Hospitals with good neighborhood search in Oklahoma State

Introduction/Business Problem

It is very challenging to find a best hospital in any state or city. There are many websites out there from which you can get hospital review and ratings like Hospital Compare, Consumer Reports, The Leapfrog Group, Quality Check, Health Grades. But even then finding a best hospital with good neighborhood that will help you out with any emergency need during hospital visits is a manual work . Finding a perfect neighborhood area can be very tedious and time consuming process.

This project attempts to help you with the search for the hospital according to your need about the neighborhood in Oklahoma State. For example, one might want to choose the area which has proximity to medical store, cafes, public transport etc. This project attempts cluster hospitals based on provided categories of neighborhood. We can visualize those clusters on the map to get better idea of area. That will make it easy to choose a certain hospital to look for and narrow down your search.

Stakeholders/Target Audience: People who are looking for a hospital with perfect neighborhood in Oklahoma State, especially the people who are visiting the particular hospital for the first time or starting any long term treatment at particular hospital.

This project will help you to get answers of some basic questions while choosing hospitals in Oklahoma state, but not all the questions. You can always use above mentioned website with this project to get all the answers.

Data

I plan to use two datasets:

1. Oklahoma State Healthcare Facilities data downloaded from https://catalog.data.gov/dataset/oklahoma-health-care-facilities . I reduced the data set to focus on 'Oklahoma' City only to decrease the number of the call made to the Four-Square API.

2. This dataset has some column which will be of no use for our project so I have dropped it from the dataset. We will mainly use Latitude, Longitude, Address to query the FourSquare API and get the desire result.

That data will be use to create clusters of neighborhood with different need.

After some basic pre-processing the data is as below :

	HEALTH CARE FACILITY NAME	ADDRESS	CITY	ZIP	PHONE1	LONGITUDE	LATITUDE	LOCATION
0	A Better Way of Care	2828 NW 57th, Suite 315	Oklahoma City	73112	(405) 819-4696	-97.563946	35.529659	(35.529659, -97.563946)
1	Abiding Home Health of Oklahoma City	5929 North May #204	Oklahoma City	73112	(405) 607-2302	-97.565848	35.531457	(35.531457, -97.565848)
2	Accentra Home Healthcare	4350 Will Rogers Parkway, Suite 500	Oklahoma City	73108	(405) 488-2222	-97.597747	35.452096	(35.452096, -97.597747)
3	Accentra Home Healthcare, Inc.	4350 Will Rogers Pkwy, Suite 500	Oklahoma City	73108	(405) 917-1094	-97.597747	35.452096	(35.452096, -97.597747)
4	Accredo Health Group, Inc.	4901 West Reno, Suite 950	Oklahoma City	73127	(405) 942-3961	-97.607725	35.464552	(35.464552, -97.607725)

3. Foursquare venues data with categories

Assumptions: Data is downloaded from DATA.GOV site. So I am assuming that provided HealthCare Facilities data for Oklahoma covers all the hospitals in the state. I am going to focus on 'Oklahoma' city only to get better results for this project.

