

# COMPARING COMMUNITY DEMOGRAPHICS TO FAST FOOD APPETITES

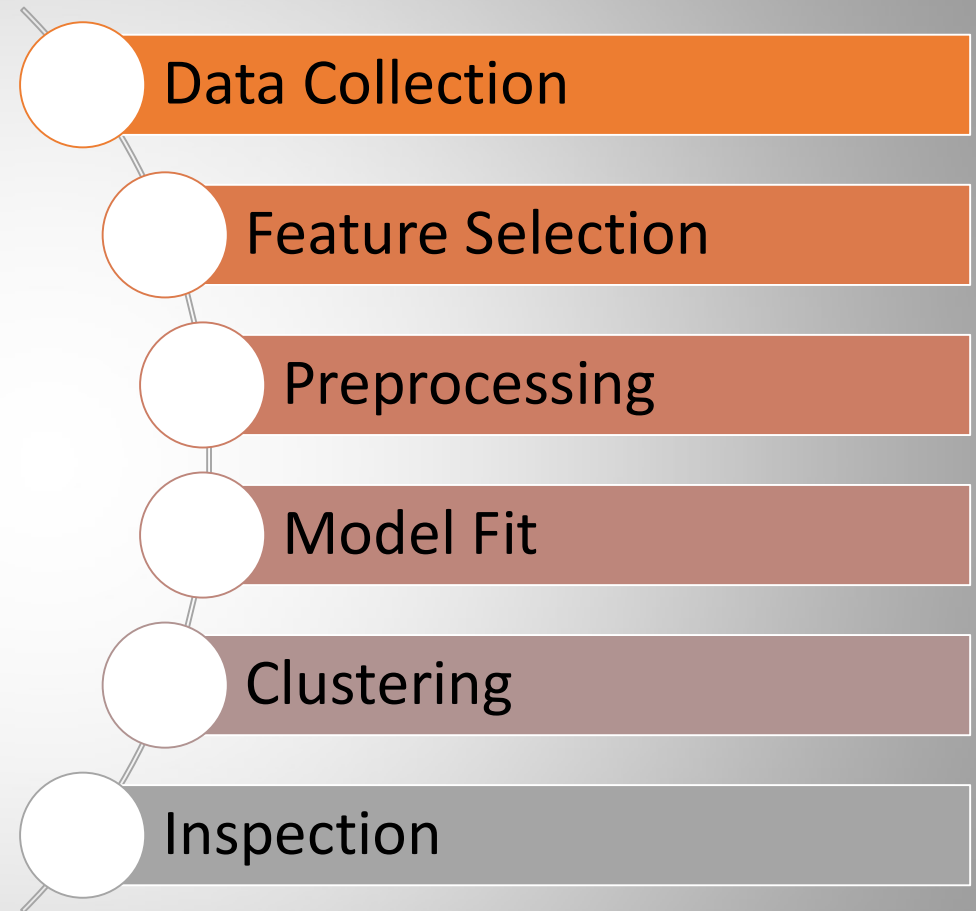
Coursera Data Science Capstone Project  
Richard Queen  
December 21, 2019

# INTRODUCTION

- BMI has become a major health issue in the United States
  - Associated with exponentially higher mortality rates
  - Creates higher risk for multiple chronic disease conditions like heart failure and diabetes
- From 1990 to 2017, BMI rose almost 176% in the US, with more than 30% of adults now being obese
- Fast Food meal portions are 4 times larger than they were in 1950s
- This study will look at correlation between Fast Food choices of a community and that community's socioeconomic demographic

# DATA METHOD

We will be utilizing this 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.



