Capstone Project: Battle of Neighborhoods
Applied Data Science Capstone by IBM/Coursera
Table of Contents

Introduction: Business Problem

In this project we will try to find a shortlist of nieghbourhooods in and around Dayton (Ohio, USA) that are preferrable for opening a Brewery.

The idea is to find nieghborhoods around population centers, particularly ones with a large population above 18 years of age, and high income brackets, to ensure a large clientele. The brewery would fare even better well if there isn't too much competition from other breweries in the immediate vicinity.

We will use data science powers to cluster and select a cluster with a few most promissing neighborhoods guided by the above criteria. Advantages of these neighbourhoods will then be expressed so that best possible final location can be chosen by stakeholders.

Data

Based on definition of our problem, factors that will influence our decission are:

- number of people in the neighbourhoods
- number of people in the negihbourhoods above a certain age (18 years)
- markers of spending ability through per capita income and median housing value
- venues in the neighborhood that increase the appeal of the neighborhoods and in turn the likelihood, of customers, to visit the brewery
- number of and distance to bars/breweries in the neighborhood, if any

For each city in the counties surrounding and containing the greater Dayton matropolitan area.

Following data sources will be needed to extract/generate the required information:

- the name of neighbourhood in the 4 counties surrounding Dayton (Montgomery, Miami, Clark, and Greene) will be obtianed from the pgeocode package in python (! pip install pgeocode)
- the US census data will be accessed via its api for demographic information in these neighborhoods (https://api.census.gov/data/)
- number of venues and their type and location in every neighborhood will be obtained using **Foursquare API**

Methodology

DataSet 1: obtain a list of zip-codes relevant to Ohio and sub-select those in the counties surrounding Dayton, Ohio

Steps:

- A list of zip-codes that that include all zip-codes in OH are generated
- The zip-codes are queried to the 'pgeocode' database to return the name of the neighborhood/city, county, state and the latitude and longitude of the center of the neighborhood.
- Those neighborhoods that belong first to Ohio and then to the 4 counties surrounding Dayton (Greene, Montgomery, Clark, Miami counties) are sub-selected

	postal_code	place_name	county_name	latitude	longitude
9	43010	Catawba	Clark	39.9991	-83.6222
2300	45301	Alpha	Greene	39.7117	-84.0233
2304	45305	Bellbrook	Greene	39.6402	-84.0824
2306	45307	Bowersville	Greene	39.5806	-83.7249
2307	45308	Bradford	Miami	40.1286	-84.4293

DataSet 2: Obtain the total population, population above 18 years, estimated income/per capita income, and median house values from U.S. census database

Steps:

- A U.S. census API is accessed (https://api.census.gov/data) after creating an apiKey
- The dataset for 2018 is selected as the most recent database which was averaged over 5 years, this was chosen to avoid errors in single year data.
- The database was scoured for relevant variable, not an easy task considering the very large number of sub-categories.
- The following categories were chosen:
    - Total population
    - Total males above 18
    - Total females above 18
    - Per-capita income
    - Median housing value

  The population size, particularly those above 18 years, was chosen to add features with insight into the size of the clientele. The per-capita income and median house value were chosen to further add information about the spending capacity of the clientele

The housing value also can be used as a surrogate for the cost of real-estate price in these neighborhoods.

	zipcode	population	Male_above_18	Female_above_18	Income	PropertyCost
0	43010	260	100	90	23990	105200
1	45384	2053	894	1014	7159	115700
2	45319	378	141	149	27630	125000
3	45361	328	142	114	20548	102300
4	45372	319	122	132	25755	88200

DataSet 3: obtain venues around the centers of the neighborhoods from FourSquare API

Steps:

- The venues around the centers of the neighborhoods are explored.
- The results are queried for the name, location and category of the venue in the neighborhood.
- Limits are set to 100 results within 2 miles of the neighborhood center
- The venue category feature is used to create dummy values and the dataframe is grouped by the neighborhood to determine the mean value of each venue category in each neighborhood

