In which we leverage data, geocoding, K-Means clustering, and the Foursquare API to aid a mobile food service start-up in prioritizing communities.

Java Truck Pilot Project

Community Prioritization Analysis

Andy Clarke

1. Introduction/Business Problem

1.1: Background

2020 has been a strange and turbulent year, and for many businesses (especially the small ones) it has been a stressful and worrisome time. Some of the hardest hit were the shops and eateries of downtown cores the world over. Recent and ongoing developments surrounding the COVID-19 Pandemic have seen an exodus of employees and activity out of characteristically bustling business centres, and as workers retreated to the safety of their homes their employers were fast in transitioning towards remote work as a means of keeping their operations up and running. This, however, was and is not a solution for those brick-and-mortar shops in these business centres who relied on the routine traffic that had come to define their metropolitan locales, and with many businesses coming to embrace the ease, efficiency, and cost savings stemming from operating remotely, many of these downtown establishments have been made to face the reality that much of customer base may never return.

Great change, however, can be accompanied by great opportunity, and a young barista with business aspirations of their own thinks they might have an opportunity to seize upon their own small share of said opportunity. This individual, henceforth referred to as 'the client', having years of experience making and serving fine craft coffee beverages and products, knows well that you can take the customer from the coffee, but you can't take the coffee from the customer. The demand is still there, and with the shift occurring so quickly and unexpectedly, that demand is almost certainly untapped. Further, with so many businesses (especially small restaurants, eateries, cafes, etc.) being forced to shut their doors the client has noted that much of the physical capital they require is now available at bargain bin prices through mean such as public auctions. For many the office is now the home, but that doesn't mean they aren't still willing to spend $5 to $10 on a fancy morning pick-me-up, and the client plans to satisfy the demand by getting ahead of the competition by emulating the tried and true business model of the ice cream truck and adapting it to the provision of fine craft coffee.

1.2: Business Problem

Though their mobile operation is expected to be much cheaper to run than a conventional brick-and-mortar operation, the acquisition and leasing of all the necessary equipment, licenses, and permits (etc) will still be costly, and they intend to finance this operation partly by means of small business loans. As a result, the client is in the process of developing a comprehensive business plan to present to potential lenders/investors, and they have requested your services in the development of a specific piece of that business plan. You have been hired for the purpose of contributing to the market research portion of their business plan. More specifically, you have been tasked with analyzing the various communities around the city of Calgary, Alberta (where they intend to operate) with the goal being the production of intuitive visual aids to help in identifying high Vs low priority neighbourhoods/communities for the client's mobile java truck.

The Business Problem which we seek here to address is outlined explicitly below…

How can the various communities of the City of Calgary be profiled/classified for the purpose of identifying those communities which could be expected to provide the highest likelihood of positive returns for our client?

1.3: Interest

As outlined in the 'Background' section, potential investors and creditors are the intended audience. However, other entrepreneurial minds may wish to replicate the analysis for their respective niche should they too wish to capitalize on the sudden changes foisted upon our economy by COVID-19 and the associated commercial fallout in a similar manner

2. Data Acquisition and Cleaning

2.1: Data Sources and Assumptions

Our data is derived from several sources. First, our geospatial data was obtained from the City of Calgary at their website here. However, at the client's request, the analysis was to be restricted to a very specific subset of Calgary communities. This came after consultation with the client from which the following determinations and assumptions were made.

- ***Business parks and Industrial regions are already the targets of many mobile food service vendors. Though these vendors produce a dissimilar product and there could be room here for future expansion, it is assumed that the highest return is to be found in residential areas.***
- ***The analysis should be restricted to residential areas since these will be the areas to which demand has transitioned away from commercial parks and the downtown core. Uniform data is also difficult/impossible to obtain for non-residential areas (such as industrial sectors) thus preventing uniform analysis.***
- ***Downtown/City Centre is to be disqualified from analysis as this is the region from which workers and the economic/activity they bring with them are transitioning away from and towards their homes in the surrounding residential areas. Further, roadways are narrower, parking is in short supply, restrictions are tighter, and the market is highly saturated downtown and assumed to be the lowest return sector of the city.***
- ***Community sub-districts are assumed to include primarily condo and apartment complexes which makes their occupants more difficult to reach/access. Data is also unavailable for these areas.***

As a result of these assumptions and determinations the geojson file defining the geographic boundaries of those communities under investigation is edited to include only residential communities outside of Calgary's City Centre (downtown). This process is described in the data cleaning section of this document. The modified geojson file and containing the geospatial data actually used for mapping and for forming a template for our 'Community' feature can be found here.

Community level profile data is obtained from the City of Calgary's website and can be viewed and downloaded here. Median Household Income and Population are derived from each community profile. After consultation with the client the following assumptions based on their personal industry expertise were made.

- ***Coffee is universally enjoyed within all communities and regions of the city, and that the primary determinants for success when it comes to Craft Coffee providers are income (many coffee beverages are quite expensive), and population relative to existing services (market saturation).***

Therefore, only income and population figures are derived from the data available in the community profiles. Further, since this data is not made available by means of accessing individual PDF files for each community, the inclusion of other features from within the profiles would have increased the time and cost required to produce the analysis, so data from these profiles was limited to these 2 features. A CSV file containing this data can be found here.

Data for venues is obtained via Foursquare's places API. A search radius of 5,000m (5km) is used to make sure each potential competitor in each neighbourhood is identified. These data are used in conjunction with population data in order to determine a measure for market saturation. Consultation with the client resulted in the following assumptions being applied to the acquisition of venue data from the places API.

- ***A potential competitor is understood as stand-alone establishments which are not part of a food court.***
- ***As well, a potential competitor Is understood as those establishments for which coffee and coffee related services/products are the primary product/service provided.***

The client's experience leads them to believe that they won't be directly competing with every single food service establishment which sells coffee or some coffee products as part of their menu. As a result, the API call is filtered on the following 2 Foursquare Places categories.

- ***Café:*** '4bf58dd8d48988d16d941735'
- ***Coffee Shop:*** '4bf58dd8d48988d1e0931735'

2.2: Data Acquisition and Cleaning

Since the data was obtained from multiple sources with each being sourced for their own specific features there was an abundance of pre-processing and cleaning that needed to be completed prior to carrying out the analysis.

Figure 1: Geospatial Data Pre-Cleaning

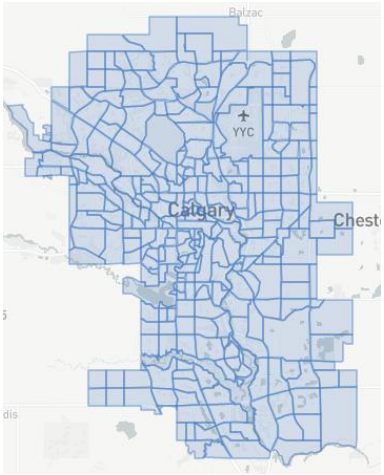
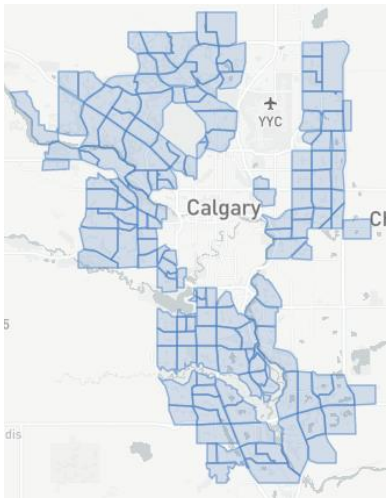


Figure 2: Geospatial Data Post-Cleaning



First the geospatial data contained in our geojson file needed to be cleaned of any entries pertaining to communities being excluded from the analysis. This meant sifting through the file and removing any keys and the corresponding data for which the 'Class' feature value was not listed as 'Residential', and as well it meant removing any keys and corresponding data for which the 'Sector' feature value was 'Centre' which denotes the City's downtown core. As stated above in the Business Problem section, the client's goal is to tap into demand that has since left the downtown core and since relocated to residential areas and suburbs where those working from home no longer have the sort of easy access that they one did to coffee shops and cafes. This actually resulted in the removal of exactly one half of the 'Community Boundaries' from the geojson file from 306 down to 153. Figures 1 Figure 2 above allow the reader to visualize the changes made. These changes were accomplished using the geojson file editor which can be found here. The labels contained within the geojson file acted as the template for the entire dataset since any mapping that included community boundaries would require that the relevant labels be perfectly matched. The strings representing our community labels needed to be perfectly harmonized with those in the geojson file, and as well all of the labels contained in the file had to be contained in the column of the dataset being keyed on. As a result, data drawn from other sources is modified to match the file in later steps.

After any communities labeled as 'Centre' or anything other than 'Residential' for their 'Sector' and 'Class' features the file was loaded into a Jupyter Notebook and processed further. The raw CSV geospatial data file which describes the initial dataframe can be found here. The table is too large to present the entirety of its features here so a list is provided instead. Initially the following features were included in the dataset. The only information directly meaningful to the analysis is contained in the feature describing the community name (see list below), however, some of the features are later used for the purpose of filtering during the cleaning and processing of venue data gathered from Foursquare in later steps.

- ***type***: This describes the entries type within the geojson file. As usual, each of our boundaries are included under the label 'Feature', thus this feature is removed as it provides nothing of value to the analysis.

