Kerry Kurcz

Professor Boudourides

MSDS 430

December 8, 2019

Introduction

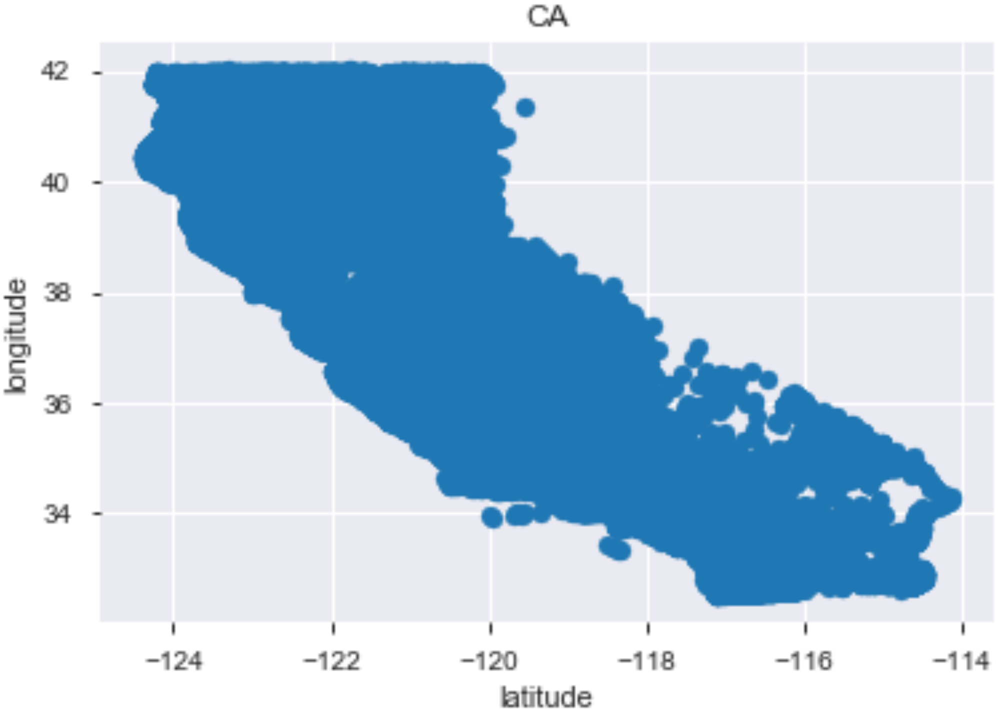
I grew up where the Woolsey fires hit California last year. It is significant to me hearing about more fires happening in the state around this time of year. Given a dataset from across the entire country of about 1.9 million fires collectively between the years 1992 and 2015 provided by Kaggle, I wanted to see the distribution of fires across the United States and how they compared by state and region according to size, year, and degree of damage.

The variables used mostly in this analysis include State (STATE), name (FIPS_NAME), size (FIRE_SIZE), statistical cause code (STAT_CAUSE_CODE), year (FIRE_YEAR), discovery day of the year (between 1 and 365; DISCOVERY_DOY), latitude (LATITUDE), and longitude (LONGITUDE). I also add two columns, region (REGION) and region code (REGION_CODE), where region code is the same as region but in numerical form for analysis. Region is based on the State column.

