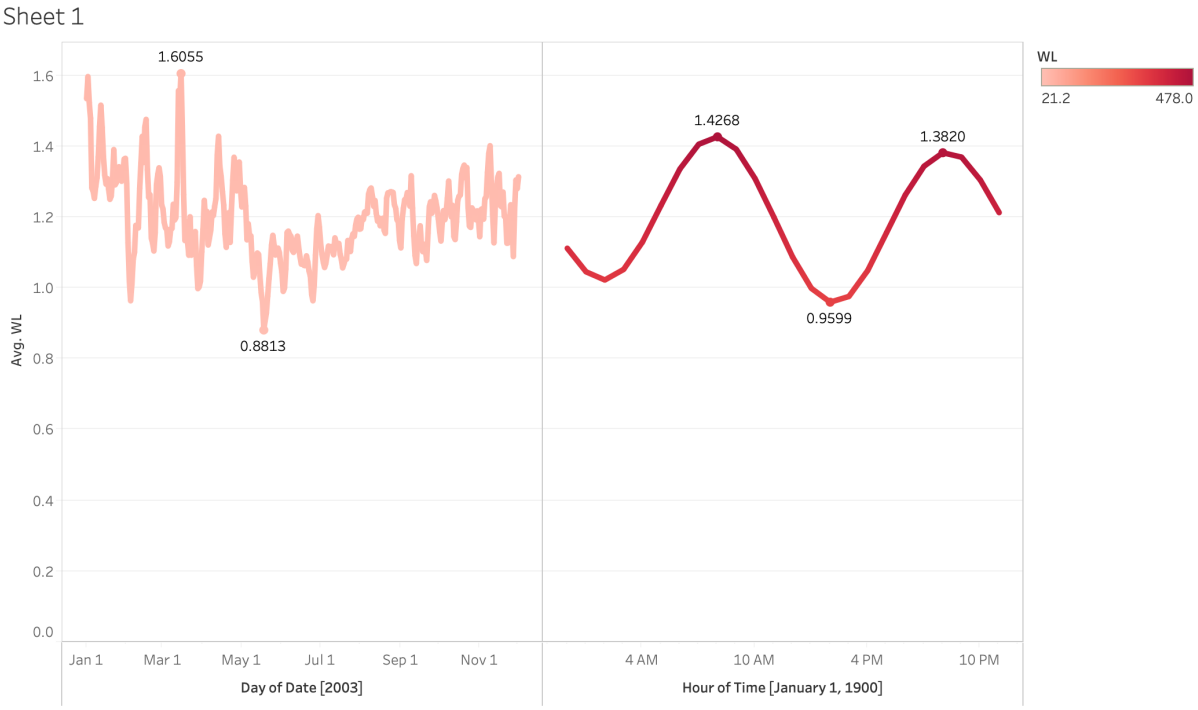
HW3
Chaonan Shi
1901412

Problem2:
2a
From Tableau:



The trends of average of WL for Date Day and Time Hour. Color shows sum of WL. The data is filtered on Time Hour, which keeps 25 of 25 members. The view is filtered on Time Hour, which keeps non-Null values only.

In this question, I set up the color based on the WATER LEVEL with red, so that deeper red represents higher WATER LEVEL;

I also set up the TWO data variables: “day of data” and “hour of time” to measure the water level as question asked;

As we can see right graph above, the highest water level happens on each day is on the 8AM with 1.4268 by average and lowest water level happens on the day time is on the 2PM with 0.9599 by average;

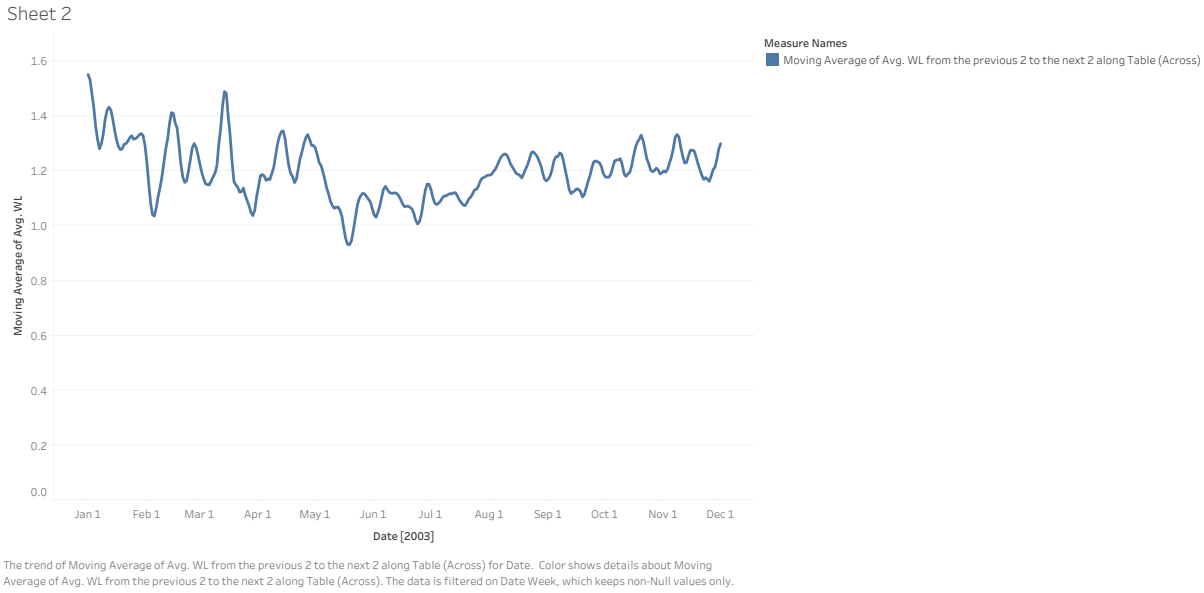
As we can see left graph above, the highest water level happens on the day of data in the year is at the March 15th with 1.6055 by average and lowest water level happens on the day of data in the year is at the May 18th by average.

