Coffee Shops and ASU Campuses

Zack Kenyon

March 24, 2020

Course: IBM Data Science: Professional Certificate

1 Introduction

1.1 Business Problem

Outside of partying, one of the most prolific college stereotypes is being wired on coffee and staying up late to finish that assignment or cramming for tomorrow's exams. Anyone who's spent time on a college campus will happily confirm that there is some truth to the stereotype. Naturally, the next question is where are these student's buying their coffee and where should we set up shop to tap into that market?
The goal of this project will be to determine potential locations one could open a coffee shop if you are wanting to take advantage of Arizona State University's (ASU) very large student population. ASU has 4 campuses: Tempe, Downtown Phoenix, West (Glendale) and Polytechnic (East Mesa). Each of these campuses offer vastly different landscapes in regards to the surrounding businesses. These areas will be analyzed to find the zipcodes within 1 mile of each campus that have a favorable amount of coffee and non-coffee related businesses.

1.2 Target Audience

The audience for a project such as this would be any aspiring restaurant owners, especially those interested in opening a coffee shop. Most restaurants close within a year of opening and while other factors certainly contribute to this, the hardest to overcome is a poor location. By understanding the areas in which you want to open your coffee shop, this issue can avoided or, at least, mitigated.

2 Data Collection

Location data will be the basis of this project. This includes geographic coordinate positions, venue names, venue categories, venue addresses, and geospatial files for zipcode mapping. The FourSquare API and Geopy's Nominatim package will be tools used to accomplish this.

2.1 FourSquare

Using Foursqaure's API, with sandbox credentials, the explore feature was used to gather all venues within a 1 mile radius from the listed address of each one of ASU's campuses. Given that Tempe's campus is nearly 1 square mile, it received a search point at both it's northeast and southwest ends to effectively capture the surrounding areas.
The explore feature yielded the venue names, categories and their latitudes and longitudes with 100 venues being located Downtown, 40 at Polytechnic, 100 at both NE and SE Tempe, and 72 at the West Campus. This information will then be used to determine within each zipcode: the 10 most common venues, the number of coffee shops, the number of restaurants and finally the frequency of coffee shops relative of other venues.
There were two important limitations regarding the data collected from FourSquare at this time. First, due to the COVID-19 outbreak, trending data could not be used. The data would have been skewed due to campus and business closures and since there is no way to pull historical trending data, it had to be removed altogether. Second, the addresses of the venues could not be called with Sandbox credentials as it would've taken an additional 6 days since user's only receive 50 address calls a day.

2.2 Geopy Nominatim

The Geopy Nominatim package converts addresses to geospatial coordinates. To get the coordinates for each campus, the five addresses were passed to the Nominatim object. Then, to overcome the FourSquare address limitation, the coordinates gathered from each venue were passed through Nominatim's reverse function, allowing the zipcode of each venue to be collected.

2.3 Data-Cleaning

After retrieving the venue information for SW and NE Tempe, they were both recombined into a single Tempe label. Duplicates were then dropped from the Tempe label by comparing both latitude and longitude values with

only the first instance being kept. In doing so, the number of Tempe venues dropped from 200 to 120.
In order to link venue zipcodes to the geojson file, some of the data points needed to truncated from the 9-digit code to the 5-digit. For example, 85004-3009 was shortened to 85004, preserving only the general 5-digit code and not the additional specifier. This narrowed the total number of zipcodes in the data set from 18 to 13.

3 Methodology

3.1 Exploratory Analysis

The first bit of exploration was in determining the total amount of coffee shops and total amount of restaurants within each zipcode. To do so, the dataframe sliced to contain only the zipcode and venue category. This slice was then searched for the category 'coffee shop', grouped by zipcode, and counted. A similar process was done for restaurants and chain coffee shop by searching the venue categories for 'Restaurant—Joint—Place' and 'Starbucks—Dutch' respectively. From this, it was seen that less than half of the zipcodes have one coffee chain and none of them have two. This was concerning, as it creates a feature of binary data type. When using K-Means clustering, binary features are typically avoided whenever possible as they create a strong separation line; it's ether 1 or 0. For this reason, it was dropped moving forward. Below is the resulting dataframe after conducting these operations.

