Battle of the Cities: Small Town USA

A Comparative Analysis of Two Colorado Competitors

Analyst: Roger Smith

Introduction to the Business Problem

All across America's Midwest there are hundreds of small and medium sized communities that are struggling to remain alive, vibrant, and relevant to the families who live there. Many of these cities were established during the frontier expansion of the United States during the 19th century. They served as the highways and supply lines carrying eastern goods across the country and for delivering farm produce to both coasts. They are the typification of the term "Small Town America". However, as the wealth, resources, and opportunities of the mid-twentieth and early twenty-first century have greatly expanded in metropolitan areas, these small towns have remained stuck with low incomes and few opportunities for the next generation that is raised there.

This report uses Four Square location data and machine learning techniques to explore and compare two of these towns in Southeastern Colorado. The objective is to identify the advantages and quality of living in both towns. The data and report will be useful to the cities' leaders as they seek to develop a more robust economy for their community.

This report is structured as a consulting product commissioned by the city leaders of Lamar, CO to assist them in competing for economic opportunities with their historical rival, La Junta, CO.

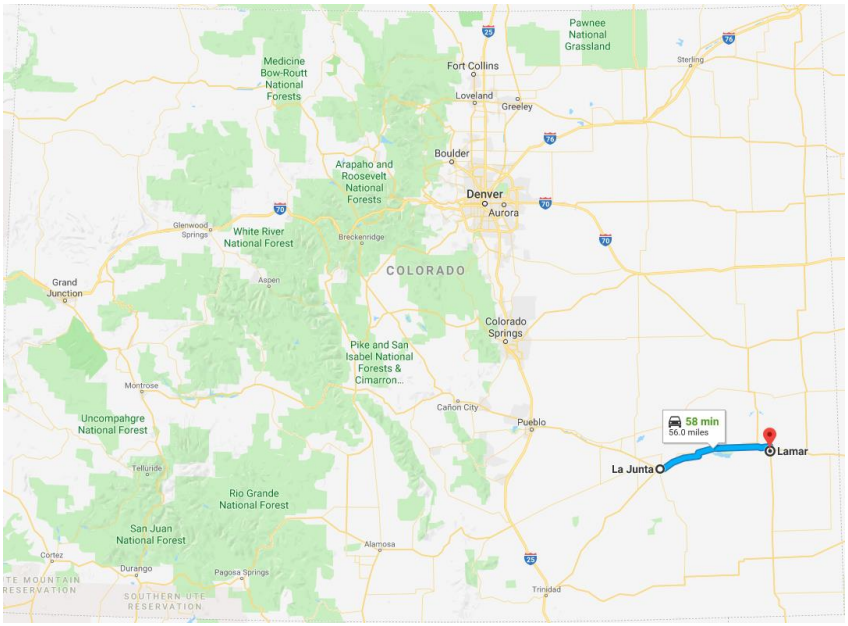


Figure 1. Colorado map showing location and distance between target towns.

Image: produced with Google Maps

Discuss Background

The two towns selected for this study are Lamar, CO, and La Junta, CO. The towns have very similar histories, locations, economies, and population sizes. Both serve as the county seat and are the largest communities in their respective counties. Both towns were established in the 1880's along the Arkansas River and the Santa Fe Trail, which was the major commercial highway from Missouri to Santa Fe New Mexico from 1820 to 1920, with its peak occurring around 1880.

