First Class Restaurant Location Analysis: Finding Optimal Locations within the United States

Sammie Haskin

Kennesaw State University

June 3, 2020

Introduction

- The creation of a successful first-class restaurant franchise can hinge on a variety of geographic and sociodemographic factors:
  1. identifying the target demographic
  2. maximizing exposure to potential customers
  3. accessibility

- As a result the selection of an optimal location is paramount in the creation of a successful first-class restaurant.
- Using multiple techniques in statistics and machine learning, the aim is to identify counties in which a first-class restaurant venture could be profitable

Data

- Data were gathered from a variety of sources across 2772 counties and 46 states.
- Using a variety of techniques and concepts in set theory, these data were integrated into one data set.
- The sources gathered are as follows:

Information Derived From County-Level Data Sets

Data Set	Information Contained
Federal Bureau of Investigation: Crimes Known to Law Enforcement (2015, 2016, and 2017)	Frequencies of various categories of crime
Office of Policy Development and Research: Free Market Rent Info (2015-2017)	Estimates of the median cost for rent in each area (for gauging inflation)
United States Census Bureau	Geographic area of each county and estimates Of the size of each county's population
United States Department of Agriculture	Employment and unemployment rates, Median household income
Bureau of Economic Analysis	Personal income by county

Data

Counties Data Dictionary

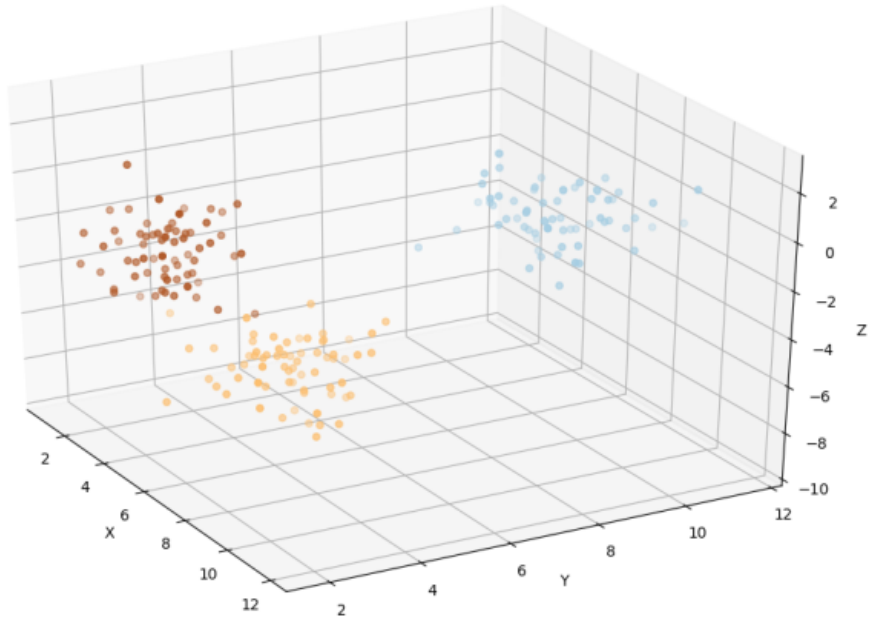
Variable Name	Description	General Type	Specific Variable Type
area	name of the county followed by the state in which the county is located	Categorical	String
crime_severity	yearly weighted sum of the number of crimes in a given county on average (2015-2017)	Numeric	Integer
state_crime_severity	yearly weighted sum of the number of crimes in a given state on average (2015-2017)	Numeric	Integer
popestimate2017	most recent estimate of the number of people within each county in 2017	Numeric	Integer
population_density_est	yearly number of people per square mile of county area on average (2015-2017)	Numeric	Float
state_population_density_est	yearly number of people per square mile of state area on average (2015-2017)	Numeric	Float
crime_severity_by_pop_density	average weighted sum of the number of crimes in a county after dividing by the average county population density estimate (2015-2017)	Numeric	Float
state_crime_severity_by_pop_density	average weighted sum of the number of crimes in a state after dividing by the average state population density estimate (2015-2017)	Numeric	Float
avg_income	average income per capita by county from 2015-2017	Numeric	Float
employment_pct	average yearly employment rate of a county from 2015-2017	Numeric	Float
avg_fmr_est	Yearly average of the mean cost of 0-4 room apartments in each area within (2015-2017)	Numeric	Float

Utilizing the resulting data set, the assessed geographic and sociodemographic factors contained would be used for a variety of purposes in this analysis.

