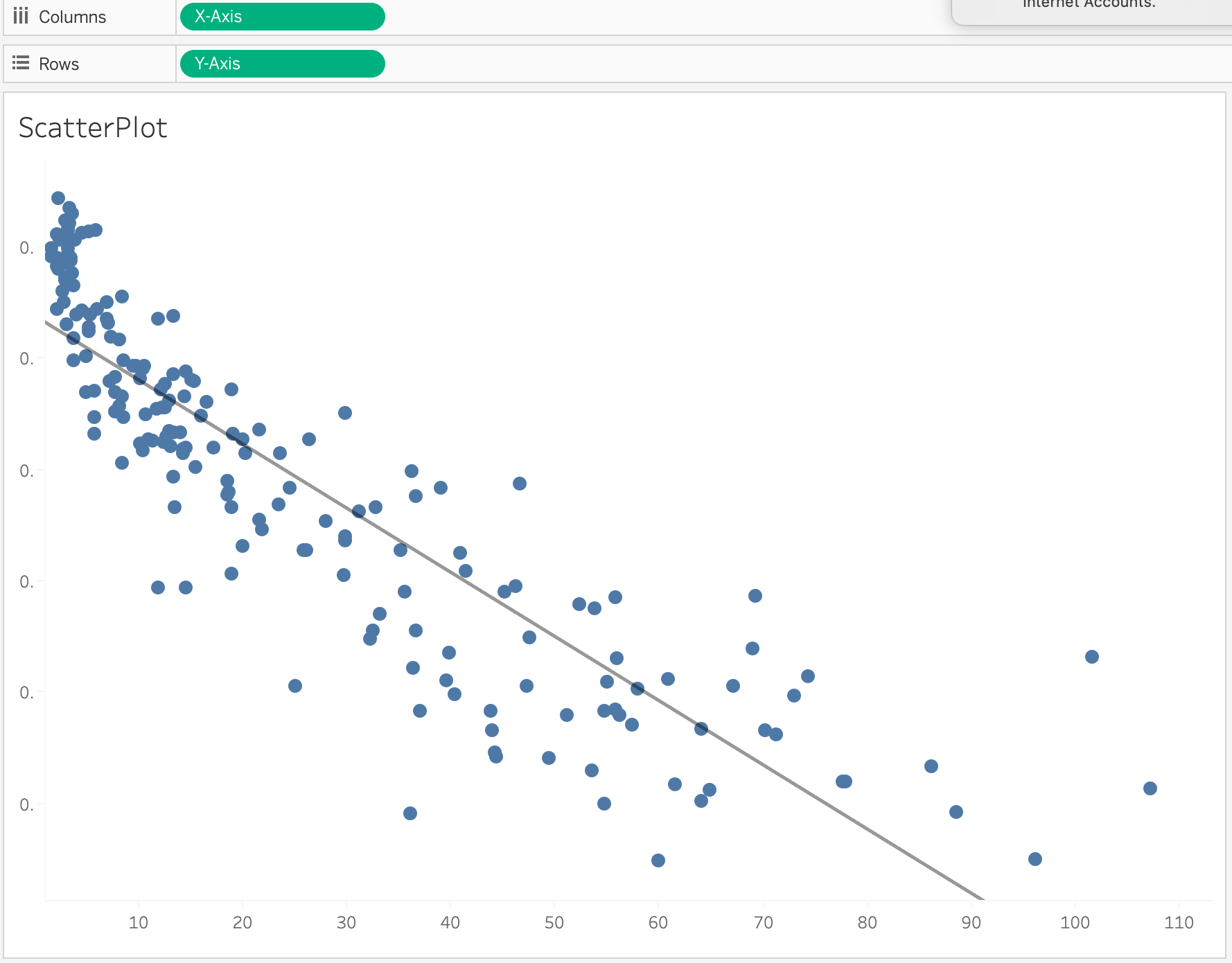
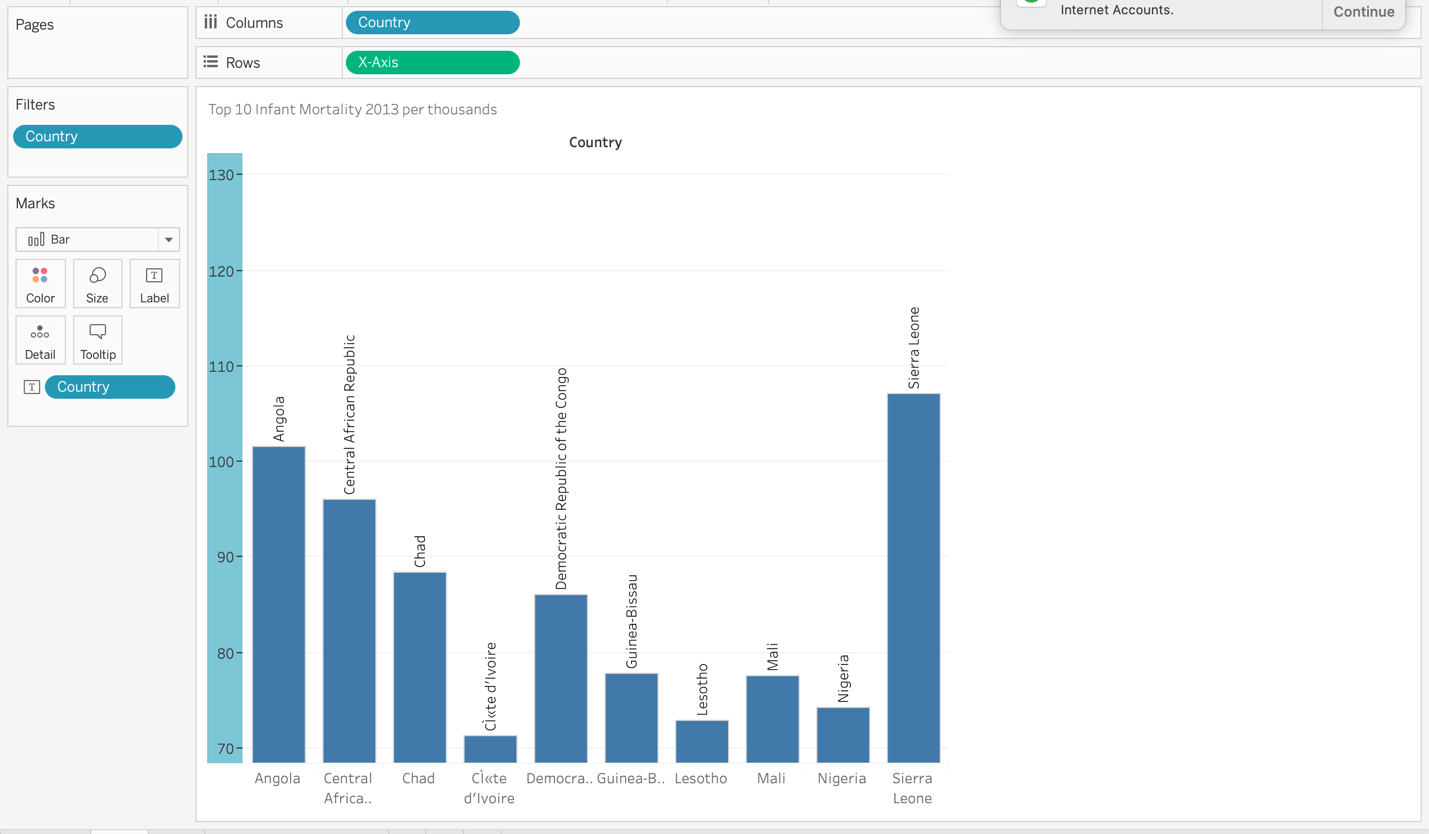