- ***properties.comm_structure:*** Describes the era or building status of the community (whether or not it has finished building out to its boundaries). This feature is also removed as there was little of vale to be gleaned from it.
- ***properties.name***: This is the community/neighbourhood name. This feature is kept and used as a template for all subsets for the purposes of mapping. This feature was renamed 'Community'.
- ***properties.sector:*** Describes the geographical subsector sector of the city broken down into N,S, E, W, NW, NE, SW, and SE. This featured is not considered in the analysis, but is used in a subset of the geospatial data for the purposes of filtering venue data from Foursquare later.
- ***properties.class.code:*** A code corresponding to the class of the case. Since at this point each entry is residential, each code Is 1. This feature is removed as it is of no value for the analysis or as a filter for later cleaning/processing.
- ***properties.srg***: A variation of the comm_structure feature which references only the development state and excludes the era or years of development. This feature is also removed.
- ***properties.class***: A variation of the class.code feature. All are residential and the feature is removed but is also used in a subset of the full geospatial file later for the purposes of filtering venue data from Foursquare.
- ***properties.comm_code:*** An acronym for the properties.name feature. We did not want to work with acronyms and wanted all visual aids to be intuitive, and so this feature is removed and the community names are favoured.
- ***geometry.type:*** Describes the shape type of the geographical data. This is not relevant to the analysis and all mapping is handled by Folium by passing the geojson file, so this feature is removed.
- ***geometry.coordinates:*** Describes the coordinate set defining the boundaries of each community. This is all contained in the geojson file and is handled by Folium during the mapping process, so this feature is removed.

A small subset of the initial file is now summarized in a dataframe (CGY_COMM) which forms the basis for the overall dataset which is to be appended and completed as data is acquired from other sourced, cleaned/processed, and appended to our dataframe. From this completed dataframe, several subsets are also derived for different sections of the analysis. A sample of this dataframe is shown below in Figure 3.

Figure 3: Initial CGY_COMM Dataframe

	Community	Sector	Class
0	ABBEYDALE	NORTHEAST	Residential
1	ACADIA	SOUTH	Residential
2	ALBERT PARK/RADISSON HEIGHTS	EAST	Residential
3	APPLEWOOD PARK	EAST	Residential
4	ARBOUR LAKE	NORTHWEST	Residential

Population and Income data were drawn, as mentioned, from the City of Calgary's Community profiles. A hyperlink is provided above to this list of community profiles at the City of Calgary's website. It is reported here that their data was sourced from the Statistics Canada's 2016 Census data, however, I could not locate this data at Statistics Canada's Community Data Project website, so I instead opted to extract each community's income and population figures directly from their respective profiles. Profiles

were missing for a total of 13 of the 153 communities considered in the analysis. However, these communities remained in the dataset since they could still be used in the heatmapping portion and for one of the choropleths maps in this analysis. Both population and income data are of type float. A sample of the compiled CSV file is included below in Figure 3. The features in the data are described in the list below

- ***Population:*** A measure of the population for each community in the dataset. This feature is used as a measure of community potential on its own during the choropleth mapping section of this analysis, but is also used to determine a measure for market saturation to be used in both the choropleth and mapping sections as well.
- ***Median Inc:*** A mean measure of the median household income for each community. Median household income was the most current measure I could find compiled at the community level for the city of Calgary.

Table 1: Community Profile Data Sample

Community	Population	Median Inc
ABBEYDALE	6150	81232
ACADIA	10435	72552
ALBERT PARK/RADISSON HEIGHTS	6640	64429
APPLEWOOD PARK	6850	84965
ARBOUR LAKE	10760	109790
ASPEN WOODS	9060	199759
AUBURN BAY	14850	136961
BAYVIEW	740	260339
BEDDINGTON HEIGHTS	11840	88241

After manually compiling all of the Community Profiles data for income and population a CSV file is saved to be read into the Notebook where the data processing and analysis is executed.

Instead of extracting community coordinates from the geojson file, it was simpler to use the Nominatim API and geocoding to assign coordinates to each community. Once this was accomplished the coordinates were appended to the CGY_COMM dataframe as can be seen in Figure 4 below.

Figure 4: CGY_COMM Dataframe with Coordinates Appended

	Community	Sector	Class	Latitude	Longitude
0	ABBEYDALE	NORTHEAST	Residential	51.058836	-113.929413
1	ACADIA	SOUTH	Residential	50.968655	-114.055587
2	ALBERT PARK/RADISSON HEIGHTS	EAST	Residential	51.044845	-113.990195
3	APPLEWOOD PARK	EAST	Residential	51.044658	-113.928931
4	ARBOUR LAKE	NORTHWEST	Residential	51.136786	-114.202355

Community population and income data is now also appended to the CGY_COMM dataframe to produce the dataframe below in Figure 5.

Figure 5: CGY_COMM with Community Profile Data Appended

	Community	Sector	Class	Latitude	Longitude	Population	Median Inc
0	ABBEYDALE	NORTHEAST	Residential	51.058836	-113.929413	6150.0	81232.0
1	ACADIA	SOUTH	Residential	50.968655	-114.055587	10435.0	72552.0
2	ALBERT PARK/RADISSON HEIGHTS	EAST	Residential	51.044845	-113.990195	6640.0	64429.0
3	APPLEWOOD PARK	EAST	Residential	51.044658	-113.928931	6850.0	84965.0
4	ARBOUR LAKE	NORTHWEST	Residential	51.136786	-114.202355	10760.0	109790.0

The next step was to acquire and process venue data from the Foursquare Places API. The API call, as mentioned previously, applied a search radius of 5 km and filtered on the categories 'café' and 'coffee shop'. Each community's coordinates are passed as a parameter in the API call and data for all coffee shops and cafes identified in the call were returned. A subset of this data was extracted and transformed into the dataframe CGY_JAVA. The top of this dataframe is displayed in the Figure 6 below.

Figure 6: CGY_JAVA Dataframe from Foursquare API Call

	Community	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ABBEYDALE	51.058836	-113.929413	Tim Hortons	51.074283	-113.957754	Coffee Shop
1	ABBEYDALE	51.058836	-113.929413	Tim Hortons	51.038402	-113.931031	Coffee Shop
2	ABBEYDALE	51.058836	-113.929413	Tim Hortons	51.056854	-113.982534	Coffee Shop
3	ABBEYDALE	51.058836	-113.929413	Starbucks	51.038989	-113.908861	Coffee Shop
4	ABBEYDALE	51.058836	-113.929413	Starbucks	51.064950	-113.956489	Coffee Shop

The features in the above CGY_JAVA dataframe are described in the list below.

- ***Community (Object):*** Refers to the community corresponding to the coordinates passed in the API call. It is important to note that this does not necessarily represent the community to which the venue belongs since the 5km search radius of the call results in overlapping search radii as the loop iterates across each community in our dataset. As a result, this feature is removed from the dataframe.
- ***Community Latitude (Float)***: Describes the latitude passed in the API call which identified this case/venue for the given search radius and categories. It is dropped for the same reasons as the 'Community' feature above.
- ***Community Longitude (Float)***: Describes the longitude passed in the API call which identified this case/venue for the given search radius and categories. It is dropped for the same reasons as the 'Community' feature above.
- ***Venue (Object)***: This is the name of the venue. This feature was not used in analysis, but was held for reference purposes.
- ***Venue Latitude (Float)***: Latitude position for the given venue. This feature is kept in the dataframe for use in all sections of the analysis to follow.
- ***Venue Longitude (Float)***: Longitude position for the given venue. This feature is kept in the dataframe for use in all sections of the analysis to follow.

- ***Venue Category (Object)***: Describes the category to which the venue belongs according to Foursquare's categorization of venues. This feature is not used in the analysis portion but is held as a filter feature for cleaning/processing the data.

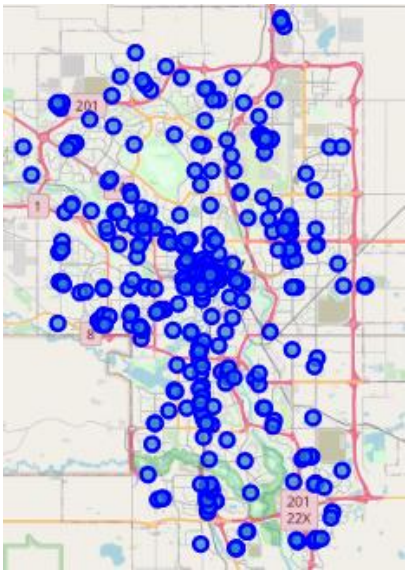
A total of 4,682 venue results were returned and translated into the initial CGY_JAVA dataframe. However, as was already explained, overlapping search radii resulting from the specified search radius are known to have resulted in the duplication of many of the returned venues. In addition, though the call filtered on the above specified categories, 138 of the venues returned possessed the Venue Category label 'Tea Room'. The duplicates were filtered from the dataframe using the 'drop_duplicates' method in Pandas by passing a subset parameter consisting of the 'Venue', 'Venue Latitude', and 'Venue Longitude' features to ensure that any distinct venues potentially sharing a space weren't incidentally dropped from the set. Those belonging to the 'Tea Room' category were simply removed by slicing the subset from the dataframe. After filtering out duplicates and unwanted categories a total 386 entries/venues remained, meaning that a total of 4,296 (or 91.8%) of the initial results were either duplicates or are assumed to not be potential competitors to the client.

Figure 7: CGY_JAVA Post Cleaning

	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Tim Hortons	51.074283	-113.957754	Coffee Shop
1	Tim Hortons	51.038402	-113.931031	Coffee Shop
2	Tim Hortons	51.056854	-113.982534	Coffee Shop
3	Starbucks	51.038989	-113.908861	Coffee Shop
4	Starbucks	51.064950	-113.956489	Coffee Shop
...	...	...	...	...

Further cleaning was also necessary due to the far-reaching nature of the specified search radius. Mapping the venue data revealed 5 venues which appeared to reside outside of Calgary. By running a simple loop passing the coordinates belonging to the indexes of these venues confirmed these suspicions. A map of the results and a list of these communities are provided below.

