Fusion Coffee Lab Denver Expansion

Jason R. Foster

March 16, 2019

1 Introduction

1.1 Business Problem

In this study I consider myself to be the owner of Fusion Coffee Lab, a successful, brick-and-mortar coffee shop near Arizona State University in Tempe, AZ and seek to find a neighborhood into which I can expand in the Denver, CO metropolitan area.

Combining demographic data with location data from a Location-Based Social Network (LBSN) I will attempt to use K-Means Clustering to find similar locations in the Denver metropolitan area that I'm hoping will allow me to limit my search areas and focus instead on a more specific location and available real-estate.

Knowing my current customers, my expansion would ideally be in an area in or near one of Denver's older but growing neighborhoods, possibly in or near one of the Top Neighborhoods for Young Professionals, not in close proximity to an international airport (which is something I don't like about my current location), have less than 50% of existing coffee shops be independent, and possibly near some or all of the following 'anchor' venues:

- A university and/or medical complex
- An arts district and/or performing arts complex
- Professional office buildings

1.2 Target Audience

While this study is limited in scope, namely the expansion of a single business, and may not have broad appeal, it does provide a working example and demonstrates the power of data in providing actionable intelligence to a business. If I were the business owner, it would help me to narrow my search to specific areas of Denver that might feel familiar and in which I could find similar customers, and help me to focus my efforts on other aspects of expansion, such as staffing, real-estate, and a modified business plan.

2 Data

In order to solve the stated problem I am going to be combining data from multiple sources.

- A list of target zip codes in the Denver metropolitan area and for Tempe, AZ
- Demographic information for all the zip codes
- Categorical, venue and attraction data from an LBSN
- A comprehensive list of franchise coffee shops

The following data sources will be used to collect the necessary data

- Zip-Codes.com for the list of zip codes in and around Denver
- HomeTownLocator.com for the demographics of the various zip codes and for Tempe, AZ
- Google Geolocation API for locating the centroids of the zip codes
- Foursquare API for the recommended venue data and category data
- Wikipedia for a list of franchise coffee shops

2.1 Data Acquisition

Note that the bulk of the code responsible for much of the work in this study, including obtaining the data, exists in my **Python Utility Module**, which is also linked at the bottom of the study.

2.1.1 Zip Codes and Demographics

The list of zip codes for Denver metro area are first obtained from Zip-Codes.com and to that list we add the zip code for Tempe, AZ (85281). This list is then used to scrape demographics from HometownLocator.com. Note that there are six zip codes in the Denver metro area for which HometownLocator has no demographics, and two for which it has demographics but the values are all zeros. When manually inspecting the pages for these zip codes, the site suggests to use `80202` as a substitute. In scraping demographics, I do fall back to that zip which creates duplicates in the results, which I then remove. This means that there are eight zip codes that are being removed from the list obtained via Zip-Codes.com.

Because of the growth in Denver and the associated inflation in the cost of real estate and therefore salaries, note that I am making an adjustment two two demographic features for Tempe. Specifically, I'm adjusting the median home value by a factor of 1.53 and the average household income by a factor of 1.21. This is based on information from Sperling's Best Places. These adjustments allow for a like-for-like comparison on the two features.

The demographics site provides 19 features per zip code which are divided into four categories: Population, Housing, Income and Households. The following table summarizes the features in each category.

Section	Notes
Population	Total population, population in families, households, density and diversity index
Housing	Total Housing Units (owner- and renter-occupied, vacant) and average home values
Income	Median and mean household income and per capita income
Households	Total households, average household size, family households and average family size

The following table displays a sample of the demographic data obtained.

Out[2]:

	ZipCode	Latitude	Longitude	Diversity Index3	Median Home Value	Population Density2
90	80302	40.038629	-105.371668	28	673618	292
91	80303	40.000538	-105.207780	37	558063	738
92	80304	40.045474	-105.283851	38	651186	3573
93	80305	39.979999	-105.248737	27	528506	2462
94	85281	33.436665	-111.940325	74	302061	5453

2.1.2 Foursquare Venues

Next, the top 100 recommended venues are obtained for each of the zip codes using the Foursquare API. Note that we allow the Foursquare API to determine the appropriate radius for our search, because it knows the density of venues in the zip code and will choose an appropriate one to accommodate our request for 100 venues.