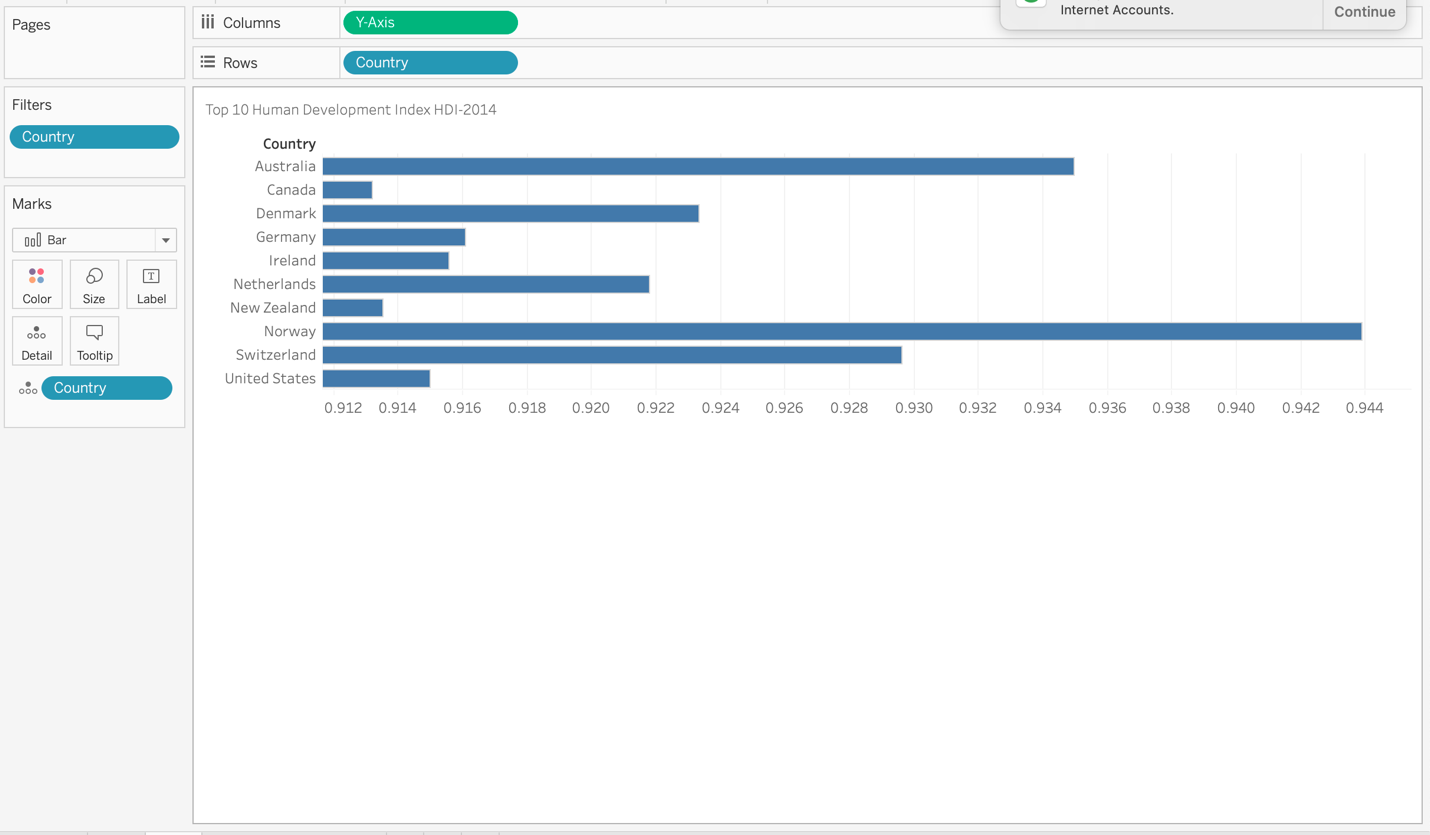
Part I: Self-service Analytics