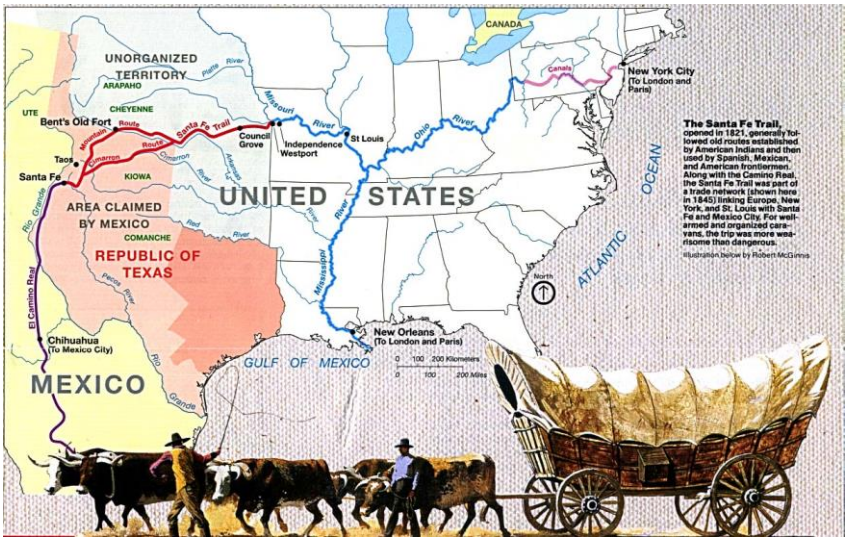


Figure 2. Historic map showing major commercial trails of the 1800's.

Image from Wikipedia: https://en.wikipedia.org/wiki/Santa_Fe_Trail

Lamar, Colorado

Lamar is the county seat for Prowers County and reports a 2017 population of 7,601. The census identifies it as 75% white and 44% Hispanic, with all other races having less than 2% each. (Note: census race categories are overlapping, so totals are always more than 100%.) Median income for the city is $31,521. Per capita income is $16,944.

Historical population		
Census	Pop.	%±
1890	566	—
1900	987	74.4%
1910	2,977	201.6%
1920	2,512	−15.6%
1930	4,233	68.5%
1940	4,445	5.0%
1950	6,829	53.6%
1960	7,369	7.9%
1970	7,797	5.8%
1980	7,713	−1.1%
1990	8,343	8.2%
2000	8,869	6.3%
2010	7,804	−12.0%
Est. 2016	7,534 [6]	−3.5%
U.S. Decennial Census[13]		

Figure 3. Historic Lamar train station and population.

Image from Wikipedia: https://en.wikipedia.org/wiki/Lamar,_Colorado

The towns largest industry is its surrounding agricultural and ranching production. Primary agricultural crops are dry land wheat, alfalfa, and corn. Primary animal products are beef with intermittent forays into pork.

La Junta, Colorado

La Junta is the county seat for Otero County and reports a 2017 population of 6,898. The census identifies it as 74% white and 44% Hispanic, with all other races having less than 2% each. (Note: census race categories are overlapping, so totals are always more than 100%.) Median income for the city is $29,002. Per capita income is $14,928.

Historical population		
Census	Pop.	%±
1890	1,439	—
1900	2,513	74.6%
1910	4,154	65.3%
1920	4,964	19.5%
1930	7,193	44.9%
1940	7,040	−2.1%
1950	7,712	9.5%
1960	8,026	4.1%
1970	8,205	2.2%
1980	8,338	1.6%
1990	7,637	−8.4%
2000	7,568	−0.9%
2010	7,077	−6.5%
Est. 2016	6,930 [5]	−2.1%
U.S. Decennial Census[15]		

Figure 4. Historic La Junta welcome station and population.

Image from Wikipedia: https://en.wikipedia.org/wiki/La_Junta,_Colorado

The towns largest industry is its surrounding agricultural and ranching production. Primary agricultural crops are cantaloupes, watermelon, and alfalfa. Primary animal products are beef with a small number of chicken facilities.

Describe Data

Initial queries using the browser at www.FourSquare.com of activities and businesses in both cities provide the following overviews. This high-level review of the data indicates that the reports from the two cities are not collected and created according to the same criteria. It appears that for Lamar, many businesses are listed as both “Food” and “Breakfast” and the same has not been done for La Junta businesses. This fact will skew mathematical analysis and should be accounted for in the final analysis and recommendations. However, it is interesting that overall, both towns seem to have equivalent numbers of businesses.

Table 1. Comparative venues in both towns from Four Square web browser interface

	Lamar, CO	La Junta, CO
Food	30	8
Coffee	4	2
Nightlife	8	4
Fun	8	3
Shopping	8	4
Breakfast	14	4
Top Picks (Trans-categorical)	30	30

Four Square maps showing the “Top Picks” in the core of each town are provided.

