Assignment 4: Advanced Analytics

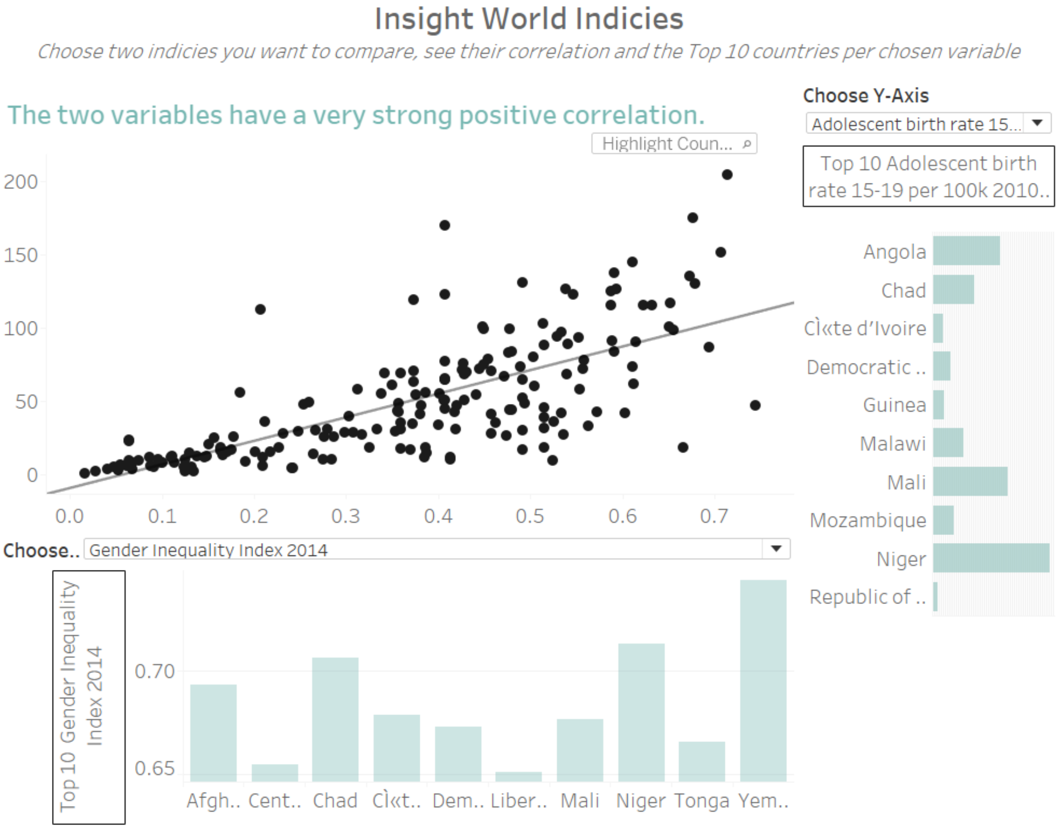
Team member: Chuyue Wu, Menghong Han, Xiaofan Sun, Yifan He

Part A

Brief Introduction of Replication:

Our replication is based on the world indices data, which contains various information of each country. According to it, we can make a scatter plot with each dot standing for one specific country to explore these indicators and find out their correlation.

Screenshot:



Insight World Indicies

Choose two indicies you want to compare, see their correlation and the Top 10 countries per chosen variable

The two variables have a very strong positive correlation.

Choose Y-Axis

Adolescent birth rate 15...

Choose.. Gender Inequality Index 2014

Plot1 – Replication of Part A

Insights:

Insight 1: We cannot simply only interpret a metric by its name, but to see its correlation with different metrics to see if it has a confusing name, and then re-interpret it in a more scientific way.

