Where to open a brand new CrossFit box in Denver, CO

Applied Data Science Certification – Capstone Project

Author: Filippo Gronchi

Project Goal and Stakeholder

The goal of this project is to identify the best neighborhood where to open a new Crossfit Box in the city of Denver, Colorado. Crossfit is a pretty new sport involving movements from different disciplines like gymnastics, running, rowing and weightlifting executed either individually or in group classes with motivational music background. This new activity will comprise a large area dedicated to sport (weights, machines and racks, hanging bars, rings, small track for running and floor mats for bodyweight gymnastic), changing rooms/showers plus a recreation and cafe area. Nice to have then a small outdoor areas for summer workouts.

Principal requirements to achieve are:

- the facility should not be close to another similar activity
- the structure should preferably be located in a low crime area
- the selected neighborhood shall present a large young/middle age population i.e. between18 and 65 (although Crossfit could be practiced also by younger and older people and the box owner is encourage to work hard to involve those people).

Possible stakeholder for this task would be a business owner/enterpreneur or a fitness company who wants to get into the CrossFit world in one of the most beautiful cities of US.

Fig. 1: City of Denver Neighborhoods

Project Data

The data to be collected for this analysis come from the several public datasets available on the web (official source www.denvergov.org).

In particular we will make use of:

- Neighborhood list – to have immediately available in a tabularar way the complete Denver neighborhood list and the correct naming convention. From this CSV file we will use the NBHD_NAME column as the baseline for the Denver Neighboorhood Dataframe.

- 2015-2020 Denver Crimes file list – to find out the top 10 Neighborhoods for Crimes and therefore exclude them from the final selection. This huge CSV needs a bit of elaboration and cleaning in order to remove the car accidents (keeping crimes only) and then grouping all the remaing entries by NEIGHBORHOOD_ID (i.e. Neighborhood name).

- 2010 Census Demographic Data – to identify the Neighborhoods with the majority of the population in the target age range. In this case data are already aggregated by NBHD_ID (i.e. Neighborhood name). The column PCT_LESS_18 and PCT_65_PLUS will be used to derive the people percentage in the range 18-65 for each neighborhood.

To quickly fill up the Neighboorhood dataframe with geographical data (i.e. latitude and longitude values) Python GEOPY package will be imported and called within a custom function.

For venue categorization analysis and neighborhood clustering, Foursquare data shall be used via avalable API. In particular we will exclude from the selection all the neighborhood having already one or more venues in the sport category.

Functional Development

Technical background

The data analysis has been completed using Python 3.8.5 on a dedicated Jupyter Notebook. This file along all the images and attached documentations have been placed under revision control using my personal GitHub repository named *Capstone_Project (https://github.com/f1li/Coursera_Capstone)*.

The Python libraries used in this project are:

- numpy
- pandas
- json
- requests
- matplotlib
- seaborn
- sklearn
- folium
- geopy

Analysis Baseline

First step of the analysys is to create a table containing all the possibile locations for the new sport structure. In this case we are talking about Denver neighborhoods which are officially 78. From the denvergov.org web site (an important and authorized repository of informations and public datasets provided by the city council) the Neighboorhood table has been imported in CSV format using pandas function *read_csv*. That dataframe is then integrated with geographical coordinates using Geopy. This step unfortunately is not fully accurate and for about 40% of the neighborhood the values provided are not correct. Therefore at this point a statistical/visual analysis is needed to detect the wrong values. Those numbers are then manually replaced with the correct one in the initial table.

Foursquare

After this preliminary clean-up I have to pull from Foursquare all the venues for each row in the Neighboorhood table. This process is done using my Foursquare "personal" account and the provided API. I get a new dataframe containing for each row a Denver venue, its Category and the Neighborhood it belongs to. Since one of the prerequisites is to avoid area where other sport/fitness are already present, all the Denver sport venues (categories "Gym", "Gym/Fitness Center", "Gym Pool", "Pool", "Athletic & Sports" and containing "Field") are identified and the respective Neighboorhoods are removed from the complete list.

