Capstone Project - The Battle of Neighborhoods

A Study of Neighborhoods of Overland Park

Huiping Qin

May 11, 2019

Overland Park: image from wikipedia (ref. 7)

Acknowledgement

Thanks to my family for their full support during the study of these data sciences courses and as always. Thanks go to all the instructors of these courses for their clear and informative video classes and hands on labs. I also deeply appreciate all the co-classmates of these courses for their timely grading of my assignments and positive feedback.

Table of Contents

1. Introduction

Why Overland Park (ref. 6) shows up frequently on the top list of best places to live in US? There are various info on the Internet, however deep neighborhoods analysis is lacking. Through data analysis and machine learning technology, we may find evidences to the answer, which may in turn help local business on recruiting marketing and attract more talents to Overland Park. It may also help people who want to find a better place to retire or to raise a family to make informed decision. I have lived in Overland Park for over 20 years and searching an answer on the question may help me to know my community better and to love it even more. I love my local community, with easy access to good education, libraries, museums, shopping, entertainment, park and trails, and so forth. But what are other neighborhoods in Overland Park like, what venues do they have to contribute to the listing of Overland Park on top places to live in the US?

Equipped with knowledge and skills just acquired from the Coursera Data Sciences classes, I hope to find the answer by using a combination of location data and machine learning to explore neighborhoods in the city of Overland Park. In this study, I am going to use the Foursquare API (ref. 4) to explore neighborhoods in Overland Park. I will use the **explore** function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters. I will use the *k*-means clustering algorithm to complete this task. Finally, I will use the Folium library to visualize the neighborhoods in Overland Park and their emerging clusters.

2. Data

2.1 Data sources

To be able to segment and cluster neighborhoods in Overland Park, I will need the list of neighborhoods in Overland Park and their corresponding geo coordinates data. However, neither is readily available. I have searched the Internet and found neighborhoods list on the Nextdoor website (ref. 1) I have used the BeautifulSoup Python library (ref. 2) to scrape the names of the neighborhoods from the Nextdoor website. To get the neighborhoods' geo coordinates, I have used the geopy library (ref. 3) to get the latitude and longitude values of the neighborhoods. For the location data such as data describing places and venues, I have used Foursquare API (ref. 4) to get info from their server. Please

find below Fig. 1 of snapshot of neighborhoods list of Overland Park on Nextdoor and Fig. 2, a sample of data from Foursquare.

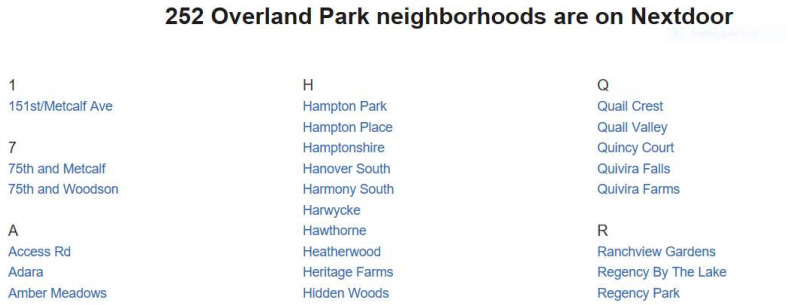

Fig. 1 Overland Park Neighborhoods on Nextdoor Snapshot

	name	categories	lat	lng
0	Downtown Mission	Historic Site	39.014909	-94.662374
1	ARC	Gym / Fitness Center	39.013159	-94.663269
2	Henhouse	Grocery Store	39.010678	-94.667633
3	98.9 The Rock!	Rock Club	39.016930	-94.666710

Fig. 2 Sample Data from Foursquare

2.2 Data Wrangling

Neighborhoods' names were scraped from Nextdoor website. Since they are local specific neighborhood names and got erroneous coordinates back per geopy query. I have added " Overland Park, Kansas" to the neighborhood names as addresses and got back reasonable geo coordinates. The geopy package did not return geo coordinates for all neighborhoods in the list, so I have to limit the analysis on only the neighborhoods with geo coordinates returned. Eventually, I have built a dataframe of neighborhoods with their top 10 most common venues in their respective neighborhood, ready for segmentation and clustering analysis.

3. Methodology

3.1 Exploratory Data Analysis

To get a sense of the neighborhoods I scraped from the Nextdoor site and the geo coordinates per geopy query, I have been able to visualize the Overland Park with the neighborhoods on it in below map.