For example, *Prison population per 100K people* and *the Expected years of schooling* have a moderate positive correlation. Similarly, the former also shows a moderate positive correlation with all metrics evaluating higher education levels such as *Population with at least some secondary education percent* and *Mean years of schooling*. Apparently, it shows a moderate negative correlation with all metrics evaluating lower education levels, or higher crime rates - such as - *Primary school dropout rate* and *Infant Mortality*. This surprised me because generally, we simply interpreted prison population as - the higher of this metric in one country, more crimes would there be in this area thus the lower education levels. But according to the graph with real data, the prison population per 100K might probably be re-interpreted as a metric which has negative relationship with crime rates.

To think more deeply, the reason behind this would be - the larger prison population might show the higher efficiency of police department in this area, revealing that this area may be relatively well-developed and rich enough to construct completed public works. This explanation could somehow be reconfirmed by the moderate positive correlation between *Prison population per 100K people* and *Human Development Index*. Society issues were too complicated to view them separately. When we have various metrics, we could generally first pick one of them as a benchmark(in this example, Prison population per 100K people), finding its relationships with different other metrics to get much more information from multi dimensions.

Insight 2: Correlation & Causation: tricky one when doing analysis based on correlation graphs

When we analyze and draw conclusions based on correlation graphs, a very typical mistake we would probably make is to confuse correlation with causation. Typically, when two sets of data have strong correlations, we could say one is independent variable and the other is a dependent variable. But the fact is, some data sets having very strong relationship, but we cannot draw any conclusions for a cause-and-effect relationship, or say, we would better not to draw a cause-and-effect conclusion indiscreetly. An example in our visualizations is, a strong correlation between *Under-five Mortality* and *Change mobile usage*. If you say the higher under-five mortality there is, the higher the changes of mobile usage only based on this graph - that is a bit

ridiculous. In analysis, we should draw conclusions based on our common sense in reality(or even research literature), rather than trusting data blindly.
To avoid confusion, we should take a review on definition of correlation more cautiously. Correlation is a mutual connection between two or more things, which is normally a statistical and mathematical connection based on counting data. But causation means the relation between something that happens and the thing that causes it. Usually, the first thing that happens is the cause and the second thing is the effect. Our subconscious sense makes us confused sometimes, when we draw a conclusion of a real-world problem, thinking more cautiously is important.

