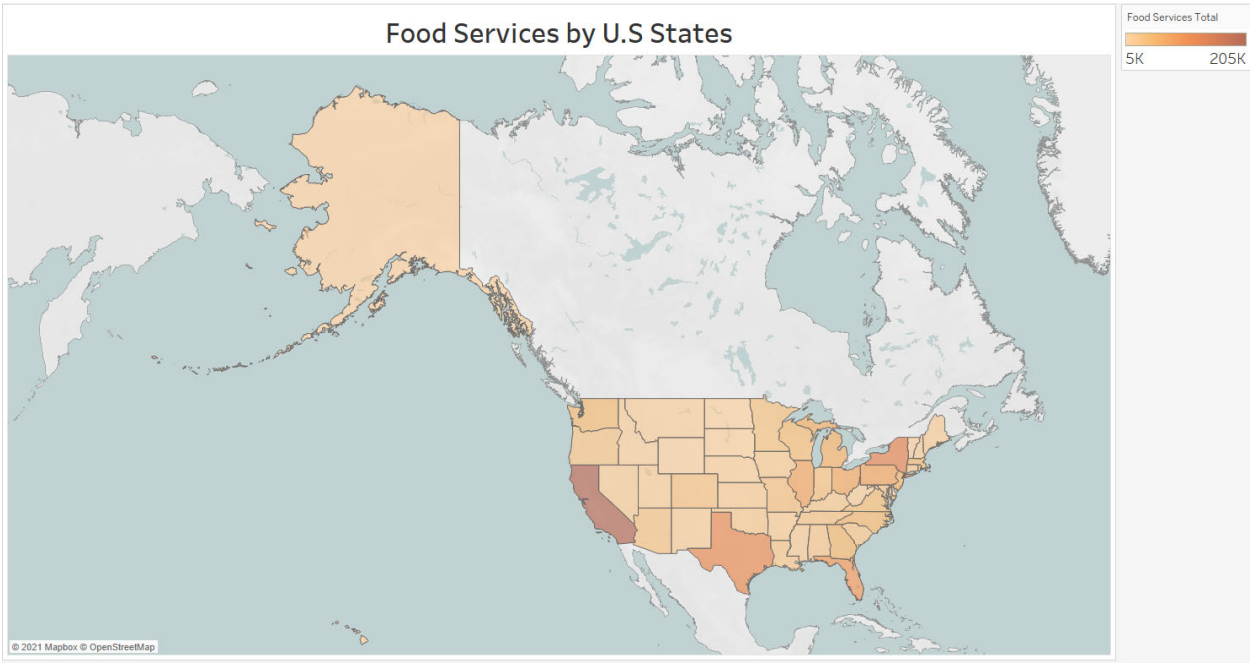
DSC 465 Homework2

Ronaldlee Ejalu

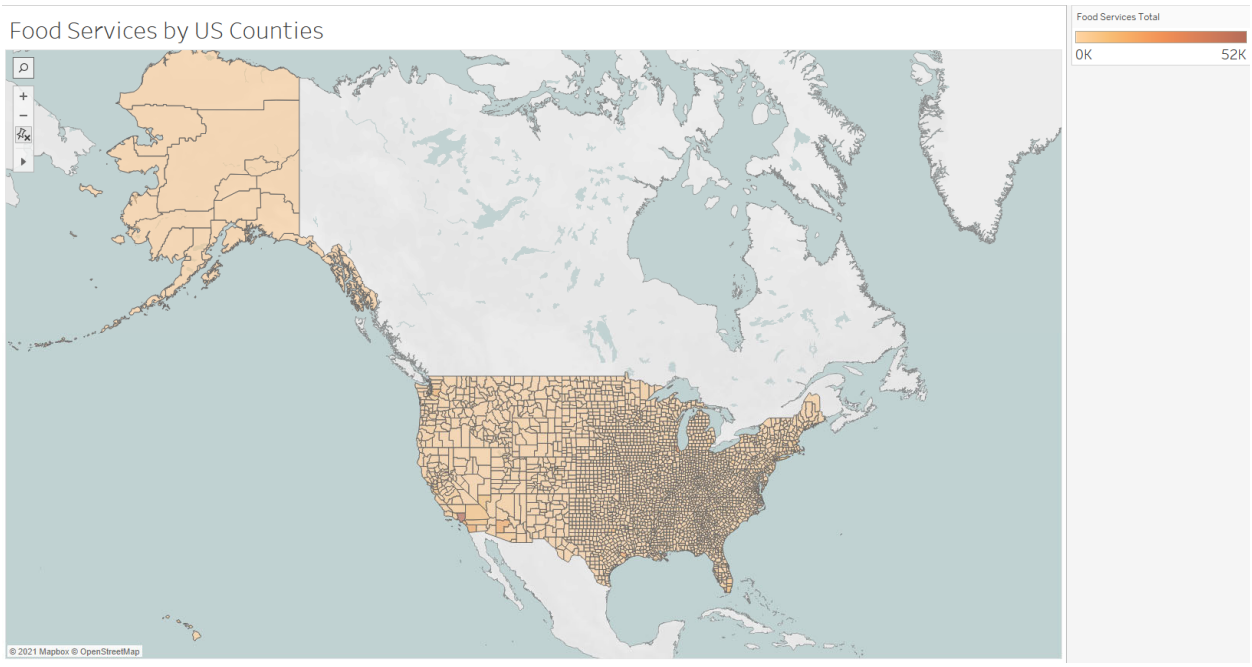
StudentId: 2020637

1. FoodSrvByCounty data set
    a. Food Services by State



In the above graph, we see different levels of food services available in different states of the United States of America. To come up with the graph, I had to first un pivot the years' columns data into the corresponding rows then computed a calculated field known as Food services total and used its aggregate measure to derive the coloring this graph. We see that California has the highest availability of food services, followed by NewYork, Texas and Florida.