	Neighborhood	Yoga Studio	ATM	Accessories Store	Airport	Airport Service	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	...	V / F
0	43010	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00	0.0	...	0
1	45305	0.0	0.000000	0.0	0.0	0.0	0.028571	0.0	0.00	0.0	...	0
2	45307	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00	0.0	...	0
3	45308	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00	0.0	...	0
4	45309	0.0	0.000000	0.0	0.0	0.0	0.095238	0.0	0.00	0.0	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...
65	45502	0.0	0.010101	0.0	0.0	0.0	0.030303	0.0	0.00	0.0	...	0
66	45503	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00	0.0	...	0
67	45504	0.0	0.010000	0.0	0.0	0.0	0.050000	0.0	0.01	0.0	...	0
68	45505	0.0	0.000000	0.0	0.0	0.0	0.022989	0.0	0.00	0.0	...	0
69	45506	0.0	0.000000	0.0	0.0	0.0	0.027778	0.0	0.00	0.0	...	0

70 rows × 257 columns

These categories are merged with the pervious 2 datasets to create a single dataset with population, demographics, income and venues of each neighborhood.

	zipcode	population	Male_above_18	Female_above_18	Income	PropertyCost	Yoga Studio	ATM	Accessories Store
0	43010	260	100	90	23990	105200	0.0	0.000000	0.0
1	45384	2053	894	1014	7159	115700	0.0	0.000000	0.0
2	45319	378	141	149	27630	125000	0.0	0.000000	0.0
3	45361	328	142	114	20548	102300	NaN	NaN	NaN
4	45372	319	122	132	25755	88200	0.0	0.000000	0.0
...	...	...	...	...	...	...	...	...	...
69	45504	17639	6675	7337	27799	118900	0.0	0.010000	0.0
70	45309	12439	4630	5150	32710	147700	0.0	0.000000	0.0
71	45369	3492	1336	1389	30400	157400	0.0	0.000000	0.0
72	45373	35989	13144	14119	31306	150600	0.0	0.047619	0.0
73	45420	25145	9618	10048	24583	82400	0.0	0.000000	0.0

71 rows × 262 columns

The top 10 most common venues in each neighborhood are determined in order to add more substance and aid in describing the neighborhoods

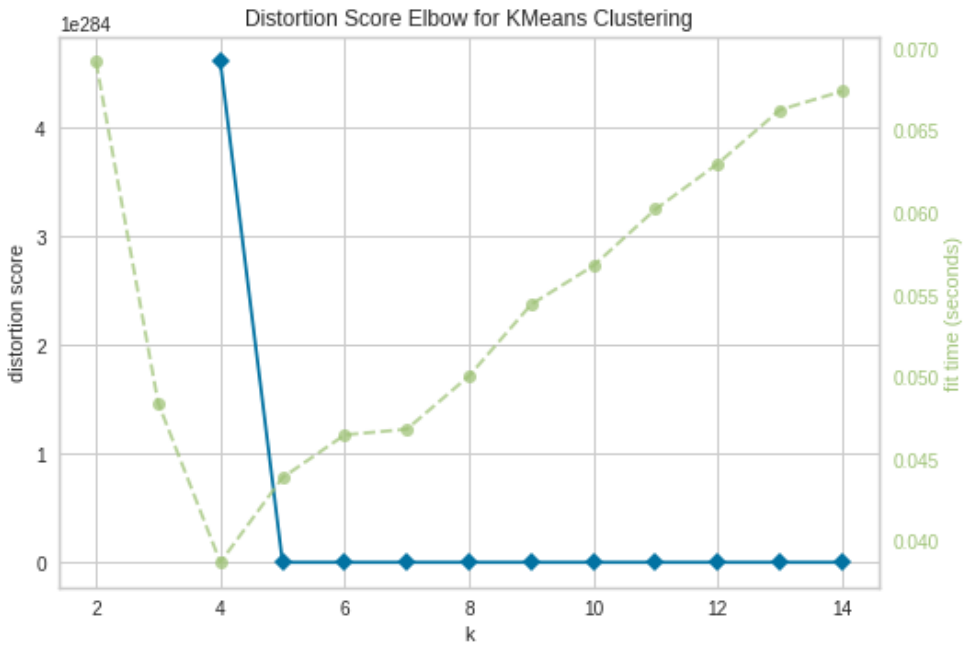
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	43010	Pub	Hot Dog Joint	Food Truck	Food Stand	Food & Drink Shop	Food	Fondue Restaurant	Flower Shop	Flea Market	Fish & Chips Shop
1	45305	Bank	Sandwich Place	Grocery Store	Pizza Place	Discount Store	American Restaurant	Fast Food Restaurant	Hardware Store	Supermarket	Mexican Restaurant
2	45307	Trail	Farm	Rest Area	Women's Store	Flea Market	Fabric Shop	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	Flower Shop
3	45308	Grocery Store	Gas Station	Discount Store	Construction & Landscaping	Pizza Place	Fabric Shop	Farm	Farmers Market	Fast Food Restaurant	Fish & Chips Shop
4	45309	Gas Station	Discount Store	Sandwich Place	American Restaurant	Fast Food Restaurant	Fried Chicken Joint	Pizza Place	Bank	Theater	Grocery Store

