Neighborhood Analysis of Denver, Colorado

Capstone Project by Megan Harrison

Coursera Course: IBM Data Science Professional Certificate

Table of Contents

Introduction

I was born and raised in Denver, but have lived out of the country for the last 5 years. In the next 6-8 months I will be moving back to Denver with my husband, but we do not know where to live, as the landscape of the metropolitan area has changed much due to the city's considerable growth.

In this notebook, I will analyze the neighborhoods in and around Denver, Colorado in order to find the best place to live, taking into consideration the following factors:

- **Neighborhood characteristics**
    - Outdoor activities (parks, trails, dog parks, etc.)
    - Restaurants & bars (coffee shops, breweries, restaurants)
- **Housing prices** (to live within our budget)
- **Proximity to downtown and to the airport** (my husband and I don't want a long commute to work, and we often fly to visit family)

This analysis could also be of use to other Denver residents who are looking to move neighborhoods, as well as individuals looking to open venues in neighborhoods with certain characteristics.

Data

To realize this analysis, I will use the following data sources:

- **Zillow data** for Denver neighborhoods and median housing prices http://files.zillowstatic.com/research/public/Neighborhood/Neighborhood_Zhvi_AllHomes.csv

	RegionID	SizeRank	RegionName	RegionType	StateName	State	City	Metro	CountyName	date	price
0	274772	0	Northeast Dallas	Neighborhood	TX	TX	Dallas	Dallas-Fort Worth-Arlington	Dallas County	1996-01-31	134517.0
1	112345	1	Maryvale	Neighborhood	AZ	AZ	Phoenix	Phoenix-Mesa-Scottsdale	Maricopa County	1996-01-31	NaN
2	192689	2	Paradise	Neighborhood	NV	NV	Las Vegas	Las Vegas-Henderson-Paradise	Clark County	1996-01-31	139741.0
3	270958	3	Upper West Side	Neighborhood	NY	NY	New York	New York-Newark-Jersey City	New York County	1996-01-31	246925.0
4	118208	4	South Los Angeles	Neighborhood	CA	CA	Los Angeles	Los Angeles-Long Beach-Anaheim	Los Angeles County	1996-01-31	134826.0

  - With this data, I will acquire the unique neighborhoods of the city ('RegionName').
  - Additionally, the 'price' column indicated median house prices of the neighborhood.

- **Foursquare data** to understand the neighborhoods' characteristics. Given that there are so many categories of variables (234 categories identified in Denver), I will realize a mapping of these categories to more comprehensive groups, thus allowing the clustering algorithm to better generalize.

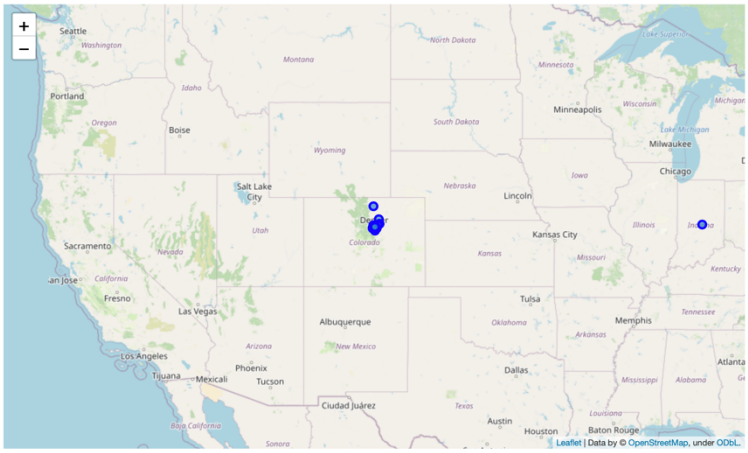
Venue	Category
Sports Club	Athletics / Sports
Boxing Gym	Athletics / Sports
Dance Studio	Athletics / Sports
Basketball Court	Athletics / Sports
Athletics & Sports	Athletics / Sports
Gym	Athletics / Sports
Yoga Studio	Athletics / Sports
Golf Course	Athletics / Sports
Recreation Center	Athletics / Sports
Gym / Fitness Center	Athletics / Sports
Volleyball Court	Athletics / Sports
Football Stadium	Athletics / Sports
Stadium	Athletics / Sports
Skating Rink	Athletics / Sports
Pub	Bar / Brewery
Wine Bar	Bar / Brewery
Beer Garden	Bar / Brewery
Sports Bar	Bar / Brewery
Distillery	Bar / Brewery
Brewery	Bar / Brewery
Bar	Bar / Brewery
Gastropub	Bar / Brewery
Dive Bar	Bar / Brewery
Insurance Office	Business

