FINDING CITIES SIMILAR TO BOISE

ISAAC TRUSSELL

BACKGROUND

I have been looking into moving from Boise, Idaho for professional and personal reasons. I am sad at the thought of leaving Boise because I really like it. This made me think that I could run an analysis of cities across the US and compare them to Boise. This will give me some ideas on some options that are similar to Boise and look into the possibility of moving there.

DATA ACQUSITION

Population Data:

- https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population
- This contains general information about population, population density, and square mileage of a city

Climate:

- https://www.infoplease.com/math-science/weather/climate-of-100-selected-us-cities
- Has information about average temperature in summer, fall, winter and spring as well as the precipitation in cities across the US

Venue:

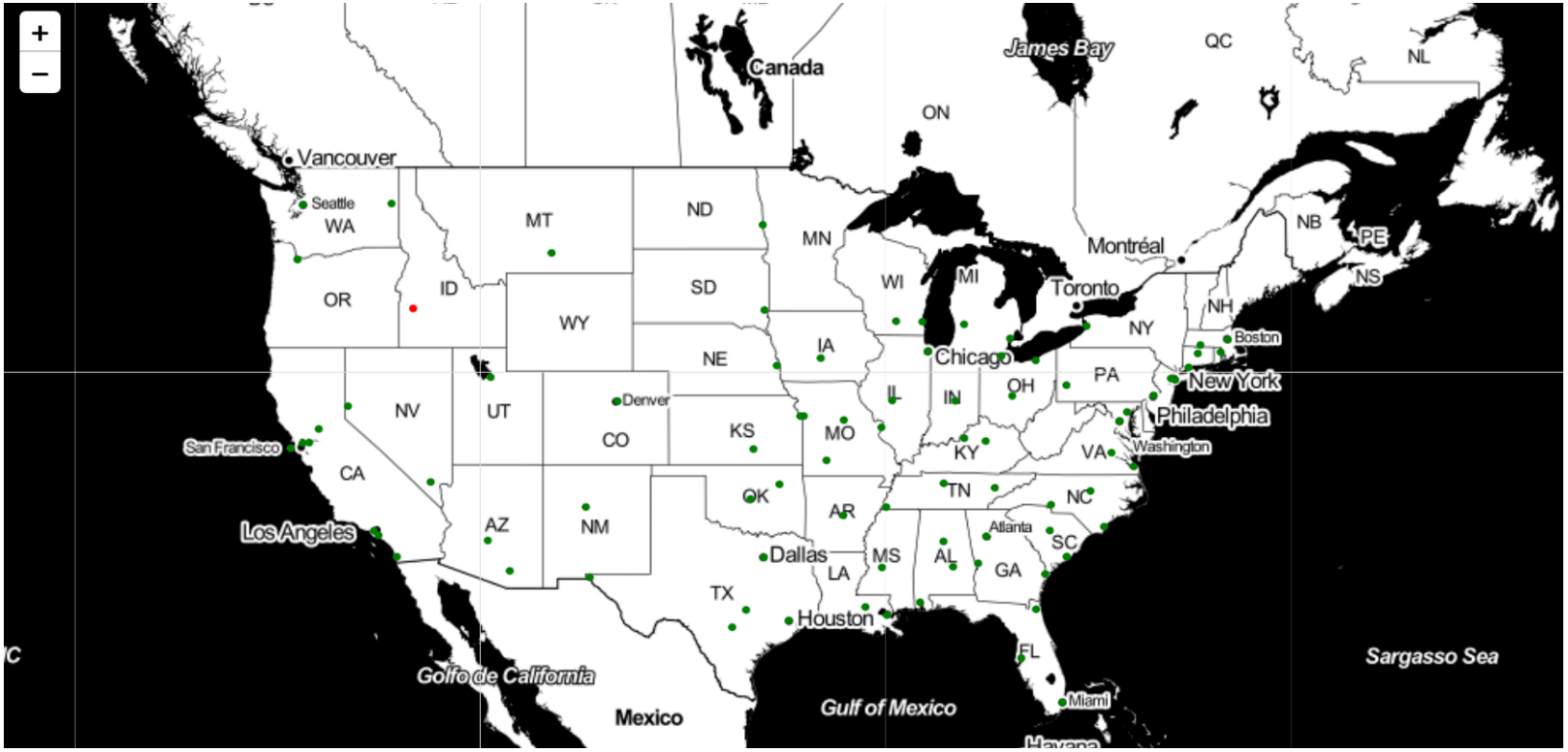
- Used the Foursquare API to get information about popular venues in the area.

