Coursera IBM Applied Data Science Capstone

Denver Neighborhoods

January 10th, 2021

Introduction

The capital of Colorado, Denver is a sprawling metropolis that attracts both tourists and new residents, leading the statewide population growth. Since 2010, the city has grown by a cumulative 21% [1]. Particularly, Denver has become a popular destination for people relocating during the COVID-19 pandemic [2]. The median age of residents in Denver is 34 [3], suggesting a large population of early career professionals who are or soon to be parents of young children. These people have a particular demand of living environment, such as the convenient access to child services and healthy grocers. In addition, these people often take into consideration of housing price when they choose neighborhoods to live in, because they often have to balance the needs between different categories of living expenses.

This project aims to provide an analysis of Denver neighborhoods by integrating the information of median house price index for each neighborhood and the level of facility abundance. The analysis results are visualized in a map, which can be used in a recommender system or personal assistance application if implemented successfully. The target audience are incoming as well as current residents who are interested in moving into Denver neighborhoods convenient to raising young children based on their willingness or capability to afford housing expenses.

Data

The following datasets are required to accomplish the objective.

1. A list of neighborhoods in Denver and their latitude and longitude coordinates
2. The housing index of these neighborhoods
3. The number of select venues in each neighborhood

The identify of a neighborhood can be determined by its name, coordinates, and sometimes zip code. After researching on the datasets of Denver neighborhoods available on the Internet, it became clear that the best way to identify each neighborhood for this project is to use a combination of name and coordinates. Specifically, the neighborhood name will be used to obtain the home value index for each neighborhood via Zillow (Zillow Home Value Index, or ZHVI, can be downloaded in csv file from https://www.zillow.com/research/data/). The coordinates of each neighborhood will be used as inputs to obtain the list of nearly venues through the Foursquare API (https://developer.foursquare.com/docs).

The name and coordinates of Denver neighborhoods were obtained as a geojson file from Kaggle (https://www.kaggle.com/broach/denverairbnb?select=neighbourhoods.geojson). The neighborhoods are polygon shaped, with the coordinates of each polygon point reported in the geojson file. The center coordinates were calculated by averaging the highest number and lowest number for latitudes and longitudes, respectively. Note it is also possible to obtain center coordinates for each neighborhood by

using Google Maps manually for free. The more automatic method of Google Maps Geocoding API charges a small amount of fee, so it was not pursued in this project. Attempt was made to explore other tools such as (https://geocoder.readthedocs.io/), but it didn't work as it was built on Google Maps API.

These center coordinates were then used to make requests in the Foursquare API to obtain a list of venues within the specified radius. Foursquare has a large number of categories, but only a few were selected in this application that are most relevant to the needs of working parents of young children. Target categories and their categoryId are listed below.

- Child Care Service (5744ccdfe4b0c0459246b4c7)
- Baby Store (52f2ab2ebcbc57f1066b8b32)
- Preschool (52e81612bcbc57f1066b7a45)
- Drugstore (5745c2e4498e11e7bccabdbd)
- Grocery Store (4bf58dd8d48988d118951735)
- Organic Grocery (52f2ab2ebcbc57f1066b8b45)
- Urgent Care Center (56aa371be4b08b9a8d573526)
- Gym / Fitness Center (4bf58dd8d48988d175941735)
- Shopping Mall (4bf58dd8d48988d1fd941735)

Methodology

After a list of Denver neighborhoods and center coordinates were obtained from processing the geojson file mentioned in the Data section, the neighborhoods data were visualized on a map using the Folium package to ensure that the calculation of center coordinates was reasonable (Figure 1). There is a total number of 78 neighborhoods in Denver.

