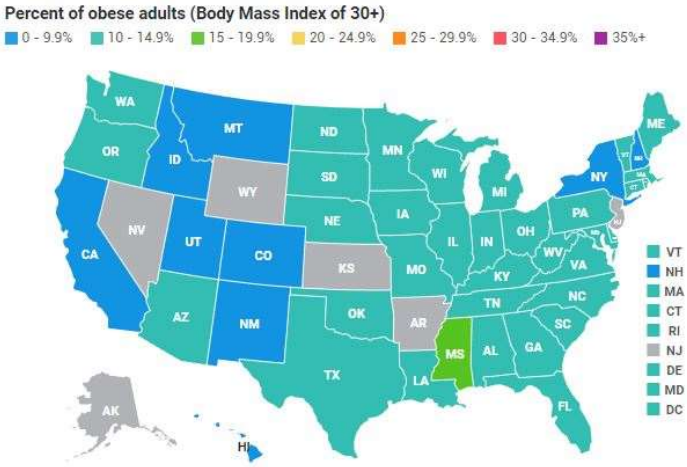
Comparing Community Demographics to Fast Food Appetites

By: Richard Queen

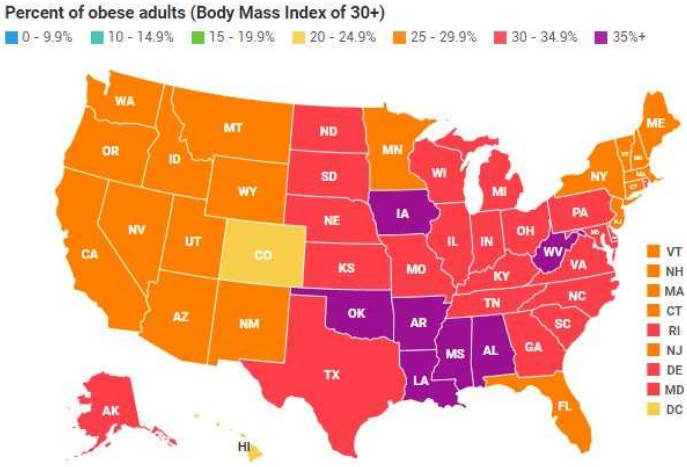
December 21, 2019

Coursera Data Science Capstone Project

1990



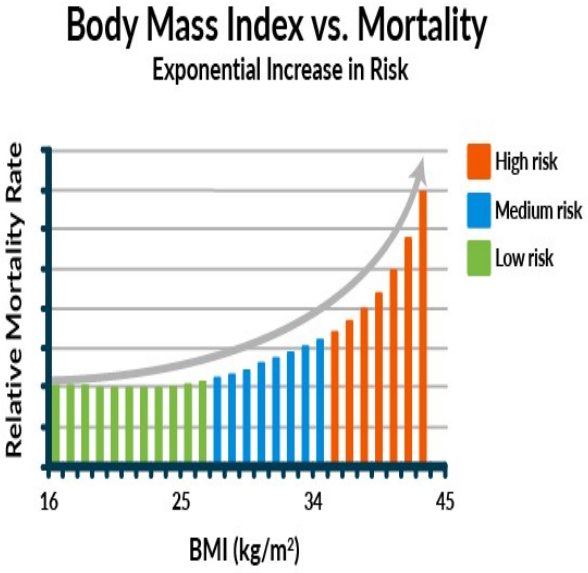
2017



I. Introduction

A. Description & Discussion

It's no secret that obesity has become a major healthcare problem in the United States. Not only does the U.S. have higher obesity rates than other countries but the prevalence of obesity has drastically increased in recent decades. As the prevalence maps in the title of this paper indicate, from 1990 to 2017, obesity rates in the U.S. rose an average of 175.7% across the country, with multiple southern states leading the way at more than a 220% increase. While the country saw an 11.1% obesity rate in 1990, today almost 31% of Americans find themselves significantly overweight. Body Mass Index (BMI) is the metric used to determine one's obesity status and higher BMI rates are directly correlated with mortality rate increases in a healthcare setting. Aside from mortality, higher BMI rates cause complications and lead to higher risks for chronic diseases such as heart failure, diabetes, and others.



BMI is considered Protected Health Information (PHI) in the U.S. so while it is not possible to study BMI at the community level, it is generally understood and accepted that Fast Food restaurants and menus play a contributing factor to society's BMI growth. Today's Fast Food portions are more than four times what they were in the 1950's and Fast Food chains have exploded in growth during that time, becoming one of America's favorite meal options.

By combining data sources from the US Census Bureau on community demographics with data from Foursquare location services, containing community member's top food choices, I would like to see what correlations exist. According to an article published by the National Institute of Health, higher obesity rates are seen in lower income populations. When studying socioeconomic factors such as poverty rates, income levels, and unemployment rates, a significant correlation was seen between the quality and amount of food (Fast Food) consumed in low income populations and their associated obesity rates.

Being a resident of Eastern Kentucky, I would like to study this correlation in my tri-state area of Kentucky, Ohio, and West Virginia. By separately clustering communities, first based on socioeconomic factors, and second based on food choices, I will analyze communities within a 150-mile radius around Eastern KY, looking at any correlation. The below table shows that all three of these states rank in the top 15 states in the U.S. for obesity rates. My assumption is that, by using unsupervised clustering methods, I will find a direct positive correlation in communities who choose more Fast Food options also having lower socioeconomic status when measuring income levels, unemployment, and other factors.

