IDENTIFYING METROPOLITAN AREAS IN THE SOUTH EASTERN UNITED STATES SIMILAR TO SILICON VALLEY

TIMOTHY BRANNON

MARCH 2019

INTRODUCTION

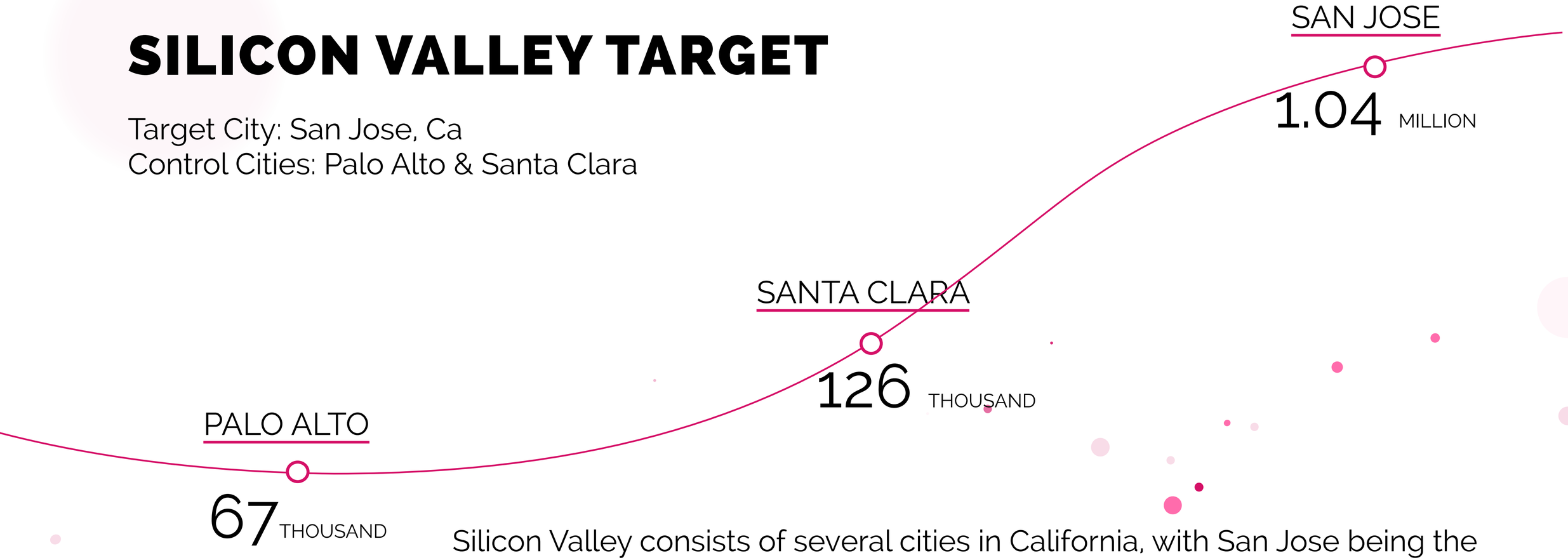
An upstart Information Technology company has grown rapidly and is looking to move into an office space that will serve as its Headquarters. The company is based in the South East and would like to identify metropolitan areas in the South Eastern United States as possible location candidates. The ideal location would be similar to Silicon Valley.

Data Question:

Which metropolitan areas in North Carolina, Virginia, Georgia, South Carolina and Eastern Tennessee are most comparable to Slicon Valley?

SILICON VALLEY TARGET

Target City: San Jose, Ca
Control Cities: Palo Alto & Santa Clara



Silicon Valley consists of several cities in California, with San Jose being the most populated. This was chosen as the target city of comparison. Santa Clara and Palo Alto were included as controls.