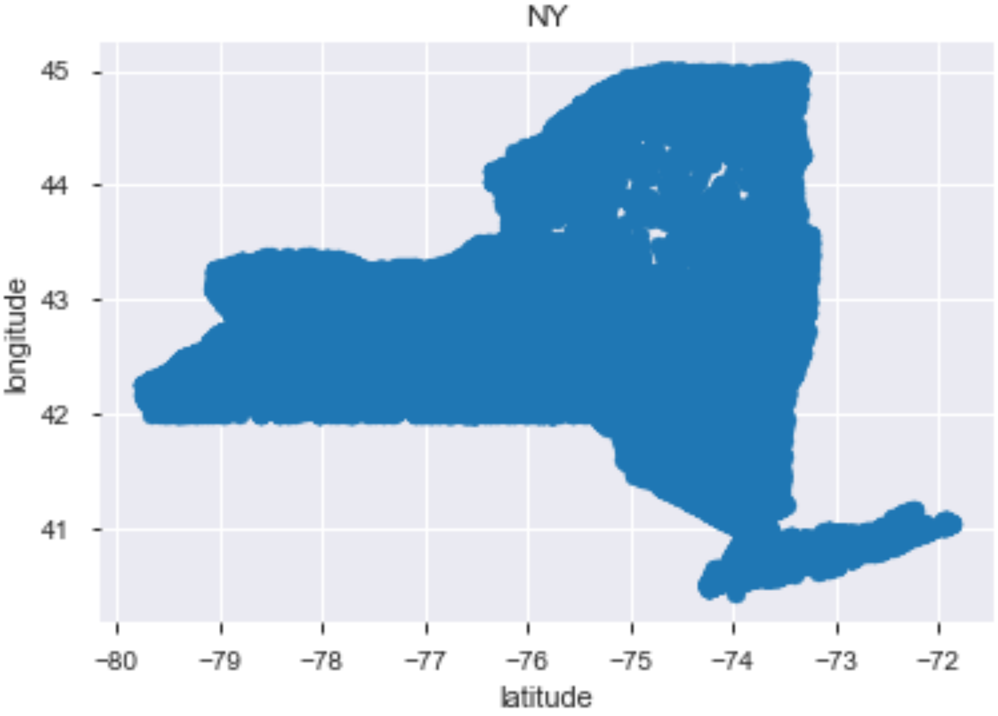
Algorithmic Process

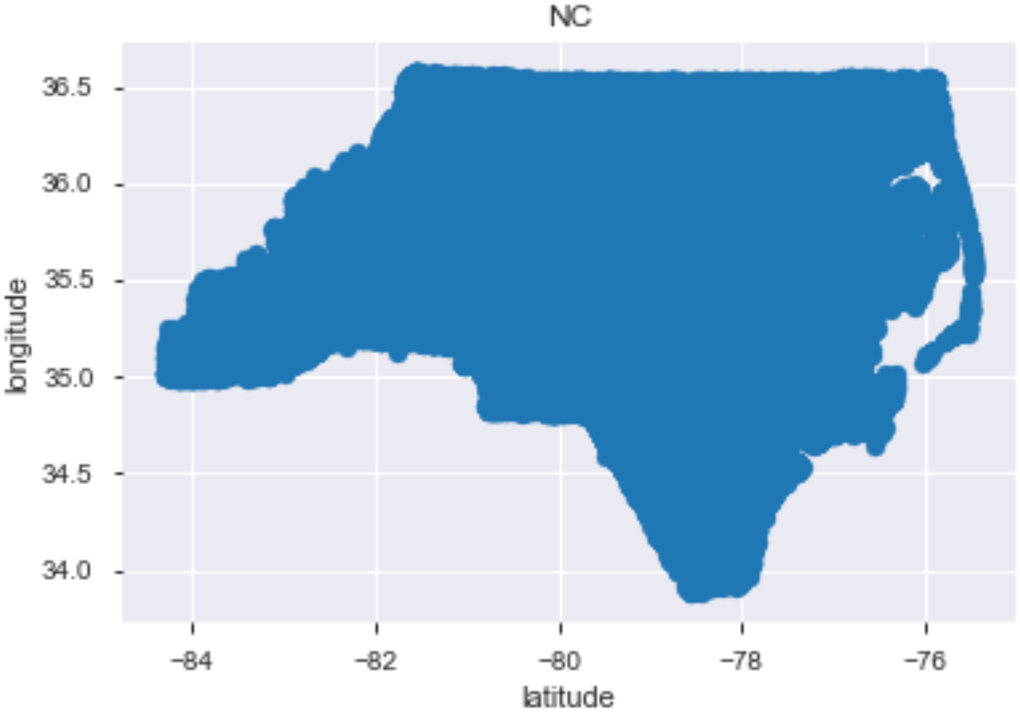
The dataset is given as a sqlite file, so I first import it to a pandas DataFrame. Then I begin with basic exploratory analysis of the dataset. The size, the variables, the index. I do this keeping in mind which variables might be the most useful for my analysis – and based on this, which analysis would be most appropriate.

While writing the code to output Pearson Correlation Matrix just for the state of California (initially my plan) using the seaborn package, I accidently included latitude and longitude and noticed that the plot output a perfect picture of the physical state of California.

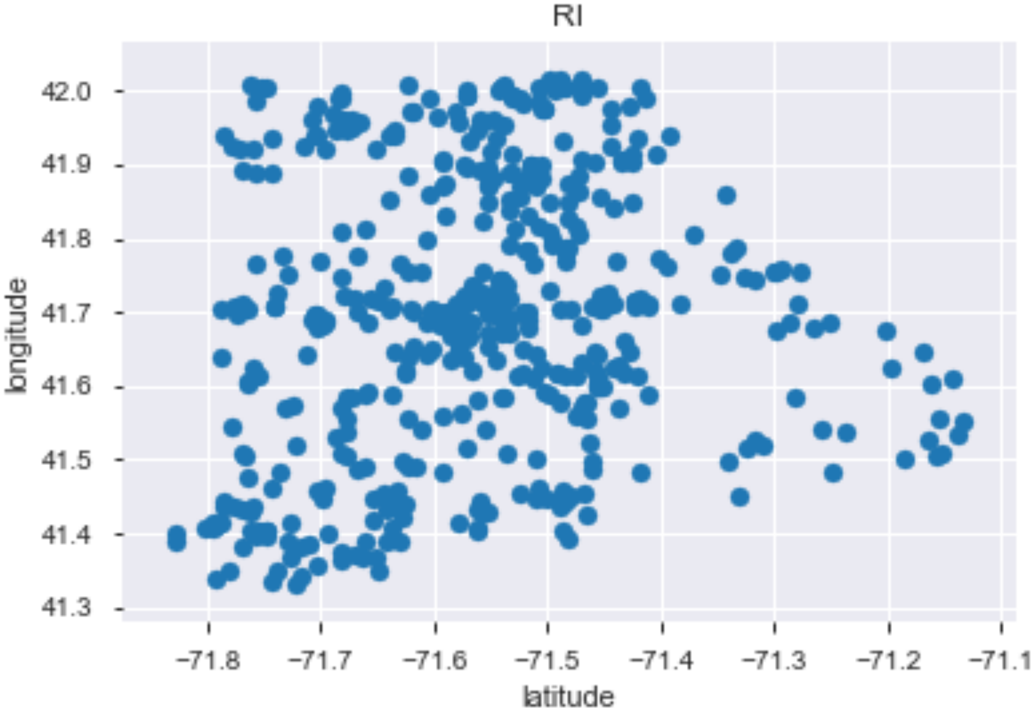


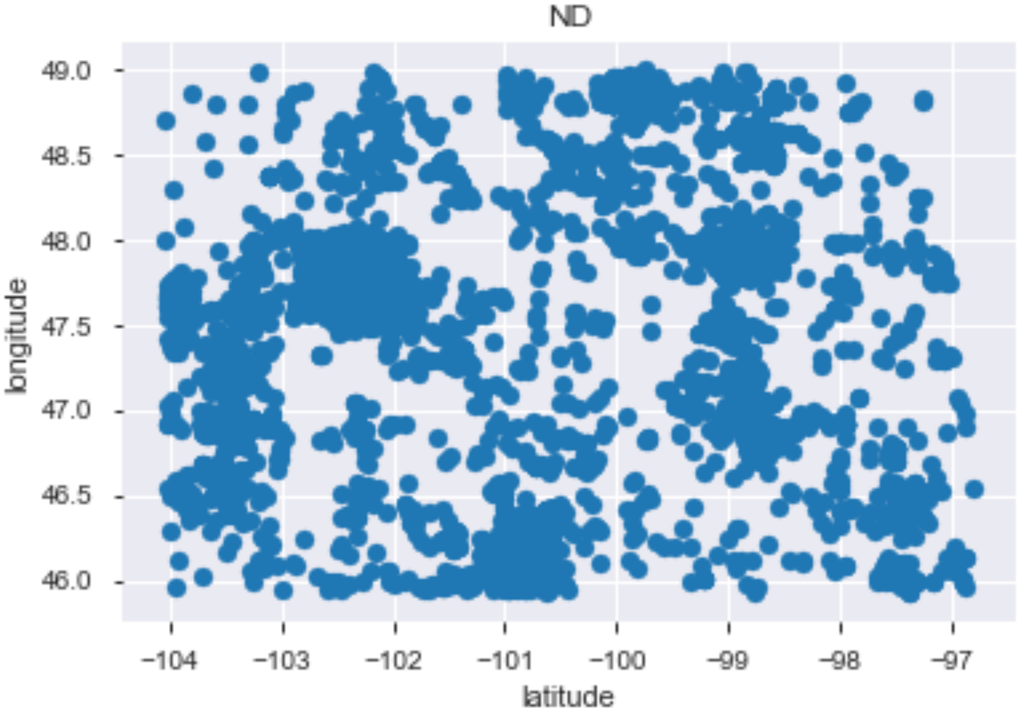
I thought this was cool, so I use the longitude and latitude coordinates to print out all of the states. Some, such as New York and North Carolina, came out very clear.



