Capstone project – Final report

Anand Lonkar

Contents

1. Introduction - Business problem

Where do I setup my next Montessori?

My sons Montessori school is in process of expanding to other areas. They already have presence in 2 zip codes within the Dallas Fort Worth metroplex and both are quite successful. Assuming that both the locations that they have share some common characteristics (demographic information, percentage of preschoolers, household income, presence of other day cares/schools nearby) I am trying to figure out the best zip code to setup a similar school in the 100 miles radius of my zipcode

The owner of the Montessori school will benefit from this analysis and hopefully will take the recommendation from this exercise. An extension of this could be beneficial to other educational franchisees like Kumon, Mathnasium etc.

2. Data Needed

The value of a Montessori school can be understood by someone who has been to one or has seen a toddler who goes to one. The method of teaching differs from other preschools and nurseries. Montessori school teachers need to undergo special training. All this contributes to these schools being costlier than other preschools. This means that the parents sending their kids to these schools should be atleast in the middle-class income bracket.

The North Texas area is rapidly expanding with new cities and subdivisions rapidly being built. More people moving in adds to the little people population and drives demands for schools and preschools. Not all communities are built equal though with some having better facilities than others and some having more schools than others. Areas with more preschools competing for same population of kids under 5 make that area less attractive for new schools.

With this background the following information was used:

- ✓ All Zip codes within 100 miles of my location - Data from https://www.zipcodestogo.com/lookups/radius-search.php
- ✓ Location data for all zip codes in Texas - https://github.com/OpenDataDE/State-zip-code-GeoJSON
- ✓ Census Data for each zip code –
- ✓ Population data – Total population and kids under 5
- ✓ Income profile for each Zip code
- ✓ Housing data for each Zip code
- ✓ Foursquare API - to find educational institutions in each of the zip codes

This data combined together was used for analysis and clustering

3. Methodology

The first step is to get the Zip codes within a 100-mile radius of the current locations. For simplicity, I took my location as the center and got data for surrounding zip-codes. This is our area of interest.

Next step is to get the location data for each of the zip codes. Fortunately this is a common problem and I was able to find a open source file giving all the required zip codes and their latitude longitude. Even though these may not be exact, they serve the purpose.

Once the location data was prepared, the next step is to get the number of preschools in a zip code. The Foursquare APIs were helpful in this context but posed several problems:

1. The APIs depend on a provided latitude and longitude and radius to find venues. Since the zip codes vary greatly in terms of area, applying this is not a one step solution. To overcome this issue, I matched the zip codes from the resultset venues to the zip code I was searching for effectively limit it to the zip code of interest
2. Another problem was the categories. Foursquare has preschools in different categories:
    a. Daycare - 4f4532974b9074f6e4fb0104
    b. Preschool - 52e81612bcbc57f1066b7a45
    c. Nursery School - 4f4533814b9074f6e4fb0107

    There is no separate classification for a Montessori school. This means that I found way more schools than expected. Though a Montessori school will compete with other Montessori schools, the target demographic is same irrespective of the type of school. So, this worked out to my advantage!

3. The data returned from foursquare was not consistent, the zip code and city fields were not returned consistently. I had to work around these and default them.

The school data was then aggregated over the zip codes.

Next was getting the census data. This is tricky since there is too much data to sift through and it is spread across multiple tables. After doing some research on the data, I finalized 3 tables each corresponding to Housing data, Income data and Age data from the 2017 estimate. I removed the unnecessary columns and renamed the columns to a more human readable format.

Merging this data was done on the zip code columns.