Top Picks: Lamar, CO

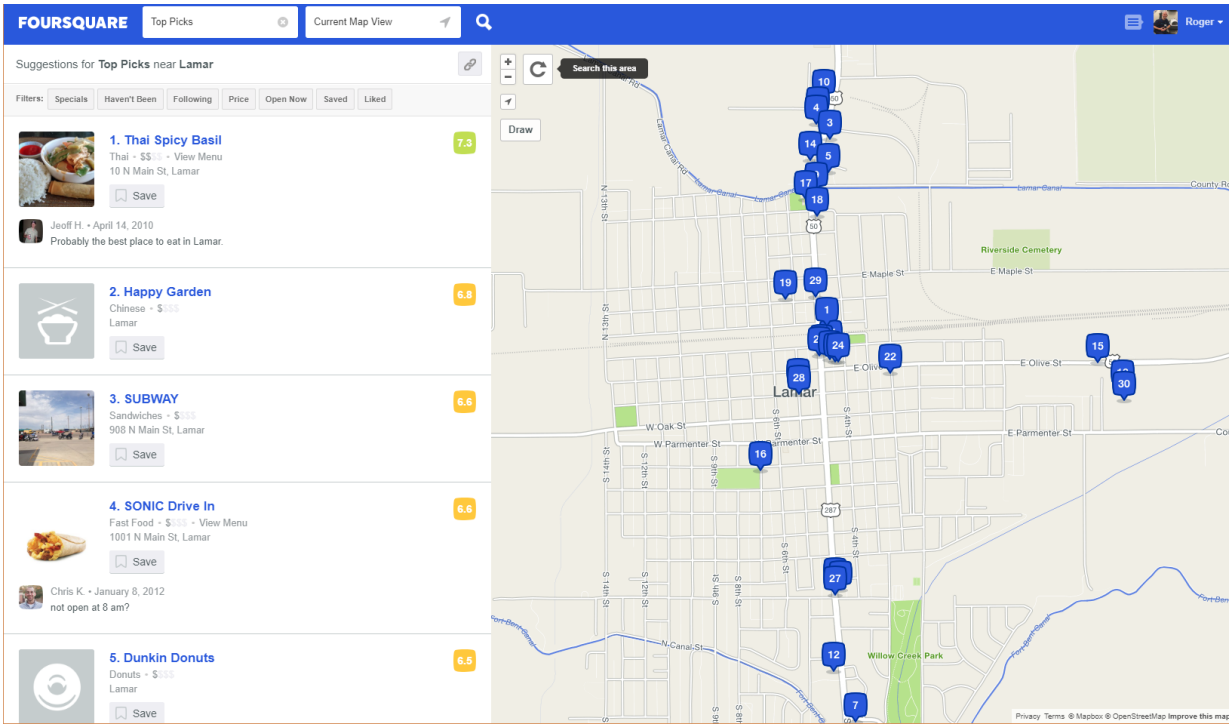


Figure 5. Four Square venues in Lamar using web browser interface

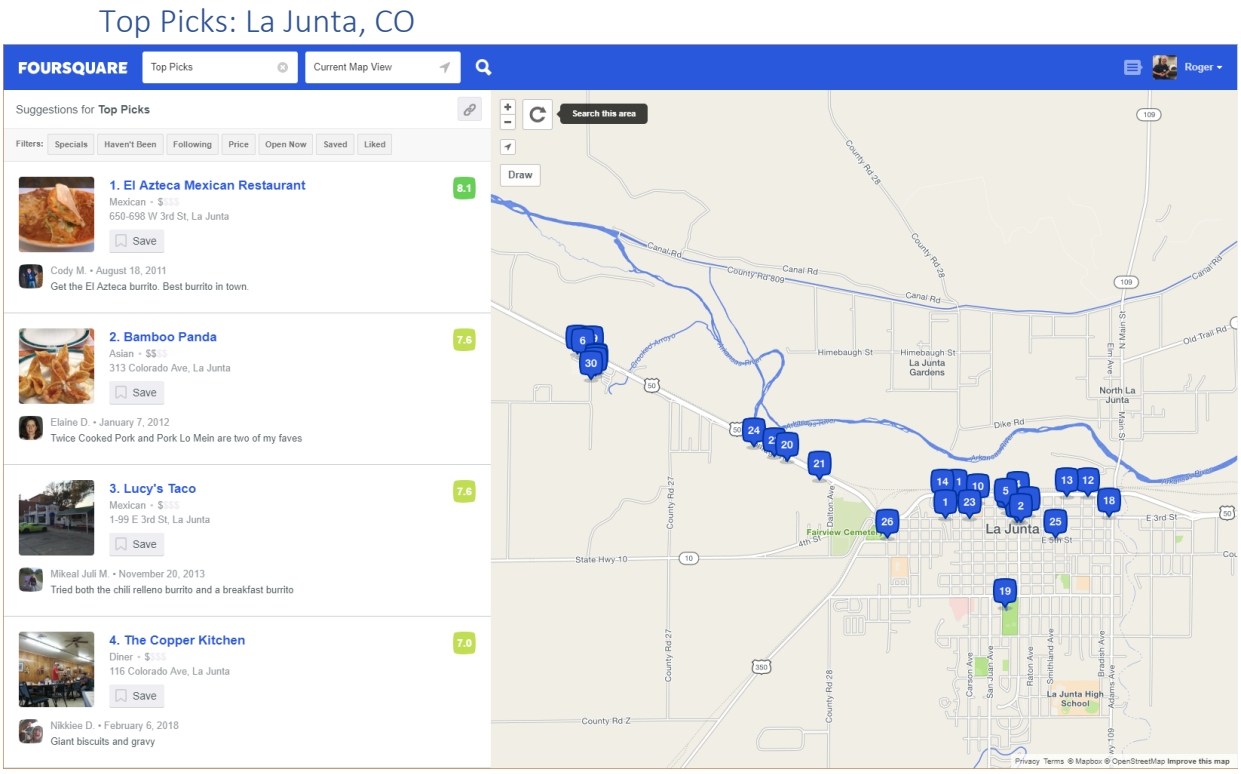


Figure 6. Four Square venues in La Junta using web browser interface

The information about these two towns will be compared using machine learning techniques in order to estimate quality of living in the two towns, to identify weaknesses or unmet needs in each town, and to recommend actions that might be taken by city leaders to stimulate growth, attract businesses, retain citizens, and improve the quality of living in the town.

Data Limitations

