Coursera IBM Data Science Professional Certificate: Capstone Project

Choosing the location of a new brewery in Denver, Colorado, USA

By: Martin Palkovic

Introduction

Craft breweries have exploded in popularity in the United States, and for good reason - they provide an exceptional product that is often paired with a fun ambiance in the brewery. The number of craft breweries in the United States has more than tripled in the last 13 years (Fig 1), and the COVID-19 pandemic aside, growth shows no signs of slowing down. In this data science capstone project, I'll be using data science to determine where a good location for a new brewery would be in Denver, Colorado.

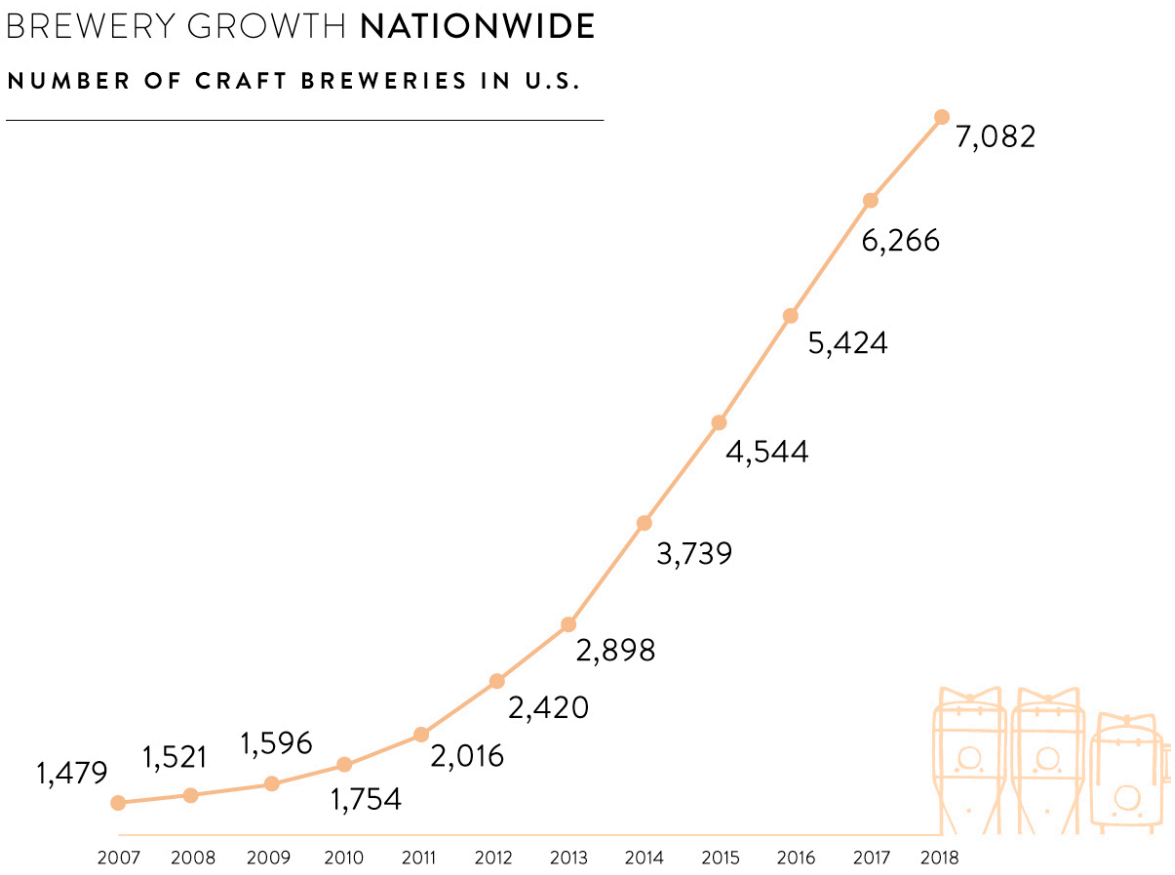

Fig. 1. Number of craft breweries in the U.S (**https://www.crresearch.com/blog/state-craft-beer**)