2b
From Tableau:

In this question, since there are too many data in table, I clean it up by smoothing the data by calculating a moving average:

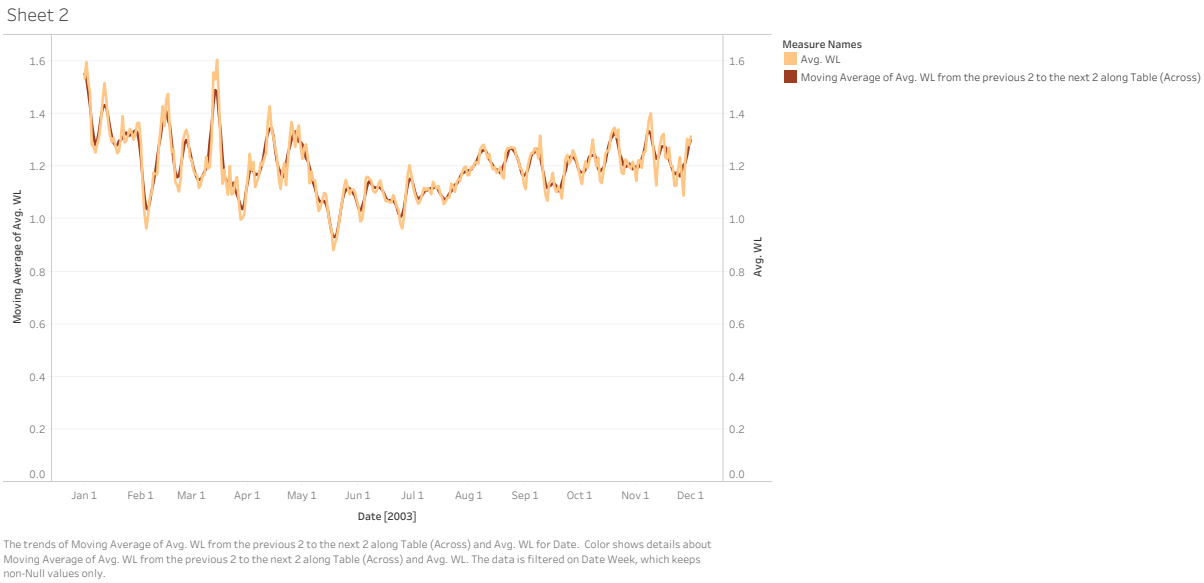
In this case, I set up the moving average of previous value from 2 to next value 2:

From Tableau:



Then use a window approach with window size that covers a range of days suitable to smoothing out the weekly variation and showing the overall trend:

From Tableau:



In this case, I changed the measure of dual axes as Average as well, since the default SUM will have different measurement which mess up the result. So that I keep the range of both two axes as fixed from 0 to 1.7.

Moreover, I set up the color based on the WATER LEVEL with red, so that deeper red represents higher WATER LEVEL;

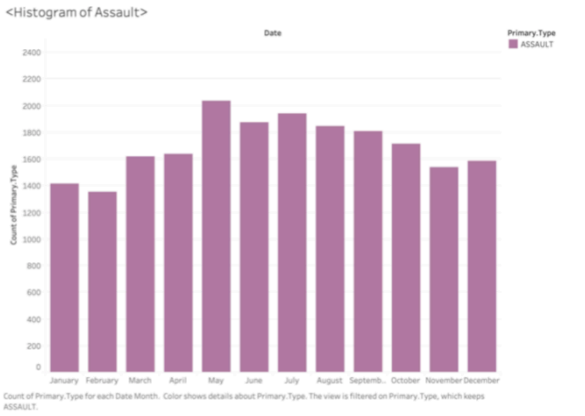
2c
Based on the above different two graphs, the first graph focus on the overall trend from dataset, and tell us the results in the big picture. In this graph, we can summarize the information and also can capture the extreme value (outliers) based on the whole dataset.