	2017Rank	2017PopObese	State	90to17Change	1990PopObese	1990Rank
0	1	38.10%	West Virginia	178.10%	13.70%	4.0
7	8	34.30%	Kentucky	170.10%	12.70%	9.0
10	11	33.80%	Ohio	199.10%	11.30%	17.0

B. Data Sources Used

Data Source 1: Census Data for KY, OH, and WV

My first data source comes from the U.S. Census Bureau, more specifically, the American Community Survey results, which are available through a web API. This data source will provide demographic information at the zip code level for the states being analyzed. Feature data items will include age, population, poverty level, education level, race, income levels, home values, and more. Features will be analyzed to determine which items to use for the clustering algorithms. I will be utilizing the latitude, longitude, and city name items to interact with the Foursquare API to retrieve top food choices for each community. Below is a preview of data available in a few West Virginia cities.

	ZIP	CITY	COUNTY	STATEID	STATE	LAT	LNG	TIMEZONE	MEDIAN_AGE	POPULATION	...	V
0	24701	Bluefield	Mercer	WV	West Virginia	37.30095	-81.20655	America/New_York	44.1	19621	...	1
1	24712	Athens	Mercer	WV	West Virginia	37.46458	-81.01405	America/New_York	24.7	2095	...	1
2	24714	Beeson	Mercer	WV	West Virginia	37.47671	-81.18917	America/New_York	60.3	232	...	2
3	24715	Bramwell	Mercer	WV	West Virginia	37.34319	-81.32865	America/New_York	52.8	484	...	4
4	24724	Freeman	Mercer	WV	West Virginia	37.33081	-81.29975	America/New_York	43.5	117	...	9

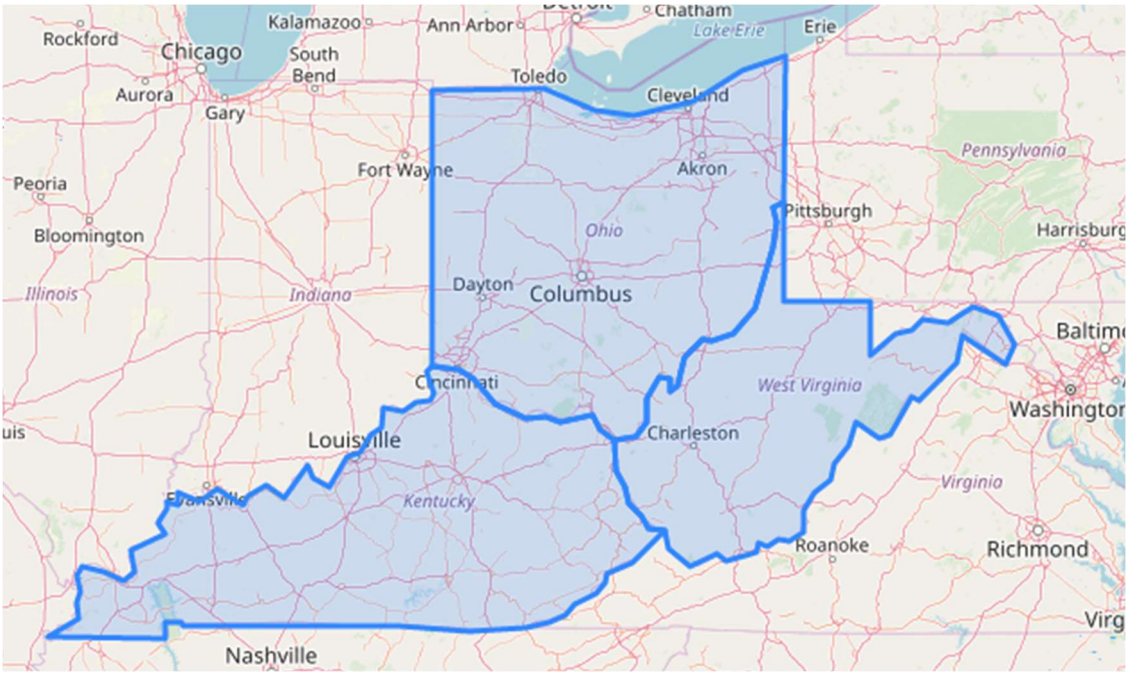
Data Source 2: Top Venues by Community - Example: Catlettsburg, KY

My second data source comes from the Foursquare locations API which offers real-time access to locations, venues, recommendations, check-ins, menus, and more. I will be utilizing the “Top Picks” in each community and looking at venue categories to analyze, not only how many food venues show up in the Top Picks, but what type of food makes an appearance as well. Below is a raw preview of the Top Picks data in the Eastern KY city of Catlettsburg.