KMeans Clustering:

The neighborhoods are analyzed to identify similar clusters.

Steps:

- The number of clusters first needs to be determined; this is done using the elbow method.



- A value of 5 is appropriate for the number of clusters

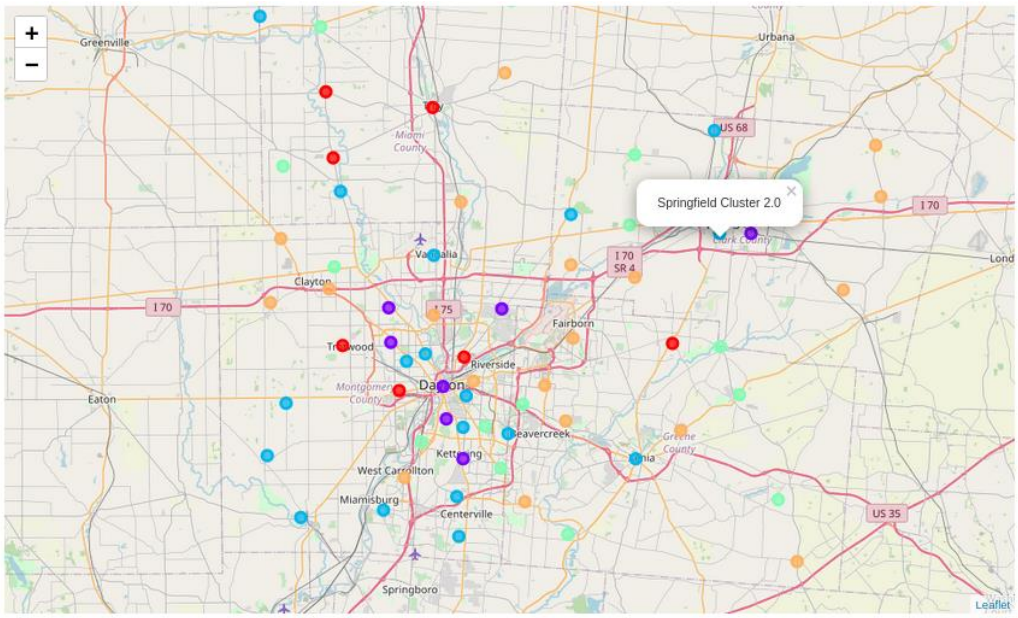
Cluster Labels	postal_code	place_name	county_name	latitude	longitude	population	Male_above_18	Female_above_18	Income	PropertyCost	1st Most Common Venue	2nd M Comm Ve
4.0	43010	Catawba	Clark	39.9991	-83.6222	260	100	90	23990	105200	Pub	Hot
4.0	45305	Bellbrook	Greene	39.6402	-84.0824	11161	4134	4253	43994	207600	Bank	Sandw P
4.0	45307	Bowersville	Greene	39.5806	-83.7249	330	119	124	18409	78800	Trail	F
2.0	45308	Bradford	Miami	40.1286	-84.4293	4139	1620	1585	26801	106900	Grocery Store	Sta
4.0	45309	Brookville	Montgomery	39.8414	-84.4165	12439	4630	5150	32710	147700	Gas Station	Disc S
...	...	...	...	...	...	...	...	...	...	...	...	...
4.0	45502	Springfield	Clark	39.9242	-83.8088	15779	6205	6420	33828	164800	Fast Food Restaurant	P P
3.0	45503	Springfield	Clark	39.9528	-83.7804	32363	11871	13617	26647	114000	Pizza Place	Fast F Restau
3.0	45504	Springfield	Clark	39.9408	-83.8343	17639	6675	7337	27799	118900	Sandwich Place	P P
1.0	45505	Springfield	Clark	39.9106	-83.7856	19919	7079	8072	19355	71900	Fast Food Restaurant	P P
2.0	45506	Springfield	Clark	39.9104	-83.8275	13868	4626	5375	20055	70600	Fast Food Restaurant	Disc S

