July 2, 2020

Applied Data Science Capstone Week5 Assignment

Peer-Graded Assignment: The Battle of Neighborhoods Week 2

Identifying opportunity for retirement business venture between college cities

Introducing the Opportunity:

A client recently shared that upon retirement, they would like to entertain the opportunity of opening a small coffee / music house in one of two college cities where they had lived while a student, to help give back to the origin communities. Those cities are Charlottesville VA and Boulder CO. Now, there are noticeably and apparent differences between the two cities, beginning with geographically and weather patterns that for this exercise will be excluded, as well as employment and consumer purchasing patterns, as well as student population levels. However, there are some commonalities which will be used to generalize through data analysis as to where may be a safe, and competitive opportunity upon retirement.

The approach:

Data sources will be acquired to help investigate the following questions:

- What areas within both cities contain the highest levels of crime?
  - Our client also voiced to focus on vehicular crimes for their customer base.
- Does population density correlate to criminal behaviors?
- Using venue data from Foursquare, what venues and locations are most common in both cities?
- Where is there a greater, and more prosperous opportunity to establish a college level coffee shop / music house in either city?

Data Sources:

To understand and explore the opportunity with greater thoroughness, we will use the following city open data and other access points, as indicated below:

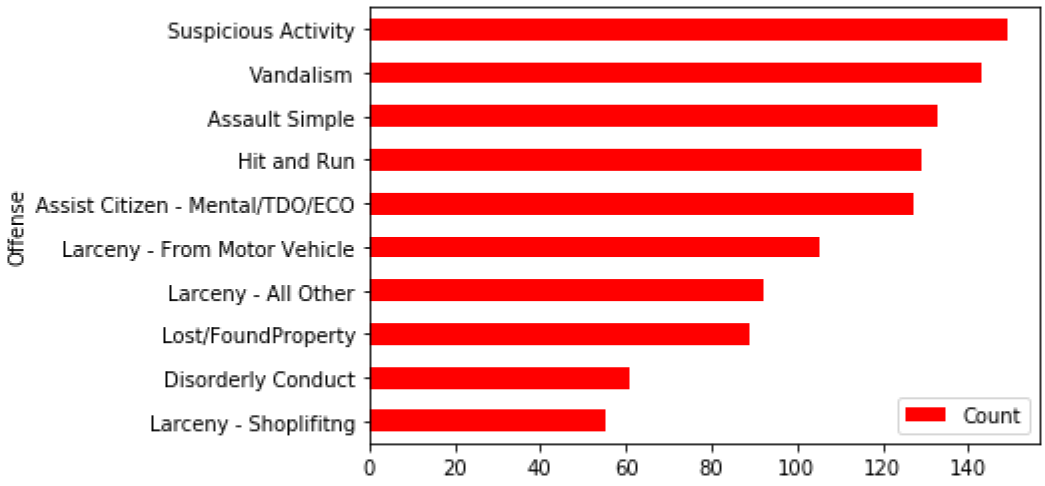
- Boulder CO Open Data Site: https://bouldercolorado.gov/open-data
- Boulder CO Neighborhood Data Site: https://www.bouldercoloradousa.com/about-boulder/boulder-neighborhoods/
- Charlottesville VA Open Data Site: https://opendata.charlottesville.org
- Charlottesville VA Neighborhood Data Site: https://opendata.charlottesville.org/datasets/f4efb475a1ca4b919fca4645b72fadd0_401/data
- Foursquare Venue Data Site: https://foursquare.com/

For this report, data was evaluated for both Charlottesville VA and Boulder CO. For the majority of the reporting to follow, examples will be provided from the Charlottesville VA data set, as the Boulder CO data set data, evaluation, and scrubbing process was virtually similar, save for minor python data functions which were required to be created based on the raw input data. Once the results and discussion section of the report are reached, we will then introduce the Boulder CO content for visualization purposes.

