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

Samme Haskin

Kennesaw State University

June 3rd, 2020

Introduction

Within the food industry of the United States, the successful expansion of a fine-dining franchise can hinge on a variety of factors including the logistics of the business, the implementation of quality service, and employment-related factors from personnel selection to retention. Subsequent to the abstraction of the logistics of restaurant functioning as well other critical aspects of business strategy, selection of optimal locations from which to expand a fine-dining restaurant franchise could potentially catalyze success or significantly inhibit future growth of a restaurant. Given the high degree of importance of location selection on the success of a new restaurant, in this analysis, the sociological and geographic characteristics of various locations within the United States are analyzed as to their viability for the facilitation of the long-term growth of fine-dining restaurant franchises that aim to expand across the United States. By leveraging location data along with open access datasets, county areas across the United States will be clustered by their various geographical factors, such as city area type as well as their distribution of venue categories within their substrata, and by their sociodemographic factors, including area affluence, crime rate, and population size. The resulting clusters will be analyzed in order to determine the most preferable locations from which fine dining restaurants could potentially expand to facilitate their long-term success.

Data

Within this analysis, geographical and sociodemographic data on counties within the United States were gathered from a variety of sources. Crime rates on a county level was gathered from the Federal Bureau of Investigation (2015-2017). From the United States Census Bureau, poulation estimates, geographical area and population density were derived from the years of 2010-2017. Information derived from the United States Department of Agriculture allowed insight into the employment rates and median household income of counties. The estimated cost of rent of apartments with a variable number of rooms was also derived from the Office of Policy Development and Research (2015-2017). The google static maps api was utilized to get precise coordinates for the location and bounds of eah county.

Using multiple methods in data merging, county data was integrated in a methodical manner. Of the data sets used to gauge the geographical census data, population size/density, unemployment rates, and crime rates, many of these data sets contained missing data for various counties and frequently contained the same set of counties named differently across data sets. To overcome these differences, various techniques were used. Using the python programming language, various informational strings such as numbers denoting captions as well as strings such as "city", "county", and "police department" were removed from each county name prior to merging data sets.

After merging data sets on 'state' and first initial of the county name, a technique known as the Levenshtein distance was used to determine similarity between words that were not exact matches. Using such a function allowed for a greater ease of finding similarly named counties in each state for the sake of merging on otherwise seemingly disparate county names. Such a technique was utilized alongside set theory to display county names that otherwise appeared to have no match from each data set. These counties were renamed accordingly such that their sociological and geographical data could be more easily merged. Similarly, when smaller numbers of counties were missing in a new data set to merge into the larger partially merged data set, regular expressions were used to search for similarly named counties in new data sets by replacing vowels and spaces with optional wildcard characters of any length. The results, regardless of whether the Levenshtein distance technique or regular expressions was used, was inspected visually to ensure the accuracy of merging on the same set of counties.

The final data set contains 2772 counties and 46 states as well as a variety of metrics that were used for the purpose of clustering and identifying suitable counties for the creation of a first-class restaurant. The metrics used within the analysis are shown in Table 1. For each county assessed in the final data set, the coordinates of each county as well as the bounds of each county were recorded and described in Table 2. Utilizing both data sets, a subset of counties that demonstrate favorable metrics could be further analyzed as to the potential foot traffic and competition within each area.

Table 1: 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

Table 2: County Coordinates 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
Google County Name	formatted name of each county for the purpose of using the google maps api to gather county coordinates and bounds	Categorical	String
Latitude	latitude of the county	Coordinate	Float
Longitude	longitude of the county	Coordinate	Float
Northeast Bound	latitude and longitude associated with the northeast corner of the county	Coordinate	Tuple
Southwest Bound	latitude and longitude associated with the southwest corner of the county	Coordinate	Tuple

Method

For the purposes of discovering and understanding clusters of counties in the data, several statistical and machine learning methods were used. Subsequent to merging counties data sets from a variety of sources, imputation procedures were performed with the aim of capturing trends across counties with the aim of identifying as many counties that may be conducive a first class restaurant venture. With the data that were available, the estimates of the frequencies of crimes in different counties were calculated when missing. As the counts of crimes of different categories can be considered a discrete variable, a generalized poisson regression was used for the process of imputation. For the purpose of clustering and grouping counties on several key metrics, several features were engineered.