Methods

For the purposes of identifying clusters of similar counties within the data, density based spatial clustering of applications with noise (DBSCAN) was utilized.

The cluster of counties exhibiting the most optimal properties would subsequently be identified and further analyzed.



Methods

DBSCAN has several benefits over the K-Means clustering algorithm in that:

1. It makes no assumptions about the shape of each cluster.

2. It allows for the identification of the number of clusters in the data without prior specification.

3. Extremely differing observations can be identified as noise without assignment to a cluster.

Methods

- The DBSCAN algorithm takes two parameters: epsilon, and minPts

  epsilon - the minimum distance of each point from a cluster to be considered a member of that cluster

  minPts - the total number of nearby points to consider the points as belonging to one cluster

- The minPts parameter was first chosen using the heuristic of taking the natural logarithm of the sample size: $\ln(2772) \approx 8$
- Utilizing principle components analysis with three components, a value for epsilon was chosen such that:
  - Unique clusters could be observed upon visualization with 3 components.
  - Observations that could be considered noise were not assigned to a cluster.
  - As necessary both parameters were revised through trial and error.

Methods

Once a cluster of similar counties exhibiting near optimal metrics was identified:

1) The inverse of crime severity was calculated such that lower scores denoted crimes of greater severity and higher scores denoted crimes of lesser severity.

2) The percentiles metric of each county was to be subsequently calculated.

3) An overall rank for each county was subsequently produced by finding the percentiles across all metrics.

Among all counties within the cluster that were above the 95th percentile, the top 10 ranked areas were to be further analyzed in terms of:

1) Proximity to businesses and tourist attractions

2) Competition from nearby groceries stores, delis and restaurants.

Results

- Using DBSCAN with an eps of .41 and a "minPts" of 10, 3 unique clusters were revealed.
- Viewing the total sample size in each cluster, the distribution sample of each cluster is as follows:

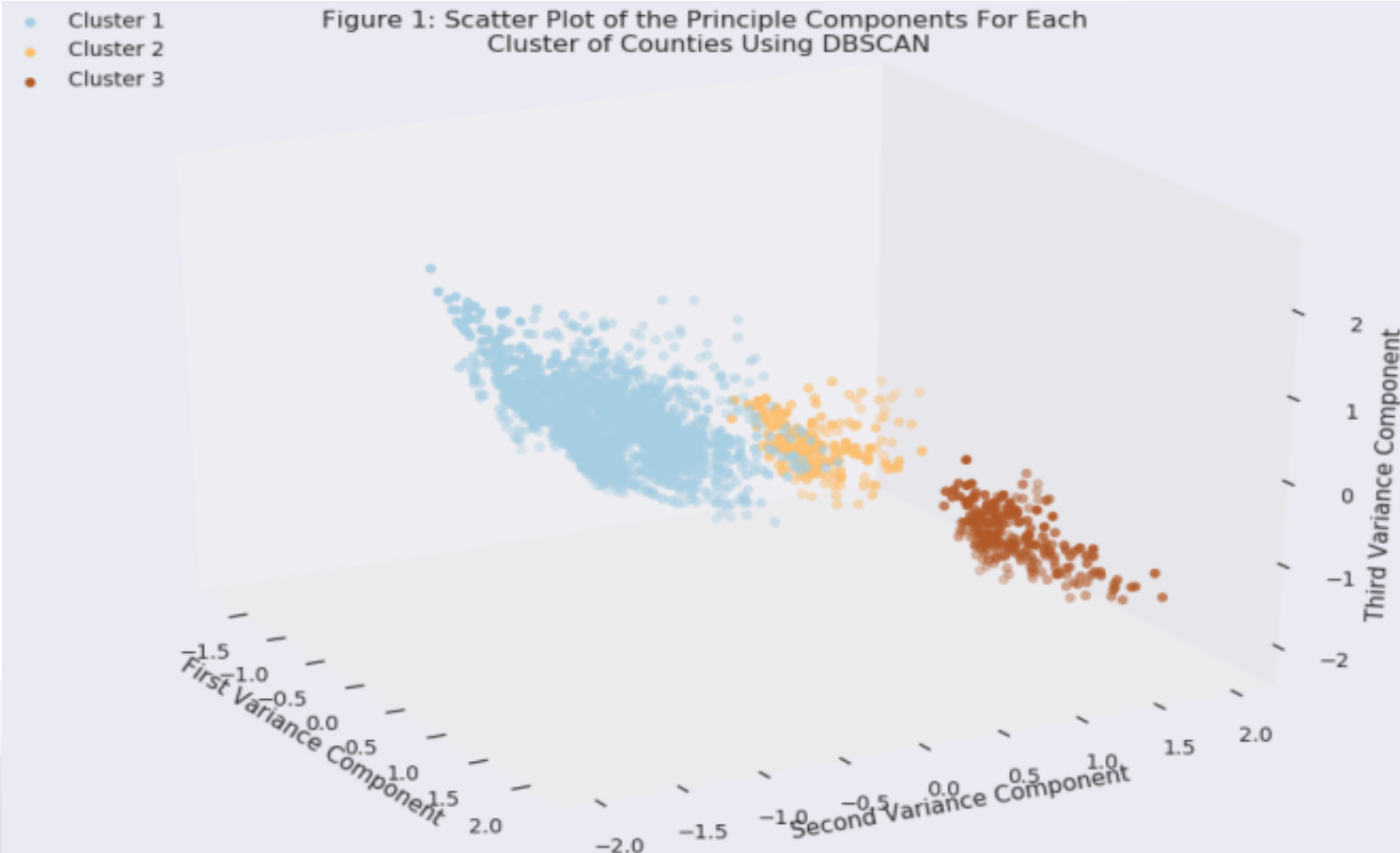
Sample Size Within Each Cluster

Cluster	*N*
1	1886
2	229
3	252
Noise	405

Results

- Utilizing Principle Component Analysis, it was discovered that 3 factors accounted for 80.62% of the variance of across all variable used for clustering.

- Viewing the resulting clusters using the principle components, the DBSCAN algorithm appeared to result in clusters that were distinct from each other and did not appear to cluster noise.



Results
Cluster 1

- Appeared to have a moderate yet variable population density.
- State and county level crime rates were extremely small within this cluster.
- Counties within this cluster had moderately large incomes per capita.
- Counties within this cluster had average employment rates of which were similar to those of other clusters.

Descriptive Statistics for the 1st Cluster of Counties

	Population Estimate (2017)	Population Density Estimate	County Crime Severity Index	State Crime Severity Index	Employment Rate	Average Income Per Capita
Sample Size	1886	1886	1886	1886	1886	1886
Mean	52485.69	81.75	33.70	1875.58	95.18%	40756.91
Standard Deviation	76743.56	105.40	29.75	959.11	1.48%	9326.30
Minimum	457.00	0.47	0.00	6.52	89.59%	21136.33
1st Quartile	12018.50	19.58	12.82	1162.96	94.28%	34707.42
Median	25924.00	43.30	25.63	1711.87	95.30%	39197.17
3rd Quartile	57212.25	94.75	44.76	2645.55	96.29%	44954.50
Maximum	747642.00	741.24	199.06	4579.55	98.22%	137133.00

Results
Cluster 2

- Contained very large population densities on average.
- State and county level crime rates were moderately large within this cluster.
- Contained significantly lower incomes per capita than other clusters.
- Counties within this cluster had slightly below average employment rates.

Descriptive Statistics for the 2nd Cluster of Counties

	Population Estimate (2017)	Population Density Estimate	County Crime Severity Index	State Crime Severity Index	Employment Rate	Average Income Per Capita
Sample Size	229	229	229	229	229	229
Mean	55653.16	100.78	50.45	5855.90	94.01%	35193.53
Standard Deviation	79636.26	115.09	35.31	97.68	1.10%	7478.99
Minimum	1628.00	1.45	3.29	5535.38	91.22%	19944.67
1st Quartile	14184.00	30.91	25.30	5803.95	93.24%	30948.67
Median	25334.00	55.99	40.25	5925.90	94.09%	33805.67
3rd Quartile	61386.00	121.19	67.81	5925.90	94.87%	38404.00
Maximum	589162.00	568.54	169.28	5925.90	95.98%	76168.00

Results
Cluster 3

- Contained very small population densities on average.
- State and county level crime rates were extremely large within this cluster.
- Contained higher incomes per capita than other clusters. The difference was negligible, however.
- Counties within this cluster had slightly above average employment rates.