    b. Food Services by county

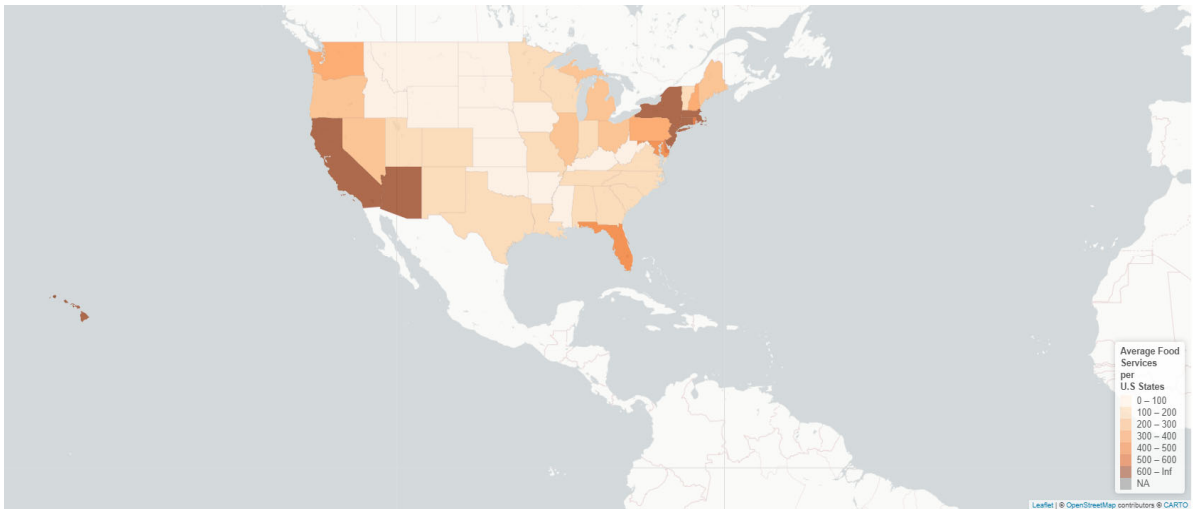


The above graph shows the different levels of food services available in the different state counties of the United States of America. I filtered out all those counties with null values in the latitude and longitude columns then using the same approach, I used the aggregate of the calculated field, food services total, to derive the coloring of the different counties within a given state.

It is shown that Los Angeles , Orange County and San Diego in California, Cook county in Illinois, Dallas and Harris counties in Texas have the highest availability of food services.

c. Extra credit

I. Average Food Services per U.S States

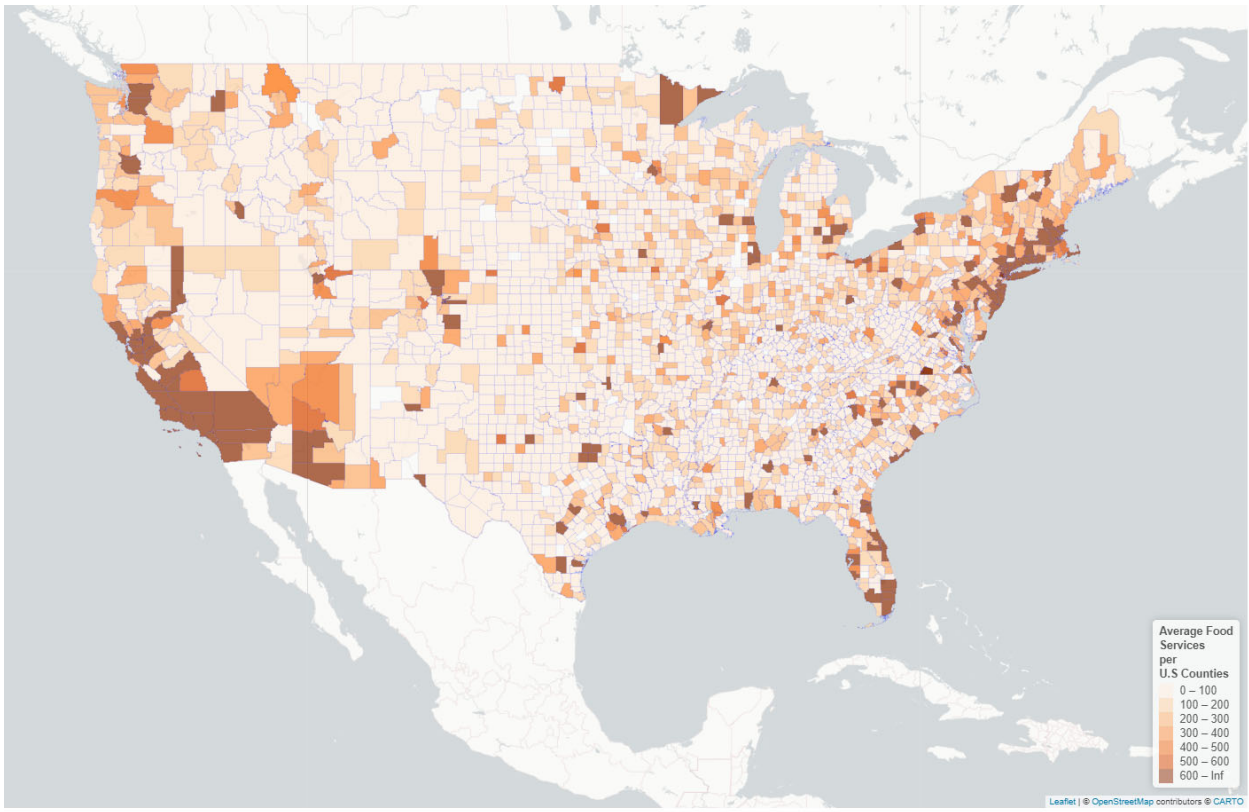


I had to cleanup the states field by downloading a states look up file which I used to clean up the states field in the food services data set. Also, I went a head to unpivot the food services years' data into rows and then used that to group the states together while finding the mean (average).

I then merged the states shape file with the food services data set and used the leaf let package to graph the map.