Coffee Shop Frequency, Total Coffee Shops, Total Restaurants per Zipcode

	Zipcode	Coffee Shop Freq	Total Coffee Shops	Total Restaurants
0	85003	0.076923	4	18.0
1	85004	0.152174	7	10.0
2	85007	0.000000	0	1.0
3	85034	0.000000	0	0.0
4	85051	0.000000	0	1.0
5	85212	0.080000	2	5.0
6	85236	0.000000	0	3.0
7	85280	0.333333	1	0.0
8	85281	0.045455	4	45.0
9	85287	0.103448	3	12.0
10	85295	0.000000	0	2.0
11	85301	0.000000	0	4.0
12	85302	0.034483	2	24.0

Figure 1: In the above table, we can see that a number of zipcodes have 0 coffee shops at all. Those same areas seem to have little to no restaurants as well. On the other end of the spectrum, 85281 has 45 restaurants which is nearly double the next highest total.

To find the frequency of coffee shops and the top venue categories within each zipcode, a dataframe containing only the features: zipcode and venue category, was converted to a one-hot-encoded matrix. The one-hot-dataframe was then grouped by zipcode using the mean function for aggregation. Doing so allowed for each row to be sorted by frequency to collect the labels of top 10 most frequent venue categories. Below is the resulting data frame from these operations.

Top 10 Venue Types per Zipcode

Zipcode	#1 Most Common Venue	#2 Most Common Venue	#3 Most Common Venue	#4 Most Common Venue	#5 Most Common Venue	#6 Most Common Venue	#7 Most Common Venue	#8 Most Common Venue	#9 Most Common Venue	#10 Most Common Venue
85003	American Restaurant	Coffee Shop	Pizza Place	Hotel	Music Venue	Park	Salon / Barbershop	Theater	Thai Restaurant	Sandwich Place
85004	Coffee Shop	Hotel	Breakfast Spot	Art Gallery	Bar	New American Restaurant	Cocktail Bar	Market	Science Museum	Sandwich Place
85007	Pizza Place	Yoga Studio	Chinese Restaurant	Coffee Shop	College Baseball Diamond	College Stadium	Comedy Club	Comic Shop	Construction & Landscaping	Convenience Store
85034	Museum	Yoga Studio	History Museum	Coffee Shop	College Baseball Diamond	College Stadium	Comedy Club	Comic Shop	Construction & Landscaping	Convenience Store
85051	Convenience Store	Park	Grocery Store	Hunting Supply	Italian Restaurant	Yoga Studio	Dive Bar	Discount Store	Dessert Shop	Deli / Bodega
85212	Rental Car Location	Airport Terminal	Airport	Hotel	Coffee Shop	Wings Joint	Convenience Store	American Restaurant	Brewery	Fast Food Restaurant
85236	Fast Food Restaurant	Home Service	Smoke Shop	Golf Driving Range	Hotel	Construction & Landscaping	Miscellaneous Shop	Paper / Office Supplies Store	Sandwich Place	Bank
85280	Coffee Shop	Pool Hall	Convenience Store	Yoga Studio	Chinese Restaurant	College Baseball Diamond	College Stadium	Comedy Club	Comic Shop	Construction & Landscaping
85281	Pizza Place	Bar	Sandwich Place	Coffee Shop	Chinese Restaurant	American Restaurant	Burger Joint	Restaurant	Breakfast Spot	Brewery
85287	Coffee Shop	Mexican Restaurant	Deli / Bodega	Sandwich Place	Fried Chicken Joint	New American Restaurant	College Stadium	Comedy Club	Middle Eastern Restaurant	Burger Joint
85295	Asian Restaurant	Sandwich Place	Big Box Store	Yoga Studio	Donut Shop	Dive Bar	Discount Store	Dessert Shop	Deli / Bodega	Ethiopian Restaurant
85301	Intersection	Bar	Chinese Restaurant	Shop & Service	Pizza Place	Garden	Thai Restaurant	Mexican Restaurant	Convenience Store	Discount Store
85302	Pizza Place	Grocery Store	Mexican Restaurant	Donut Shop	Gas Station	Bar	Pharmacy	Chinese Restaurant	Coffee Shop	Intersection

Figure 2: In the above figure, we can see the top 10 venues per zipcode. One thing to notice is that zipcodes 85280, 85007 and 85034 have nearly identical category listings even though they are not geographically close.

Looking at the identical zipcodes, what likely happened is that after all the venues fell into the first few categories, the remaining categories were filled in alphabetical order, all having a frequency of 0. When looking at the total venues for 85007, 85034, 85280 this was confirmed as they have 1, 1, and 4 venues respectively.

