Coursera
Capstone Final Project Report

- HIMA ARE

Restaurant Location Recommend System

Coursera Capstone Final Project – Report

- Hima Are

Apr – 20 - 2020

Report Contents

1. Problem Description

2. Data Section.

3. Evaluation Method to Solve the Problem.

4. Getting and Preparing Data sets.

5. Evaluating the Results.

6. Conclusion.

1. Problem Description

One of My Friend and his Colleague are working in a MNC in DFW area and they want to
enter into small business area by starting a restaurant in DFW Metroplex area,
to provide food service to growing Indian/Asian community in the area.
But they are facing difficulty to identify the right location with the limited resources.
My Friend told me about the problem that they are facing and asked for help in solving their problem
to find out the right location to start their restaurant business and they want to
start an Indo Chinese Fusion restaurant with more Indian flavors.

There are more than 100+ localities based on the zip codes and DFW metro plex is highly diversified area
and in some area Asian population is more and, in some areas, Asian community existence is very low.

As the partners are planning to start the Indochinese fusion restaurant, planning to start the restaurant
where the Asian community is existing.

As the primary targeted customers are Asian community,
We need to identify the location (i.e. city, zip code) where the Asian community existing and with less or no
competition along with other possible customers in DFW area.

2. Data Section.

1) All the cities and zip codes located in Texas state can be captured from
   https://www.zipdatamaps.com/list-of-zip-codes-in-texas.php

df_tx_zip.head()

	Zip Code	Zip Code Type	Zip Code Name	County
0	73301	Unique	Austin	Travis
1	73344	Unique	Austin	Travis
2	73960	NaN	Texhoma	Sherman
3	75001	Non-Unique	Addison	Dallas
4	75002	Non-Unique	Allen	Collin

2) the latitude, longitude and demographic information associated with Asian community can be captured from https://www.zipdatamaps.com/<zip code>

df_dfw_ll_pop.head(3)

	ZipCode	Total_Population	Asian_Population	AsianPop_Percent	Latitude	Longitude
0	75001	12414.0	1284.0	10.343161	32.960049	-96.838417
1	75002	63140.0	5616.0	8.894520	33.091141	-96.606972
2	75006	46364.0	3315.0	7.149944	32.953411	-96.901871

2. Data Section.

3) Getting Venue data.

Planning to use FourSquare "Search Venue" api to get the business/venue information for specific category located in or nearby that zip code.

1) Indian / Indo Chinese Restaurants in that zip code to evaluate the competition, i.e. negative impact

2) Asian Restaurants (i.e. Korean/Thai/Chinese/Japanese/Asian) in that zip code to evaluate the competition, i.e. negative impact.

3) Shopping Places (i.e. Asian Groceries/Groceries/Warehouse/Malls etc.) in that zip code to evaluate the potential buyers to visit the area.

4) Residential Complex in that zip code to evaluate the number of customers in small place (i.e. customer density is more in residential complex areas.) and also possible impact during dinner hours.

5) Offices / Hotels / Universities in that zip code, to evaluate the possible impact during lunch hours.

2. Data Section.

3) Getting Venue data (continued from previous ..).

Following category IDs will be used to search for specific venues in the location.

1) Indian / Indo Chinese Restaurants :

Indian Restaurant :   4bf58dd8d48988d10f941735
South Indian Restaurant :   54135bf5e4b08f3d2429dfde
Indian Chinese Restaurant :   54135bf5e4b08f3d2429dfdf

2) Asian Restaurants :

Korean Restaurant :   4bf58dd8d48988d113941735
Chinese Restaurant :   4bf58dd8d48988d145941735
Thai Restaurant :   4bf58dd8d48988d149941735
Asian Restaurant :   4bf58dd8d48988d142941735
Japanese Restaurant   4bf58dd8d48988d111941735

2. Data Section.

3) Getting Venue data (continued from previous ..).

   3) Shopping Places :

      Grocery Store :   4bf58dd8d48988d118951735 / 4bf58dd8d48988d1f6941735
      Supermarket :   52f2ab2ebcbc57f1066b8b46 / 50be8ee891d4fa8dcc7199a7
      Warehouse Store : 52e816a6bcbc57f1066b7a54
      Shopping Mall :   4bf58dd8d48988d1fd941735

   4) Residential Complex :

      Residential Building (Apartment / Condo) :   4d954b06a243a5684965b473

   5) Offices / Hotels / Universities :

      Office :   4bf58dd8d48988d124941735
      Hotel :   4bf58dd8d48988d1fa931735
      University :   4bf58dd8d48988d1ae941735

3. Evaluation Method to Solve the Problem.

Step-1: Getting cities/localities and Zip codes in DFW area
    1) Get all the cities and Zip code data available in state of Texas.

    2) filter data to select only the zip codes available in DFW area.
        i.e. zip codes available in following counties.
        'Collin', 'Dallas', 'Denton', 'Ellis', 'Hood', 'Hunt', 'Johnson', 'Kaufman', 'Rockwall', 'Somervell', 'Parker', 'Tarrant', 'Wise'

    3) Filter data associated with Mail boxes (i.e. zip codes associated with mail boxes)

step-2: Get the co-ordinates and demographic information available for the zip codes.
    1) use zip code to get following attributes.
        a) Latitude
        b) Longitude
        c) Total Population
        d) Asian Population
        e) Percentage of Asian Population
    2) Filter the data to have minimum existence of Asian Community in the Zip Code.
    in our process we are taking 1000 as minimum number and 3.5% percentage in total population.