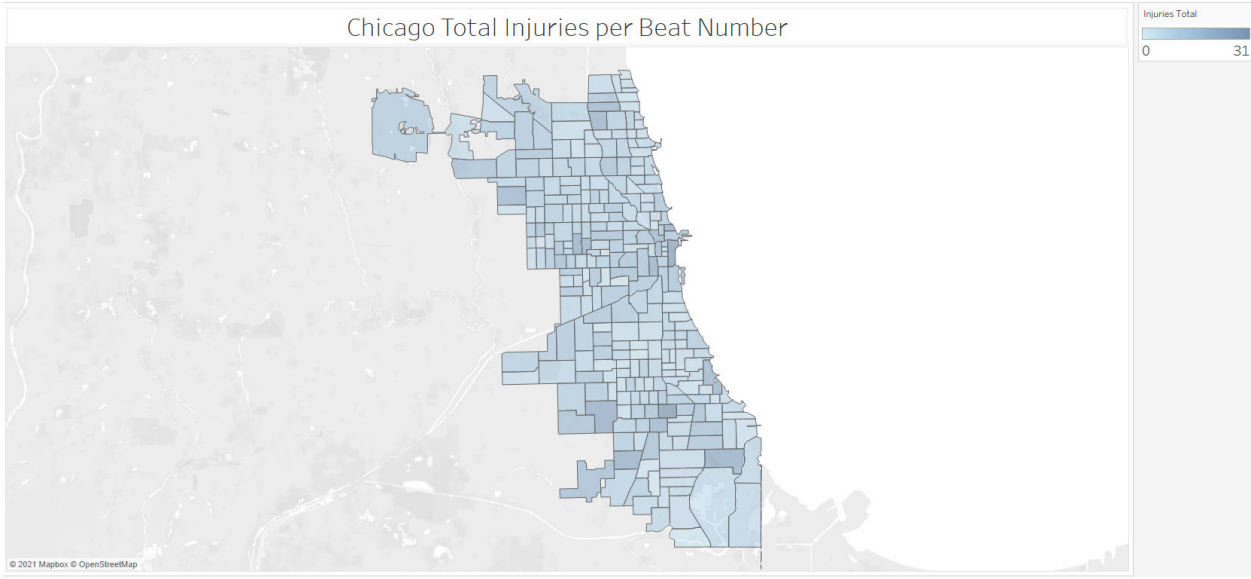
As demonstrated in the graph, California has the highest availability of food services and followed by New York.

II. Average food services per U.S counties



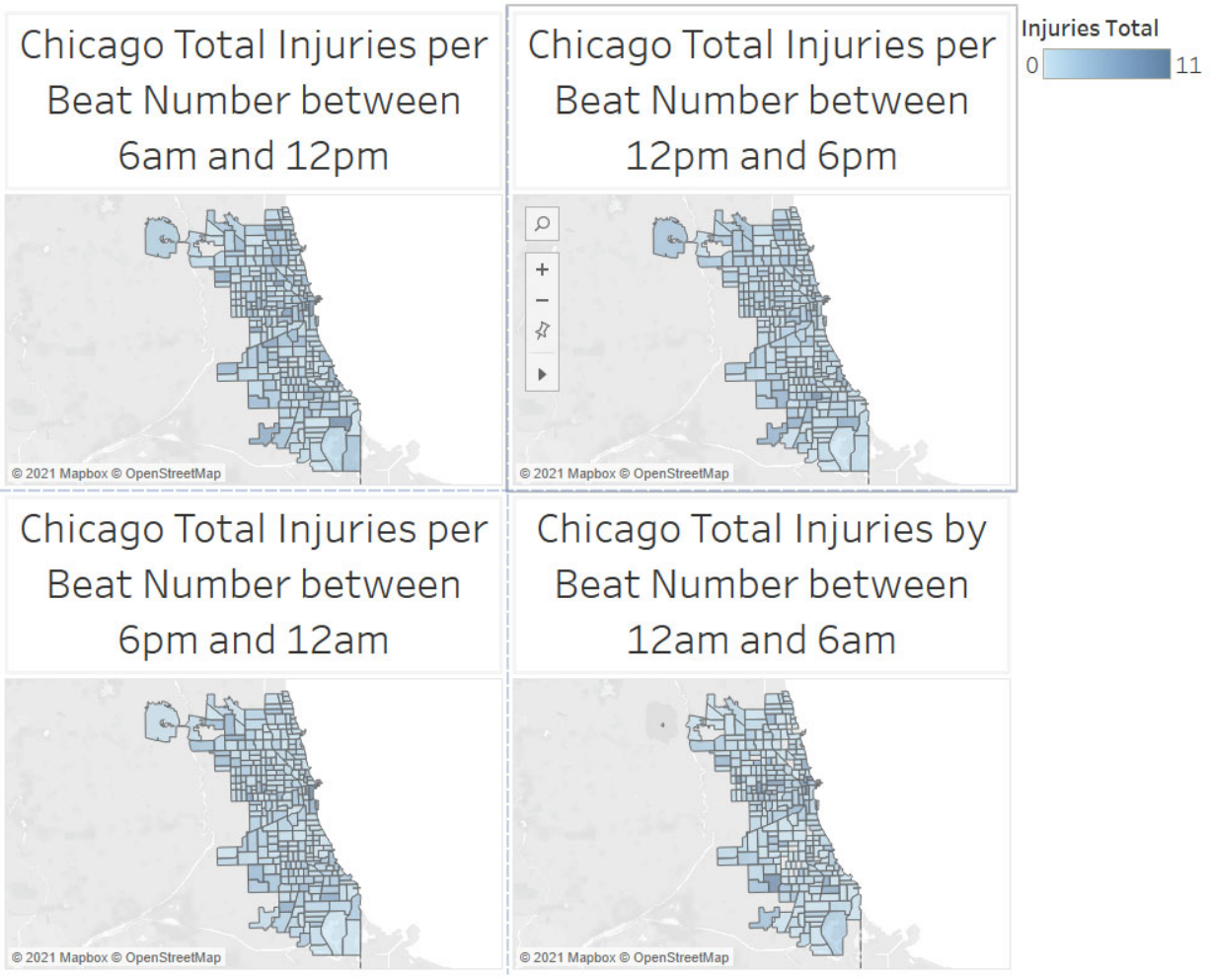
It is shown that Los Angeles , and San Diego in California, Cook county in Illinois, Harris county in Texas have the highest availability of food services.

2. Chicago_crashes data set
    a. Where total injuries occur



In the above the graph, I chose the latitude and longitude geographical variables to draw the map, used total injuries for coloring plus Beat number and geometry for details on the map. Beat Number 623 had the highest number of total injuries in Chicago followed by Beat Number 114.