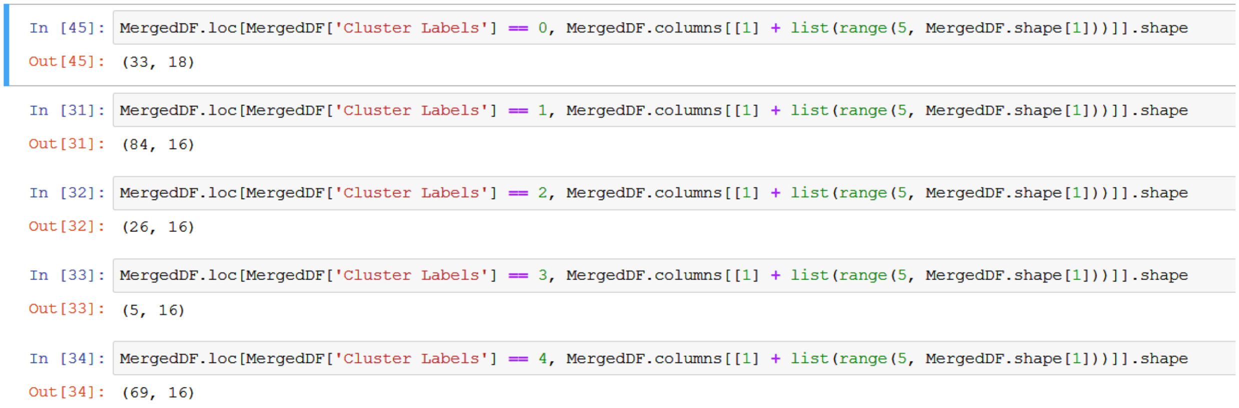
Once merged, K-means was used to cluster the data together. Some zip codes that did not have the necessary data were removed from the dataset.

After the first set of clustering, both the locations fell into different clusters. I used the K-means elbow clustering to make sure the current schools are in the same cluster.

Once this was done, I used the ratio of number of kids under 5 to the number of preschools was used to select the best zip code to open a new Montessori in the area of interest

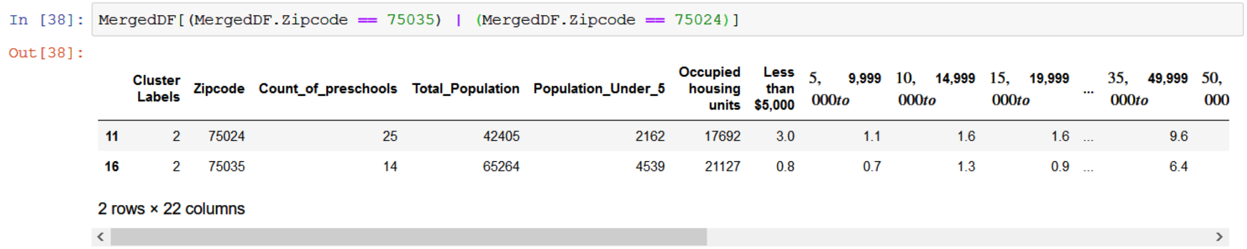
4. Results and discussion

Out of the 217 different zipcodes that participated in the battle of these neighborhoods, 5 clusters were formed. The number of clusters was chosen arbitrarily.



```
In [45]: MergedDF.loc[MergedDF['Cluster Labels'] == 0, MergedDF.columns[[1] + list(range(5, MergedDF.shape[1]))]].shape
Out[45]: (33, 18)

In [31]: MergedDF.loc[MergedDF['Cluster Labels'] == 1, MergedDF.columns[[1] + list(range(5, MergedDF.shape[1]))]].shape
Out[31]: (84, 16)

In [32]: MergedDF.loc[MergedDF['Cluster Labels'] == 2, MergedDF.columns[[1] + list(range(5, MergedDF.shape[1]))]].shape
Out[32]: (26, 16)

In [33]: MergedDF.loc[MergedDF['Cluster Labels'] == 3, MergedDF.columns[[1] + list(range(5, MergedDF.shape[1]))]].shape
Out[33]: (5, 16)

In [34]: MergedDF.loc[MergedDF['Cluster Labels'] == 4, MergedDF.columns[[1] + list(range(5, MergedDF.shape[1]))]].shape
Out[34]: (69, 16)
```

I verified that clusters are formed in such a way that both the locations are in same cluster

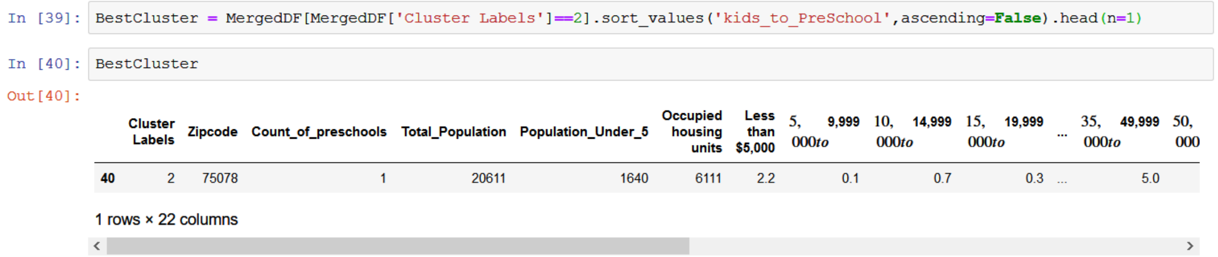


```
In [38]: MergedDF[(MergedDF.Zipcode == 75035) | (MergedDF.Zipcode == 75024)]
Out[38]:
```