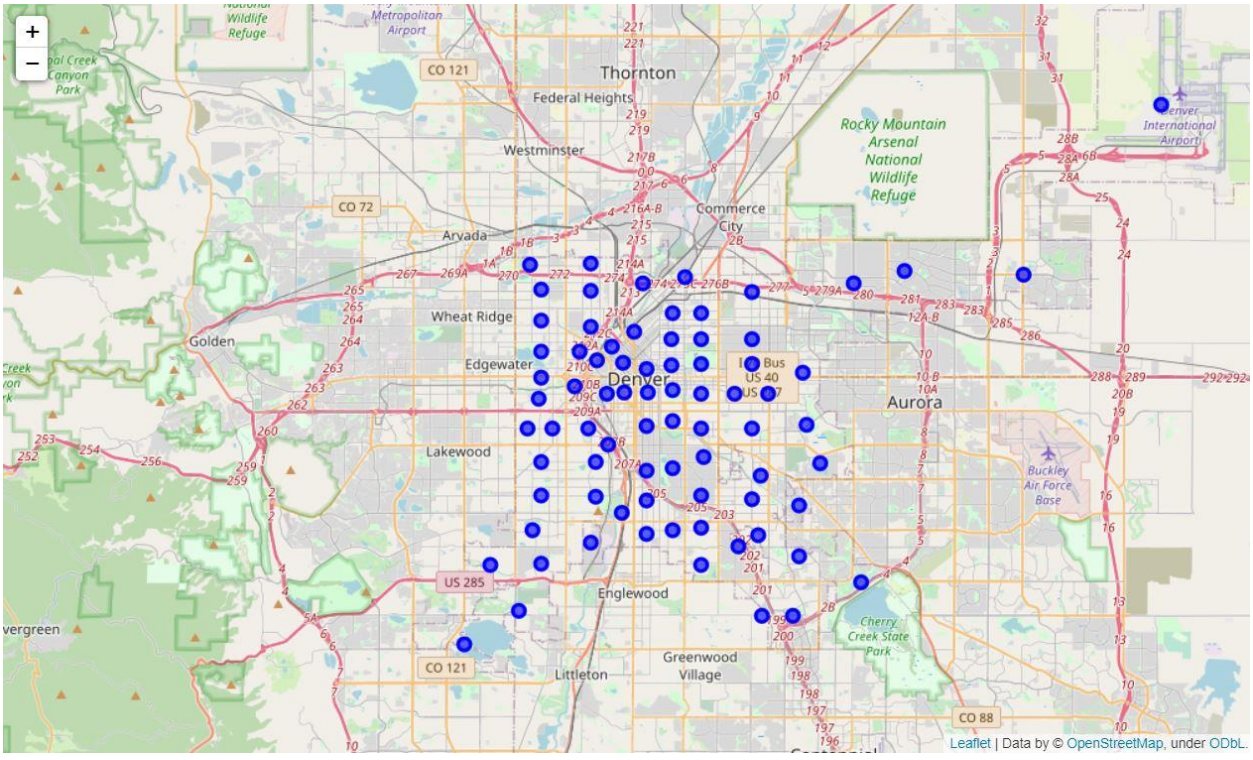

Figure 1. Visualization of Denver neighborhoods using calculated center coordinates.

To obtain the housing data for Denver neighborhoods, the csv file from Zillow was first read into a pandas dataframe, followed by extracting the entries where City is Denver. Counting the unique value of RegionName revealed that data of 77 neighborhoods are available, missing the Kennedy neighborhood. Considering that Kennedy is adjacent to and only to Hampden, the data of Hampden were copied to fill in the missing data of Kennedy. This assumption may not be valid. Further details of Kennedy and Hampden should be collected to assess the assumption. After the missing data and other inconsistencies in the name of neighborhoods were addressed, the most recent home value index data (column "2020-11-30") were selected for further use. By using the cleaned dataframe and the geojson file, a choropleth map was plotted to visualize the home value index (Figure 2).