71 rows × 20 columns

Results

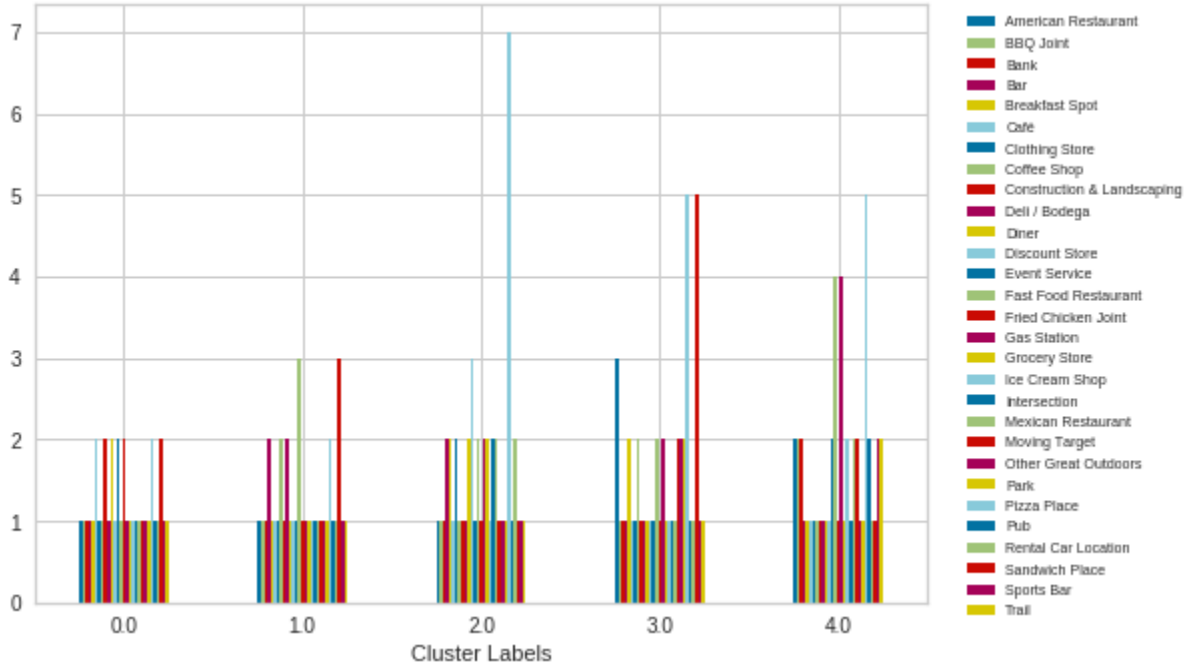
Mapping for visualization:

Folium is used, centered on the mean of all neighborhood centers, to visualize the location and cluster of each neighborhood



Predict cluster properties and labels:

The number of venues in each category for each neighborhood in the cluster are summed for each cluster.