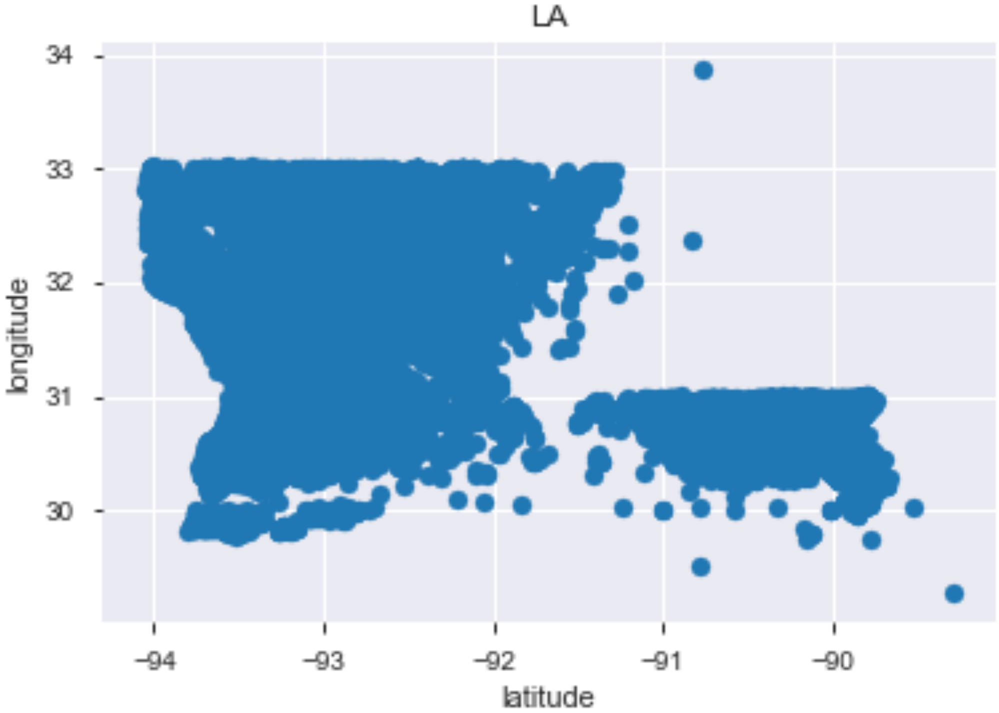


Other states, such as Rhode Island and North Dakota, did not come out clear at all. Is this because such states do not have many fires, or these states are not as good as reporting them?

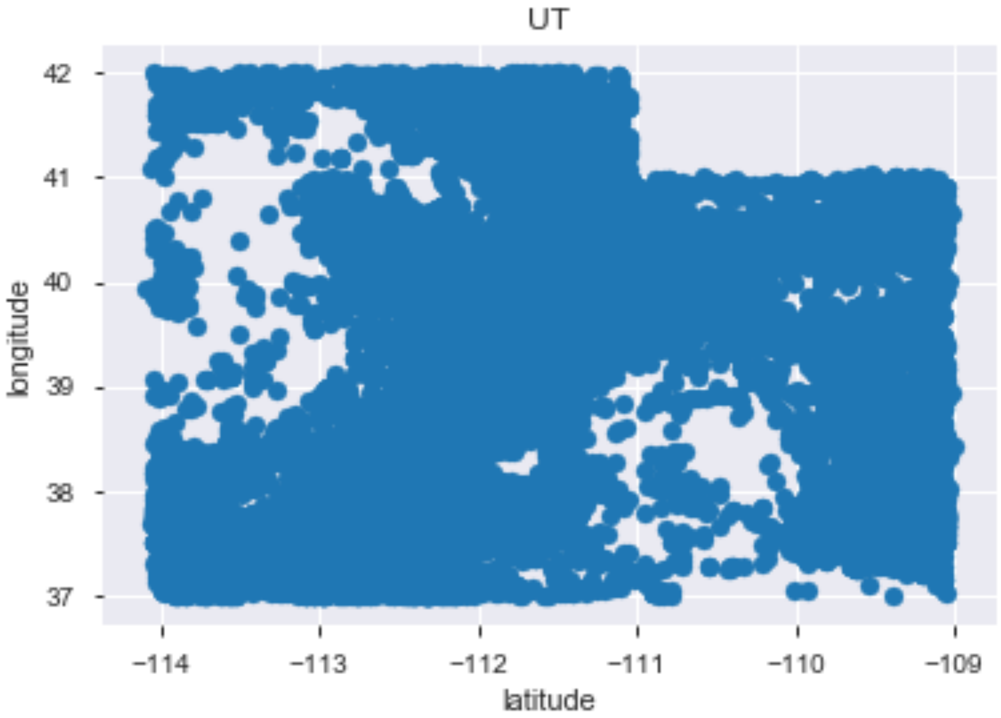




Lousiana also looks very interesting. Is it that the state is actually quite swamp-like toward it’s southern boarders, that it is impossible to be on fire?



That being said, take a look at Utah.

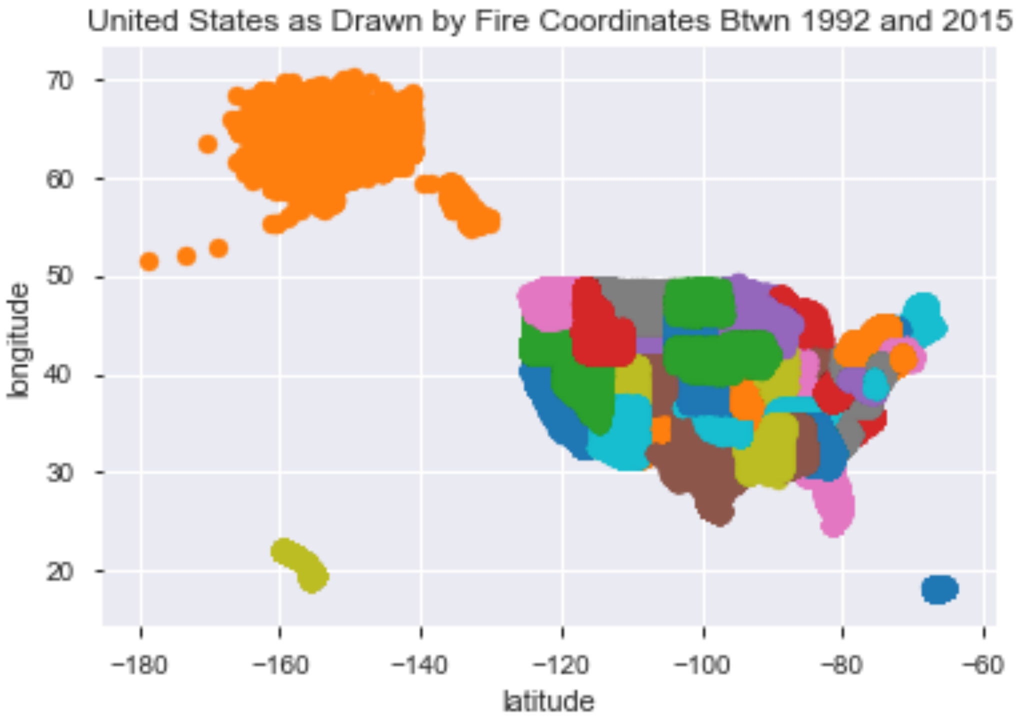


Now take a look at what Utah *really* looks like:

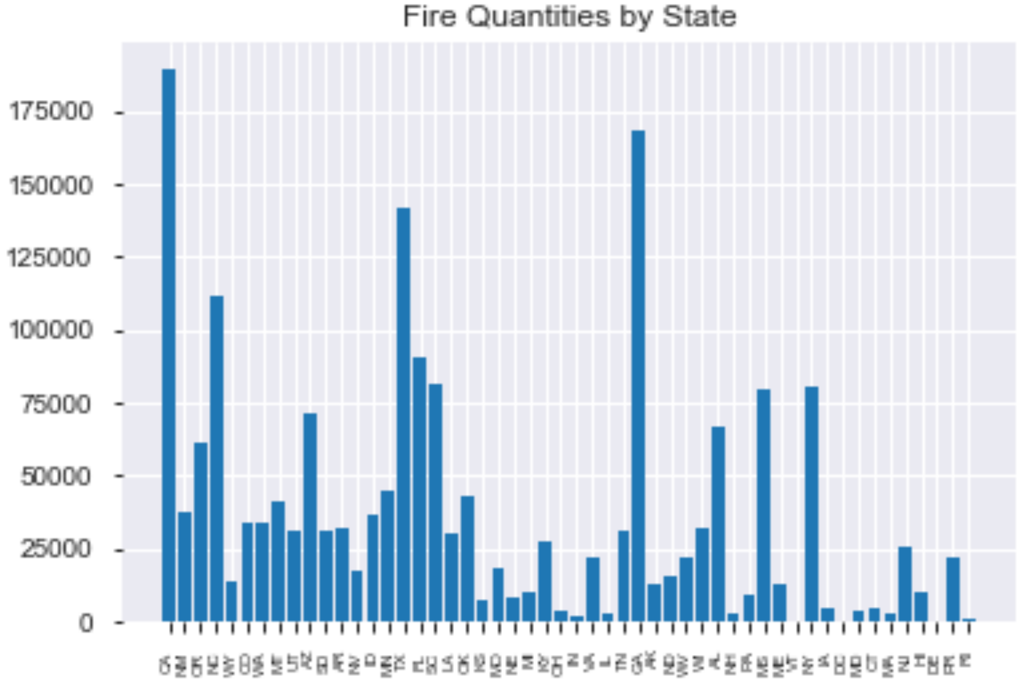
Ogden
Salt Lake City
Sandy
Provo
UINTAH AND OURAY RESERVATION
UTAH
Granc
Junctic
St. George

It looks as if the coordinates were input backwards.

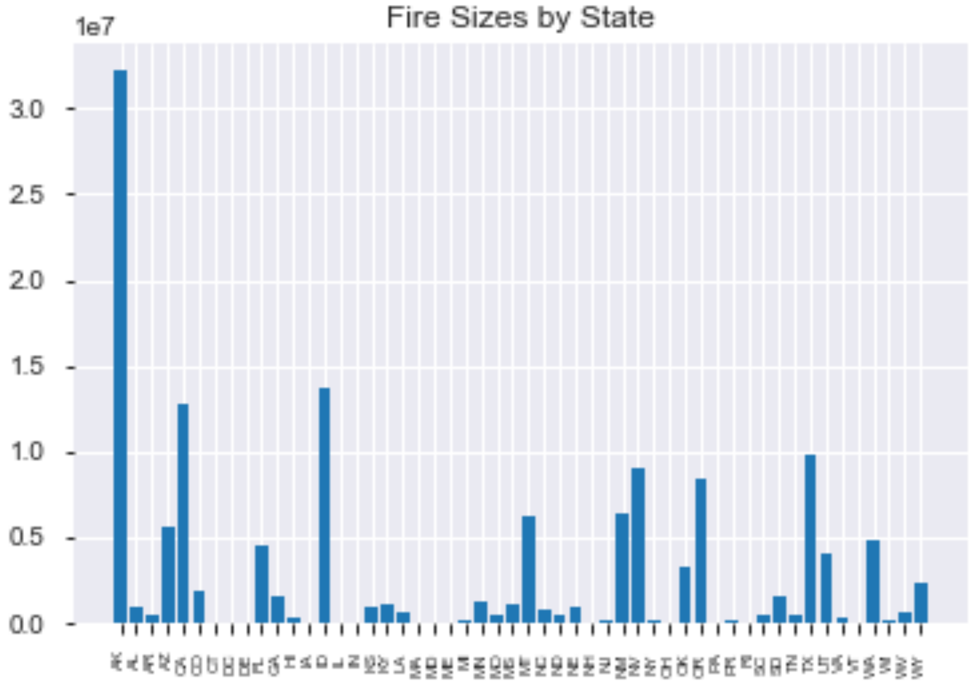
I also used the entire dataset to draw the United States as drawn by fire coordinates. It kind of looks like a child's drawing.



I am sure if I graphed quantities of fires, larger states such as California and Texas would be quite high, and smaller states (especially Rhode Island, since we saw it did not really have a lot of data points visually) would be quite low. This turns out to be quite true. In fact, California has the highest number of fires in all.