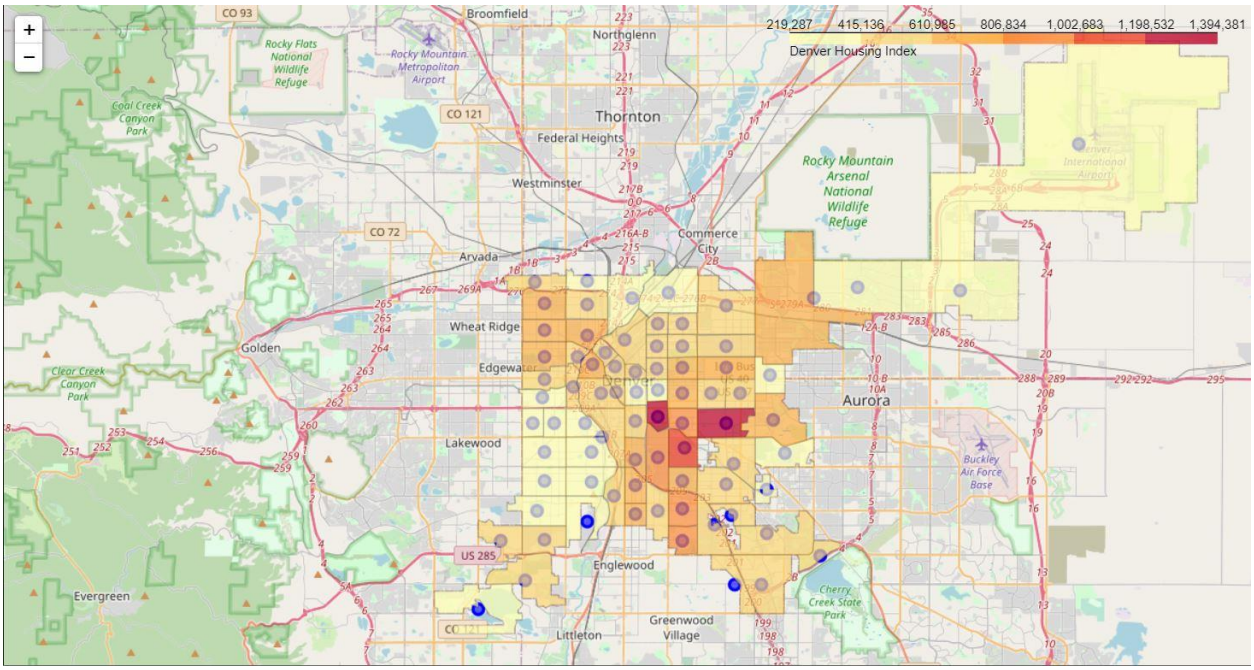

Figure 2. Visualization of Denver neighborhoods home value index.

The venues data for all the neighborhoods were retrieved from Foursquare API by passing in the center coordinates of each neighborhood to query the venues of specified categoryId using a loop. Foursquare returns the data in json format. Select information such as neighborhood and venue name as extracted and further grouped by neighborhood to count the number of target venues in each neighborhood. It was noticed that only 77 neighborhoods were reported. No data was retrieved for DIA neighborhood. Considering that the venue categories were pre-determined to cater to the needs of the target audience, it is more interesting to evaluate the facility abundancy than to understand the similarity among these neighborhoods.

It is possible to rank the neighborhoods based on the total number of venues, but instead the neighborhoods were grouped on a few levels (e.g., high-medium-low level of facility abundancy), which provides more flexibility in choosing neighborhoods and enables effective decision making (i.e., choosing from three groups instead of from >70 options). The clustering of neighborhoods was performed by

using k-means clustering, a popular unsupervised machine learning algorithm. The number of clusters was pre-determined to be three.

Results

K-means clustering successfully clustered Denver neighborhoods into three groups. From Table 1 below, it appears that the more abundant the facilities are, the larger number the label has. Therefore, we can safely assign label "0" to DIA neighborhood, which returned zero venues.

Table 1. Results of Denver neighborhoods clustering

	Neighborhood	VenueCount	Label
0	Athmar Park	4	0
1	Auraria	32	2
2	Baker	5	0
3	Barnum	5	0
4	Barnum West	2	0
...	...	...	...
72	West Colfax	4	0
73	West Highland	28	2
74	Westwood	2	0
75	Whittier	8	0
76	Windsor	6	0

77 rows × 3 columns

The results of clustering are also visualized in the map, with cluster 0 in red, cluster 1 in purple, and cluster 2 in cyan. Pop-ups for each neighborhood contain the information of neighborhood name and cluster label (Figure 3).