DATA STRATEGY

STEP ONE

Identify
Prospective Cities

STEP TWO

Compile
Demographic
data

STEP THREE

Compile Local
Venue data

STEP FOUR

Analyze, Review
& Feature Select

STEP FIVE

Scale & Preprocess
Feature Set

STEP SIX

K-Means Clustering
of Feature Sert

STEP SEVEN

Review Cluster
Groupings

STEP EIGHT

PROVIDE RECOMMENDATIONS

PROSPECTIVE LOCATION CITIES

21 Moderate to large cities were identified within the South Eastern Region of the United States as prospective location cities. They are as follows:

- Raleigh, NC
  Pop: 479,332
- Charlotte, NC
  Pop: 859,035
- Atlanta, GA
  Pop: 486,290
- Charleston, SC
  Pop: 138,036
- Winston Salem, NC
  Pop: 244,605
- Augusta, GA
  Pop: 197,166
- Columbia, SC
  Pop: 133,114
- Durham, NC
  Pop: 267,743
- Richmond, VA
  Pop: 223,170
- Norfolk, VA
  Pop: 244,703
- Alexandria, VA
  Pop: 160,035
- Florence, SC
  Pop: 38,317
- Savannah, GA
  Pop: 146,444
- Asheville, NC
  Pop: 91,902
- Roanoke, VA
  Pop: 99,837
- Knoxville, TN
  Pop: 187,347
- Greensboro, NC
  Pop: 290,222
- Cary, NC
  Pop: 165,904
- Athens, GA
  Pop: 125,691
- Albany, GA
  Pop: 73,801
- Chattanooga, TN
  Pop: 177,571

DATA SOURCES

Information for Target and Prospective cities is compiled by webscraping source html code, and API quiries with the python request library from the sources below.

- WORLD POPULATION REVIEW
  Web scraped HTML resource providing Population Denisty information
  www.worldpopulationreview.com
- WIKIPEDIA
  Web scraped HTML resource providing Population information
  www.wikipedia.com
- AREA VIBES
  Web scraped HTML resource providing General Demographic data
  www.areavibes.com
- FOUR SQUARE
  Venue Search API based on Latitude/Longitude Location
  www.foursquare.com

COMPILED DATA

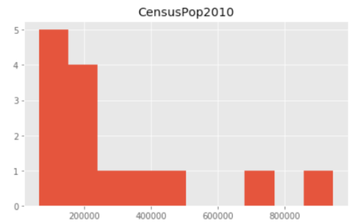
Webscraping and API queries produces a list of compiled demographic and venue category informaiton for each city. Compiled variable data points are displayed below:

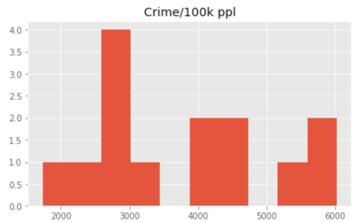
WORLD POPULATION REVIEW	WIKIPEDIA	AREA VIBES		FOUR SQUARE
POPULATION DENSITY km/sq	LATITUDE	CRIME PER 100 K	MEDIAN INCOME	VENUE NAME
	LONGITUDE	VIOLENT CRIME PER 100 K	UNEMPLOYMENT RATE	VENUE LATITUDE
	ESTIMATE POPULATION	PROPERTY CRIME PER 100K	POVERTY LEVEL	VENUE LONGITUDE
	2010 CENSUS POPULATION	INCOME PER CAPITA		VENUE CATEGORY

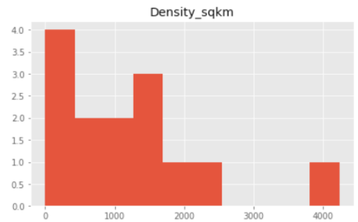
DATA DISTRIBUTION ANALYSIS

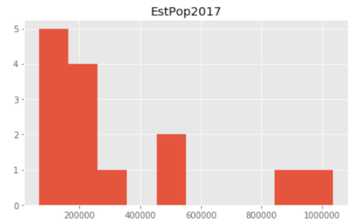
Review compiled demographic variable data distribution for target, control and prospective cities to better understand the structure of compiled data.