The following table displays a sample of the venue data.

Out[3]:

	ZipCode	Venue	Venue Main Category	Venue Top-Level Category
0	80202	Denver Union Station	Train Station	Travel & Transport
1	80202	Whole Foods Market	Grocery Store	Shop & Service
2	80202	TAVERNETTA	Italian Restaurant	Food
3	80202	Wynkoop Brewing Co.	Brewery	Nightlife Spot
4	80202	Hopdoddy Burger Bar Denver	Burger Joint	Food

Simpson's Diversity Index

The method used in the course labs for clustering based on the top venues and their categories generated a lot of features. Clustering high-dimensional data comes with its own set of problems and usually requires techniques other than K-Means. In our case, this is true because the concept of 'distance' between points is based on the means of a lot of zeros and ones, which over a large number of features, becomes almost meaningless. In my own clustering work for this class, based on silhouette analysis, I discovered repeatedly that the best number of clusters was 2.

Because of this, I'm using what is called Simpson's Diversity Index (SDI) which is normally meant for measuring biodiversity, but it does allow us to measure the *richness* and *evenness* of populations. It is easily adaptable to venues and their categories. In my approach, I'm collecting the tree of Foursquare Categories and using the top-level categories as a sort of 'species'. This approach allows me to preserve the information in the types and numbers of venues and significantly reduce the dimensions in the dataset, and results in just 28 features in the final dataset.

The following table displays a sample of the SDI data.

Out[4]:

	ZipCode	Food	Nightlife Spot	Outdoors & Recreation	Shop & Service
80	80302	22.084507	5.400000	4.200000	7.142857
81	80303	14.440000	2.117647	7.714286	15.206897
82	80304	21.591837	2.571429	5.062500	11.636364
83	80305	12.000000	4.000000	3.700508	13.444444
84	85281	17.808696	2.941176	4.545455	7.000000

One-Hot Encoding

For the last dataset we will utilize one-hot-encoding to generate dummy variables for the myriad of venue categories returned form Foursquare and then use those categories to generate the top ten most recommended venues for each of the target zip codes. This data will be used later in the study when we examine the similarity of the clusters we generate.

The following table displays a sample of the top venues returned from Foursquare for each zip code.

Out[5]:

	ZipCode	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
80	80304	Park	Trail	Mexican Restaurant
81	80305	Trail	Coffee Shop	Sandwich Place
82	80401	Trail	Brewery	Coffee Shop
83	80403	State / Provincial Park	Trail	Park
84	85281	Coffee Shop	Pizza Place	Sandwich Place

3 Methodology

Given the primary sources of data, I will be performing the following top-level steps in this study

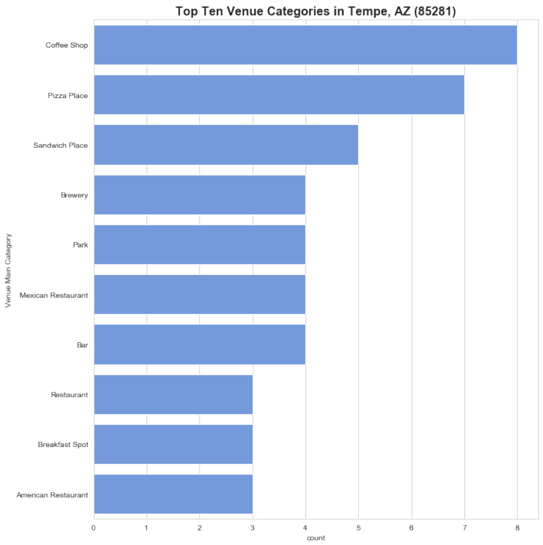
1. Perform exploratory analysis of the main data sources to identify any patterns and to help with possible feature reduction
2. Perform K-Means Clustering to identify candidate zip codes, which are those that are clustered with Tempe
3. Explore the cluster containing Tempe, and identify the best zip codes, based on the stated criteria
4. Provide guidance on the top zip codes, based on the mix of franchise vs. independent coffee shops in the candidates