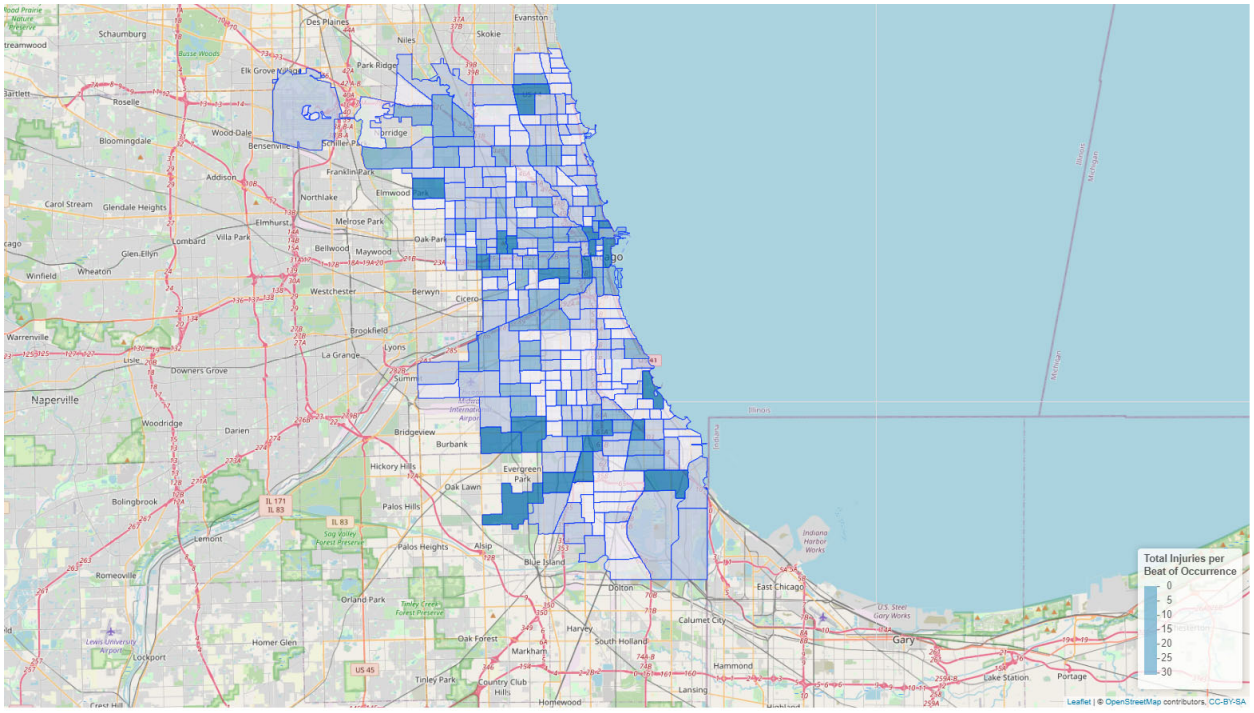
b. Common crashes in different parts of the city based on the time day.



From the above tableau dashboard, most of the accidents occur between 12pm and 6pm; there are lots of injuries that occur in the South side of Chicago around this time. Here we can differentiate the total injuries per given Beat Number based on the time of the day.

In R.