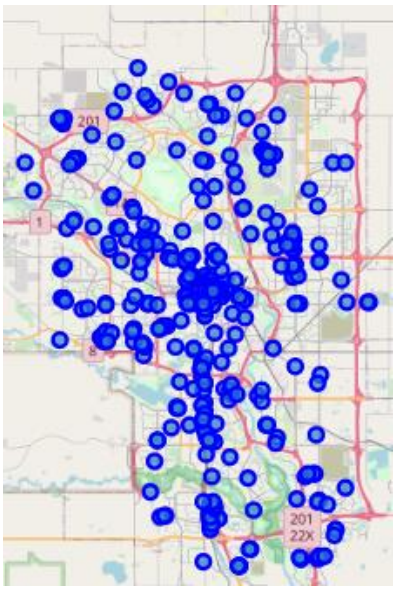
Figure 8: Mapped Venue Data Post-Filtering



- ***Shell, 293105, CrossIron Lane, Balzac, Rocky View County, Alberta, T4A 0V1, Canada***
- ***CrossIron Mills, Bass Pro Way, Balzac, Rocky View County, Alberta, T4A 0V1, Canada***
- ***CrossIron Mills, Bass Pro Way, Balzac, Rocky View County, Alberta, T4A 0V1, Canada***
- ***CrossIron Mills, Bass Pro Way, Balzac, Rocky View County, Alberta, T4A 0V1, Canada***
- ***CrossIron Mills, Bass Pro Way, Balzac, Rocky View County, Alberta, T4A 0V1, Canada***

These observations were then removed from the venues set (CGY_JAVA) as they are too far from the city to be considered worthwhile to the client. Au updated map is provided below.

Figure 9: Mapped Venue Data Post-Cleaning



Now that the venue data has been obtained accurate community labels need to be assigned to each of the venues. As explained in the description of the data the existing community labels produced by our API call loop pertained to the community to which the coordinates passed in the URL belonged and did not necessarily pertain to the community to which the actual venue coordinates belong. To obtain accurate community labels for each of the venues returned by the Foursquare API call(s) reverse geocoding was applied via a loop very similar to that which provided us with community coordinates in an earlier step. The results of this were then incorporated with the venues dataset. With accurate community label assignments included the venue set now contains all of the necessary data and it is deemed complete. A visual sample of the venue set is provided below in Figure 10. As may have already been assumed, the number of unique community results returned (130) is fewer than that of the number of communities passed in our API call and as well is fewer than that contained in our community profile set (CGY_COMM). This is simply a result of not every community containing a venue that matched our search criteria.

Figure 10: Venue Dataset Including Community Labels

	Community	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Albert Park/Radisson Heights	Tomoka Coffee	51.037764	-113.982446	Coffee Shop
1	Altadore	Monogram Coffee	51.010967	-114.100054	Coffee Shop
2	Applewood Park	Tim Hortons	51.038402	-113.931031	Coffee Shop
3	Applewood Park	Starbucks	51.038989	-113.908861	Coffee Shop
4	Applewood Park	Tim Hortons	51.038459	-113.910788	Coffee Shop

With accurate community labels assigned to each venue the venue data could be grouped according to the Community in order to provide a count of the total number of venues present in any given community. This subset is then to be merged with the community profiles set in order to complete it. However, before accomplishing this the community labels produced from the reverse geocoding process needed to be harmonized with those contained in our geojson data to make sure that consistency was maintained across all of our data sources so that any mapping could be completed without issue. As the reader may already have assumed or deduced from the maps above the venue data also contains venues belonging to communities in those regions of the city which were filtered from our initial data (those classed as non-residential or those within the centre sector). To remedy this a filter dataframe (FILTER_DF) was created from the unmodified geojson file and a function was devised to cross reference both the grouped and ungrouped venue data. From the ungrouped venue data (CGY_JAVA) a grouped subset was derived and both were referenced against the filter dataframe to detect any discrepancies. A screenshot of the code for the function is provided for the readers reference below, as is a capture of the grouped subset of the venue data which is to be appended to our community profile set following harmonization. The test argument is a list of community labels to be harmonized and the target argument is a list of community labels being referenced against. The returned dataframe (CROSS_DF) is a dataframe for which a 'False' bool value is returned after passing an community from the test list and checking for an equivalent string in the target list. This dataframe represents those community labels which are either not present or are mislabeled.

Figure 11: Cross Reference Function for Label Harmonization

```
def CrossRef(test,target):

    #cross_dict = dict.fromkeys(['Community','Bool'])
    cross_dict = {'Community':[],
                  'Bool':[]}

    for comm in test:

        cross_dict['Community'].append(comm)
        cross_dict['Bool'].append(comm in target)
        CROSS_DF = pd.DataFrame(cross_dict)
        CROSS_DF = CROSS_DF[CROSS_DF['Bool'] == False]

    return (CROSS_DF)
```

Figure 12: Grouped Venue Set (CGY_JAVA_GROUPED) Subset for Cross Referencing

	Community	Venue Count
0	ALBERT PARK/RADISSON HEIGHTS	1
1	ALTADORE	1
2	APPLEWOOD PARK	3
3	ARBOUR LAKE	5
4	ASPEN WOODS	3

For both the grouped and ungrouped venue data an obvious discrepancy can be noted from visual inspection of the community labels contained in the filter data and those against which it is being referenced. The geojson data (which we are harmonizing to) contains capitalized labels whilst the labels return by the reverse geocoding process are a mix of upper- and lower-case characters. This is adjusted for prior to generating lists for each and passing through the cross reference function.

The grouped venue data is passed though the function first and the following dataframe is returned signalling that two labels either aren't present in the filter or data or are mislabeled. A search of string characters in the filter data confirmed that the issue was one of inconsistency in community labels and the labels were adjusted. Afterwards the function was run again to ensure that all labels in our venue set were consistent with those in the geojson file. The dataframe returned by the function is shown below for the readers reference in Figure 13.

Figure 13: Grouped Cross Reference Return DF

	Community	Bool
81	PARKHILL/STANLEY PARK	False
100	SCARBORO/SUNALTA WEST	False

Before appending the grouped data to our community profile set unwanted community classes as well as any communities belonging to the 'Centre' sector of Calgary needed to be filtered from the data. To accomplish this the now adjusted grouped venue subset was merged with the filter set to produce the dataframe described by the snippet below in Figure 14.

Figure 14: Grouped Venue Filter Dataframe

	Community	Venue Count	Sector	Class
0	ALBERT PARK/RADISSON HEIGHTS	1	EAST	Residential
1	ALTADORE	1	CENTRE	Residential
2	APPLEWOOD PARK	3	EAST	Residential
3	ARBOUR LAKE	5	NORTHWEST	Residential
4	ASPEN WOODS	3	WEST	Residential
5	AUBURN BAY	6	SOUTHEAST	Residential
6	AURORA BUSINESS PARK	1	NORTH	Industrial
7	BANFF TRAIL	1	CENTRE	Residential
8	BEDDINGTON HEIGHTS	1	NORTH	Residential
9	BELTLINE	6	CENTRE	Residential

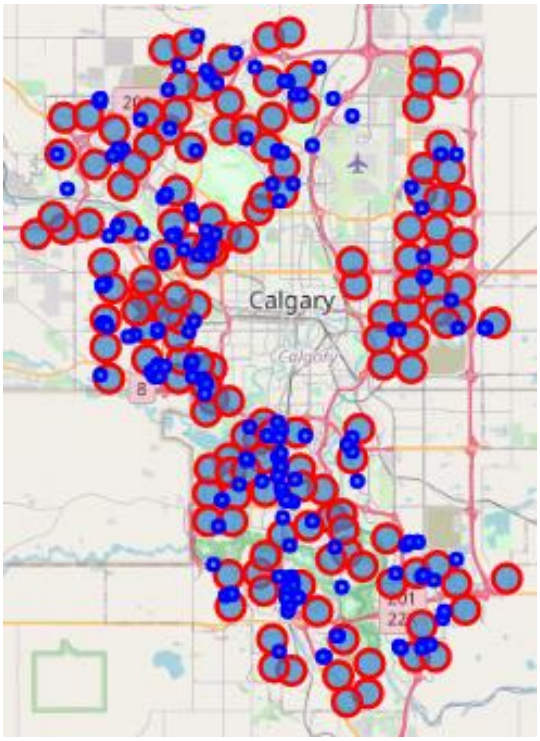
Using the ‘Class’ and ‘Sector’ labels present in the grouped venue data filter dataframe shown above a fully harmonized list of communities and their respective venue counts was produced. After doing so a total of 75 communities remained in the set, implying the same number of unique community labels would exist for the ungrouped venue set which would serve as its own standalone dataset after undergoing the same harmonizing and filtering process. Having explained the process harmonization/filtering process in detail for the grouped venue set (CGY_JAVA_GROUPED) I won’t repeat it again for the ungrouped set since the same process was repeated and the same results obtained (which could be assumed since the grouped set is merely a subset of the ungrouped set).

With the grouped venues set fully harmonized with our community profile and geospatial data the community profile set (CGY_COMM) was completed by merging the two to produce the dataframe represented by the Figure 15 below. This dataframe is referred to as CGY_COMM_FULL.

Figure 15: CGY_COMM_FULL Community Profile Dataset

	Community	Sector	Class	Latitude	Longitude	Population	Median Inc	Venue Count
0	ABBEYDALE	NORTHEAST	Residential	51.058836	-113.929413	6150.0	81232.0	NaN
1	ACADIA	SOUTH	Residential	50.968655	-114.055587	10435.0	72552.0	NaN
2	ALBERT PARK/RADISSON HEIGHTS	EAST	Residential	51.044845	-113.990195	6640.0	64429.0	1.0
3	APPLEWOOD PARK	EAST	Residential	51.044658	-113.928931	6850.0	84965.0	3.0
4	ARBOUR LAKE	NORTHWEST	Residential	51.136786	-114.202355	10760.0	109790.0	5.0

Figure 16: Map of Complete Community and Venue Datasets



A visual summary of the completed base datasets (CGY_COMM_FULL and CGY_JAVA) is provided in Figure 16 above. by mapping the communities as red markers and all of the venues as blue markers on a map of the city of Calgary.

