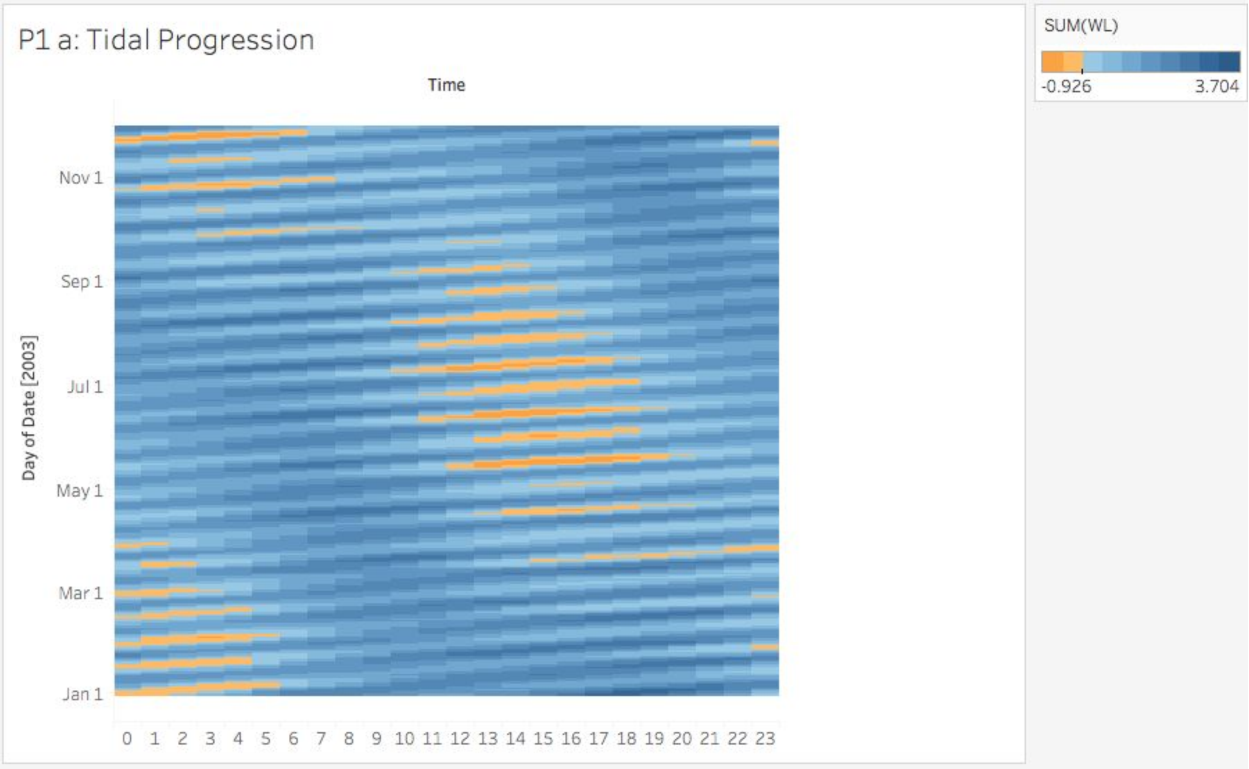
Alex Teboul
DSC 465: Assignment 3
Due: 5/22/2019

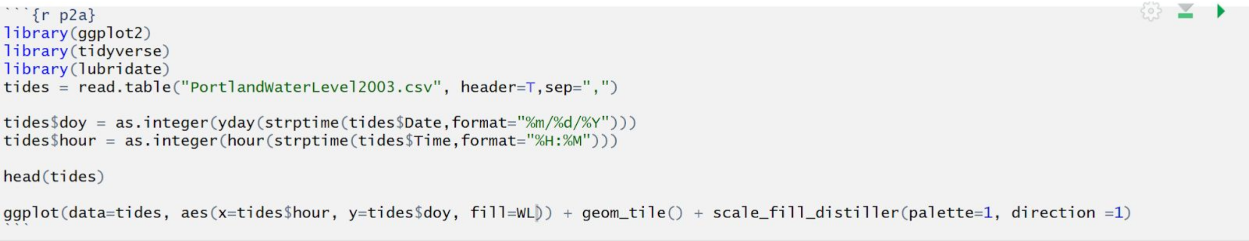
Problem 2 - Tableau & R
2) Download the Portland Water Level dataset and explore it by creating the following visualizations of the time series from the techniques described in lecture. Use both R and Tableau for at least one graph here. They should, of course, adhere to the design criteria that we've learned, and should clearly display the information described in each part.

Then write a single paragraph outlining the differences between the information that each graph communicates.

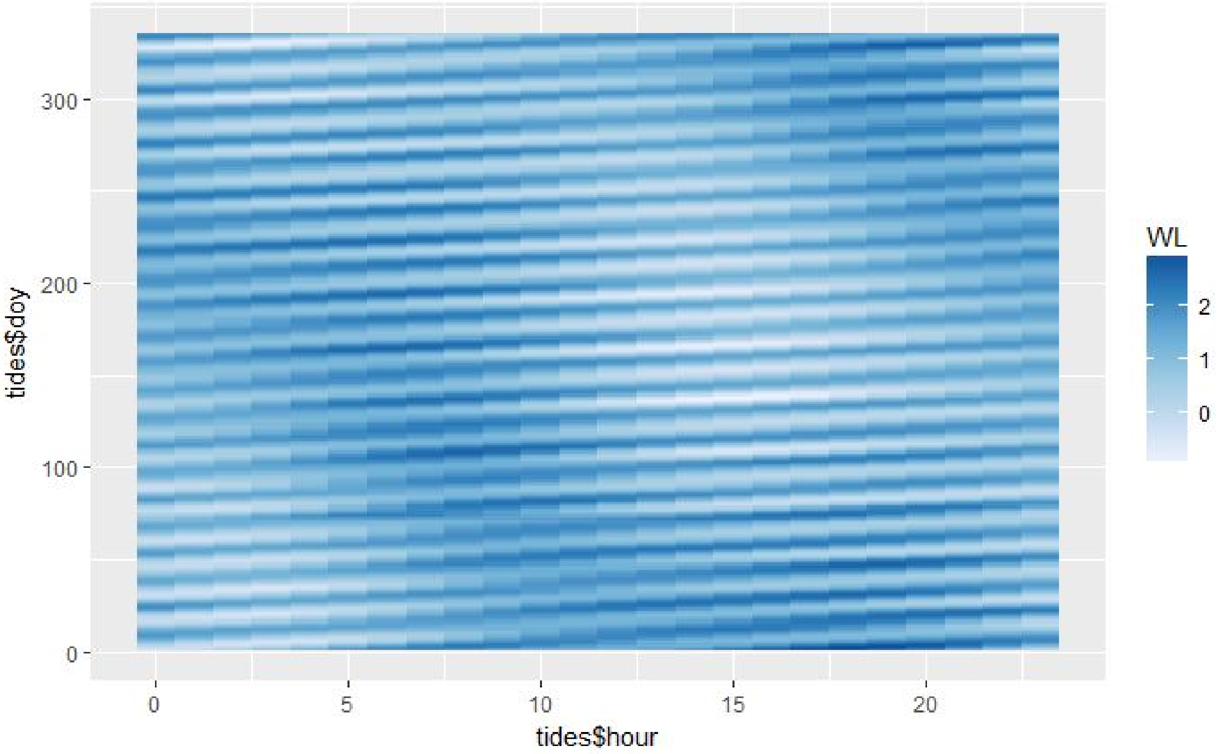
a. This data contains a year of data with water level (WL) measurements every hour as a function of Time (i.e. 365 x 24 data points!). Graph the cycles that happen each day (because of tides) as a level plot with hours (extracted from the Time field) on the x-axis and Date on the y. This can be done quite easily both in Tableau and R as shown in class. The plot should show the progression of the tides over the days of the year. Work with the color scheme to make it communicate the trend of high tides most clearly, and explain in your analysis below your reasoning behind the choice.