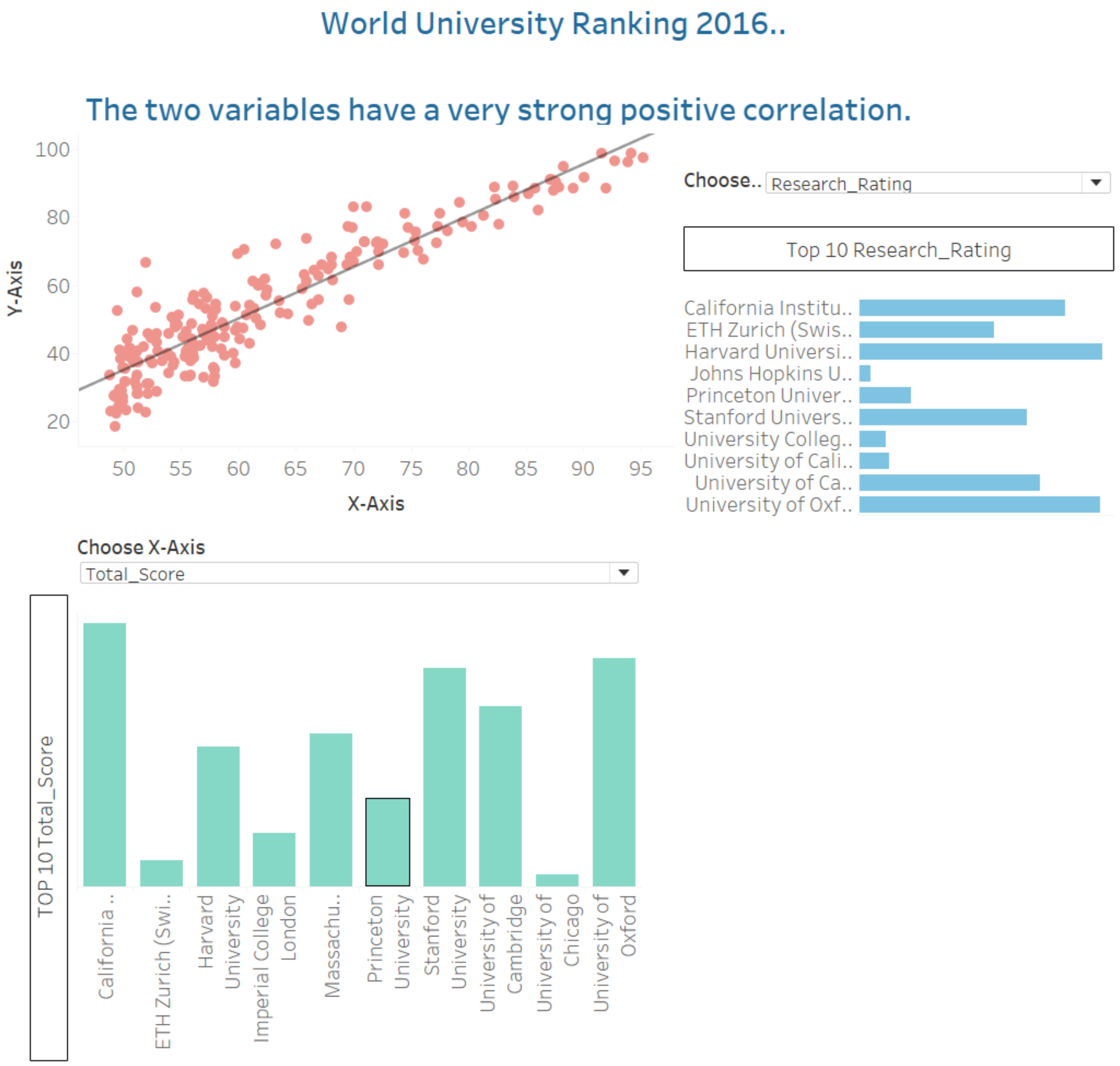
Brief Introduction of Application:

Here is our application. We used *World University Ranking 2016* data in data.world website (https://data.world/hhaveliw/world-university-ranking-2016), filtered data of 2016, chose top-200 rankings' universities, and imported it into Tableau. Then we repeated the process in the previous case and got the application dashboard.

Analogous Features:

The world indices dataset contains various information of each country, according to which we can make the scatter plot with each dot standing for one specific country to explore these indicators and find out their correlation. When it comes to the new data, it is the same story. It lists top 200 universities over the world and provides their ranking and rating scores based on different criteria (such as teaching rating, industry income rating, research rating and so on). In the application part, we can also transform the new dataset into scatter plot, finding out which indicators are related to each other.

Screenshot:



World University Ranking 2016..

The two variables have a very strong positive correlation.

Choose.. Research_Rating

Choose X-Axis Total_Score

Plot2 – Application of Part A

Insights:

From the interactive correlation table, we can get a clear view of how the 9 independent variables (teaching, Research_Rating, Citations_Rating, Student/Staff_Ratio, Inter_Students_Rate, Inter_Outlook_Rating, Industry_Income_Rating, Female_Students_Rate, Num_Students) correlated with the total score of universities. The relationships can be divided into 3 groups according to the degree of correlation : the group with very strong or strong correlation, moderate correlation, and weak or no correlation.