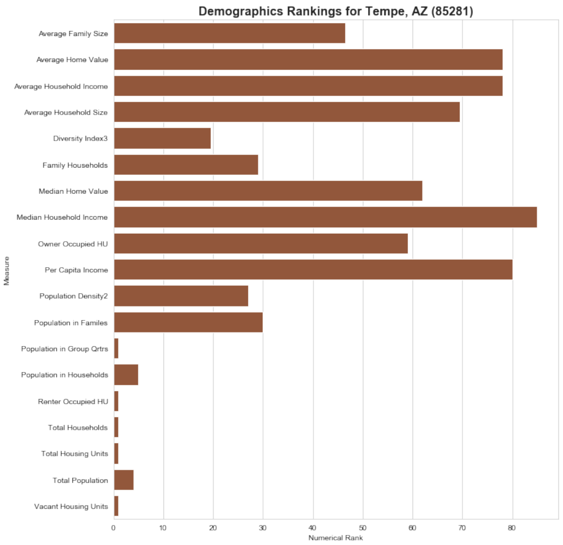
3.1 Exploratory Analysis

Some exploratory analysis of the data is performed to identify any patterns or issues in the data that must be addressed prior to finding similar zip codes.

3.1.1 Tempe

Below we examine both the top ten venues in Tempe and the rank of each of the demographic attributes.





One other item of interest is the top venues visited after visiting my shop. The following code utilizes Foursquare's NEXTVENUES endpoint to find the top five venues most likely to be visited after my shop.

Out[9]:

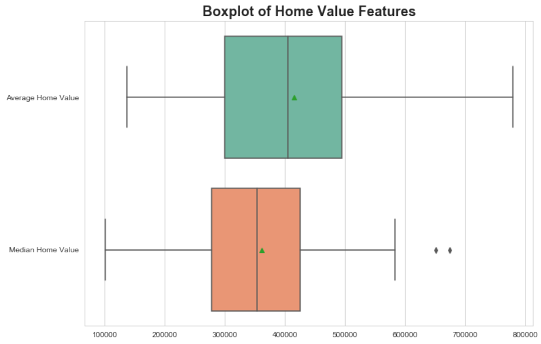
	Venue	Venue Main Category	Venue Latitude	Venue Longitude
0	Casey Moore's Oyster House	Seafood Restaurant	33.420793	-111.942643
1	Buffalo Exchange	Thrift / Vintage Store	33.421699	-111.943222
2	Otto Pizza and Pastry	Pizza Place	33.421443	-111.942493
3	Dutch Bros. Coffee	Coffee Shop	33.416966	-111.926024
4	Four Peaks Brewing Company	Brewery	33.419517	-111.915911