Descriptive Statistics for the 3rd Cluster of Counties

	Population Estimate (2017)	Population Density Estimate	County Crime Severity Index	State Crime Severity Index	Employment Rate	Average Income Per Capita
Sample Size	252	252	252	252	252	252
Mean	35849.49	37.53	59.71	9496.48	95.71%	41549.56
Standard Deviation	57669.80	60.53	45.93	31.25	0.97%	8838.88
Minimum	134.00	0.18	0.00	9482.73	92.81%	25054.67
1st Quartile	5343.00	3.73	25.86	9482.73	95.18%	36402.75
Median	13759.50	15.22	42.96	9482.73	95.83%	40116.67
3rd Quartile	40241.50	43.90	83.66	9482.73	96.34%	45207.67
Maximum	362457.00	371.13	222.06	9567.22	97.97%	89232.00

Results

Summary of Each Cluster

Summaries of Each Cluster

Cluster 1	Cluster 2	Cluster 3
• Modest yet highly varying population densities	• Highest population density	• Extremely low population density
• Extremely low crime rates	• Modest crime rates	• Extremely high crime rates
• Relatively high income per capita	• Lowest income per capita	• Highest income per capita but negligibly so
• Relatively high employment rates	• Lowest employment rates, but negligibly so	• Highest employment rates but negligibly so

Results

Refining the List of Preferable Counties

- Overall, counties within Cluster 1 exhibited a preferable combination of favorable characteristics
- Containing 1882 counties, it was imperative to reduce the list of possible counties to only the most preferable.
- Subsetting the list of counties from Cluster 1, 10 counties that generated incomes per capita above the 95th percentile were chosen for further analysis on the basis of their average percentile across each factor.

Results:

Refining the List of Preferable Counties

- The 10 counties that exhibited the most favorable metrics are as follows:

Table 7: Geographic and Sociodemographic Factors of Counties Exhibiting Favorable Metrics

County Name	Average Percentile	Population Estimate (2017)	Population Density Estimate	Income Per Capita	Estimated County Crime Severity Index	Estimated State Crime Severity Index	Employment Rate
Rockingham, New Hampshire	96.50%	306363	382.52	$67,687.00	0.43	6.52	96.83%
Chittenden, Vermont	95.66%	162372	261.00	$57,592.00	0.21	52.10	97.50%
Hunterdon, New Jersey	92.12%	125059	286.41	$83,532.33	0.00	30.60	96.15%
Burlington, New Jersey	90.64%	448596	547.17	$57,716.00	0.00	30.60	95.37%
Grafton, New Hampshire	88.24%	89386	50.97	$57,647.00	1.90	6.52	97.57%
Oldham, Kentucky	87.32%	66415	333.02	$57,642.33	7.33	736.69	96.37%
Washington, Minnesota	86.63%	256348	599.33	$61,734.33	8.33	1711.87	96.84%
Delaware, Ohio	85.74%	200464	430.34	$68,748.00	15.10	1162.96	96.45%
Dallas, Iowa	85.52%	87235	141.91	$61,719.00	9.18	1251.73	97.45%
Lake, Illinois	84.59%	703520	514.78	$73,865.67	13.47	594.04	95.01%

Results:

Analysis of Area Attractions and Competition

Using the Foursquare api along with the coordinates of each county, the following information was gathered:

- the number of nearby venues
- the number of nearby restaurants
- metrics as to the competitiveness (total likes and rating out of 10) of each restaurant within the area

The following data combined with the previous metrics into these counties provide insight into the usefulness of each location in the creation of a first-class restaurant:

Estimates for the Number of Venues and Restaurants Within Each County

	Total Restaurants	Total Venues	Total Highly Rated Restaurants	Average Restaurant Rating	Average Restaurant Likes
Rockingham, New Hampshire	238	371	22	8.19	35.22
Hunterdon, New Jersey	206	309	3	7.75	23.94
Washington, Minnesota	142	282	11	8.18	34.73
Lake, Illinois	112	172	39	8.62	82.97
Oldham, Kentucky	101	192	4	7.62	27.16
Delaware, Ohio	59	137	0	7.08	14.69
Burlington, New Jersey	48	103	1	7.41	11.54
Chittenden, Vermont	9	42	1	7.64	17.22
Grafton, New Hampshire	1	5	0	7.68	4
Dallas, Iowa	1	4	0	7.67	1

Results:

Analysis of Area Attractions and Competition

- As expected, densely populated counties with high income per capita also housed the most venues.

Such areas include: - Rockingham, New Hampshire - Hunterdon, New Jersey

- Washington, Minnesota - Delaware, Ohio - Lake, Illinois

Of these:

- All contained a total of at least 170 venues of any type and at least 59 restaurants, food deli's, or bakeries.
- Rockingham, Washington, and Lake Counties each contained at least 10 highly rated restaurants with ratings of at least 9 out of 10.
- Delaware and Hunterdon contained 0 and 3 highly rated restaurants, respectively.

