TOP VENUES IN SOCIOECONOMIC CLUSTERS

By: Jakub Polanowski

Will Socioeconomics Affect the Types of Venues Present in Clusters of Neighborhoods?

- Socioeconomic factors are important features used to gauge the economic well-being of a neighborhood
- For a city government, is it possible to get an idea of the types of venues that will be most common within clusters of neighborhoods with similar socioeconomic factors?
  - Could be used to determine which types of venues would thrive within clusters
  - What type of venues are most important to residents within the cluster due to their socioeconomic factors
- Are there types of venues that are particular affected by high or low socioeconomic status?

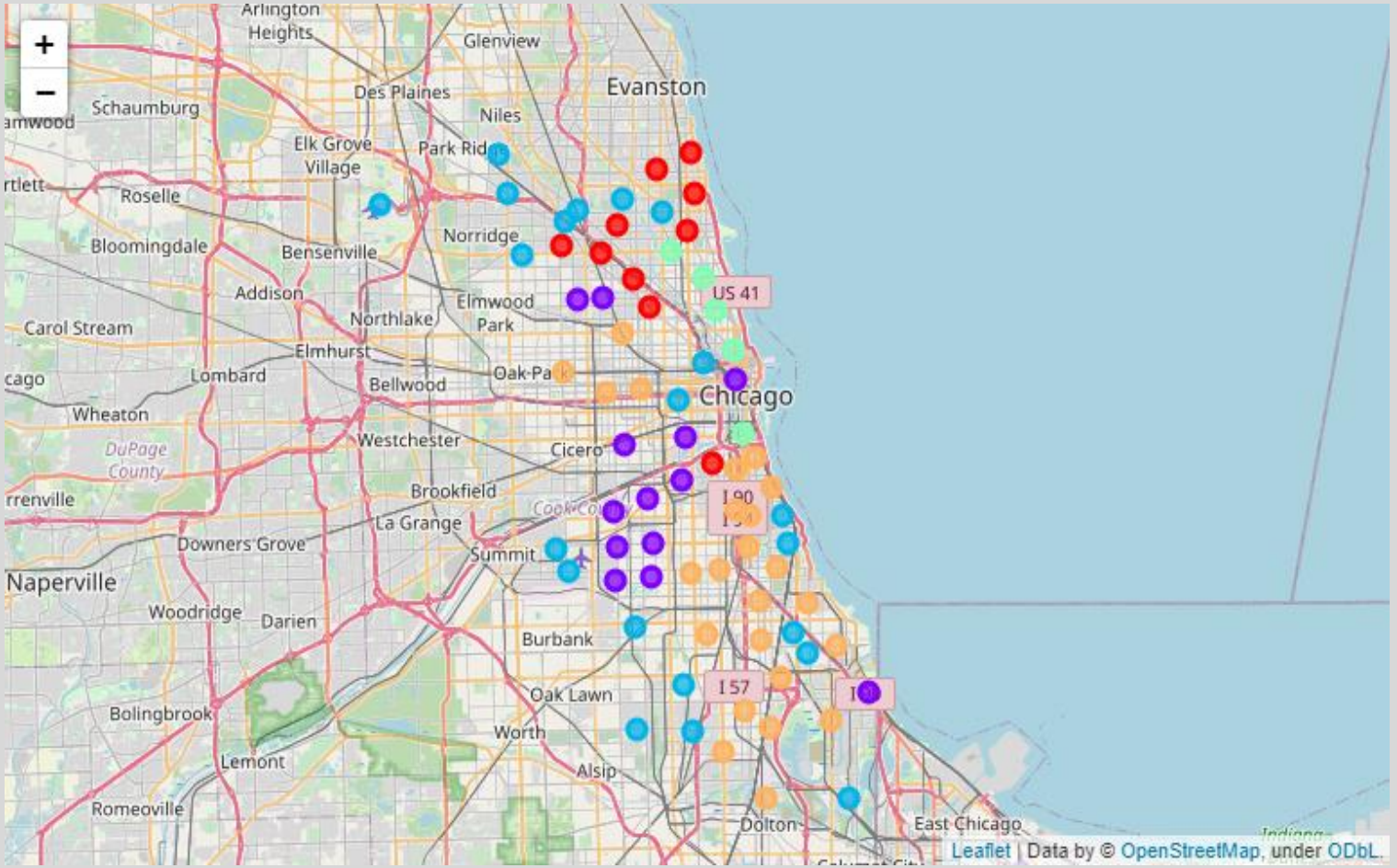
Data

- Data was source from the City of Chicago Socioeconomics 2017 Database
  - 9 features
  - 2 irrelevant features dropped (community area number' and 'percent aged under 18 or over 64
  - Missing values for any features resulted in the neighborhood being dropped from the dataset
- Geolocation data for each Neighborhood obtained by performing geocoding
  - Arcgis geocoding resolver used
- Lists of venues obtained form FourSquare API via the geolocation coordinates of neighborhoods
  - Radius of 500 ft used to approximately get all venues within the neighborhood

Methodology

- Socioeconomic features normalized by converting values to standardized scalars of between 0 and 1
- Clusters calculated using K Nearest Neighbors
  - Number of clusters set to 5
  - Used K Nearest Neighbors implemented KMeans function from scikit-learn
  - Model fitted to normalized socioeconomic features
- Top 5 Venues Determined for Each Cluster
  - Cluster Identifier Applied to Each Venue based on which Neighborhood Venue was in
  - Onehot encoding applied to get counts of each Venue Category
  - Onehot encoded data grouped by cluster identifier
  - Top 5 most common venue types columns added to socioeconomic dataset

Map Of Clusters



Cluster	Color
1	Red
2	Purple
3	Blue
4	Green
5	Orange

Economic Rankings of Clusters

- Cluster 4 is most affluent
- Cluster 2 has highest hardship index and lowest per capita income
- Cluster 5 has highest percent households below poverty

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

Top Venues in Each Cluster

- Poorest two clusters do not have bars within top 5, others do
- Cluster with highest % households below poverty only cluster to have Bus Station within top 5 venues
- Coffee shops and Pizza Places present within top 5 of 4/5 clusters eachs

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

Conclusion

- Coffee Shops and Pizza Places appear to be approximately universal among cluster, their prevalence appears to be independent of socioeconomic status
- Bars and Bus Stations appear to be linked with socioeconomic factors, in which bus stations only appear in the top 5 for the cluster with the highest % households below poverty and bars only in the wealthy to moderately wealthy clusters
- Remainder of Venue Types appear to be approximately random and no indication of a trend with socioeconomic factors was observed.