Where to live? A personalized neighborhood identifier for expats

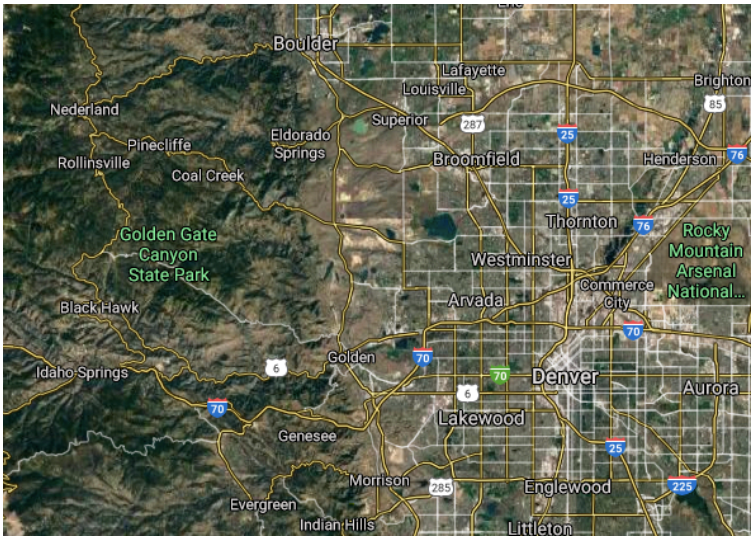
Capstone Project report

1. Introduction

Where you live to a high degree determine the life you will be living. The real estate prices determine how much money you'll have left for things more fun than rent. The shops, restaurants, parks, recreation opportunities, etc., all have an influence on what your everyday life will and can look like.

It can be hard enough to choose a neighborhood or city in your own state or country. If you are moving abroad for a new job, determining where to settle down becomes all the more difficult, as you will likely have very limited experience and information about the characteristics of different cities or neighborhood.

My husband and I are experiencing this firsthand. We used to live in a downtown neighborhood of Copenhagen (Denmark), but my husband recently got a new job across the Atlantic Ocean at the outskirts of Denver, Colorado. We know little to nothing about Denver and the nearby cities, so where do we start our search for a new home? Which city and neighborhood would we be most happy in?



This data science analysis sets out to create a model that can help an expat determine where to start the search for a new place to live when moving to a new country. The title says neighborhood, but depending on the size of the cities or towns in question, it might as well be called a city locator guide. Our own situation is used as a case study, but the study should be replicable for other situations.

Evaluation criteria

We have one criterion for our new home location, which is non-negotiable. It must be within a reasonable transportation time to work.

Constraint

- Reasonable transportation time home-workplace

We have also made a list of things that we believe will make us appreciate a local neighborhood. However, we don't know how to prioritize them.

Preferences

- Affordable rental prices
- Nature/trails in the surrounding area
- Presence of coffee shops, breweries and restaurants
- Parks (dog walking opportunities)
- Opportunity to do sports

It should be noted here, that the evaluation criteria used to determine whether an expat would like to live in a location or not, will of course vary widely depending on the expat's preferences, life situation, etc. The criteria listed above are not meant to represent any generalization of what characterizes a good place to life.

Expected outcome

As has been mentioned above, this is a case study, and therefore focused on our situation, location and preferences. Furthermore, this study does not aim at predicting the optimal place to live for us - in reality, the perfect apartment in the third-best neighborhood may be a much better choice than an average-apartment in the best neighborhood. Instead, the outcome of this analysis will be a list of the top neighborhoods for us in the specified area, giving us a much better starting point for our search for a new home.

2. Data

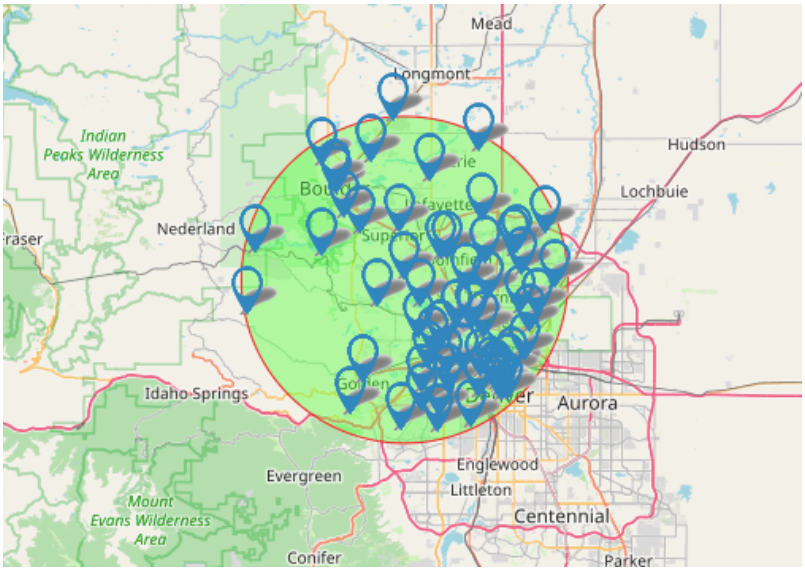
The data for the analysis can be divided into three parts: Data regarding the locations to investigate, data regarding housing prices and data regarding