Results:

Analysis of Area Attractions and Competition

- Of the 10 counties, those that averaged no more than $58,000 income per capita contained a variable amount of venues as well as a sparsity of highly rated restaurants.

These areas include:

- Grafton, New Hampshire - Oldham, Kentucky - Chittenden, Vermont - Burlington, New Jersey

Of these:

- Grafton, New Hampshire was sparsely populated and contained only 5 recorded venues and 1 of which was a restaurant.
- Oldham, Kentucky contained a moderate population size within a densely populated area. This area contained a large amount of venues of any kind (192) and 101 of which were restaurants. Only four restaurants within the county were rated with a score of at least 9 out of 10.
- Burlington, New Jersey was densely populated, containing a moderate amount of venues and restaurants, one restaurant of which received a rating greater than 9 out of 10.
- Chittendenen, Vermont contained a moderately large population density, high the employment rates, and a moderate amount of venues (42) and a low amount of restaurants (9). Only one restaurant within the area was highly rated.

Results:

Analysis of Area Attractions and Competition

The county of Dallas, Iowa, unlike the other 9 counties, contained:

- A moderate size at 87235 persons.
- A population density of 141 people per square mile.
- An extremely low amount of restaurants and venues of any kind with totals of 1 and 4, respectively.
- No restaurants that were rated as a 9 out of 10

Discussion

The counties of Rockingham, Lake, and Washington exhibit preferable metrics across:

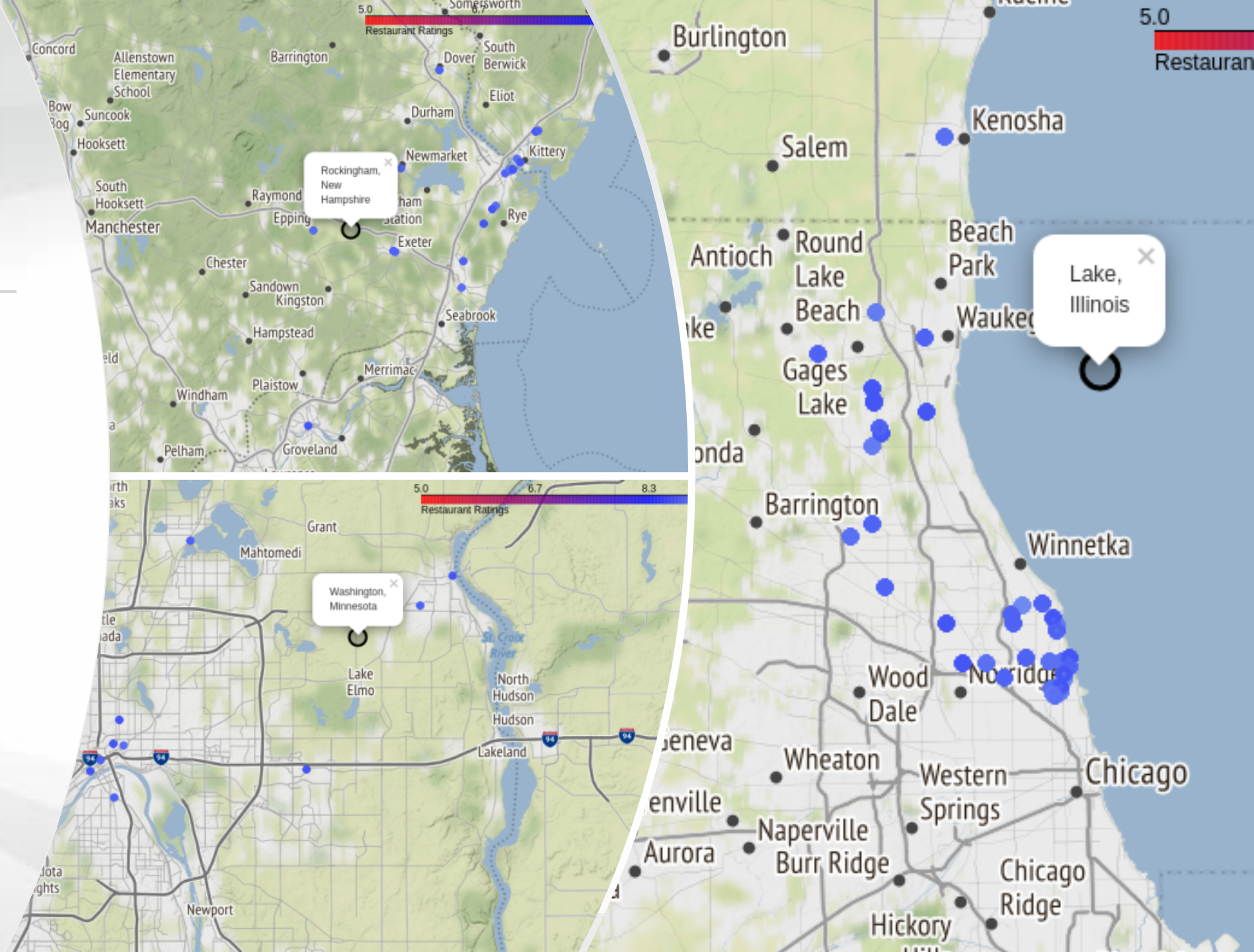
- crime rates
- population size and density
- income per capita
- potential foot traffic as gauged by the number of businesses in the area.