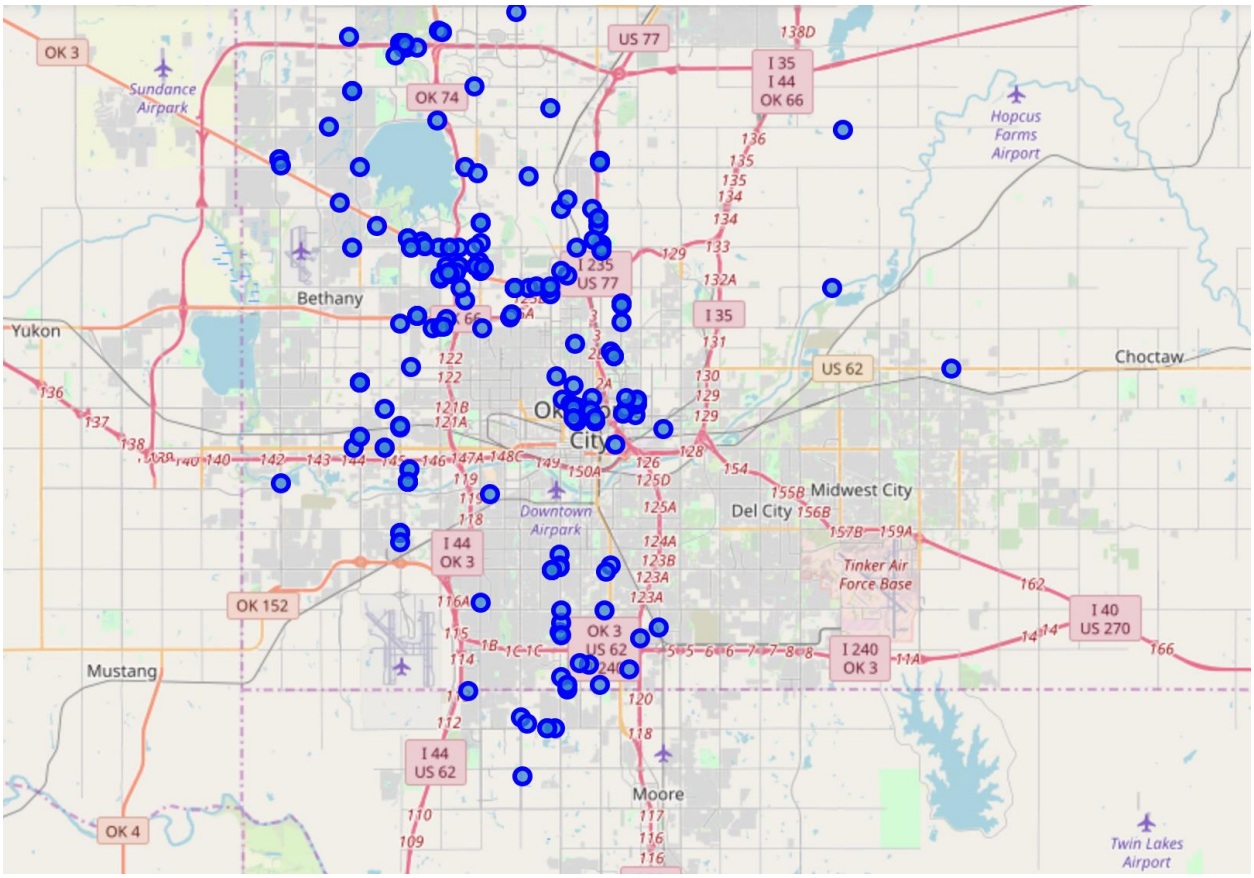
Methodology

This section is divided into three main parts data wrangling, Data analysis and clustering algorithm used to solve the problem.

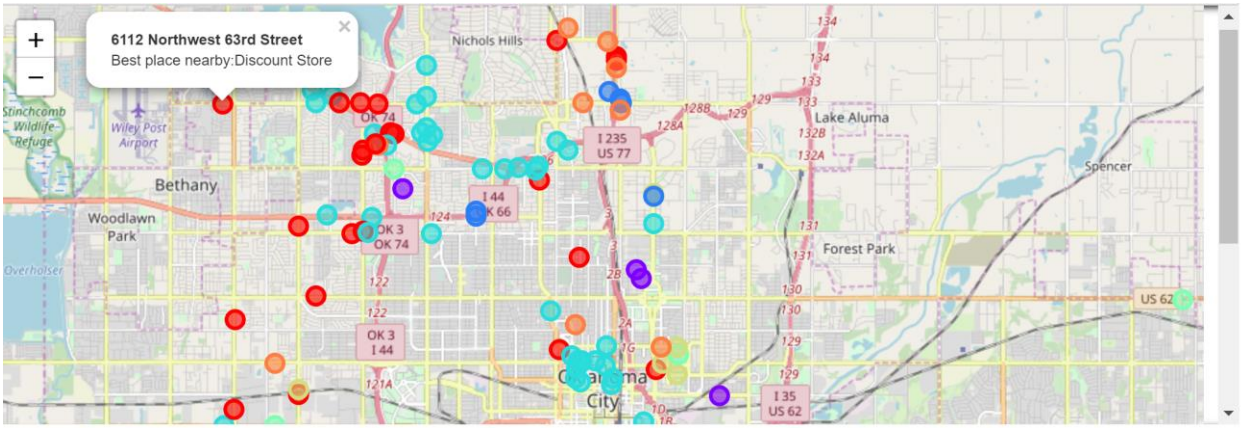
- **Data Wrangling** : Data obtained from Data.Gov for Oklahoma HealthCare Facility has many columns but I concentrated on few columns only (City, Latitude, Longitude and Address). By analyzing data on first hand I found out that it has many address repetition. So I have selected only unique address from the location data. Also, I filtered the data for Oklahoma City to minimize the calls to Four square API and for better result.