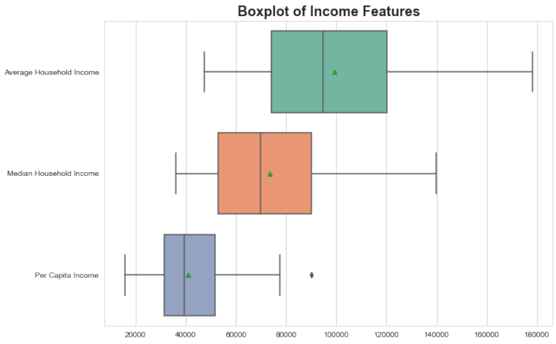
3.1.2 Demographics

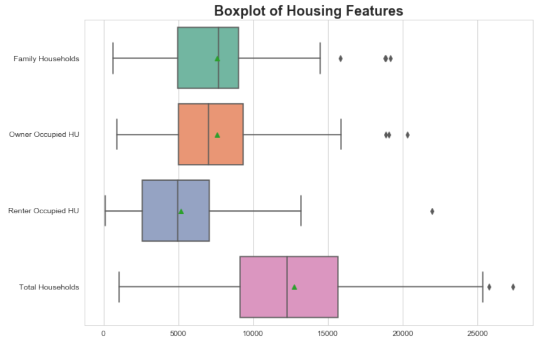
As we have seen from displays of the dataset, the demographic data contain a great deal of variety in terms of features. Specifically the data contain

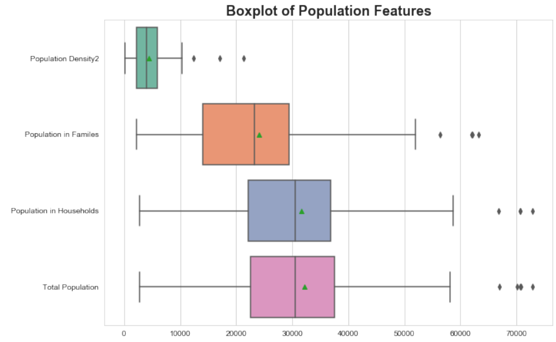
- Monetary values for things like income and property values, which may be very different in some cases
- A Diversity Index, which is essentially a probability that two persons, chosen at random from the zip code, belong to different race or ethnic groups. This is essentially another SDI.
- Quantitative values, such as the number of owner-occupied housing units and various population counts
- Measures such as population density which is the total population per square mile

Next I want to look at the descriptive statistics that summarize the central tendency, dispersion and shape of a dataset's distribution, and to see if there are values that look like outliers. The following boxplots display some of the dataset's features, and are separated into groups of 'like' features.


