Introduction: IRS Data and Foursquare Amenities

Eric Roberts

For the final capstone project of the Applied Data Science unit, it's required to use the Foursquare API and some other data source to analyze a problem.

Using the FourSquare API, I'm interested in amenities in rich and poor neighborhoods in the United States and to see if there is a measure of commonality or divergence and if so, how this commonality or diverence can be expressed.

Data used to solve the problem

It's necessary pull in some data about the wealth of the US in order for this analysis. The Internal Revenue Service (IRS) publishes some data about the tax returns based on ZIP codes.

This information is located here at the IRS - Individual Income Tax Statistics

The average Adjusted Gross Income (AGI) of the taxpayers within a ZIP code will be used as a metric of wealth.

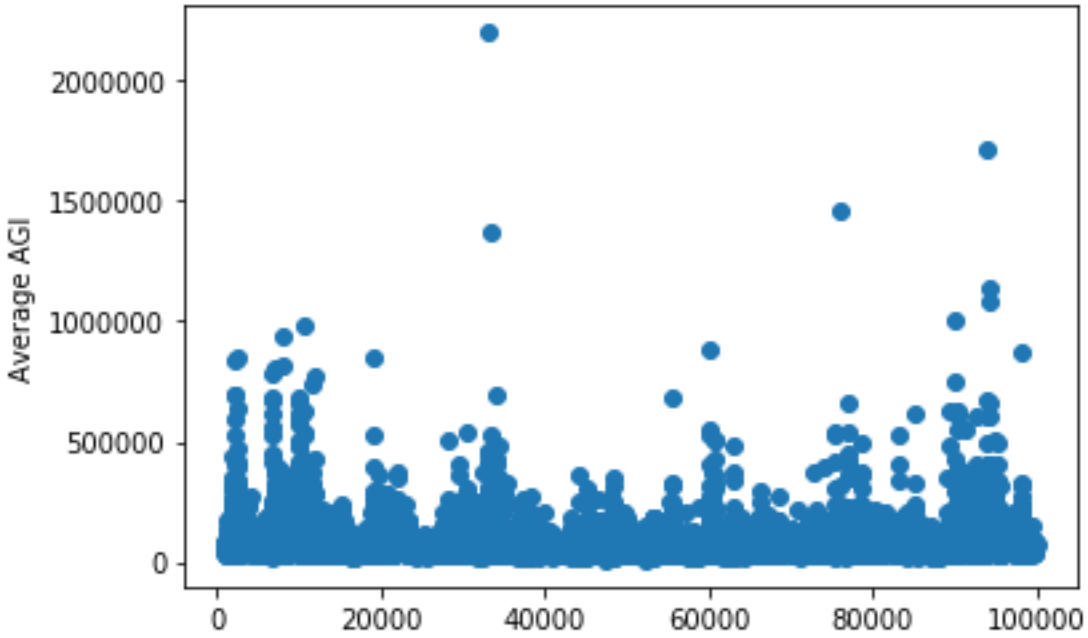
ZIP Codes are not geographical, so it'll be necessary to lookup each ZIP code and get Geocode pairs. I will use the Geocoders API or a lookup via a table like this (OpenData)

The FourSquare API can be used to lookup amentities given a latitude, longitude and tagged, for example input is Lexington, MA, 42120 (42.452895, -71.21619)

FourSquare will report the number and type of amentities given this data.