DATA CLEANING AND PREP

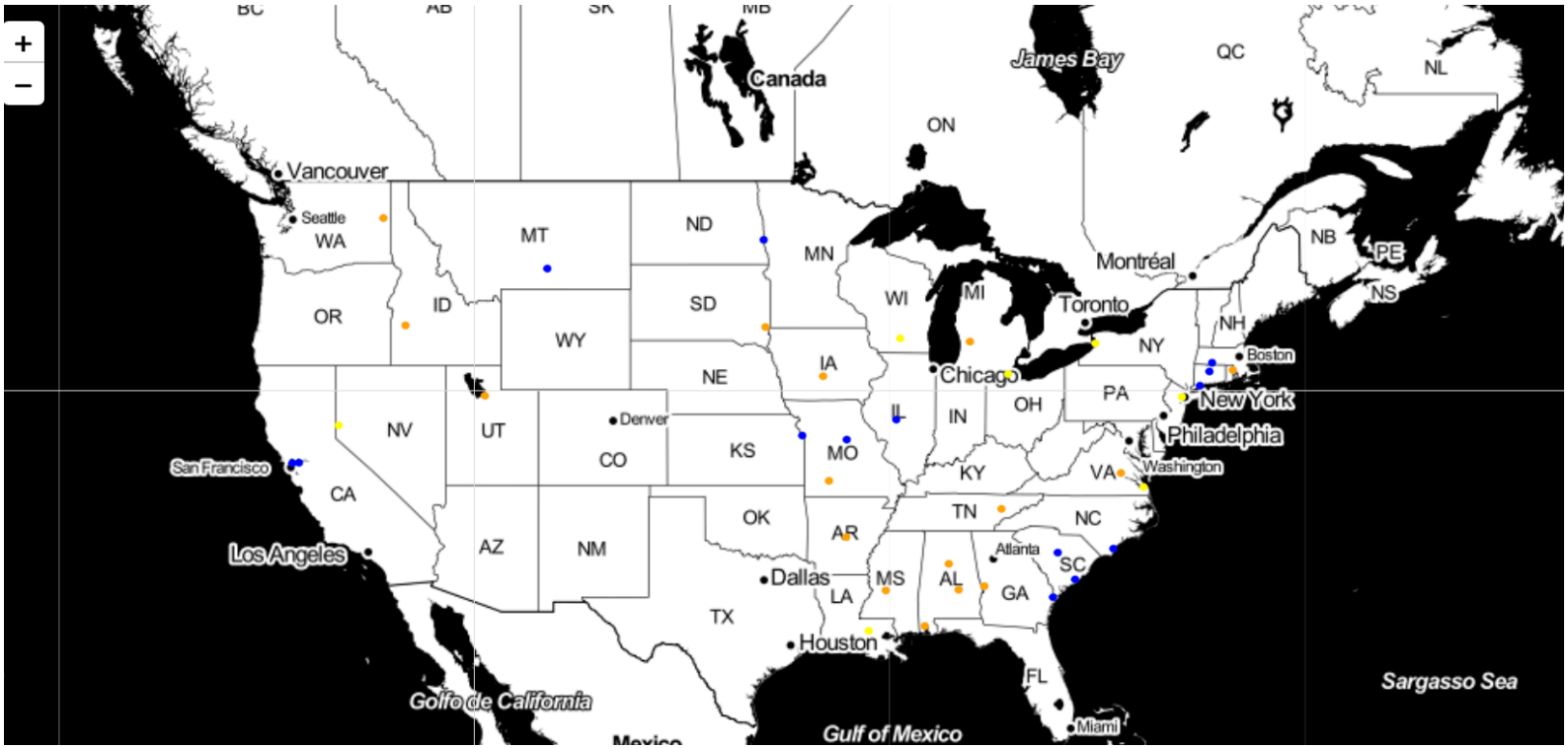
A lot of the data came in with units attached, therefore I had to remove a lot of characters on columns as well as change them from object dtypes to something that could be usable later by sklearn.

The climate dataset had fewer cities than that of the population dataset so I was forced to drop some cities. I used the .join method to combine the 2 datasets and then used the .dropna method to remove all the rows that contained missing data.