# DATA COLLECTION SOURCES

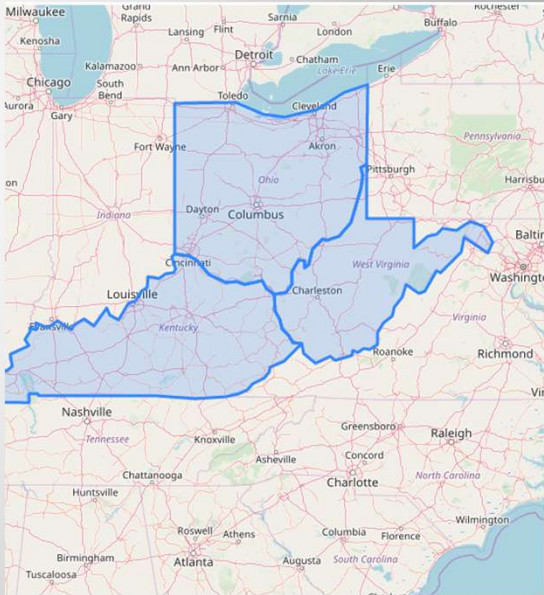
US CENSUS  
DEMOGRAPHICS BY ZIP  
CODE

STATE	LAT	LNG	TIMEZONE
West Virginia	37.30095	-81.20655	America/New
West Virginia	37.46458	-81.01405	America/New
West Virginia	37.47671	-81.18917	America/New
West Virginia	37.34319	-81.32865	America/New
West Virginia	37.33081	-81.29975	America/New

FOURSQUARE API “TOP  
PICK” VENUES

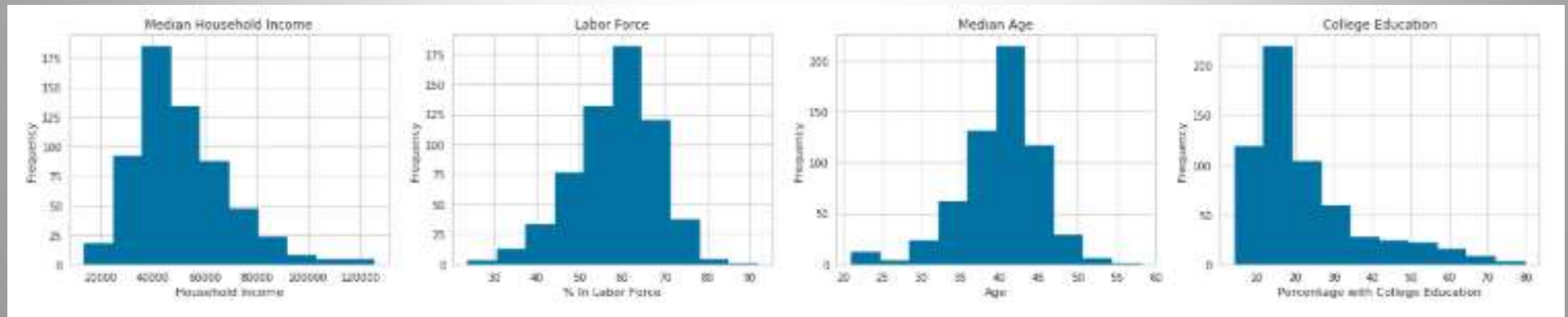
hasPerk	id	location
False	4bd1b70777b29c744a0c8d82	{‘address’: ‘341 St’, ‘lat’: 38.40
False	4cc09acd97bc721e27768c67	{‘address’: ‘34 St’, ‘lat’: 38.40
False	5d76ac0f6ced760008e028e6	{‘address’: ‘351 St’, ‘lat’: 38.40
False	4c4e23c19932e21eadb243cd	{‘address’: ‘351 St’, ‘lat’: 38.40
False	5568ea79498e8665411c0460	{‘lat’: 38.404477771 ‘lng’: -82.6011