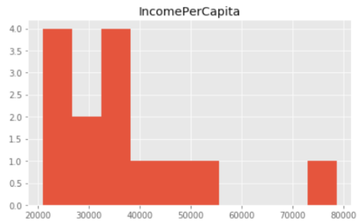
The set of compiled variables show a diverse set of distributions with some uniform, right and left skewed distributions. Where Population and Med Income were the most skewed.

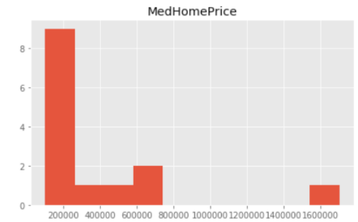


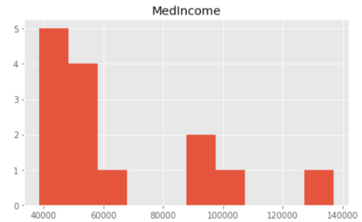


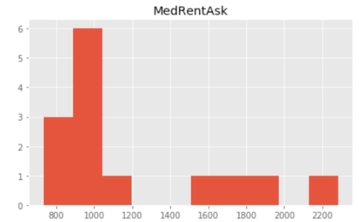


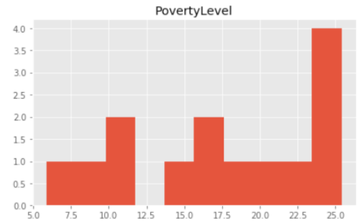


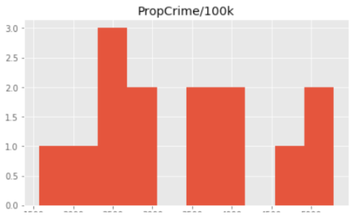


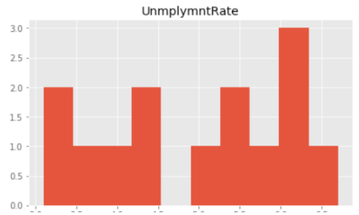


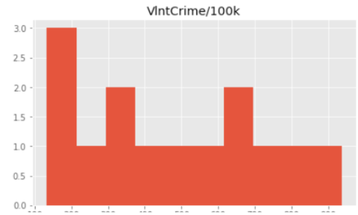












CORRELATION ANALYSIS

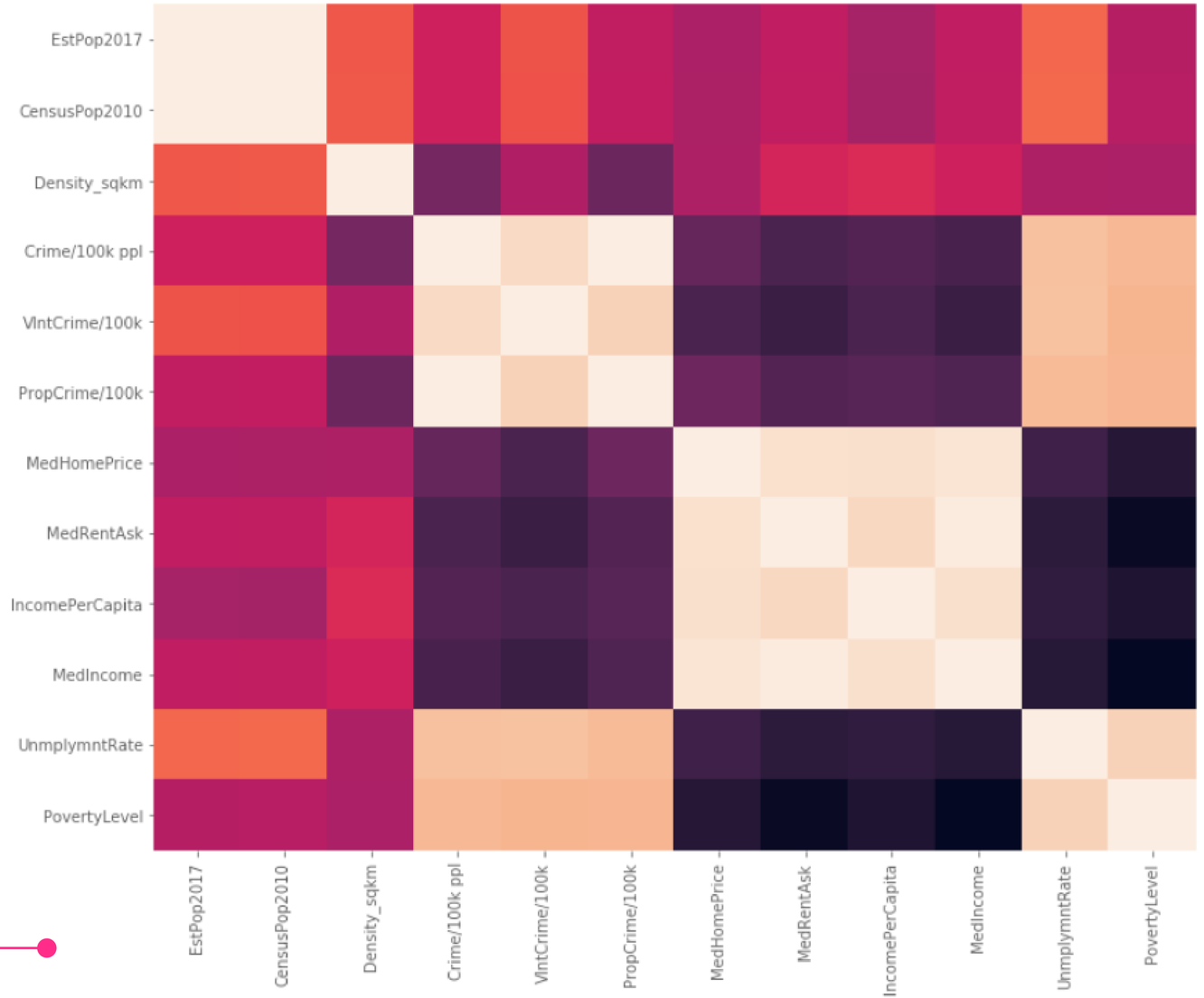
Review compiled demographic variable correlation to make feature determinations.

Strong Positive Correlations:

- Population, and 2010 Census Population
- Crime per 100k people & Violent Crime per 100k/ Property Crime per 100k

Strong Negative Correlations:

- Poverty Level and Median Income
- Poverty Level and Median Rent Ask



FEATURE SELECTION

Review of visualization and analysis led to Estimate Population and Census Population to be replaced by Growth Rate. Crime per 100k was removed from the data set. Selected demographic features are scaled with the min/max standard scaler and merged with venue category data.

GROWTH RATE	INCOME PER CAPITA	VENUE CATEGORY AVERAGE
POPULATION DENSITY km/sq	MEDIAN INCOME	
VIOLENT CRIME PER 100 K	UNEMPLOYMENT RATE	
PROPERTY CRIME PER 100K	POVERTY LEVEL	

Growth Rate:

Derived by subtracting 2010 Census Population from Estimate Population, and dividing the value by Estimate Population

Crime p/100k: Removed

Venue Category Average:

Venues are Onehot encoded for the Category variable, The sum of onehot category count is divided by total venues. The count proportion is derived for each category

K-MEANS CLUSTERING

Given the small sample set cluster size is set to 3 distinct groupings with a random starting point value of 0. The K-Means Clustering algorithmn assigned groupings as follows:

Group 0

San Jose, CA

Palo Alto, CA

Santa Clara, CA

Alexandria, VA

Cary, NC

Group 1

Columbia, SC

Augusta, GA

Florence, SC

Knoxville, TN

Albany, GA

Chattanooga, TN

Group 2

Raleigh, NC

Charlotte, NC

Atlanta, GA

Winston Salem, NC

Durham, NC

Richmond, VA

Norfolk, VA

Savannah, GA

Asheville, NC

Roanoke, VA

Greensboro, NC

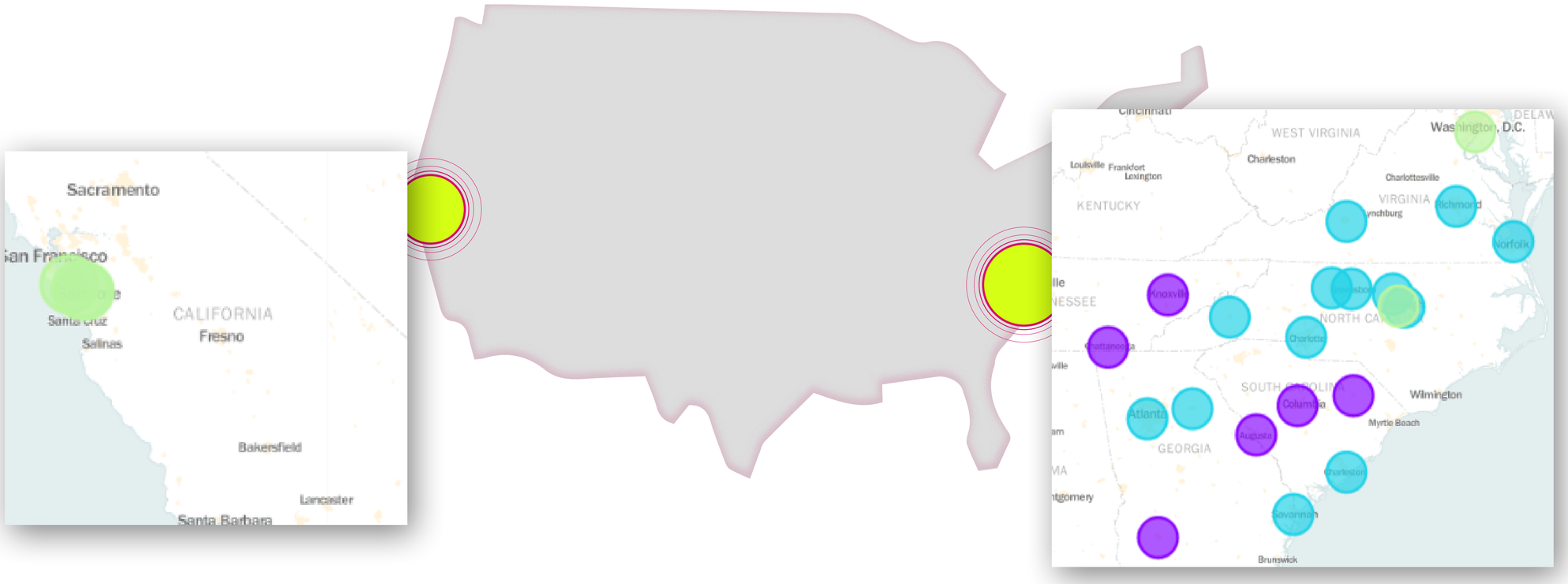
Athens, GA

12.

Target Group Clusters

Target Group Clustering:
Cluster Assignments are visualized with the Folium map plotting library.

Prospective Location Reccomendation:
Cary, NC or Alexandria, VA



RECOMMENDATIONS

With all features and venue catgeroies considered the Cary, NC and Alexandria, VA cities would be the recommendation for prospective locations similar to Silicon Valley.

Note: Additional iterations of the K-means algorithmn were conducted to test the recommended results with 2 & 4 possible groupings. in both instances Cary, NC and Alendria, VA were clustered with the Target City.

CONSIDERATIONS

- Control Cities with varying demographics were also grouped with the Target city and Recommended Cities.
- Given the small size of the sample set K-means was set to group cities into 3 clusters as the ideal scenario
- There are a large number of additional demographic data points that could also be included that may affect clustering results.
- Feature selection and scale strategy may affect clustering results.

14.

CONTACT

TIMOTHY BRANNON

Raleigh, NC 27610 | brannon.timothy@gmail.com