Heights'
 'Carolina Pines' 'Coachmans Trail' 'Country Club Hills' 'Crossgate'
 'Dominion Park' 'Durant Trails' 'Falls Church' 'Fayetteville Street'
 'Glenwood South' 'Glenwood Village' 'Glenwood-Brooklyn' 'Hayes Barton'
 'Hedingham' 'Hickory Hills' 'Lake Park' 'Lake Wheeler' 'Laurel Hills'
 'Leesville' 'Longview Gardens' 'Madonna Acres' 'Maiden Lane'
 'Moore Square' 'Mordecai District' 'North Carolina State University'
 'North Hills' 'North Pointe' 'Oak Park' 'Olde Raleigh' 'Parkland'
 'Pinecrest' 'Quail Hollow' 'Quail Ridge' 'Raleigh Country Club'
 'Renaissance Park' 'Rochester Heights' 'Skycrest Village' 'Southall'
 'Springdale/Leesville' 'Stonehenge' 'Summerfield' 'Summit Ridge'
 'Swift Creek' 'Tadlock Plantation' 'Timberlake' 'Trailwood'
 'Trinity Woods' 'Tysonville' 'Umstead' 'University Park' 'Vanguard Park'
 'Victoria Place' 'Village on the Green' 'Walnut Creek' 'Wayland Heights'
 'Westlake' 'Westover']
```

```
#Cluster 1
cluster_1 = raleigh_merged['Neighborhood Name'].loc[raleigh_merged['Cluster Labels'] == 1].unique()
print(cluster_1)
```

```
['Brentwood Estates']
```

```
#Cluster 2
cluster_2 = raleigh_merged['Neighborhood Name'].loc[raleigh_merged['Cluster Labels'] == 2].unique()
print(cluster_2)
```

```
['Capitol District' 'Fairfax Hills' 'Hi-Mount' 'Historic Oakwood']
```

```
#Cluster 3
cluster_3 = raleigh_merged['Neighborhood Name'].loc[raleigh_merged['Cluster Labels'] == 3].unique()
print(cluster_3)
```

```
['Battery Heights' 'Depot District' 'Lakemont' 'South Park' 'Southgate'
 "Wilder's Grove"]
```

```
#Cluster 4
cluster_4 = raleigh_merged['Neighborhood Name'].loc[raleigh_merged['Cluster Labels'] == 4].unique()
print(cluster_4)
```

```
['Anderson Heights' 'Cameron Park' 'Drewry Hills' 'Falls River/Bedford'
 'Five Points Historic Neighborhoods' 'North Ridge'
 'North Ridge Country Club' 'Quail Meadows' 'Roanoke Park' 'Woodcrest']
```

5. Results

We found that most of the neighborhoods either belonged to Cluster 0 or 4. There were 65 neighborhoods in Cluster 0 and 10 in Cluster 4. After our kmeans algorithm was trained on neighborhoods of Raleigh, we utilized it to predict cluster label for Vijay Nagar and Old Palasia to find their similarity with neighborhoods of Raleigh.

```
#Displaying Cluster Labels for Indore Neighborhoods - Vijay Nagar and Old Palasia
indore_grouped_cluster
```

	Neighborhood Name	Cluster Labels	Chinese Restaurant	Indian Restaurant	Hotel	Fast Food Restaurant	Dessert Shop	Clothing Store	Coffee Shop	Pizza Place	Shopping Mall	Hot Dog Joint	Plaza	Sandwich Place	Restaurant	Gre Res
0	Old Palasia	2	0	4	0	0	2	1	1	0	1	1	2	1	1	1
1	Vijay Nagar	0	1	3	1	1	0	0	0	1	0	0	0	0	0	0

While Old Palasia belonged to cluster 2, Vijay Nagar belonged to cluster 0.

Thus, we were able to find neighborhoods similar to both Vijay Nagar and Old Palasia by analyzing the other members of their cluster.

A person residing in Vijay Nagar, Indore will find the similarity in the following neighborhoods if he visits Raleigh:

```
print(similar_neigh_vijay_nagar)
```

```
['Capitol District' 'Fairfax Hills' 'Hi-Mount' 'Historic Oakwood']
```

Whereas, a person residing in Old Palasia, Indore will find the similarity in the following neighborhoods if he visits Raleigh:

```
print(similar_neigh_old_palasia)
```

```
['Anderson Heights' 'Biltmore Hills' 'Bloomsbury' 'Cameron Park'
 'Drewry Hills' 'Falls River/Bedford' 'Five Points Historic Neighborhoods'
 'Laurel Hills' 'North Ridge' 'North Ridge Country Club' 'Quail Meadows'
 'Roanoke Park' 'Vanguard Park' 'Woodcrest']
```

6. Future Work

The current scope of the project only considers two locations from Indore, i.e., Vijay Nagar and Old Palasia. The inclusion of all neighborhoods from Indore will make us capable of comparing two cities and suggesting similar neighborhoods for each neighborhood from Indore. Furthermore, the system could become capable of running between any two or more cities and suggesting similar neighborhoods between them. Once the system is capable of dynamically adapting to multiple cities it could be used as a platform for finding similar neighborhoods. A user would be allowed to enter current city and destination cities. The system would then utilize the current city and suggest similar neighborhoods in the destination cities.