GEOJSON MAP LAYOUT  
FILES



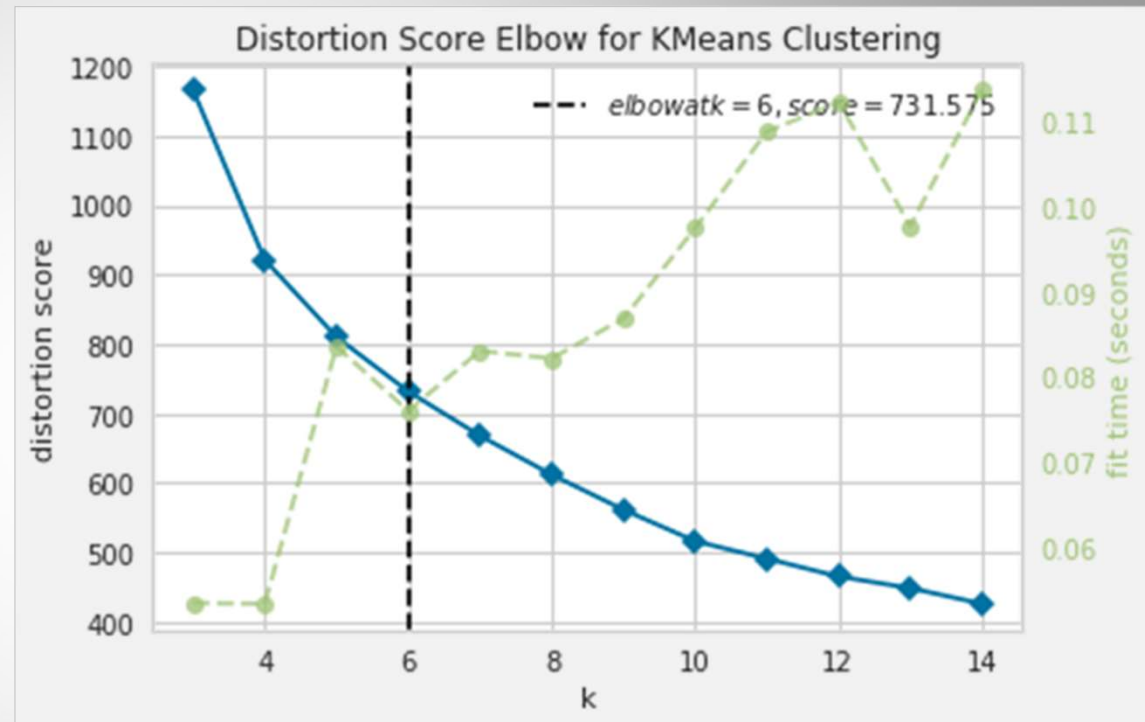
# FEATURE SELECTION

- We first select features to cluster zip code communities on socioeconomic status
- We focus on key metrics of median household income, percentage in labor force, median age, and education level



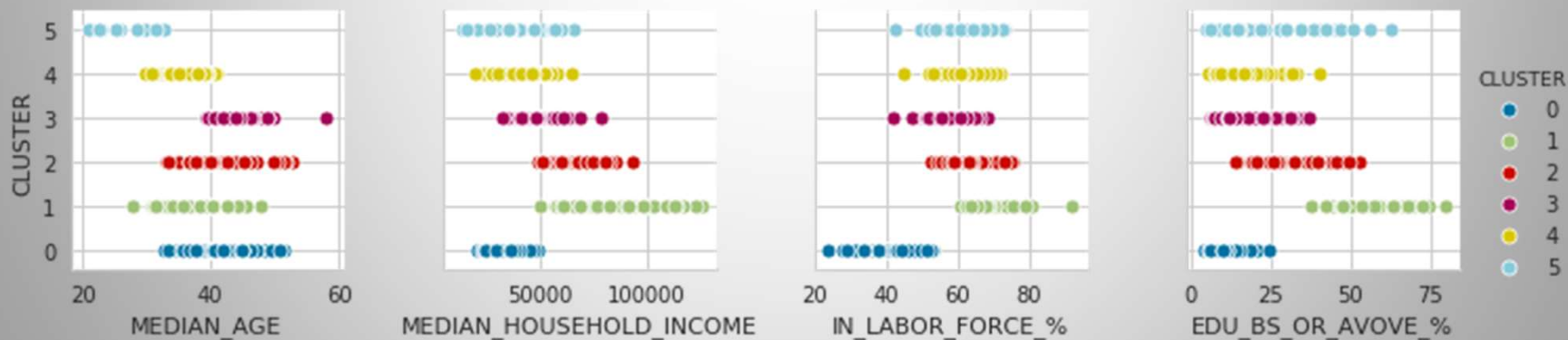
# ELBOW METHOD

The elbow method measures the distortion score at various values of K in K Means clustering. Our analysis tells us that 6 is the optimal level of K. This means we will cluster our community zip codes into 6 distinct groups or clusters.