Using the recorded and estimated rates of different crimes, a weighted crime severity score was calculated. For this purpose, the weighted severity of each crime as devised by Blumstein (1974) were utilized, and a severity score for each year was calculated as a linear combination of each crime category with its assigned weight. The weights utilized for crimes of different categories are shown in Table 3.

Table 3: Weights For the Severity of Various Types of Crimes

Crime Category	Severity Score
Murder and non-negligent manslaughter	33.29
Rape	15.33
Robbery	6.43
Aggravated Assault	9.74
Property Crime	2.45
Burglary	2.64
Larceny Theft	2.26
Motor Vehicle Theft	2.29
Violent Crime	9.74

In addition, an estimate of the population density in each county was derived by calculating the population size per square mile. The crime severity scores to be used in the clustering analysis were subsequently adjusted to account for the population density within each county and state. For this analysis, the average of the metrics of state and county crime severity, income, and unemployment rate were calculated across the years of 2015-2017 for the purpose of performing a clustering analysis.

Subsequent to the process of engineering features, selecting uncorrelated features,and standardizing the features of population density, county and state crime severity index, and employment rate, a clustering algorithm known as density based spatial clustering of applications with noise (DBSCAN) was implemented to cluster counties. DBSCAN is an algorithm that allows researchers to cluster data points with similar features together while accounting for noise in the data where data points do not appear to fit in any cluster. The parameter of "minPts", or the total number of nearby points to consider the group of points as a cluster was first estimated using a heuristic of taking the natural logarithm of the sample size. The parameter of "eps", or the minimum distance of each point from a cluster to be considered a member of that cluster was chosen by visualizing the final resulting clusters using principle components analysis with 3 components. As necessary, these parameters were revised to form clusters that were distinct and did not appear to cluster points that could be considered as noise.

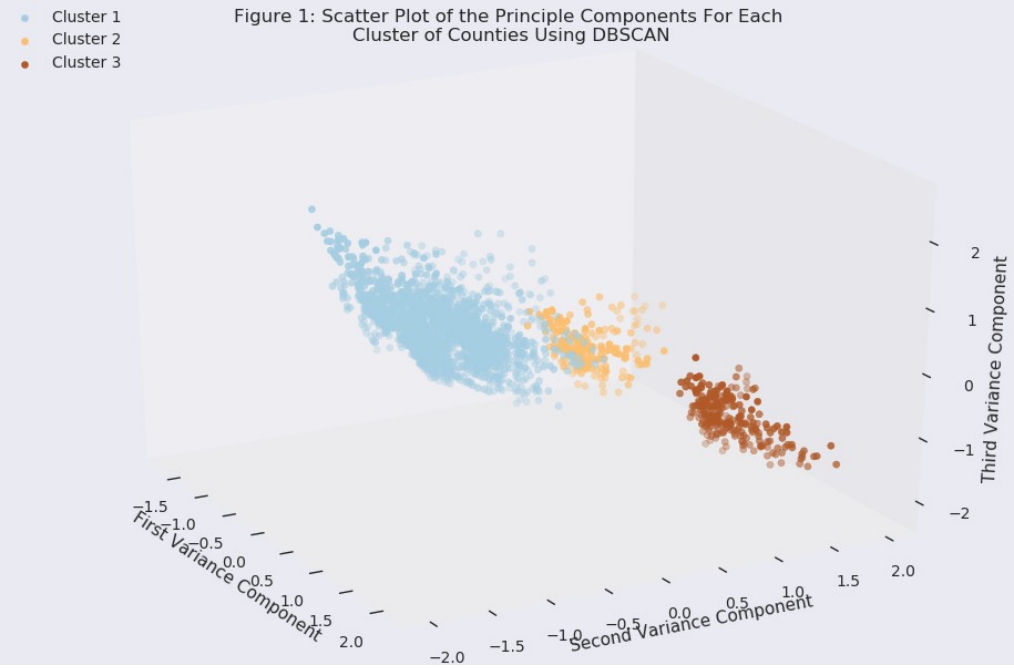
After analyzing the clusters formed with the DBSCAN algorithm , a subset of the cluster of counties that exhibited optimal properties across several metrics and an income per capita above the 95th percentile were to be further analyzed. To do so, the inverses of the metrics of county and state crime severity were found by multiplying each metric by -1 such that lower scores denoted greater crime severity and larger scores denoted a lower severity of crime within the area. Afterward, each county of the cluster were ranked from smallest to largest on each metric. By finding the average of each rank, an overall score across the metrics of crime severity, income, population density, and employment rate could be produced for the purposes of comparison and identification the top 10 high income counties given the data. Among the counties that were observed to exhibit the most favorable characteristics across multiple metrics, the degree of potential foot traffic, as assessed by the number of nearby businesses as well as county population density, and the degree of competition, as measured by the total number of nearby highly rated restaurants, could be gauged.