3.2 K-Means Clustering

With information collected regarding each zipcode's coffee shop frequency, total coffee shops, total restaurants, and, top 10 venues, K-Means clustering was used to determine clusters of similar zipcodes. Using a k-value of 5, the model was fit to one-hot-dataframe to give 5 zipcode clusters. The number of clusters was assigned to be 5 after testing 3, 5 and 7 clusters. When using 3 clusters, the model failed to yield enough variance 8 zipcodes falling in one cluster, and 4 in another and 1 in the last. 7 clusters divided the zipcodes too much as it results in three 1 zipcode clusters. Ultimately, 5 clusters struck a nice middle ground in that respect and upon further look, there seemed to be a easy 5 way split for both the features: total restaurants and frequency of coffee shops.

Zipcode, Clusters and their Venue Information

Zipcode	Cluster	Coffee Shop Freq	Total Coffee Shops	Total Restaurants	#1 Most Common Venue	#2 Most Common Venue	#3 Most Common Venue	#4 Most Common Venue	#5 Most Common Venue	#6 Most Common Venue	#7 Most Common Venue	#8 Most Common Venue	#9 Most Common Venue
85003	0	0.076923	4	18.0	American Restaurant	Coffee Shop	Pizza Place	Hotel	Music Venue	Park	Salon / Barbershop	Theater	Thai Restaurant
85302	0	0.034483	2	24.0	Pizza Place	Grocery Store	Mexican Restaurant	Donut Shop	Gas Station	Bar	Pharmacy	Chinese Restaurant	Coffee Shop
85007	1	0.000000	0	1.0	Pizza Place	Yoga Studio	Chinese Restaurant	Coffee Shop	College Baseball Diamond	College Stadium	Comedy Club	Comic Shop	Construction & Landscaping
85034	1	0.000000	0	0.0	Museum	Yoga Studio	History Museum	Coffee Shop	College Baseball Diamond	College Stadium	Comedy Club	Comic Shop	Construction & Landscaping
85051	1	0.000000	0	1.0	Convenience Store	Park	Grocery Store	Hunting Supply	Italian Restaurant	Yoga Studio	Dive Bar	Discount Store	Dessert Shop
85280	1	0.333333	1	0.0	Coffee Shop	Pool Hall	Convenience Store	Yoga Studio	Chinese Restaurant	College Baseball Diamond	College Stadium	Comedy Club	Comic Shop
85295	1	0.000000	0	2.0	Asian Restaurant	Sandwich Place	Big Box Store	Yoga Studio	Donut Shop	Dive Bar	Discount Store	Dessert Shop	Deli / Bodega
85281	2	0.045455	4	45.0	Pizza Place	Bar	Sandwich Place	Coffee Shop	Chinese Restaurant	American Restaurant	Burger Joint	Restaurant	Breakfast Spot
85004	3	0.152174	7	10.0	Coffee Shop	Hotel	Breakfast Spot	Art Gallery	Bar	New American Restaurant	Cocktail Bar	Market	Science Museum
85287	3	0.103448	3	12.0	Coffee Shop	Mexican Restaurant	Deli / Bodega	Sandwich Place	Fried Chicken Joint	New American Restaurant	College Stadium	Comedy Club	Middle Eastern Restaurant
85212	4	0.080000	2	5.0	Rental Car Location	Airport Terminal	Airport	Hotel	Coffee Shop	Wings Joint	Convenience Store	American Restaurant	Brewery
85236	4	0.000000	0	3.0	Fast Food Restaurant	Home Service	Smoke Shop	Golf Driving Range	Hotel	Construction & Landscaping	Miscellaneous Shop	Paper / Office Supplies Store	Sandwich Place
85301	4	0.000000	0	4.0	Intersection	Bar	Chinese Restaurant	Shop & Service	Pizza Place	Garden	Thai Restaurant	Mexican Restaurant	Convenience Store

Figure 3: The above table has been sorted by cluster. We can see that 2 zipcodes are in cluster 0, 5 in cluster 1, 1 in cluster 2, 2 in cluster 3 and 3 in cluster 4.