3. Evaluation Method to Solve the Problem. (continued..)

step-3 Get the Venue data from foursquare "Search Venues" API for following categories.

1) a) Indian / Indo Chinese Restaurants.

b) Asian Restaurants (i.e. Korean/Thai/Chinese/Japanese/Asian)

c) Shopping Places (i.e. Grocery Store / Department Store / Supermarket / Warehouse / Shopping Mall)

d) Residential Complex (i.e. Residential Building (Apartment / Condo))

e) Offices / Hotels / Universities

2) While getting venues is located with same zip code or if the venue is with different zip code, but if it is available in < 2 km from zip code co-ordinates.

step-4) display of the venues for each category.

3. Evaluation Method to Solve the Problem. (continued..)

step-5) Evaluate the net score for the impact of each category.
    1) Used following weightage to calculate the net score

        a) Indian / Indo Chinese Restaurants: - 1.5 (negative impact due to competition)

        b) Asian Restaurants: -1 (negative impact due to competition possible customer split.)

        c) Shopping Places: 0.5 (positive impact)

        d) Residential Complex: 1.5 (positive impact due to more available customers in small area

            and may impact dinner hours)

        e) Offices / Hotels / Universities: 1.5 (positive impact during lunch hours)

        f) Asian Population: 1 for every 1000 people. (positive impact due to targeted Asian Community)

    2) then aggregate the data set to get the number of venues for each category.

    3) calculate the net weight score using the weightage defined in the above step.

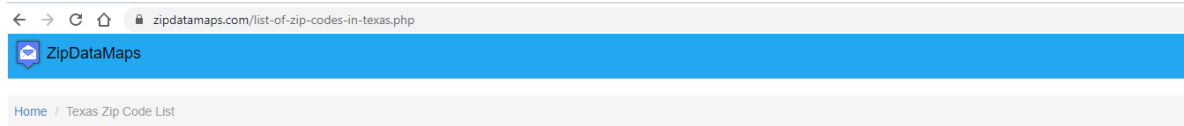   4) sort the dataset using the net weight score in descending order.

3. Evaluation Method to Solve the Problem. (continued..)

step-6) display the possible location with all categories.

1) get the to 2 or 3 records for evaluation and display each record individually with all
   categories in one map.

2) Evaluate and finalize the possible best location with available opportunities to run the business.

4. Getting and Preparing Data Sets

Texas Zip Code data available @ https://www.zipdatamaps.com/list-of-zip-codes-in-texas.php

Read data using read_html() in pandas library.

4. Getting and Preparing Data Sets (continued..)

Latitude and Longitude and Community information is available @ https://www.zipdatamaps.com/<zip code>
ex: https://www.zipdatamaps.com/75063

After capturing latitude, longitude and community data , Data set is merged into Initial data set with City and Zip Code information.

```
: df_dfw_all_data = pd.merge(df_dfw_zip, df_dfw_ll_pop , on='ZipCode' , how='left' )
```

```
: df_dfw_all_data.head()
```

	ZipCode	City	County	Total_Population	Asian_Population	AsianPop_Percent	Latitude	Longitude
0	75001	Addison	Dallas	12414.0	1284.0	10.343161	32.960049	-96.838417
1	75002	Allen	Collin	63140.0	5616.0	8.894520	33.091141	-96.606972
2	75006	Carrollton	Dallas	46364.0	3315.0	7.149944	32.953411	-96.901871
3	75007	Carrollton	Denton	51624.0	7822.0	15.151867	33.009960	-96.896088
4	75009	Celina	Collin	8785.0	61.0	0.694365	33.325680	-96.736870

4. Getting and Preparing Data Sets (continued..)

➢ Latitude and Longitude and Community information is available @ https://www.zipdatamaps.com/<zip code>
    ex: https://www.zipdatamaps.com/75063

➢ After capturing latitude, longitude and community data , Data set is merged into Initial data set with City and Zip Code information.

➢ Final Asian community data set will be prepared by filtering / excluding the data where the Asian population is < 1000 and the percentage of Asian population < 3.5 in total population.

```python
df_dfw_all_data = pd.merge(df_dfw_zip, df_dfw_ll_pop , on='ZipCode' , how='left' )
```

```python
# Create a DataFrame with Asian Community is more than 1000 and >3.5% of Total Population.
df_dfw_asian = df_dfw_all_data[(df_dfw_all_data['Asian_Population']>1000) & (df_dfw_all_data['AsianPop_Percent']>3.5)]
```

```python
df_dfw_asian.head()
```

|	ZipCode	City	County	Total_Population	Asian_Population	AsianPop_Percent	Latitude	Longitude
0	75001	Addison	Dallas	12414.0	1284.0	10.343161	32.960049	-96.838417
1	75002	Allen	Collin	63140.0	5616.0	8.894520	33.091141	-96.606972
2	75006	Carrollton	Dallas	46364.0	3315.0	7.149944	32.953411	-96.901871
3	75007	Carrollton	Denton	51624.0	7822.0	15.151867	33.009960	-96.896088
5	75010	Carrollton	Denton	21607.0	6081.0	28.143657	33.054829	-96.871742