- In Tableau chose to show high tides in darker shades of blue as one would expect with deepening water, and low tides as a orange-brown to mimic the colors of sand and dirt that would be visible in low tides.

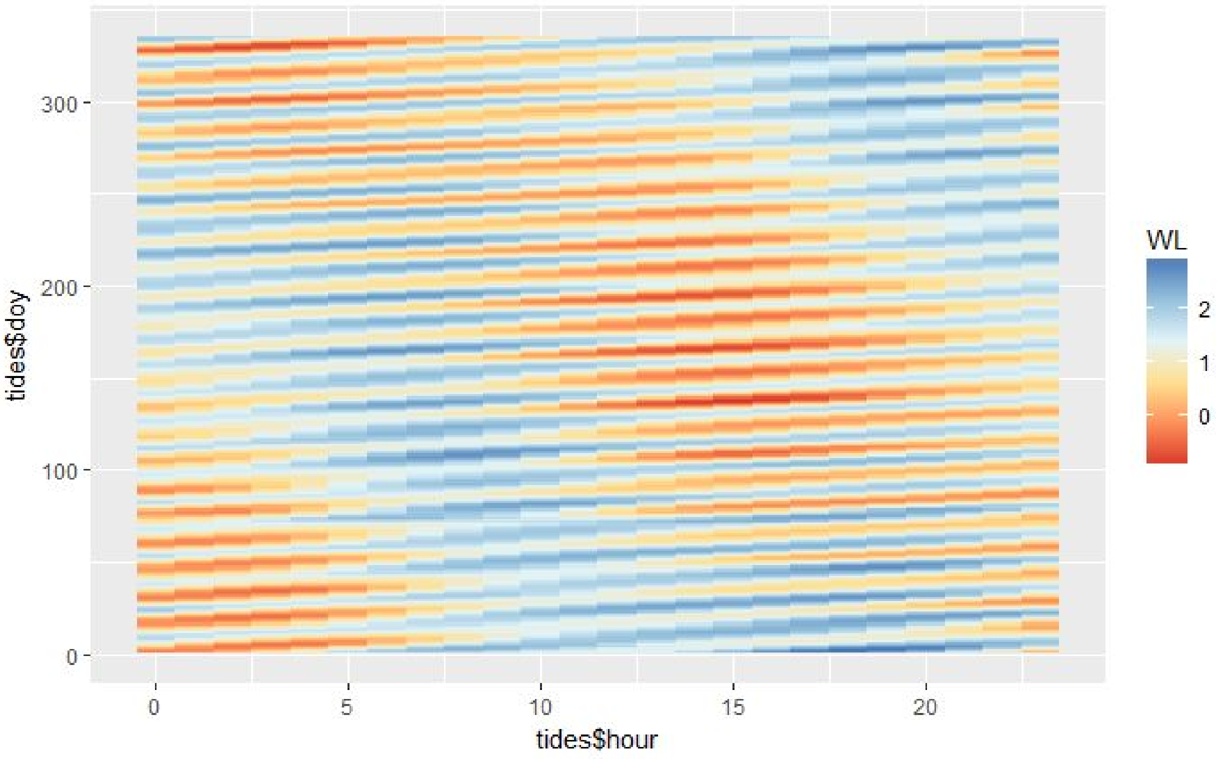


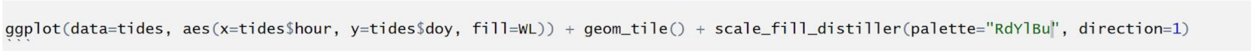
~~~
```{r p2a}
library(ggplot2)
library(tidyverse)
library(lubridate)
tides = read.table("PortlandWaterLevel2003.csv", header=T,sep=",")

tides$doy = as.integer(yday(strptime(tides$Date,format="%m/%d/%Y")))
tides$hour = as.integer(hour(strptime(tides$Time,format="%H:%M")))

head(tides)

ggplot(data=tides, aes(x=tides$hour, y=tides$doy, fill=WL)) + geom_tile() + scale_fill_distiller(palette=1, direction =1)
```
~~~



- In R, I first went for the simpler color scheme. Dark shades of blue indicate deeper water (i.e. higher water level values), and the light shades are for shallower lower water levels.

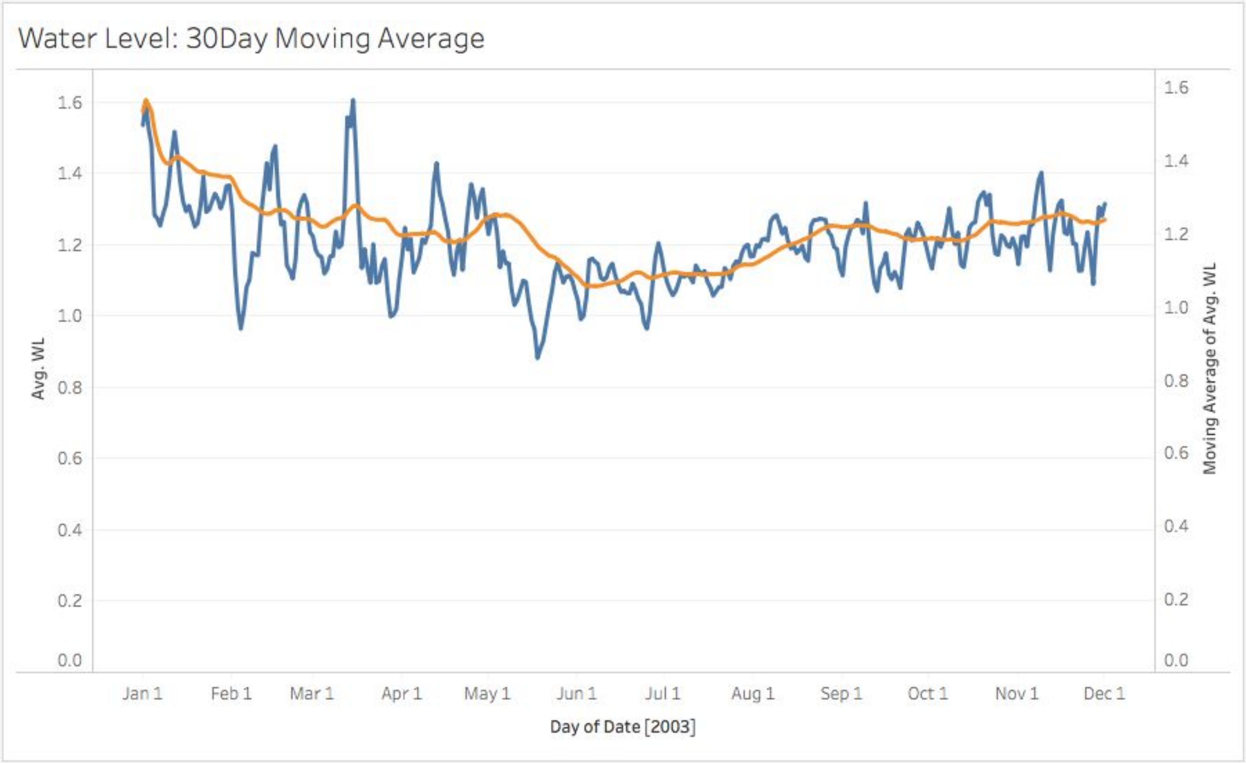
- Then I changed the color scheme to suggest that lower water levels were red, then sea level yellow, and above sea level in blue to show deep water. In this new graph shown below, it is very clear when the high water levels appear as they are in blue. So this final chart is the one I think best conveys the intended message on the time of high water levels throughout the year.





```
ggplot(data=tides, aes(x=tides$hour, y=tides$doy, fill=WL)) + geom_tile() + scale_fill_distiller(palette="RdYlBu", direction=1)
```

b. For this part, first aggregate the data by calculating the daily mean of the water level and then plot the result as a line graph. This can be done in Tableau quite easily, and in R if you use the dplyr “summarize” function as I showed in class in the first couple of weeks. Even with this, there is still a lot of data, clean it up by smoothing the data by calculating a moving average (Quick table calculation in Tableau or the rollmean function in R). 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. Graph the smoothed result over the previous result (e.g. dual axes in Tableau).