In this visualization case, we try to visualize the correlation among variables of an identity. We first create two parameters: x-axis and y-axis and generate calculated fields for each parameter based on all variables of the identity, namely the country, and then pull the two parameters to row and column respectively. After that, we can visualize the correlation of any two variables by selecting them.



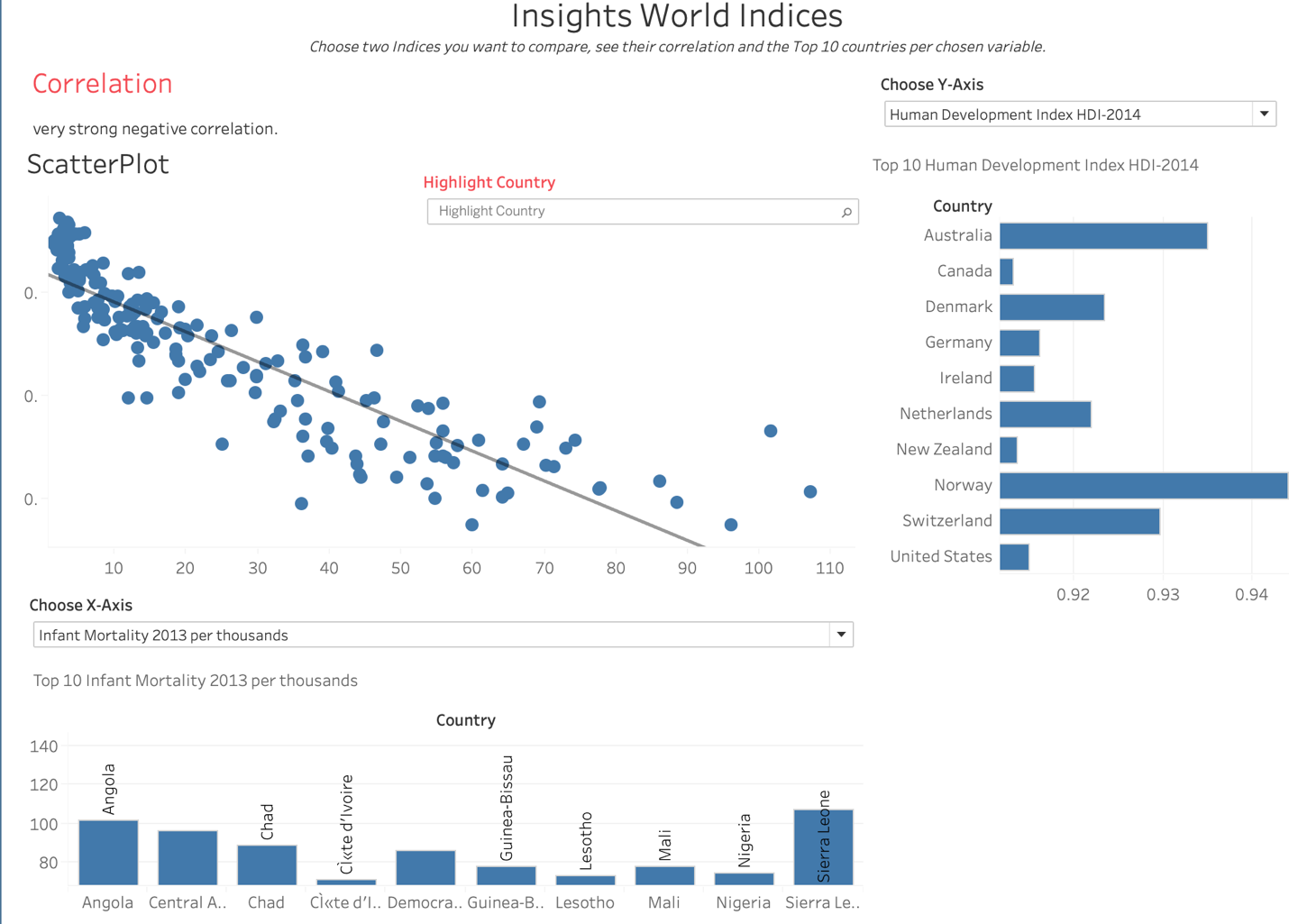
The result looks like this for two variables. Moreover, we create two bar charts for two variables on the y-axis and x-axis. Each bar chart represents the top 10 countries with highest value on that particular variable. In addition, we also create a text box to tell readers of the chart if the two variables show correlation.