1. Narrowing down the field - areas to investigate

As the first step, I will define our list of prospect home locations by zipcodes. Defining areas by zipcodes is chosen because the study covers cities and suburbs of different sizes, making it difficult to obtain a full list of neighborhoods to investigate.

The study is limited by our constraint that it must be within reasonable transportation time from home to workplace. Based on this requirement, I have chosen the city of Westminster as the starting point for generating a 15-mile radius to cover our study. This will ensure that the study covers the cities of Denver, Golden and Boulder and that no location is either very far from work nor from the mountains.

The web page Freemaptools.com provides list of zip codes within a certain radius. For this analysis, a list of the 87 zip codes to be found within a 15 miles radius of the zip code 80021 (Westminster) will be used.



It would be nice to connect the zip codes to cities as well as to their geographical coordinates. The Federal Government of the US have created a database for all US zip codes, which can be accesed from the data sharing platform OpenDataSoft. Before retrieving the data, I have restricted the results to only show zip codes in the state of Colorado, which yields 2.487 zip codes. The data can be found here

Merging the two datasets based on zip codes will generate a dataset of all the zip codes within our investigation radius, including the city and latitude and longitude.

2. Budget constraints - housing prices

In order to evaluate the relative affordability of an area, I want to compare the rental prices of the different areas. Rental price is chosen over real estate price, as we don't want to buy anything right know, even if the two would be expected to be highly correlated.

Data on rental prices and median home value by zip codes and cities has been provided by Zillow, an online real estate database company. The data on rental prices is updated monthly and lists rental prices either by house type (for example House 1 bedroom) or by square meter for different zip codes. The data on median home values is updated annual and shows the price per sqm for a home in a given city. The data sets are retrieved from OpenDataSoft and sorted to only include zip codes in Colorado. The fact that the data is already sorted by zip codes makes it easy to compare rental prices across zip codes. The data on rental prices can be found here, and the data on median home values here.

3. Facilities and surroundings - Foursquare

In order to explore the characteristics of the different neighborhoods, such as parks, trails, coffee shops, etc., I will use data provided by Foursquare. Foursquare started as a social networking media, where you could easily let your friends know where you were and what you were doing. It has now developed into *"the most trusted, independent location data platform for understanding how people move through the real world"* according to its own webpage. Foursquare is similar to Yelp and TripAdvisor, and its users rate and recommend everything from restaurants to trails and airport lounges. Foursquare has data about the location of a wide range of different category of 'venues' (locations providing opportunities for shopping, eating, drinking or other activities) as well as knowledge about the popularity of each of these among its users.

Using the Foursquare API, I can collect information regarding venues located in each zip code. This information will be the main data used for the analysis.

3. Methodology

Data collection and preparation

To get the geolocation data, I retrieved the CSV file from OpenDataSoft, as described above. To get the relevant zip-codes, I went to https://www.freemaptools.com/find-zip-codes-inside-radius.htm and searched for zip codes within 15 miles of 80021 (Westminster). The result was a list of zip-codes, which I converted into a dataframe, that I could merge with the dataset on geolocation data in such a way, that the resulting dataset would only include zip codes within my radius of search.

Housing prices
I then took a look at the rental price data, which was a much larger data set (>22,000 rows). I investigated the different house types and price per unit and the available data for each. Using the data per square meter seemed reasonable, as the different sizes of the house units are then taking into account. However, as we would like a house with 1 bedroom, I'd also want to be able to compare zip codes on the rental price of a 1-bdr house. Therefore, I reduced the dataset to a dataset that returns the mean rental prices per sqm and the mean rental price of a 1-bdr house for each zipcode (where available).

The data also revealed that it has been collected over a range of years. While I'm mainly interested in new data, I don't need to predict the rental price, I just want to be able to compare. My data may be biased if the relative price between the zip codes have changed a lot in recent years.