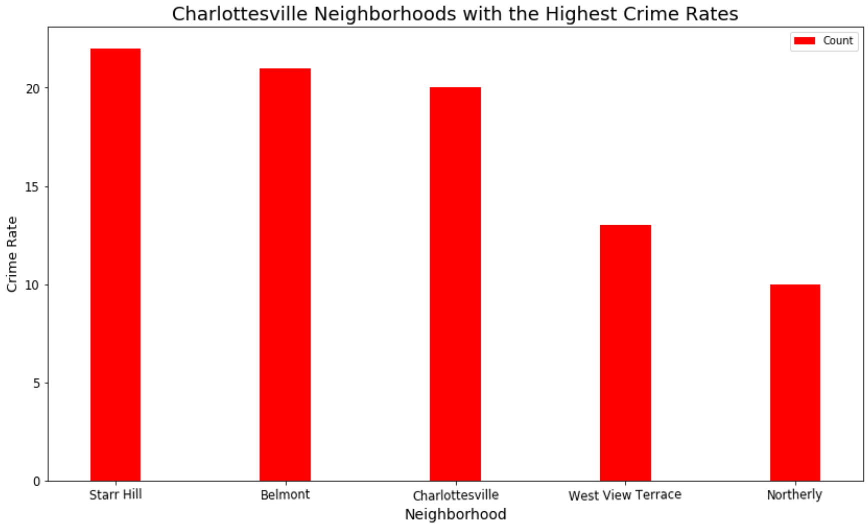
Data Scrubbing

All data sources were retrieved separately and prepared independently. Starting with Charlottesville, the crime data was retrieved from the city open data source and immediately appended the year of each occurrence derived from the actual reporting date. Upon completion, the data set was further filtered to only represent the current year. Street address information data attributes were also scrubbed, for example house numbers or block numbers were modified to integer from float data types. Incidents were categorized based on the offense, and then passed through a custom reverse function that provided additional latitude / longitude coordinates based on the offense street address. Additionally, if the address was incomplete and the function returned an incomplete or incorrect data set, the offense crime record was discarded as an outlier.

Once the data set appeared to be in more proper order, simple bar graphs were used to help visualize offenses both by category and by neighborhoods. In the examples below, the client wanted to examine in greater detail those offenses that dealt with vehicular breaking and theft. Once identifying such offense from the master categories, these were then filtered and transposed back into the neighborhood groupings to better identify those neighborhoods in question where this activity was considered the most repeated.



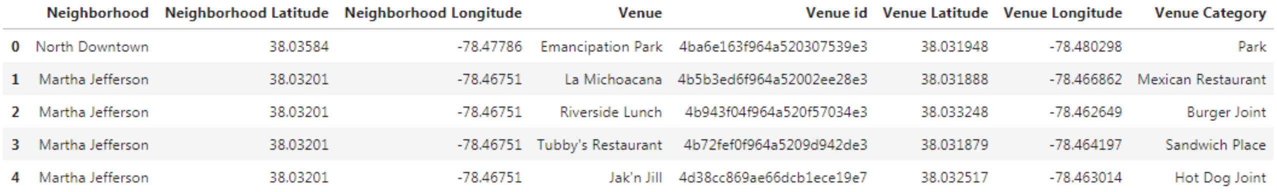
Taking specifically the larceny and theft criminal events into greater consideration, we then went back to the neighborhood data and depicted those current year activities to understand which neighborhoods in question posed the greatest risk.



Data Modelling:

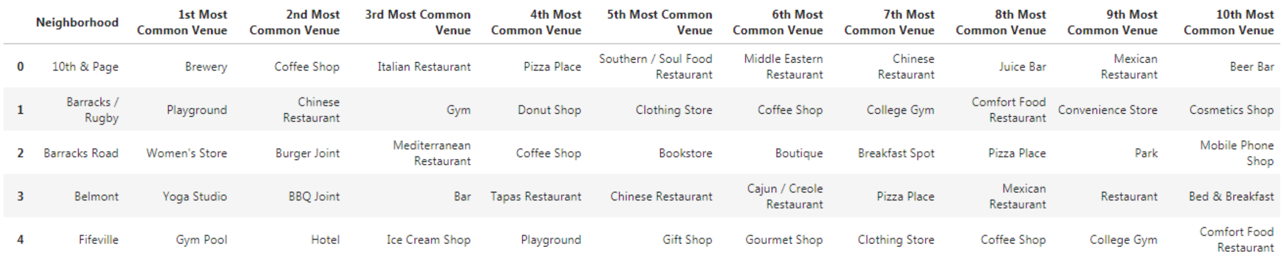
Using the dataframes that help construct the general neighborhood master data, and coupling with the foursquare location data, new merged dataframes per city were created that depict the venue detail data for each corresponding neighborhood. In addition, columns were added to include supporting content such as venue names, venue latitude and longitude, as well as venue category.

Below is an example taken from the Charlottesville VA analysis:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue id	Venue Latitude	Venue Longitude	Venue Category
0	North Downtown	38.03584	-78.47786	Emancipation Park	4ba6e163f964a520307539e3	38.031948	-78.480298	Park
1	Martha Jefferson	38.03201	-78.46751	La Michoacana	4b5b3ed6f964a52002ee28e3	38.031888	-78.466862	Mexican Restaurant
2	Martha Jefferson	38.03201	-78.46751	Riverside Lunch	4b943f04f964a520f57034e3	38.033248	-78.462649	Burger Joint
3	Martha Jefferson	38.03201	-78.46751	Tubby's Restaurant	4b72fef0f964a5209d942de3	38.031879	-78.464197	Sandwich Place
4	Martha Jefferson	38.03201	-78.46751	Jak'n Jill	4d38cc869ae66dcb1ece19e7	38.032517	-78.463014	Hot Dog Joint