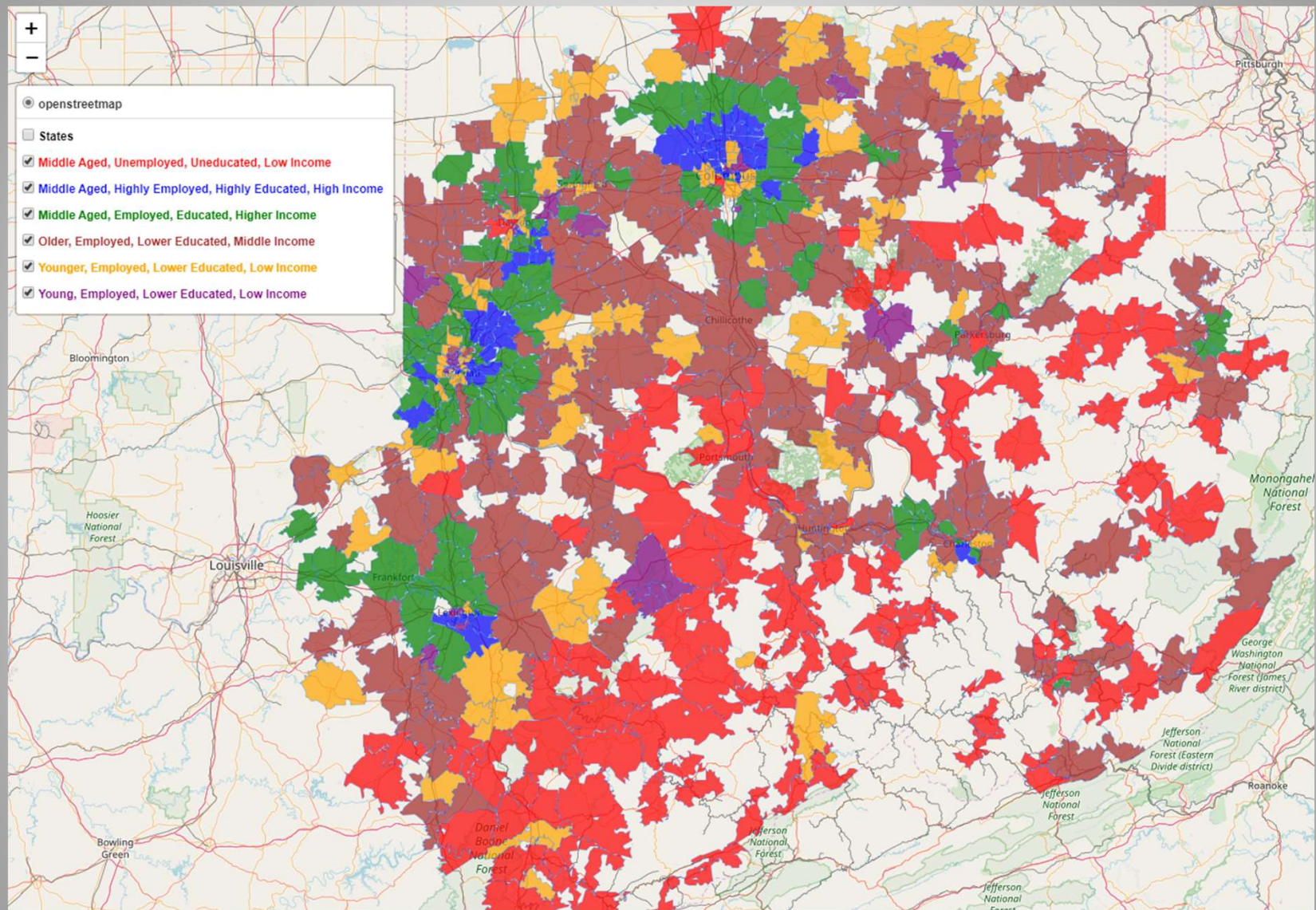


# CLUSTER CHARACTERISTICS

- 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









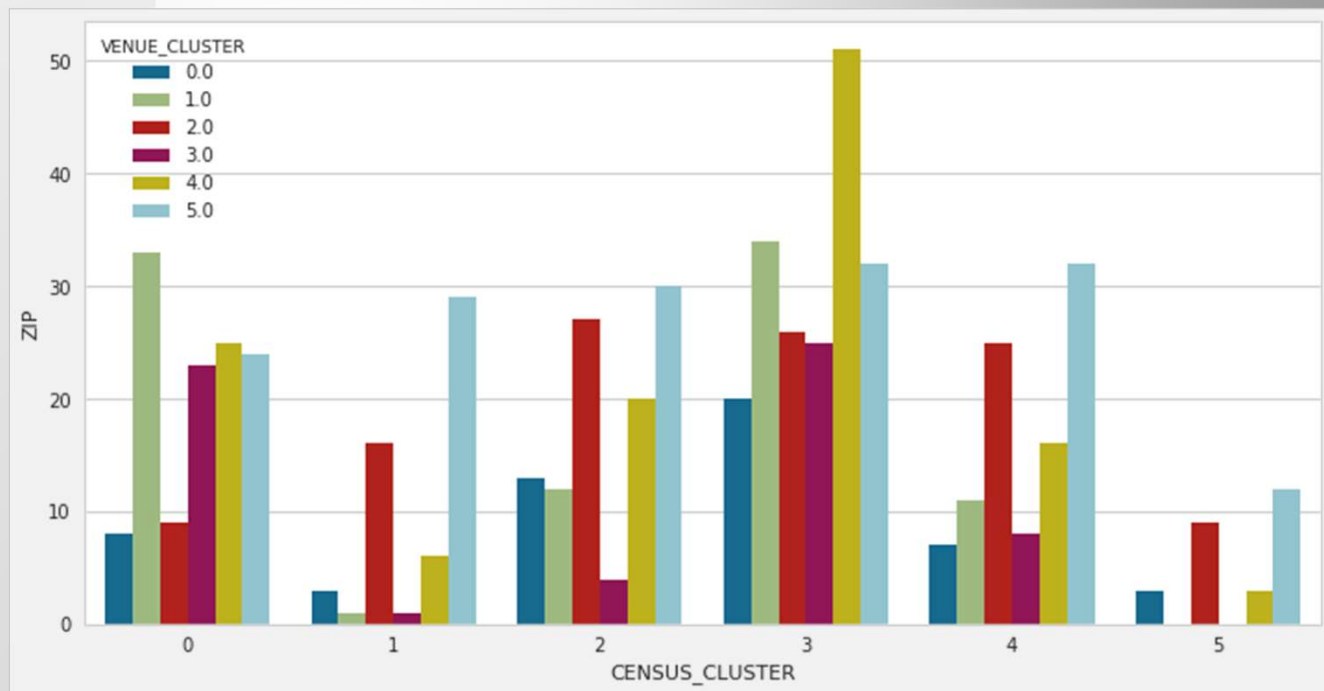
# FOURSQUARE TOP VENUES PER CITY