Boxplot of Home Value Features


Boxplot of Income Features


Boxplot of Housing Features

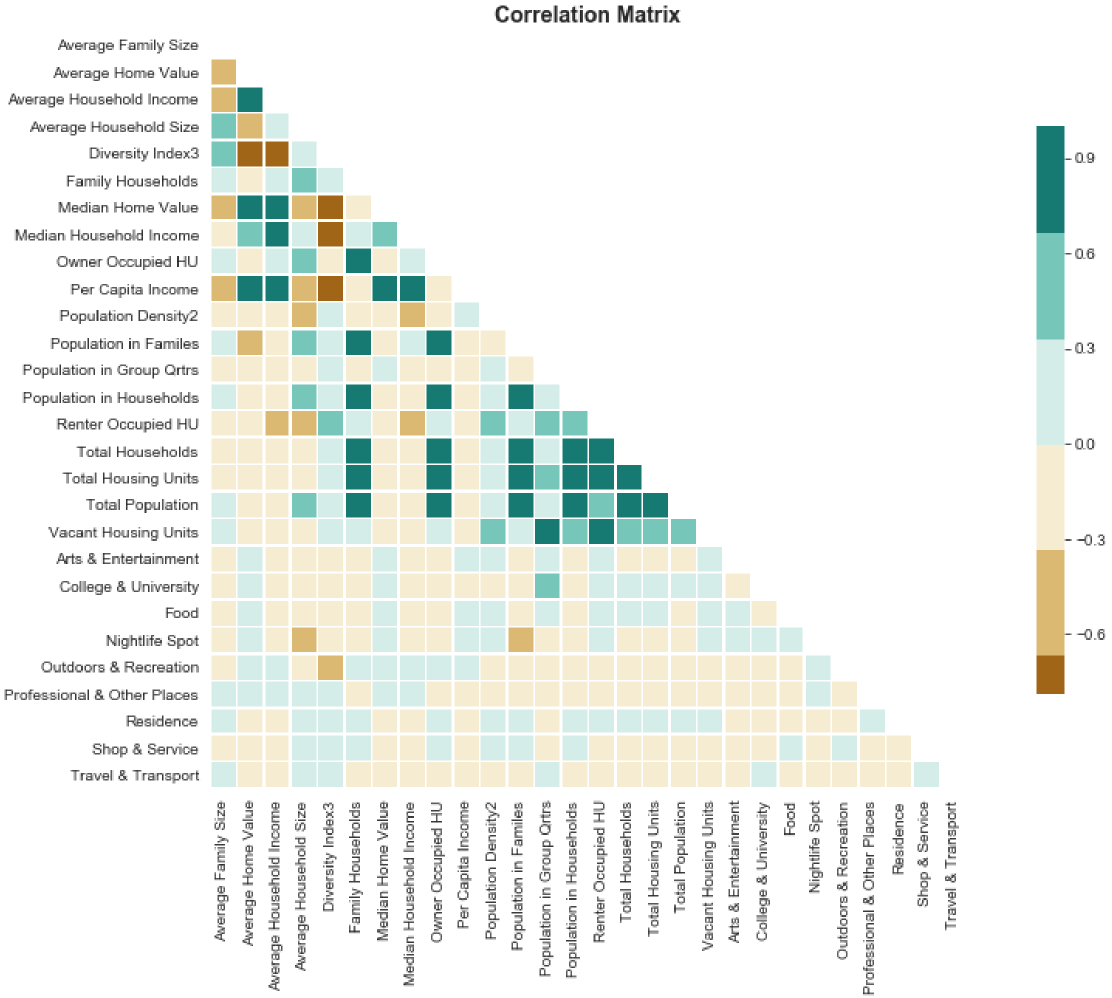

Boxplot of Population Features

Observations on Demographics Features

The boxplots show some values for some of the features as outliers. As an example, Total Population has some very notable and large outliers, one of which happens to be Tempe, the other of which is 80219. The latter has a population density, however, that isn't an outlier, which indicates that the population is perhaps a function of another variable, possibly the total land area of the zip code. Similarly, the outliers in population density are both very small counties in Denver and suggest that perhaps there are a higher number of multi-family housing units in those zip codes. Please note that I looked for data that had a breakdown of the types of housing units in each zip code, but could find no reliable source. I conclude from what I've seen and the research I've done that some of the features I'm using are likely affected by other features that I cannot reliably collect or determine.

That being said, I do not believe I can qualify any of the values in the dataset as true 'outliers'. In a sense, the values are what they are and are intrinsic measurements of a specific area. K-Means clustering is sensitive to outliers, so I need to choose a method of scaling the data that will appropriately represent the outliers. `Scikit-learn` includes a preprocessor named `RobustScaler` which removes the median and scales the data according to the quantile range and is therefore not influenced by a few number of very large marginal outliers. I will use this preprocessor to scale the data prior to clustering.

Next, I examine the correlation matrix of the entire dataset to see if we can find any patterns that might help with further dimensionality reduction.



Observations on Correlation Matrix

From the matrix, there are some correlations that make sense. For home values, the average and median values are strongly positively correlated, as are median/average home values and average household income. Households and housing units are obviously strongly correlated because, while a housing unit is a house, an apartment, a mobile home, a group of rooms, or a single room occupied (or if vacant, intended for occupancy) as separate living quarters, a household includes all the people who occupy a housing unit. So, these two things would vary depending on just sheer numbers of housing units.

A couple of interesting correlations are also revealed. Average home value is strongly negatively correlated to both diversity index and average household size, which means larger households occupy lower-valued homes and that neighborhoods containing a high-degree of a single ethnicity (regardless of the ethnicity) contain lower-valued homes. Similarly, per capita income is also strongly negatively correlated with diversity index, meaning that more diverse neighborhoods have lower incomes than less diverse ones. Lastly, the venue category Nightlife Spot was fairly strongly negatively correlated with Population in Families, possibly suggesting that families don't choose to live near nightlife spots.