Once established, the process of hot encoding as instructed earlier within the curriculum was performed on the data. This approach provides a representation of the category variables as binary vectors. The data was grouped again by neighborhood, and the frequency and mean values of each category were also calculated. The last step was to accumulate by descending order the top 10 most common venues by neighborhood. The final results were placed in a new dataframe. An example of the Charlottesville VA data frame and selected neighborhoods follow:



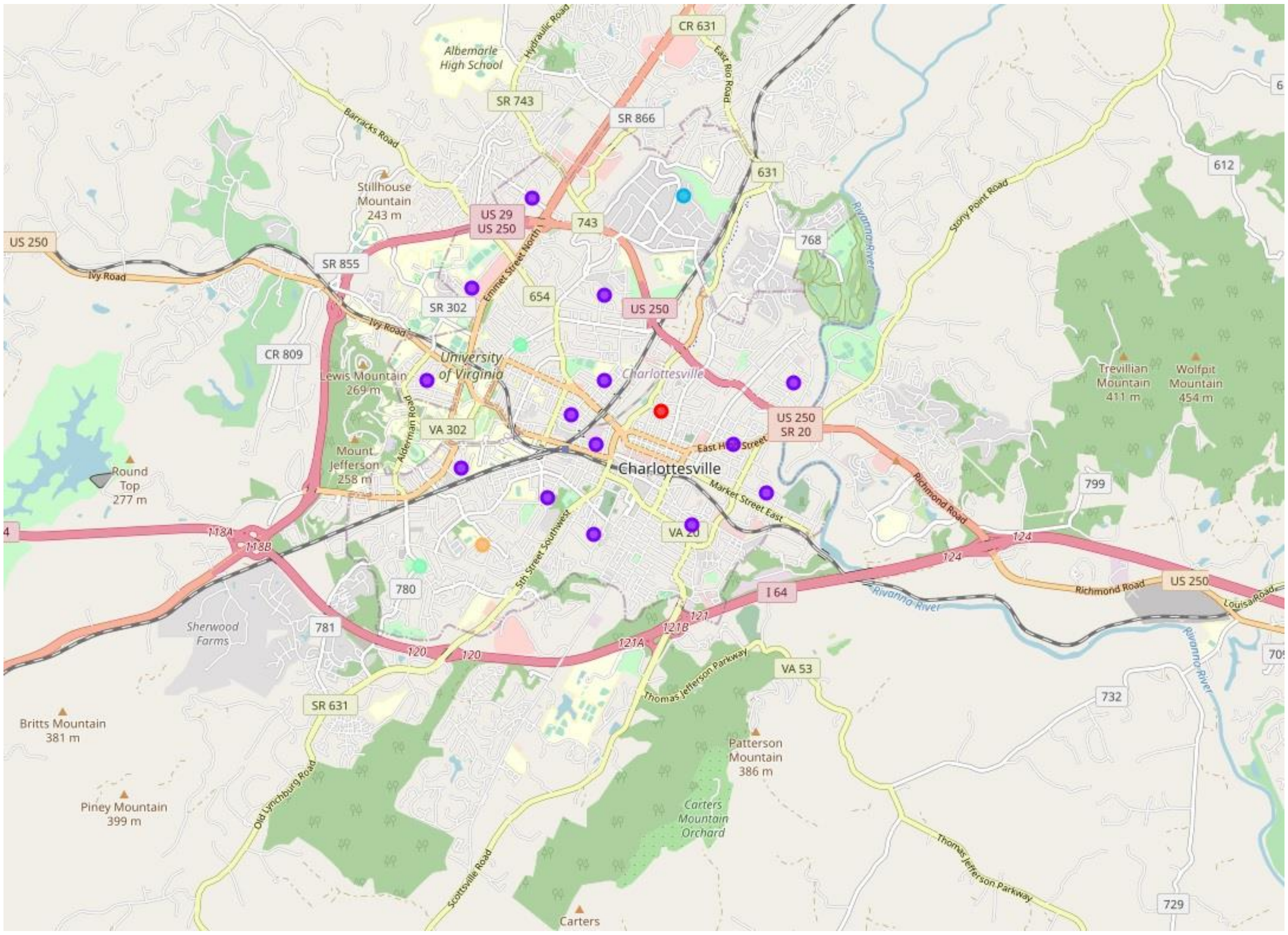
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	10th & Page	Brewery	Coffee Shop	Italian Restaurant	Pizza Place	Southern / Soul Food Restaurant	Middle Eastern Restaurant	Chinese Restaurant	Juice Bar	Mexican Restaurant	Beer Bar
1	Barracks / Rugby	Playground	Chinese Restaurant	Gym	Donut Shop	Clothing Store	Coffee Shop	College Gym	Comfort Food Restaurant	Convenience Store	Cosmetics Shop
2	Barracks Road	Women's Store	Burger Joint	Mediterranean Restaurant	Coffee Shop	Bookstore	Boutique	Breakfast Spot	Pizza Place	Park	Mobile Phone Shop
3	Belmont	Yoga Studio	BBQ Joint	Bar	Tapas Restaurant	Chinese Restaurant	Cajun / Creole Restaurant	Pizza Place	Mexican Restaurant	Restaurant	Bed & Breakfast
4	Fifeville	Gym Pool	Hotel	Ice Cream Shop	Playground	Gift Shop	Gourmet Shop	Clothing Store	Coffee Shop	College Gym	Comfort Food Restaurant