- In order to smooth out weekly variations I used the 30 day moving average overlayed on top of the original daily mean water levels. The moving average is show in orange.

c. Write a reflection paragraph on what each of the two graphs tells you about the data. What are the strengths of each graph?

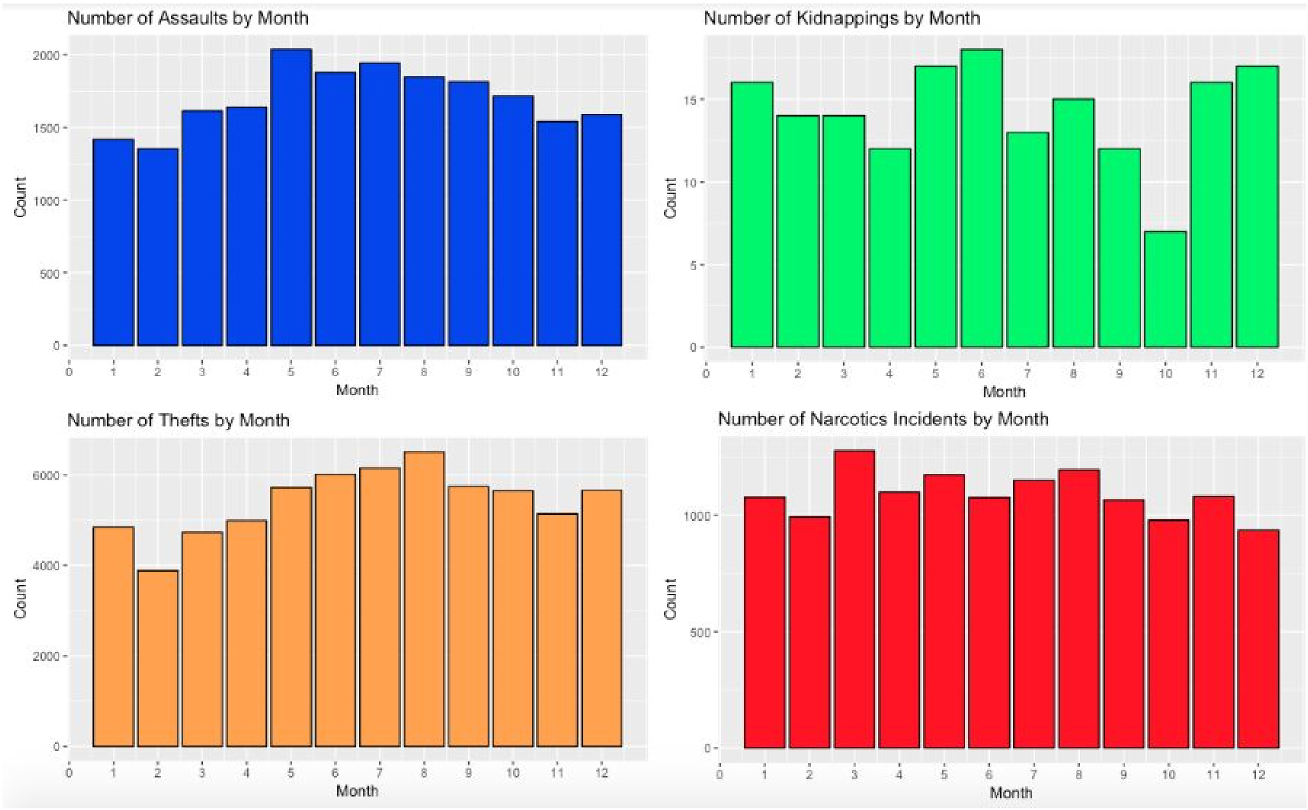
- The level plot tells us about the variation of water levels by hour throughout the year. Displaying it visually in the form of a level plot with colors that brings attention to the changing water level works well for this dataset. We can see that each day there are high and low water levels corresponding with the tides. At a higher level, we see that at the beginning of the year low tides take place in the early hours of the day (close to midnight), and as the year progresses, these tides gradually shift in a cycle that brings them back to early hour low tides roughly by the end of the year. The strength of the level plot is in in seeing the hourly water levels for the whole year and visually representing the water level scale with colors.

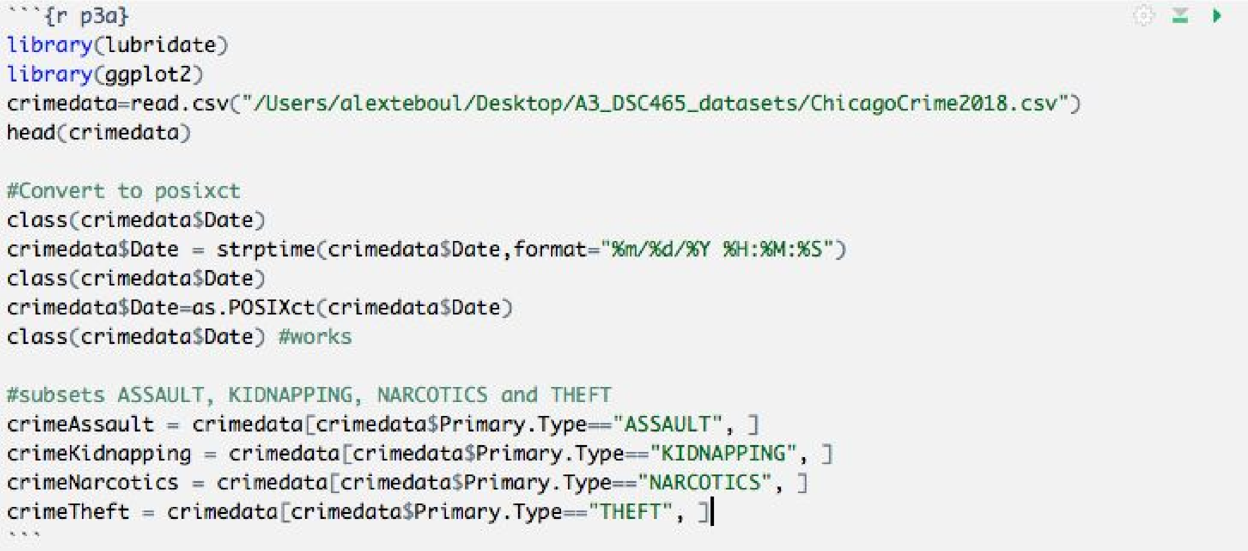
- The line graph tells us more clearly how average daily water levels fluctuate throughout the year. By superimposing the 30 day moving average we can see the year's trend without weekly variation. It makes sense that there is only a small variation in daily average water level throughout the year. Each day there are high and low tides, with the average level recorded in the graph, and it varies by about a foot throughout the year on average. The strength of this graph is to show the general water level trend that exists in the data over the course of the year.