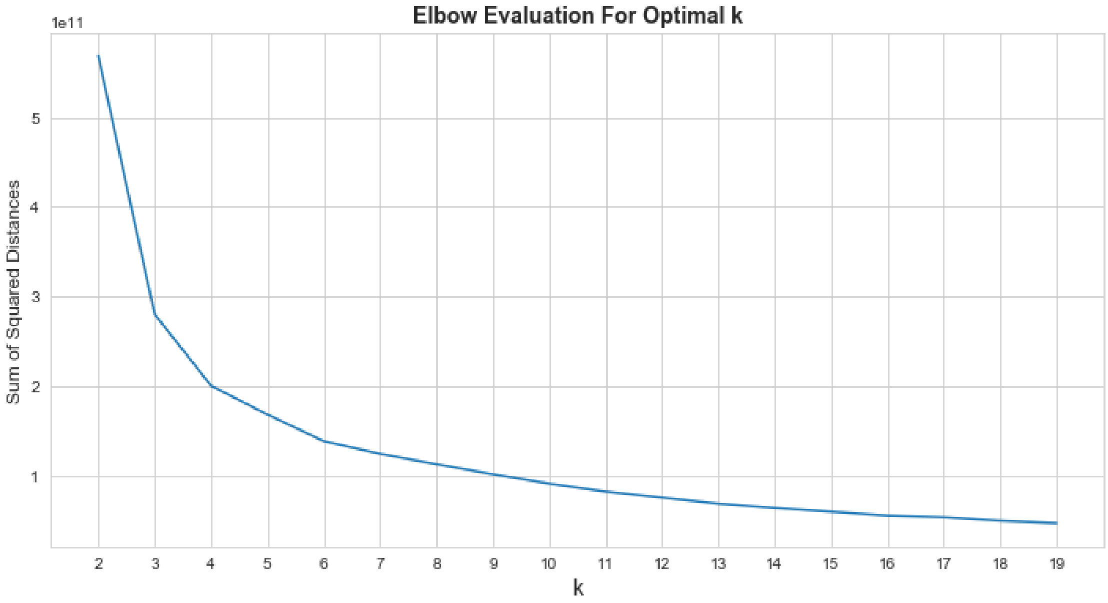
3.2 Dimensionality Reduction

Based based on the correlations in the matrix and an understanding of the underlying data, I am going to be removing several features. Total population is a good candidate for removal, since we have a similar and more rich feature, Population Density that is more descriptive because it takes into account the size of a zip code, which is data I cannot obtain. Since I have made a modification to Median Home Value and Average Home Value is strongly correlated to that, I'm removing Average Home Value. I'm also removing the features related to housing units, because we have a similar measure, Total Households, to which it is also strongly positively correlated. I'm also removing one feature from the venues set, 'Residence', because I do not consider Foursquare to be a good source of residence information.

The resulting DataFrame contains 85 observations with 23 features.

3.3 Identifying Candidate Zip Codes

We will use our reduced-dimension dataset to cluster all the data to find any Denver metro zip codes that are in the same cluster as Tempe. We first scale the data using RobustScaler and then attempt to find an optimum number of clusters using the 'elbow' methodology.



Based on the results there is a distinct elbow around 6 or 7 and another at about 16. For this analysis, I'm choosing 7 as the number of clusters to build from the dataset.

4 Results

From the seven clusters generated, Tempe was placed into Cluster 6. With this in mind, we can now start to examine the cluster and the data we have to see if we can find from the candidates any suitable locations to pursue for my expansion. Below is a brief summary of the top venues in each of the zip codes in Cluster 6.