Incomes across USA

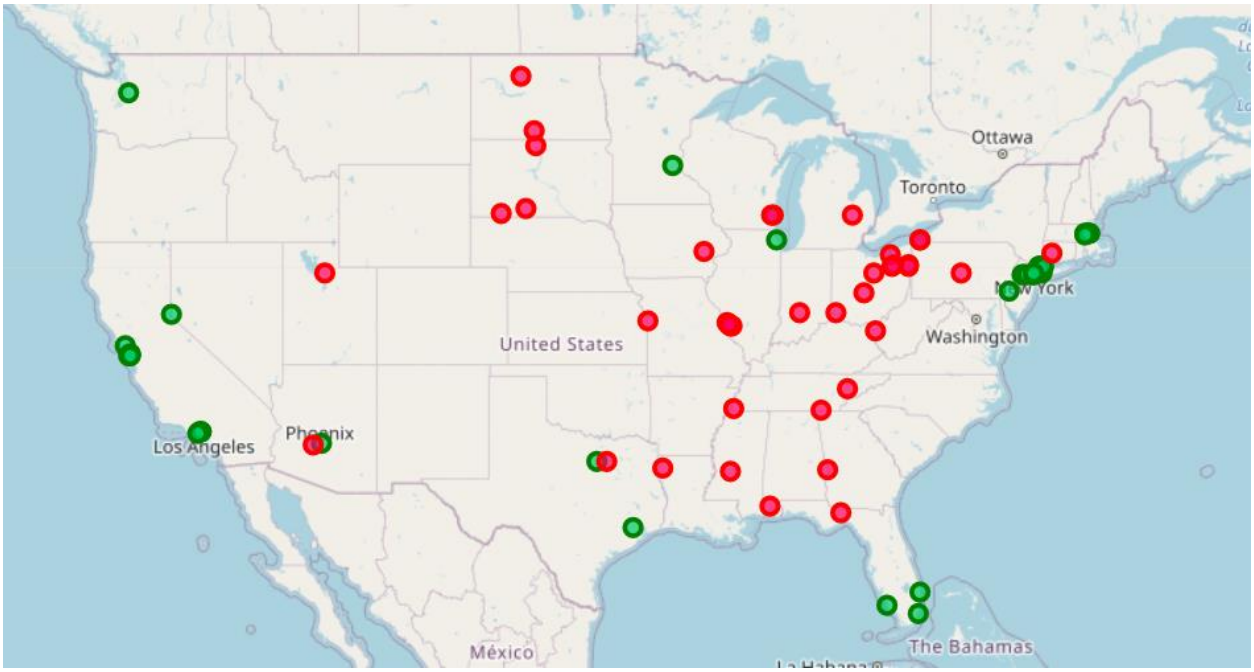
A weighted average of AGI (Adjusted Gross Income) reported by taxpayers in each zipcode can be computed. There's an impressive outlier in zipcode 33109, which is Palm Beach, Florida.



The minimum income zipcode is Bloomington, Indiana, but I suspect that this is because it covers a college area and true "wealth" is not represented.

Interactive Map

A map of the top 40 income “rich” ZIP Codes and lowest 40 income “poor” ZIP codes can be created:



Querying Foursquare

The OpenData data can be used to find a center of longitude and latitude for each zip code. Then Foursquare can be queried.

zipcode	City	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
tag							
poor	360	360	360	360	360	360	360	360
rich	1134	1134	1134	1134	1134	1134	1134	1134

As we might expect, there's a lot more amentities (at least covered by Foursquare) in rich areas than poor ones.

The frequency of each amenity was calculated.

	Poor Areas. Lowest 40 income Areas			Rich Areas. Highest 40 income Areas		
	venue	freq	common	venue	freq	common
0	Fast Food Restaurant	0.05	0	Coffee Shop	0.05	0.05
1	Pizza Place	0.05	0	Italian Restaurant	0.04	0
2	Coffee Shop	0.04	0.04	Hotel	0.04	0
3	Bar	0.04	0	American Restaurant	0.03	0.03
4	Sandwich Place	0.03	0.03	Park	0.03	0
5	American Restaurant	0.03	0.03	Steakhouse	0.02	0
6	Pharmacy	0.02	0	Bakery	0.02	0
7	Café	0.02	0.02	Restaurant	0.02	0.02
8	Convenience Store	0.02	0	Burger Joint	0.02	0
9	Diner	0.02	0	Café	0.02	0.02
10	Discount Store	0.02	0	Sushi Restaurant	0.02	0
11	Gas Station	0.02	0	Seafood Restaurant	0.02	0
12	Hookah Bar	0.01	0	Sandwich Place	0.02	0.02
13	Middle Eastern Restaurant	0.01	0	Dessert Shop	0.01	0
14	Restaurant	0.01	0.01	Deli / Bodega	0.01	0
	total	0.39	0.13		0.37	0.14

There's commonality between about 13% of the amenities across rich and poor groups.

Conclusions

- We analyzed the top 40 richest zipcodes with AGIs ranging from ($2.1M$,2.1M,182K) and the lowest 40 poorest zipcodes with AGIs ranging from ($8k$,8k,22k).
- Rich zipcodes have about 3 times the amount of amentities according to Foursquare
- There is commonality between some amentities in rich and poor neighborhoods, but most are different.
    - There's commonality between 13% of amentities
    - Rich and poor Americans alike like coffee, sandwiches, and American restaurants.
    - If we consider pizza to be Italian food, there's a like of Italian food across income segments.
    - Further analysis using this data could be used at the middle - segmenting "K" areas of the country with similar income and amentity availability.

References

All code can be found on Github.

https://github.com/projekt888/Coursera_Capstone/tree/master/Final