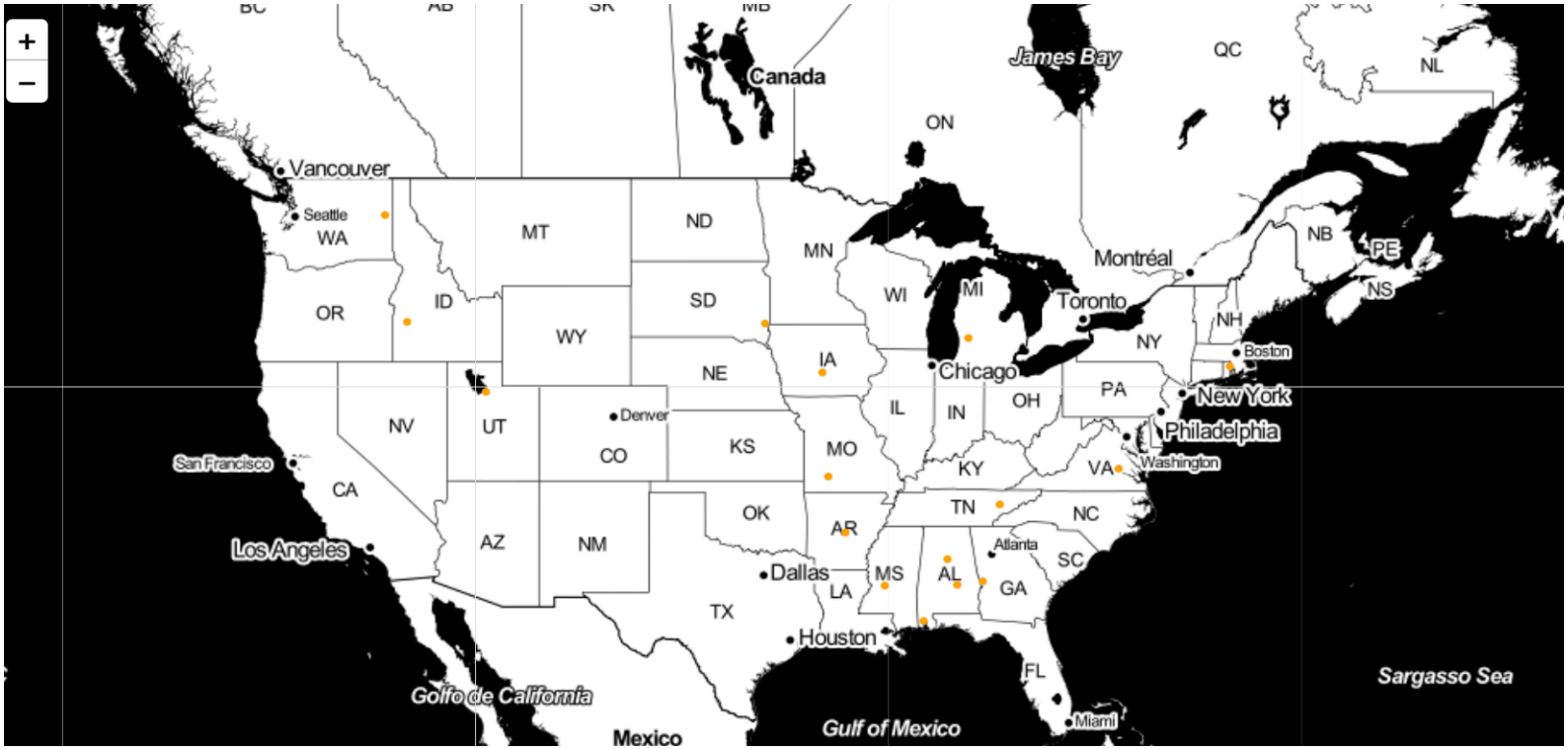
MAP OF ALL CITIES INCLUDED IN ANALYSIS



CITIES SEGMENTED INTO 7 GROUPS



CITIES SIMILAR TO BOISE CLUSTERED INTO 3 GROUPS



CITIES GROUPED IN WITH BOISE

- Richmond
- Spokane
- Des Moines
- Birmingham
- Salt Lake City
- Grand Rapids
- Montgomery
- Little Rock
- Colombus
- Mobile
- Knoxville
- Sioux Falls
- Providence
- Springfield
- Jackson

INTERPRETATION

Looking at the final grouped dataset, population seems to be the only column that has any pattern to it to me. Overall population seems to be the biggest factor in deciding the clusters of cities. The final dataset only has one row missing in the range of 55 – 71. I would be interested in doing a deeper analysis on how much each column contributed to the final results of the model. I could potentially have used a principal component analysis (PCA) to do this. This might help decide which features could be dropped and help create a cleaner model.