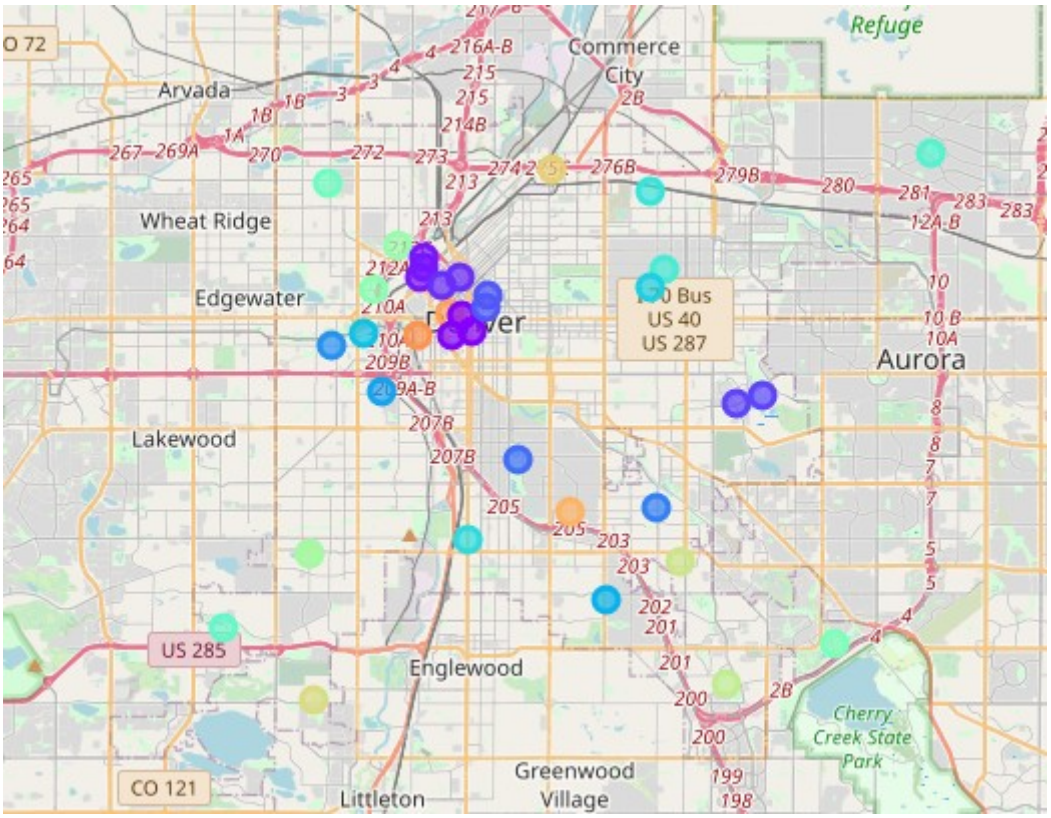

Fig. 2: Denver Sport Venues grouped by Neighboorhood

Using the data retrieved from Foursquare I built a complex Pandas dataframe containing all the Denver venues. For a quick and immediate EDA (exploratory data analysis) of this database I grouped all the venues by Neighboorhood and then sorted the result. A simple bar chart shows the 2 top areas for venues number (windsor and central business distric). Maximum Venue number is topped to 100 because in the Foursquare calls I kept the default limit (100).