When I merged the rental price data set with the data on geolocation, it turned out that I only had price per sqm data on 36 out of 86 zip codes, and only price of a 1-bdr house for 27 out of 86 zipcodes. Therefore, I decided to use the median home price per sqm instead. This dataset provides data on a city level rather than on a zip code level, which means that it is less detailed, but covers all of the areas we want to look at. Of course, another possible approach would have been to guess the rental price dependent on neighboring available data, but with so little data available, I decided for the median home value instead, which left me with 83 zip codes with geolocation and housing price information.

Foursquare
The next step was to gather information on the facilities and opportunities for each area. Using the Foursquare API, I retrieved information on venues for each of the 83 zip codes using their

geolocation with a radius of 1.5 kilometers (resulting in a diameter of 3 km, which is within the distance that I consider reasonable for being part of my neighborhood). This yielded a little more than 3,000 venue results, with the name and venue category for each, along with their location.

In order to be able to analyze the data, I used hot encoding to create dummy variables for each venue category and then summing them up per zip code, allowing me to see, for example how many coffee shops exist in a given zip code.

Data analysis

Exploring
I first explored my data by looking at the datasets. I had 83 zip codes with 277 different venue categories. I plotted the zip codes on a map, which showed me that many of the zip codes were located in Denver, and that the further out of the city you got, the less dense were the markers on the map, which is as expected.

As part of my exploratory analysis, I created 8 'preference groups' by using a search function on the list of venue categories. For example, I don't care much whether I go to a coffee shop or a café for coffee, so if a venue had either of the words café, tea or coffee in them, they would be added to my preference group 'cafe'. If I *really* loved pizza, I would have created a group just for pizza, but instead I created one for all kinds of restaurants. In my sports group, I chose only the sports I like to do (it really doesn't matter to me whether there is a golf course or not). My personal list of preference groups with their corresponding key words is as follows:

- cafe = 'café|coffee|tea '
- restaurant = 'restaurant|pizza|food|mexican|steakhouse|burger|breakfast'
- bar = 'brewery|beer|bar|pub'
- grocery = 'grocery|convenience'
- shopping = 'shop|store'
- park = 'park'
- nature = 'nature|trail|mountain|river'
- sports = 'yoga|tennis|gym|bike'

Next step was to create a new dataset containing with the preference groups, where the amount of venues within each preference group for each zip code could be seen. An example is given below:

	Area	House price sqm	cafe	restaurant	bar	grocery	shopping	park	nature	sports
0	(80001) Arvada	342	0	0	0	0	0	0	0	0
1	(80002) Arvada	342	3	22	4	4	16	1	0	3
2	(80003) Arvada	342	1	13	2	2	5	1	0	2
3	(80004) Arvada	342	2	8	3	4	5	0	1	1
4	(80005) Arvada	342	0	2	0	1	1	0	1	1

Filtering and sorting
With the preference groups, I could filter my results according to my preferences. I tested different combination, e.g. adding up as many of my preference venues as possible, setting criteria to

minimum 3 cafes, etc. However, the filter I liked the best was the filter requiring at least 1 venue in each preference group. Only 7 zip codes passed this test.

Finally, I also wanted to find the ten most common venues for each of the seven areas, to allow me to explore them a bit more. This would also capture venue groups that I did not include in my preference groups and thus provide me with a more objective picture of the different neighborhoods.

To do this, I simply created a new dataframe, where the venues were sorted based on how frequent they were in each neighborhood.

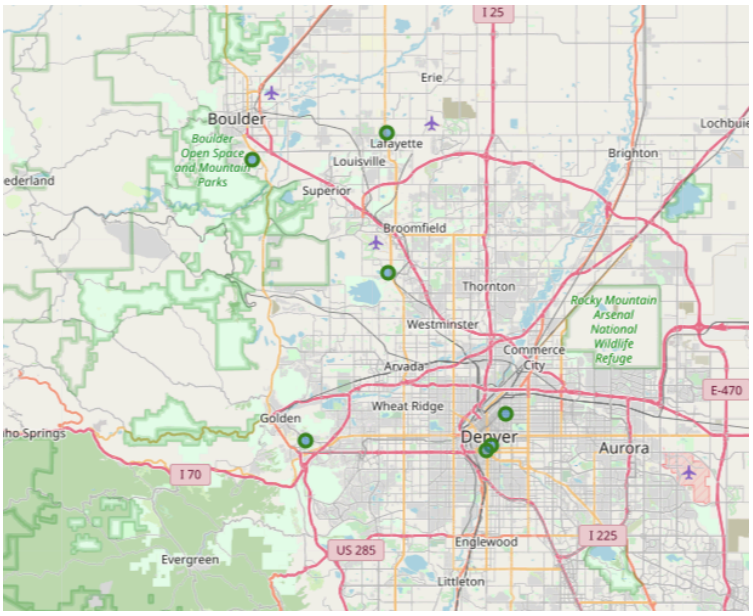
4. Results

The seven areas which passed my criteria are summarized in the table below.