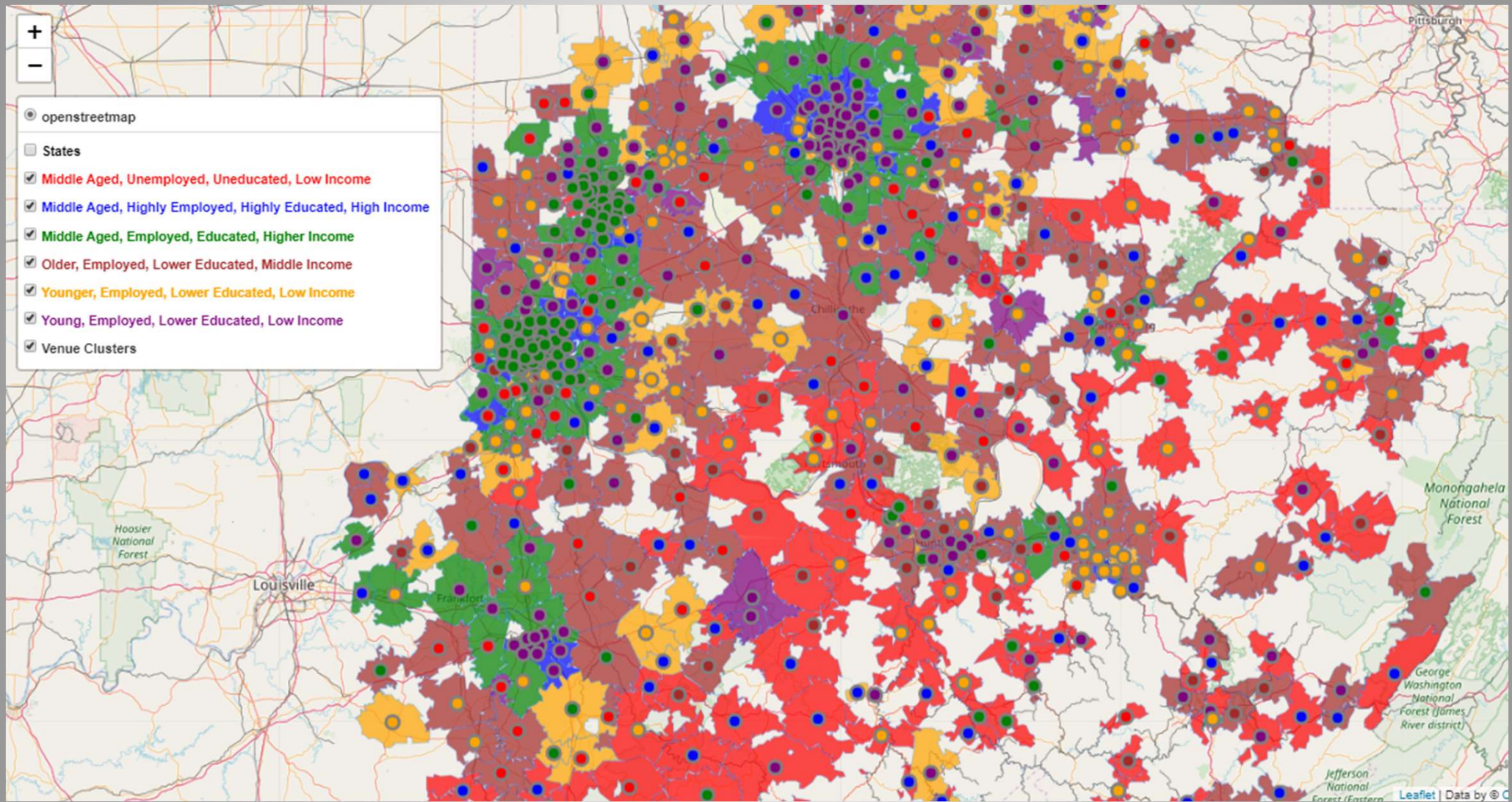
- We are able to pull top 10 venues per city in order to cluster our communities a second way and then look for correlation

	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 Most Common Venue
328	Pleasureville	KY	Auto Dealership	City Hall	Sandwich Place	Gas Station	Discount Store	General Entertainment	Farm	Elementary 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 Food Restaurant