4. Getting and Preparing Data Sets (continued..)

DFW Locality Map with Asian Community existence.
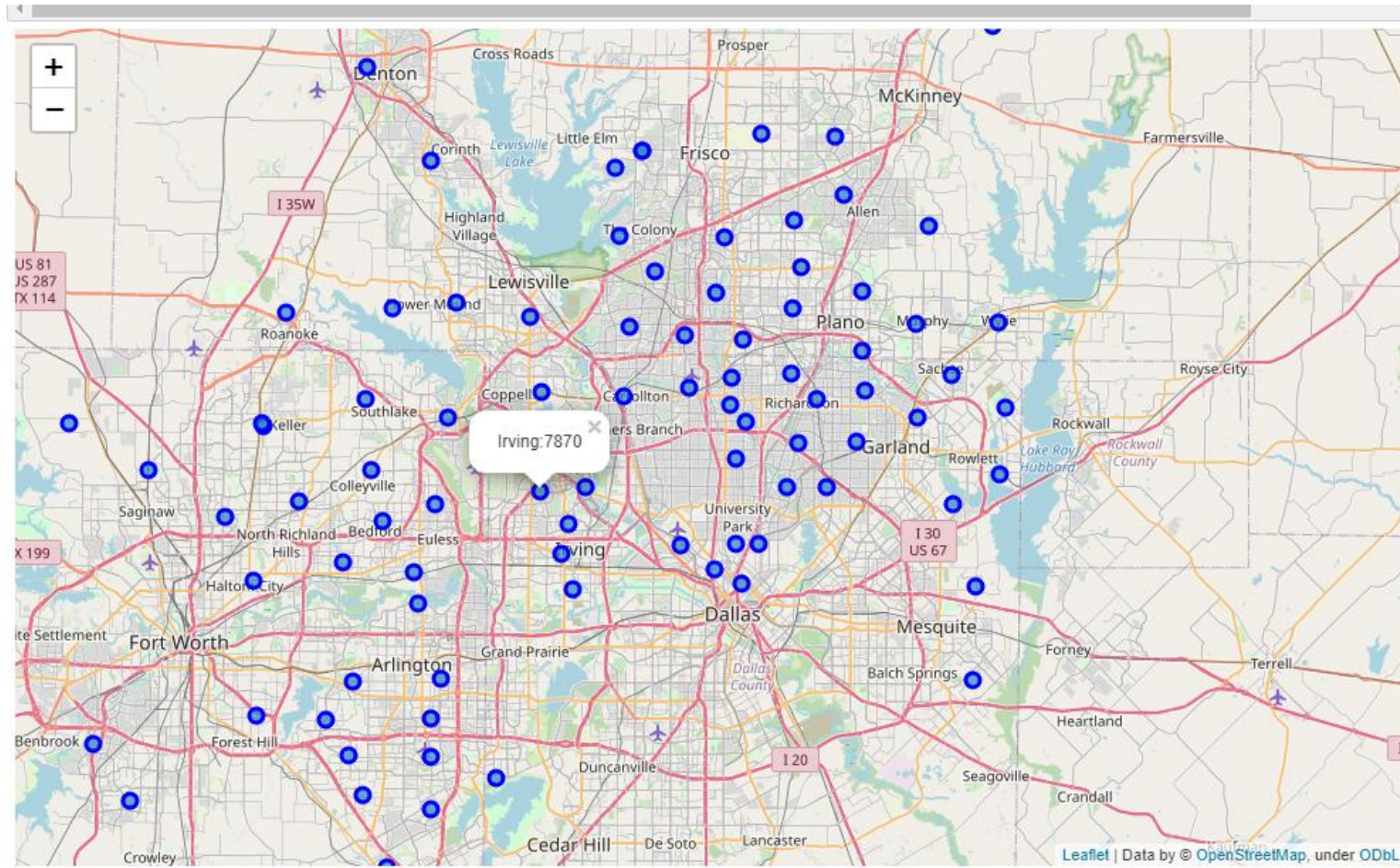
4. Getting and Preparing Data Sets (continued..)