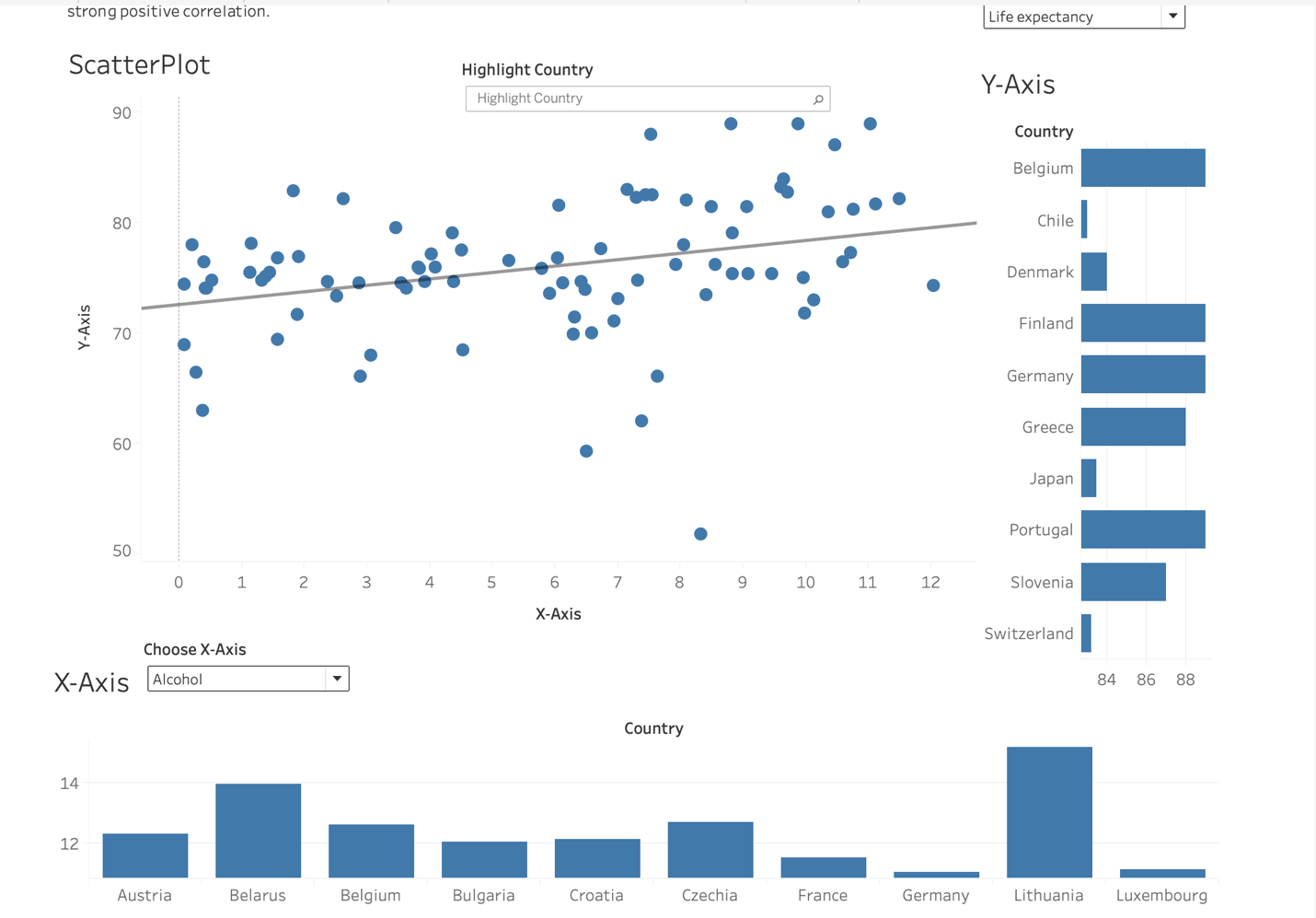




Finally, we put all four parts on the same dashboard, so that we can select which two variables we would like to test for their correlation and all four parts will adjust to the selection accordingly. The dashboard is shown below.



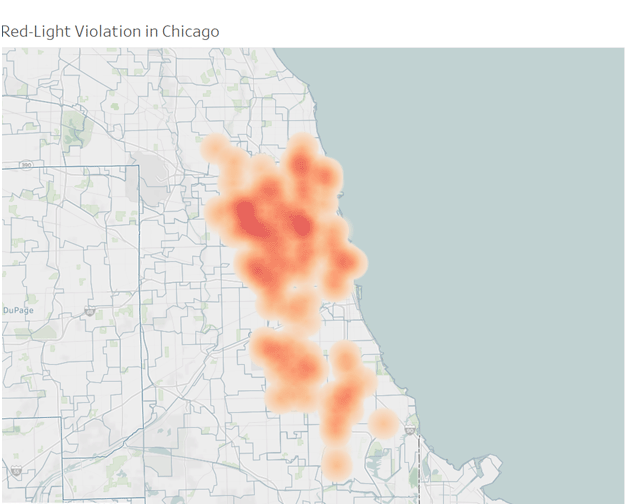
In addition, we find another dataset, which contains some health condition indicators for all countries. We try to visualize correlation among these indicators, and so we replicate the process and create a dashboard using this dataset, which is shown below.



In this case we would like to see if how life expectancy is correlated with alcohol consumption. Though more alcohol consumption leading to shorter life expectancy sounds more like a common sense statement, when we actually plot the scatterplot of the two variables, it appears that the contrary is true. More alcohol consumption actually increase life expectancy! But when we think about it a little further and look at the plot more closely, we find that the countries with highest alcohol consumption concentrate in Europe and can be categorized as developed countries for the most part. It is also these countries that have the longest life expectancy. It seems that economic development appears to be the hidden factor to account for both high alcohol consumption and life expectancy. i.e. the linear regression of the two variables is strongly biased because there is an omitted variable that is correlated with both dependent and independent variables. It is the bar charts that show top 10 countries that imply the idea. So it is really important to have the bar charts in the same dashboard so that we can examine the correlation/regression more closely and identify any abnormalities.

Part II: Geo-spatial Analytics

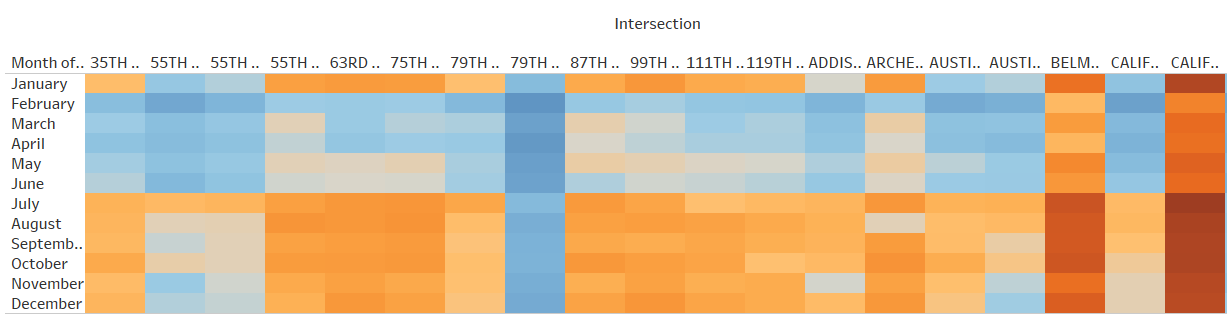
After we merged the data and graphed it on the map, the density map looks like this.