	categories	hasPerk	id	location	name	referralId
0	[{'id': '4bf58dd8d48988d16e941735', 'name': 'F...	False	4bd1b70777b29c744a0c8d82	{'address': '3404 Court St', 'lat': 38.4055664...	Wendy's	v-1576011796
1	[{'id': '4bf58dd8d48988d1ca941735', 'name': 'P...	False	4cc09acd97bc721e27768c67	{'address': '3416 Court St', 'lat': 38.405021,...	Little Caesars Pizza	v-1576011796
2	[{'id': '52dea92d3cf9994f4e043dbb', 'name': 'D...	False	5d76ac0f6ced760008e028e6	{'address': '3500 Court St', 'lat': 38.4045, '...	Family Dollar	v-1576011796
3	[{'id': '4bf58dd8d48988d10f951735', 'name': 'P...	False	4c4e23c19932e21eadb243cd	{'address': '3501 Court St', 'lat': 38.404499,...	Rite Aid	v-1576011796
4	[{'id': '4bf58dd8d48988d118951735', 'name': 'G...	False	5568ea79498e8665411c0460	{'lat': 38.404477771725105, 'lng': -82.6011671...	IGA - Craycraft	v-1576011796

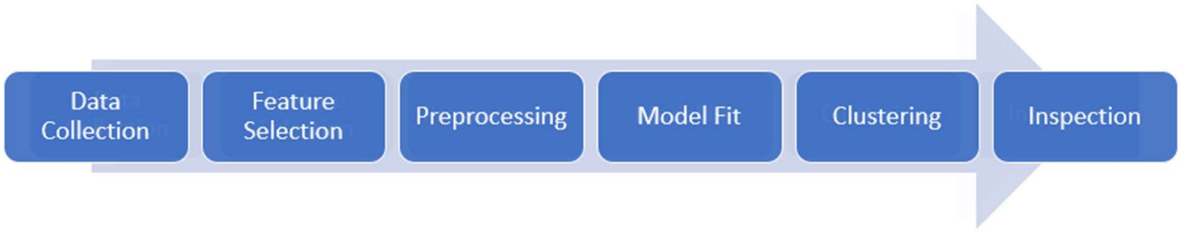
Data Source 3: Geographical Information for Display on Map

Lastly, I will be utilizing GeoJSON files by state and by zip code to display community clustering information directly onto a geographical map, allowing visual inspection of clustering results in a meaningful manner. Below is a sample map indicating that we will be working with information in the states of Kentucky, Ohio, and West Virginia.



II. Methodology

Census Demographic Data Clustering



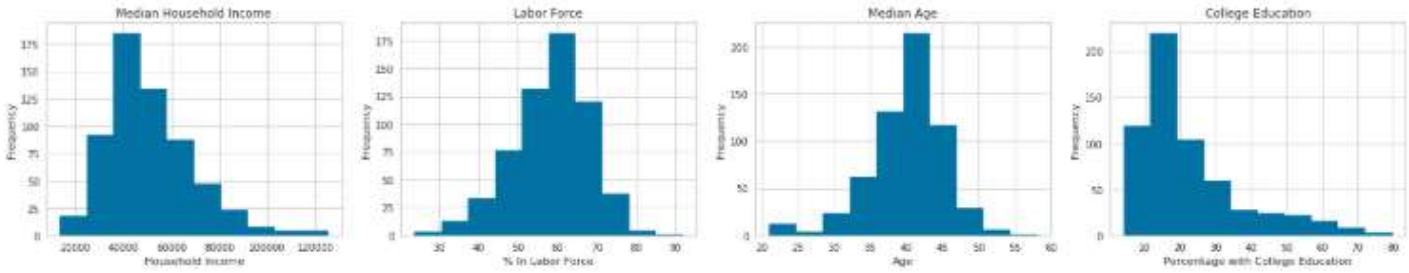
We will be utilizing the above data flow pipeline to pull in our data, understand it, preprocess and standardize it, cluster, and then analyze the results. I will explain each step of the process as we move through this pipeline with our data.

We will begin by restricting our Census data set to only those communities (zip codes) within 150 miles of the Eastern Kentucky zip code of 41101. In order to bring back meaningful results from Foursquare, we will also limit our communities to only those zip codes with more than 2,500 people in total population.

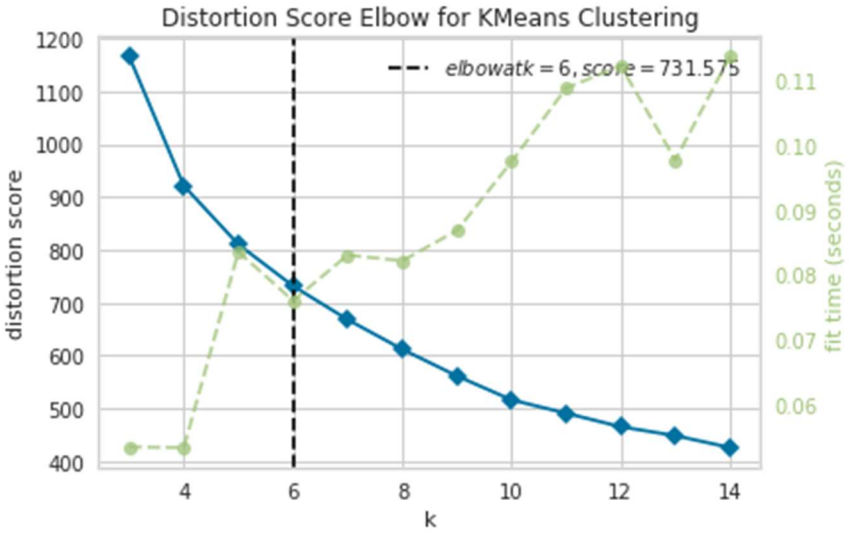
To do this, we will first create a function allowing us to pass a set of coordinates for each zip code and calculate its distance to zip 41101. The distance in miles will be returned from the function. Therefore, we can loop through our Census data set row by row, first checking to see if the total population is greater than 2,500 people. If it is, we pass the coordinates of that zip code to our newly created function and the distance to zip 41101 is returned. If the distance is less than 150 miles, we store that zip code for future use. We are creating a list of all zip within 150 miles with greater than 2,500 people. This will be the list of zip codes which we analyze their demographics and begin to cluster for similarity.