Throughout this course we have seen how extensive the databases are on major cities like New York and Toronto. It is also clear that managing and analyzing this volume of data is significantly aided by machine learning techniques. However, within the United States there are only 10 cities with a population of over one million. There are over 16,000 with a population of under 10,000. Therefore, it is important to understand to what degree the data and techniques from this course can be applied to these smaller communities.

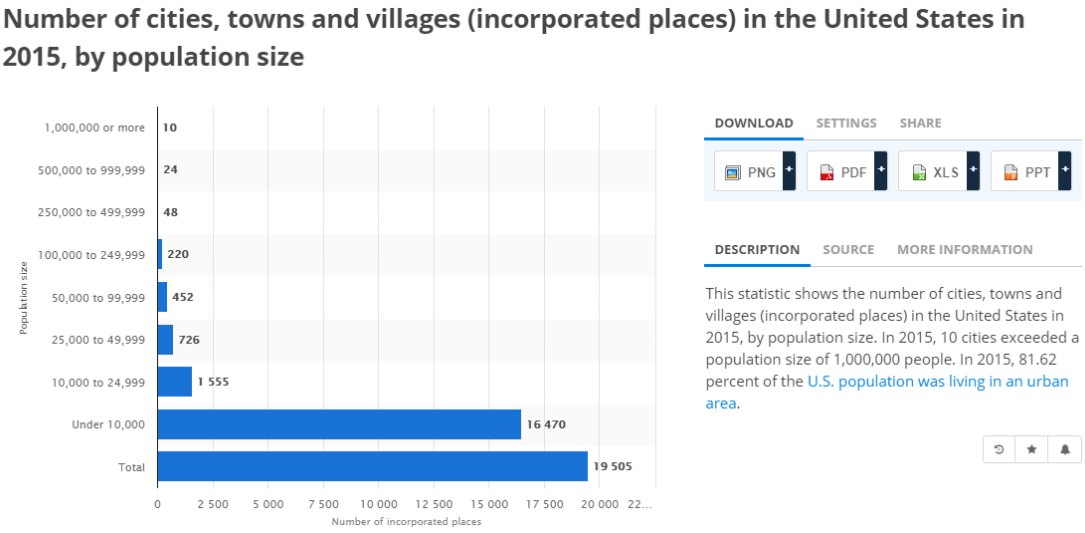


Figure 7. Sizes of cities and towns in the USA.