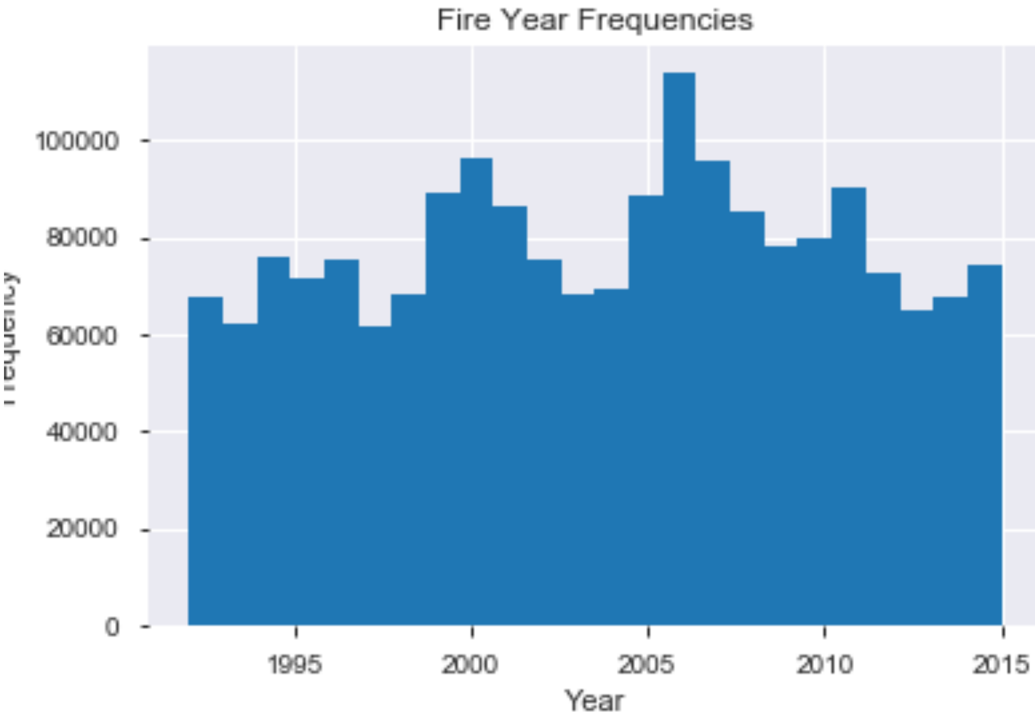


How do these fires compare by size? Interestingly, Alaska has it the worst.

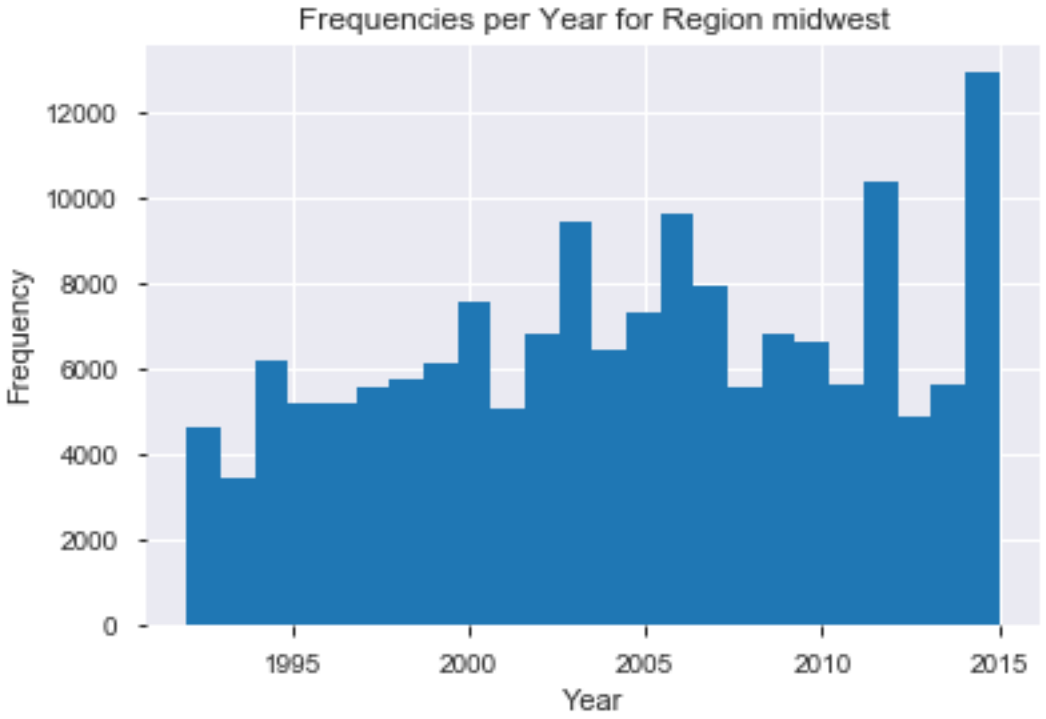


While California has the highest number of fires, it is only third highest in the amount of ground those fires cause. Still, I think Alaska, while having over 30,000,000 acres of ground covered by fire over the 13 years, did not experience as much turmoil as other states with more civilization such as California, Texas, and maybe even Idaho.

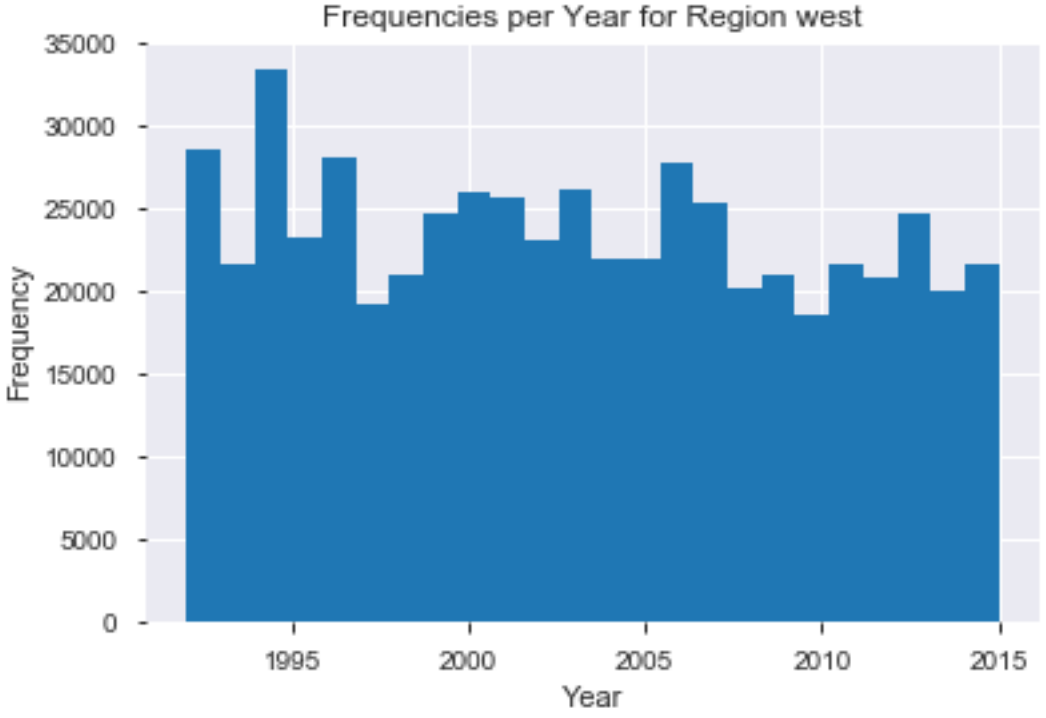
Fire frequencies by year was also interesting to look at. For all states, I see a slight increase over time. This could be simply due to better reporting mechanisms.