Once the venues by neighborhood have been ordered in descending order, the k-means clustering algorithm was applied to the overall neighborhood content. This algorithm identifies ‘k’ number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is considered one of the more simple and popular unsupervised machine learning algorithms easily utilized today. For our exercise, we chose to categorize our venues into 5 clusters based on the frequency of occurrence. The resulting cluster groups give us the ability to identify which neighborhoods have higher concentrations of similar venues, and to help guide us toward the answer as to which may be best suited for our client’s business needs.

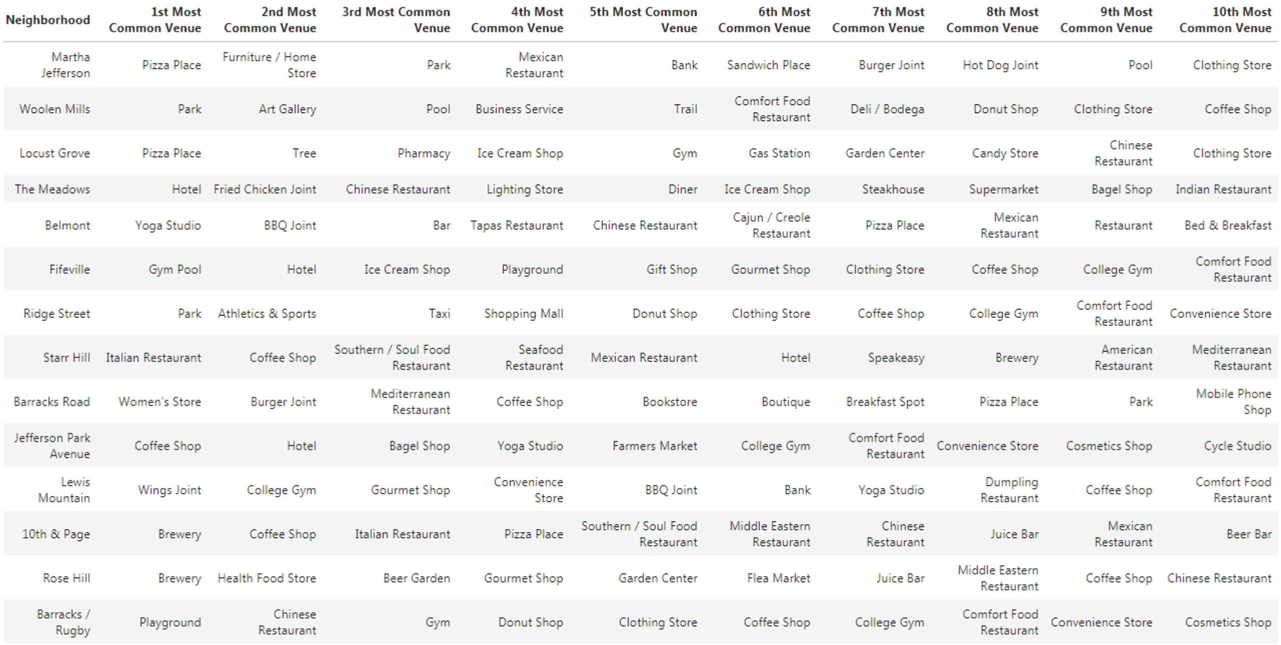
Results Section

Charlottesville VA

Visualizations always help support and tell the story of the underlying data process. In order to better summarize the findings, a geo topographical map generated using the folium library was created and depicted within the map are the color-coded markers for each neighborhood cluster. In the example below, we can see that the purple clusters are the most dominant grouping, indicating that those neighborhoods are very similar in nature to each other in terms of venue offerings. This cluster groups 13 of the 19 neighborhoods being evaluated, while the remaining 5 appear quite independent, 3 of each being slated into its own cluster.

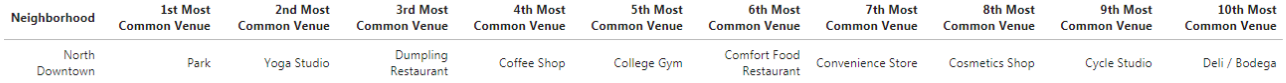


When we investigate these clusters and examine more closely, we first note the purple cluster above, and also depicted in detail below. This is the largest cluster and contains the majority of neighborhoods, including 14 of the 19 neighborhoods under evaluation. Examining the common venues, we can determine that the most common locales in these neighborhoods are food services such as restaurants, pizza shops, bars, gourmet, and national cuisine. This cluster appears to be oriented primarily towards those who choose to experience outside activities, social gatherings, eating out, or attending large gatherings that can be hosted either at the local university or hotel complexes.