Image: from https://www.statista.com/statistics/241695/number-of-us-cities-towns-villages-by-population-size/

Methodology

The locations, populations, and numbers of businesses in the two towns are very similar. The objective is to identify the businesses districts that exist in both towns and to use this to make informed recommendations on where businesses might choose to locate.

The author is very familiar with the town of Lamar and very unfamiliar with La Junta. Therefore, data analysis, statistics, machine learning, and mapping tools are used to create a picture of Lamar which can be compared with the author's knowledge. This will reveal both the capabilities and the limitations of the techniques in generating understanding of the community. Then the same techniques will be applied to La Junta and should be able to reveal information that is unknown, but which will be comparable to the knowledge about Lamar.

Analysis of both cities used the following steps:

1. General a blank map of the city.
2. Select and plot the centroid for the search for venues in Four Square.
3. Identify the total number of venues for each town in Four Square.
4. Select 100 venues for analysis in each town.
5. Plot 100 venues on the map of the town.
6. Perform basis statistical analysis of the venues collected.

7. Generate basic histogram/bar charts of the types of businesses.
8. Apply K-Means cluster analysis machine learning to the venues.
9. Identify the various business districts (clusters) in both towns.
10. Draw conclusions about the structures of the businesses based on both the Four Square data, maps, and clusters.
11. Report conclusions and recommendations.

Throughout each of these steps various data were displayed to guide the analysis and to inform revisions and rerunning of the Python code.

Figure X provides comparatively scaled maps of both towns with the centroids plotted. In both cases, the street address of City Hall was selected as the centroid for the search of venues in Four Square. The address of city hall was found on each town's municipal web site.

The street address of city hall was converted to latitude and longitude coordinates using the Python geopy.geocoders library.

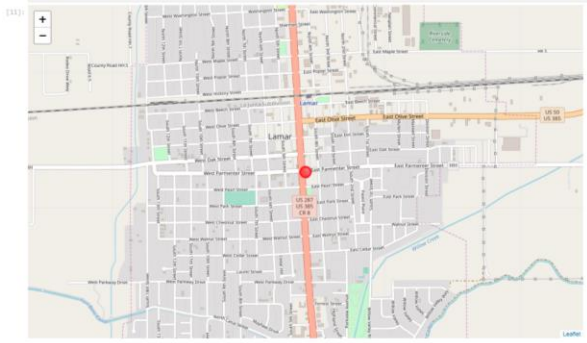
Lamar – City Hall	La Junta – City Hall
	

Figure 8. Comparative maps of both towns with centroids plotted.

Describe Exploratory Analysis

Multiple runs of the Python code were performed with variations in order to get a feel for the type, volume, and quality of data available in Four Square.

Radius of Search

The radius of the search was varied over several hundred meters to determine size of the area in which venues were located. We finally settled on a radius of 1600 meters, which is approximately one mile. Larger numbers do not add additional data.

Number of Venues

We learned that the Four Square database contained 103 venue records for the town of Lamar and 115 for the town of La Junta, surprisingly similar numbers. For the analysis we chose to limit the number of venues from each town to exactly 100. Initially we discovered that Four Square contains 62 and 61 unique categories for these businesses. This is far too many to make useful groups. Later you will see the smaller number of more general categories that we applied to the entries.

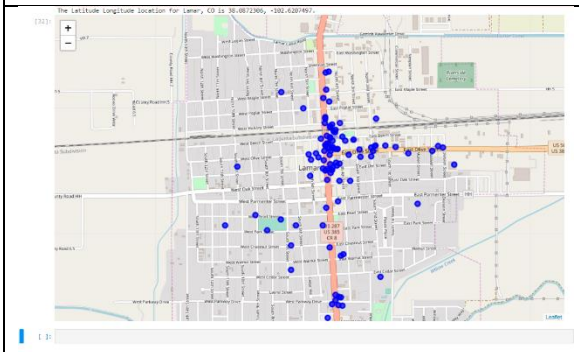
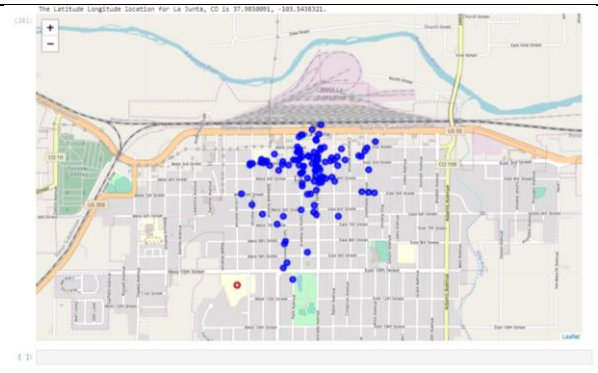
Lamar	La Junta
	