The group with very strong or strong correlation with total score includes teaching, Research_Rating and Citations_Rating, and the relationships are all positive correlations. The group with moderate correlation includes Student/Staff_Ratio and Inter_Students_Rate, and Student/Staff_Ratio has a moderate negative correlation with total score, which is consistent with common sense. In addition, we found that most of the universities with the high Student/Staff_Ratio are from Germany. It seems against our common knowledge that generally Germany universities are famous for small class scale. We dug further and found out that all the outliers are public universities, which may somehow break the stereotype of Germany university teaching system that maybe only German private universities have smaller class size. The group of weak or no correlation includes Industry_Income_Rating, Female_Students_Rate, Inter_Outlook_Rating and Num_Students. The result that there is weak or even no correlation between total score and Industry_Income_Rating is unexpected, since we thought the knowledge transformation rate should be strongly correlated with the total score because it suggests the extent to which businesses are willing to pay for research and institution's ability to attract funding in the commercial marketplace, which is a useful indicator of institutional quality, therefore, we found the result a little weird. Maybe Industry_Income_Rating is a strong variable for predicting the total score, but the marking system Times Higher Education rankings weighted this dimension lower than we imagined. We found out the marking system of Times Higher Education rankings and the weight of this factor is as low as 2.5% as expected. Furthermore, we dug further and found there is a moderate correlation between Research_Rating and Industry_Income_Rating which makes sense.

Reflection on replication and application:

First, through the process of finding proper dataset, we learned that valuable data is the one that can tell a story. For example, at beginning, we planned to use the Spotify data that compiles a playlist of the songs streamed most often over 2018. However, most of the correlations between attributes in the data fail to have a real-world meaning.

Second, before making visualizations, we are supposed to drop useless variables. Take the new dataset as an instance: if we regress world rank on total score, a very strong positive correlation will be printed out. However, this result cannot reach any useful insight, since the world rank of each university is totally based on its total score.

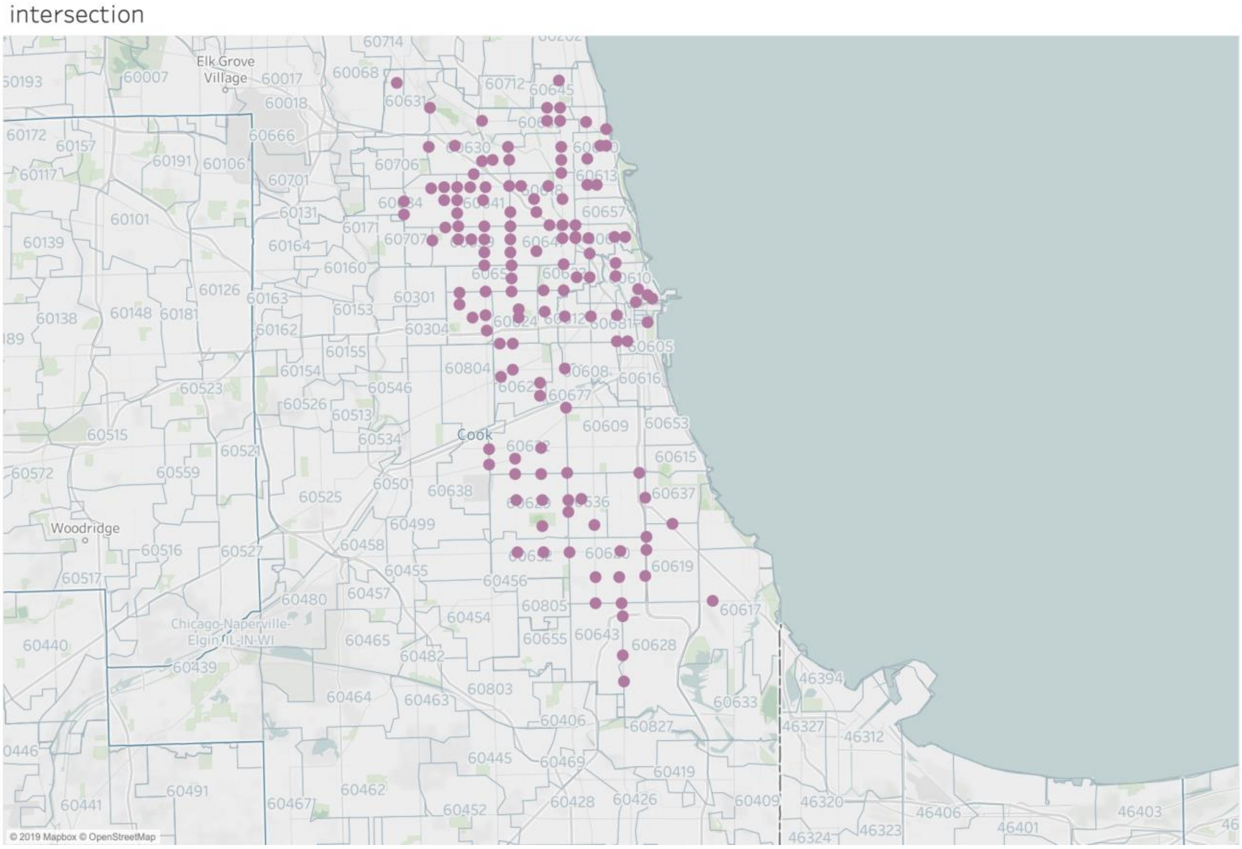
Finally, arranging dashboard to make visualization clear and easy to interpret is essential.