Problem 3 - R
3) Download the ChicagoCrime.csv dataset which contains a list of crimes committed in the city of Chicago in 2018. To prepare the dataset for visualization, you will need to perform the following cleanup steps

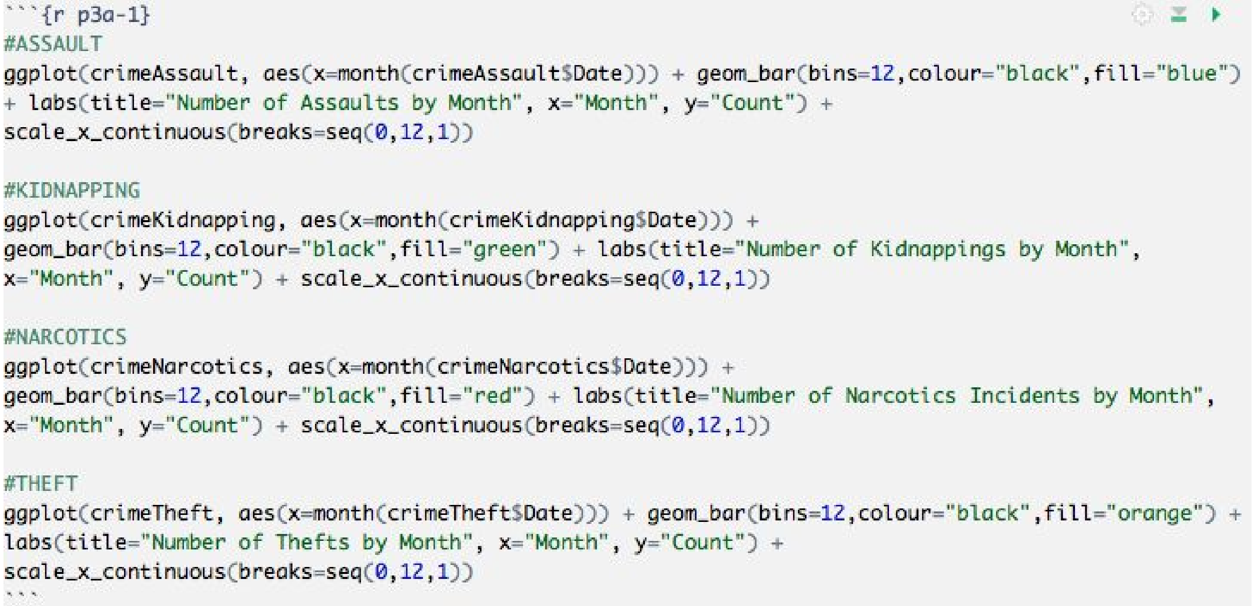
With this dataset, create the following visualizations

a. Subset separately on four of these crimes (ASSAULT, KIDNAPPING, NARCOTICS and THEFT) and create histograms of these by the Month. In R, the record count will be the default for geom_bar() and if you load the "lubridate" package, you can extract the month from a DateTime POSIXct field with "month(DateTime)", in Tableau, you can select the month of the year independently of what date it is in from the drop-down menu on the date in columns and selecting "month" in the first set of date options (not the second set that has the year attached in the example on the right side of the menu option), then use the "number of records" measure. Color the bars in each graph a different color and then compose these four into a single 2x2 display of graphs (you may do this in the software, or you may copy the images out of the software and compose them in Word). This will make a nice small-multiples plot.



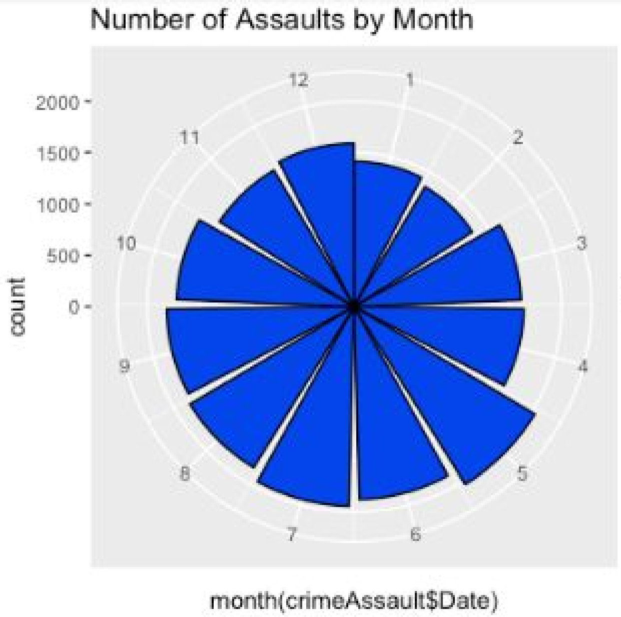


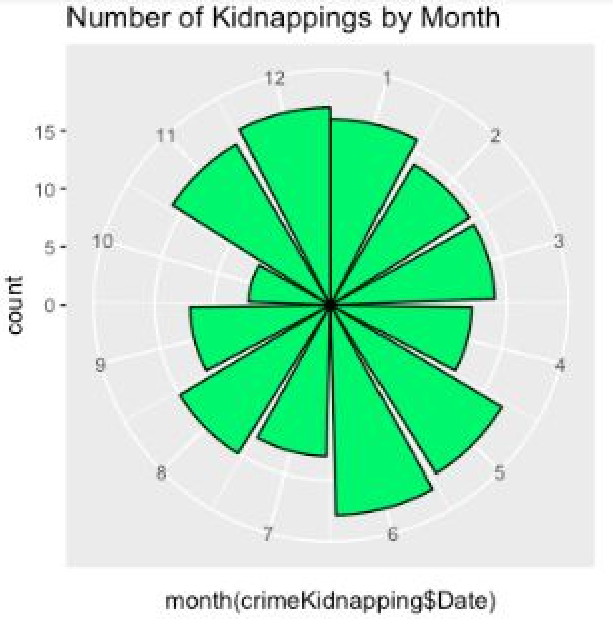
~~~
```{r p3a}
library(lubridate)
library(ggplot2)
crimedata=read.csv("/Users/alexteboul/Desktop/A3_DSC465_datasets/ChicagoCrime2018.csv")
head(crimedata)

#Convert to posixct
class(crimedata$Date)
crimedata$Date = strptime(crimedata$Date,format="%m/%d/%Y %H:%M:%S")
class(crimedata$Date)
crimedata$Date=as.POSIXct(crimedata$Date)
class(crimedata$Date) #works

#subsets ASSAULT, KIDNAPPING, NARCOTICS and THEFT
crimeAssault = crimedata[crimedata$Primary.Type=="ASSAULT", ]
crimeKidnapping = crimedata[crimedata$Primary.Type=="KIDNAPPING", ]
crimeNarcotics = crimedata[crimedata$Primary.Type=="NARCOTICS", ]
crimeTheft = crimedata[crimedata$Primary.Type=="THEFT", ]
```
~~~

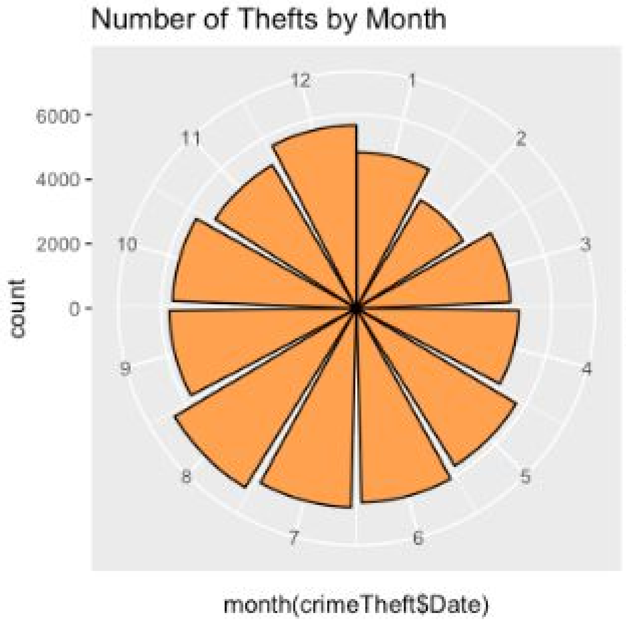


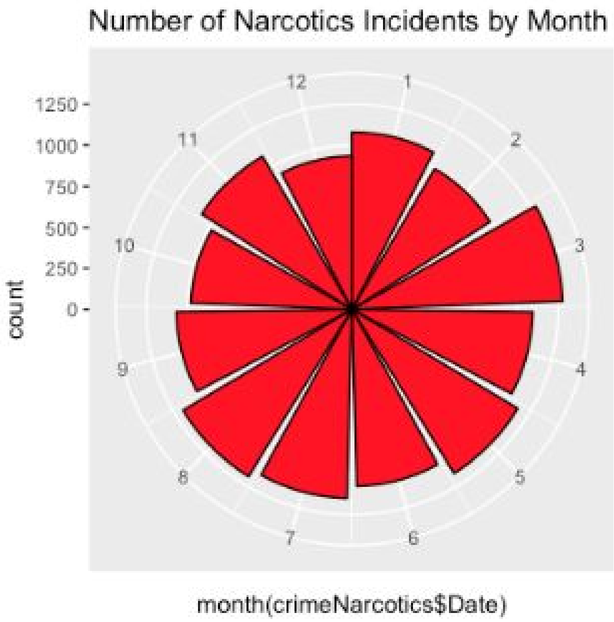
~~~
```{r p3a-1}
#ASSAULT
ggplot(crimeAssault, aes(x=month(crimeAssault$Date))) + geom_bar(bins=12,colour="black",fill="blue")
+ labs(title="Number of Assaults by Month", x="Month", y="Count") +
scale_x_continuous(breaks=seq(0,12,1))