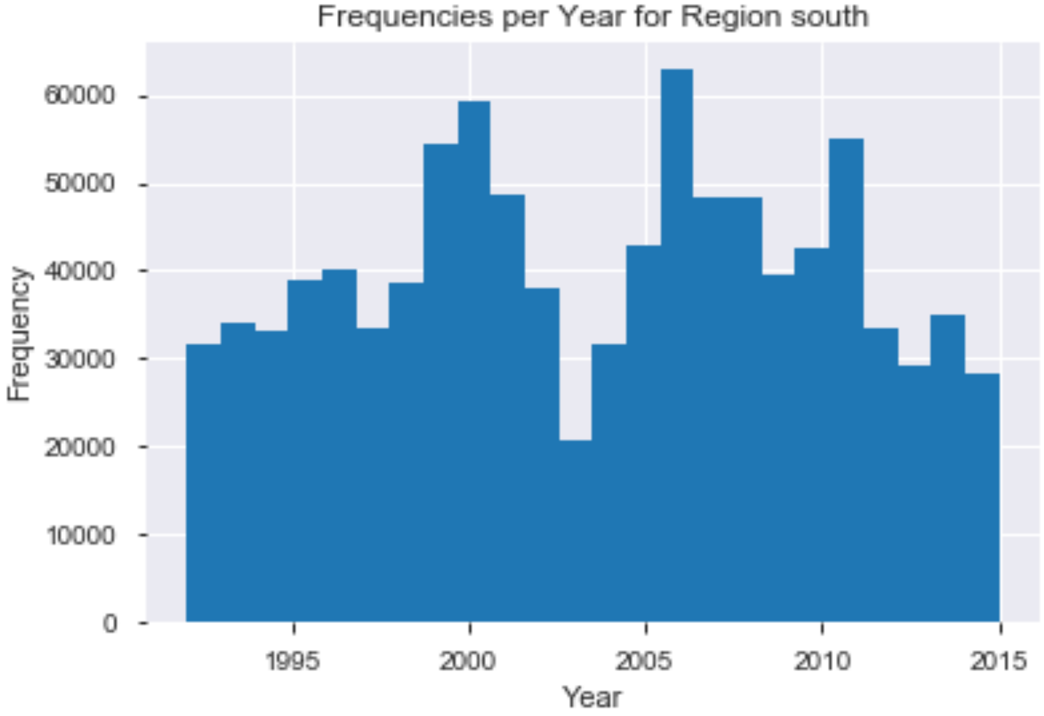
Anyway, partially because the states look so strange in the U.S. map shown earlier, and partially because 52 data points (for 50 states plus D.C. and Puerto Rico) is much more difficult to handle than 5 (for 5 regions, west, midwest, northeast, south, and other), I decided to do my analysis by region.