Part B

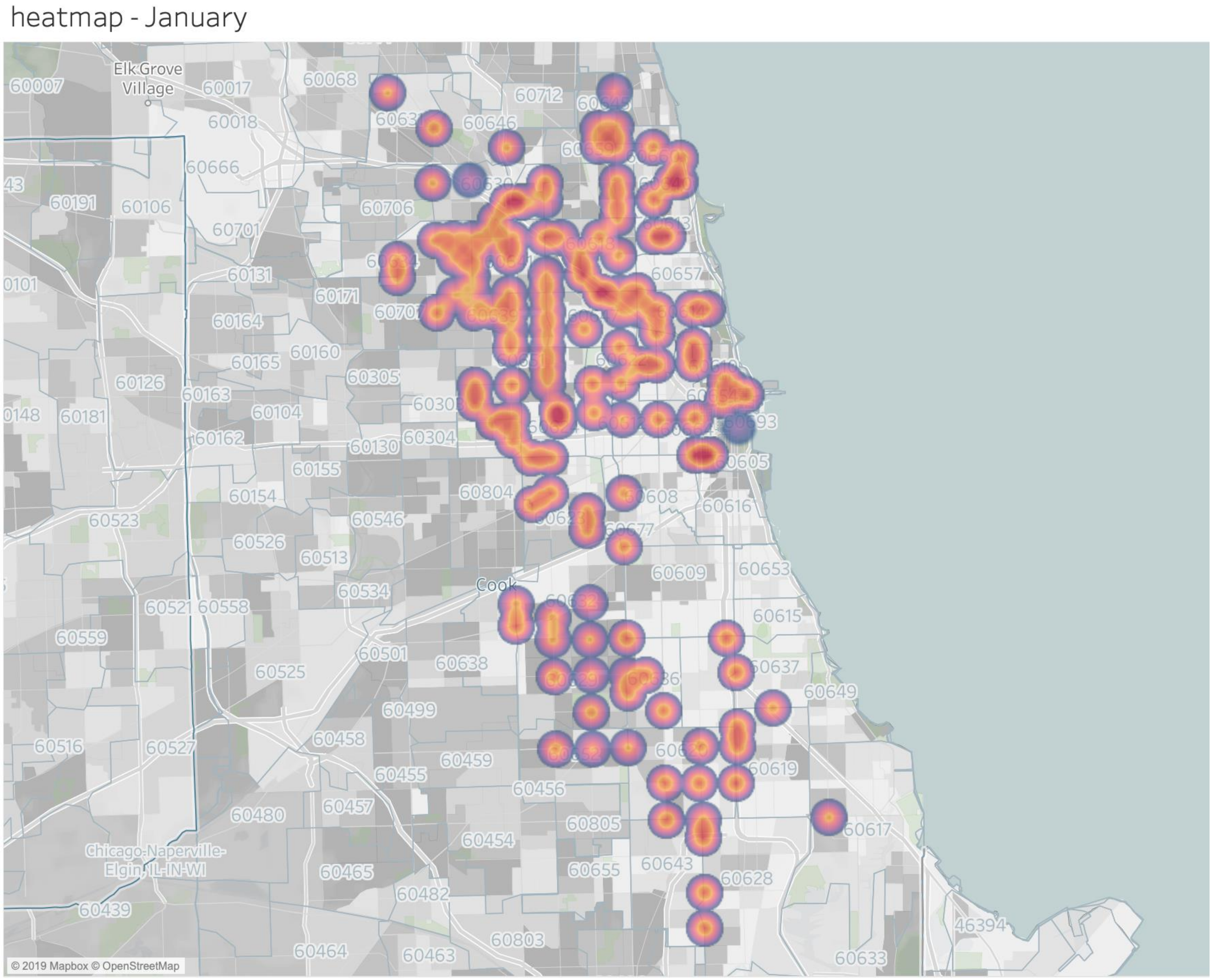
Brief Introduction of Replication:

In this part, we use longitude and latitude as columns and rows, then we adjust map layer to show many other indicators, such as land cover, coastline, and county borders.

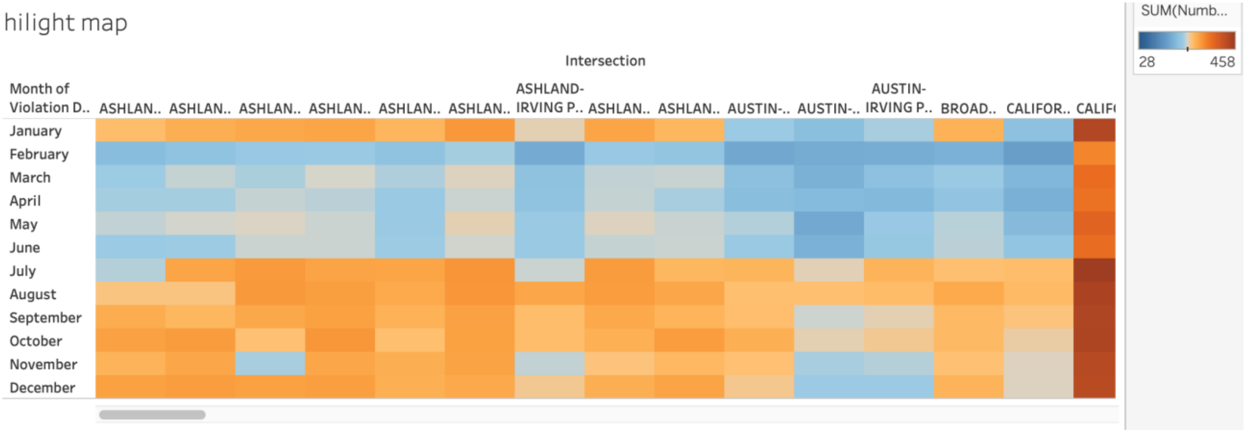
Screenshot:



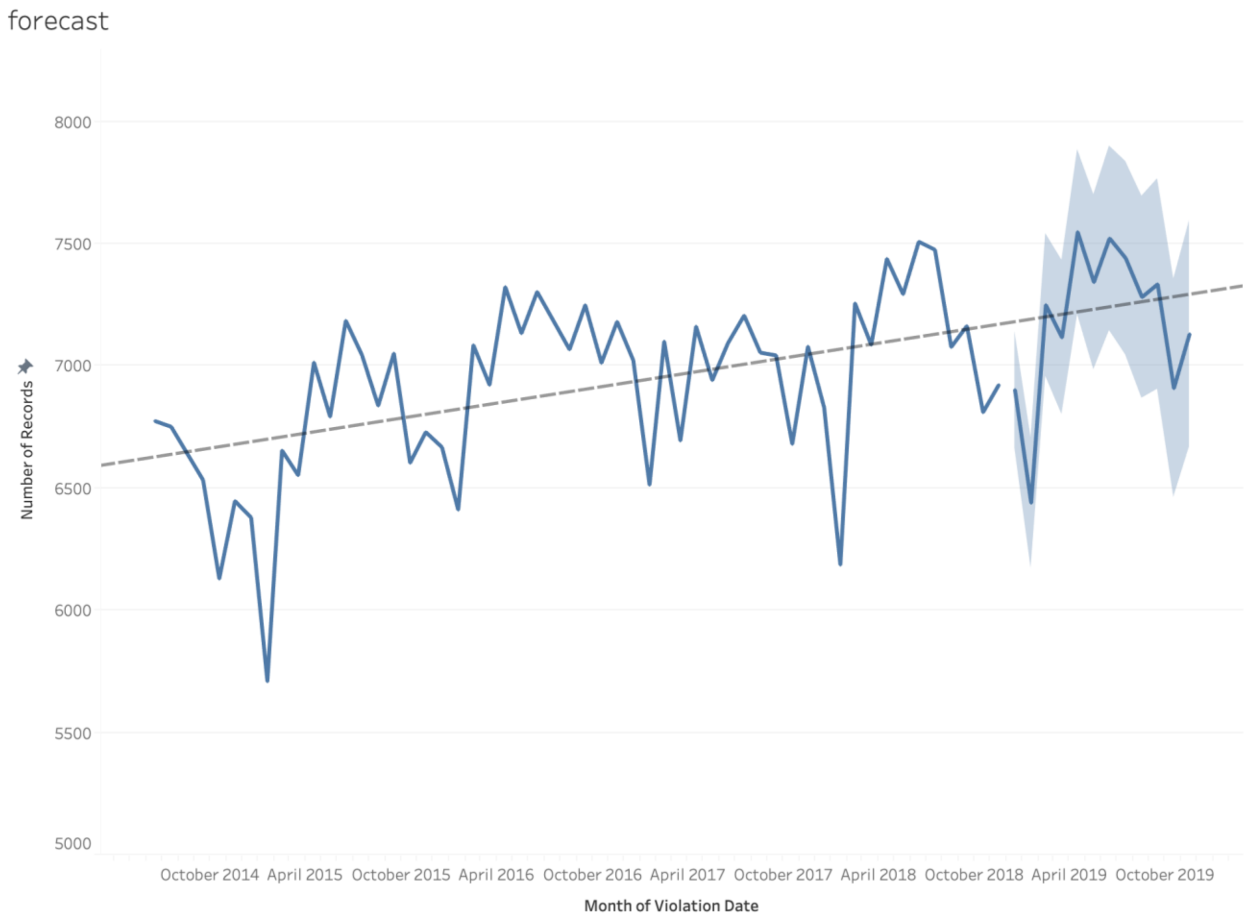
Plot3.1 – Replication of Part B(1)