3: Methodology/Analysis

Analysis of the now cleaned and processed data is handled in 3 separate stages. Through discussion with the client it was determined that intuitive visual aids in the form of maps would be the ideal product of the analysis for the following reasons.

- ***This prioritizes the preferences of the audience. Visualized data is easy to consume and interpret and would thus be the ideal/referred means of presenting the results to potential investors and creditors.***
- ***Maps could also constitute valuable tools during the initial stages of operation as different communities and routes are experimented and actual data is gathered for consideration at a later date.***

The three components of the Analysis are as follows…

1. ***Heat Mapping:*** A simple heat map is derived from the ungrouped venues data (CGY_JAVA) which provides a visual representation of competitor 'density' throughout the various communities of Calgary. By avoiding these dense pockets of competing establishments, the client might find that they develop a consistent customer base of individuals who are willing to await the arrival of their java truck over travelling to a brick=and-mortar café/coffee shop.
2. ***Choropleth Mapping:*** The choropleth maps are utilizing as a means of visually summarizing the various features of the community profiles datasets by mapping their magnitudes along a continuous color gradient for each community. This is why it was necessary to ensure consistency in community labels across each dataset, as generating choropleth maps in Folium requires that all labels contained in the geojson data are both present and consistently defined in the mapping data. Four separate Choropleth maps are produced, with each describing a different metric of interest to the client. From the client's experience, the most relevant factors in determining the success or failure of their operation are
    a. *Community Affluence: Which is approximated using the 'Median Household Income' feature. This assumption is supported by the article here.*
    b. *Population Density: An abundance of potential customers to draw from is essential. Though this is a nominal metric it is still potentially valuable to have an idea of how many people reside within a given community.*
    c. *Market Saturation: This choropleth map creates a new metric from the data. A ratio is calculated using community population as the numerator and the venue count as the denominator to produce a measure for Market Saturation. This is assumed to be more informative and powerful than nominal standalone measures such as Venue Count and/or Population. It is understood that the more saturated the market/community the harder it will be to establish a customer base in that community*.
    d. *Community Potential Index: This choropleth is representative of a unique metric derived from each of the totality of the relevant features in the community profiles set. It is a weighted index of the Median Household Income and Market Saturation figures after*

normalizing both items. The client, informed by their expertise, felt that community affluence seems to contribute more to a craft coffee establishments' success, and so a %60-%40 split favoring income was settled on.

3. ***Community Segmentation/Clustering:*** *The descriptive tools available to the customer are to be furthered by a clustering analysis using the K-Means algorithm. The client felt that, for both operational use and for presentation to potential investors/creditors, the assignment of labels to each community as a function of the relevant feature set would be a valuable addition to the business plan. The optimal k-value (number of clusters) is determined using the Elbow Method, a K-Means model is fit and cluster labels assigned to each community, and then the clusters are provided with descriptive labels of their own by means of visualizing the model assignments on a scatter plot. The communities are then plotted to a map of the city and color coded for ease of reference.*

3.1 Heat Mapping

The heat map is fairly straight forward. No data subsets needed to be generated. Venue data is passed as a parameter through Folium's heat map plugin and markers for each community are provided as well. No changes were made to the data, as null values are simply interpreted as zeros. The client can utilize the map to determine communities and areas of communities that are potentially underserviced with regards to coffee shops/cafes. Even a saturated community could possess areas/pockets without a nearby vendor and this could represent an opportunity for the client to appeal to the potential customer's preference for service and convenience when choosing who patronize.

A portion of the map is displayed below in the *Results* section of the report, but it is recommended that the reader view the full map within the Jupyter Notebook used to execute this analysis so that they can view the entirety of the data represented as a heat map. If the full map is shown much of the information is lost, and the information is best viewed with full control over the zoom level and position of the map.

3.2: Choropleth Mapping

As detailed above the choropleth mapping portion of the analysis is accomplished in four separate parts. As a result, different subsets of data needed to be produced for the different metrics represented in some of the choropleth maps. Each map and its constituent data are described below. Note that for the *Market Saturation* and *Community Potential Index* choropleth maps some null values needed to be replaced.

3.2.1: Population Choropleth

Not unlike the heat map, no data subset was generated for the population choropleth map. Null values remain null and those communities lacking data are blacked out on the map. While the heat map informs the customer of where dense pockets of competitors are, it fails to say anything about the community's demographics. A region that appears underserviced on the heat map might not contain enough potential customers or the sort of affluent demographic being targeted (which is to be represented by the choropleth map representing income labels). The *Yellow* → *Orange* → *Red* color gradient is used for this map.

3.2.2: Median Household Income Choropleth

Similar to the choropleth mapping of nominal community population, this choropleth map details nominal median household income levels for each community, thus filling a hole in the demographic mapping that is lacking in both the heat map and the choropleth describing community population levels. There is no need to create a data subset as the nominal float type values contained in the 'Median Inc' feature is all that is being mapped. The *Purple → Blue → Green* color gradient is for this map.

3.2.3: Market Saturation Choropleth

This map requires the creation of a data subset which includes the creation of a 'Saturation' feature which represents population per venue for each community. An issue arises in that no incumbent competing establishments were identified for over 70 of the communities in the dataset, and as a result any attempt at calculating a ration using the 'Venue Count' feature results in a null value. In addition, there are 13 communities lacking population data. In order to preserve derive as much information as possible it was important to distinguish between null 'Saturation' values occurring as a result of a null 'Venue Count' and those occurring as a result of a null 'Population' value. A null 'Population' value represents a lack of data, while a null 'Venue Count' value simply suggests that there are no existing competing establishments within the community which can be translated to a value of zero. A zero will likewise return a null value when calculating a value for saturation though. A community lacking any competing establishments, in this case, takes on its own unique form by being the only the case in which a null value can be returned. This is accomplished by filling all null values for the 'Population' feature with the population mean. In doing so an approximation of saturation is determined for those values lacking data, and communities without existing venues are color coded magenta so that they clearly stand out against the *Purple → Blue* color gradient describing communities containing existing establishments.

This does unfortunately mean that the map and the values it is based on may be less reliable for those communities for which the population was approximated using the mean. For this reason the dataframe containing all communities for which data was not available is provided below for the readers and the client's reference. At a later date it may be worthwhile to try to find appropriate data to populate these fields with so that the results are more meaningful for these specific communities. A sample of the data subset containing the new 'Saturation' feature used to generate the map is provided as well.

Figure 17: Market Saturation Choropleth Subset

	Community	Population	Venue Count	Saturation
0	ABBEYDALE	6150.0	NaN	NaN
1	ACADIA	10435.0	NaN	NaN
2	ALBERT PARK/RADISSON HEIGHTS	6640.0	1.0	6640.000000
3	APPLEWOOD PARK	6850.0	3.0	2283.333333
4	ARBOUR LAKE	10760.0	5.0	2152.000000

Figure 18: Communities Lacking Income and Population Data

	Community
9	BELMONT
10	BELVEDERE
17	CARRINGTON
30	CORNERSTONE
61	HASKAYNE
65	HOMESTEAD
66	HOTCHKISS
75	LIVINGSTON
101	PINE CREEK
118	SETON
139	UNIVERSITY OF CALGARY
149	WOLF WILLOW
152	YORKVILLE

3.2.4: Community Market Potential Index

This choropleth map is derived from a measure referred to here at the 'Community Potential Index' which was calculated as a means of obtaining an overall picture of the market potential contained within a community based on our limited feature set. A subset of the community profile data (CGY_COMM_FULL) is generated here which also includes a 'Saturation' feature, however, for the purposes of calculating the index any null or zero values for the 'Venue Count' feature are replaced with 1. A null (or zero) value resulting from a lack of establishments in a given community was much more meaningful for the simple market saturation measure and the resulting choropleth than it would be for the 'Potential' feature being generated here which is the feature containing the community potential index score for each community. A null or zero value for 'Venue Count' for the community potential index score produces either a null or zero result which, unlike in the prior choropleth, provides little to no valuable information about the overall profile for the given community. After discussion with the client the following assumption was considered appropriate to apply in this context.

- ***Though a community might not possess a competing venue within its own boundaries it is reasonable to expect, given the varying size and shape of Calgary's communities, that a coffee shop or café is still relatively accessible and possibly closer than for some households in communities which contain such an establishment.***

The client was agreed that a null or zero 'Value Count' value could be potentially misleading in it's own right (due to the varying size and shape of communities) and that whatever accuracy is lost in prescribing 1 to null values for 'Venue Count' is worth the potential return expected from the community market potential index score.

After filling null values for the 'Venue Count' feature with 1 a saturation feature is generated for the subset. This time, however, null values for both 'Population' and 'Median Inc' are allowed to remain since we only miss out on calculating an index score of 13 communities as a result of missing data.

Further, since both population and income are considered in the calculation of the index score, it is understood that replacing null values for both features with some approximation of their real value (such as the feature mean) introduces substantial uncertainty regarding the validity of the results, and so it was determined best to allow these cases to remain null and to appear as such in the choropleth map to follow.

Once the 'Saturation' feature is determined for the subset, the two features to be multiplied in the calculation of the index score are normalized using Z-Score normalization. The 'Potential' feature, which represents the market potential index score for each community, is then calculated by weighting the normalized income and saturation measures according to the %60-%40 split (which the client believed represented a meaningful approximation of the relative importance of the two variables in the determination of the expectation of customer traffic for a competitive establishment) and then determining the product of the weighted variables. A sample of the subset including the 'Potential' feature is included below.