Location of existing Indian restaurants in DFW area. (for the Indian / Indo Chinese Restaurants categories listed in page 7)
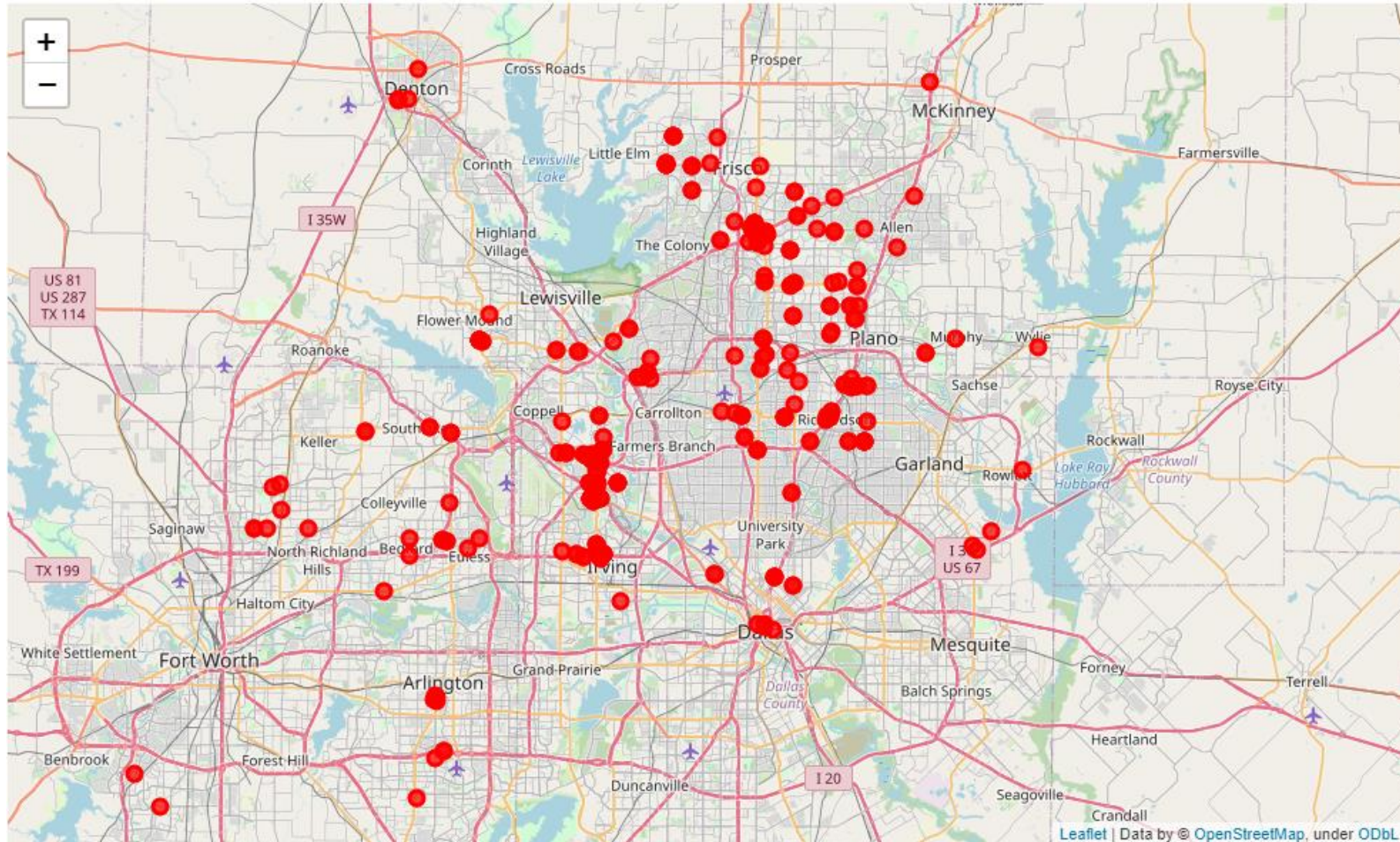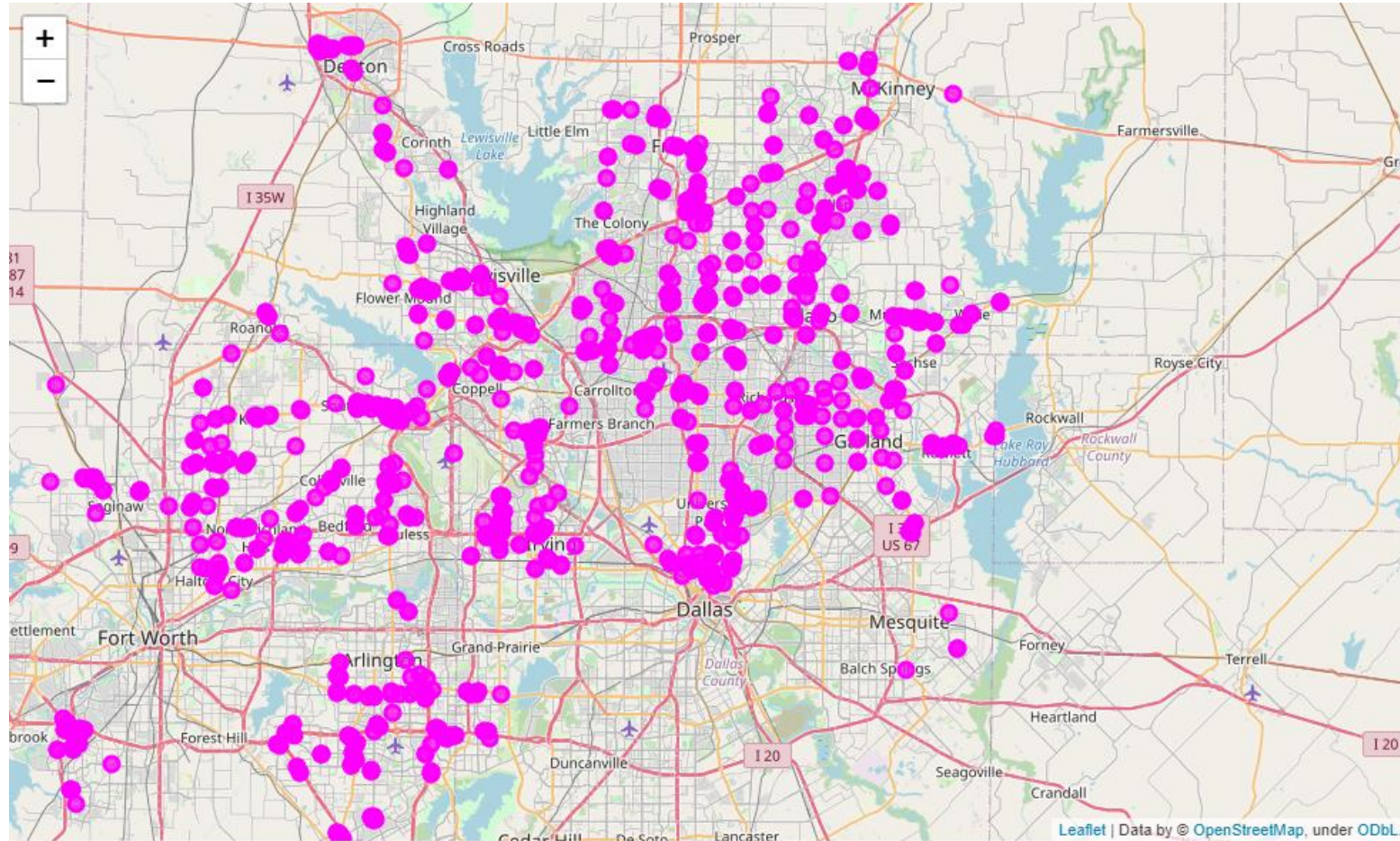
4. Getting and Preparing Data Sets (continued..)