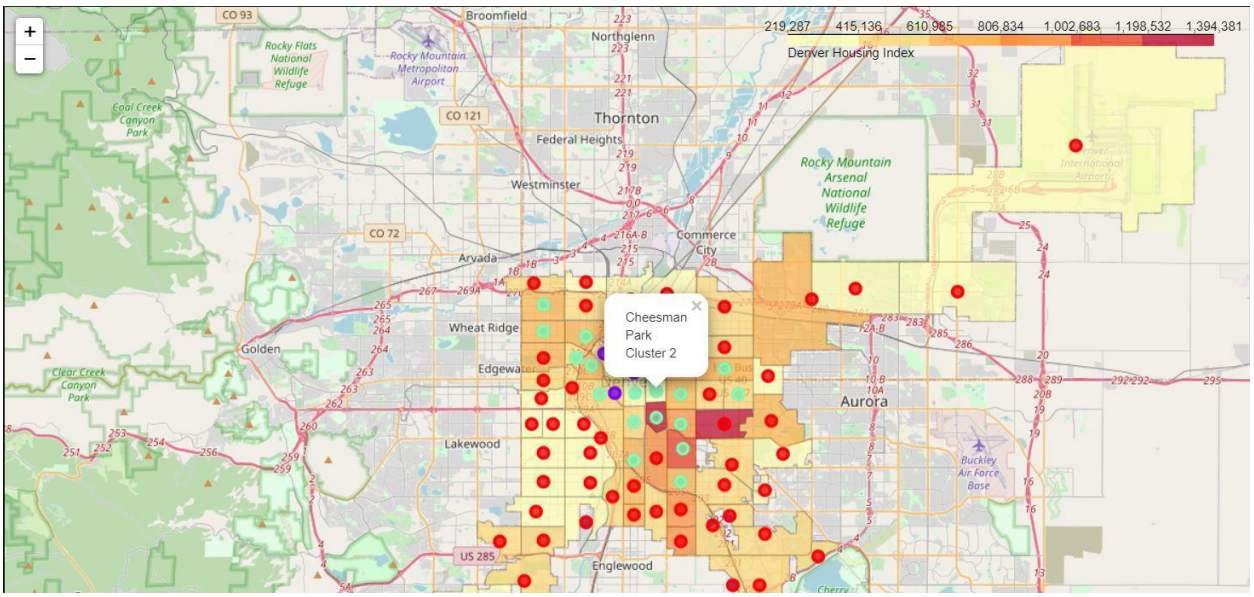

Figure 3. Visualization of Denver neighborhoods home value index and clustering results (pop-up refers to one example neighborhood of both high facility abundancy and high housing affordability)

Discussion

Neighborhoods with convenient access to facilities important to raising young children (e.g., child care service, baby store, preschool) are mostly in the center and northwest of the city. Many of these neighborhoods are in the areas with low-to-medium home value index, particularly Cheesman Park (Figure 3) and Capitol Hill, which is friendly to young parents who have limited budget on housing expenses. However, the availability of housing in these areas can be limited depending on the time of the year and market situation. Alternatively, people can look into the red-dotted neighborhoods close to the cyan-dotted neighborhoods. These neighborhoods are still viable choices given that the families have easy access to transportation. For those who would prefer to live in high home value areas, they may be interested in looking into Country Club and Becaro neighborhoods.

Conclusion

In this project, an analysis of Denver neighborhoods was conducted to support decision making for people who are raising young children or planning for it in the near term. This analysis integrated two pieces of information for all 78 Denver neighborhoods, i.e., house price and the level of facility abundance. The first piece of information was visualized by using a choropleth map, and the second by clustering the neighborhoods based on the number of target venues. The workflow comprised of raw data acquisition, data cleaning/wrangling, machine learning K-means clustering, and map visualization. This analysis can be used in a recommender system or personal assistance application if implemented successfully. Future work should include other considerations such as criminal data to add more perspectives and allow for wholistic decisions.

References

[1] https://www.denverpost.com/2020/03/25/denver-colorado-2019-census-population-growth-estimates/

[2] https://www.denverpost.com/2020/12/16/denver-most-moved-to-cities-covid-pandemic/

[3] https://www.mymove.com/city-guides/co/denver/