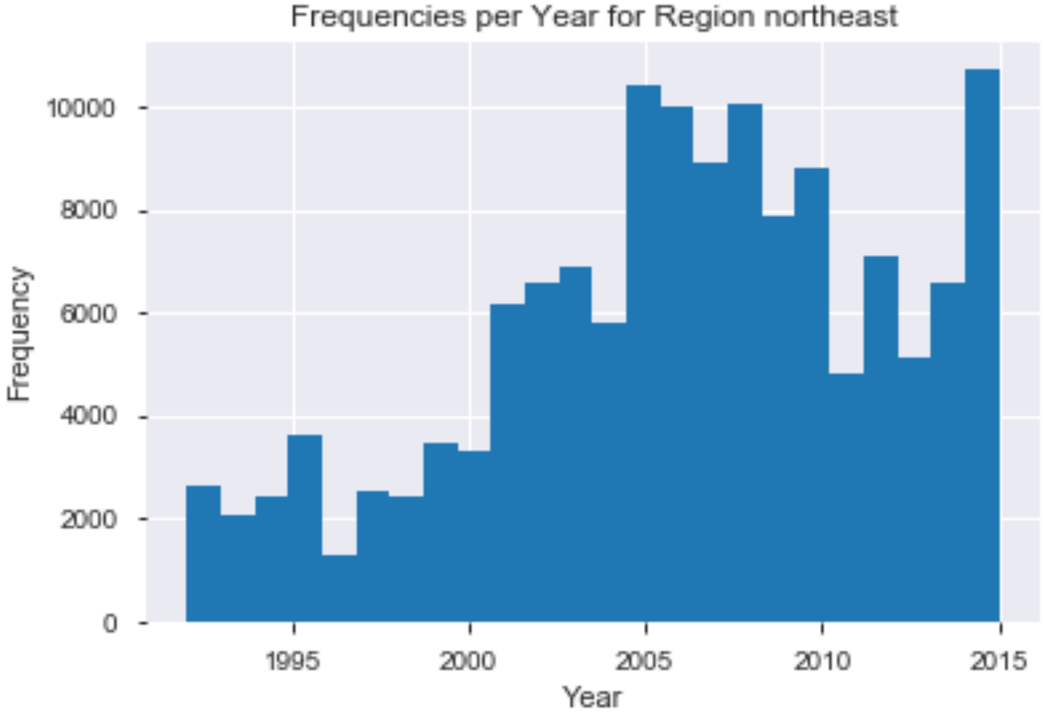


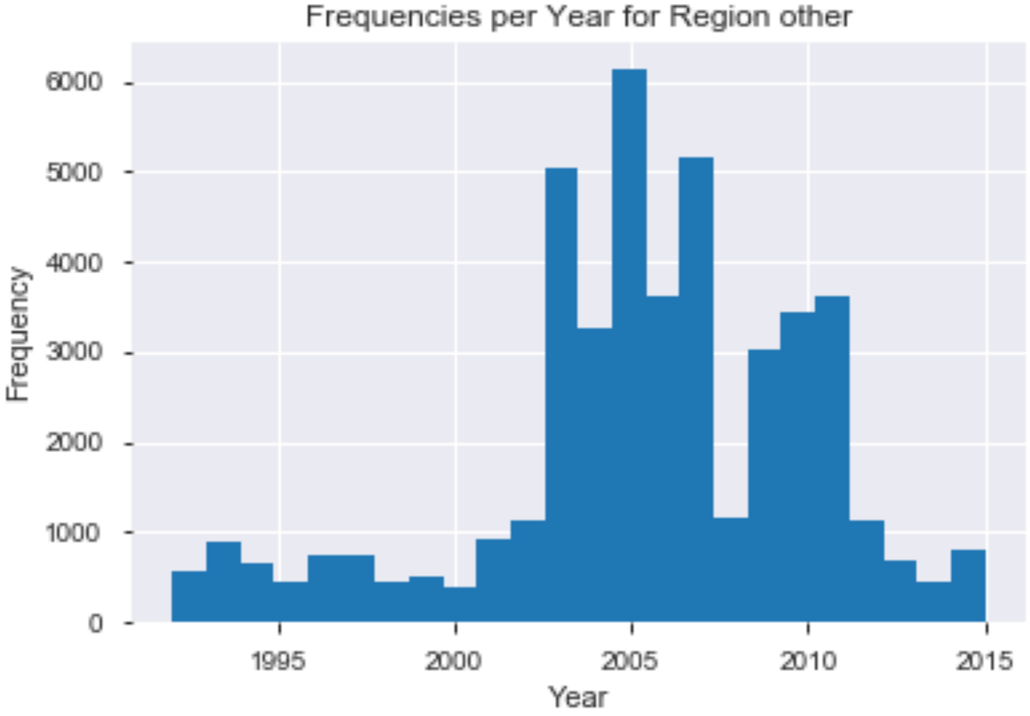
Frequencies per Year for Region midwest



Frequencies per Year for Region west

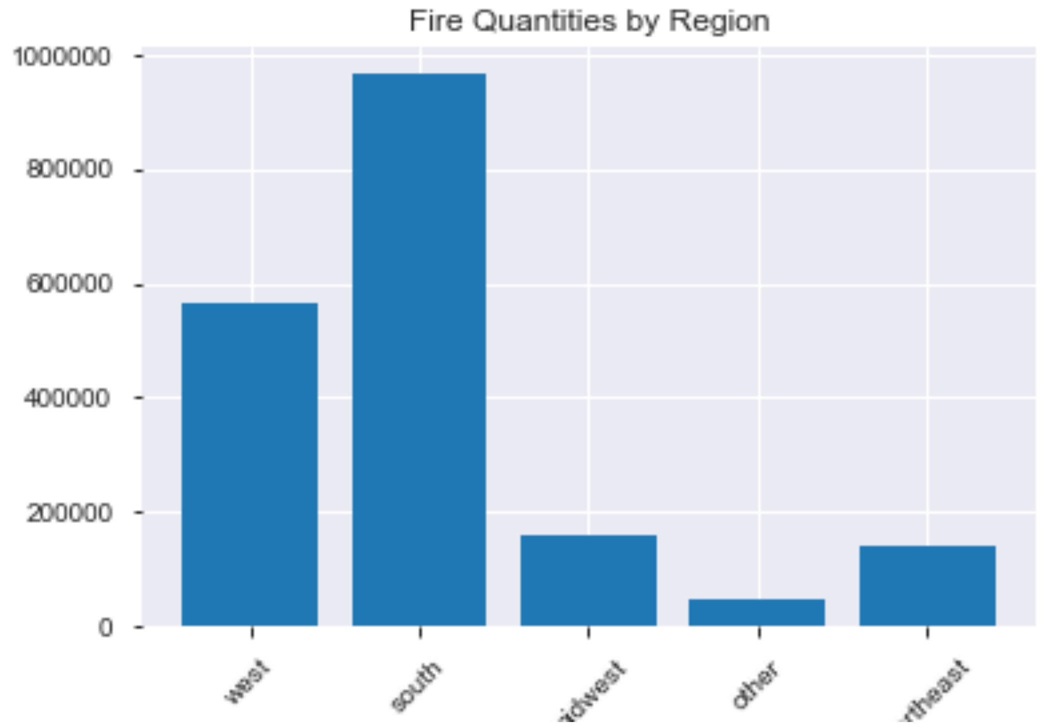

Frequencies per Year for Region south
Frequency
60000
50000
40000
30000
20000
10000
0
1995
2000
2005
2010
2015
Year


Frequencies per Year for Region northeast
Frequency
10000
8000
6000
4000
2000
0
1995
2000
2005
2010
2015
Year



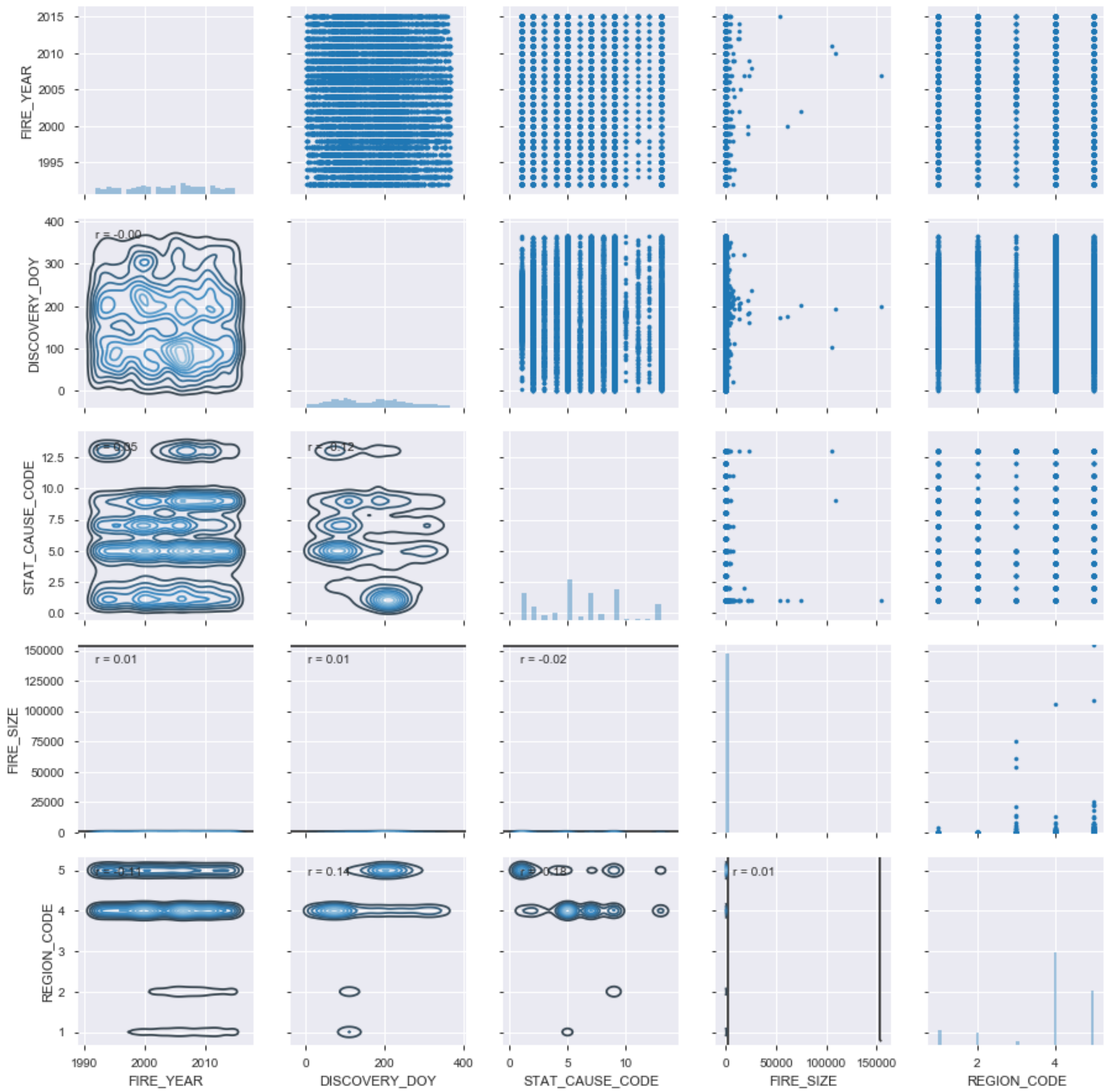
Where it is dry in the midwest, there's a pretty consistent high number of fires. Northeast and midwest regions have definitely increased over time. The south had a dip somewhere around 2005, then spiked back up.