	Cluster Labels	Zipcode	Count_of_preschools	Total_Population	Population_Under_5	Occupied housing units	Less than $5,000	5, 000*to* 9,999	10, 000*to* 14,999	15, 000*to* 19,999	...	35, 000*to* 49,999	50, 000
11	2	75024	25	42405	2162	17692	3.0	1.1	1.6	1.6	...	9.6	
16	2	75035	14	65264	4539	21127	0.8	0.7	1.3	0.9	...	6.4	

2 rows × 22 columns

This cluster was then sorted by the kids to preschool ratio and the winner was zip code 75078 -a nearby suburb of Propser.



```
In [39]: BestCluster = MergedDF[MergedDF['Cluster Labels']==2].sort_values('kids_to_PreSchool',ascending=False).head(n=1)

In [40]: BestCluster
Out[40]:
```

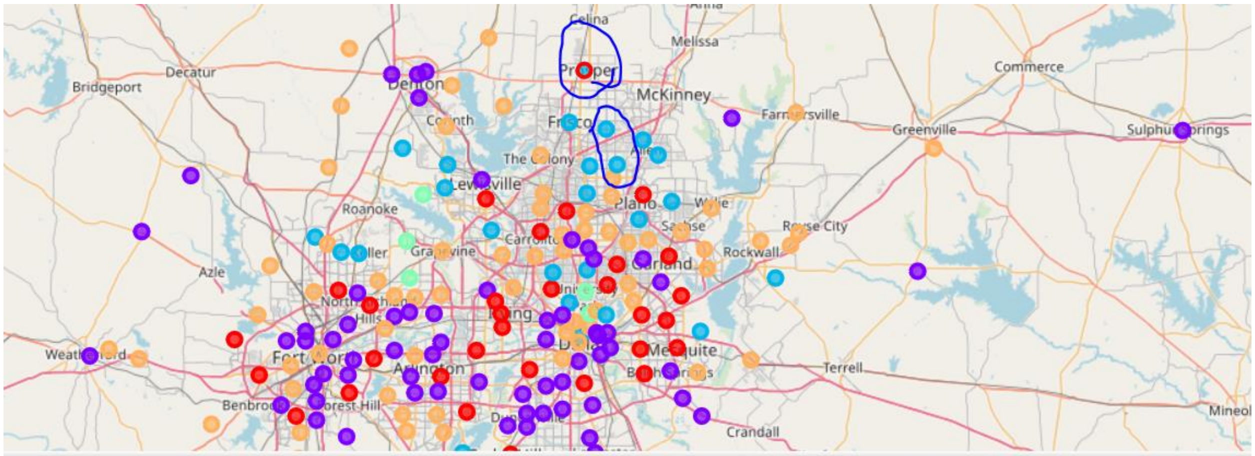
	Cluster Labels	Zipcode	Count_of_preschools	Total_Population	Population_Under_5	Occupied housing units	Less than $5,000	5, 000*to* 9,999	10, 000*to* 14,999	15, 000*to* 19,999	...	35, 000*to* 49,999	50, 000
40	2	75078	1	20611	1640	6111	2.2	0.1	0.7	0.3	...	5.0	

1 rows × 22 columns

Prosper has many similarities with both Plano and Frisco which I know about anecdotally and is now proved by this exercise.

- ✓ Frisco and Prosper are new and upcoming locations
- ✓ All 3 are high income cities mostly formed by expats from other states
- ✓ Frisco and Plano are home to some of the new transplant companies from across US while Frisco and Prosper also serve as sleeper cities to the working class

Highlighted below are existing locations in the bottom and the suggested location on the top.



5. Conclusion

After collecting the data about the general characteristics of the areas of interest, I used the Kmeans clustering to identify like zip codes. With some expected results there were some unexpected zip codes across the region that were similar.

Finally the areas with high similarity and least competition won!