Types of Venues Most Prevalent within Socioeconomic Factors

Jakub Polanowski

June 2, 2020

1. Introduction

Socioeconomic features, per Capita Income, Hardship Index, and so forth, are important measurements of the status of neighborhoods within a city. Looking at Chicago in particular, form the point of view of a City Government Official, there is the question of how those socioeconomic features affect the prominence of different types of venues within clusters of neighborhoods based on the socioeconomic features of those neighborhoods. The question specifically being, will there be specific types of venues that are present in only particular clusters, will there be specific types of venues present in all/most clusters or will there be no indication of any relationship?

2. Data

2.1 City of Chicago Dataset Cleaning

The city of Chicago dataset, containing useful socioeconomic data, also contained irrelevant data. The irrelevant data, ‘Community Area Number’ and ‘Percent Aged Under 18 or over 64’ was drop from the dataset, since these were not considered to be relevant to classifying the neighborhoods within socioeconomic clusters. Any neighborhoods that were missing values for any of the renaming feature categories were dropped since missing values could skew the cluster.

2.2 Geolocation of Each Neighborhood

For the proposes of using the FourSquare API and mapping the clusters, the geolocation (longitude and latitude) was required. This was performed using the geocoder package using the arcgis resolver, in which the name of the neighborhood was added to a string in the format of ‘<neighborhood>, Chicago, Illonios’ as the search query for the resolver. The resulting longitudes and latitudes were then added to the dataset table.

2.3 Venue Lookup via FourSquare API

Nearby venues were found by sending queries with the longitude and latitude of the neighborhoods, a radius of 500 ft was used, which would approximately cover each neighborhood. The only feature of interest from the results was the ‘Venue Category’ itself, all other data was dropped.

3. Methodology

3.1 Socioeconomic Data Normalization

Socioeconomic data was standardized within the standardscalar function form sklearn.preprocessing. The goal of performing this was to convert each socioeconomic feature into a standardized scalar that could be used to properly cluster the neighborhoods, without performing this normalization, features with larger magnitudes would have larger influence on the clustering results.

3.2 Calculating the Clusters

Clusters were calculated using the K nearest neighbors algorithm implemented in the KMeans function from sklearn. The number of clusters to generate was set to 5, the rational for this being that it would provide a decent range of clusters cluster with varying social economic factors, a smaller number of clusters would result in neighborhoods of significantly different socioeconomic factors being lumped together whereas a higher number of clusters would provide more but less distinct clusters – which would ultimately fail to address the questions set out in the analysis.