Figure 19: Community Market Potential Index Score Subset

	Community	Population	Median Inc	Venue Count	Saturation	Median Inc Norm	Saturation Norm	Potential
0	ABBEYDALE	6150.0	81232.0	1.0	6150.000000	-0.715152	0.046505	-0.410489
1	ACADIA	10435.0	72552.0	1.0	10435.000000	-0.949323	0.924074	-0.199964
2	ALBERT PARK/RADISSON HEIGHTS	6640.0	64429.0	1.0	6640.000000	-1.168466	0.146857	-0.642337
3	APPLEWOOD PARK	6850.0	84965.0	3.0	2283.333333	-0.614443	-0.745388	-0.666821
4	ARBOUR LAKE	10760.0	109790.0	5.0	2152.000000	0.055289	-0.772285	-0.275741

3.3: Community Clustering/Segmentation

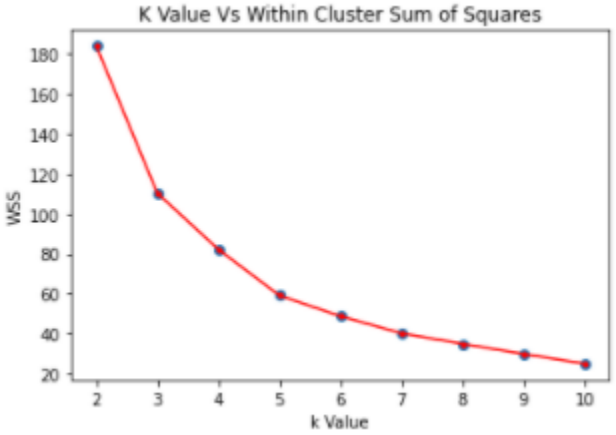
Clustering of the communities for in our dataset is accomplished by means of the K-Means algorithm. Since the presence of more than 2 variables being considered in the algorithm makes visualizing and interpreting the results much more difficult a subset of the data is produced which is limited to only measures of income and market saturation. This is convenient since it still incorporates all of our relevant features while avoiding the application of techniques such as Principle Component Analysis (PCA). In order to avoid having to drop over 70 communities from consideration the same assumption that justified the replacement of null 'Venue Count' values in the construction of the market potential index score is applied here and null values for 'Venue Count' are filled with 1. This limits the loss of communities considered to only those lacking income and population data. Null values are dropped from the dataset and the subset upon which the K-Means model is to be fit is completed. A portion of that subset is depicted below in Figure 20.

Figure 20: K-Means Clustering Subset

	Median Inc	Saturation
0	-0.715152	0.046505
1	-0.949323	0.924074
2	-1.168466	0.146857
3	-0.614443	-0.745388
4	0.055289	-0.772285

With the data subset determined the next step was to determine the optimal K-value for the model. This was accomplished using the *Elbow Method*. A loop was devised which iterated over a range of K-values, fit a model for each, extracted the Within-Cluster Sum of Squared Errors (WSS), and plotted these to a simple scatter plot which is depicted below in Figure 21.

Figure 21: WSS Vs K-Value for Application of the Elbow Method



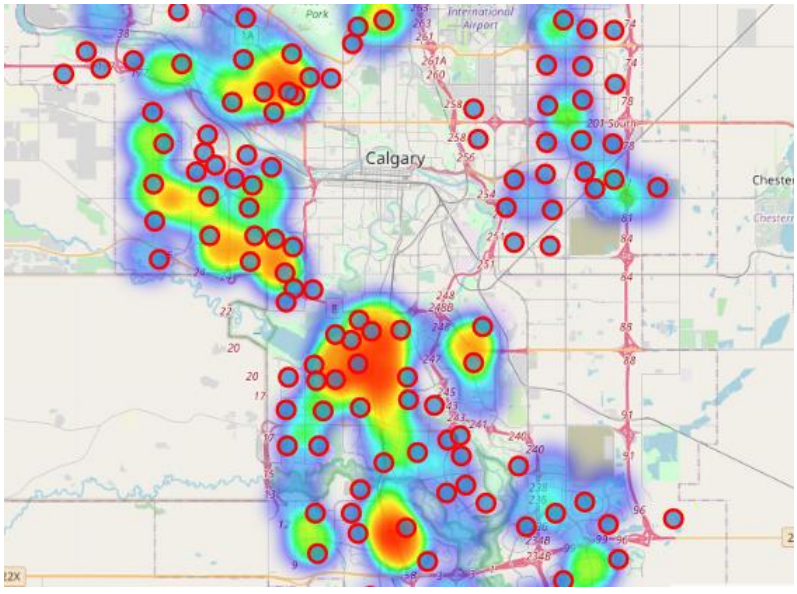
The elbow of the curve is determined to correspond with a K-value of 5, and thus the optimal K-value for the analysis determined to be 5. With this determined the model was ready for fitting and cluster labels can be assigned to each of the communities under consideration.

4: Results

The results section is handled in the same order and in the same format as the Methodology/Analysis section. Discussion of the results is handled in its own section.

4.1: Heat Mapping

Figure 22: Heat Map Sample



The results of heat mapping the 'Venue Count' feature for the purposes of identifying dense pockets of competing establishments are presented above in Figure 22. As described previously in the Methodology section it is recommended that the mapping results be viewed in the Jupyter Notebook for this project as control over the zoom function is important for viewing the mapping results, and this is especially true for heat maps.

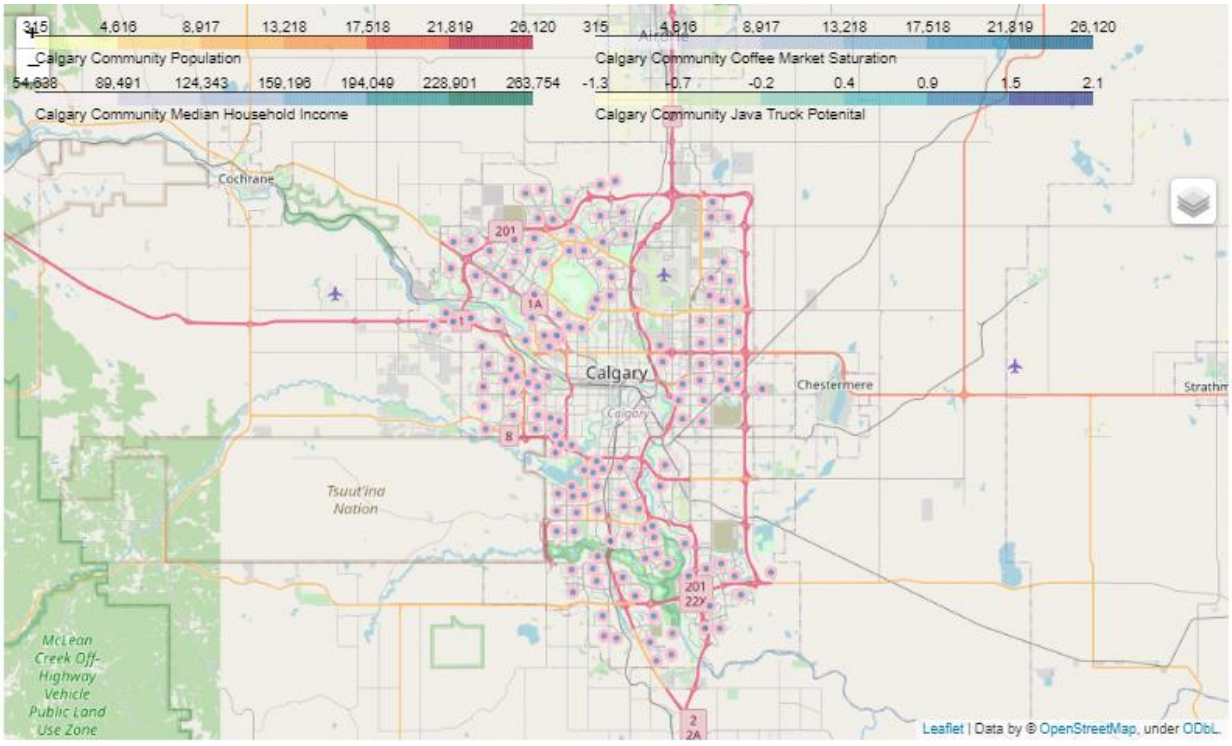
4.2: Choropleth Mapping

As was the case for the heat map, the choropleth maps are also best viewed within the notebook as the user can control the zoom level to better distinguish between community boundaries, and can as well adjust the layer control to switch between the choropleth map being viewed. Each were added as layers in the same map object, which is why all four legends are visible for each of the Figures below.