4 Results

4.1 Choropleth

Choropleth of Total Coffee Shops in Each Zipcode

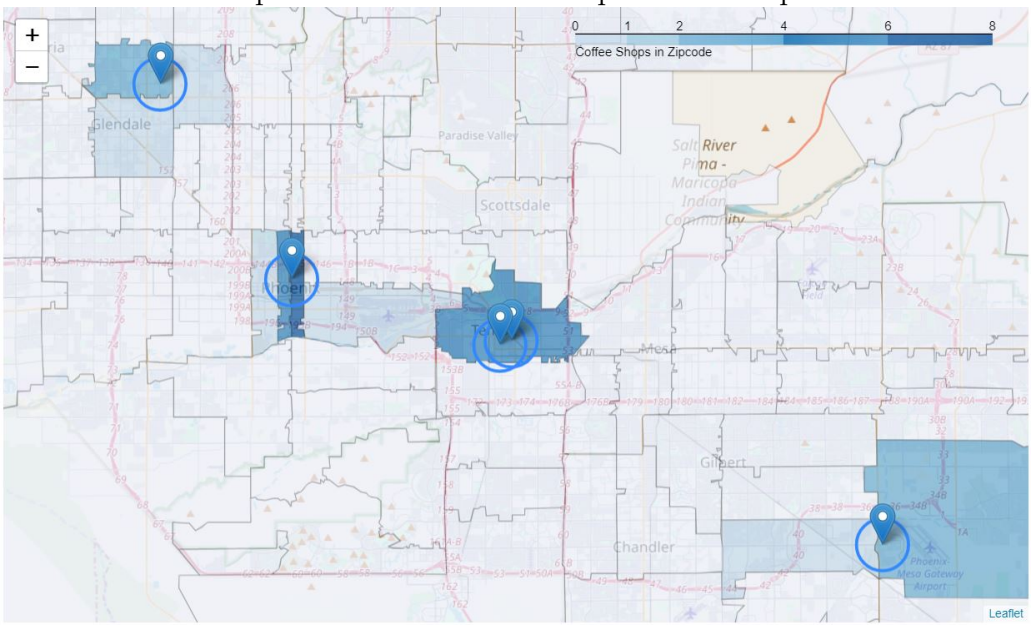

Figure 4: In the above figure, each campus is displayed along with the 1 mile search radius used to collect the data. The campuses are, in order of top left to bottom right, West, Downtown, Tempe, and Polytechnic. The zipcode with the most coffee shops is just east of the Downtown campus, followed by the west side of Downtown and Tempe.

4.2 Cluster Maps

Downtown Campus Clusters

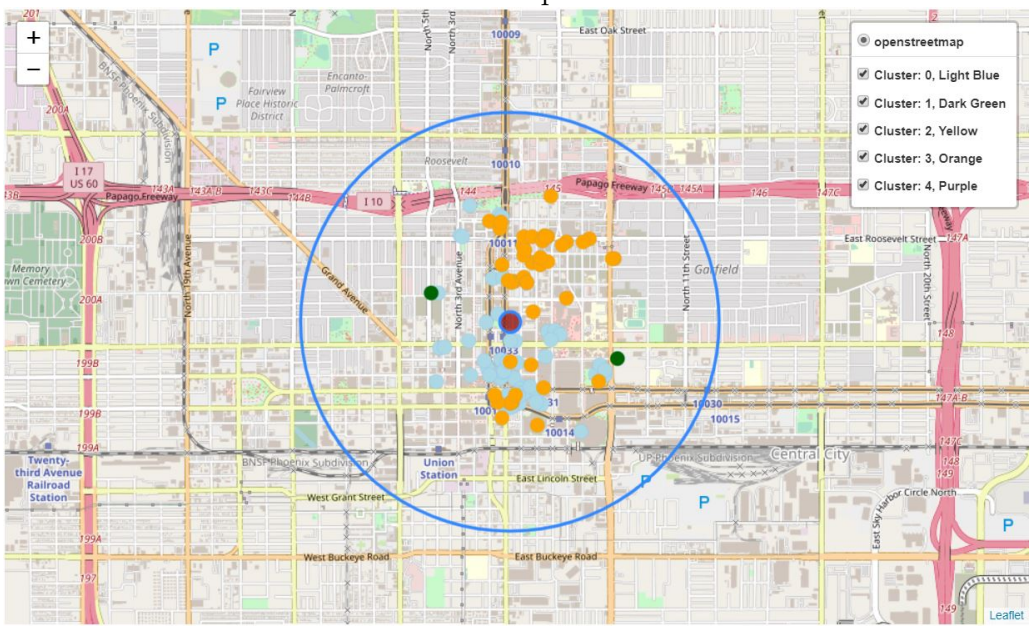

Figure 5: In the above figure, we can see the Downtown campus, indicated by the red circle in the center. The campus is surrounded by the 1 mile search radius and the cluster regions have been 'painted' on the map using the venues in each cluster to shade the map. The Downtown campus is mainly divided into Cluster 0, light blue and Cluster 3, orange.