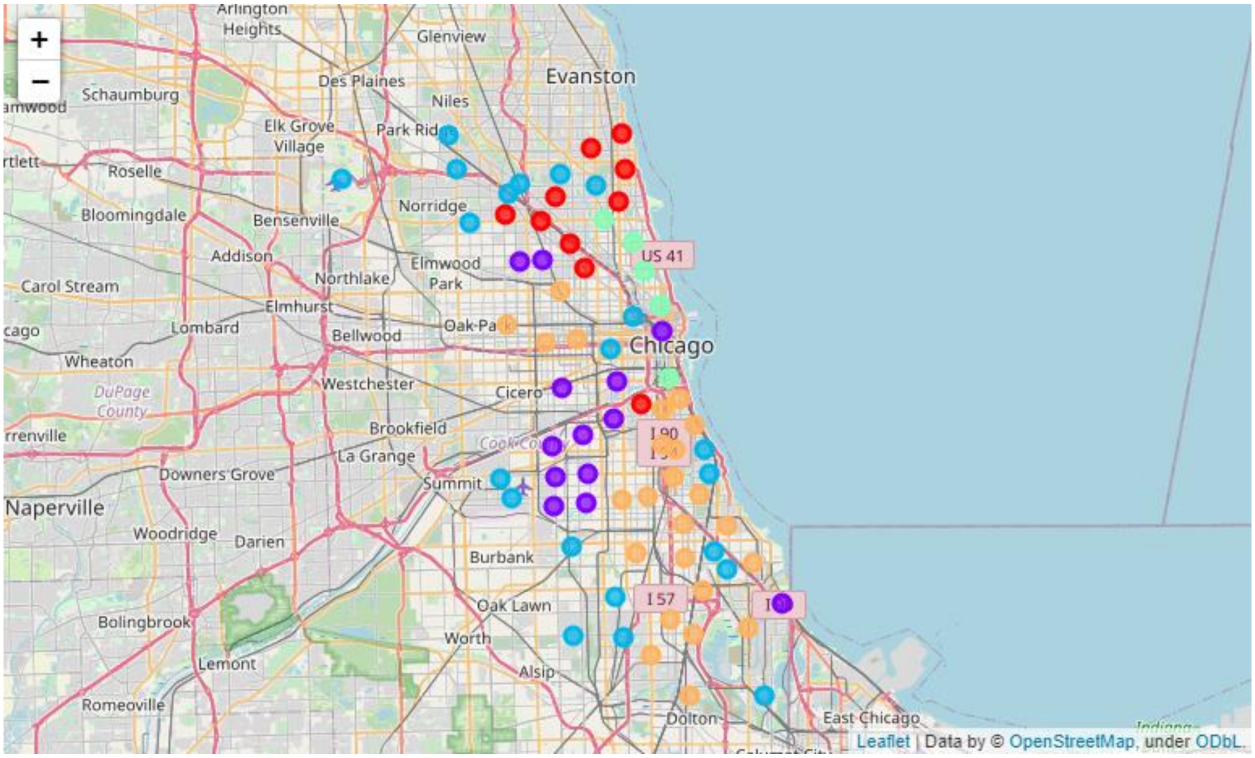
3.3 Determination of Top 5 Venue Types

Once the data was fitted, the cluster identifier number was merged with the venue dataset to identify which cluster the venue belonged to in each neighborhood. Onehot encoding was then performed, then the onehot encoded data was grouped by the cluster identifier, with onehot counts been computed as averages.

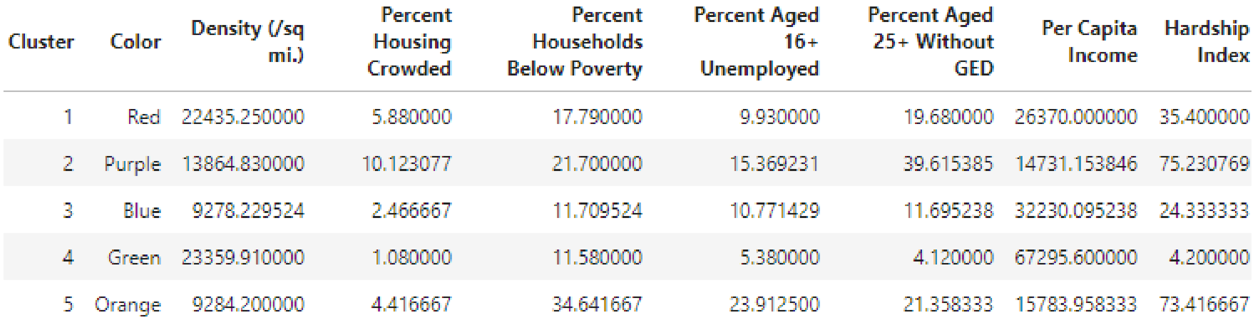
For each cluster, the top 5 most common (based on highest average) were calculated and then added to the socioeconomic data and cluster identifiers.

4. Results

4.1 Mapping of Socioeconomic Clusters



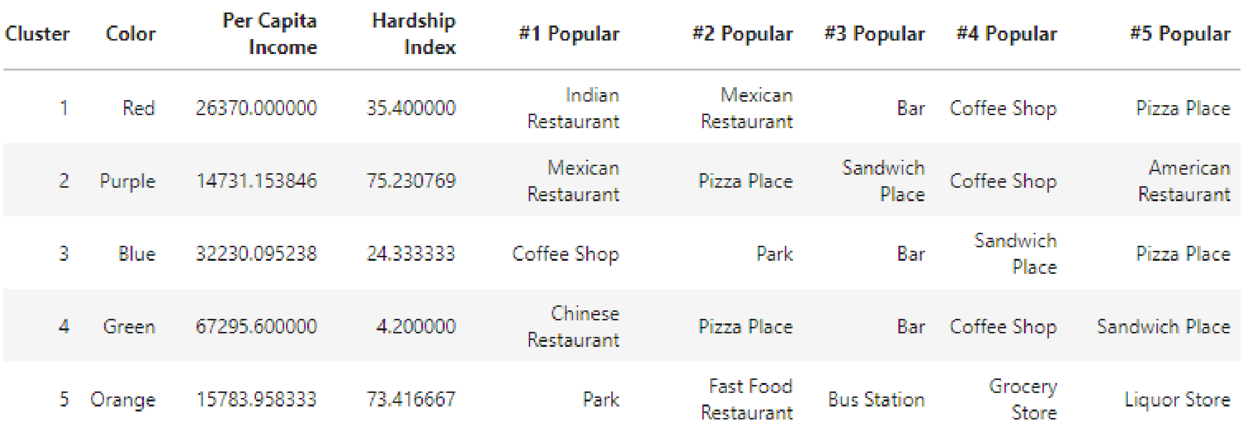
Shown above is the mapping of each neighborhood within Chicago, colored by cluster, with the clusters based on the socioeconomic of each neighborhood, with the averages per cluster being shown below.