Now that we've found our list of 602 relevant zip codes, let's begin some data exploration to determine which columns to utilize for our clustering. Given that the study quoted in our Introduction focused on income metrics and unemployment, we'll utilize those features, along with education, and age. Instead of the unemployment rate, we'll be using an inverse of that metric, showing the percentage of the population in the labor force.

We'll use the 'describe' method of our data frame to look at the statistical distribution of each of our columns, ensuring no null values and also visually checking for reasonable min and max values in our dataset so we ensure we have meaningful data to use for clustering. Next, we will plot multiple histograms to look at the distribution of values within a few selected columns. Once we are comfortable with our data, we will standardize and transform the data into consistent range values so that ranges across the multiple feature columns become directly comparable, making for better fitting of the clustering algorithm.

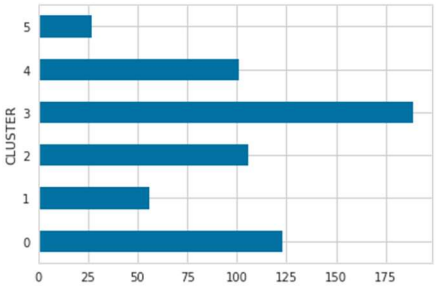


Now that our data has been cleaned and transformed, we are ready to begin the model fitting. We will utilize the Elbow Method for K Means clustering to identify what value of K gives us the best clustering of our data, achieving the lowest distortion score while giving reasonable timing on the clustering algorithm.

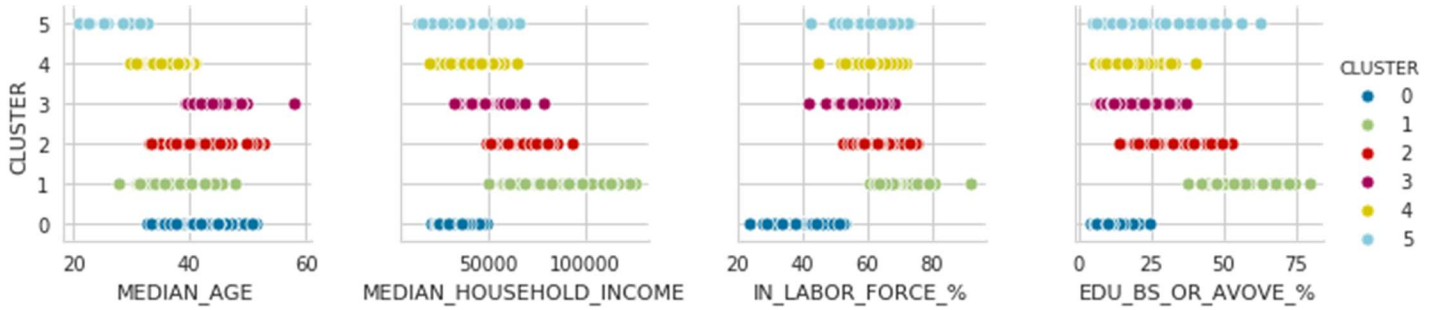


Although a clear 'elbow' is not identified here, we see 6 is returned as the optimal value of K, giving us the best balance of distortion score with time taken to fit the model. We will use this K value of 6 to now fit our K Means model and cluster our 602 zip codes into 6 individual clusters based on our selected demographics. We will then analyze the clusters to see their identifying characteristics, attempting then to name the clusters so they are useful once we combine them with our venue data from Foursquare. Below shows the results of clustering with the table on the left showing the average mean feature characteristics of each cluster, and the chart on the right showing how many zip codes fall into each cluster category.

	MEDIAN_AGE	MEDIAN_HOUSEHOLD_INCOME	IN_LABOR_FORCE_%	EDU_BS_OR_AVOVE_%
CLUSTER				
0	42.065854	34005.967480	44.665854	11.996748
1	37.860714	84606.571429	71.235714	56.862500
2	40.625472	66521.773585	66.051887	30.488679
3	43.161905	48814.539683	57.378307	16.648148
4	36.255446	42962.336634	61.782178	16.793069
5	26.477778	36458.296296	60.507407	29.396296



Now that we have the communities clustered, let's see how many communities fall into each cluster. We will turn the table above into a visual representation of the feature characteristics. Our goal is to name the individual clusters with more meaningful names instead of cluster numbers. Later, we will visualize these on a map for even better interpretation.