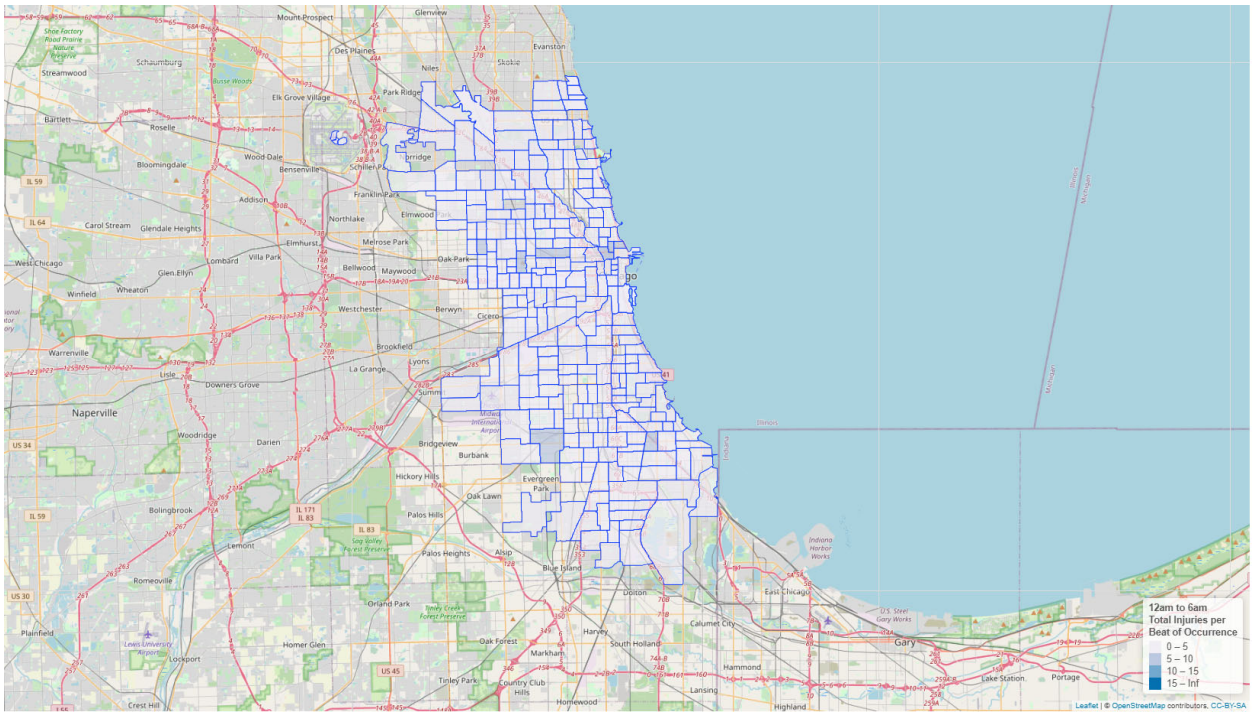
I. Where total injuries occur



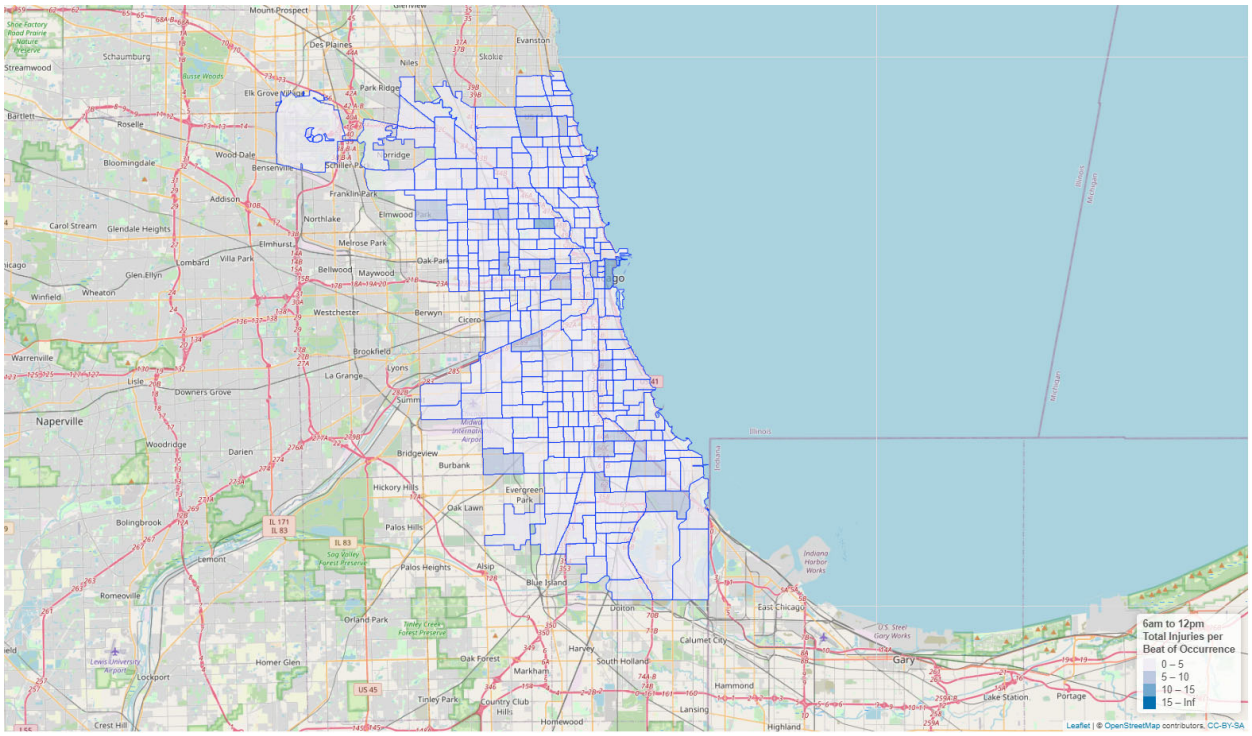
Before graphing this map, using html with R, I generated beat labels which I used as tool tip to display information about a particular colored square with the total injuries and the respective beat number corresponding to that square so that when the user hovers over with their mouse, that information is displayed.

II. Common crashes in different parts of the city based on the time of the day

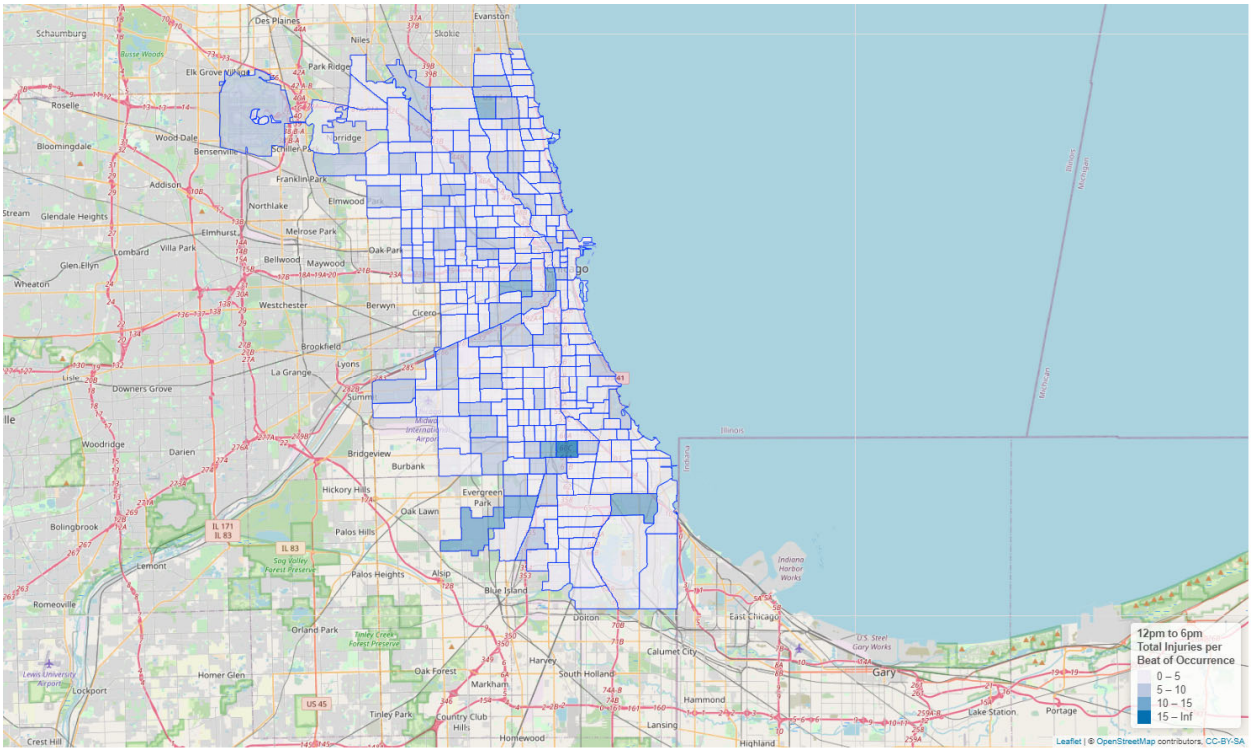
- 12 am to 6 am Total Injuries per Beat of Occurrence