Cluster	Color	Density (/sq mi.)	Percent Housing Crowded	Percent Households Below Poverty	Percent Aged 16+ Unemployed	Percent Aged 25+ Without GED	Per Capita Income	Hardship Index
1	Red	22435.250000	5.880000	17.790000	9.930000	19.680000	26370.000000	35.400000
2	Purple	13864.830000	10.123077	21.700000	15.369231	39.615385	14731.153846	75.230769
3	Blue	9278.229524	2.466667	11.709524	10.771429	11.695238	32230.095238	24.333333
4	Green	23359.910000	1.080000	11.580000	5.380000	4.120000	67295.600000	4.200000
5	Orange	9284.200000	4.416667	34.641667	23.912500	21.358333	15783.958333	73.416667

Some key takeaways that cluster #4 is the wealthiest cluster, having the highest Per Capita Income, lowest unemployment of people age 16+ and also lowest percentage of households below poverty, also having the lowest hardship index. The poorest cluster would be cluster #5, which has the highest percentage of households below poverty and also the highest unemployment, however it does not have the lowest Per Capita Income or the highest hardship index, that would be cluster #1, although the difference is relatively small.

4.2 The Top 5 Venue Types in Each Cluster



Cluster	Color	Per Capita Income	Hardship Index	#1 Popular	#2 Popular	#3 Popular	#4 Popular	#5 Popular
1	Red	26370.000000	35.400000	Indian Restaurant	Mexican Restaurant	Bar	Coffee Shop	Pizza Place
2	Purple	14731.153846	75.230769	Mexican Restaurant	Pizza Place	Sandwich Place	Coffee Shop	American Restaurant
3	Blue	32230.095238	24.333333	Coffee Shop	Park	Bar	Sandwich Place	Pizza Place
4	Green	67295.600000	4.200000	Chinese Restaurant	Pizza Place	Bar	Coffee Shop	Sandwich Place
5	Orange	15783.958333	73.416667	Park	Fast Food Restaurant	Bus Station	Grocery Store	Liquor Store

Looking at the most popular, there is no commonality among the results, however for the 2nd most popular (popular in terms of commonality), 'Pizza Place' is the 2nd most popular in both the poorest and the wealthiest clusters. Additionally. 'Pizza Place' does appear as the 5th most popular for both cluster 1 and cluster 3, so it is common among all clusters. It would therefore appear that pizza places are rather socioeconomically neutral, which is unsurprising as the type of food served by Pizza Places (primarily pizzas) is something that can come in a wide range of prices (depend on quality) and therefore it is unsurprising that it would be common amongst the clusters. The same appears to apply coffee places which appear in 4/5 clusters, 3/5 as 4th most popular while #1 most popular for cluster 3.

However what's interesting is the clusters in which bars are among the 5 top, which all are 3rd most popular within clusters 1,2,4. This is particularly of note because this excludes the poorest clusters, neither cluster 2 or 5 have bars within their top 5 most popular venue types. This could potentially indicate that the lower socio-economic factors of those clusters' prices them out of bars. More so, the 2nd to poorest cluster has a Liquor stores as 5th most popular (as opposed to a bar) whilst the poorest has no venue that is centered around alcohol, indicating that they are like priced out. That being said, this is not to say there are no bars or no liquor stores within the poorest cluster, however it is likely that the poor socioeconomic status of the cluster has resulted in a decreased prominence of such venues.

5. Discussion

In looking at the effects of socioeconomic factors, the main consequence of socioeconomic factors on types of venues within the top 5 appears to be tied to alcohol centric venues. Specifically, bars were absent from the poorest 2 clusters, and the poorest cluster had no alcohol centric (bar, liquor store) venue whereas the second poorest had a liquor store. An indication here is that increased wealth of a cluster of neighborhoods results in a greater prominence of bars within the area.

On the other hand, pizza place and coffee shop venues appear to be independent of socioeconomic status of clusters indicating that these venues are universal among socioeconomic status of neighborhoods.

Another interesting find is that of all venues within the top 5, the only nonfood related venues were parks and bus stations. Bus stations only appear in the top 5 of 1 cluster, this cluster being the 2nd poorest cluster. Since this is the only public transportation type venue, it would indicate that public transport is important to this poorer neighborhood. The reason why this appear in cluster 5 rather than cluster 2 (poorest based on per capita income) is perhaps due to cluster 5 having the highest percentage of households below poverty (34.6%), which could contribute the importance of public transportation given that more households are below poverty and therefore less are able to afford other types of transportation.

6. Conclusion

In conclusion, the effect of socioeconomic status on the prevalence of specific types venues is dependent on the type of venue itself. Venues like Pizza Places and Coffee Shops appear universally among clusters, however venues like Bars, appear only in the more affluent clusters being absent in the top 5 of the poorest two clusters. The number of households below poverty would also indicate the importance of public transportation to a cluster, in which the cluster with the highest percentage of households below poverty was the only cluster which had a venue related to public transportation, Bus Stations, within the top 5 venues.