- **Data Analysis**: Oklahoma HealthCare Facility dataset has 149 unique city's data and I filtered on a single city as to get better result namely Oklahoma. First I have plotted the full map of Oklahoma State with all the hospitals listed with their address to get the overview. I also plotted the Oklahoma City's hospitals only on the map. After that I have generated the neighborhood venues for all the hospitals in the list and then I created seven different clusters based on the top four venue categories provided by Foursquare API data. I also explored top five venues in each cluster.

Oklahoma City map with all the listed hospitals :



Oklahoma City Hospitals according to the neighborhood type cluster :



Since our objective is to ease the search for people who have preferences in terms of nearby venues, I tried to explore all the hospitals under Oklahoma city using Four square

API and obtained multiple venues names, their geographical coordinates and category for each sector within 500m radius and put a limit on the number of venues as 100 for each sector.

From four square API, I obtained 211 different venue categories like Café, Pubs, Public transport, Park, Nature Reserve, Vintage stores etc.

- **Machine Learning**: I decide to use K means Clustering for the problem because after the data analysis it was obvious that we can categorize locations into different homogenous clusters based on the venue categories. Further on, we need to create our dataset in a format where rows will represent the hospital address and all columns will be venue categories and we will calculate the score for every category in each row as per their presence or absence in the area. We sort the categories for each area in order of their occurrence, for example if a area has more cafes and pubs, the area will get higher score for the cafes and pubs and less for other categories. The previous step will be sufficient to execute data mining technique I chose to accomplish this task(K mean clustering). K means clustering will segment the data into groups where groups are similar among themselves but different from other groups in terms of occurrence of venue categories. For our dataset, K mean clustering will append another column to the dataset which will depict the cluster number, similar areas will be grouped together.

Result

I obtained 7 clusters from K means clustering executed in the previous step which are explained below:-

- Cluster 0. Contains hospital with proximity to Medical Center, Park, Gym, Sandwich place etc.
- Cluster 1. The cluster highlights all hospitals in the proximity to Steakhouse, Wings Joint, Mediterranean Restaurant etc.
- Cluster 2. The cluster highlights all hospitals in the proximity to Fast Food Restaurant, Discount Store, American Restaurant, Grocery Store etc.
- Cluster 3. It contains hospitals with proximity to Bar, Coffee Shop, Gym, Gas Station, etc.
- Cluster 4. It contains hospitals with proximity to Mediterranean Restaurant, Medical Center, Gift shop, etc.
- Cluster 5. It contains hospitals with proximity to Pizza Place, Lawyer, Museum, Doctor's Office, etc.

- Cluster 6. It contains hospitals with proximity to Hotel, Breakfast Store, Clothing shop, etc.

The above clusters segment the data into 7 homogeneous groups which exhibit similarities among themselves in terms of venue categories are being heterogenous from the other groups.

Conclusion

- As demonstrated in the Result section, we can effectively simplify the hospital search process for our customers by clustering the area based on the neighborhood venue categories and filtering the cluster to narrow down their search area. Customers can then prefer to focus their hospital search by analyzing various trade offs. Those who are confused can efficiently compare various areas and choose the one which is most suitable for them.