Location of existing Asian restaurants in DFW area. (for the Asian Restaurants categories listed in page 7)
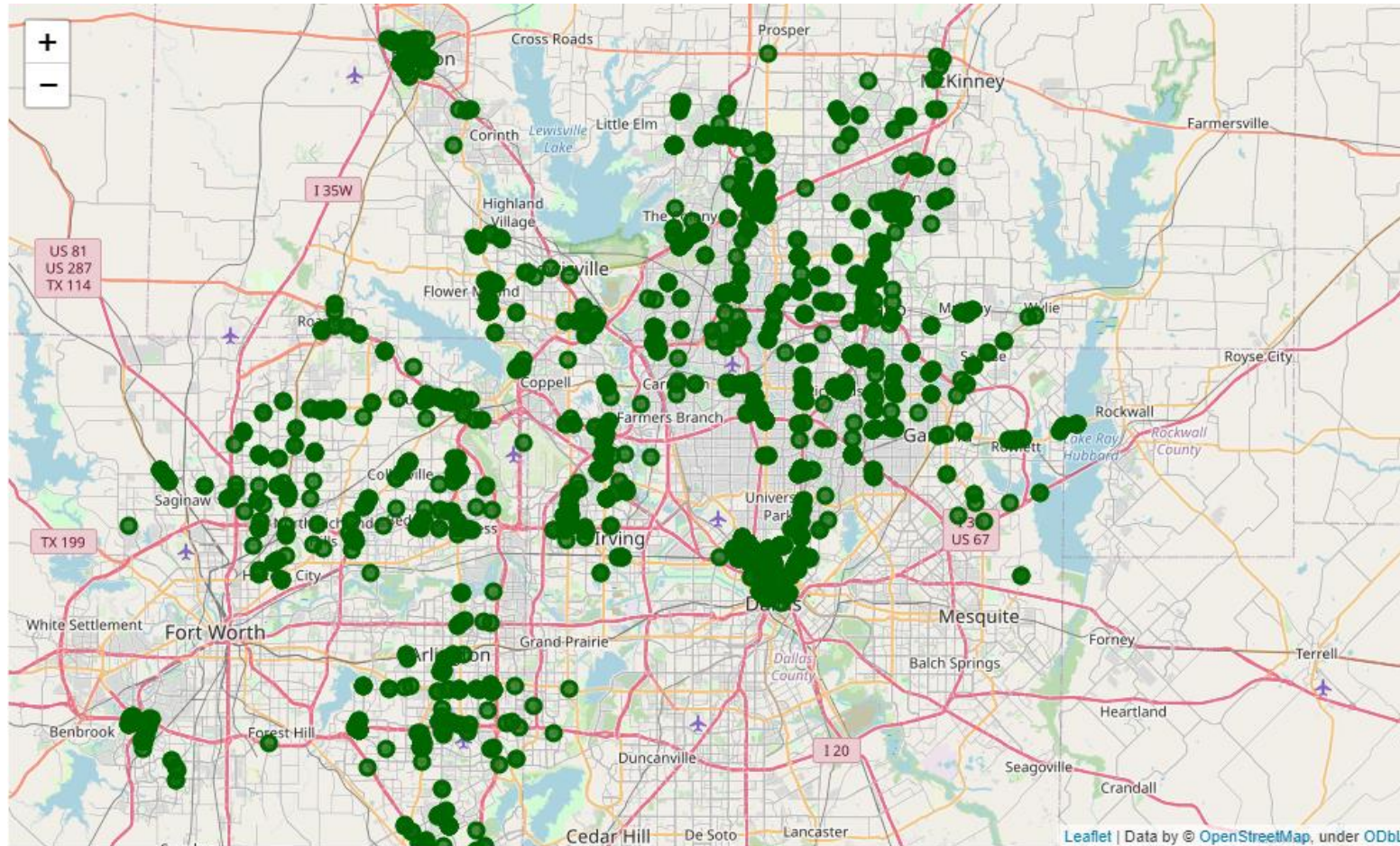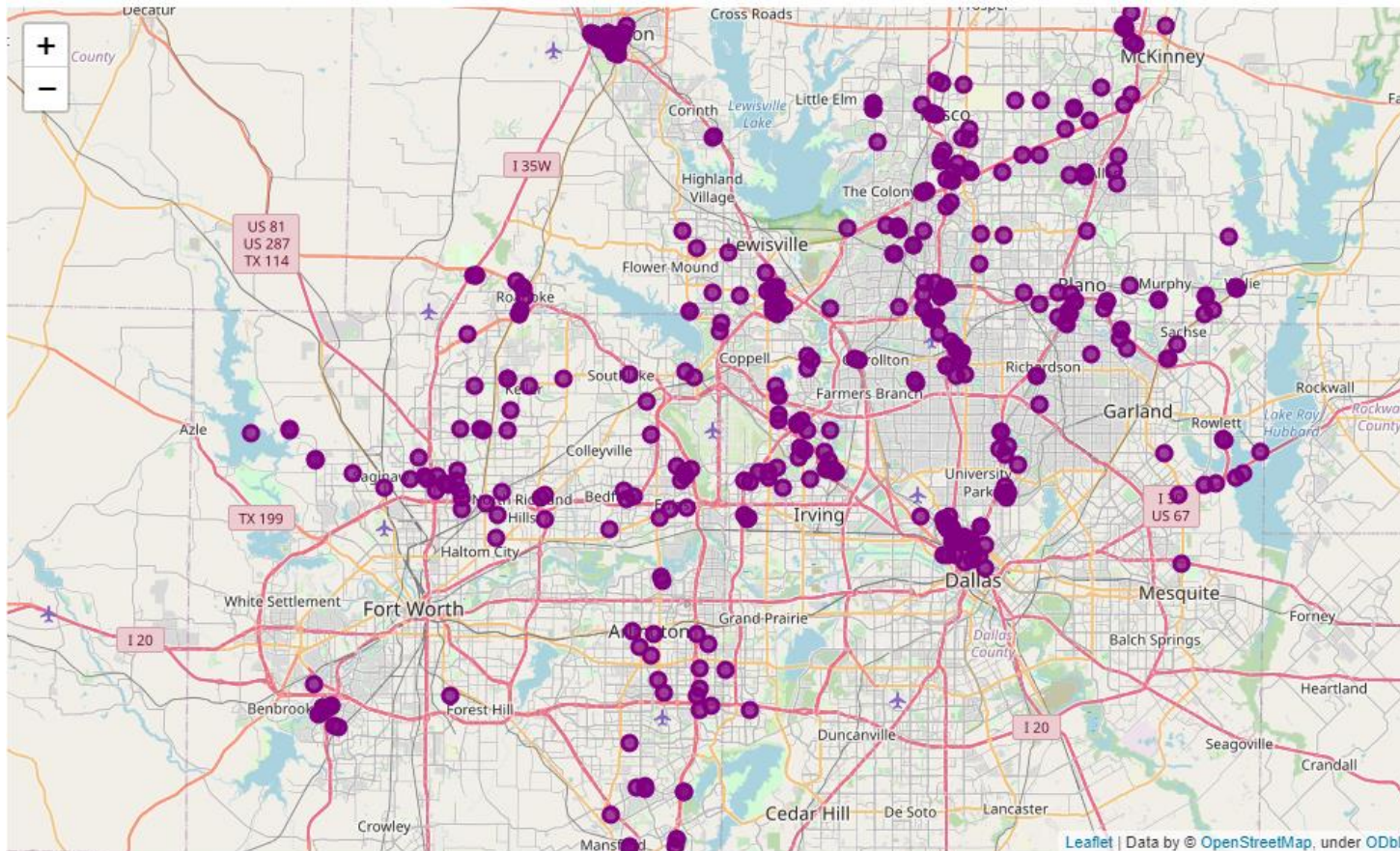
4. Getting and Preparing Data Sets (continued..)