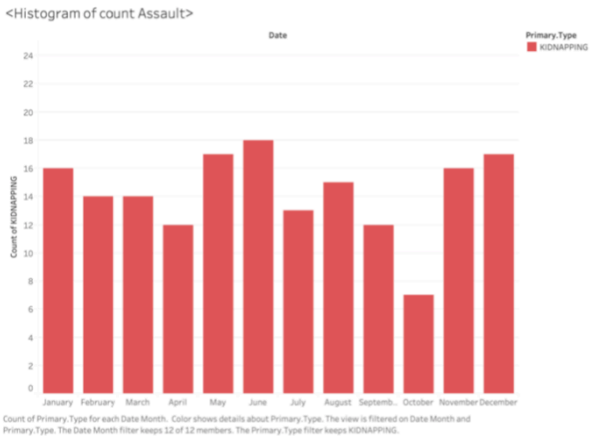
However, the trend is actually changing all the time in some cases especially for the time-series dataset. For example the weather, stock price, and the water level in this case. So the second graph smooth the dataset by using the window at some particular time period to calculate the average number based on that time range. In this case, since chart patterns can be difficult to read given the volatility in water level movements, moving averages can help smooth out these erratic movements by removing day-to-day fluctuations and make trends easier to spot.

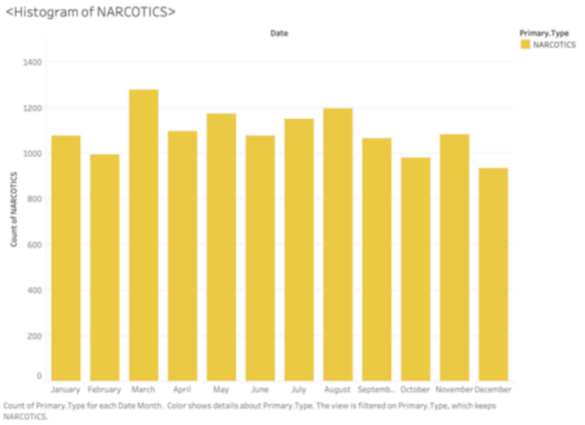
Problem3:

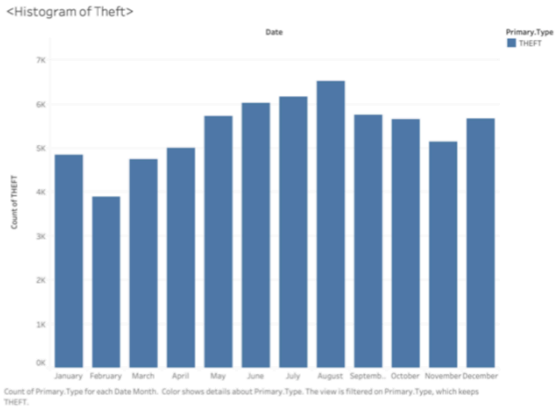
3a:

From Tableau:



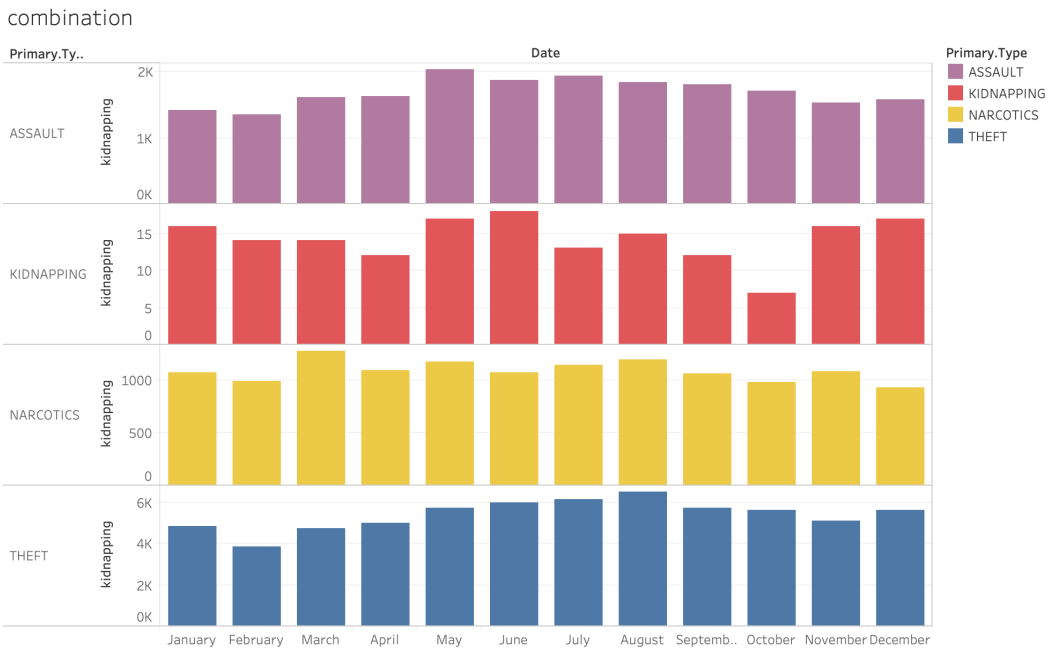




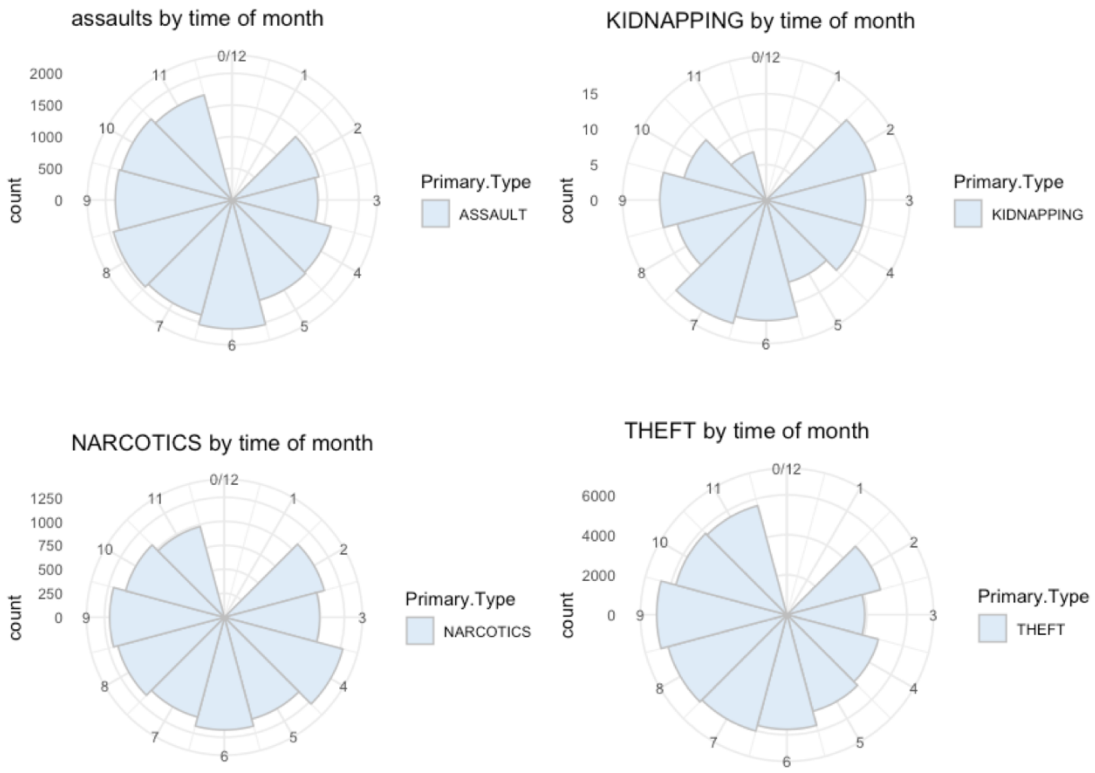


For assault, I fixed the range from 0 to 2500 for better display;
For kidnapping, I fixed the range from 0 to 25 for better display;
For NARCOTICS, I fixed the range from 0 to 1500 for better display;
For theft, I fixed the range from 0 to 7500 for better display;

Moreover, I use Tableau to combine those four different graphs by editing the filter with Y-axe as "*Independent axis ranges for each row or column"* so that they will display their own histogram graph into one graph with date in X-axe:



3b:
From RStdio:



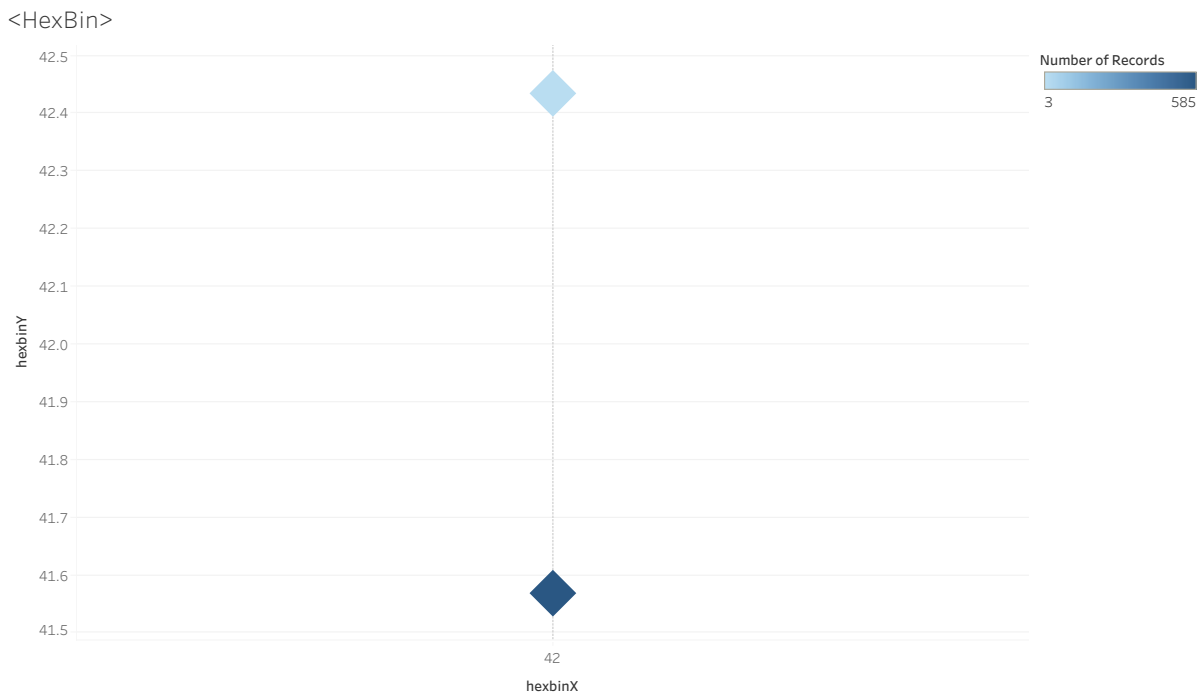
3c:
For the bar graph, since it is time-series dataset, so that data distributed followed by the actual time (month) in this case, it also helps us to see relationships quickly. However, bar graphs can be difficult to read accurately. A change in the scale in a bar graph may alter one's visual perception of the data. For example, in this question, the scale of each categories are quite different, it may misleads people to read the different bar by same scale.

Moreover, if bars are close to each other, then it is hard to observe the difference between each period. For instance, the length of bar of assault in Jan and Feb are quite similar, we are hard to differential them. On the top of that, since the bars sorted by time-series, even make it harder to compare with each other cross the bar.

For the rose plot, it makes more sense in this question comparing with bar graph. The rose-plot compact the time as a circle or ring, using the radius to illustrate the value or differences, so it not only makes bars easier to read, but also saves space as well. For the time-series data set, it makes more sense to people that gather time data as “clock” and observe the difference among with different period.

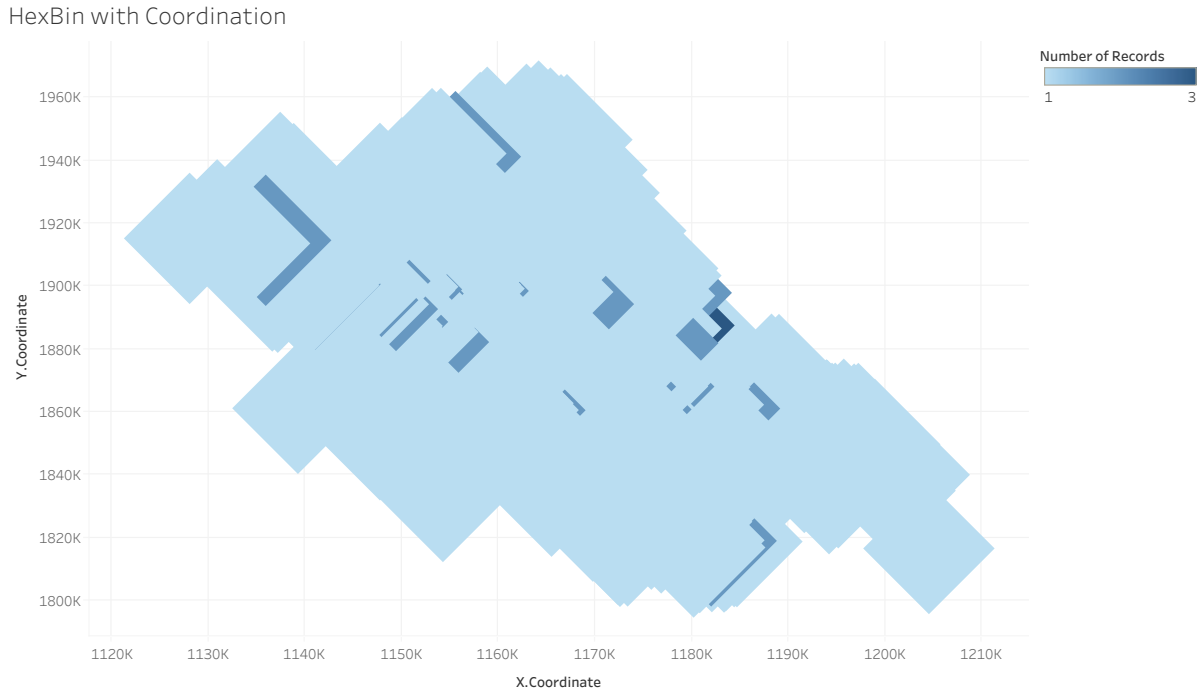
However, the rose-plot also have negative aspects. For example, if we put too many information in the rose-plot, the graph will mess up and hard to read. Second, the scale is also limited for rose-plot, if we get too many levels or scales, the graph would not display clearly, even tick marks overlapping each other.