Although these areas may attract potential customers:

- The number of highly rated restaurants is multitudinous within these counties
- As a result, nearby competition must be accounted for if creating a first-class restaurant within one of these counties.

Discussion

- The following plots display the locations of restaurants with ratings of at leest 9 out of 10.
- Observing the locations of nearby competition:
  - Many high rated restaurants are centered close to large bodies of water.
  - It is possible that such areas may provide aesthetics that result in higher ratings than otherwise.
  - Careful analysis of the benefits and drawbacks of creating first class restaurants in proximity to these bodies of water



Discussion

- In contrast, the remaining counties appear to have a lack of competition among highly rated restaurants:
  - Each area contains either no highly rated venues or no more than 3 highly rated venues.
  - Of these counties, the high income per capita, the busy economy, and the dearth of highly rated restaurants may prove profitable for the creation of a first-class restaurant within Hunterdon, New Jersey or Delaware, Ohio.
- Careful thought must be given if setting a first-class restaurant within Chittenden, Grafton, Dallas, or Oldham, however:
  - Each of these areas contains either an income per capita below $60,000 or a relatively low amount of businesses that could boost foot traffic in the vicinity.

Conclusion

- The DBSCAN algorithm was used for the identification of a cluster of potentially suitable areas for a first class restaurant.
- Within this cluster, statistical procedures were used to derive and assess the suitability of 10 potentially viable counties.
- Through the implementation of these procedures, a subset of counties exhibiting optimal properties were identified. As data is ever changing, such an analysis can be applied to re-evaluate preferable areas during the expansion of a successful franchise.

References

- Blumstein, A. (1974). Seriousness Weights in an Index of Crime. American Sociological Review, 39(6), 854-864. Retrieved from http://www.jstor.org/stable/2094158
- Bureau of Economic Analysis (2018). Personal Income by County, Metro, and Other Areas [Data file] Retrieved from https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas
- * Federal Bureau of Investigation (2010). Crime in The United States. Retreived from https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/violent-crime
- Federal Bureau of Investigation (2015). Offenses Known to Law Enforcement by State by Metropolitan and Nonmetropolitan Counties [Data file]. Retrieved from https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/offenses-known-to-law-enforcement/offenses-known-to-law-enforcement
- Federal Bureau of Investigation (2016). Offenses Known to Law Enforcement by State by Metropolitan and Nonmetropolitan Counties [Data file]. Retrieved from https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-6/table-6.xls/view//
- Federal Bureau of Investigation (2017). Offenses Known to Law Enforcement by State by Metropolitan and Nonmetropolitan Counties [Data file]. Retrieved from https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/downloads/download-printable-files
- Foursquare (n.d.) PlacesAPI. Retrieved from https://developer.foursquare.com/
- Google. (n.d.). Maps Static API. Retrieved from https://maps.googleapis.com/maps/api/
- Office of Policy Development and Research (2017). County Level Fair Market Rents [Data file]. Retrieved from https://www.huduser.gov/portal/datasets/fmr.html#2019_data
- United States Census Bureau (2010). Population, Housing Units, Area, and Density: 2010 - United States -- County by State; and for Puerto Rico 2010 Census Summary File 1 [Data file]. Retrieved from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk
- United States Census Bureau (2017). City and Town Population Totals: 2010-2017 [Data file]. Retreived from https://www.census.gov/data/datasets/2017/demo/popest/total-cities-and-towns.html
- United States Department of Agriculture (2018). Employment, Unemployment, and Median Household Income [Data file]. Retrieved from https://www.ers.usda.gov/data-products/county-level-data-sets/

* referenced but not used