Figure 9. Initial plot of 100 venues in each town.

Errors in Four Square

During the exploration of the data from both towns we found what we believe to be an error in the storage of data in Four Square. One venue, specifically Jake's Gym, in La Junta is actually located in downtown La Junta near the City Hall. A Google search for the business gives its location within 5 city blocks, or 0.3 miles, of City Hall in La Junta. It is returned in the list of venues for the town and within the 1600 meter radius used.

However, the latitude and longitude coordinates returned by Four Square place the gym 11 miles away in the downtown area of Rocky Ford, CO. The "distance" field which is returned by Four Square for the distance between the venue and the centroid of search accurately identifies this distance as 16500 kilometers. Therefore, based on the Lat/Lon this business should not be returned by the search even though its true physical location is within the search area. We are not able to fully explain this behavior.

For purposes of this study, the correct latitude and longitude coordinates for Jake's Gym were found using Google and those coordinates were inserted into the Python Pandas dataframe so it could be included in the analyses.

Statistical Testing Performed

The venues in each town were collected into a small number of categories for comparison. These categories differed from those on the Four Square web site interface because the data set was much more diverse than the categories provided on the web site.

Since the two towns have approximately the same number of registered venues (Lamar: 103, La Junta: 115), the categories are roughly the same size.

However, this categorization revealed a number of errors or anomalies in the data stored in Four Square. These included:

1. Churches. Each town actually possesses several dozen churches. But only 3 and 4 (La Junta and Lamar, respectively) are registered for each town in Four Square.
2. Education. Most of the schools in both towns are not included in the Four Square databases.
3. Wrong Category. A coffee shop named "The Brew House" is registered as a church.

4. Unusual Venues. In Lamar a historical statue known as “Madonna of the Prairie” is registered as a building. In La Junta a street intersection is registered as a building.

These kinds of errors emphasize that data science requires active intervention by the human scientist to check the quality of the data, to adjust the algorithms, and to make interpretations of the results.