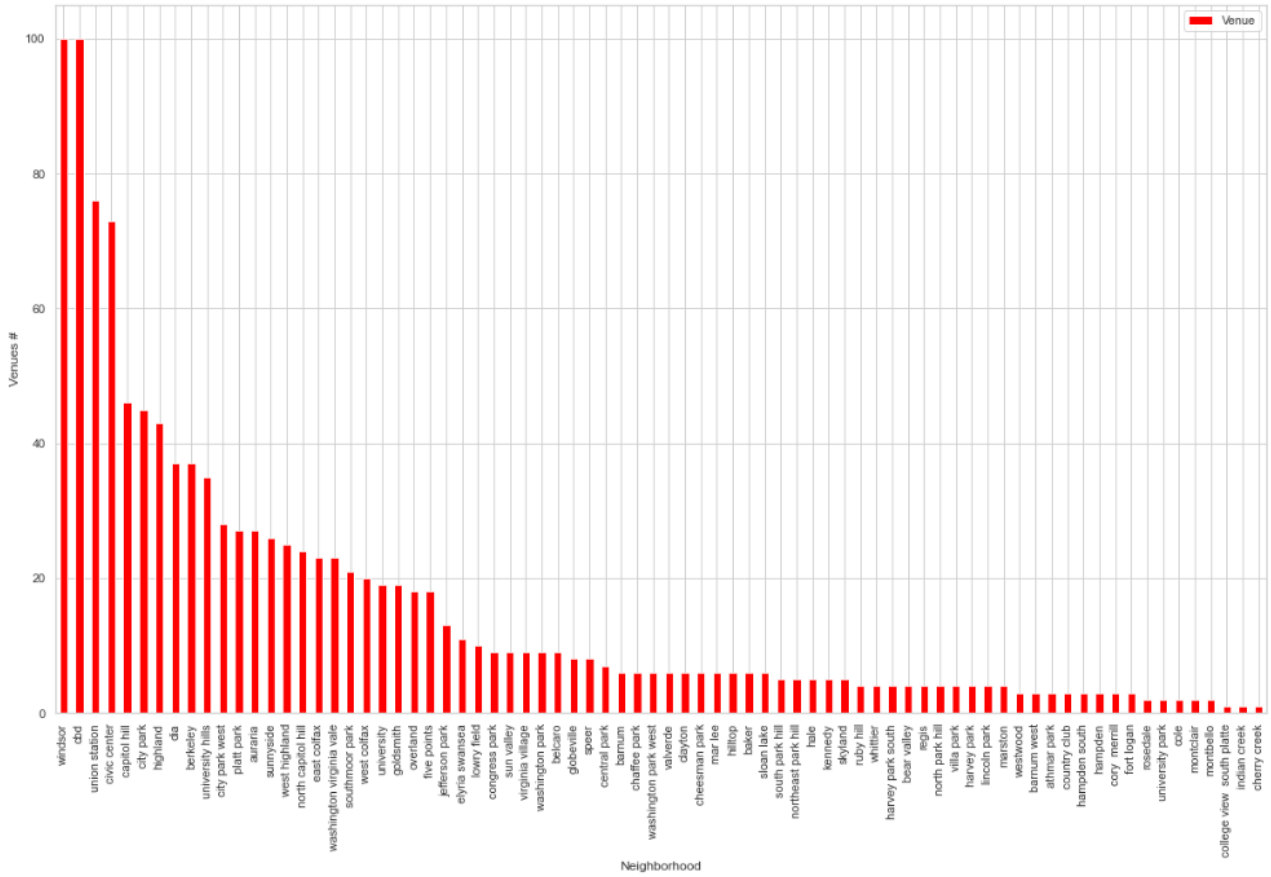

Fig. 3: Denver Venues grouped by Neighboorhood

The other interesting observation to be done on this chart is that only 27 Neighboorhoods (roughly 1 out of 3) present 10 or more venues.

Machine Learning (Clustering)

In order to prepare the Features Matrix for the Neighborhood clustering the Denver venues dataframe is encoded and aggregated by Neighborhood transforming all the features from categorical to numerical. A K-means model is then fed with this data in order to group all the neighborhoods into **5** clusters. The result shows 3 very small sets which shall be excluded from our analysis while focusing on the other big 2. From a quick look to the most frequent categories in each Neighboorhood it's easy to find a common aspect/tipology for each cluster.