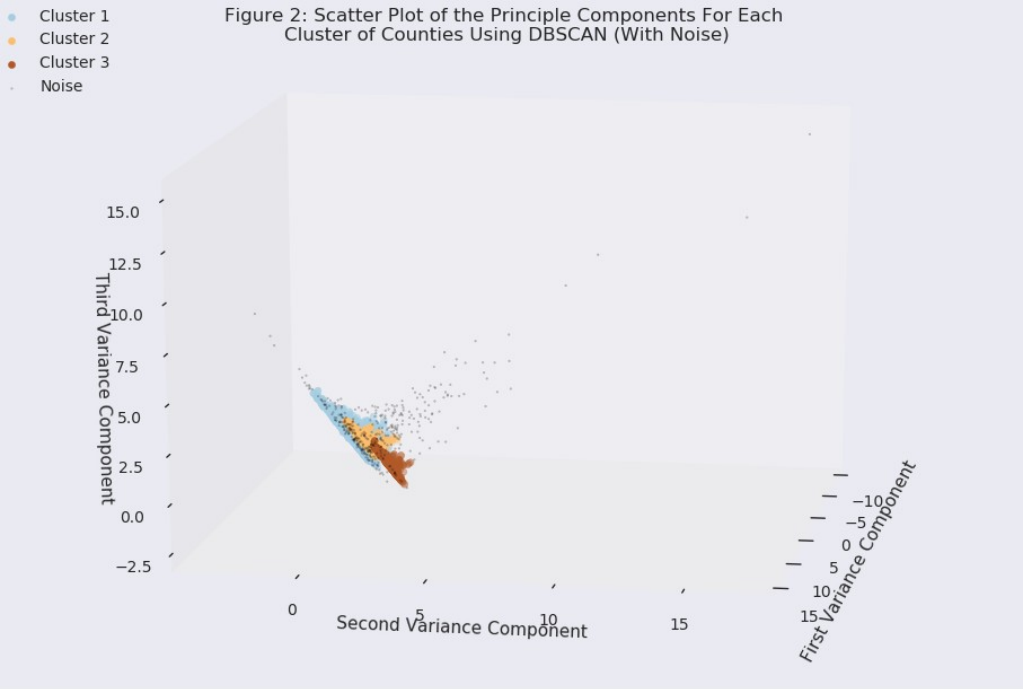
Areas with several venues within counties and a high income per capita may exhibit a high degree of foot traffic consisting of a population that may be able to afford to dine in a first-class restaurant. To identify counties in which this may be the case, the Google Static Maps api was used to record information pertaining to each county's coordinates and borders (Google, n.d.). Using the Foursquare api, information pertaining to the number, names, and ratings/likes of venues were acquired

within the estimated radius of each county (Foursquare, n.d.). From this information, a list of counties that could be optimal for the start-up of a first-class restaurant venture can be generated.