By sheer number, the south has the most fires. The south is also the largest region.



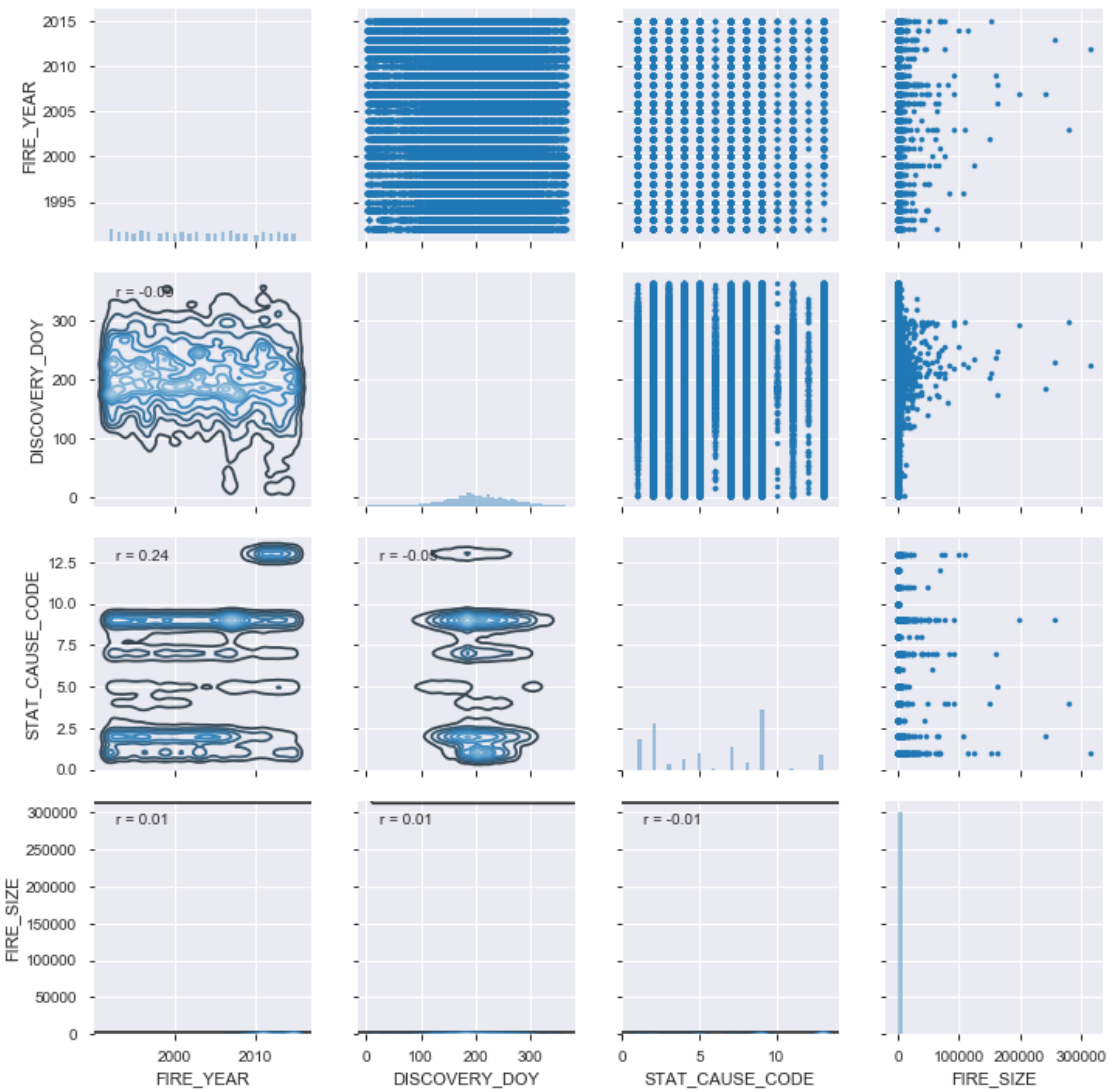
Analysis

Prior to running the analysis, I want to make sure there are no correlating variables. After narrowing it down to just Fire Year, Discovery Day of the Year, Statistical Cause Code, Fire Size, and Region Code, I still have so many data points that it takes the seaborn package hours to run on my local machine. So for brevity, I take a sample from the dataset of about 20,000 fires (which turns out to only be about 1% of the data, which is n’t great). This is the resulting grid:



What is demonstrated above is that there is really not a lot of correlation among these variables at all (all values for r are very close to 0), which is good, but it might not properly represent the entire dataset.

California alone is still about 200,000 fires, which takes a while to output seaborn's Pearson Correlation matrix (including the numerical values for r), but not as long as the entire 1.9 million dataset.



In the California subset, there is some correlation between Statistical Cause Code (STAT_CAUSE_CODE) and Fire Year (FIRE_YEAR). Based on the plot between Discovery Day of Year (DISCOVERY_DOY) and Fire Size (FIRE_SIZE), it looks like we see larger fires in the middle of the year versus the beginning or end of the year, which makes sense, since spring and summer happens in the beginning/middle.

I ran a multiple linear regression analysis, trying out 2 different target variables. First I tried predicting Fire Size based on year, discovery doy, and stat cause code; then I tried predicting Stat Cause Code based on year, discovery doy, and fire size.

With an R^2 of about 5%, the better of the two models is still not great. However, I was able to get more familiar with the python syntax of the sklearn package.

Conclusions

Alaska is on fire a lot. California is, too, but its fires don't cover as much ground. The northeast has seen an increase in number of fires between 1992 and 2015, as had the Midwest.

It's possible that the regions chosen are too large to capture the truth; the regions chosen leave too high for a margin of error.

Moreover, some datapoints could be missing from the dataset, based on the cutoff maps shown when plotting latitude and longitude coordinates of each state.

Multiple regression is probably not the best to use on this dataset due to the performance of the variables differing in time. Meanwhile, predicting statistical cause code is easier to do given fire size, discovery doy, and fire year compared with predicting fire size based on the other variables.

Finally, the seaborn package was useful to plot the correlation coefficients, but took a long time to run. I let it go for about 5 hours for the entire dataset before killing it. For about 20,000 rows, it took around 10 minutes. For 200,000 rows (California only), it took about 25 minutes. All of these experiments were using somewhere between 4-6 columns. Thus, having a means to an optimized plotting mechanism would be useful.