Business Problem

The objective of this project will be to determine a good location for a new brewery in the city of Denver, Colorado. Denver sits on the arid great plains at the base of the Rocky Mountains in Colorado. Denver is a popular city for outdoorsy young people, and is consistently considered one of the best places to live in the US (https://realestate.usnews.com/places/rankings/best-places-to-live?src=usn_pr). Additionally, the brewery scene is thriving here. There are currently 148 breweries in Denver (https://www.denvermicrobrewtour.com/what-are-top-breweries-in-downtown-denver/#:~:text=Denver%20(and%20the%20surrounding%20metro,the%20city%20of%20Denver%20alone.). Despite the fact that a few neighborhoods are saturated with breweries, there are many other neighborhoods that could likely sustain additional breweries. I'm hypothesizing that a detailed, data science based/machine learning analysis of the current venues in the city will help me clarify which neighborhoods are good candidates for a new brewery.

Target Audience of this project

The target audience for this project would be anyone interested in opening a venue in the city of Denver. Although I'll only be focusing on breweries, the workflow presented herein would be applicable for any venue, be it a restaurant, coffee shop, brewery etc. The data science methodology I use will remain the same for any other venue.

Data

-City of Denver Statistical Neighborhoods (https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-statistical-neighborhoods). The city of Denver has many great, free datasets available for download through their Open Data portal. This is a dataset of polygon features for all 76 named neighborhoods in Denver.

- Foursquare API (https://foursquare.com/). Foursquare is a location data company, with data similar to that of Yelp or Google Maps. The one main advantage of Foursquare is its excellent API, which allows programmers to pull data for their own apps, projects etc. I'll be downloading venue data (restaurants, coffee shops, nail salons, bars etc are all considered venues) for all of Denver, and then I'll clean the data to only display the breweries.

Methodology

First, I began by importing all of the necessary libraries. These included the following:

- Pandas (Data analysis library)¶
- GeoPandas (Data analysis library for geospatial data)¶
- Matplotlib (Data visualization library)¶
- NumPy (Scientific computing library)¶
- Scikit-learn (Machine learning library)¶
- JSON (Reading and writing JSON files)¶

- Requests (used when reading JSON files)¶
- Random (Generates random values)¶
- Folium (Geospatial/mapping library)¶
- Geocoder (retrieves latitude/longitude coordinates for an address)

I then downloaded the 'statistical neighborhoods' dataset from the city of Denver's website as a shapefile, converted the shape file to a json file for ease of use in the Folium library, cleaned the data, and plotted the data in Folium. Then, I defined a function to call the Foursquare API, which downloaded data for venues in Denver by their neighborhood. After that, I employed a technique referred to as 'One Hot Encoding' - this quantifies the percentage of venues that a particular venue occupies in a given neighborhood. For example, if a neighborhood has a one hot value of 0.1 for the 'Coffee Shop' column, that means that 10% of the venues in that neighborhood are coffee shops. Once I have these 'one hot' values, I cleaned the data to produce a new data frame for just the breweries in a neighborhood. Lastly, I run the K-Means machine learning algorithm (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html) on the data to group the data into clusters, plot the data in Folium, and generate a new data frame.

Results

Neighborhoods were grouped in to 5 clusters, ranging from 0 to 4. Neighborhoods falling within clusters 0 or 4 have a large number of existing breweries. Breweries in clusters 0 and 4 encompass 10-14% of total venues in those neighborhoods. Neighborhoods in cluster 2 and 3 have a lower percentage of breweries, while all neighborhoods in cluster 1 (with the exception of 'Union Station') have no breweries (Tables 1-5).

Cluster_0

	Neighborhood	Brewery	Cluster Labels	Longitude	Latitude
26	Five Points	0.1324	0	-104.9834	39.7591
50	Overland	0.125	0	-104.9931	39.6814
65	Valverde	0.1429	0	-105.0152	39.7172

Table 1. Dataframe for cluster 0. These three neighborhoods have the highest number of breweries, comprising ~12-14% of total venues in those neighborhoods.