4.2.0: Choropleth Map Community Markers

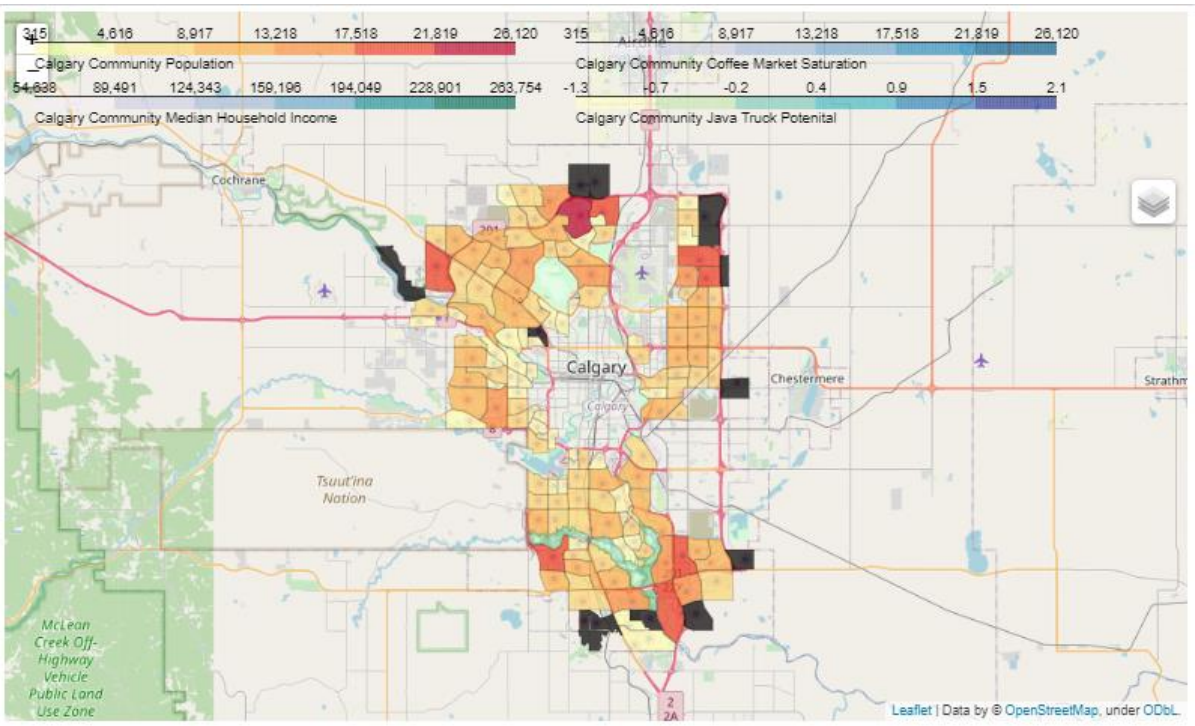
By using the layer controls and turning off each of the Choropleth map layers community markers can be accessed to allow the user to identify which community belongs to each boundary contained in the map. Without any of the choropleth layers active the map appears as in Figure 23 below. Access to each choropleth layer is controlled by hovering over the object in the right upper corner beneath the legends and toggling specific map layers on or off. I am inexperienced with Folium and suspect there is a way to highlight specific communities and their labels by simply hovering over them without the need to turn off the choropleth layers to access markers, but due to time constraints this format was used for now.

Figure 23: Choropleth Map Base



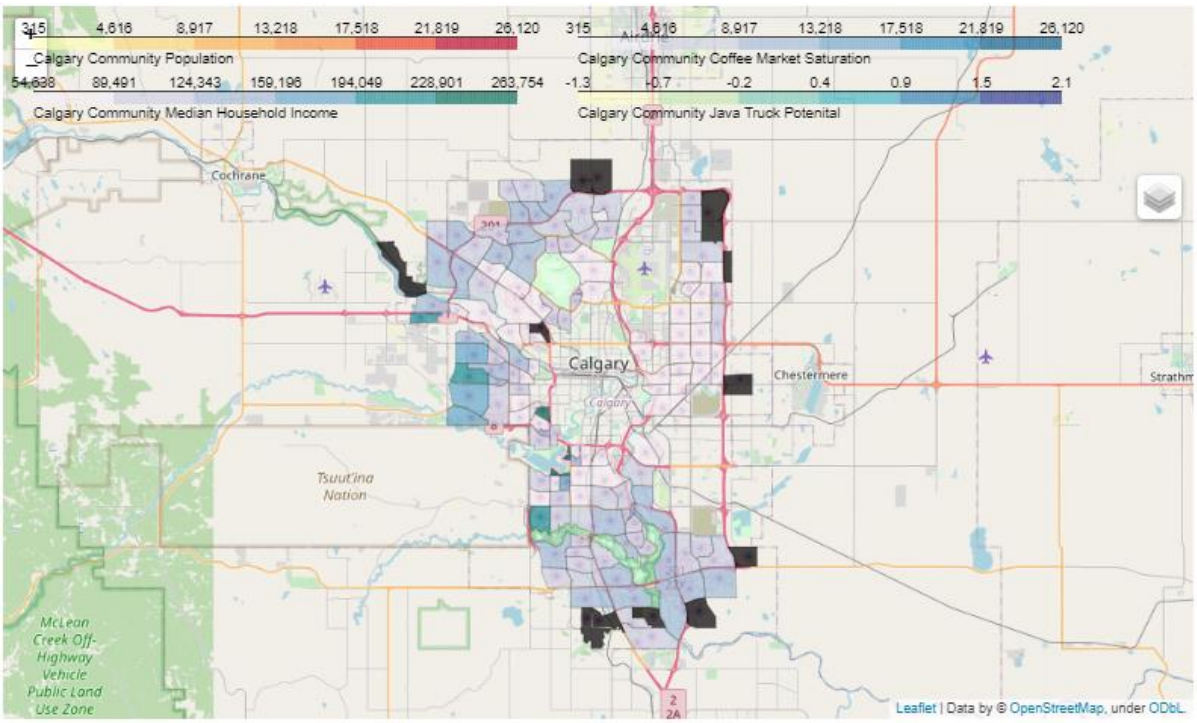
4.2.1: Population Choropleth

Figure 24: Population Choropleth Map



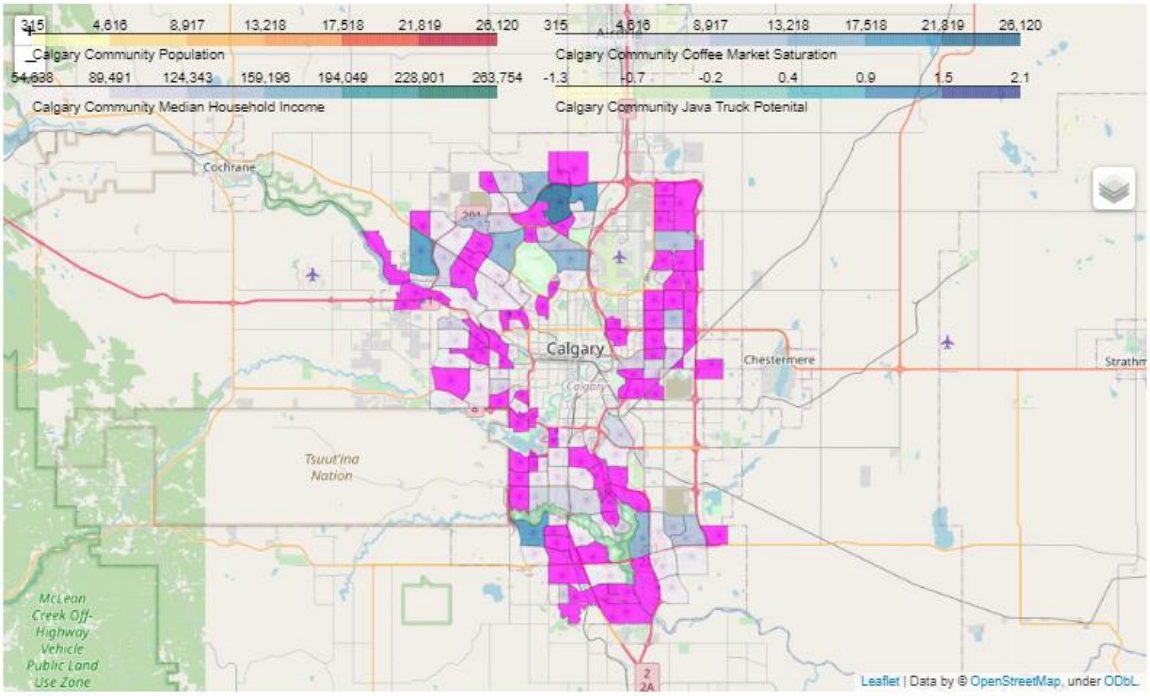
4.2.2: Median Household Income Choropleth