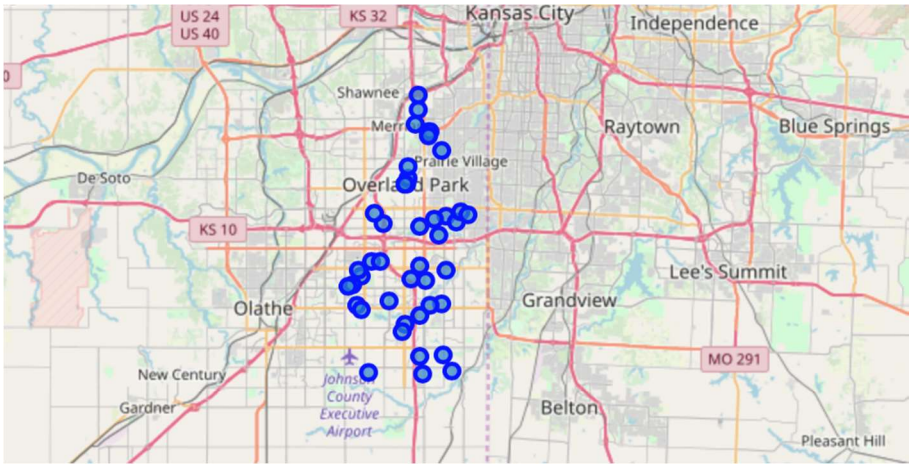

Fig. 3 Overland Park neighborhoods on Map

Combining the geo coordinates with data from Foursquare, I was able to build a data frame for venues for the neighborhoods and their corresponding geo coordinates as shown below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Access Rd	39.013294	-94.663835	Downtown Mission	39.014909	-94.662374	Historic Site
1	Access Rd	39.013294	-94.663835	ARC	39.013159	-94.663269	Gym / Fitness Center
2	Access Rd	39.013294	-94.663835	Henhouse	39.010678	-94.667633	Grocery Store
3	Access Rd	39.013294	-94.663835	98.9 The Rock!	39.016930	-94.666710	Rock Club
4	Adara	38.885210	-94.730738	Mai Thai	38.884119	-94.726609	Thai Restaurant

Fig. 4 Sample neighborhoods with venues info and geo coordinates

Similar with as above, I was able to get the full list of venues for all neighborhoods in Overland Park and the unique categories. I was also able to print out each neighborhood along with the top 5 most common venues as sampled below.

```
----Highland Village----
          venue  freq
0   Coffee Shop   0.2
1  Dance Studio   0.2
2   Pizza Place   0.2
3  Liquor Store   0.2
4      Pharmacy   0.2


----Indian Creek Village----
                venue  freq
0  Korean Restaurant  0.09
1             Bakery  0.09
2       Dance Studio  0.09
3              Café  0.09
4      Grocery Store  0.09
```

Fig. 5 Two sample eighborhoods and their top 5 most common venues

By applying one hot encoding, I was able to build a dataframe with neighborhood and their top 10 most common venues in their respective neighborhood, ready for segmentation and clustering analysis.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Access Rd	Gym / Fitness Center	Rock Club	Historic Site	Grocery Store	Women's Store	Fried Chicken Joint	Food Truck	Food	Fast Food Restaurant	Fabric Shop
1	Adara	Gym / Fitness Center	Basketball Court	Salon / Barbershop	Kids Store	Sports Club	Thai Restaurant	Women's Store	Furniture / Home Store	Fried Chicken Joint	Food Truck
2	Apple Valley Estates	Gas Station	Sushi Restaurant	Pizza Place	Arts & Crafts Store	Grocery Store	Liquor Store	Garden	Furniture / Home Store	Fried Chicken Joint	Food Truck
3	Brittany Park	Playground	Park	Health & Beauty Service	Gym / Fitness Center	Women's Store	Fabric Shop	Furniture / Home Store	Fried Chicken Joint	Food Truck	Food
4	Caenen	Gym / Fitness Center	Gas Station	Gym	Cosmetics Shop	Coffee Shop	Salon / Barbershop	Fast Food Restaurant	Smoothie Shop	Gift Shop	Sports Club

Fig. 6 First 5 rows of the dataframe with top 10 most common venues in each neighborhood

3.2 K-means Cluster

The K-means is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data, which is a form of **unsupervised** machine learning. K-means will partition the dataset into preselected number of groups as clusters. The item in each cluster are similar to each other in terms of the features included in the dataset. K-means tries to minimize the intra-cluster distances and maximize the inter-cluster distances.

Neighborhood venues are unstructured and unlabeled data. K-means clustering is a perfect way to provide some insights to the neighborhoods in an area in a quickly manner.