Cluster_1

	Neighborhood	Brewery	Cluster Labels	Longitude	Latitude
17	Clayton	0.0	1	-104.9501	39.7674
35	Harvey Park South	0.0	1	-105.0399	39.661
14	City Park	0.0	1	-104.9502	39.7456
13	Cherry Creek	0.0	1	-104.9493	39.7194
12	Cheesman Park	0.0	1	-104.9664	39.7346

11	Chaffee Park	0.0	1	-105.012	39.7876
10	Central Park	0.0	1	-104.8791	39.7766
9	Capitol Hill	0.0	1	-104.9802	39.7336
8	CBD	0.0	1	-104.9932	39.7451
6	Belcaro	0.0	1	-104.9505	39.705
5	Bear Valley	0.0	1	-105.0655	39.6611
4	Barnum West	0.0	1	-105.0465	39.7185
18	Cole	0.0	1	-104.9662	39.7672
19	College View - South Platte	0.0	1	-105.0132	39.6724
20	Congress Park	0.0	1	-104.9503	39.7329
34	Harvey Park	0.0	1	-105.0411	39.6753
33	Hampden South	0.0	1	-104.8969	39.6407
32	Hampden	0.0	1	-104.8908	39.661
31	Hale	0.0	1	-104.9316	39.7329
29	Globeville	0.0	1	-104.9868	39.7805
28	Gateway - Green Valley Ranch	0.0	1	-104.7708	39.7849
27	Fort Logan	0.0	1	-105.048	39.6416
30	Goldsmith	0.0	1	-104.9147	39.6738
24	East Colfax	0.0	1	-104.8945	39.7407
23	DIA	0.0	1	-104.6867	39.8579
22	Country Club	0.0	1	-104.966	39.7223
21	Cory - Merrill	0.0	1	-104.9498	39.6905
25	Elyria Swansea	0.0	1	-104.9576	39.7825
75	Windsor	0.0	1	-104.8858	39.7033
16	Civic Center	0.0	1	-104.9912	39.7354
37	Hilltop	0.0	1	-104.9243	39.7187
3	Barnum	0.0	1	-105.0324	39.7184
63	University Hills	0.0	1	-104.9326	39.6642
49	Northeast Park Hill	0.0	1	-104.9233	39.7745
48	North Park Hill	0.0	1	-104.922	39.7565
47	North Capitol Hill	0.0	1	-104.9809	39.7436
46	Montclair	0.0	1	-104.913	39.7329

45	Montbello	0.0	1	-104.8373	39.786
0	Athmar Park	0.0	1	-105.0104	39.7036
44	Marston	0.0	1	-105.0789	39.6275
43	Mar Lee	0.0	1	-105.0391	39.6895
42	Lowry Field	0.0	1	-104.8906	39.7194
41	Lincoln Park	0.0	1	-105.0049	39.7325
40	Kennedy	0.0	1	-104.8635	39.6539
36	Highland	0.0	1	-105.0123	39.7633
52	Regis	0.0	1	-105.044	39.7876
53	Rosedale	0.0	1	-104.9805	39.6731
54	Ruby Hill	0.0	1	-105.0112	39.69
38	Indian Creek	0.0	1	-104.8975	39.6848
73	Westwood	0.0	1	-105.0392	39.704
56	Sloan Lake	0.0	1	-105.0398	39.7516
57	South Park Hill	0.0	1	-104.9221	39.7456
55	Skyland	0.0	1	-104.9502	39.7565
60	Sunnyside	0.0	1	-105.0119	39.7765
61	Union Station	0.0111	1	-105.0004	39.7534
62	University	0.0	1	-104.9664	39.6745
58	Speer	0.0	1	-104.9803	39.7192
74	Whittier	0.0	1	-104.9666	39.7564
66	Villa Park	0.0	1	-105.0392	39.7305
67	Virginia Village	0.0	1	-104.9243	39.6869
68	Washington Park	0.0	1	-104.9663	39.7012
69	Washington Park West	0.0	1	-104.9799	39.7026
70	Washington Virginia Vale	0.0	1	-104.9153	39.7015
64	University Park	0.0	1	-104.9502	39.6759

Table 2. Dataframe for cluster 1. These neighborhoods have no breweries with the exception of the Union Station neighborhood, where breweries are ~1% of the total venues in that neighborhood.