	Area	House price sqm	cafe	restaurant	bar	grocery	shopping	park	nature	sports
8	(80021) Broomfield	269	1	7	2	3	6	2	3	2
10	(80026) Lafayette	481	7	13	6	2	6	1	2	1
18	(80203) Denver	446	5	32	14	2	12	1	1	5
20	(80205) Denver	446	7	20	18	4	9	3	1	1
53	(80273) Denver	446	8	29	17	2	17	1	1	6
64	(80305) Boulder	778	4	7	3	2	8	2	2	3
66	(80401) Golden	725	2	18	2	2	14	1	1	2

As can be seen from the table above, I narrowed down my home search to homes located in Broomfield, Lafayette, Denver, Boulder and Golden. It is obvious (and expected) that the Denver areas score higher on bars and restaurants that the smaller cities closer to the mountain. What surprises me, is that Denver scores equally well on nature as the cities at the foothills of the mountains. I expect this is a result of Foursquare being more oriented towards restaurants and shops than nature experiences.

Next, I wanted to see my proposed living locations, so I plotted my results on a map:



Finally, I produced the table of the 10 most common venues for each area:

	Area	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	(80021) Broomfield	Trail	Home Service	Fast Food Restaurant	Pizza Place	Convenience Store	Shipping Store	Gym / Fitness Center	Sandwich Place	Park	Liquor Store
1	(80026) Lafayette	Coffee Shop	Mexican Restaurant	Pizza Place	Sandwich Place	Salon / Barbershop	Trail	Brewery	Mediterranean Restaurant	Bagel Shop	Department Store
2	(80203) Denver	Sandwich Place	Coffee Shop	Pizza Place	Yoga Studio	Mexican Restaurant	Italian Restaurant	Nightclub	American Restaurant	Marijuana Dispensary	Brewery
3	(80205) Denver	Bar	Brewery	Coffee Shop	Zoo Exhibit	Marijuana Dispensary	Liquor Store	Park	Burger Joint	Convenience Store	Pharmacy
4	(80273) Denver	Coffee Shop	American Restaurant	Brewery	Nightclub	Yoga Studio	Salon / Barbershop	Mexican Restaurant	Italian Restaurant	Marijuana Dispensary	Bar
5	(80305) Boulder	Bus Stop	Coffee Shop	Trail	Indian Restaurant	Café	Park	Pizza Place	Pub	Brewery	Lake
6	(80401) Golden	Pizza Place	Fast Food Restaurant	Mexican Restaurant	Pet Store	Pharmacy	Coffee Shop	Liquor Store	Sandwich Place	Hardware Store	Doctor's Office

5. Discussion

Data quality and availability
As I have already hinted above, the lack of detailed data on rental prices and nature opportunities could mean that my results do not give me a completely accurate picture. The study could definitely be improved by for example combining the data from Foursquare with data on trails or similar.

Lack of an optimisation model
This study has focused more on gathering, sorting and filtering data in a clever way, than on developing a complicated model. The results of this study were not a specific recommendation for the best neighborhood, but rather a set of recommendations for areas to look into. This was chosen, because the available data and my own uncertainty regarding my preferences would not have allowed me to specify a good optimisation model. However, had I known more about trail and rental data (for example rental prices of available homes in the listed areas), an optimisation function could have been developed to suggest the best area.

6. Conclusion

The tables and the map in the result section concludes the job of the data scientist and leaves me with a much better starting point for my house-hunting process. My husband and I are now left with a map of seven potential places to start searching for a new home, all fulfilling our preference criteria. We can use the tables to explore them in more detail, but whether we want to weigh housing prices higher or lower than the amount of coffee shops and nature areas are up to us.

From the table on most common venues, we can for example learn that the most common venue in Boulder i a bus stop, indicating good public transport. Also, Boulder has an abundance of coffee shops and trails, which I really love. Denver provides better opportunities in terms of dining and going out. Broomfield has trails, home service and fast food as top three, indicating that 'city-life' with restaurants and cute coffee shops may be harder to find in Broomfield. Golden has a lot of eating options, but a surprising lack of nature and sports, most likely due to unreliable data on the nature aspects. Oh, well, Foursquare cannot know everything about the world... yet, at least ;)