Results

The DBSCAN clustering algorithm was implemented using an "eps" of .41 and a minPts of 10, and the chosen features included county population density, county and state crime severity by population density, and employment percentage. More factors were not used for clustering, because at least one of these factors correlated moderately with average income per capita, free market rent prices, and state population density. From this clustering procedure, 3 clusters were revealed. Cluster 1 contained 1882 data points, while clusters 2 and 3 contained sample sizes of 227 and 247, respectively. 416 points were unassigned to a cluster. For the purposes of observing the formed clusters, principle components analysis was utilized in which 3 components were observed to account for 80.62% of the variance of the standardized metrics. As indicated by Figure 1 and Figure 2, the clusters formed from the analysis, it can be seen that distinct clusters were formed while counties that appeared to be noise were not assigned to a cluster. As indicated by Tables 1, 2 and 3, several distinctive features of each cluster were evident.





Counties from cluster 1 each appeared to have a modest population density of approximately 81.8 people per square mile on average. With a standard deviation of approximately 105.4, the population density was significantly variable across these counties however. This cluster also appeared to have the lowest county and state crime severity scores averaging 33.7 and 1875.6, respectively when accounting for population density (or persons per square miles). The average employment rate for counties within this cluster was 95.1%. This cluster consisted of a higher percentage of counties (4.3%). that were above the 95th percentile in average income in comparison to other counties, but this difference in the aggregate was negligible compared to those of other clusters. Only 4 (.21%) out of the 1886 counties within this cluster fell below the 5th percentile in income per capita however.

Table 4: 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

Cluster 2 had the highest population density at approximately 100.8 people per square mile on average. Similar to the other two clusters, the average employment rate of these counteis was approximately 94.0%. When comparing the county and state crime severity index after accounting for population density in the area, it can be seen that this cluster had a significantly more crime of greater severity than those observed from cluster 1. When accounting for population density, areas within this cluster had a mean county crime severity score of approximately 50.4 while having a state crime severity index of approximately 5855.9. While only 2.2% of the areas within this region were above the 95th percentile in income earned, approximately 3.5% of counties fell below the 5th percentile in median income per capita.

Table 5: 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

Counties within cluster 3 had the lowest population density at approximately 37.5 people per square mile on average. Although this set of counties had the highest employment rate at 95.7, the difference in this circumstance was once again negligible in comparison to the employment rates of counties in the other two clusters. Whilst having the highest rate of employment, it can still be seen that, when compared to those of the other two clusters, these counties had a slightly higher than average county crime severity index of 59.7 on average after taking into account the population size per square mile with a mean. With a very low variability in the state crime severity index of each of these counties, it can be observed that these counties were located in states with the highest weighted state crime severity scores of 9496.47 on average after accounting for state population density. No counties within this cluster below the 5th percentile in income per capita. Although approximately 4.0% of counties within this area were above the 95th percentile in income per capita, the high crime rates and low population density could make the prospect of setting a first class restaurant within these counties less profitable venture as a result.

Table 6: 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

Overall, cluster 1 consisted of counties that exhibited the most favorable metrics across demographic and geographical factors. As this cluster also consisted of 1886 counties, various techniques were used to further identify a subset of counties from this cluster that exhibit the most preferable metrics for a the creation of a first class restaurant. After calculating the percentiles of each metric for each county within cluster 1, a summary statistic that is the average of the percentiles was produced.
The list of counties was further subset through the identification of the top 10 counties that were also above the 95th percentile in income per capita, and the statistics for these counties are shown in Table 7.