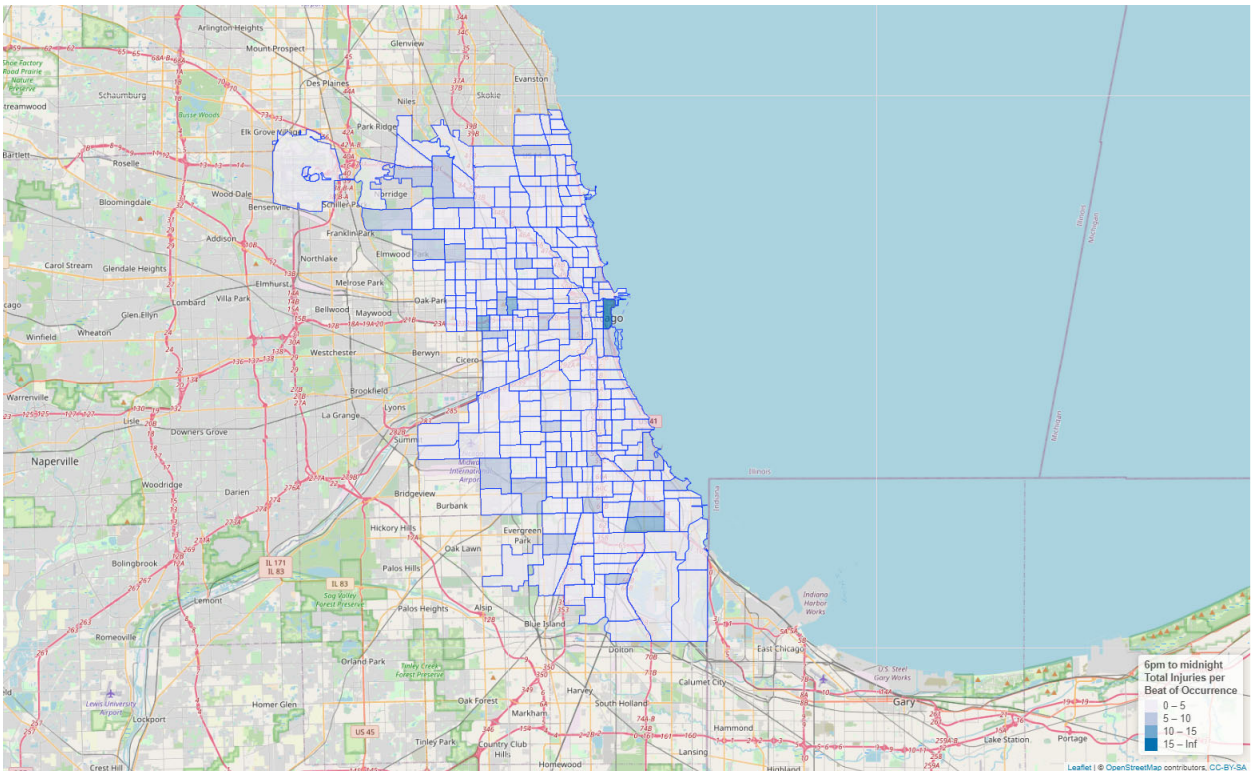
- 6 am to 12pm Total Injuries per Beat of Occurrence



- 12pm to 6pm Total Injuries per Beat of Occurrence.



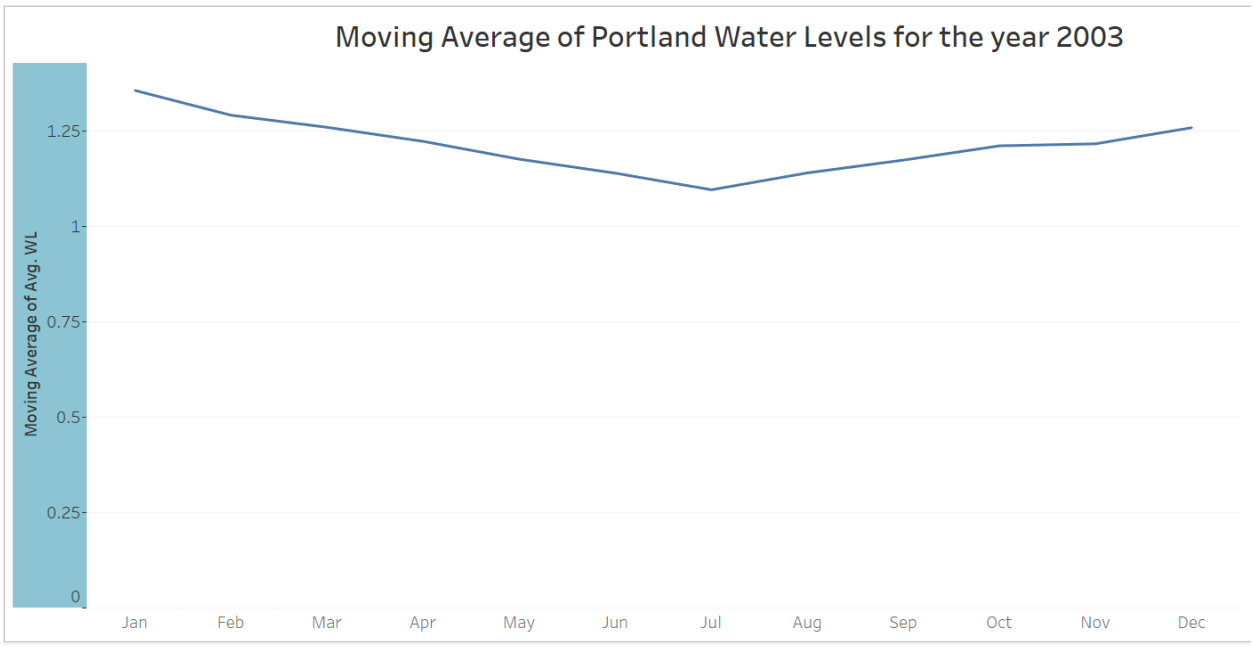
- 6pm to midnight Total Injuries per Beat of Occurrence.