Location of existing Shopping Venues in DFW area. (for the Shopping Places categories listed in page 8)

4. Getting and Preparing Data Sets (continued..)

Location of existing Apartments / Multi Family Residential Complexes in DFW area. (for the Residential Complex categories listed in page 8)

4. Getting and Preparing Data Sets (continued..)

Location of existing Office / Hotel / University venues in DFW area. (for the Office / Hotel / University categories listed in page 8)
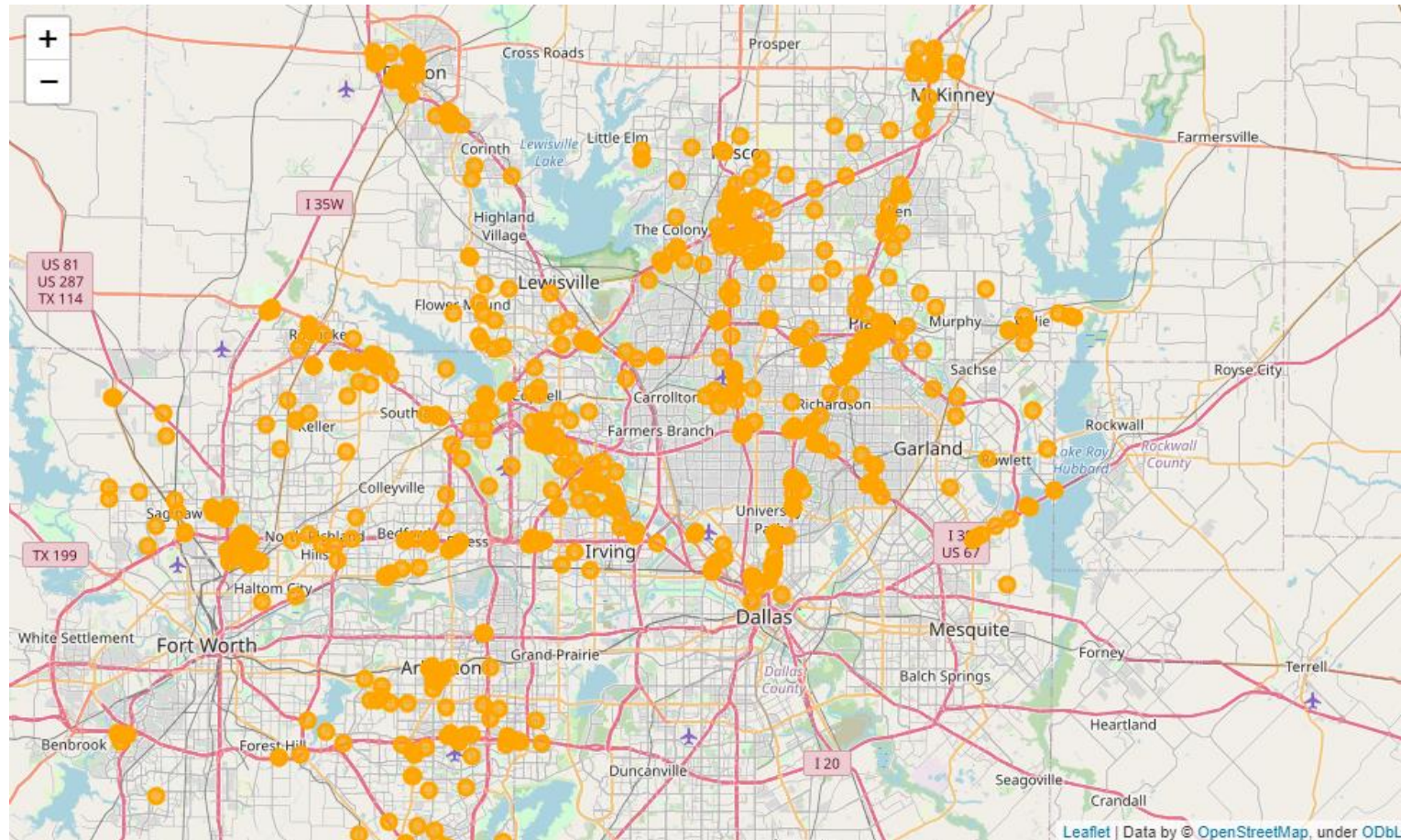
5. Evaluating the Results

After getting the aggregated count for each category group by zip code the data set is as listed below and ready to measure the weight score for each location.

```
df_dfw_ir_weight_score.head(10)
```

	City	ZipCode	County	Asian_Population	Indian_Restaurants	Asian_Restaurants	Shopping_Places	Residential_Complex	Office_Hotels_Places
0	Addison	75001	Dallas	1284.0	2.0	9.0	6.0	5.0	9.0
1	Allen	75002	Collin	5616.0	2.0	17.0	19.0	3.0	3.0
2	Carrollton	75006	Dallas	3315.0	3.0	22.0	10.0	2.0	2.0
3	Carrollton	75007	Denton	7822.0	3.0	43.0	16.0	1.0	4.0
4	Carrollton	75010	Denton	6081.0	3.0	14.0	2.0	4.0	0.0
5	Allen	75013	Collin	6752.0	2.0	20.0	15.0	4.0	7.0
6	Coppell	75019	Dallas	7747.0	3.0	6.0	13.0	3.0	7.0
7	Flower Mound	75022	Denton	2836.0	4.0	9.0	4.0	2.0	4.0
8	Plano	75023	Collin	5449.0	10.0	33.0	12.0	1.0	3.0
9	Plano	75024	Collin	12997.0	21.0	44.0	8.0	3.0	26.0

5. Evaluating the Results (continued..)

Used the weight scores listed in page-11 to measure the score for each zip code listed in the dataset and added as new column i.e. "WtScore" to dataset after that sorted the data on WtScore in descending order to list the places with highest score in top.