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

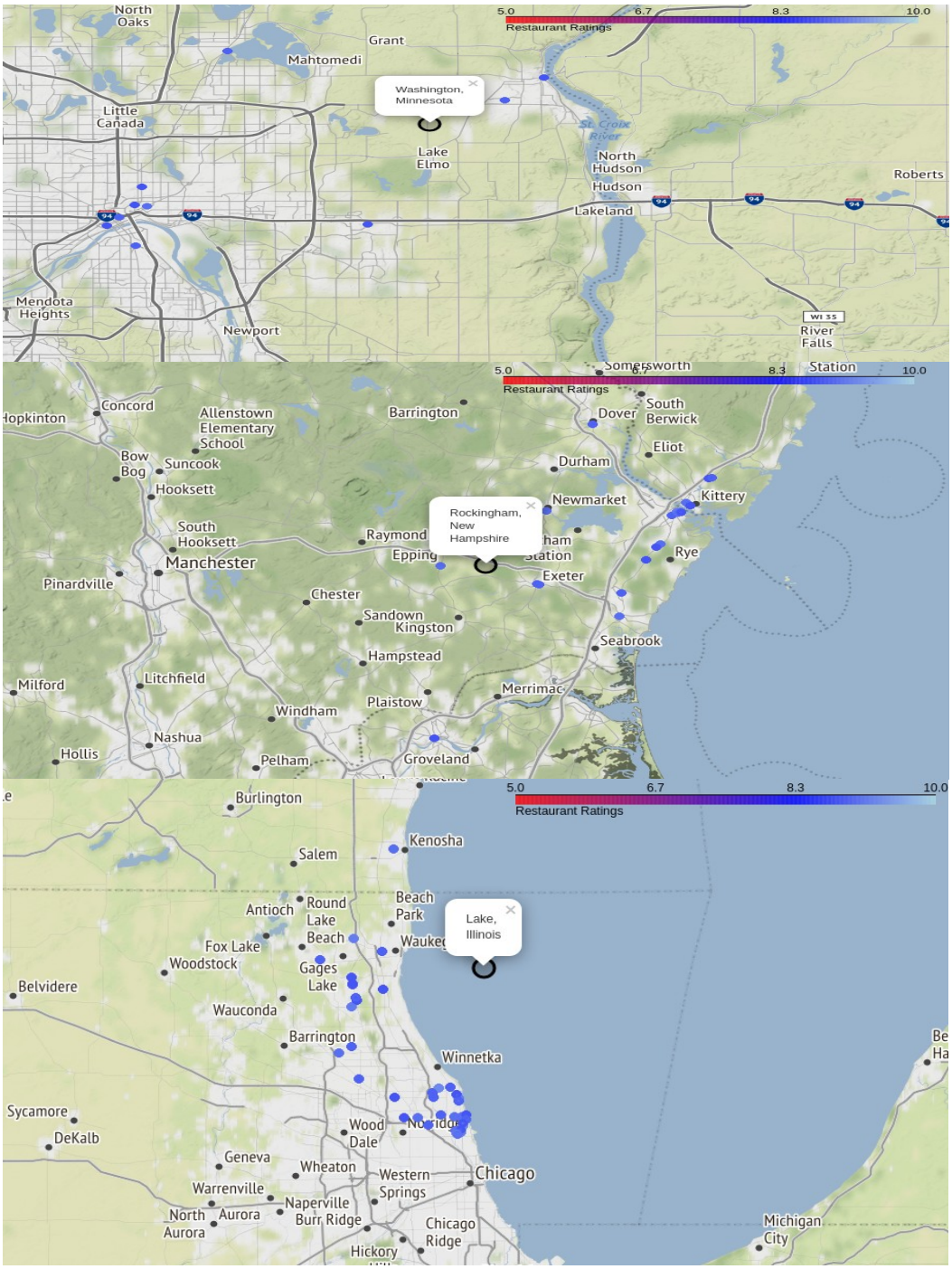
County Name	Average Percentile	Population Estimate (2017)	Population Density Estimate	Income Per Capita	Estimated County Crime Severity Index	Estimated State Crime Severity Index	Employment Rate
Rockingham, New Hampshire	96.50%	306363	382.52	$67,687.00	0.43	6.52	96.83%
Chittenden, Vermont	95.66%	162372	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%