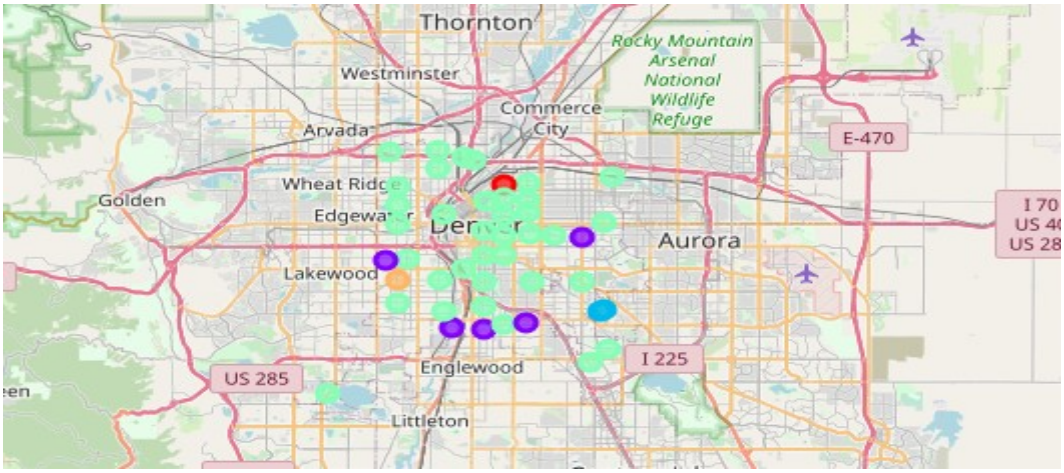

Fig. 4: Denver venues grouped into 5 clusters (different colors)

In particular:

- cluster RED: Restaurants, Stores (1 neighborhood, to be excluded)
- cluster PURPLE: Parks, Food (5 neighborhoods)
- cluster CYANO: Foods (2 neighborhoods, to be excluded)
- cluster GREEN: Food, Recreation, Shops (36 neighborhoods)
- cluster ORANGE: Parks, Zoo, food (1 neighborhood, to be excluded)

Crime dataset

Next step involves the analysis of Crimes in the Denver area in the last 5 years. From the same source a huge CSV (about 536000 rows) is downloaded in my notebook using Pandas. Then with the usual methods (describe, info, shape) a quick EDA is performed and all the not relevant informations are removed. Only the crimes entries are kept (all the car accidents shall be removed). Reported Date field is semplified extracting only the Crime Year. Finally I aggregated the remaing data by Neighboorhood. The Crime distribution chart highlights the outlier Neighboorhood (Five Points).

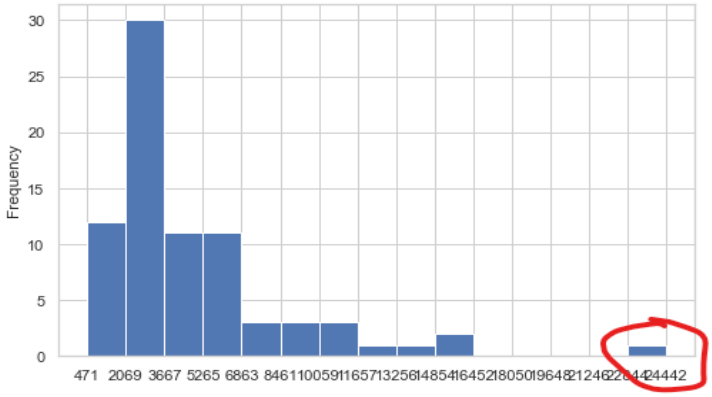

Fig. 5: Crimes distribution

The top 10 Neighboorhoods for Crimes in 2015-2020 are then shown in a bar chart and on the city map.