Out[17]:

	ZipCode	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
2	80204	Brewery	Mexican Restaurant	Coffee Shop
19	80236	Mexican Restaurant	Coffee Shop	Convenience Store
24	80249	Coffee Shop	Airport Lounge	Steakhouse
35	80003	Coffee Shop	Mexican Restaurant	Pizza Place
44	80214	Coffee Shop	American Restaurant	Mexican Restaurant
49	80232	Mexican Restaurant	Coffee Shop	Spa
56	80013	Convenience Store	Park	Pizza Place
57	80014	Mexican Restaurant	Coffee Shop	Pizza Place
62	80110	Mexican Restaurant	Coffee Shop	Grocery Store
84	85281	Coffee Shop	Pizza Place	Sandwich Place

This shows that for all but one zip code in the cluster, a Coffee Shop is in the top 3 venues recommended, and for that one where it isn't, that a Coffee Shop is still in the top 5. So, people are going to these zip codes for coffee, which is at least promising. Lets take a look at the demographics to see how those line up.

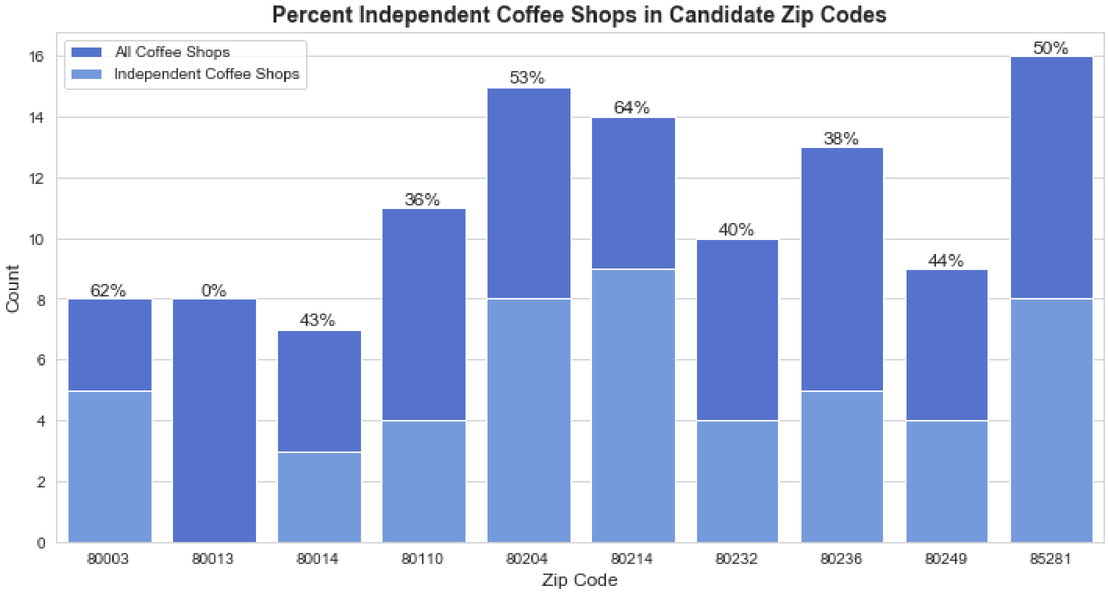
Out[18]:

	ZipCode	Median Home Value	Diversity Index3	Food	Shop & Service
2	80204	282512	81	9.573427	9.800000
19	80236	265220	77	14.857143	9.920635
24	80249	273947	86	12.235955	4.500000
35	80003	249237	56	17.893617	12.461538
44	80214	278333	73	13.630037	12.521739
49	80232	272250	62	18.150838	13.000000
56	80013	247916	69	11.226804	10.714286
57	80014	230819	65	20.787709	16.666667
62	80110	232251	67	12.255319	14.695652
84	85281	302061	74	17.808696	7.000000

4.0.1 A Look at Existing Coffee Shops

Lastly, I want to dig a little into the existing coffee shops in the candidate zip codes, see what types they are (meaning are they franchise or independent) and where they are located. I'm also only concerned with actual coffee shops, so because the Foursquare 'coffee' section returns things like tea rooms, cafés, and convenience stores (all of whom serve coffee), those venues aren't really like my shop. I also wanted to distinguish between franchise and independent coffee shops. We retrieve the top 100 recommended venues in the 'coffee' category, remove the venues that are not coffee shops, and determine whether each is a franchise or not using a list of coffee franchises obtained from Wikipedia.