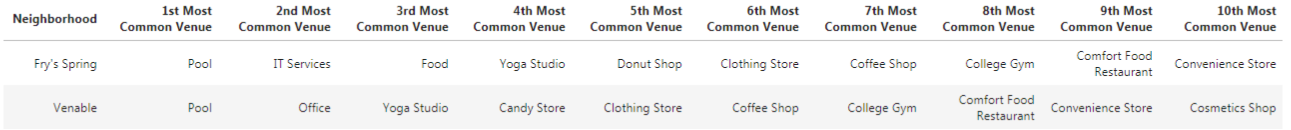
Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Martha Jefferson	Pizza Place	Furniture / Home Store	Park	Mexican Restaurant	Bank	Sandwich Place	Burger Joint	Hot Dog Joint	Pool	Clothing Store
Woolen Mills	Park	Art Gallery	Pool	Business Service	Trail	Comfort Food Restaurant	Deli / Bodega	Donut Shop	Clothing Store	Coffee Shop
Locust Grove	Pizza Place	Tree	Pharmacy	Ice Cream Shop	Gym	Gas Station	Garden Center	Candy Store	Chinese Restaurant	Clothing Store
The Meadows	Hotel	Fried Chicken Joint	Chinese Restaurant	Lighting Store	Diner	Ice Cream Shop	Steakhouse	Supermarket	Bagel Shop	Indian Restaurant
Belmont	Yoga Studio	BBQ Joint	Bar	Tapas Restaurant	Chinese Restaurant	Cajun / Creole Restaurant	Pizza Place	Mexican Restaurant	Restaurant	Bed & Breakfast
Fifeville	Gym Pool	Hotel	Ice Cream Shop	Playground	Gift Shop	Gourmet Shop	Clothing Store	Coffee Shop	College Gym	Comfort Food Restaurant
Ridge Street	Park	Athletics & Sports	Taxi	Shopping Mall	Donut Shop	Clothing Store	Coffee Shop	College Gym	Comfort Food Restaurant	Convenience Store
Starr Hill	Italian Restaurant	Coffee Shop	Southern / Soul Food Restaurant	Seafood Restaurant	Mexican Restaurant	Hotel	Speakeasy	Brewery	American Restaurant	Mediterranean Restaurant
Barracks Road	Women's Store	Burger Joint	Mediterranean Restaurant	Coffee Shop	Bookstore	Boutique	Breakfast Spot	Pizza Place	Park	Mobile Phone Shop
Jefferson Park Avenue	Coffee Shop	Hotel	Bagel Shop	Yoga Studio	Farmers Market	College Gym	Comfort Food Restaurant	Convenience Store	Cosmetics Shop	Cycle Studio
Lewis Mountain	Wings Joint	College Gym	Gourmet Shop	Convenience Store	BBQ Joint	Bank	Yoga Studio	Dumpling Restaurant	Coffee Shop	Comfort Food Restaurant
10th & Page	Brewery	Coffee Shop	Italian Restaurant	Pizza Place	Southern / Soul Food Restaurant	Middle Eastern Restaurant	Chinese Restaurant	Juice Bar	Mexican Restaurant	Beer Bar
Rose Hill	Brewery	Health Food Store	Beer Garden	Gourmet Shop	Garden Center	Flea Market	Juice Bar	Middle Eastern Restaurant	Coffee Shop	Chinese Restaurant
Barracks / Rugby	Playground	Chinese Restaurant	Gym	Donut Shop	Clothing Store	Coffee Shop	College Gym	Comfort Food Restaurant	Convenience Store	Cosmetics Shop

The next cluster, which was categorized as an independent neighborhood, is represented by the red marker in the map. This cluster appears to distinguish recreational, physical, and convenience within minimal travel distance. Coffee shop competition however is discovered to be relatively strong within this cluster so it has been tagged as one to avoid for the purposes of the clients business venture.



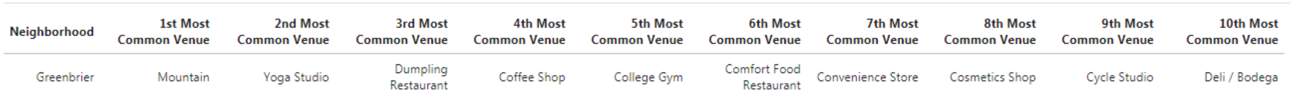
Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
North Downtown	Park	Yoga Studio	Dumpling Restaurant	Coffee Shop	College Gym	Comfort Food Restaurant	Convenience Store	Cosmetics Shop	Cycle Studio	Deli / Bodega

The third cluster indicated by the light green marker groups two recreational venues together. Noticeably, they also are in close proximity to the university campus and facilities.