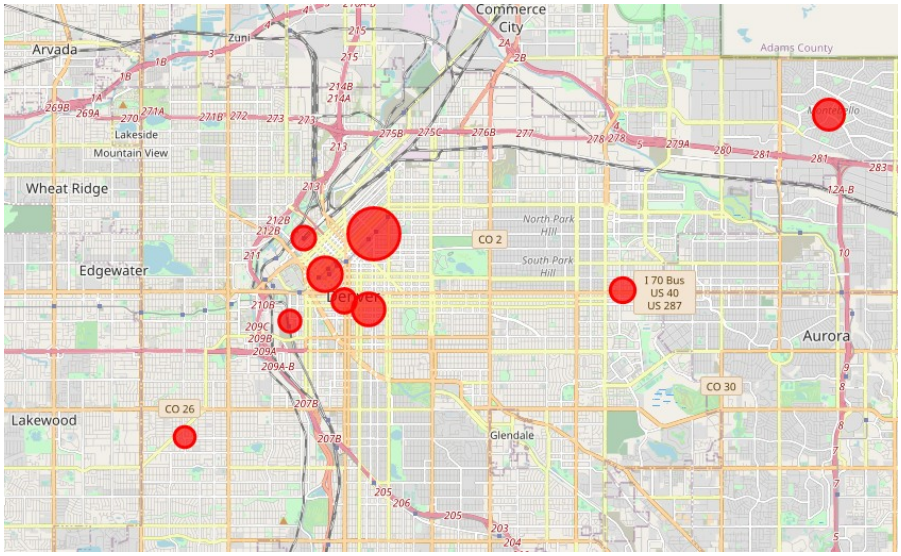

Fig.6: Top 10 Neighboorhoods for 2015-2020 Crimes

These are the 10 neighborhoods **better to avoid**:

- five points
- cbd
- capitol hill
- montbello
- east colfax
- civic center
- union station
- lincoln park
- westwood
- gateway green valley ranch

Demographic dataset

Demographic analysis is based on 2010 Census data from denvergov.org. Pandas as usual is of great help to download CSV and perform basic EDA. Data are already aggregated by Neighborhood. In this case I created a calculated columns to show the Citizens percentage in the range 18-65 years and selected Top 10 neighbordhood to be considered.

The top 10 neighborhoods for citizen age between 18 and 65 **to consider** are:

- auraria
- cbd
- capitol hill
- civic center
- north capitol hill
- university
- speer
- five points
- cheesman park
- washington park west