Cluster_2

	Neighborhood	Brewery	Cluster Labels	Longitude	Latitude
51	Platt Park	0.0357	2	-104.9811	39.6866
39	Jefferson Park	0.0526	2	-105.0192	39.7522
2	Baker	0.0417	2	-104.9962	39.7152
71	West Colfax	0.0455	2	-105.0386	39.74
72	West Highland	0.0323	2	-105.0392	39.764
1	Auraria	0.037	2	-105.0083	39.7458

Table 3. Dataframe for cluster 2. These neighborhoods have a low to moderate number of breweries.

Cluster_3

	Neighborhood	Brewery	Cluster Labels	Longitude	Latitude
59	Sun Valley	0.0714	3	-105.0211	39.7358

Table 4. Dataframe for cluster 3. This neighborhood has a moderate number of breweries.

Cluster_4

	Neighborhood	Brewery	Cluster Labels	Longitude	Latitude
7	Berkeley	0.1081	4	-105.0394	39.7767
15	City Park West	0.1034	4	-104.9666	39.7454

Table 5. Dataframe for cluster 4. These neighborhoods have a high number of breweries.

Discussion

Living in the city of Denver and knowing most of these neighborhoods, I think these results can be analyzed in two different ways. One could look at this data and think, 'if I were to open a brewery, I should do it in a neighborhood with NO existing breweries'. However, I think this would lead to poor business at the brewery location, as some of these neighborhoods don't really ‘feel' like prime real estate for a brewery location. In other words, I would posit that there is a reason these neighborhoods don’t have any existing breweries - a brewery simply wouldn’t do well there. Always remember to incorporate qualitative observations into your analysis as

well - the numbers never tell the whole story. If we consider the neighborhoods in clusters 2 and 3, these are all up and coming neighborhoods that are popular enough to sustain new brewery businesses, but are not yet saturated with breweries to the point that neighborhoods in clusters 0 and 4 are. Real estate is likely on the cheaper side in these neighborhoods as well, which will be good for a new business owner. I would exclude the neighborhoods of Sun Valley (the only neighborhood in cluster 3) and Auraria from the conversation - Sun Valley is a largely commercial area that incorporates the Denver Football stadium, I'd consider this a poor option. Auraria is adjacent to downtown, contains Metropolitan State University of Denver, and likely does not have any real estate available.

Lastly, I wouldn't totally write off the neighborhoods in cluster 0 and 4, there's a reason these neighborhoods have a lot of breweries. They're simply cool, hip, well established neighborhoods. While real estate is likely to be expensive in these locations, I'd consider these a good second tier option if none of the neighborhoods in cluster 2 work out.

Conclusion

If you're planning to open a new brewery in Denver, I'd choose one of the following neighborhoods (Table 6). These are all up and coming, popular neighborhoods that could easily sustain a quality brewery, and some of these will likely have lower commercial real estate prices than neighborhoods in cluster 0 or 4.

Cluster_Final

	Neighborhood	Brewery	Cluster Labels	Longitude	Latitude
51	Platt Park	0.0357	2	-104.9811	39.6866
39	Jefferson Park	0.0526	2	-105.0192	39.7522
2	Baker	0.0417	2	-104.9962	39.7152
71	West Colfax	0.0455	2	-105.0386	39.74
72	West Highland	0.0323	2	-105.0392	39.764

Table 6. Martin's final list of recommended neighborhoods to build a new brewery in Denver, Colorado using data science and machine learning techniques.

References

Best Places to Live methodology. "The 150 Best Places to Live in America, Ranked." *U.S. News & World Report*, U.S. News & World Report, 2020, realestate.usnews.com/places/rankings/best-places-to-live?src=usn_pr.

"Complete Guide To The Top Breweries in Downtown Denver." *Denver Microbrew Tour*, 4 Mar. 2020, www.denvermicrobrewtour.com/what-are-top-breweries-in-downtown-denver/.

"The State of Craft Beer." *C + R Research*, www.crresearch.com/blog/state-craft-beer.