We've now clustered our community zip codes into 6 distinct groups, with defining characteristics as seen above. Let's name each cluster and add the cluster name back to our data frame so we have a more meaningful way of looking at our community zip codes. Below are named clusters in descending order based on the number of zip codes in each.

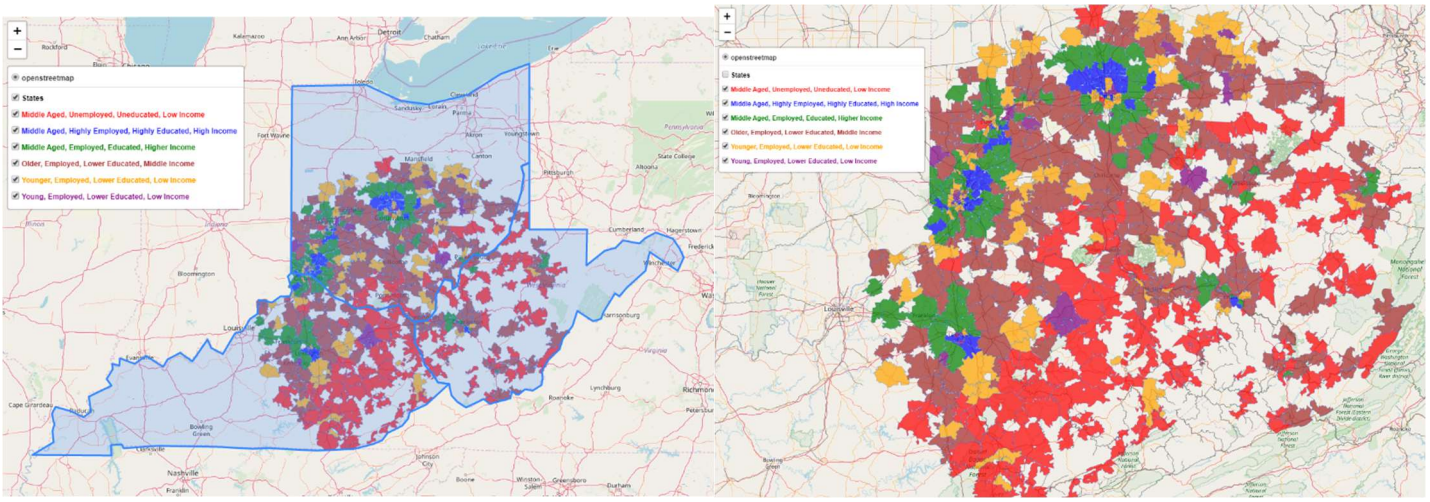
- Cluster 3: Older, Employed, Lower Educated, Middle Income
- Cluster 0: Middle Aged, Unemployed, Uneducated, Low Income
- Cluster 2: Middle Aged, Employed, Educated, Higher Income
- Cluster 4: Younger, Employed, Lower Educated, Low Income
- Cluster 1: Middle Aged, Highly Employed, Highly Educated, High Income
- Cluster 5: Young, Employed, Lower Educated, Low Income

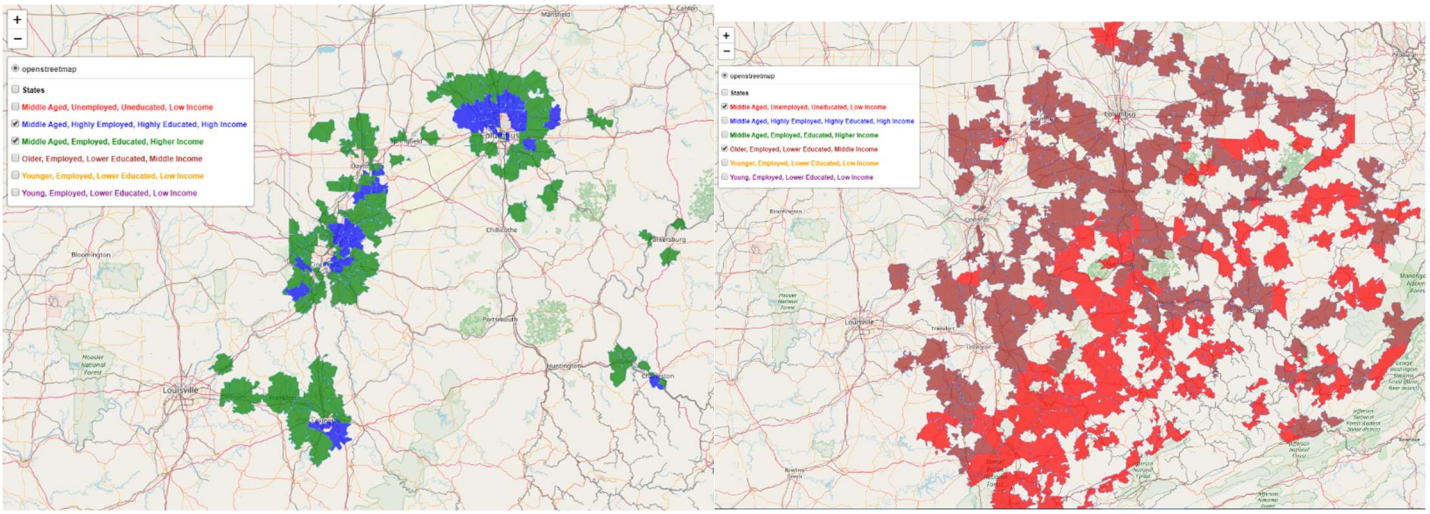
Now that we have our cluster names, we assign them as a column in our census data frame.