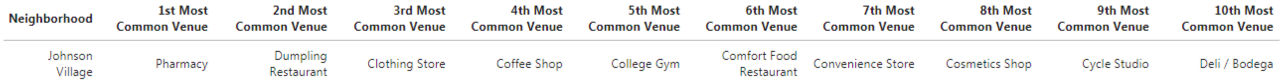
Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Fry's Spring	Pool	IT Services	Food	Yoga Studio	Donut Shop	Clothing Store	Coffee Shop	College Gym	Comfort Food Restaurant	Convenience Store
Venable	Pool	Office	Yoga Studio	Candy Store	Clothing Store	Coffee Shop	College Gym	Comfort Food Restaurant	Convenience Store	Cosmetics Shop

The fourth cluster, highlighted with the blue marker on the map, can be slightly misleading at first. Although it still appears to be close to the city boundaries, the first most common venue, Mountains, is actually close to the entryway towards the Appalachian trail. Coffee shops again appear common within this neighborhood cluster, so it will be as well tagged for the client as one that may not be the best consideration.



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
Greenbrier	Mountain	Yoga Studio	Dumpling Restaurant	Coffee Shop	College Gym	Comfort Food Restaurant	Convenience Store	Cosmetics Shop	Cycle Studio	Deli / Bodega

Finally, the last cluster as indicated with the gold marker on the map is part of a quaint, close quartered community in the south region of the city. It appears to offer many venues suitable for all ages, and again noticing the abundance of established coffee shops in the region may not be the ideal location for the client.