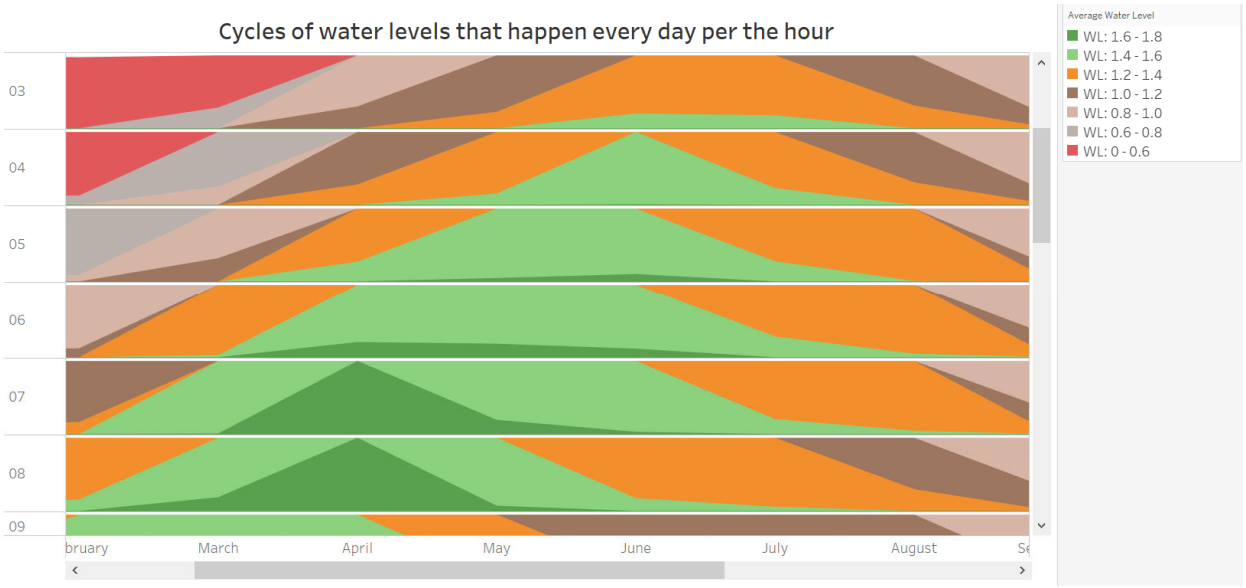
Before graphing these maps, I created 4 bins of 6-hour differences, such as 12am to 6 am, 6am to 12pm , 12pm to 6pm and 6pm to midnight, which I used as buckets to allocate the injuries that occurred within a given hour window. I went a head to group the beats together and sum up the total injuries, and then merged the shape file with the aggregated injuries data set of the respective beat numbers grouped together based on time of the day.

Based on the maps above, a majority of injuries occur between 12pm and 6pm implying that there are quite a number of accidents that occur between 12pm and 6pm in the city of Chicago based on the numbers that are reported.

3. Portland Water Level data set
    a. In Tableau



b. Horizon graph showing the cycles of water levels that happen every day per the hour



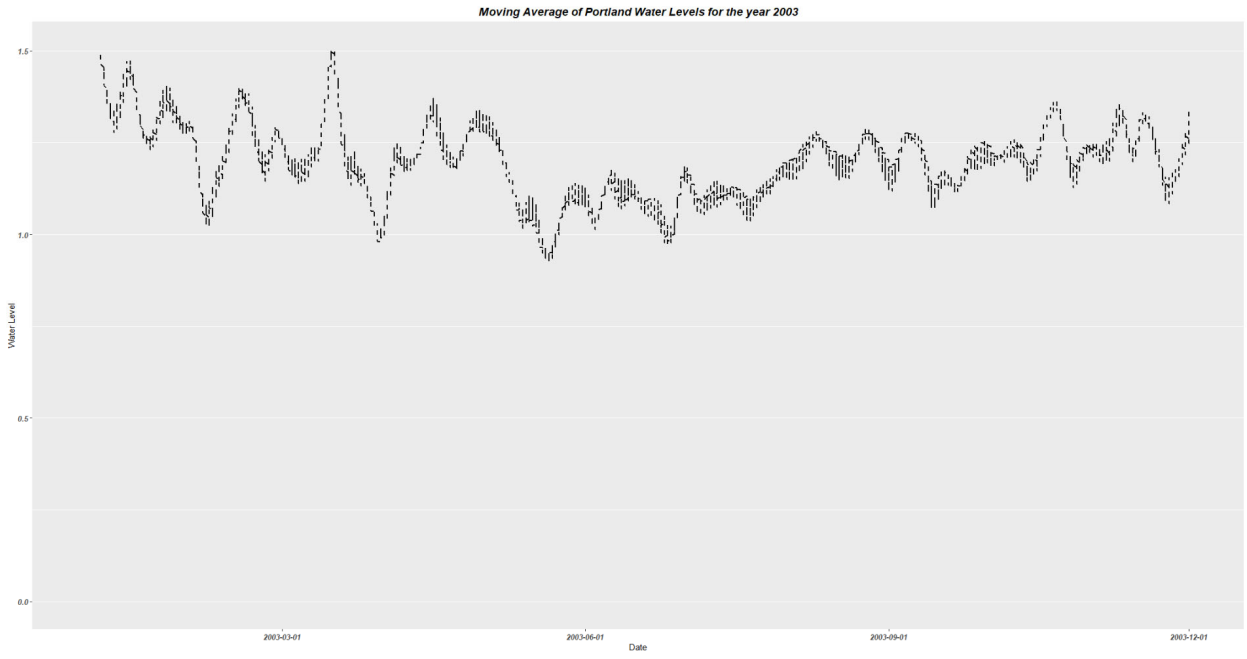
c. In the first graph, I used Date and WL variables construct the graph. When I dragged the Date field from the Data Pane to the Columns, I selected months. Also, on dragging the WL field on the rows from Data pane, I selected the Average measure as the aggregate measure and used the quick tools tab to derive the moving average for this graph.
As demonstrate above in the graph, the water levels keep on dropping between January and July. They start rising again in July when it is summertime.

In deriving the horizon graph, I put the date field on the column and used month as the function to display on the x -axis . On the rows, I dragged the calculated field, Event Hr, which I used on y -axis. I used the average water level to derive different average water level ranges which I used for coloring in Tableau and to provide the lowest level of detail, I used the average water level. Using multiple-selection, I dragged all the average water level calculated fields from the data pane to the rows to derive the horizon chart. This graph shows the comparison of water levels in different period of time in a day in a month within given hour.