For each of the 10 counties of interest, the number of businesses, restaurants, and restaurant ratings were recorded using the Foursquare api for the purpose of assessing the potential foot traffic and the competition within each area. These metrics are displayed in Table 8. Furthermore, the locations of counties as well as the restaurants and food bakeries with ratings at or above a 9 out of 10 within these locations were visualized as shown in Figure 3. Utilizing these metrics and geographical factors in tandem, informed decisions as to where a first class restaurant may thrive can result.

Table 8: 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

Map of the Potential Competition within Counties
Exhibiting Favorable Geographic and Sociodemographic Factors



Discussion

Of the 81 counties identified as having an income per capita of over the 95th percentile, the metrics of the ten counties previously displayed within Table 7 also indicated low crime rates, potentially thriving economies, and a preferable degree of potential foot traffic as measured by population size and density. As expected, the densely populated, high income counties of Rockingham County, New Hampshire, Hunterdon County, New Jersey, and Washington County, Minnesota housed several businesses and restaurants alike. Although the potential for foot traffic in these areas is likely, it can be also be seen that Rockingham and Washington also contain a high degree of competition within the area. In the counties of Rockingham and Washington, 22 and 11 restaurants with a rating of 9 or above, respectively.

In contrast, the counties of Lake, Oldham, Delaware and Burlington contained a significantly lower amount of restaurants and businesses within their area. While Oldham, Delaware, and Burlington contained 4, 0, and 1 highly rated restaurants, respectively, Lake County, Illinois contained the most at 39 highly rated restaurants, deli's, and bakeries combined. The counties of Chittenden, Grafton, and Dallas have even fewer venues and restaurants within the area. Of these three counties, only Chittenden, Vermont contained a restaurant/bakery with a rating of 9 or above. Within the areas of Oldham, Delaware, Burlington, Chittenden, and Grafton, more careful thought would be needed for the purposes of choosing locations that accrue a high degree of foot traffic given the dearth of potentially high quality competition and business of any kind within these areas.

For the areas of the prospect of creating a first-class restaurant in the busier counties of of Rockingham, Washington, and Lake County may require accounting for the surplus of highly rated restaurants within each area. Observing the locations of potential competitors, Figure 3, as shown previously, indicates that the vast majority of highly rated restaurants, bakeries, and deli's are significantly closer to bodies of water such as lakes and rivers. Care in selecting locations that could provide aesthetic views and scenery while minimizing the risk of customer attrition to nearby competition could be paramount for consideration if a first-class restaurant is to be situated in any of these 3 counties.

Conclusion

Utilizing cluster analysis with the DBSCAN algorithm in addition to other statistical procedures, locations that could facilitate the success of a first-class restaurant startup were identified. The DBSCAN algorithm revealed a cluster of similar counties with low crime rates, relatively high employment rates on average, a large population sizes, and relatively successful economies in which many could afford the monetary cost of fine dining. Within this cluster, counties were ranked according to their percentiles across multiple metrics such that a subset of ten high-income counties exhibiting favorable metrics could be identified. Within the highly populated counties of Rockingham, Washington, Lake County, Oldham, Delaware, Burlington, Chittenden, Grafton, and Dallas, the higher income of county residents, relatively low crime rates in surrounding areas after adjusting for population density, and high employment rates all indicate the potential for recurrent customers if a first-class restaurant is to be situated in any of these areas. With these favorable metrics arises the need to account for foot traffic within these areas as well as the higher degree of competition within nearby areas. By taking each of these factors into account, a successful first-class restaurant could result.

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