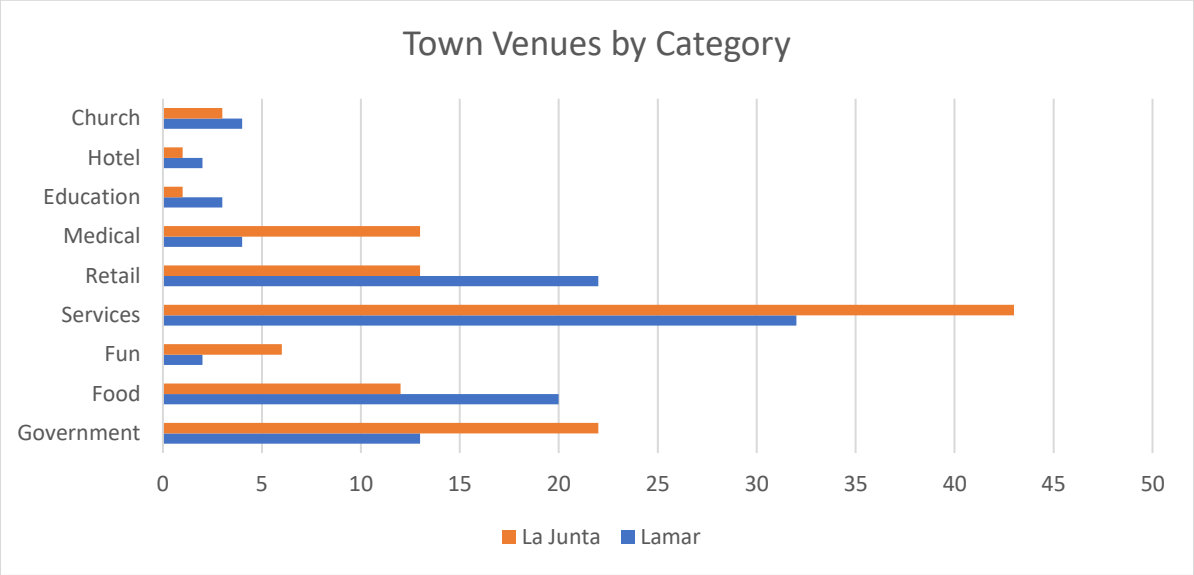


Figure 10. Venue category comparison.

Machine Learning Technique Used

We relied on the unsupervised learning technique, k-means clustering to organize the venues into business districts. This fit well with our goal to understand where customer traffic, complementary, and competitive businesses would be located.

Experiments were performed for the number of clusters to be used. Based on knowledge of the venues in Lamar, it was expected that at least four clusters would be needed. The code was run with clusters from 4 through 10 to determine which provided the most useful information. IN both towns, five clusters were found to be sufficient for the venues and locations.

Results

The results of this analysis reveal that both towns possess distinct business or venue districts which lie along the major corridors of traffic. Lamar’s clusters generally follow state highway 285 and state highway 50. Highway 50 is also “Main Street” in this town. Lamar possesses five clusters, four of which follow these highways. The fifth cluster is made up of multiple individual venues which are usually schools, stadiums, and small grocery stores embedded in neighborhoods in the southwest section of town.

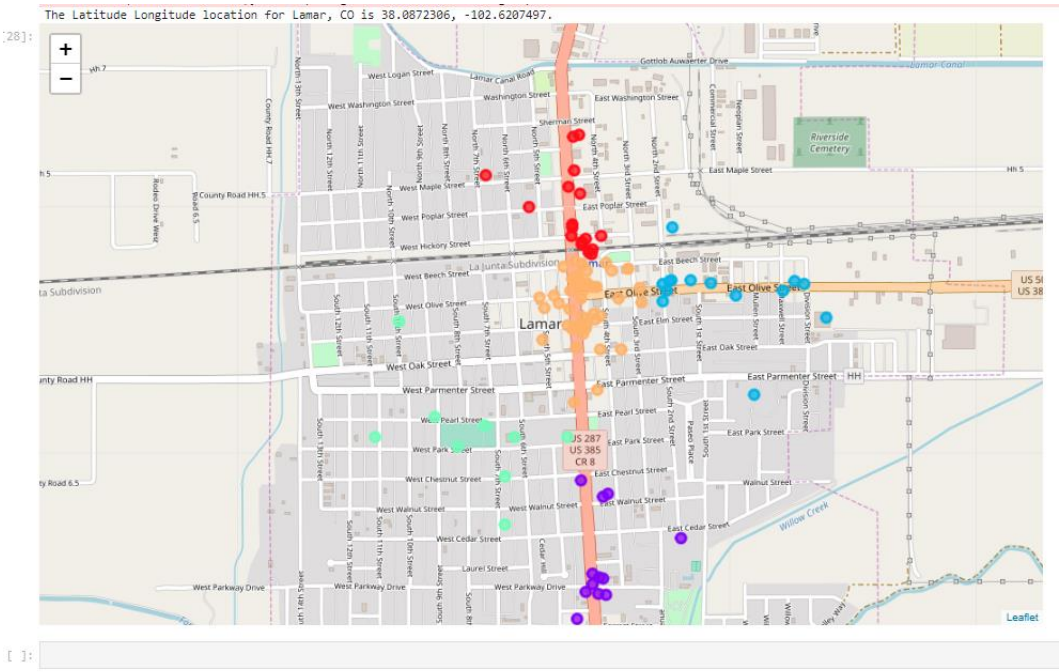


Figure 11. Venue clusters in Lamar, CO

La Junta has a similar pattern f just four venue districts. Three of these are aligned with highway 50 and Colorado Avenue. The former is known as “First Street” and the latter is effectively “Main Street” but does not use that name. The fourth district is also composed of schools, stadiums, and small convenience stores in the mid-southern region of the town. La Junta is bounded to the north by a large railroad yard and a river. These features have resulted in a town in which almost all living and working areas are south of this barrier. Lamar, on the other hand has a much smaller railroad yard which runs through the middle of the town, so venues exist in all directions from it.