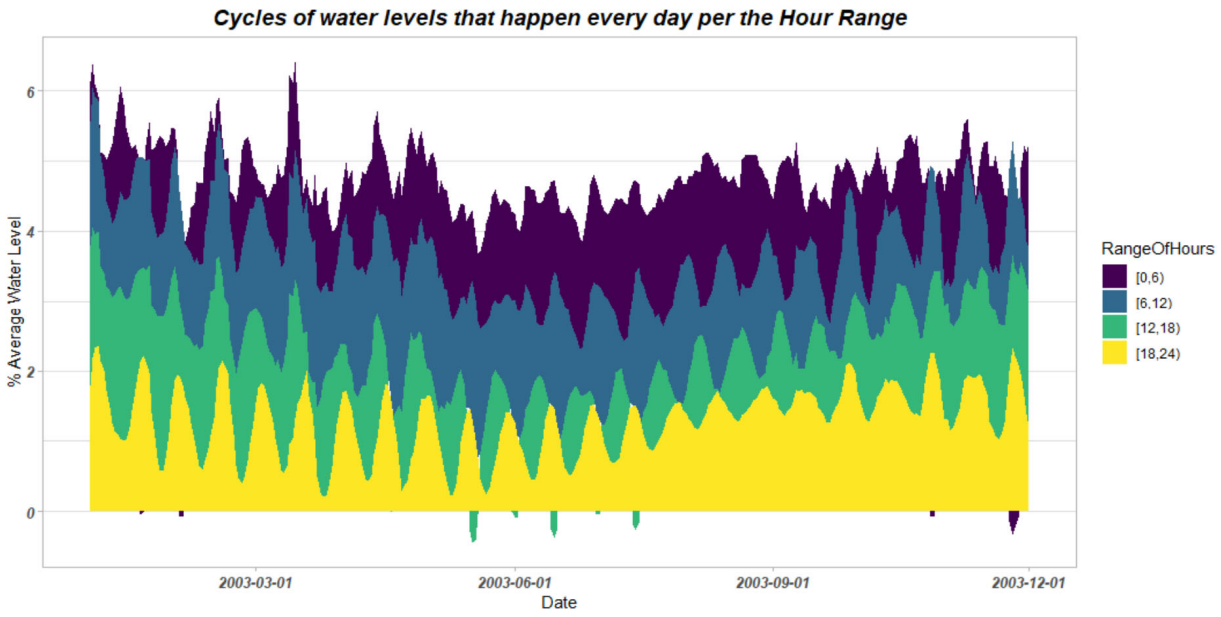
The major difference here is the graph in a shows the water level as a moving average in each month of the year from which we can understand the trend in water levels in the year of 2003 whereas the graph in b shows the moving average during specific times on the different days of the months. Furthermore, the graph in b shows the hourly changes in the water levels during the given month

In R

a. Moving Average



b. Horizon chart



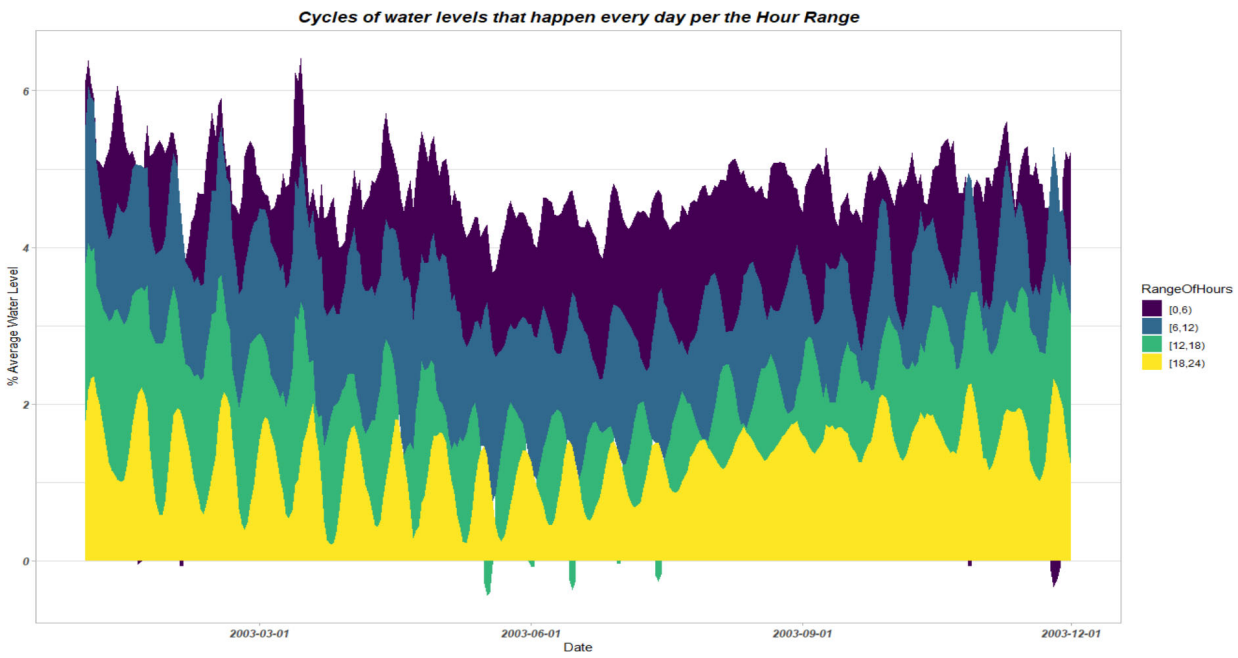
c. The first graph in r, I used the ggplot() function with geom_ma() layer to derive the simple moving average .

In second graph, I created 4 bins of 6 hours differences then grouped the data by Date and Range of Hours and derived the average of the water level.

I used ggplot() function with the geom_area() layer to graph the area map, used Range of Hours field as a color fill field. In order serve color blinded people, I used the scale_fill_viridis() to provide color maps that are perceptually uniform in both color and black-and-white.

The major difference here is the graph in a shows the water level as a moving average in each month of the year from which we can understand the trend in water levels in the year of 2003 whereas the graph in b shows the moving average during specific times on the different days of the months. Furthermore, the graph in b shows the hourly changes in the water levels during the given month. The graph in b clears shows that during the colds of months of Nov, Dec, Jan, Feb 2003 when the hours are in the range of 0 to 6, which is 12midnight to 6am, the average water levels are below 0 percent.

4. Divergent Color Scale with Water levels at the mid-point



The above graph was constructed using R using date and the average water level on my x and y aesthetics. I used ggplot() function with the geom_area() layer to graph the area map, used Range of Hours field as a color fill field. In order serve color blinded people, I used the scale_fill_viridis_d() to provide color maps that are perceptually uniform in both color and black-and-white. The divergent color scales were used to show the water levels that occur within the various hourly intervals throughout the year of 2003. It is shown that water levels drop below 0 percent between 12midnight to 6 am between Jan and March and also around Nov and Dec, which explains that around that time of year it is always cold with snow and ice.
Furthermore, water level drop below zero percent between May and July when it starts to warm up between 12pm to 18pm, which explains that around that there is a lot of humidity which is caused by the high temperatures.