During the early phase of the project, we threw around ideas about what programs we could code that would give us a better understanding of the data. We began with the idea of simply mapping the locations of all the wells in the wells1-wells7.csv files (we will refer to these as the singular wells dataset). The program that Jonathan wrote to accomplish this goal was a simple and straightforward starting point for what we could later expand on. This starting point of the project code consisted of a scatterplot with points representing wells mapped by their latitude and longitude coordinates. From this, we could make general observations regarding the general spread and shapes that the points created, leading us to the realization that our latitude and longitude were backwards, so we switched them and could then see that the data showed a distinguishable outline of the United States.

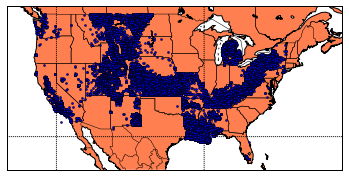
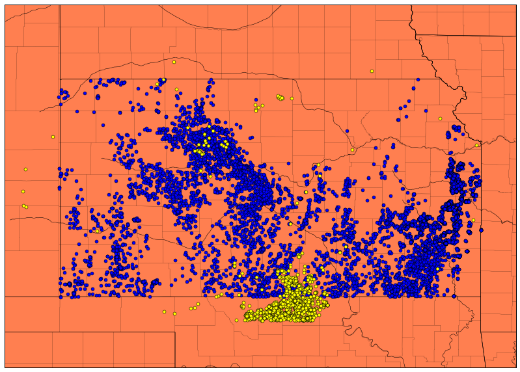
After learning what we could from the data in this form, we decided to improve the results of the code to give us a better visualization that included more context. Jonathan then implemented a solution using the basemap toolkit for python. Using this, he was then able to map the points on the United States, complete with the context of a map (see fig. 1). With this more useful visualization, we could distinguish areas where we could focus our attention and look for any patterns. Then, to further the functionality of the visualization, we added earthquake location data found by Keegan. However, due to limitations of the source, we were unable to pull earthquake data for the entire country. So, we considered the possibility of narrowing our scope to specific areas of the country. This would allow us to pull smaller chunks of location data for earthquakes and still analyze them alongside the well location data. 

Figure 1 (blue dots are wells)

Upon mapping the wells in the United States, we made an unfortunate discovery; Texas had no wells in it, and neither did Oklahoma. There were other states for which the data was missing; however, most of the states lacking data are outside the region on which we are focusing. We discovered that the data for Texas was not in the dataset at all, which made no sense because the article which cited the dataset, presented conclusions based on data from Texas. To make matters worse, the Oklahoma data – though present in the dataset – was not visible on the map we had made. This was later resolved as Ryan discovered an error in Jonathan’s algorithm that was keeping it from reading the data contained in wells7.csv.

 Upon considering the aforementioned issues we encountered, we decided to limit the scope of the project to a less broad area. Doing this allowed us to cut down the amount of data needed, and it enabled us to move forward in the project. Our focus was then narrowed down to Kansas.

So, Keegan pulled earthquake data specifically for Kansas, and Jonathan adjusted the code to plot the earthquake data in conjunction with the wells data for the same area. Additionally, the changes to the code filtered out wells which aren’t being used for fracking or wastewater disposal (see fig.2).

Figure 2 (blue dots are wells, and yellow dots are earthquakes)

With the data sufficiently focused, Jonathan was able to implement more useful features into the code that would allow us to investigate the data further. The most useful of these was the ability to set time windows for earthquake occurrences and spud dates (the point at which drilling the well begins) for the wells. This feature allowed us a more coherent way to observe correlations between the positioning and timing of both earthquakes and fracking wells.

Unfortunately, we ran into an issue regarding this approach; though we have the Oklahoma data, that data does not contain spud date information. Therefore, we would be unable use the data in an effective way that would assist in us reaching any conclusions regarding it. This is another reason we decided to focus particularly on Kansas; the data is rich with useful features that would then allow us to draw more helpful insights from the data.

[Discuss data source and inconsistencies]