The top 3 venues are used as the label for each cluster:

	top 3 venues
0.0	[Fried Chicken Joint, Event Service, Sandwich ...
1.0	[Sandwich Place, Fast Food Restaurant, Pizza P...
2.0	[Pizza Place, Discount Store, Mexican Restaurant]
3.0	[Sandwich Place, Pizza Place, American Restaur...
4.0	[Pizza Place, Fast Food Restaurant, Gas Station]

The clusters are further analyzed to determine the mean demographics:

Cluster Labels	latitude	longitude	population	Male_above_18	Female_above_18	Income	PropertyCost	Brewery
0.0	39.887771	-84.214229	14377.000000	4999.428571	5868.285714	25566.000000	119757.142857	0.001429
1.0	39.836550	-84.122938	17874.125000	6784.000000	7333.250000	25022.125000	107725.000000	0.013186
2.0	39.788233	-84.180872	15765.833333	5757.555556	6425.277778	30336.111111	127150.000000	0.005811
3.0	39.864500	-84.052288	11296.000000	4243.823529	4599.941176	28317.882353	132917.647059	0.005742
4.0	39.827940	-84.024500	10938.100000	4143.350000	4424.450000	28762.650000	131860.000000	0.004361

Discussion:

Select most appropriate cluster:

Cluster number 2 is found to be the most appropriate for opening a brewery because of the highest per capita income and among the highest populations.

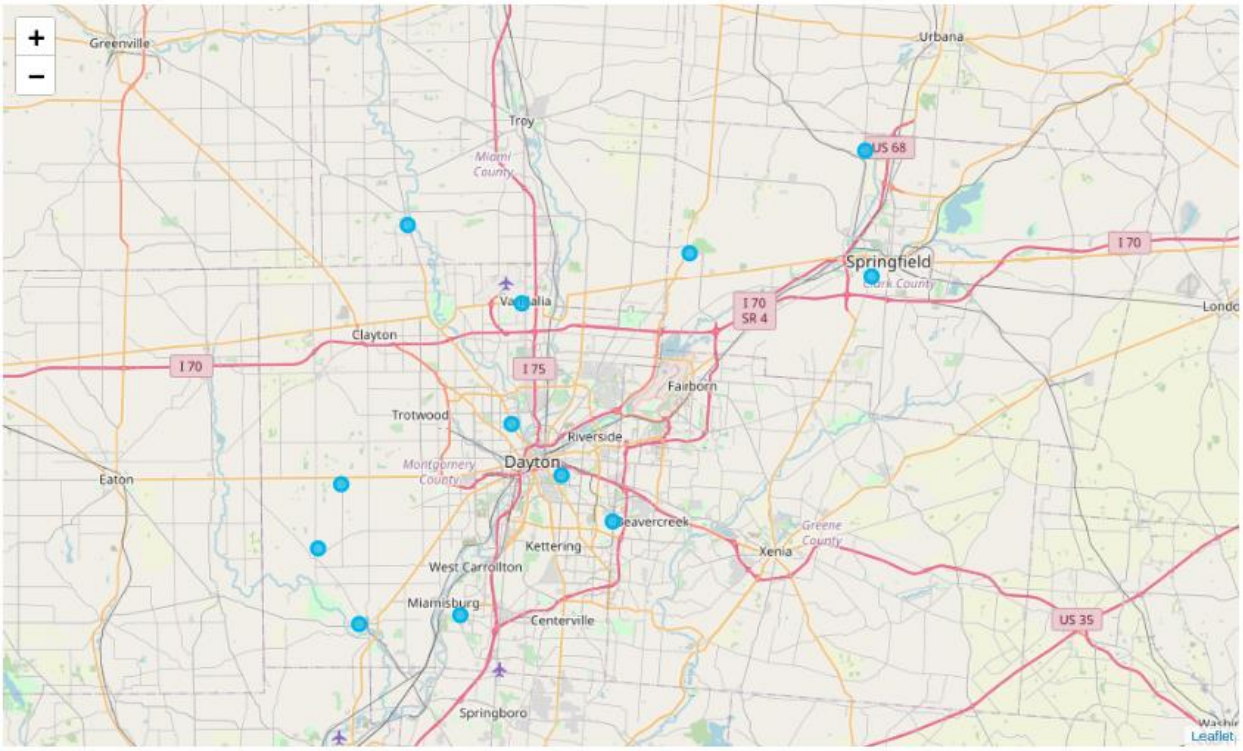
For example: the neighborhood with the highest income is in cluster 2