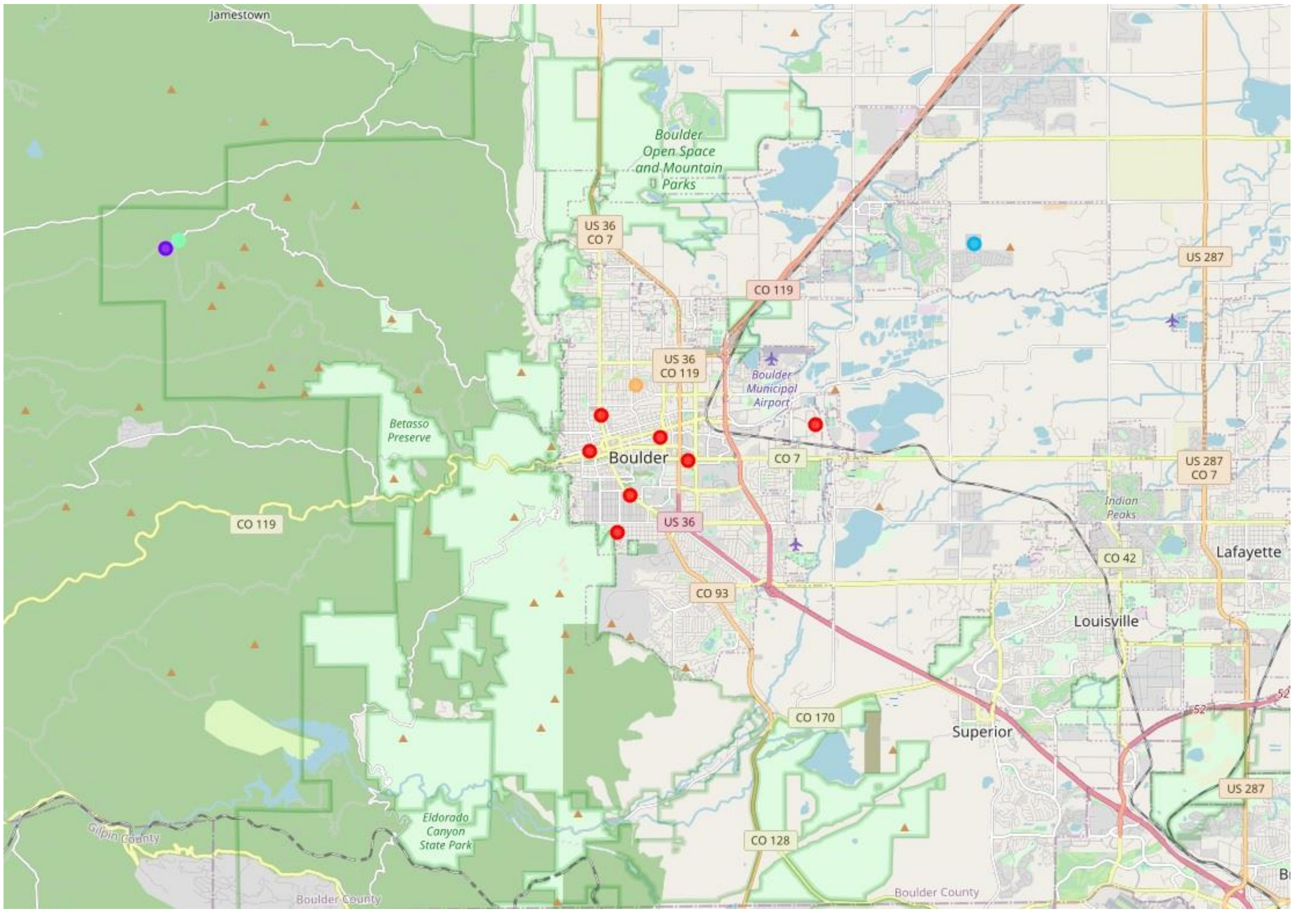


Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Johnson Village	Pharmacy	Dumpling Restaurant	Clothing Store	Coffee Shop	College Gym	Comfort Food Restaurant	Convenience Store	Cosmetics Shop	Cycle Studio	Deli / Bodega

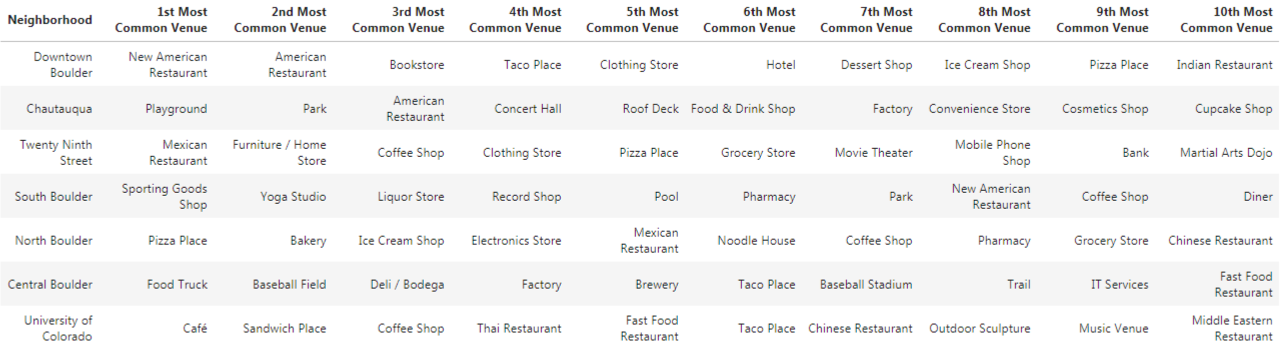
Boulder CO

Here, we will turn our attention to the neighboring city and its results. Similarly, when evaluating the landscape and venue similarities within Charlottesville VA, and again applying the same K-Means clustering algorithms onto the folium topographical map, we can quickly visualize that the 11 neighborhoods being evaluated are grouped, predominately into the red cluster containing 7 of the total neighborhoods.

Similarly, when evaluating the landscape and venue data found within Boulder CO, and again applying the same K-Means clustering algorithms onto the folium topographical map, we can quickly visualize that the 11 neighborhoods being evaluated are grouped predominately into the red cluster containing 7 of the total neighborhoods. In the example below, we can see that the red clusters is the most dominant grouping, indicating that those neighborhoods are very similar in nature to each other in terms of venue offerings. This cluster groups 7 of the 11 neighborhoods being evaluated, while the remaining 4 were so independent each merited its own cluster.



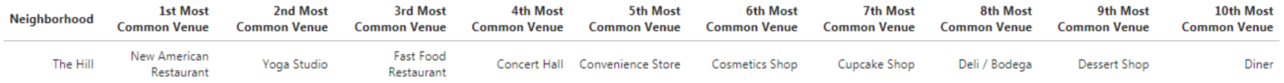
Performing a deeper evaluation, we examine the red cluster represented below. This is the largest cluster and contains represented by the red majority clusters the 7 of the 11 neighborhoods under evaluation. Examining the common venues, we can determine that the most common locales in these neighborhoods are food services such as restaurants, food trucks, sandwich places, and pizza shops. This cluster appears to be oriented primarily towards those who choose to regularly eat out or venture into common public places.



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
Downtown Boulder	New American Restaurant	American Restaurant	Bookstore	Taco Place	Clothing Store	Hotel	Dessert Shop	Ice Cream Shop	Pizza Place	Indian Restaurant
Chautauqua	Playground	Park	American Restaurant	Concert Hall	Roof Deck	Food & Drink Shop	Factory	Convenience Store	Cosmetics Shop	Cupcake Shop
Twenty Ninth Street	Mexican Restaurant	Furniture / Home Store	Coffee Shop	Clothing Store	Pizza Place	Grocery Store	Movie Theater	Mobile Phone Shop	Bank	Martial Arts Dojo
South Boulder	Sporting Goods Shop	Yoga Studio	Liquor Store	Record Shop	Pool	Pharmacy	Park	New American Restaurant	Coffee Shop	Diner
North Boulder	Pizza Place	Bakery	Ice Cream Shop	Electronics Store	Mexican Restaurant	Noodle House	Coffee Shop	Pharmacy	Grocery Store	Chinese Restaurant
Central Boulder	Food Truck	Baseball Field	Deli / Bodega	Factory	Brewery	Taco Place	Baseball Stadium	Trail	IT Services	Fast Food Restaurant
University of Colorado	Café	Sandwich Place	Coffee Shop	Thai Restaurant	Fast Food Restaurant	Taco Place	Chinese Restaurant	Outdoor Sculpture	Music Venue	Middle Eastern Restaurant