The KMeans class has many parameters that can be used, but we just used two here: the number of cluster was set to 5 and the random_state was 0.

```
Run *k*-means to cluster the neighborhood into 5 clusters.
```

```
# set number of clusters
kclusters = 5

op_grouped_clustering = op_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(op_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([0, 0, 0, 3, 0, 0, 4, 3, 0, 0], dtype=int32)
```

Fig. 7 K-means to 5 clusters

4. Results

For neighborhoods in Overland Park, K-means clustering provided 5 clusters. I was able to use folium map to visualize the resulting clusters on an interactive map. You can zoom in and out. Each neighborhood will pop up its cluster label when you click on each of the neighborhood circle.

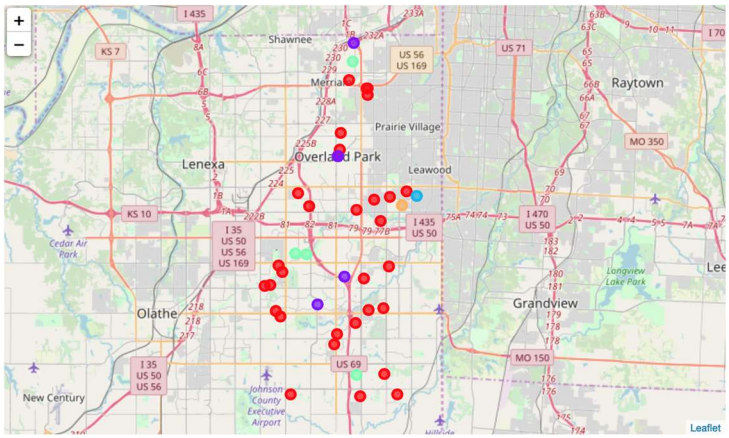

Fig. 8 Neighborhood clusters on map

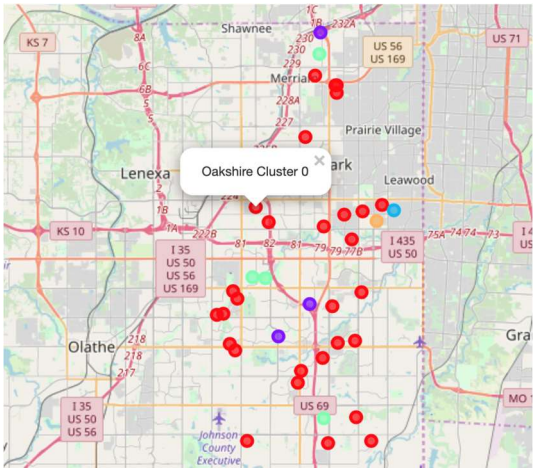

Fig. 9 Sample neighborhood in cluster 0 popup on map

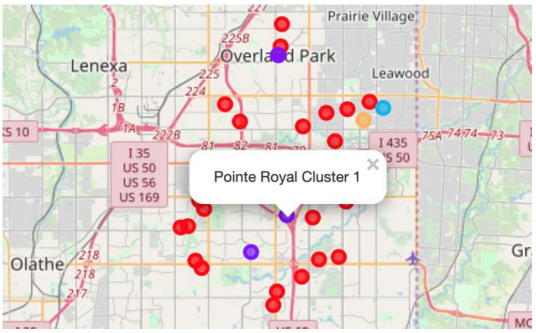

Fig. 10 Sample neighborhood in cluster 1 popup on map

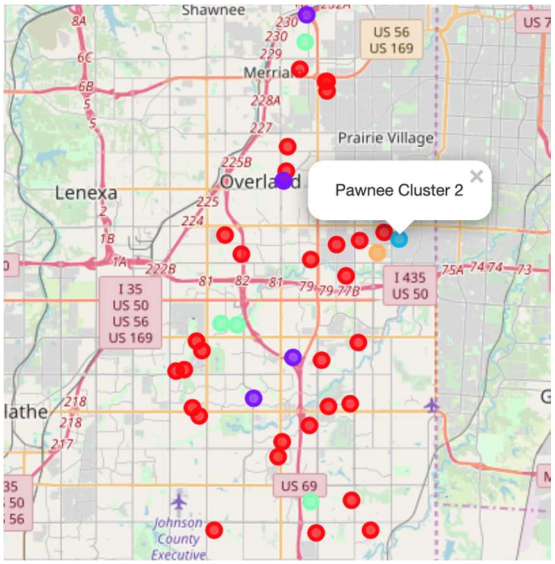

Fig. 11 Sample neighborhood in cluster 2 popup on map

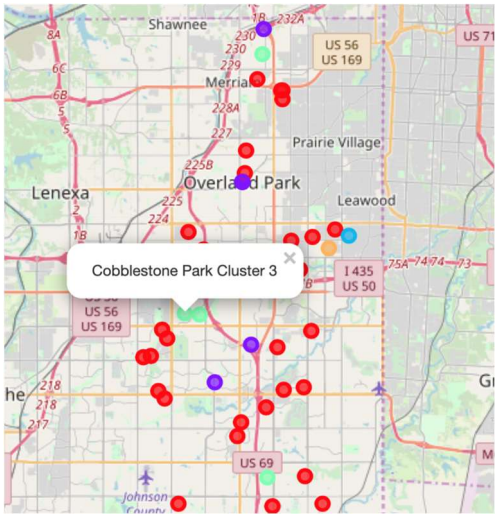

Fig. 12 Sample neighborhood in cluster 3 popup on map

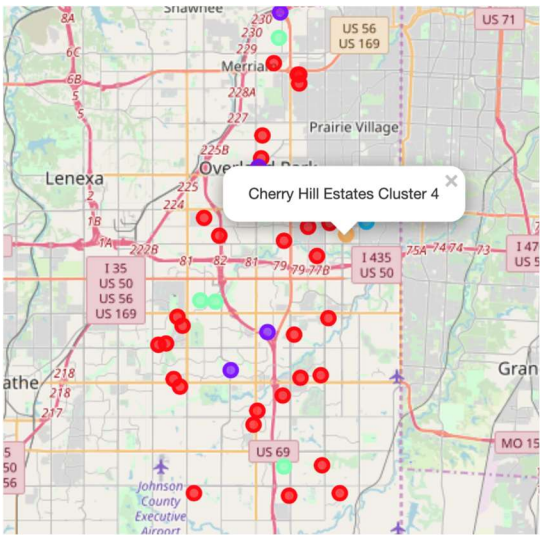

Fig. 13 Sample neighborhood in cluster 4 popup on map

Considering the common characteristics of each cluster, we can create a profile for each cluster. For example, I have named the 5 clusters as below:

- Cluster 1: Living Haven
- Cluster 2: Pizza and Shopping Haven
- Cluster 3: Refill Haven-Both Human and Cars
- Cluster 4: Leisure Haven
- Cluster 5: Happy Hour Place

5. Discussion

It is interesting that most neighborhoods in Overland Park fall into clusters 1 and 2, with cluster 1 as Living Haven, which pretty much has access to all living related venues; while cluster 2 is a Pizza and shopping Haven. With cluster 3, you eat at restaurants and fill the car tank conveniently too. For cluster 4 and 5, I label them as Leisure Haven and Happy Hour Place since its top venue is Bar. From the analysis there are 134 unique categories of venues in Overland Park, even though it may seem fewer than big cities like NY City or Toronto, it is still quite many for a city like Overland Park. Also, from the visualization of neighborhoods on map, one can see that the neighborhoods scattering all around Overland Park belongs to cluster 0, which I labeled as Living Haven. That means that in every corner of