Methodology

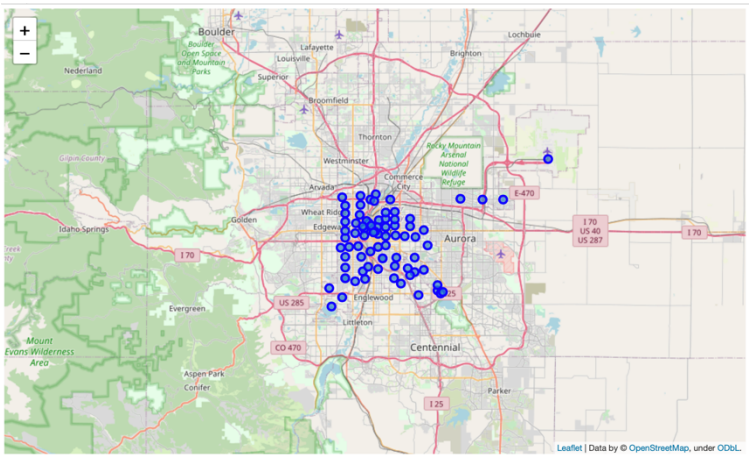
Load and Prepare Data

Zillow data:

- Filter for Denver data only.
- Use geolocator function to identify the coordinates for each neighborhood. For those not identified, I had to add the coordinates manually.
- Upon plotting the coordinates on a map, I noticed that the geolocator function identified several coordinates incorrectly:



- I identified which coordinates were not correct and updated them manually:

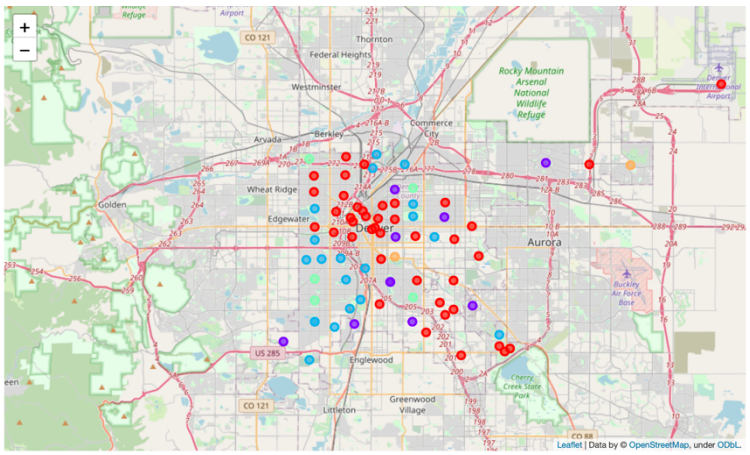


Foursquare data:

- Retrieve the *venues* for the Denver neighborhoods identified in the Zillow dataframe
- Given that 234 unique categories of venues were identified, I decided to group them into similar categories, and so that the algorithm can better generalize. In a separate spreadsheet, I realized this grouping and then updated my dataframe.
- Next, I prepared the dataframe to apply the clustering algorithm with one hot encoding.

Primary Application of Clustering Algorithm

On the data previously prepared, I applied the k-means clustering algorithm, setting the number of clusters to 5. I plotted the results on the Denver map:



A deep dive into each cluster neighborhood to see the 10 most common venues in each neighborhood, by cluster. Given the characteristics of the clusters, we are most excited about Cluster 2, because of its abundance of outdoor recreational activities as well as restaurants.

Cluster 2

```
In [31]: 1 neighborhoods_venues_sorted_clustered.loc[neighborhoods_venues_sorted_clustered['Cluster Labels'] == 2, neighbor
```

Out[31]:

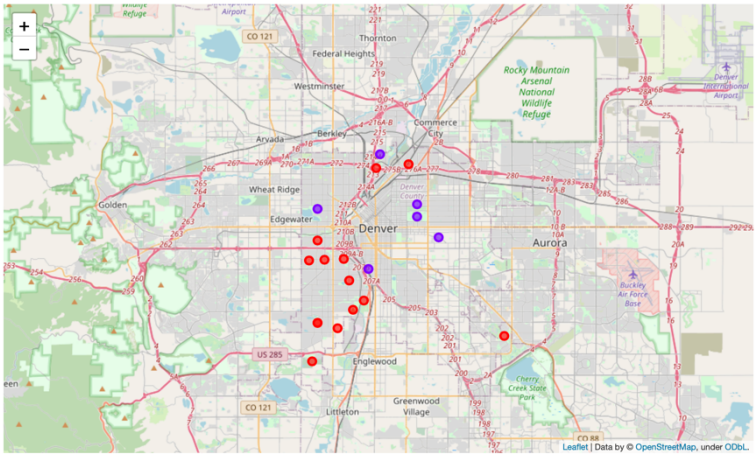
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Athmar Park	Business	Outdoor recreation	Entertainment	Transit	Shopping	Restaurant	Other	Nightlife	Landmark	Hotel
2	Baker	Bar / Brewery	Transit	Restaurant	Other	Nightlife	Shopping	Outdoor recreation	Landmark	Hotel	Health / Personal care
3	Barnum	Food store	Business	Nightlife	Transit	Shopping	Restaurant	Outdoor recreation	Other	Landmark	Hotel
4	Barnum West	Shopping	Outdoor recreation	Business	Transit	Restaurant	Other	Nightlife	Landmark	Hotel	Health / Personal care
13	City Park	Entertainment	Outdoor recreation	Restaurant	Shopping	Landmark	Nightlife	Health / Personal care	Food store	Transit	Other
18	College View	Transit	Food store	Shopping	Business	Athletics / Sports	Restaurant	Outdoor recreation	Other	Nightlife	Landmark
24	Elyria Swansea	Restaurant	Other	Food store	Transit	Outdoor recreation	Hotel	Health / Personal care	Business	Athletics / Sports	Shopping
26	Fort Logan	Other	Transit	Shopping	Restaurant	Outdoor recreation	Nightlife	Landmark	Hotel	Health / Personal care	Food store
28	Globeville	Bar / Brewery	Transit	Shopping	Restaurant	Outdoor recreation	Nightlife	Other	Landmark	Hotel	Health / Personal care
31	Hale	Food store	Restaurant	Outdoor recreation	Health / Personal care	Business	Transit	Shopping	Other	Nightlife	Landmark
34	Harvey Park	Outdoor recreation	Business	Bar / Brewery	Athletics / Sports	Transit	Shopping	Restaurant	Other	Nightlife	Landmark
35	Harvey Park South	Outdoor recreation	Business	Bar / Brewery	Athletics / Sports	Transit	Shopping	Restaurant	Other	Nightlife	Landmark

Secondary Application of Clustering Algorithm

Therefore, I have chosen the Cluster 2 to continue my analysis, adding additional features:

- *Normalized mean housing prices* (obtained from Zillow)
- *Normalized distance from each neighborhood to downtown* (using geopy's distance function)
- *Normalized distance from each neighborhood to Denver International Airport* (using geopy's distance function)

On these new features, I again once again applied the k-means clustering algorithm, setting the number of clusters to 2 (given the reduced size of neighborhoods). I plotted the results on the Denver map:



Results

Given these clustering analyses, we have come to the conclusion that Cluster 2.1 has our ideal characteristics: abundance of bars, restaurants and outdoor spaces as well as proximity to downtown Denver and the airport. We will have to consider that housing prices for this Sub-cluster are higher than those for Sub-cluster 2.0, but perhaps choosing the specific neighborhoods with lower normalized prices (e.g. Baker, Skyland) would solve this issue.

Cluster 2.1

```
denver_final.loc[denver_final['Sub Cluster Labels'] == 1, denver_final.columns[[0] + list(range(13,16)) + list(ra
```

	Neighborhood	Normalized Price	Normalized Distance Downtown	Normalized Distance Airport	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
1	Baker	0.590954	0.006789	0.448596	Bar / Brewery	Transit	Restaurant	Other	Nightlife	Shopping	Outdoor recreation	Landmark	Hotel
4	City Park	0.976423	0.000000	0.067545	Entertainment	Outdoor recreation	Restaurant	Shopping	Landmark	Nightlife	Health / Personal care	Food store	Transit
9	Hale	0.731874	0.163077	0.019909	Food store	Restaurant	Outdoor recreation	Health / Personal care	Business	Transit	Shopping	Other	Nightlife
15	Skyland	0.608967	0.040752	0.036994	Business	Shopping	Food store	Transit	Restaurant	Outdoor recreation	Other	Nightlife	Landmark
16	Sloan Lake	1.000000	0.186884	0.565382	Shopping	Restaurant	Outdoor recreation	Food store	Bar / Brewery	Transit	Other	Nightlife	Landmark
17	Stapleton	0.918158	0.266979	0.135911	Business	Athletics / Sports	Food store	Shopping	Restaurant	Health / Personal care	Transit	Outdoor recreation	Other

Discussion

There are several observations I made during this project that I would like to highlight, with the hope of helping others realize similar analyses:

- The geolocator function is wonderful for finding coordinates, but users should be careful to review, as some coordinates are incorrect, and many times not found. This makes the potential for automation of this type of analysis impossible.
- Foursquare categorization of venues is very granular, and I found that created a more high-level categorization was helpful for a more effective application of the k-means clustering algorithm

Conclusion

I thoroughly enjoyed performing this neighborhood analysis on my hometown of Denver, Colorado, especially given the need my husband and I have for choosing neighborhoods to search for our next house. This analysis could be useful for others who are also searching for their ideal location to buy a house within Denver, or also for businesses looking for a location to open. I hope that you have found this analysis interesting, as I have!