	Cluster Labels	postal_code	place_name	county_name	latitude	longitude	population	Male_above_18	Female_above_18	Income	...	1st Most Common Venue	Most Common Venue	3rd Cor
52	2.0	45419	Dayton	Montgomery	39.7155	-84.1637	15097	5269	6011	46363	...	Pizza Place	Bar	

The neighborhoods are also reassessed for existing breweries and any with existing breweries are excluded from the recommendation.

12 neighborhoods have been determined:

	Cluster Labels	postal_code	place_name	county_name	latitude	longitude	population	Male_above_18	Female_above_18	Income	...	2nd Most Common Venue	3rd Most Common Venue
15	2.0	45325	Farmersville	Montgomery	39.6867	-84.4205	2266	832	999	31675	...	Pizza Place	Park
17	2.0	45327	Germantown	Montgomery	39.6244	-84.3764	9356	3355	3604	30026	...	Grocery Store	Discount Store
22	2.0	45342	Miamisburg	Montgomery	39.6321	-84.2675	36659	13634	14722	33143	...	Fast Food Restaurant	Sandwich Place
23	2.0	45344	New Carlisle	Clark	39.9300	-84.0217	16104	5802	6149	26279	...	Pharmacy	Intersection
24	2.0	45345	New Lebanon	Montgomery	39.7398	-84.3956	6458	2217	2613	24014	...	Grocery Store	Fast Food Restaurant
34	2.0	45372	Tremont City	Clark	40.0139	-83.8333	319	122	132	25755	...	Airport	Electronics Store
36	2.0	45377	Vandalia	Montgomery	39.8883	-84.2023	15280	5933	6386	33001	...	Fast Food Restaurant	Pizza Place
37	2.0	45383	West Milton	Miami	39.9531	-84.3242	6912	2709	2853	26005	...	ATM	Grocery Store
44	2.0	45405	Dayton	Montgomery	39.7899	-84.2135	16845	6067	7166	20185	...	Fast Food Restaurant	Pharmacy
47	2.0	45410	Dayton	Montgomery	39.7474	-84.1600	16079	6327	6174	20368	...	Pizza Place	Sandwich Place
57	2.0	45430	Dayton	Montgomery	39.7092	-84.1048	7289	2785	2881	40521	...	Ice Cream Shop	Coffee Shop
70	2.0	45506	Springfield	Clark	39.9104	-83.8275	13868	4626	5375	20055	...	Discount Store	Sandwich Place



From the results discovered and presented, the following observations and recommendations can be made:

- Based on the criteria given by the investor group and the cluster data, the main neighborhood recommendation would be for those in cluster 2. This cluster was selected because of high income and generally high population levels as compared to other clusters.
- Additionally, from the sub-selection, it becomes clear that neighborhoods are close to the main population centers in the counties.

Conclusion

Purpose of this project was to identify neighborhoods in Ohio close to Dayton, with a large number of drinking age people who can afford good beer from a local brewery, in order to aid stakeholders in narrowing down the search for optimal location for a new Brewery.

By determining the population and demographics, including an estimate of spending power on the basis of income and housing price, we have identified 12 neighborhoods from all found in the 4 surrounding counties (~90 neighborhoods).

The aid the stakeholders with getting an additional sense of the type of neighborhoods we have included the 10 most common venues in the neighborhoods. The house price can be used as a surrogate to estimate real estate costs in the neighborhoods.

Final decision on optimal brewery location will be made by stakeholders based on specific characteristics of neighborhoods and locations among the 12 shortlisted, taking into consideration additional factors like attractiveness of each location (proximity to park or water), accessibility, real estate availability, prices etc.