Figure 25: Median Household Income Choropleth Map



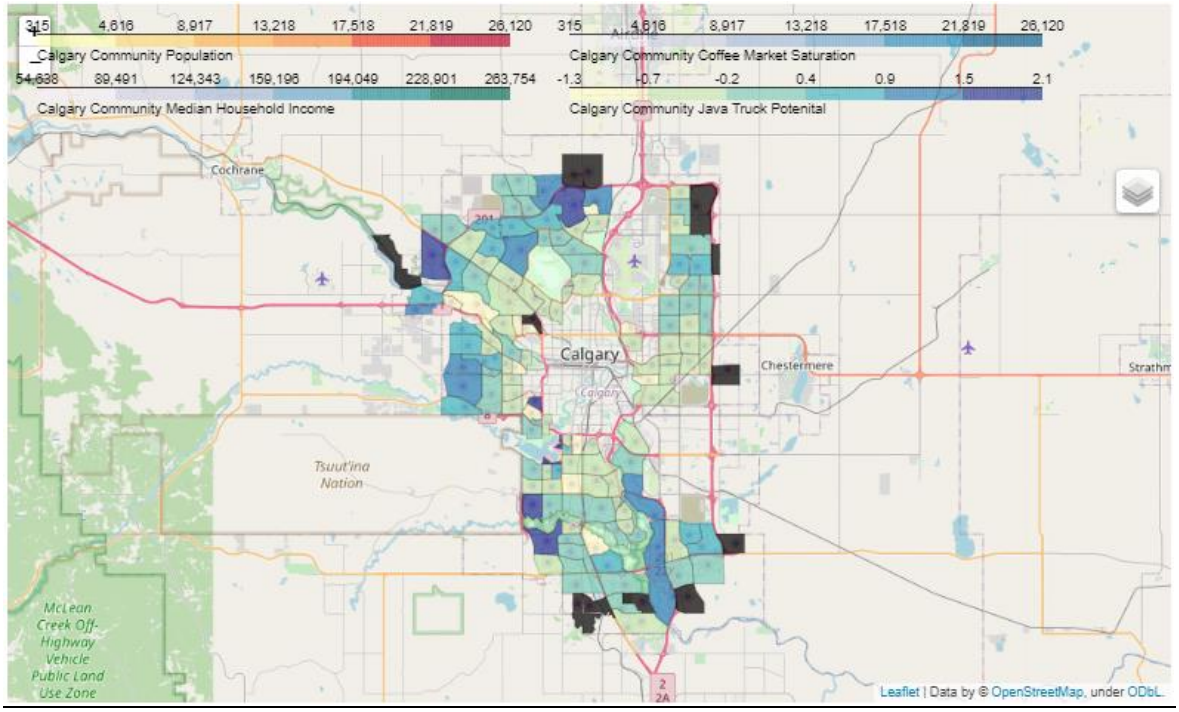
4.2.3: Market Saturation Choropleth

Figure 26: Market Saturation Choropleth Map



4.2.4: Community Market Potential Index Choropleth

Figure 27: Community Market Potential Index Score Choropleth

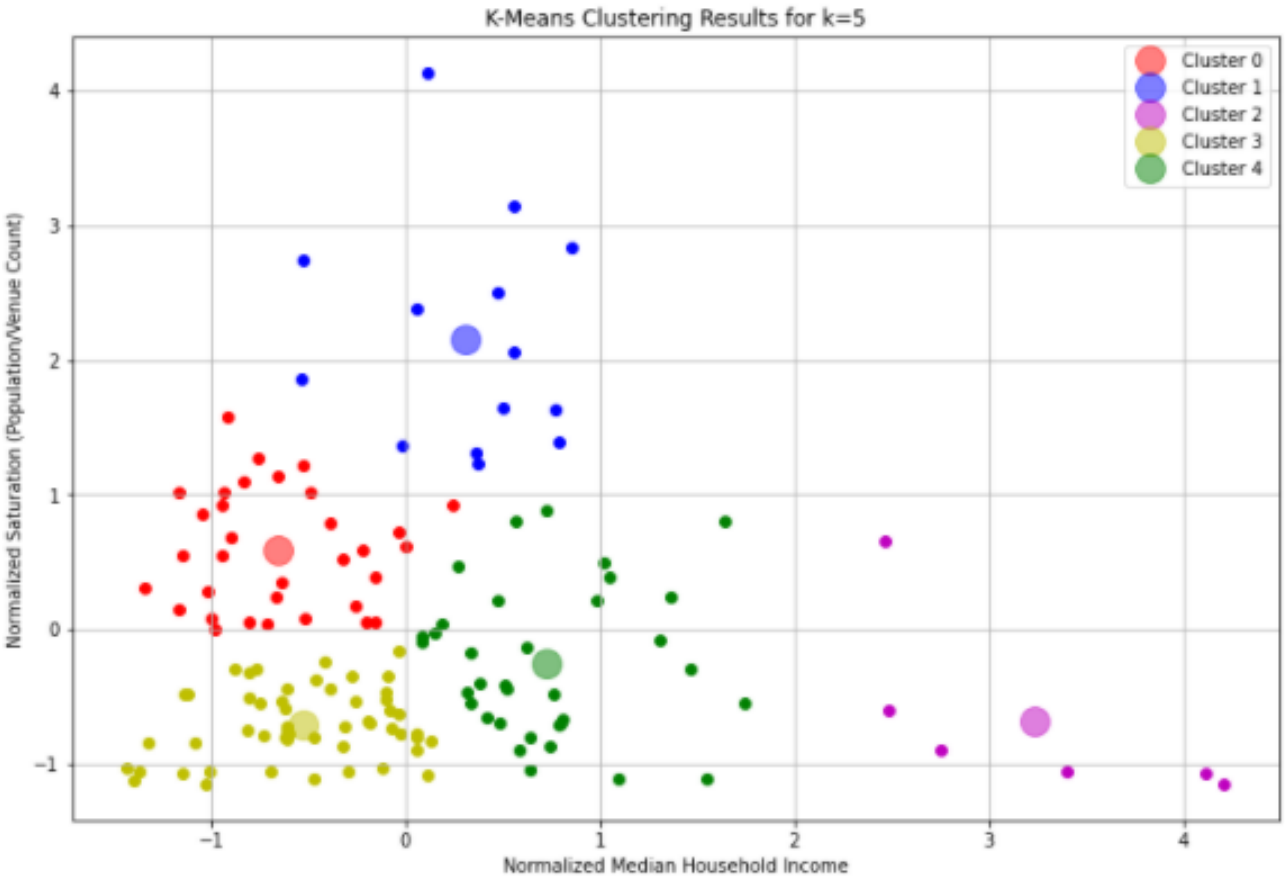


4.3: Community Clustering/Segmentation

There are 3 components to the results of the clustering portion of the analysis. The first is a scatterplot visualization of the results of the K-means model given a K-value of 5. This will be used later to prescribe descriptive titles to each of the communities according to their cluster assignment. The second is an updated dataframe including a new feature 'Cluster Labels' which denotes the cluster label assigned to each community by the model. The third component visualizes the results of the clustering model by generating a map which color codes markers for each community according to their cluster assignment. A description of the color's significance and that of the cluster assignment as it pertains to that cluster's income and market saturation levels is included in the label for each marker.

Figure 28 below visualizes the clustering results of the K-Means model by assigning each cluster a color and then plotting and color coding it's constituent observations and cluster centroids appropriately. In the *Discussion* section to follow these colors will then be used to identify each community's cluster assignment and describe it's relevance on a map of the city.

Figure 28: Scatterplot Visualization of the K-Means Clustering Results for K=5



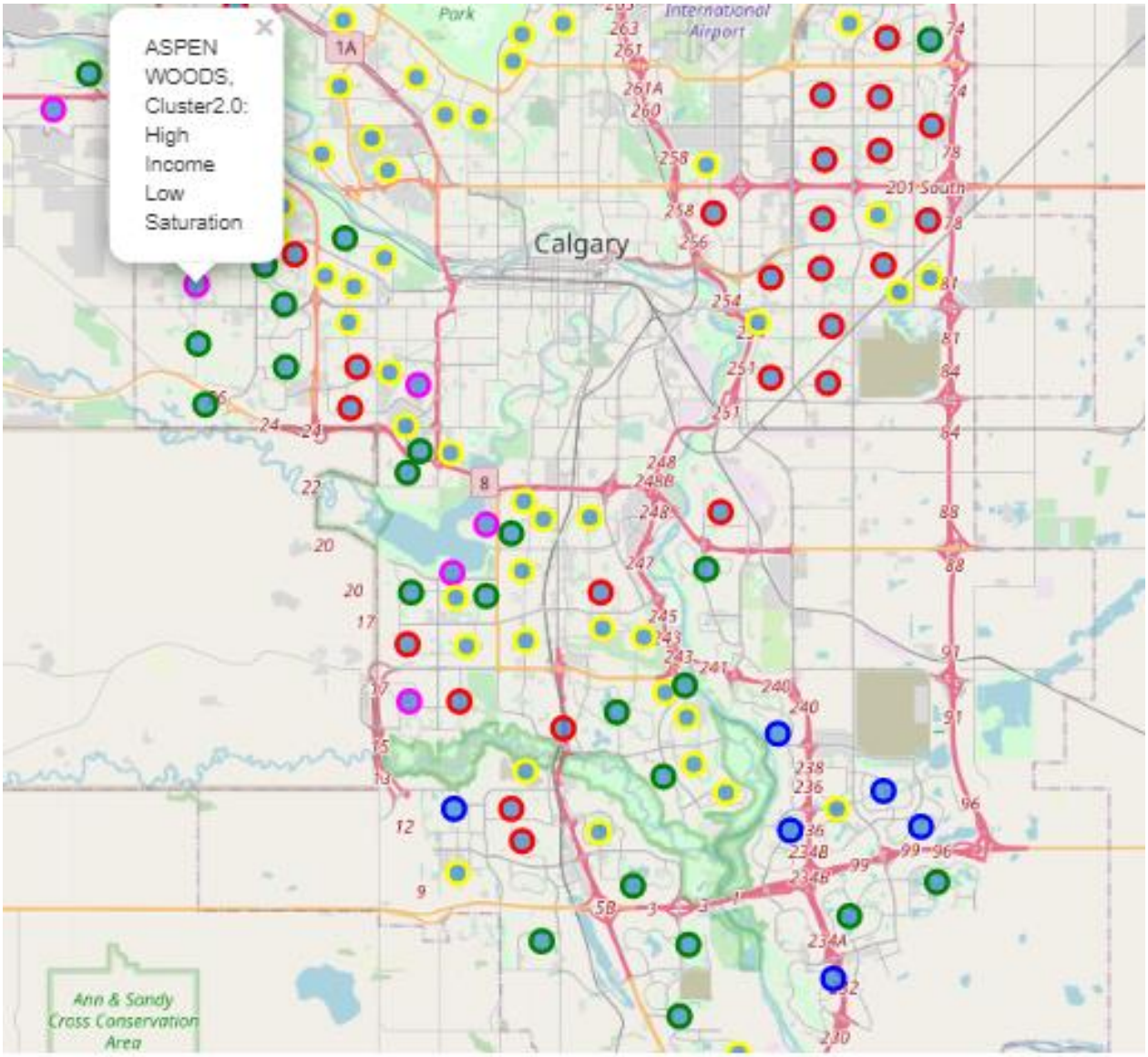
During the creation of the data subset for this portion of the analysis the index was not reset (see above) thus a community label is easily attributed to the appropriate cluster label by appending the list of cluster assignments output by the fitting of the K-Means model to the community profiles dataset CGY_COMM_FULL). This produced a new dataframe titled CGY_COMM_CLUSTERED. The top of this dataframe is provided below for the reader's reference in Figure 29.