Polytechnic Campus Clusters

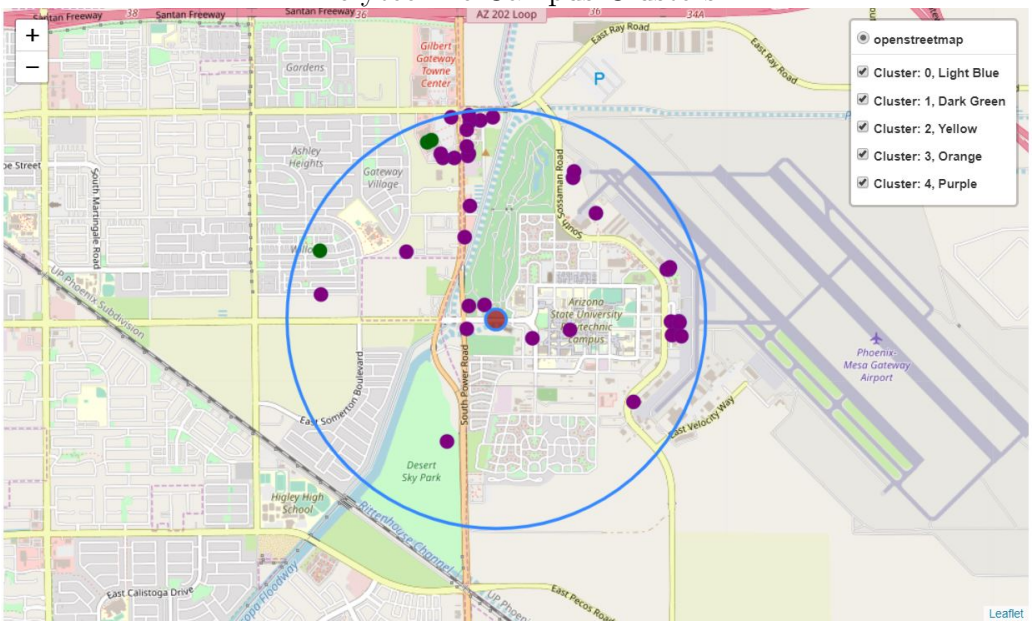

Figure 6: In the above figure, we can see the Polytechnic campus, indicated by the red circle in the center. The campus is surrounded by the 1 mile search radius and the cluster regions have been 'painted' on the map using the venues in each cluster to shade the map. The Polytechnic campus is surrounded primarily by Cluster 4, as indicated by the purple markers.