Visualize Clustered Zip Codes

With our communities clustered and named, we can now visualize them on our Folium map previewed in our Data Sources discussion above. We will segregate the six clusters and assign GeoJson dictionary files for each, allowing us to add each of them as separate groups to our Folium map and assess the results.

As you will see from the below series of pictures, our clustering offers interesting results and is validated by my knowledge of this tristate area. Most of the area is an economically depressed region with below national average income levels and education levels. This is validated by the two largest clusters belonging to lower-middle-class populations. The upper-middle- and upper-class clusters are centralized around the largest city centers of Columbus, Dayton, Cincinnati, Lexington, and a couple of others, further adding reasonableness and validation to our results.





Utilize Foursquare API to get Top Venues by Community

Now that we've fully clustered, analyzed, and visualized our zip code communities by socioeconomic status, it's time to move on to the Foursquare API which will allow us to look at the top commercial venues in each community. From that data, we will look at the top venue choices by category and then cluster our communities a second time based on this venue information.

We will start by pulling in all the raw data and cleaning it up into a data frame for easier use and manipulation. For each zip code community in our census data, we will loop through and make a call to the Foursquare API to return the top 10 "Top Picks" from each community. We will then look at the venue category types, the frequency with which they appear, and begin to cluster the communities on similar venue category preferences.

Let's now perform "one hot" encoding on the data set, converting the unique categories to column names and passing 1's and 0's to each column based on whether a zip code has that category in their Top Pick list or not. A '1' indicates that the category appears in the community results. A '0' indicates its absence. By converting to numerical values, we can utilize statistical methods to rank the categories by each community.

With this ranking ability in place, we will create a data frame by city that lists the top 10 categories per city.

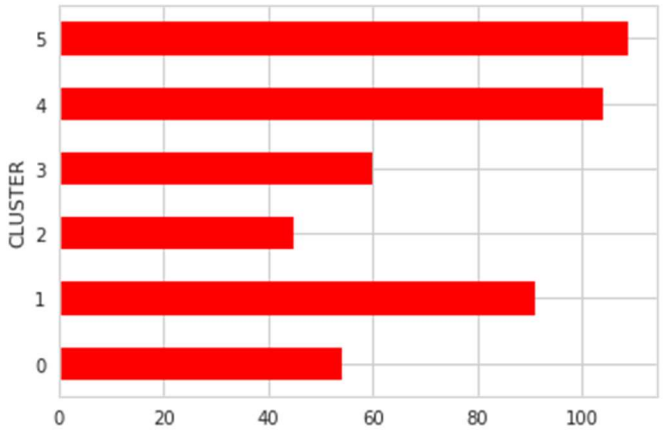
	CITY	STATEID	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Mos Commo Venue
328	Pleasureville	KY	Auto Dealership	City Hall	Sandwich Place	Gas Station	Discount Store	General Entertainment	Farm	Element School
36	Bethesda	OH	Gas Station	Spiritual Center	Financial or Legal Service	Motorcycle Shop	Fast Food Restaurant	Park	Farm	High School
84	Cleves	OH	Flower Shop	Sporting Goods Shop	Library	Gas Station	Bar	Bank	Funeral Home	Deli / Bodega
2	Alexandria	KY	Automotive Shop	Church	Rental Service	Rental Car Location	Hardware Store	School	Doctor\'s Office	Donut Shop
382	South Shore	KY	Financial or Legal Service	Hardware Store	Medical Center	Grocery Store	Train Station	Mobile Phone Shop	Gas Station	Fast Foo Restaura

Cluster Communities by Venue Choices

To begin clustering, we'll use our grouped data frame showing the mean value of each venue category by zip. We'll drop the columns for zip, city, and state so that we are left only with numerical columns for each venue category type. This time, we will skip the elbow method and use 6 as our K to match our analysis above with the census clustering.

Once we have the data clustered, we will add all the data frames back together so that we have a total picture of all demographics, venue choices, and both cluster groupings all in one data frame. From there, we will do some investigation to analyze our results.