Overland Park, there are many amenities and venues that people can easily access. This provides evidence that Overland Park deserves to be on the top lists to live in US.

During neighborhood data acquiring and analysis, I noticed that geo coordinates were not returned for some neighborhood address queried. It needs some further investigation. Also NaN kmean_labels were returned for couple of rows of neighborhoods, which also need some looking into. In addition, the neighborhood lists on Nextdoor may also need to be verified if missing any.

Even though this neighborhoods of Overland Park segmentation and clustering study provided some evidence and support to answer the question from the very beginning, further study would provide even more solid and rich evidences. For example, we can explore trending venues around Overland Park. These are venues with the highest foot traffic at the time a regular call to the Foursquare API is made. This type of study and analysis may provide insights to developing trends thus may reveal some business opportunities in Overland Park.

6. Conclusion

In this study, I analyzed the neighborhoods of Overland Park by segmenting and clustering with K-means machine learning algorithm. The analysis provided insights and evidence to answer the question why Overland Park shows up frequently on the top list of best places to live in US. In addition, from this study, I get to know my city better and know that many neighborhoods in it are just like mine, a Living Haven. Hope this study and report would be able to provide local business evidence support to attract more talents to Overland Park and anyone who may explore the possibility to move to Overland Park.

7. References

1. Nextdoor website: https://nextdoor.com/city/overland-park--ks/
2. https://www.crummy.com/software/BeautifulSoup/bs4/doc/
3. https://geopy.readthedocs.io/en/stable/
4. Foursquare.com: https://foursquare.com/
5. JupyterNotebook: https://jupyter.org/
6. https://www.opkansas.org/
7. https://en.wikipedia.org/wiki/Overland_Park,_Kansas