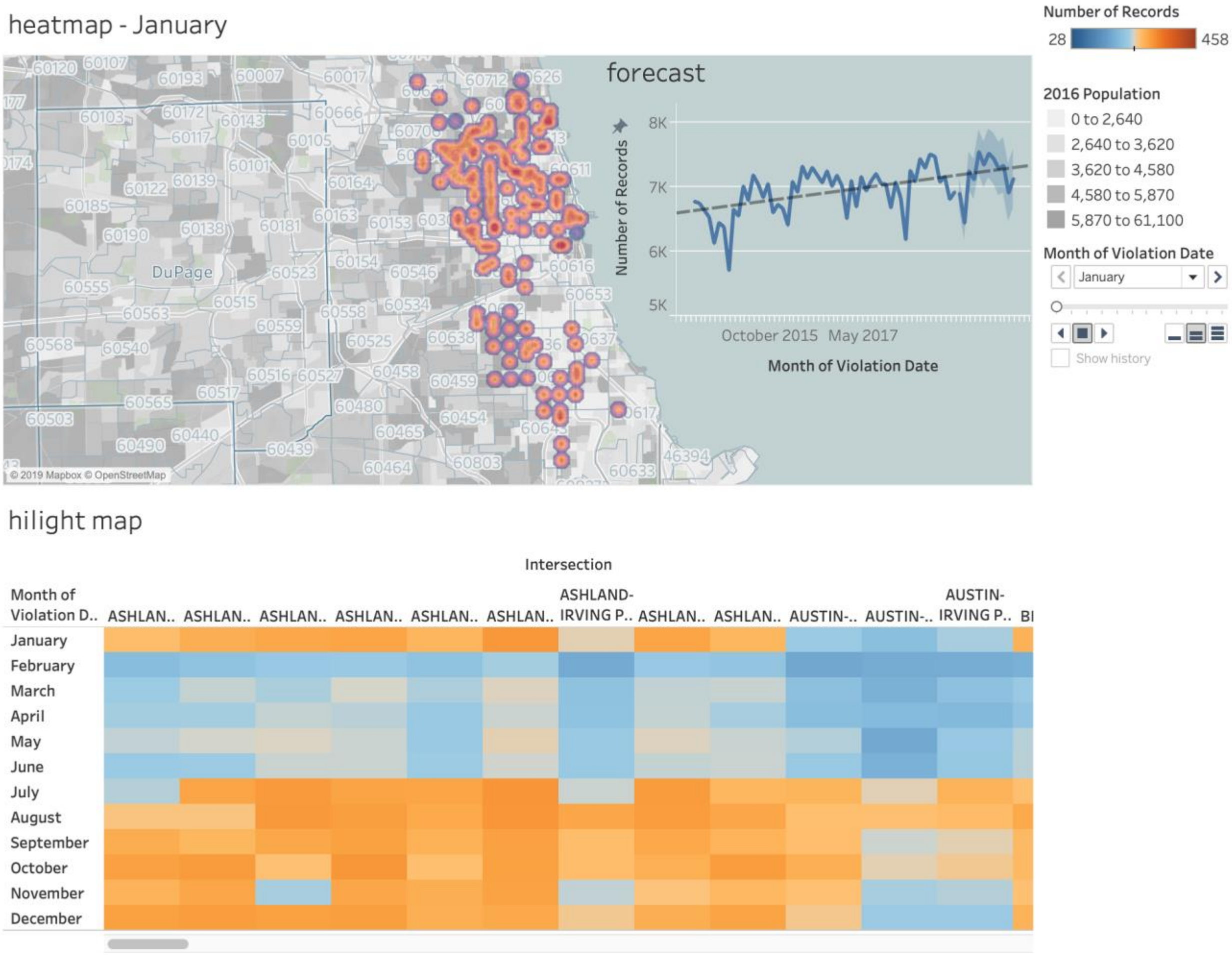
Plot3.2 – Replication of Part B(2)



Plot3.3 – Replication of Part B(3)



Plot3.4 – Replication of Part B(4)



Plot3.5 – Replication of Part B(5)

Insights:

For the heatmap (plot3.2), we use monthly data to indicate violation images of different months. On the basis of previous map, we add some more Marks and Filters. There is an In/Out (Top 20) filter to see whether one intersection is among the top20 violations. We use Density function to show the frequency that violations happen in one spot. Besides, map shows blocks of grey colors, which show population of different areas.

For the hilight map (plot3.3), we use intersection as column data and month(violation) as row data. Color is from blue to red, indicating frequency of violation from less to severe. From this plot, we can also see the monthly change throughout the year, that is, the second half year generally has more violations than the first part.

From the line chart (plot3.4) we can see change through 4 years. There is a periodical fluctuation in these years. The lowest point is generally February and time of period is about 1 year. Also, we can see the whole trend across 4 years, which is to increase at a relatively constant rate.

Finally, through the dashboard (plot3.5), we add interaction between different charts. Once we click on a dot on heatmap, we can see the exact name of intersection and violation times in that particular month. This allows us to comprehend more about severity of violations at different intersections in different time period. Also, we make forecasting chart transparent on top right corner of heatmap, which is clearer to predict future trends.