I examined the correlation between accident rate and population density. First, I downloaded all the requisite modules and Keys, then I imported the data provided from the dataframe and looked it over for useful information. I then thought about what variables would be useful to investigate the question with; I thought to use city name as my any access to a resource that segments the country into areas of equal population, but realized that the US contains a huge number of individual cities and this could lead to many small towns with single incidents. I decided to substitute zip codes as a means of segmenting the country into different areas, as these tend to be more correlated with population and thus fewer instances of extremely small incident rates. Then, I called the Census API and used it to find the population of each zipcode. Simply examining which zipcodes would have the most accidents would likely just result in the outcome of “most accidents occur where the most people live”, so I decided to use the rate of accidents per 10,000 population. To get this I grouped the data by the zip code, took a count of the number of accidents in each, and divided the number of accidents occurring in that zip code by the number of people living there. I then ran a heat map on this data, as seen here, and ran a linear regression of the population of each area’s a.

However, despite using zip codes instead of city names, I still found that there were many small towns with individually tiny incident rates which yields a concern; an area with a tiny population and can end up having a high incident rate merely by having 5 incidents, while the areas next to it could have zero and thus be off the list, leading to a spiky representation of any given area. To compensate, and to put the data into a more comprehensible tabular form, I grouped the data into a number of bins. While I was originally concerned at the potential size of some of the bins, I realized that the range of population sizes represented by any zip code in the country caps out at about 120,000. As a result, I simply binned the data by 5,000 population sized bins, and then grouped the data by these bins. This had the intended effect of significantly smoothing out the data and allowed for reasonable plotting, though at the highest sizes there were some extreme values.

Unfortunately, the data seemed to not yield many interesting results. A regression analysis of yielded an r-value of .22, which is not particularly exciting. A graph of the final table with the binned values showed a potential issue in the correlation; while the accident rate seems to go up and then go down, in reality this may be due to simply to lower car ownership in areas with higher population in each zip code (cities).

Finally, and a general issue with the dataset as a whole, is that despite its size, it doesn’t really have all the accident data in the US. How it was gathered could yield problems with sampling; it could be that the nature of how the data is reported could cause some places to underreport, or cities could be seeing an underrepresentation due to the comparatively minor nature of city driving accidents (cars are rarely totaled, instead getting into fender benders). That said, the methodology applied during my and my teams analysis could easily be applied to a more complete dataset, and thus forms a framework for finding more valuable results.