#KIDNAPPING
ggplot(crimeKidnapping, aes(x=month(crimeKidnapping$Date))) +
geom_bar(bins=12,colour="black",fill="green") + labs(title="Number of Kidnappings by Month",
x="Month", y="Count") + scale_x_continuous(breaks=seq(0,12,1))

#NARCOTICS
ggplot(crimeNarcotics, aes(x=month(crimeNarcotics$Date))) +
geom_bar(bins=12,colour="black",fill="red") + labs(title="Number of Narcotics Incidents by Month",
x="Month", y="Count") + scale_x_continuous(breaks=seq(0,12,1))

#THEFT
ggplot(crimeTheft, aes(x=month(crimeTheft$Date))) + geom_bar(bins=12,colour="black",fill="orange") +
labs(title="Number of Thefts by Month", x="Month", y="Count") +
scale_x_continuous(breaks=seq(0,12,1))
```
~~~

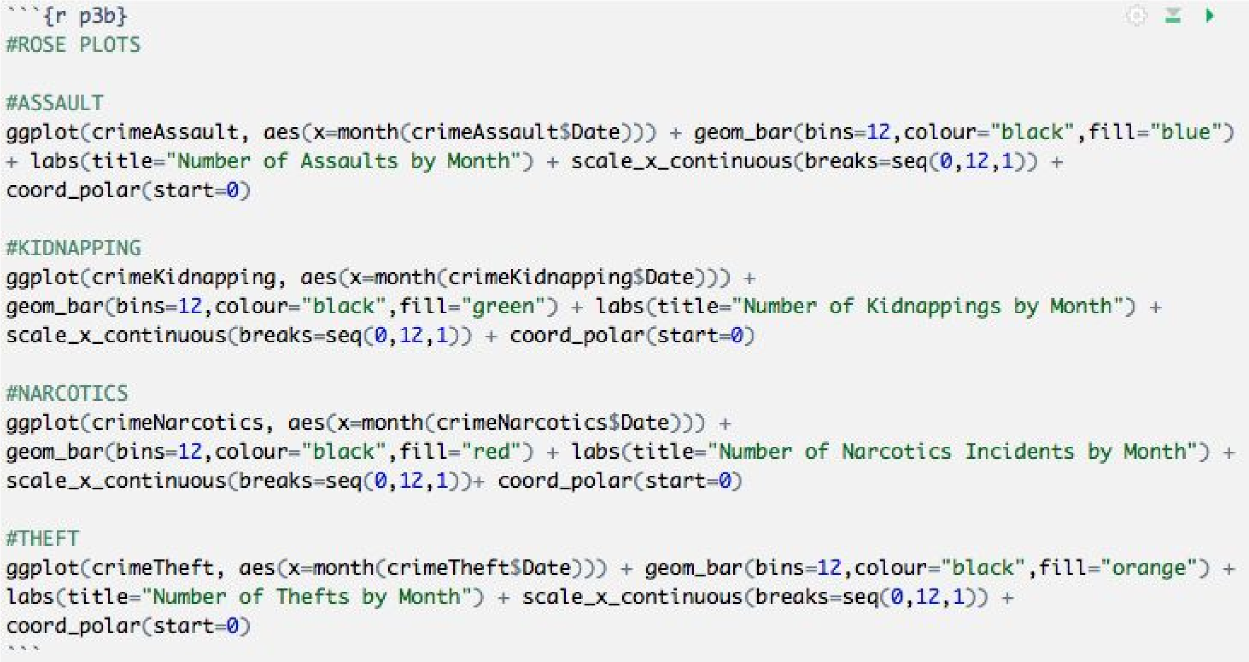
b. You will need to do this in R, as Tableau does not have a Rose-Plot feature. For the 4 monthly graphs in a), create a Rose-Plot of each graph. You can do this in R by creating the bar graph and then adding "coord_polar()" to it. You will also need to set the xlimits to keep the bars at the top from touching. Finally, play with the y-limits to either create a full rose plot or a "ring". See which communicates better. Again, color each plot by category and compose them in Word into a 2x2 display (ggplot's "facet_grid" and "facet_wrap" features don't play nice with polar plots.


Number of Assaults by Month
count
2000
1500
1000
500
0
12
1
2
3
4
5
6
7
8
9
10
11
month(crimeAssault$Date)


Number of Kidnappings by Month
count
15
10
5
0
12
1
2
3
4
5
6
7
8
9
10
11
month(crimeKidnapping$Date)


Number of Thefts by Month
count
6000
4000
2000
0
12
1
2
3
4
5
6
7
8
9
10
11
month(crimeTheft$Date)


Number of Narcotics Incidents by Month
count
1250
1000
750
500
250
0
12
1
2
3
4
5
6
7
8
9
10
11
month(crimeNarcotics$Date)