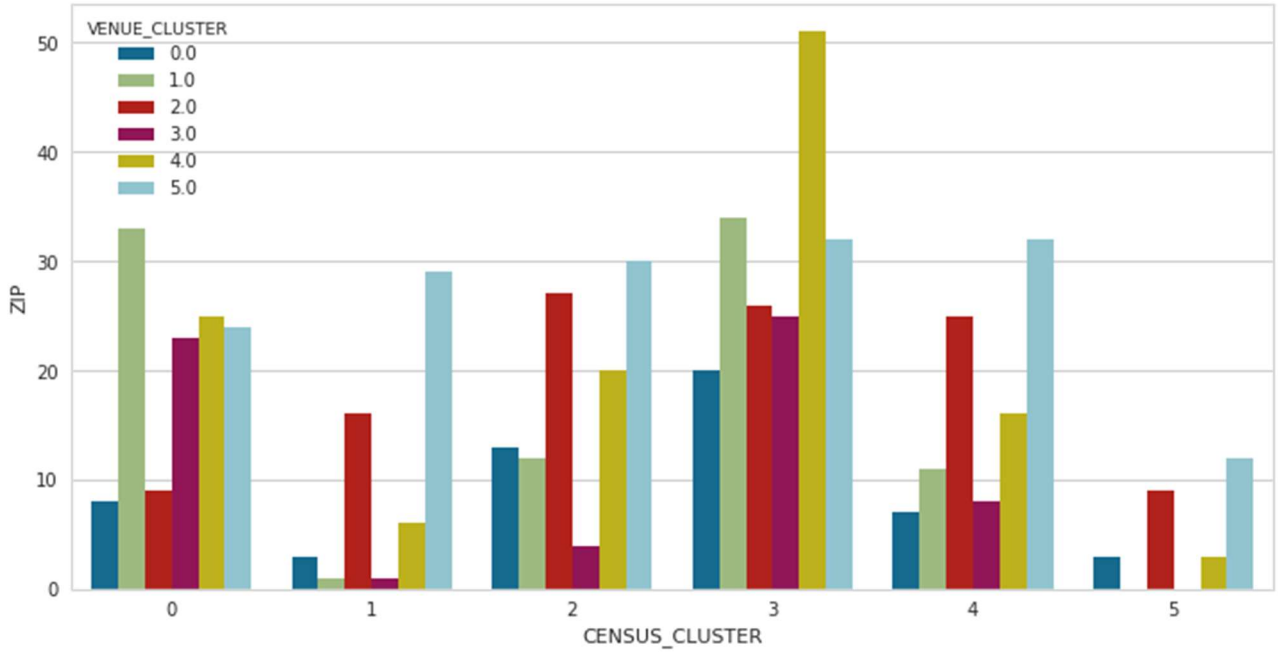
The below bar chart shows the number of zip code communities which fall into each of our venue clusters. Immediately below is a table showing a sample of communities. With a total data frame now in place we see a few zip codes with both their census cluster as well as their venue cluster. We will look at this on a broader scale to analyze correlation and test our hypothesis.



	ZIP	CITY	STATEID	CENSUS_CLUSTER	CLUS_NAME	VENUE_CLUSTER
409	45040	Mason	OH	1	Middle Aged, Highly Employed, Highly Educated,...	5.0
149	40456	Mount Vernon	KY	0	Middle Aged, Unemployed, Uneducated, Low Income	1.0
299	43102	Amanda	OH	3	Older, Employed, Lower Educated, Middle Income	1.0
577	45662	Portsmouth	OH	0	Middle Aged, Unemployed, Uneducated, Low Income	5.0
495	45305	Bellbrook	OH	1	Middle Aged, Highly Employed, Highly Educated,...	4.0

III. Results

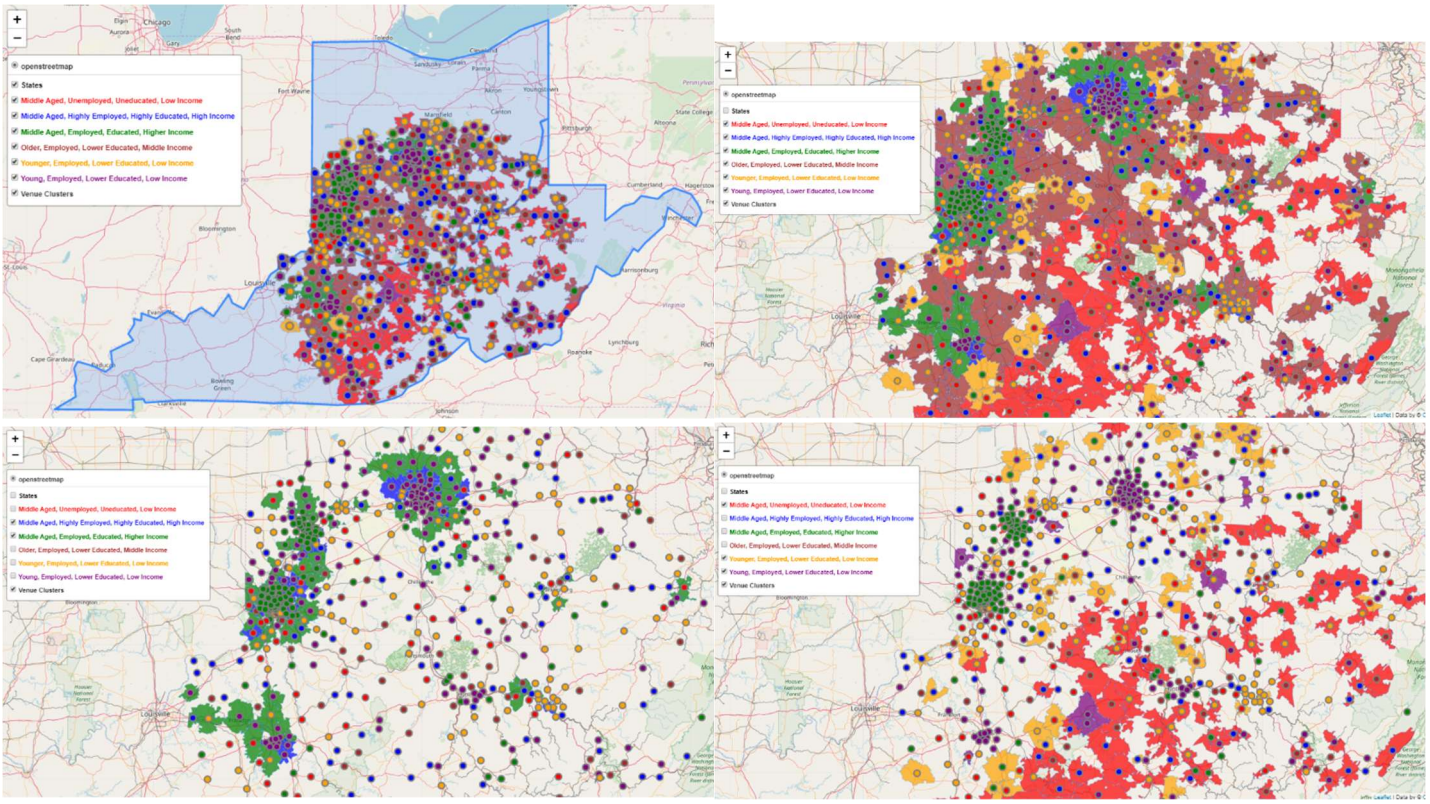
It is now time to analyze our results. We'll begin by visualizing how each community falls into both the census data clustering as well as the venue choice clustering. We'll look for any correlation and we will also calculate correlation statistics. Lastly, we will add the venue clustering information to our folium map from before so that we have all analytics on a single visual map to display.