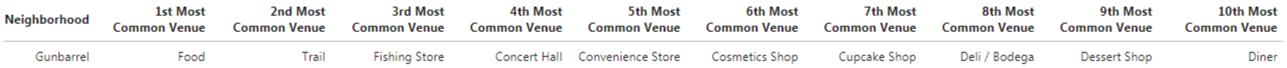
The remaining clusters each contain only one individual neighborhood, and each are presented below for further examination.

The green cluster containing The Hill area is represented as an outdoor, remotely located western neighborhood, that although has many offerings available, could possibly be too far removed from the central Boulder city or the university situated to the South-East.



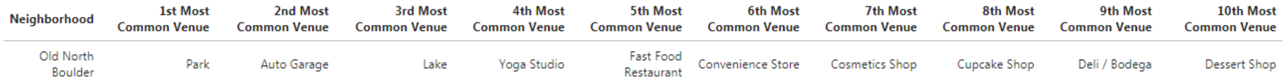
Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
The Hill	New American Restaurant	Yoga Studio	Fast Food Restaurant	Concert Hall	Convenience Store	Cosmetics Shop	Cupcake Shop	Deli / Bodega	Dessert Shop	Diner

The blue cluster containing Gunbarrel is also rural in definition, providing a host of outdoor activities and remote destinations. Considered one of the safer neighborhoods, again it's location may be too far to attract a steady stream of customers.



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
Gunbarrel	Food	Trail	Fishing Store	Concert Hall	Convenience Store	Cosmetics Shop	Cupcake Shop	Deli / Bodega	Dessert Shop	Diner

The orange cluster, representing the older portion of the city, provides a little of everything. Industry, recreation, convenience and shopping, Old North Boulder is a common congregating location that also appears to have relatively low competitors for coffee shop ventures. However, this particuar neighborhood did not achieve the most favorable crime statistics, and with industry as part of the demographics, can be a deterant for new business opportunities.



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
Old North Boulder	Park	Auto Garage	Lake	Yoga Studio	Fast Food Restaurant	Convenience Store	Cosmetics Shop	Cupcake Shop	Deli / Bodega	Dessert Shop

Discussion

Referring to the introduction section, the aim of this effort was to help determine the best neighborhoods in regard to both crime and competition for a new business venture. For both cities, we can say with a high level of confidence those neighborhoods to be Rose Hill within Charlottesville VA, and South Boulder within Boulder CO. Both have relatively low crime rates and moderate to minimal competitors allowing the potential to open a new business venture, with minimal criminal activity, minimum to moderate competitors, and most importantly, moderate to steady customer traffic based upon the neighboring, complimentary venues closely available. In addition, alternate proposed neighborhoods will be suggested during the presentation portion of the project and can be found in the summary deck.

Conclusion

This exercise as presented from the perspective of a soon to be retiree who wishes to reclaim and revisit one of their two college towns to give back to the community has certainly leaned upon the skills taught throughout this curriculum, but brought into a practical use case to gain familiarity through trial and error in order to achieve the goals set initially in the prior week. We can recount that in each of the steps applied using the common data science methodology, the goals have been achieved. First, determining the safest neighborhoods situated in both college cities, Boulder and Charlottesville based on current year crime data. Then through clustering methodologies for common features between neighborhoods, we were able to visualize and advise the most appropriate location in both cities to the client. However, due to the academic approach and limited timeframe for this micro-project, we only considered a few factors – particularly crime and frequency of other common business ventures in similar neighborhoods. This exercise did not account for other impacting data sets, such as seasonal student population, tax rates, which include personal and business, nor does it take into consideration population mean net income, consumer buying habits, or xxx. Future research would certainly help devise additional methodologies to better estimate such data points in order to thus provide more measurable and accurate results.

Data Appendix:

- Boulder CO Open Data Site: https://bouldercolorado.gov/open-data
- Boulder CO Neighborhood Data Site: https://www.bouldercoloradousa.com/about-boulder/boulder-neighborhoods/
- Charlottesville VA Open Data Site: https://opendata.charlottesville.org
- Charlottesville VA Neighborhood Data Site: https://opendata.charlottesville.org/datasets/f4efb475a1ca4b919fca4645b72fadd0_401/data
- Foursquare Venue Data Site: https://foursquare.com/