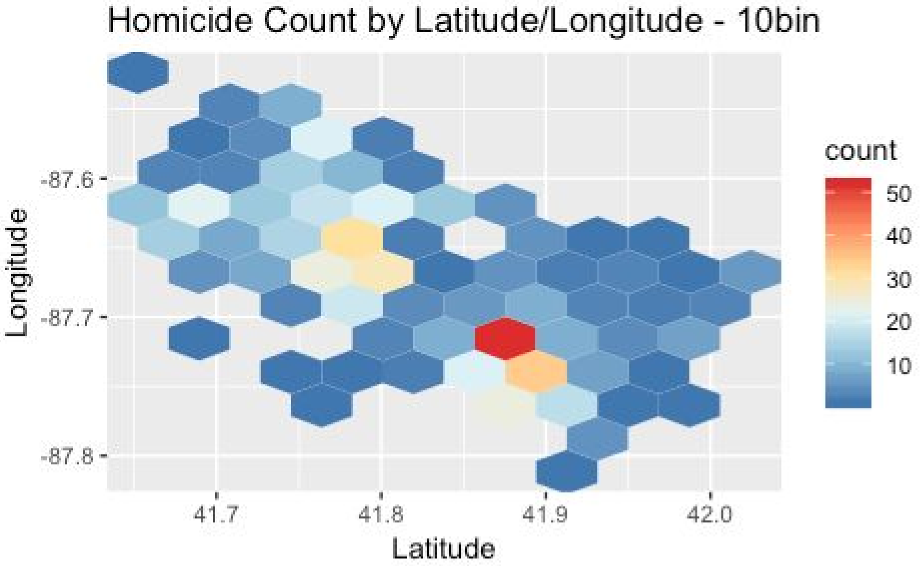


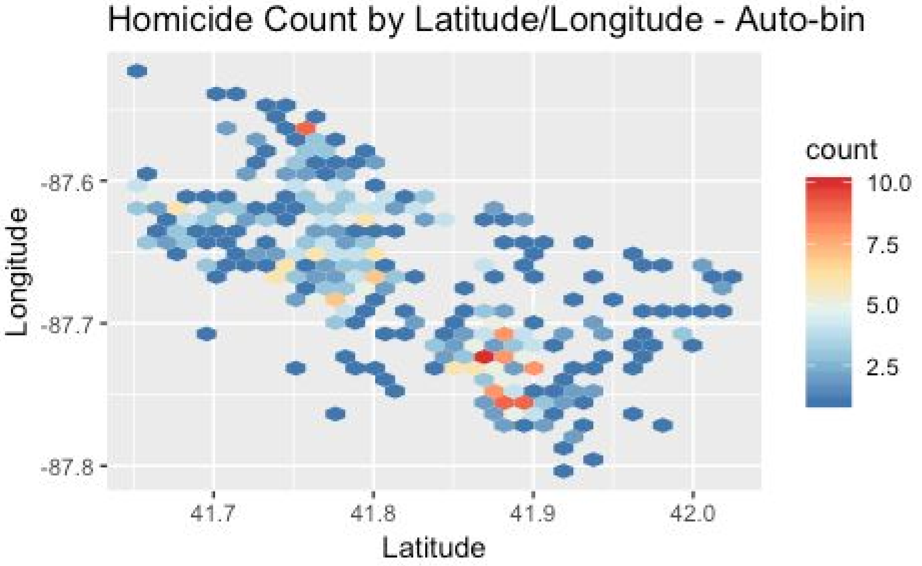
~~~
```{r p3b}
#ROSE PLOTS

#ASSAULT
ggplot(crimeAssault, aes(x=month(crimeAssault$Date))) + geom_bar(bins=12,colour="black",fill="blue")
+ labs(title="Number of Assaults by Month") + scale_x_continuous(breaks=seq(0,12,1)) +
coord_polar(start=0)

#KIDNAPPING
ggplot(crimeKidnapping, aes(x=month(crimeKidnapping$Date))) +
geom_bar(bins=12,colour="black",fill="green") + labs(title="Number of Kidnappings by Month") +
scale_x_continuous(breaks=seq(0,12,1)) + coord_polar(start=0)

#NARCOTICS
ggplot(crimeNarcotics, aes(x=month(crimeNarcotics$Date))) +
geom_bar(bins=12,colour="black",fill="red") + labs(title="Number of Narcotics Incidents by Month") +
scale_x_continuous(breaks=seq(0,12,1))+ coord_polar(start=0)

#THEFT
ggplot(crimeTheft, aes(x=month(crimeTheft$Date))) + geom_bar(bins=12,colour="black",fill="orange") +
labs(title="Number of Thefts by Month") + scale_x_continuous(breaks=seq(0,12,1)) +
coord_polar(start=0)
```
~~~

c. Analyze the differences between the bar graphs in b and the rose plots in c) analyze them for how well they communicate the patterns and differences between the categories. What are the strengths and weaknesses of each?

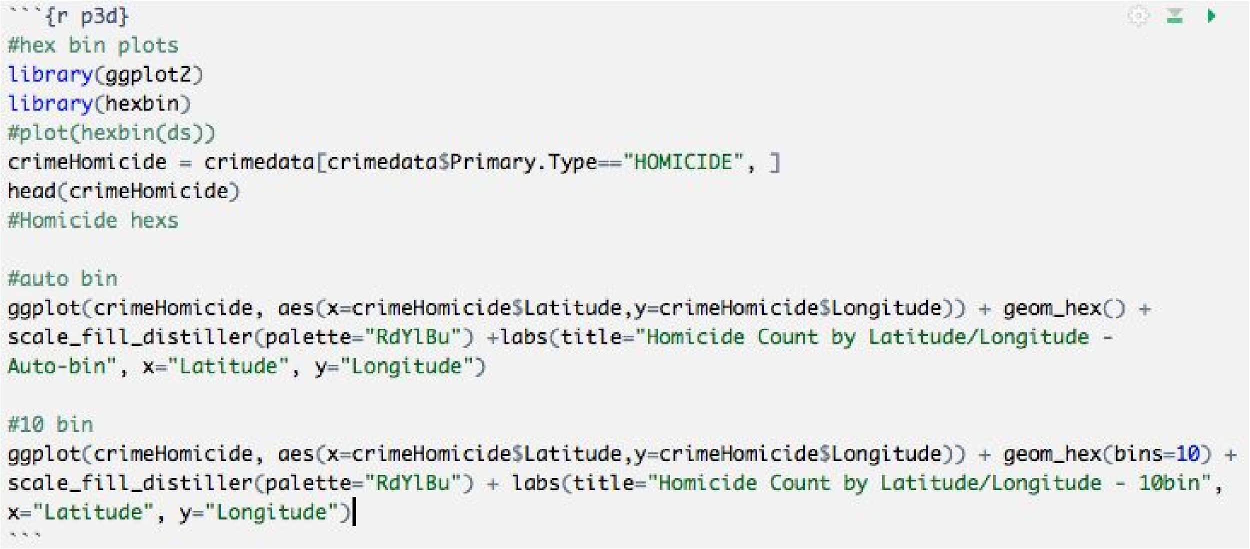
- **Bar Graphs**
  - **Strengths**: The bar graphs do a good job of displaying the counts by month in a quickly readable format. If you want to know October's count of say kidnappings, it is easy to read out 7 from the bar graph. Bar graphs are familiar to most audiences, and they convey counts in a very accessible manner.
  - **Weaknesses**: The main drawback in this case is that the periodic nature of the calendar year is harder for the viewer to see. One this point, the rose plot manages to provide a better visualization for seeing a general count trend as it related to the period of the year.
- **Rose Plots**
  - **Strengths**: The rose plots manage to capture the cyclical/periodic nature of the calendar year while still delivering the counts of each crime incident per month.
  - **Weaknesses**: The counts can be difficult to ascertain and area of the wedges is not related to the actual counts.
- **General Notes**
  - Thefts are the most common crime, followed by assaults, then narcotics and finally kidnappings as observed in the bar graphs. The rose plots help inform us of patterns like the fact that kidnappings are at their lowest numbers between July and October.

d. Create a hexbin plot of the HOMICIDE data by latitude and longitude. Choose a color scheme that highlights the area with high homicide rates.





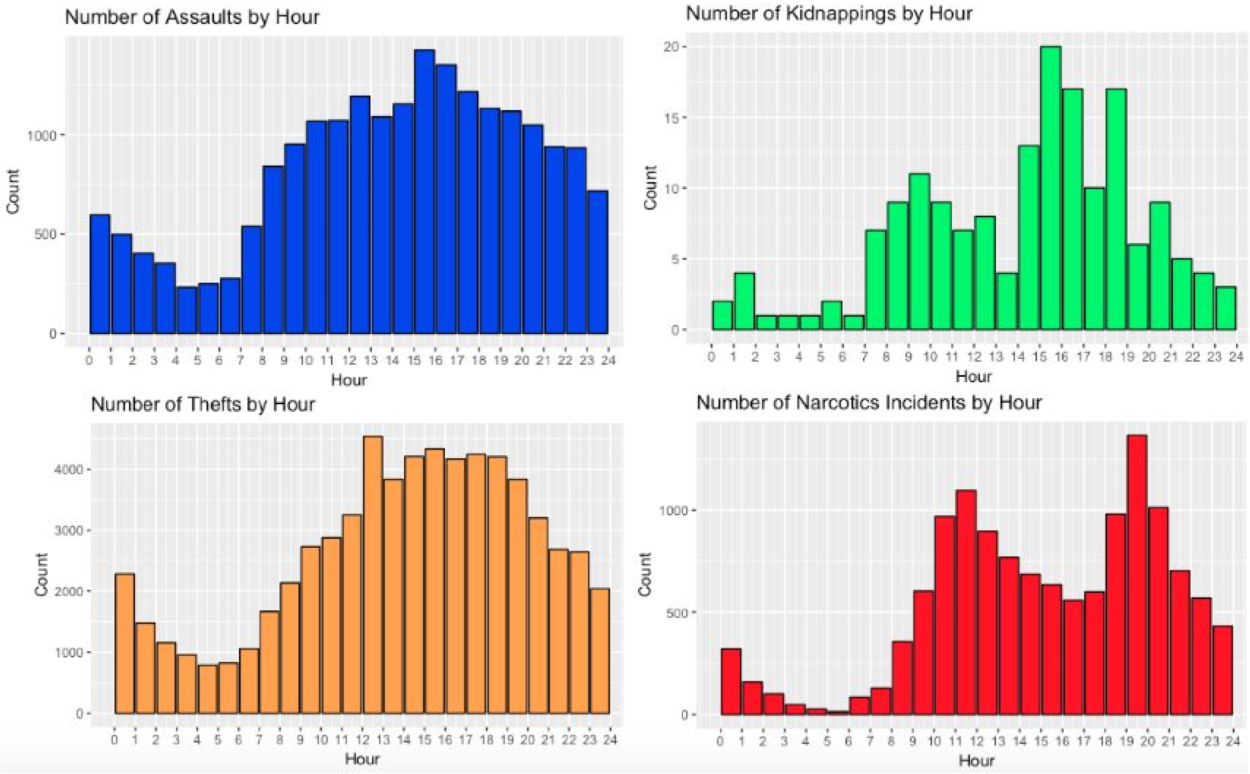
- The above two hexbin plots show the areas of high Homicide counts by Latitude and Longitude. The first plot has only 10 bins so it gives a more general representation of where more homicides took place. We see a pocket of high counts around longitude=-87.75 and latitude=41.875. The second plot is a more refined view with more bins to see less generalized counts. The color palette used was RdYlBu to highlight the high counts in Red.



~~~
```{r p3d}
#hex bin plots
library(ggplot2)
library(hexbin)
#plot(hexbin(ds))
crimeHomicide = crimedata[crimedata$Primary.Type=="HOMICIDE", ]
head(crimeHomicide)
#Homicide hexs