Tempe Campus Clusters

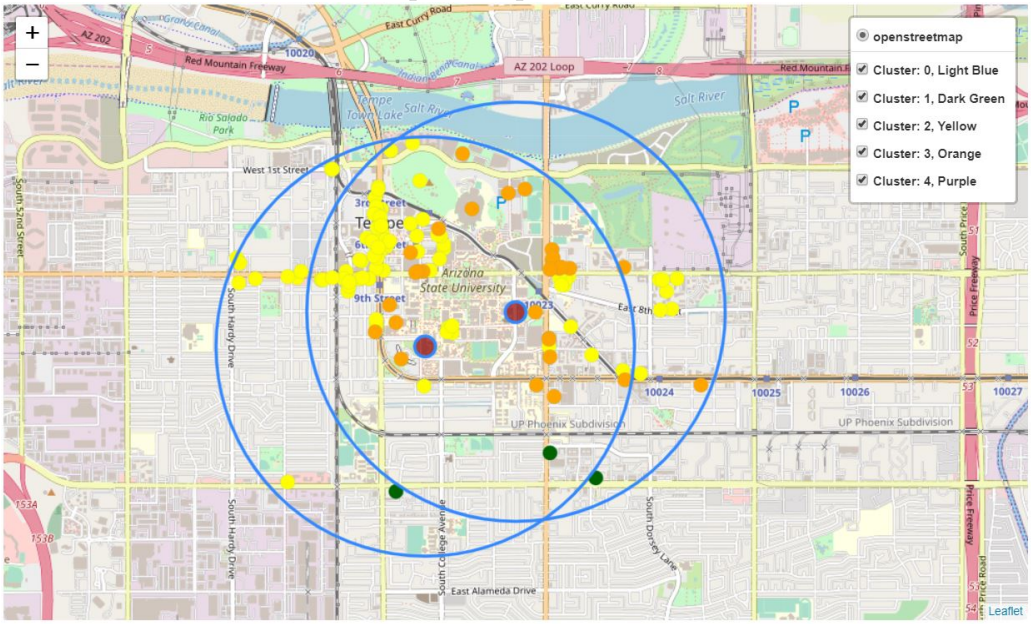

Figure 7: In the above figure, we can see the Tempe campus, indicated by the red circle in the center. The campus is surrounded by the 1 mile search radius and the cluster regions have been 'painted' on the map using the venues in each cluster to shade the map. The Tempe campus is made up of primarily Clusters 2 and 3 indicated by the yellow and orange markers respectively.

West Campus Clusters

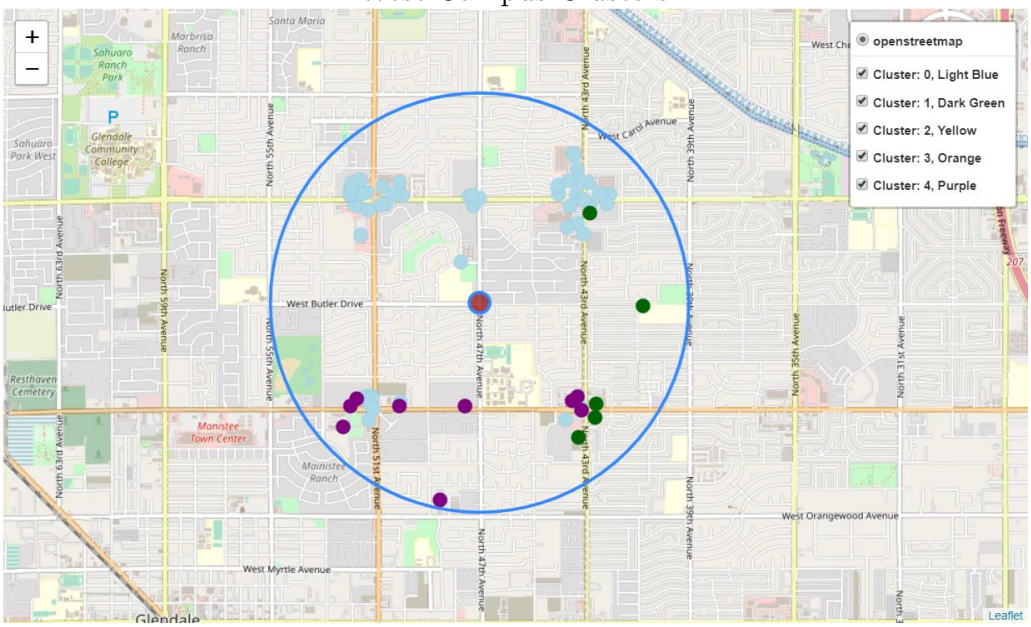

Figure 8: In the above figure, we can see the West campus, indicated by the red circle in the center. The campus is surrounded by the 1 mile search radius and the cluster regions have been 'painted' on the map using the venues in each cluster to shade the map. Cluster 0 mainly has venues on the north side of campus while the south side of campus is divided amongst Clusters 0, 4, and 1.

5 Discussion

5.1 Choropleth

When looking at figure 4, the choropleth, it becomes clear that the majority of the ASU coffee shops are surrounding the Downtown and Tempe campuses. This give a base understand of where the market is the currently most saturated. To avoid direct competition, it would be smart to stay away from the east side of the Downtown campus and the Tempe campus. The downside to this however, is that the student population decreases drastically when going to the West or Polytechnic campuses. An immediate issue with this choropleth map that could be improved upon is the outdated Tempe zipcodes. It appears that the government Geojson file used to populate the regions only includes 85281 and not 85280.

5.2 Cluster 0

Cluster 0 was the most common in terms of venues but only consisted of 2 zipcodes located primarily on the southwest side of the Downtown campus and on the north side of the west campus. The 2 zipcodes contained in Cluster 0 contain 18 and 24 total restaurants. As mentioned previously, there appears to be a couple obvious tiers in that regard. The top 10 venues only have two categories in common: Pizza Place (1st and 3rd) and coffee shop (2nd and 9th) respectively.