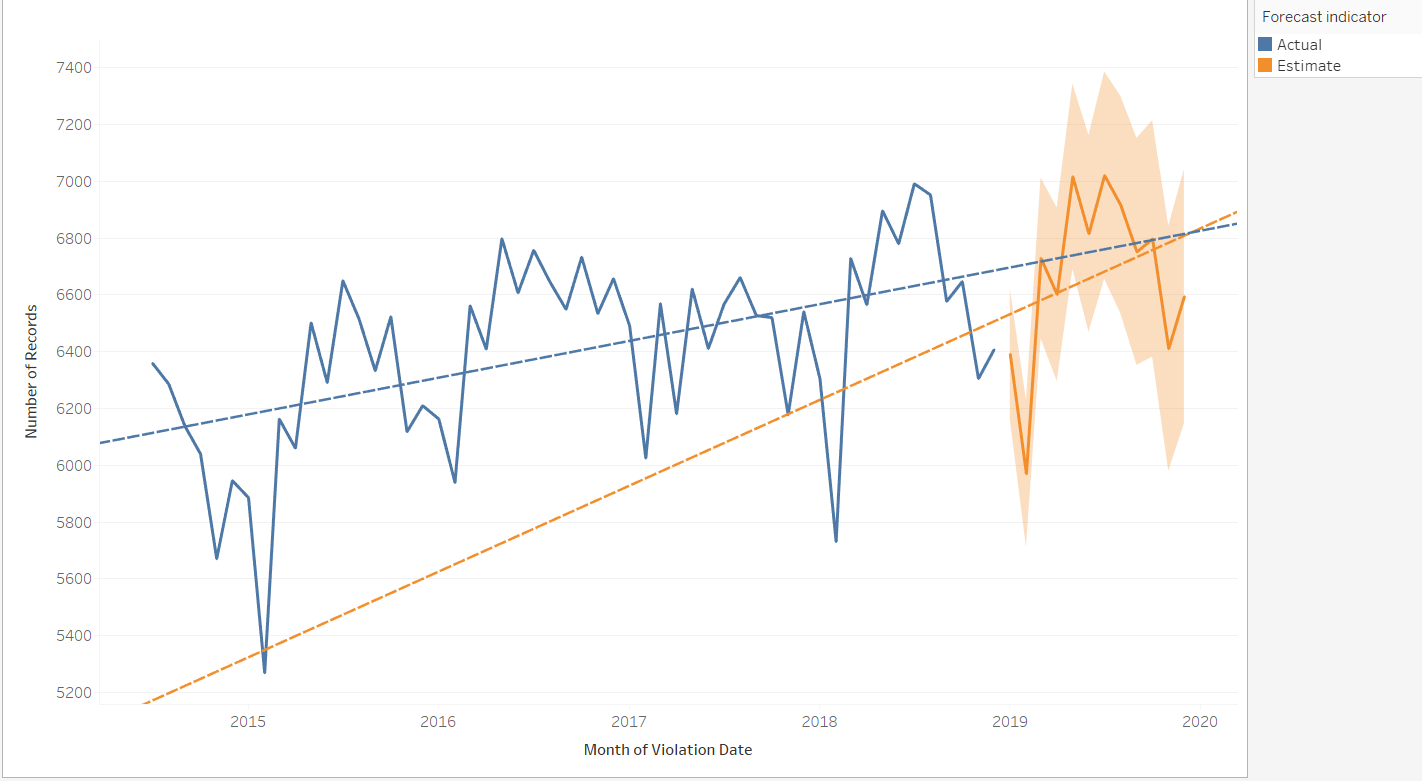
However, although the animation shows that the violations are constantly changing, it does not show us a specific trend of violation.

Then, a Heatmap becomes very useful here. By using a heatmap, we are able to detect the trend immediately. With red color indicates high level of violation and blue indicates low level, the heatmap shows a clear boundary between first half and second half of the year. From January to June, the color is closed to blue. While in July to January, the pattern of red and orange demonstrates a much higher red-light violation rate.