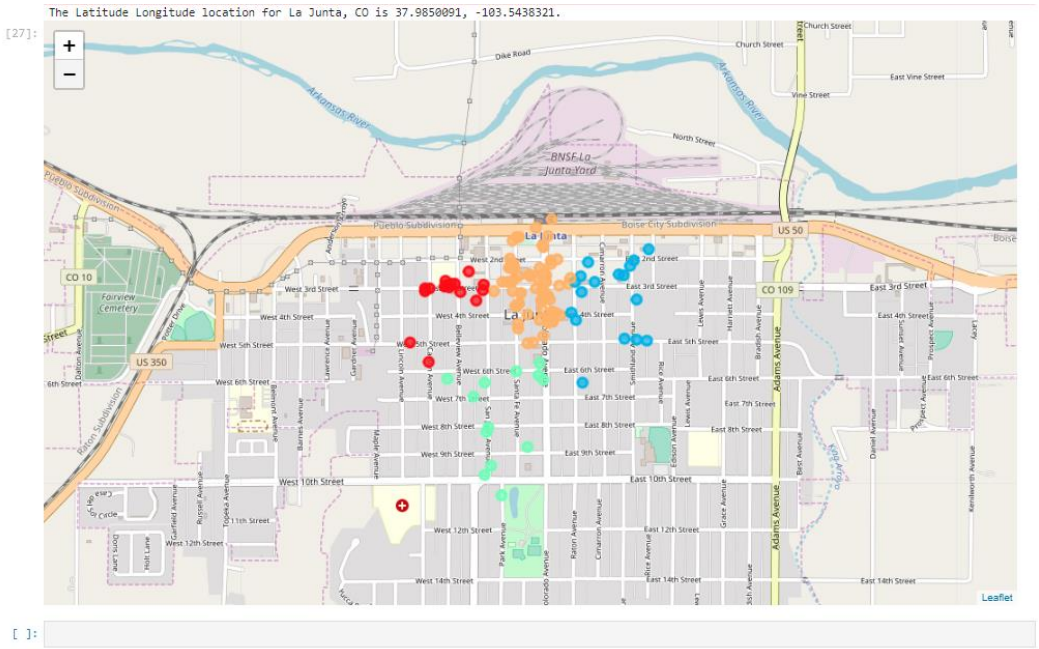


Figure 12. Venue clusters in La Junta, CO

Discussion

This analysis showed that the social and business venues in the two towns is very similar. They have roughly equal number of venues and similar numbers in each category. The clusters of business districts are similar, though influenced uniquely by the presence of large geographical features. These venues themselves would not appear to provide specific incentive for a business or a family to locate in one town of the other. The geographical maps indicate that La Junta is better suited for businesses that rely on the railroad. Lamar is better suited for business that can take advantage of semi-tractor or long-haul shipping because it is at the intersection of two state highways, as opposed to one in La Junta.

What did I notice?

Based on the analysis performed I learned that the power of databases like Four Square and tools like Python and machine learning can be very helpful in studying cities, towns, and neighborhoods. However, they cannot perform all of the necessary analysis automatically. For example, in this case, the digital data did not indicate the importance of the railroad yard of the highways. These features were evident to a human analyst who could reason on their impact.

We also learned that real-world data may contain error that cannot be handled well by automatic analysis. These errors must be addressed by the human analyst.

Recommendations

The recommendations from this study are that most businesses could perform equally well in either town. La Junta provides a better railroad system, while Lamar provides a better highway system.

Conclusion

This type of small town comparison can be performed for hundreds or even thousands of small towns across the USA and other countries. Since no one person can have a deep familiarity with this many towns, data science can be applied to create a beneficial understanding and comparison, so that decisions can be made across large datasets much more easily.

However, a data scientist should not expect the programming tools to do all of the work for her. She should be very diligent about finding anomalies in the data and observing information that is difficult for databases to quantify and analyze.

Author's Note

The author grew up in this area and experienced the various rivalries between these two towns for business, agriculture, government offices, external investment, and in high school sports.