5.3 Cluster 1

Cluster 1 was the most common cluster with 5 zipcodes but it was least populated with venues at with only 13. Each zipcode in Cluster 1 has 0-1 coffee shops and less than 5 restaurants within them. With 5 zipcodes in this cluster, it does appear at all four campuses. Its top 10 venues are nearly identical amongst 3 of the 5 zipcodes with major venues being museums, yoga studios and college stadiums. The remaining two are likewise very similar amongst its 5-10 venues, sharing commonality with the rest of the group in the presence of yoga studios.

5.4 Cluster 2

Cluster 2 was the second most populated cluster with 88 venues packed into only 1 zipcode located at the Tempe campus. It has 4 coffee shops and 45 restaurants giving it more restaurants than clusters 1, 3 and 4 combined. The majority of Cluster 2 is located on the northwest side of the Tempe campus making that location arguably the most competitive for opening a restaurant but not necessarily a coffee shop.

5.5 Cluster 3

Cluster 3 contains 75 venues spread across 2 zipcodes. The majority of cluster 3 is located on the east side of the downtown campus with some scattered venues at the Tempe campus. Both zipcodes of cluster 3 have a coffee shop as their most common venue with breakfast and deli shops following close behind. A coffee shop in cluster 3 would yield the most direct competition.

5.6 Cluster 4

Cluster 4 is similar to cluster 1 in that it has very few coffee shops and restaurants spread out amongst its 3 zipcodes. It is located primarily at the

Polytechnic campus, a few at West, having 46 venues with only 2 of them are coffee shops and 12 are restaurants. Further, the top 10 venues have next to no commonality making a decision regarding viability on this cluster a bit difficult.

5.7 Recommendation

Now, with an understanding of the types of venues within each cluster a determination can be made as to which regions an aspiring coffee shop owner should be looking into. If you are interested in a maximizing student turnout while minimizing direct competition, then Cluster 2 is the way to go. There are lots of restaurants and bars in the area but only 4 coffee shops and the Tempe campus has the largest student population. There could however be some indirect competition with the sheer amount of choices for restaurants.

If the number of other restaurants is a concern, the cluster 0 would be the best bet. Specifically out near the West campus. Cluster 0 at the Downtown campus is in direct competition with the coffee shops of cluster 3, which you should avoid, given their close proximity. Cluster 0 at the west campus contains 24 restaurants, less than half that of cluster 2, while having just 2 coffee shops. Further, a coffee shop is only the 9th most common venue. This cluster is also the only one of the group that contains a grocery store, gas station and pharmacy in its top 10 venues indicating a more suburban area thus the greater potential for a more diverse client base to coincide with the ASU West population.

Finally, Clusters 1, 3, and 4 should be avoided. Cluster 3 specifically will provide the most direct competition to a coffee shop and in its common venues, it also contains breakfast spots and Delis which would like provide even more competition for the morning crowd. Clusters 1 and 4 should be avoided for different a reason: they are wild card areas. For section 1, the primary connection seems to be that they are relatively empty areas. Cluster 4 is near a small airport and only has 12 restaurants spread out amongst 3 zipcodes. There's a line between no competition and no market that Cluster 4 appears to be crossing.

6 Conclusion

In conclusion, the objective of project was to identify possible locations for opening a coffee shop near ASU campuses. FourSquare and Geopy were utilized to collect venue information within a 1 mile radius of each campus. This

information was then used to make clusters of zipcodes in which these venues preside in. Five clusters were created with a K-Means clustering algorithm. From these clusters, two options were presented as recommendations. The first was the cluster 2, at the Tempe Campus, for its large student population, lots of nearby restaurants and relatively few coffee shops. The second recommendation was cluster 0, at the West campus, for having a bit less restaurants and coffee shops, and a more suburban location while avoiding the direct competition provided by cluster 3 at the downtown campus.

7 References

FourSquare API: https://enterprise.foursquare.com/products/places
Goepy.Nominatim: https://geopy.readthedocs.io/en/stable/#nominatim
Folium Mapping Package: https://python-visualization.github.io/folium/modules.html
AZ Zipcodes: https://github.com/OpenDataDE/State-zip-code-GeoJSON