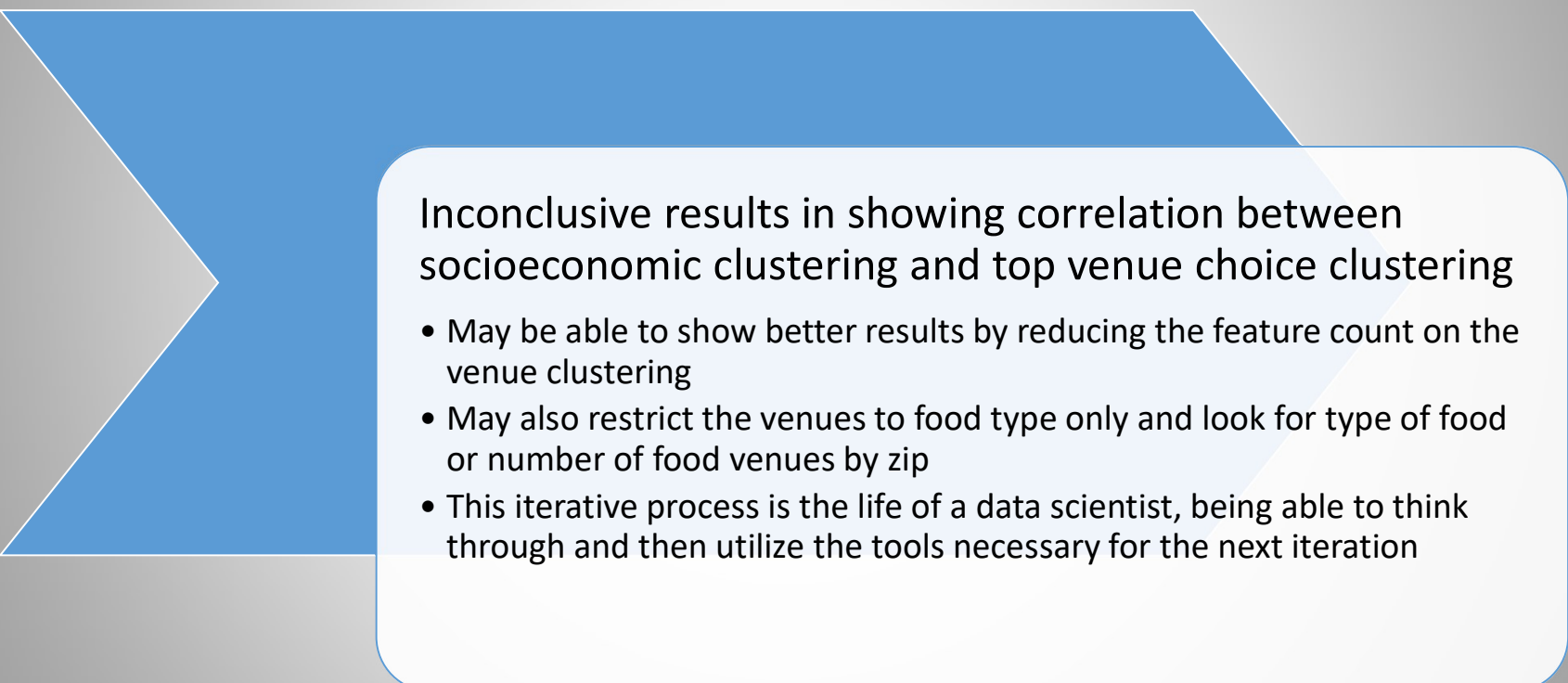
# RESULTS

Unfortunately, in this data pull from the Foursquare API, we do not see a high level of correlation between our demographic clustering and our top venue clustering.





# CONCLUSION & NEXT STEPS



Inconclusive results in showing correlation between socioeconomic clustering and top venue choice clustering

- May be able to show better results by reducing the feature count on the venue clustering
- May also restrict the venues to food type only and look for type of food or number of food venues by zip
- This iterative process is the life of a data scientist, being able to think through and then utilize the tools necessary for the next iteration



## REFERENCES



[ProCon: US Obesity Levels by State](#)

[FourSquare API](#)

[Census.gov API](#)

[OpenDataDE Zip Code Level GeoJSON](#)