3d:
From Tableau:



HexbinX vs. hexbinY. Color shows sum of Number of Records. The data is filtered on Primary.Type, which keeps HOMICIDE.

At first, I create HexbinX and HexbinY by using Latitude and Longitude, then put them as column and row (X axe and Y axe) respectively. However, since the Longitude is negative number, so it is hard to observe the pattern by using Latitude and Longitude directly. In this case, I decide to use the X and Y Coordinate variables as my x-axe and y-axe.

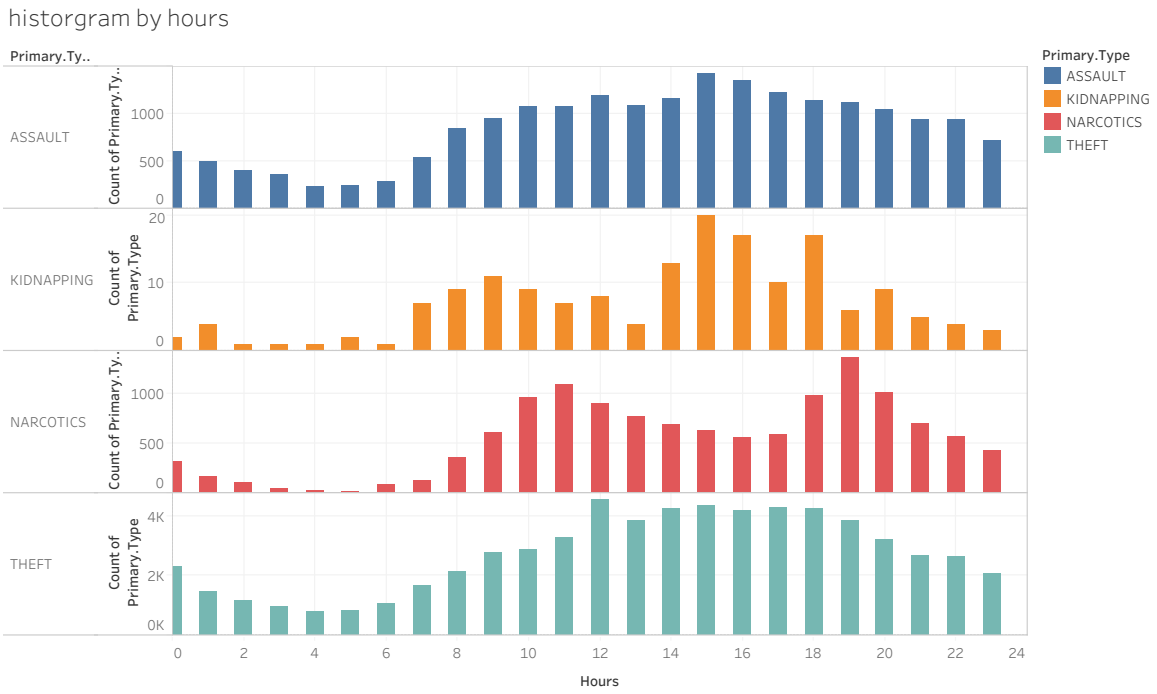


X.Coordinate vs. Y.Coordinate. Color shows sum of Number of Records. The data is filtered on Primary.Type, which keeps HOMICIDE.

In this case, coordination variables performance better than the Latitude and longitude. I also enlarge the size of observations so that we can observe the density of area with high homicide rates.

As we can see, the area of 1140k at x-coordinate with 1920k at y-coordinate, 1155k at x-coordinate with 1890k at y-coordinate, and 1180k at x-coordinate with 1890k at y-coordinate have high homicide rates.