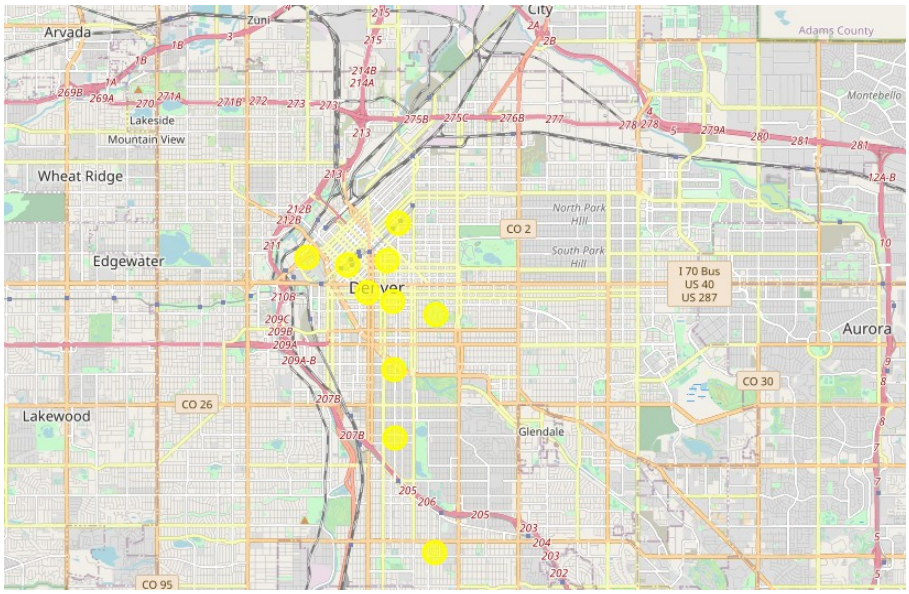

Fig.7: Top 10 Neighboorhoods for 18-65 age % range population

Conclusion

Last step of this analysis is to find which of the 78 Denver neighbordhoods meet all the criteria established with the considerations done above.

Therefore final choice should be done between the following areas.

- **auraria**
- **cheesman park**
- **speer**
- **university**
- **washington park west**

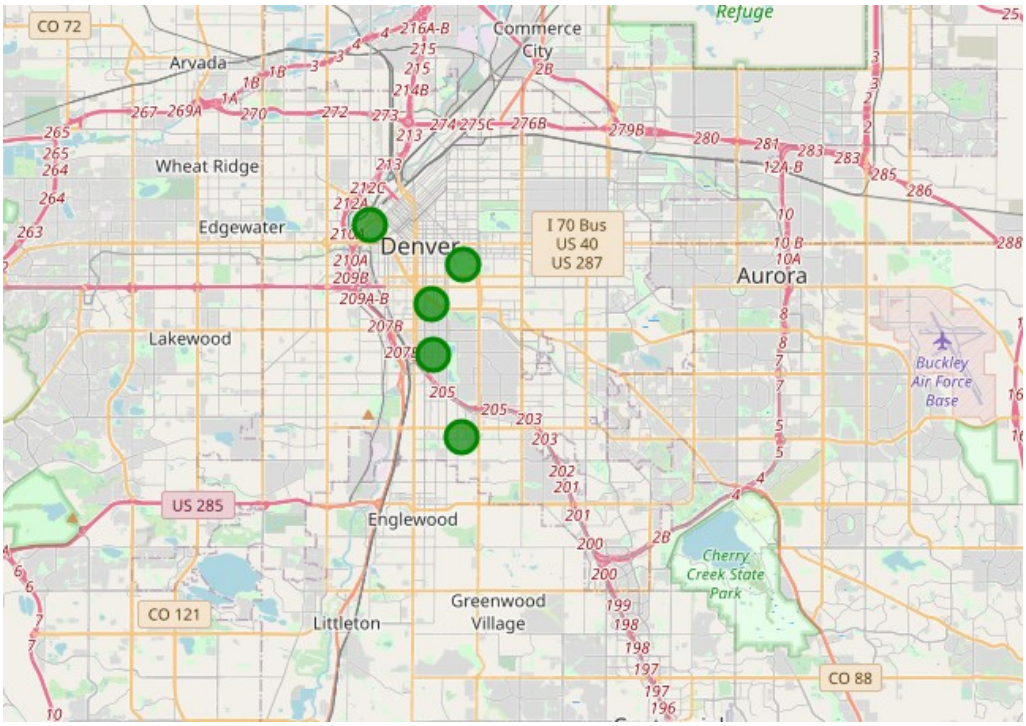

Fig.8: Final 5 Selected Neighboords

As final cross check I pulled manually from the Crossfit HQ web site the full list of all the 19 affiliated structures in Denver. They are located in Highland, Jefferson Park, Five Points, Stapleton, Indian Creek, South Park Hill, East Colfax, Virginia Village, Cbd, Lincoln Park, Civic Center, Athmar Park and Overland. These neighboorhoods are not present in the indicated areas above as expected.

Future Improvements

This analysis represents only the initial approach to understand if such an important investment would be feasible or not. Next investigations shall regard the availability of the foreseen space for the sport facility in those areas in case of new facility or the presence of already suitable buildings to be purchased or rented. In this case a further analisys on the land/rent price has to be performed.