#auto bin
ggplot(crimeHomicide, aes(x=crimeHomicide$Latitude,y=crimeHomicide$Longitude)) + geom_hex() +
scale_fill_distiller(palette="RdYlBu") +labs(title="Homicide Count by Latitude/Longitude -
Auto-bin", x="Latitude", y="Longitude")

#10 bin
ggplot(crimeHomicide, aes(x=crimeHomicide$Latitude,y=crimeHomicide$Longitude)) + geom_hex(bins=10) +
scale_fill_distiller(palette="RdYlBu") + labs(title="Homicide Count by Latitude/Longitude - 10bin",
x="Latitude", y="Longitude")
```
~~~

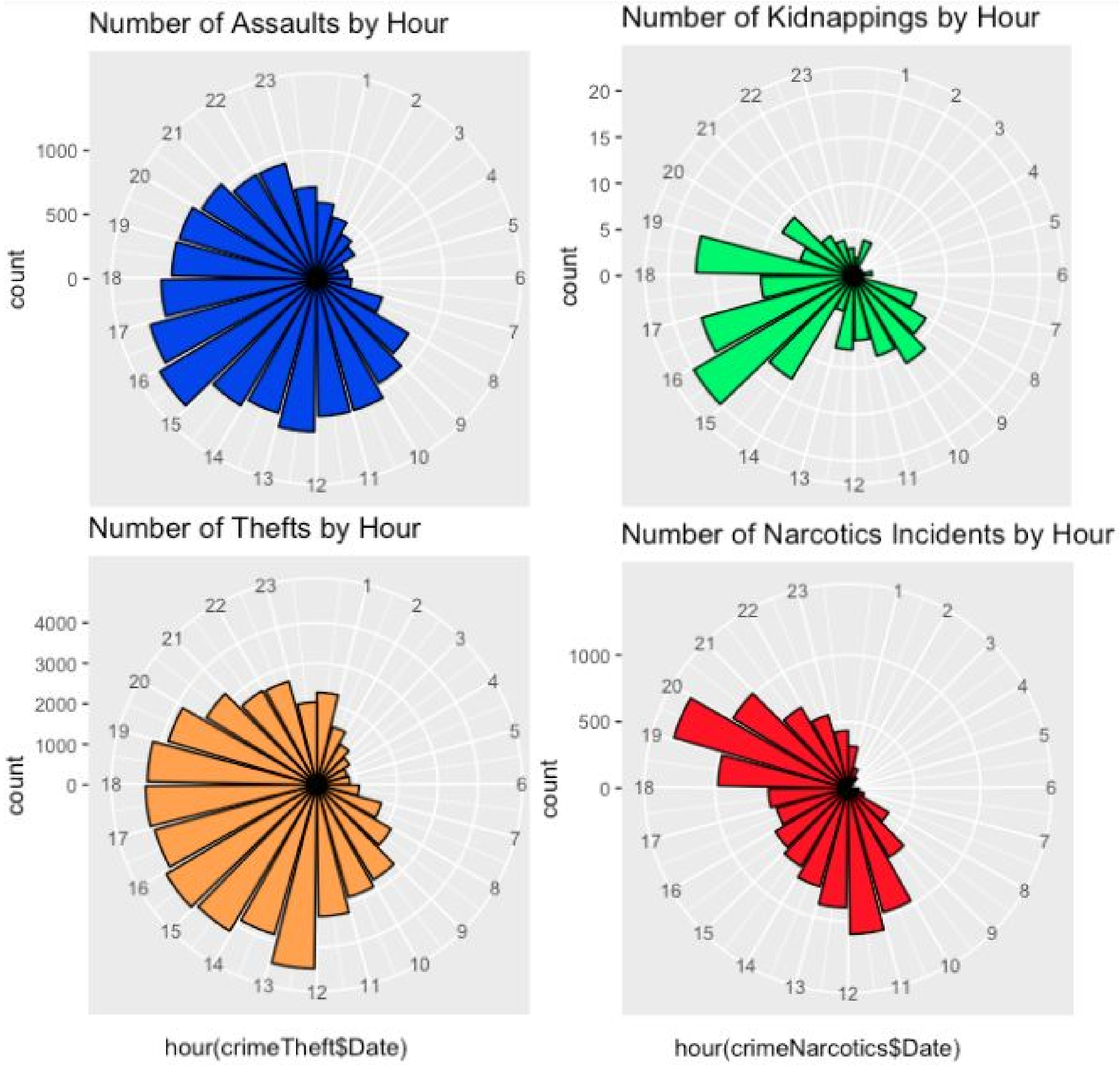
e. (Extra Credit, 5 pts) Do the same as in a) and b) for the hour of the day, rather than the month of the year. Here you may use the "hour(DateTime)", the only thing you will need to do is to find out how to get it to give you the 24 hour version instead of the 12 hour as the data was originally in AM/PM with 0-12 in hours.

Bar Graphs by Hour



- This gives an interesting view of the data that demonstrates lower crime rates in the early hours of the day (midnight to 8am). Seems that criminals must sleep as well - or at least crimes fewer crimes are reported during these hours.