Figure 29: CGY_COMM CLUSTERED Dataframe

	Community	Sector	Class	Latitude	Longitude	Population	Median Inc	Venue Count	Cluster Labels
0	ABBEYDALE	NORTHEAST	Residential	51.058836	-113.929413	6150.0	81232.0	NaN	0.0
1	ACADIA	SOUTH	Residential	50.968655	-114.055587	10435.0	72552.0	NaN	0.0
2	ALBERT PARK/RADISSON HEIGHTS	EAST	Residential	51.044845	-113.990195	6640.0	64429.0	1.0	0.0
3	APPLEWOOD PARK	EAST	Residential	51.044658	-113.928931	6850.0	84965.0	3.0	3.0
4	ARBOUR LAKE	NORTHWEST	Residential	51.136786	-114.202355	10760.0	109790.0	5.0	3.0
...	...	...	...	...	...	...	...	...	...

Finally, the new dataframe is used to generate a map of the city identifying each community and color coding the markers according to their respective cluster assignments. Information pertaining to the characteristics of each cluster is contained in each community label as previously explained.

Figure 30: Calgary Community Map with Cluster Assignments and Color Coding



Only a portion of the map is shown here, as was the case for the heat map as well, since the zoom level required to properly capture the detail crops out some of the surrounding areas. It is, again, recommended that the reader view the maps in the project's Jupyter notebook.

5: Discussion

In this section the results above are discussed. This section follows the same format and order as the previous two sections. However, since such analysis is an iterative process, an additional section is included to discuss the weaknesses of the analysis and considerations for future/future analysis. Note that any blacked out community on the maps denotes null value/measures.

5.1: Heat Mapping

The heat map geographically depicts dense pockets of established competition throughout the city of Calgary. Though the heat mapped "Venue Count' data is somewhat useful in it's own right for avoiding saturated or already well-serviced areas, the inclusion of community markers and the inclusion of their respective populations in their marker labels adds an additional dimension and allows the user to identify areas that are not only potentially underserviced in terms of the craft coffee market, but as well it allows them to identify underserviced communities which are also densely populated.

The heat map shows that, in general, as one moves inward from the periphery of the city and into the older and more established communities the heat gradient intensifies suggesting a higher density/prevalence of existing competition. This provides a useful tool for the client in two ways.

1. It substantiates the validity of their business model in a way that is intuitive and easy to understand for potential creditors. It can be shown that, with thousands of employees relocating their work activities from office to home office, there conceivably exists an abundance of untapped consumer demand in the form of people looking to maintain that part of their weekday routine involving their morning (and after) coffee/coffee beverage (among other products). Their mobile service can reach out to several of these pockets of untapped demand and establish a customer base (and possibly expand) before conventional brick-and-mortar competitors can gain a foothold.
2. It serves as a potentially valuable tool for day to day operations during the early stages of the Startup by giving the client a simple and intuitive means of identifying higher potential/underserviced regions to prioritize.

The primary weakness of this map is it's having mapped a nominal measure that says little about the overall profile of the area and the constituent elements (denoted by community markers) being mapped. However, this is offset by the inclusion of additional maps containing data derived measures such as market saturation and the community potential index.

5.2: Choropleth Mapping

5.2.1: Population Choropleth

As is also true for the *Median Household Income Choropleth* to be described later, this gradient mapping of nominal population is perhaps the least productive or useful of all of the components of the analysis. Both the population and income choropleth maps suffer from the same weakness as the heat

map in that they only map a nominal measure. However, used in conjunction with the heat map or the income choropleth the additional dimension can help to make quick observations regarding population relative to present competition, or to observe dense/sparse pockets potential customers alongside higher/lower income levels. It provides a useful tool for the client in the same ways as the heat map above (and the income choropleth to be described below).

The population choropleth shows that population levels for each community tend to be fairly evenly distributed with, as the legend(s) denote, darker shades indicating greater magnitudes in the data being mapped. However, given the inconsistent size and shape of different communities, a rough gauge of overall population density can be obtained and used to help determine community prioritization in the early stages of the client's operation. The map does show that communities with greater populations tend towards the exterior, which is valuable information given the observation made while discussing the heat map. The value of this observation is outlined below.

1. It further substantiates the validity of the business model by associating larger populations with underserviced areas/communities of the city. This is of value both for appealing to potential investors/creditors as well as for use in the prioritization of communities.

5.2.2: Median Household Income Choropleth

This map serves the client in essentially identical fashion to that of the population choropleth, so please refer to that section of the discussion for details regarding what is/can be gleaned from the Median Household Income choropleth map. Like population, income levels tend to grow as one moves outward from the more central communities which can be applied in conjunction with those observations made above about the validity of the business model and the applicability of the map(s) as useful tools in day to day operations.

5.2.3: Market Saturation Choropleth.

By replacing null values for the 'Population' feature with the feature's mean value this map allows for the explicit isolation of those communities void of competition since the only null regions on the map become limited to those for which 'Venue Count' is zero. These are the communities containing the unique 'magenta' fill color. Several key observations can be made from mapping the data this way.

1. One can easily visualize just how many communities lack competitors for the client's proposed Java truck. As mentioned earlier, the Foursquare API call returned null or zero results for over 70 communities throughout the city.
2. These underserviced areas are typically localized to the outskirts of the communities considered.

This substantiates the business model even further while, like the maps already discussed, provides a useful tool for the prioritization of communities by further supporting the notion that those communities on the outskirts of Calgary's suburbs should be targeted early on.

We can, however, make an observation from this map that can't be made from the population or saturation choropleth maps on their own. One can see that the Southern sector of the city contains both outskirt communities lacking competing establishments (magenta fill) but as well a core of lightly

shaded communities for which it is observed that there are few existing establishments for the given population. This value of this observation is summarized below.

3. The Southern region of the city is relatively unsaturated and potentially underserviced in comparison to the Northern region of the city.

5.2.4: Community Market Potential Index Choropleth

This is perhaps the most contextually valuable of the four choropleth maps in that it allows the client to visualize the measure of comprehensive measure of market potential on a community by community bases. An interesting observation drawn from this map is that it appears to argue against the notion that the South is a higher value target than the North of the city which was presumed from the *Market Saturation* choropleth. This is due to the index calculation having weighted the affluence of a community over population when ascertaining the market potential score for each community. We can see that, with the exception of the Northeast sector of the city, the data supports the business model's foundational assumption that, given recent geographical shifts in the location of demand for prepared craft coffee products as a result of the fallout of the COVID-19 pandemic, an opportunity exists for our entrepreneurial client to tap into underserviced markets with a mobile java truck service. The value here is summarized below.

1. This map visually depicts the index scores for each community suggesting to potential investors/creditors that the client has done their due diligence in performing relevant market research.
2. It demonstrates to investors/creditors that recent developments have generated substantial potential value in the outlying regions of the city for which the client's proposed business model is best suited to address/capitalize on, thus driving confidence in their proposition and in the belief that their concept can service it's liabilities and/or generate returns for investors.
3. It pinpoints those communities which, based on the data and methodology, present the highest individual potential for the client and their business, thus allowing them to target those higher value communities specifically or to hone in on higher potential clusters.
4. By prioritizing high potential communities and high potential groups of communities the client can presumably generate returns faster and can thus presumably establish core operating routes/communities/earlier on and as well can entertain the possibility of expanding earlier before similar outfits begin to exploit the same model in those area the client is not addressing.

5.3 Clustering

The K-Means clustering results provide a tool that is very similar to the *Community Market Potential Index Score* and it's corresponding choropleth map, and should be used in conjunction with the map so that another more surgical measure is available to offset those instances in which one appears to fall short in it's ability to predict high potential communities and regions. Using the scatterplot of the results (Figure 28 of the results section) each color and its respective color are prescribed descriptive titles/labels based on their income and market saturation characteristics. The clusters labels and their respective colors are categorized as follows…

- ***Cluster 0 (Red) 'Low Income Moderate Saturation'*****: Perhaps the least attractive of the clusters as incomes are low and saturation is higher than all but cluster 4 (green).**

- ***Cluster 1 (Blue) 'Moderate Income High Saturation':*** **Highest saturation of all of the community clusters, but moderate income.**

- ***Cluster 2 (Purple/Magenta) 'High Income Low Saturation'*****: These communities may be the communities our client wants to visit first. Their communities have fewer incumbent coffee shops and cafes, and the median household income is much higher than the rest.**

- ***Cluster 3 (Yellow) 'Low Income Low Saturation'*****: Similar in terms of income to cluster 3 (black), but perhaps higher potential due to lower saturation.**

- ***Cluster 4 (Green) 'Moderate Income Low Saturation'*****: Second highest potential community cluster based on the model and data.**

5.4 Additional Considerations

As is true in all data science applications, this is to be an iterative process. This is only the first stage in this analytical process, should the client wish to pursue the business question further for this given approach and its constituent parts.

The most obvious next step is for the client to actively accumulate operational data for each community. In addition, it may be worthwhile to expand the analysis above to incorporate some measure of the age of a communities residents into future analyses, as it was noted by the client that this was relevant (however, as mentioned earlier the scope was to be limited to income and population with regards to data retrieved from the community profiles.)

6: Conclusion

To conclude, the business question is first reiterated in order to remind the reader of the specific objective of the analysis.

How can the various communities of the City of Calgary be profiled/classified for the purpose of identifying those communities which could be expected to provide the highest likelihood of positive returns for our client?

Having obtained community level data from the City of Calgary's Community Profiles, the Foursquare Places API, and through the process of geocoding we were able to develop a concise profile for each community consisting of measures for the community's income, population, and the current number of competing establishments. From here features were developed to describe each communities level of market saturation and an overall measure of it's market potential for the client based on the data and methodology presented. These were converted into visual tools in the form of heat and choropleth maps to allow the customer to visually profile the communities, determine where high potential regions exist, and to establish a means of prioritizing communities. The K-Means clustering algorithm was used to explicitly segment the communities into their own clusters based on their income and saturation profiles. This information was as well translated into a map so that it can be used by the client during operations.