*Heatmap*



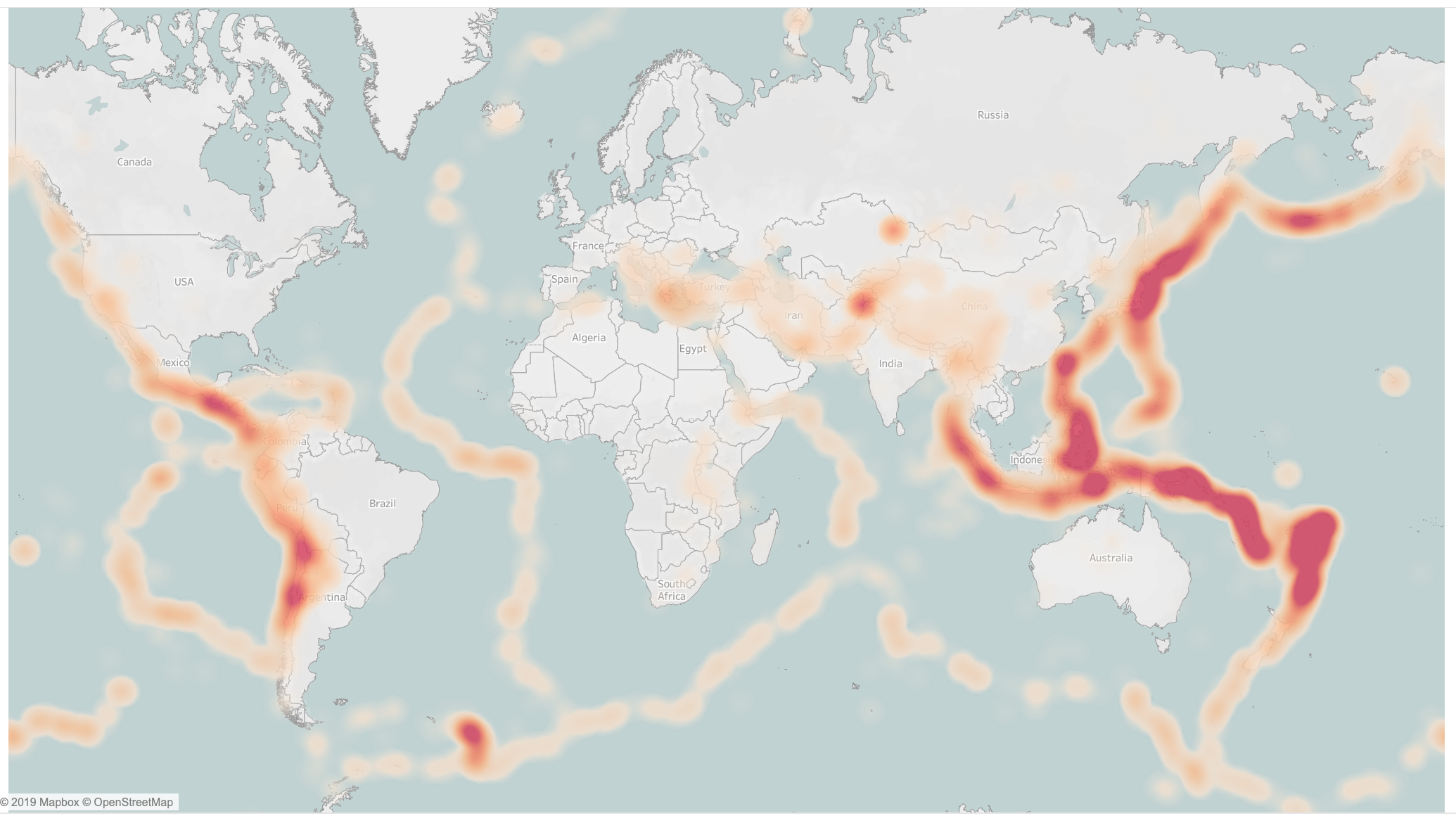
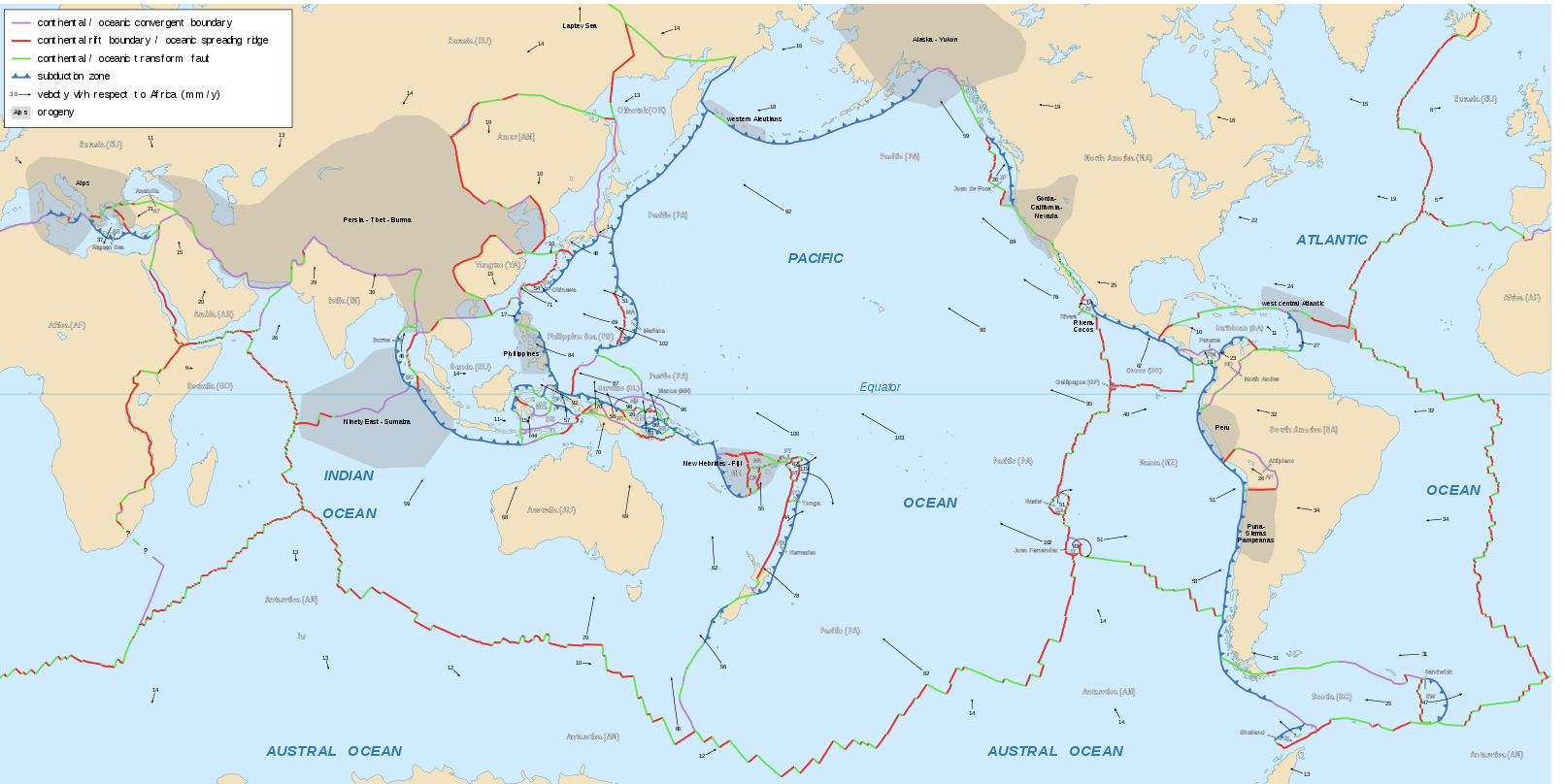
Then, we want to see the trend of red-light violations in recent years. So, we make a forecast to predict the future violation happens in Chicago.