3e (Extra credit):
(1)
From RStudio:

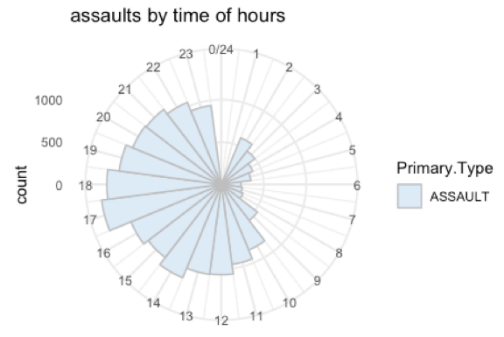


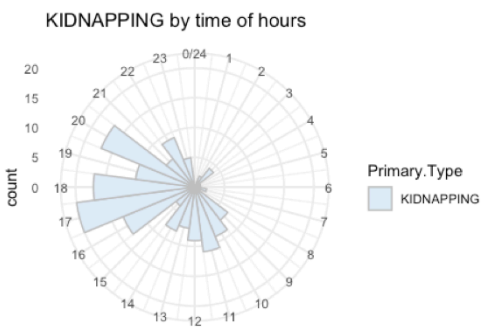
The plot of count of Primary.Type for Hours broken down by Primary.Type. Color shows details about Primary.Type. The view is filtered on Primary.Type, which keeps ASSAULT, KIDNAPPING, NARCOTICS and THEFT.

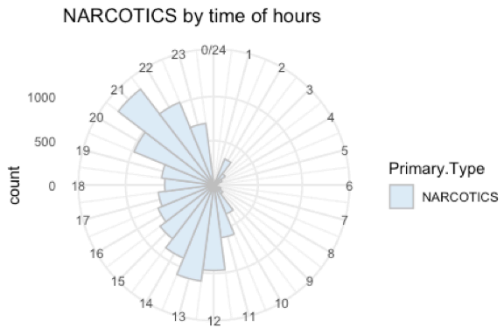
Analysis:

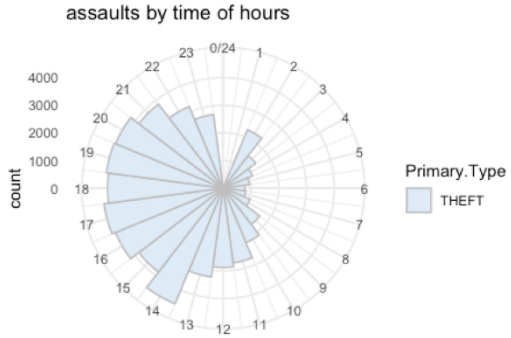
As we can see, the four types of crime happens more frequent from afternoon to night, which is 8:00 to 22:00.
The peak of ASSAULT happens at 15:00;
The peak of KIDNAPPING happens at 15:00;
The peak of NARCOTICS happens at 19:00;
The peak of THEFT happens at 12:00;

(2)
From RStudio:

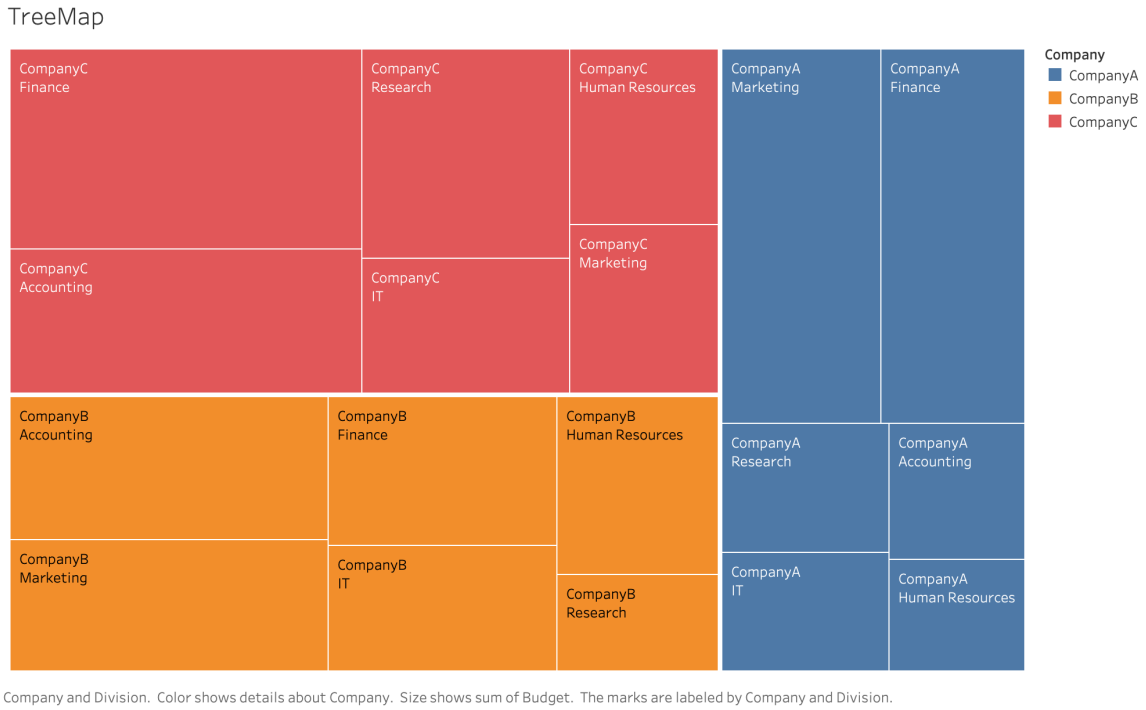








4a:
From Tableau:



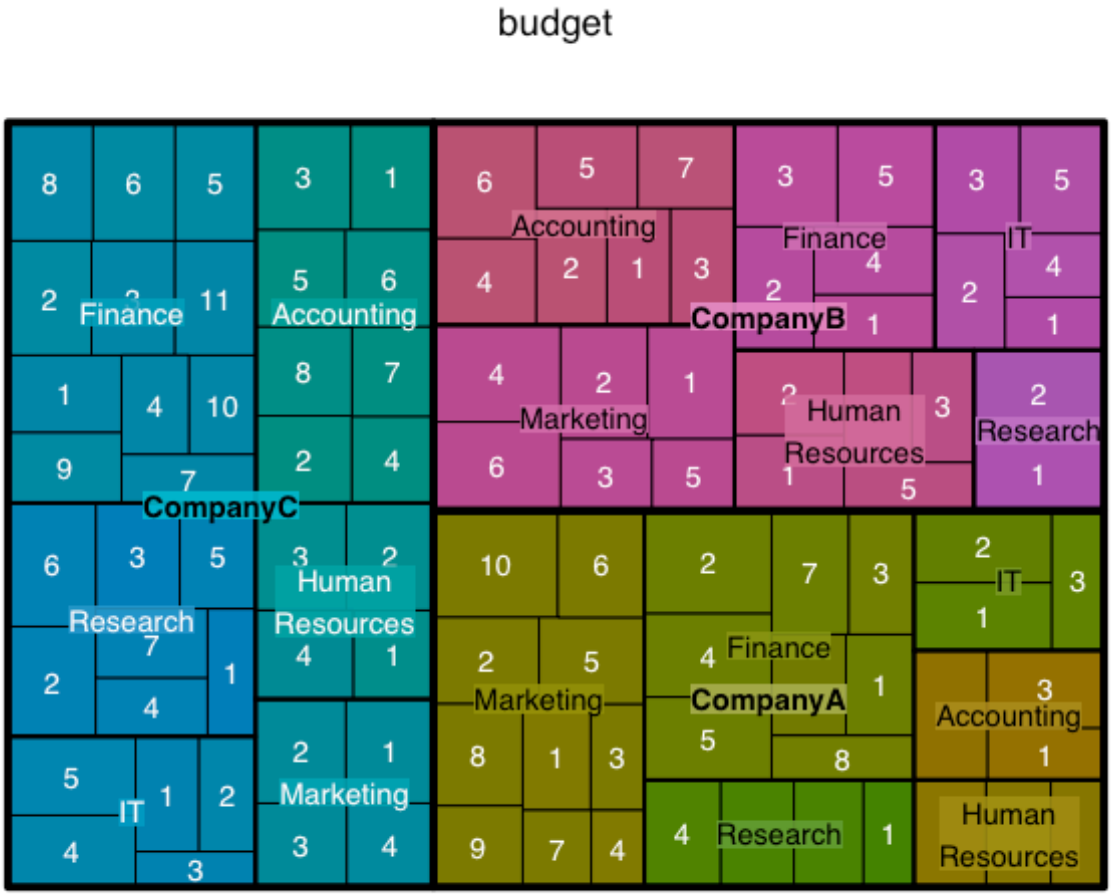
Analysis:
Set up the color by the company and also the division cells should be sized by the total budget for the division.

In the summary, the company C with finance division owns highest budget compare with division of Marketing owns lowest.

The company B with Accounting division owns highest budget compare with division of Research owns lowest.

The company A with finance division and marketing division have almost same budget own more budget compare with division of Human Resources owns lowest.

4b:
From RStudio:



Analysis:

In this case, I set up the size by “budget” for each “office”, also set up the color by “company”. So that each company owns their own color but divergent by divisions;

As we can see, the company C divergent by “blue”; company B divergent by “purple”; and company A divergent by “green”.

4c:

To compare b with a, the three levels tree-map obviously gives more information and detail about the dataset. For example, the second tree-map not only gives the information about the “company”, “division”, and “budget”, but also the exact office information: which division owns how many offices, even the distribution of offices as well, for instance, the finance from company C owns more budget compare with others, also the the mode is 8.

However, it might not feasible for all three levels with a tree-map. For instance, it is easy to mess up the graph by setting wrong color options, especially when we do not specify the color, and it is also not good at fine distinctions, exact measurements, and communicating proportional differences. Intuitively, different subtleties colors also give us different sense of readability and interpretability.