The following summarizes the results of this analysis.



5 Discussion

The results of the clustering revealed nine other zip codes that were similar in nature to Tempe, and while each of these zip codes might warrant further investigation I'd like to choose a small set of them to focus on, using our stated criteria.

5.1 Deeper Look at Candidates

5.1.1 Exclusions

Of the nine other zip codes in the cluster, the following are being excluded from further consideration: 80003, 80013, 80014, 80214, and 80249. Zip code 80013 has no independent coffee shops, but the most commonly visited venue there is a convenience store, which I assume is a top venue for coffee. Zip code 80249 is near Denver International Airport. Zip codes 80003, 80214, and 80204 already have more than 50% independent coffee shops. Further, based on other investigation, many contain no 'anchor' venues, and despite some of their proximity to parks simply do not meet the stated criteria.

5.1.2 Final Candidate Zip Codes

The following zip codes show promise as possible expansion zip codes. Further analysis would be needed to determine overall feasibility which should include appropriate and available real estate (including proximity to any existing coffee shops), further demographic analysis to include breakdowns of age groups and ethnic diversity, crime rates and other relevant criteria.

80204

This area is just to the west of downtown Denver with close proximity to popular venues such as the Colorado Convention Center, Denver Center for the Performing Arts, and the trendy LoDo neighborhood. It also includes both Mile High stadium and the Pepsi Center, which host Denver's professional sports teams and A-List concerts. It encompasses both the Auraria Campus and the West Colfax and Lincoln Park neighborhoods. It is also host to the Santa Fe Arts District and has good proximity to Sloan's Lake Park. One shortcoming is that it already contains a relatively high number of independent coffee shops, approximately 53%. Coffee shops are also the third most common venue visited in this zip code.

80236

This is largely a residential neighborhood that contains Mullen High School, a prestigious private Catholic preparatory school. It is comprised the Harvey Park South neighborhood, which contains homes built in the 1950s, like Tempe. Charms and challenges in Denver's best kept "mid-mod" secret describes some of these homes by saying, "That these are special homes was far from secret in 1955."*This design was so well-thought-out that several of the buyers were young architects who had been trained after World War II on the G.I. Bill*" It is also about a ten minute drive to the University of Denver and has close proximity to Marsten Lake, which is a popular birding location. It has the third lowest percentages of independent coffee shops, at about 38%. Coffee shops are the second most common venue visited in this zip code.

80110

This area is a mix of commercial and residential almost directly south of downtown Denver. It has very close proximity to Cherry Hills Village, which is a very affluent neighborhood with a very prestigious golf club. It also has good proximity to Swedish Medical Center and Craig Hospital in Englewood as well as to the Gothic Theater, which is an up-close live music venue converted from an art deco movie house. It also has close proximity to Arapahoe Community College in Littleton and the University of Denver. It has the second lowest percentages of independent coffee shops, at about 36%. Coffee shops are the second most common venue visited in this zip code.

6 Conclusion

In this study I sought to identify some target zip codes in the Denver metro area that could potentially be a site for the expansion of my coffee shop in Tempe, AZ. I combined demographic data for selected zip codes in the Denver metro area with venue and category data from Foursquare and used the data as the basis for clustering to identify zip codes that are most similar to 85281. Based on the clustering nine candidate zip codes were found, of which three were identified as the most suitable for further investigation and planning. Before considering any expansion those three would need further, detailed analysis to identify a suitable location.

7 References

The following sources were used in the creation of this study.

- IBM Applied Data Science Course Materials and Labs
- **My Python Utility Module**
- Sperling’s Best Places
- Foursquare API Documentation
- scikit-learn Documentation
- Pandas Documentation
- Anytree Documentation
- Simpsons Diversity Index
- Google Geocoding API
- Top Neighborhoods for Young Professionals
- Charms and challenges in Denver’s best kept “mid-mod” secret
- List of Coffeehouse Chains