The final results after data sort is as listed below.

```
df_dfw_ir_weight_score = df_dfw_ir_weight_score.sort_values(by=['WtScore'], ascending=False)
```

```
df_dfw_ir_weight_score.head()
```

City	ZipCode	County	Asian_Population	WtScore	Indian_Restaurants	Asian_Restaurants	Shopping_Places	Residential_Complex	Office_Hotels_Places
Roanoke	76262	Denton	1093.0	66.5	0.0	19.0	7.0	11.0	43.0
Denton	76201	Denton	2177.0	41.5	4.0	57.0	22.0	28.0	33.0
Grand Prairie	75052	Dallas	8610.0	40.0	0.0	10.0	30.0	8.0	10.0
Mansfield	76063	Tarrant	2875.0	37.5	0.0	34.0	28.0	10.0	27.0
Richardson	75082	Collin	5260.0	37.5	2.0	8.0	3.0	9.0	19.0

5. Evaluating the Results (continued..)

Display of the location with highest score (i.e. Roanoke:76262) with all the categories used in evaluation.
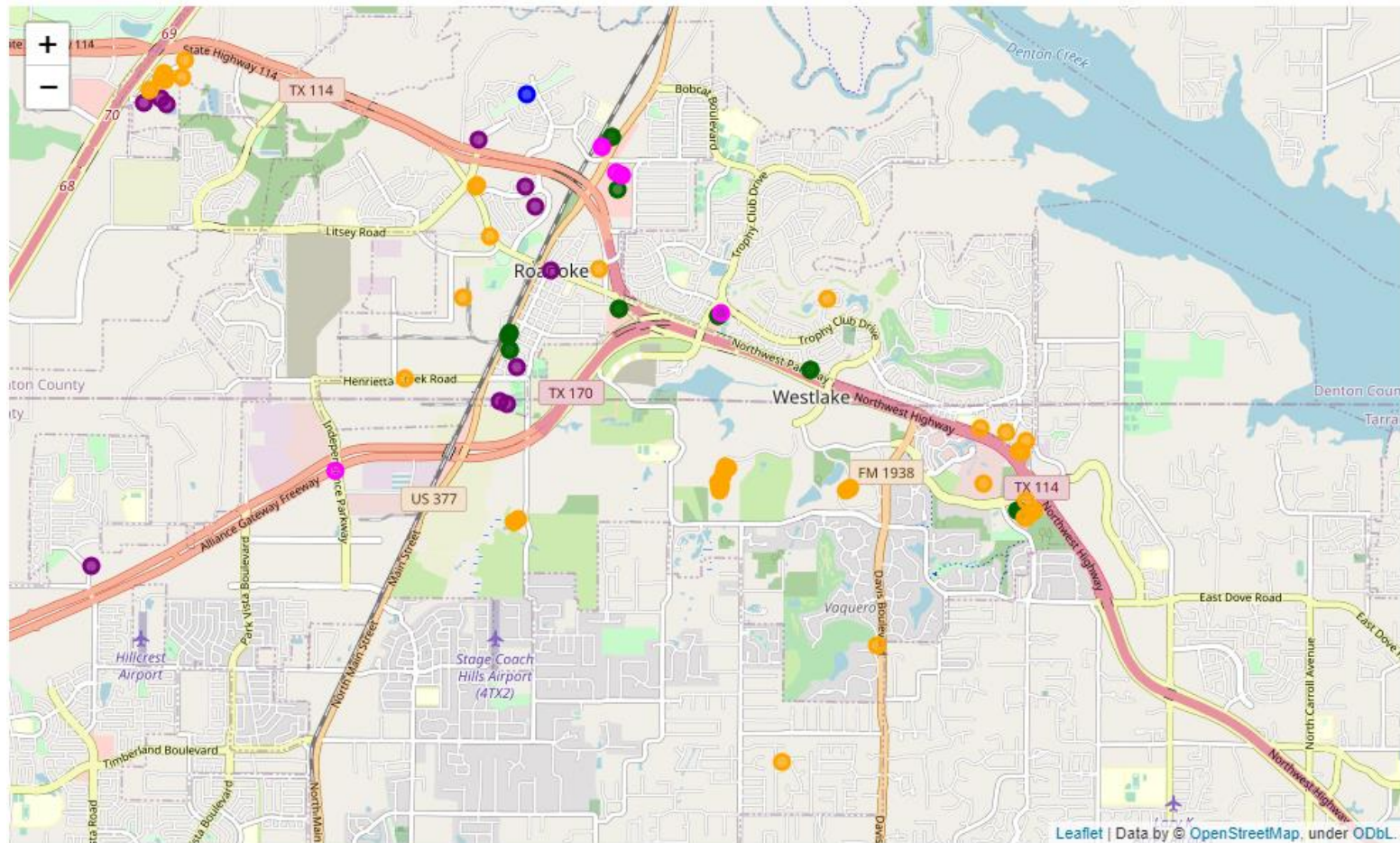
5. Conclusion

Display of the location with highest score (i.e. Roanoke:76262) with all the categories used in evaluation.

After reviewing first 3 location, we can consider Roanoke might be the best possible location to start a Indian Restaurant.
With minimum competition.

Pros:
1. No Indian / Indo Chinese Restaurants
2. Big Office complex (i.e. Charles Schwab, Fidelity, Sabre etc) and shopping places are near by this area
3. Considerable number of Residential complexes are there in the area
4. As the Property Taxes are less in this area compare to other places in DFW area and Property prices are high in other places, now Asian community is growing considerably in this area.

Cons:
1. Asian Community is less compared to other locations.

Note: We can increase the accuracy of the model by adding more input parameters (i.e. population age group wise and by adding categories in more granular level and Also by providing the customer reviews for existing restaurants and crime rate of the localities etc.)

By making some changes we can use same model as location recommender for restaurants with other cuisine.