The below table shows the original census cluster names, helping us better interpret the chart above.

		ZIP
CENSUS_CLUSTER	CLUS_NAME	
0	Middle Aged, Unemployed, Uneducated, Low Income	123
1	Middle Aged, Highly Employed, Highly Educated, High Income	56
2	Middle Aged, Employed, Educated, Higher Income	106
3	Older, Employed, Lower Educated, Middle Income	189
4	Younger, Employed, Lower Educated, Low Income	101
5	Young, Employed, Lower Educated, Low Income	27

With a calculated correlation coefficient between the two cluster methodologies of 0.02, we do not see a strong relationship between our two clustering methodologies. The map samples below update our visualizations from above, now with colored circular markings indicating the six different venue categories for each zip code. Visually, it is easy to see that we do not get marker color clustering onto the same zip code cluster colors.



IV. Discussion

We've now come through the full process of pulling multiple data sources, clustering multiple ways, and visualizing results to look for correlation and draw any conclusions. We utilized multiple tools in the data scientist tool belt to load and clean data, transform into usable format, visualize for key features, cluster using unsupervised learning techniques, and then again visualize for results and conclusions. This process is invaluable to any data scientist looking to analyze real-world problems, analyze results, and implement solutions.

Specifically, our project focused on utilizing census demographic information centered around four key metrics: median household income, education level, percentage of population in the workforce, and median age. We looked at these metrics for each zip code within a 150-mile radius from Ashland, KY and only zip codes with at least a 2,500

total population. By using K Means clustering to cluster these communities on the four key metrics, we saw interesting and expected results, where higher educated, higher employed, and higher income earners clustered around major populated cities. In our case, these cities included Columbus, Cincinnati, Dayton, Lexington, and to a lesser extent, Charleston. Our visualization of these results on a folium web map made interpretation much easier.

We next utilized the Foursquare API to pull in the top 10 venue picks of each city. By analyzing categories such as fast food, yoga studios, museums, and other places of interest, we again clustered our communities based on similar interests. We mirrored the same clustering algorithm used for our census data, employing K Means clustering with 6 target groups. While our clustering algorithm performed its function, our results were not as easily interpretable as our demographic clustering. By visualizing the venue clusters against our demographic clusters, we were able to analyze for our initial hypothesis - do lower socioeconomic demographic communities have a higher appetite for fast food type venues.

V. Conclusion

Unfortunately, our results were not clear enough for us to draw finite conclusions. As we want to be careful about drawing causation from correlation, we will simply state that we were unable to prove our hypothesis. This does not mean there is no real-world correlation, but it does mean that our model did not have the right feature selection in order to adequately portray this. As a data scientist, this is unfortunately a common place to find ourselves. We utilized our tools correctly, had good data sources, and walked through the process efficiently. That does not always guarantee, however, that we will get the expected results, or even results that we will be able to do anything with or draw conclusions from. The important thing is learning how to use the tools at our disposal and then how to go back and try multiple iterations, different feature selection, different algorithms, etc. until we find a model that works for the problem at hand. While this was an interesting topic and we utilized our resources effectively, no correlation was shown to prove our hypothesis.

As a next iteration, it may be interesting to restrict our Foursquare API call to only "fast food" category types and see how many venues are returned from each community. We could then look for correlation in the number of fast food restaurants with the socioeconomic status of the community. With so many variables at play in real world scenarios, data scientists must demonstrate patience and analytical skills to stay focused and think through the most efficient and effective model setup.

VI. References

1. ProCon: US Obesity Levels by State
2. FourSquare API
3. Census.gov API
4. OpenDataDE Zip Code Level GeoJSON