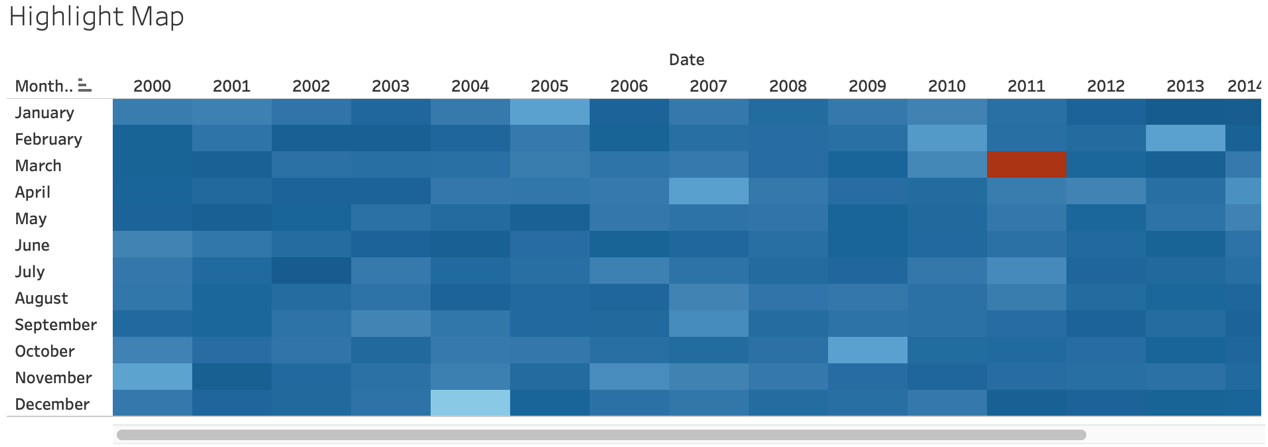


From the forecast, we see that both the actual and estimate have an upward trend while the slope of estimate is larger than the slope of actual. This means that in both short term and long term, number of red-light violations will increase and, in a speed, faster in the short run.

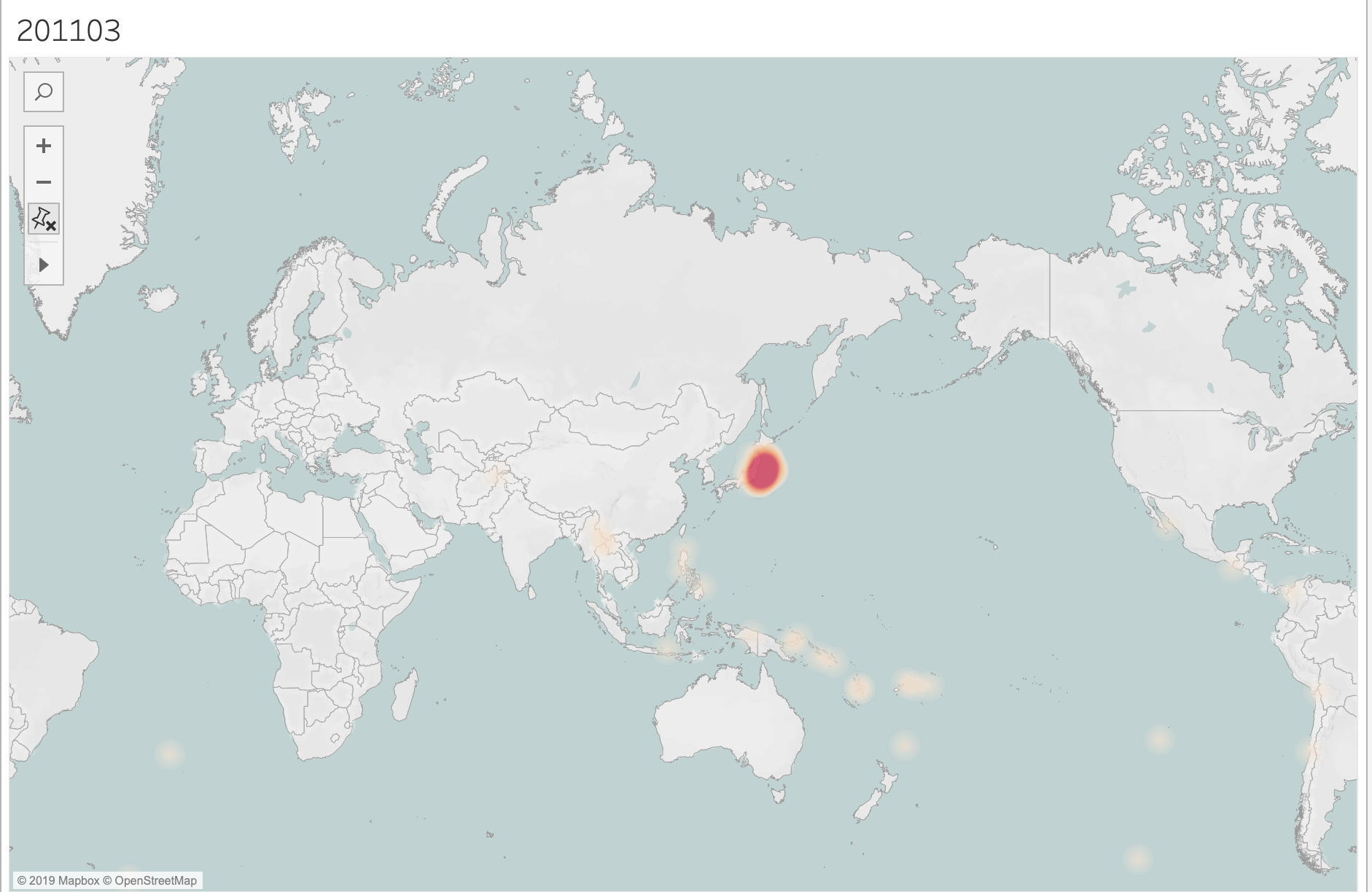
**Earthquake Prediction**

I applied the earthquake dataset from 2000 to 2016 to see the frequency and locations of earthquakes around the world. As we can tell from the Heatmap, the places where earthquakes occur most often are the convergent plate boundaries, especially Eurasian Plate and Pacific Sea Plate.





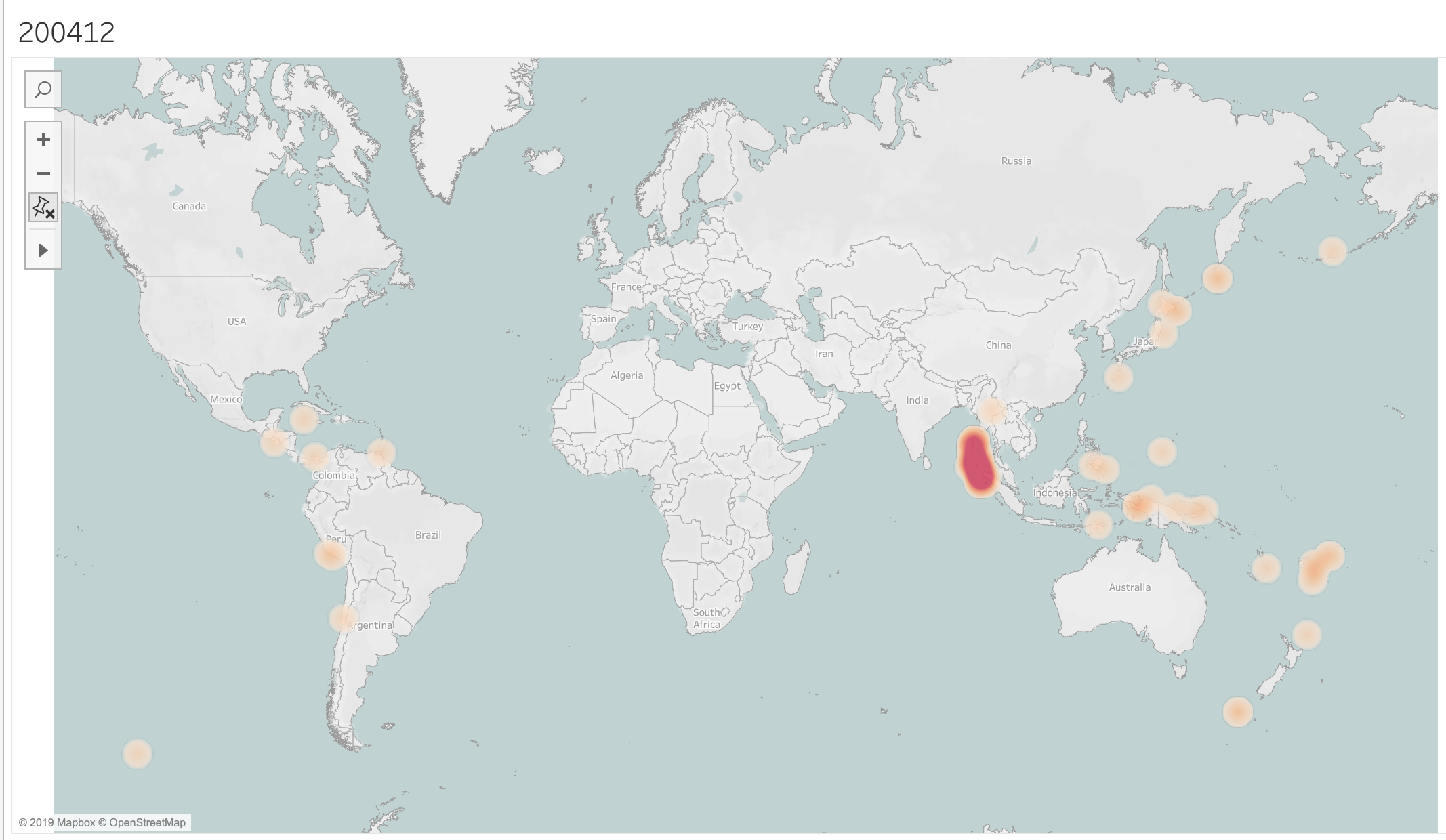
Next, we are going to take a look of the Highlight Map, which will show the numbers of earthquakes in each Month from 2000 to 2016. From the graph shown above we can see that March, 2011 has the color of red, which means during this month there was a great number of earthquakes occur. When we saw the year and month shown in the graph, we recalled that there was a disaster happened in that month. I believe that everyone knows that on 11, March, 2011 there was a magnetite 9 earthquake struck Japan northeast, triggering a massive tsunami and caused 16,000 died or listed as missing. This was the largest earthquake in 50 years.



To verified our assumption, we looked for the information on the internet and filter the Heatmap to check the data in March,2011 and found that the month matched our assumption. There was high frequency of earthquakes happened near Japan in that month.

Back to the Highlight Map, we also can see that the color of Dec.,2004 is the second darkest around these years. We also remembered that there was a serious earthquake called Indian Ocean earthquake and tsunami in 2004.

To verified the assumption, we filtered that “date” again and got the answer of our assumption.



Last, we want to predict the frequency of earthquakes using the past data. As we can tell from the graph below, the trend of earthquake is relatively stable, except for the two month I mentioned above; however, no trend means that it is difficult for us to use history data to make prediction in the future. As a result, we will conclude that the accuracy of this forecasting will not be high. Also, plate movements are relatively difficult to predict, and more advanced instruments are needed to detect plate activities.