Rose Plots by Hour



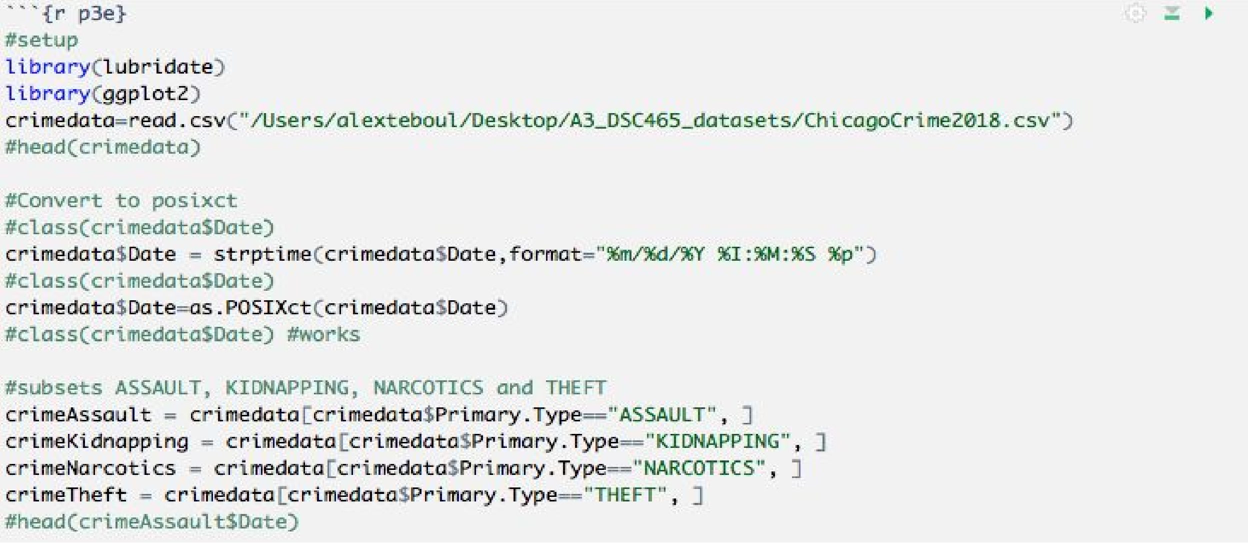
Number of Assaults by Hour

Number of Kidnappings by Hour

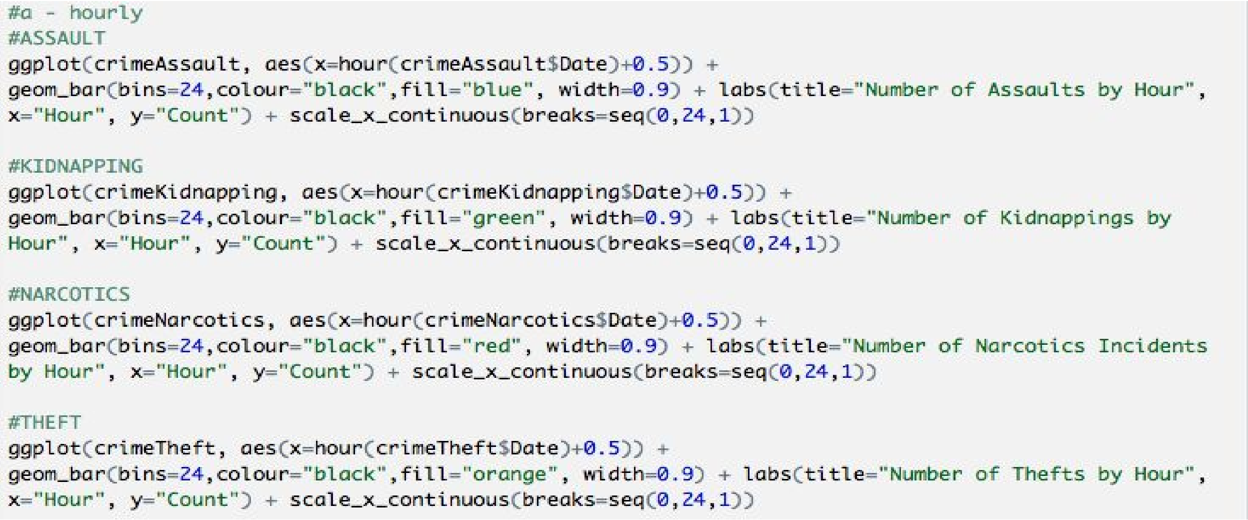
Number of Thefts by Hour

Number of Narcotics Incidents by Hour

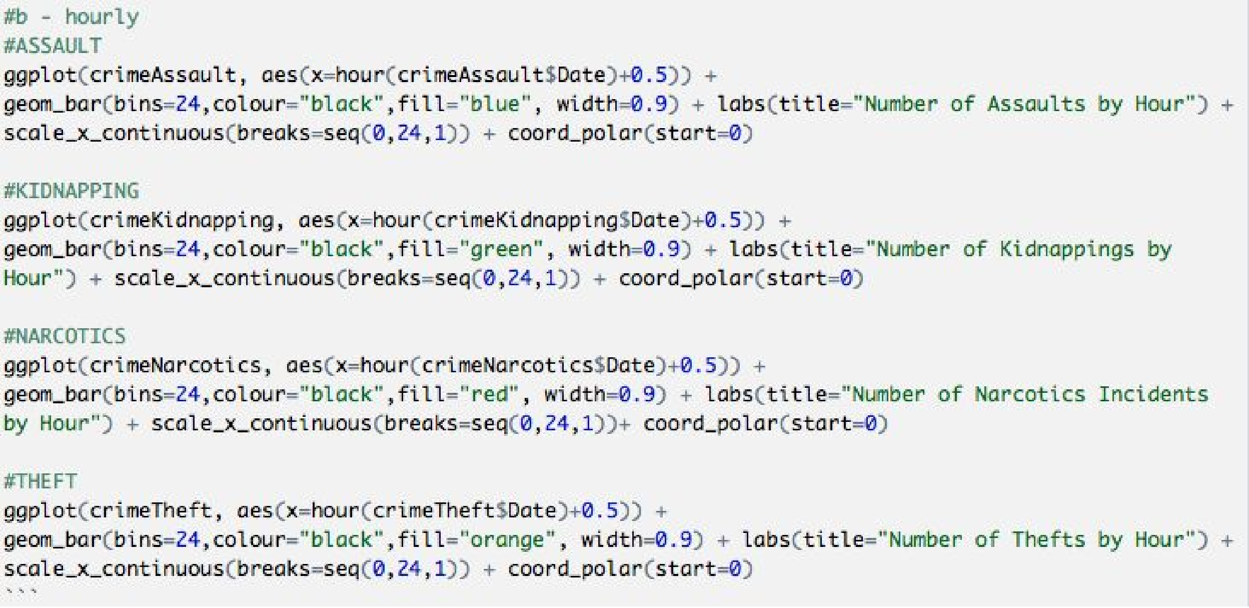
- The rose plot shows an interesting pattern for each of the crime types. The hourly counts show that for Theft, Assault, and Kidnappings - most take place during daylight hours. For Narcotics, it seems that peak hours are around 11:30 AM and 7:30 PM.



~~~
```{r p3e}
#setup
library(lubridate)
library(ggplot2)
crimedata=read.csv("/Users/alexteboul/Desktop/A3_DSC465_datasets/ChicagoCrime2018.csv")
#head(crimedata)

#Convert to posixct
#class(crimedata$Date)
crimedata$Date = strptime(crimedata$Date,format="%m/%d/%Y %I:%M:%S %p")
#class(crimedata$Date)
crimedata$Date=as.POSIXct(crimedata$Date)
#class(crimedata$Date) #works

#subsets ASSAULT, KIDNAPPING, NARCOTICS and THEFT
crimeAssault = crimedata[crimedata$Primary.Type=="ASSAULT", ]
crimeKidnapping = crimedata[crimedata$Primary.Type=="KIDNAPPING", ]
crimeNarcotics = crimedata[crimedata$Primary.Type=="NARCOTICS", ]
crimeTheft = crimedata[crimedata$Primary.Type=="THEFT", ]
#head(crimeAssault$Date)
~~~



```
#a - hourly
#ASSAULT
ggplot(crimeAssault, aes(x=hour(crimeAssault$Date)+0.5)) +
geom_bar(bins=24,colour="black",fill="blue", width=0.9) + labs(title="Number of Assaults by Hour",
x="Hour", y="Count") + scale_x_continuous(breaks=seq(0,24,1))

#KIDNAPPING
ggplot(crimeKidnapping, aes(x=hour(crimeKidnapping$Date)+0.5)) +
geom_bar(bins=24,colour="black",fill="green", width=0.9) + labs(title="Number of Kidnappings by
Hour", x="Hour", y="Count") + scale_x_continuous(breaks=seq(0,24,1))

#NARCOTICS
ggplot(crimeNarcotics, aes(x=hour(crimeNarcotics$Date)+0.5)) +
geom_bar(bins=24,colour="black",fill="red", width=0.9) + labs(title="Number of Narcotics Incidents
by Hour", x="Hour", y="Count") + scale_x_continuous(breaks=seq(0,24,1))

#THEFT
ggplot(crimeTheft, aes(x=hour(crimeTheft$Date)+0.5)) +
geom_bar(bins=24,colour="black",fill="orange", width=0.9) + labs(title="Number of Thefts by Hour",
x="Hour", y="Count") + scale_x_continuous(breaks=seq(0,24,1))
```

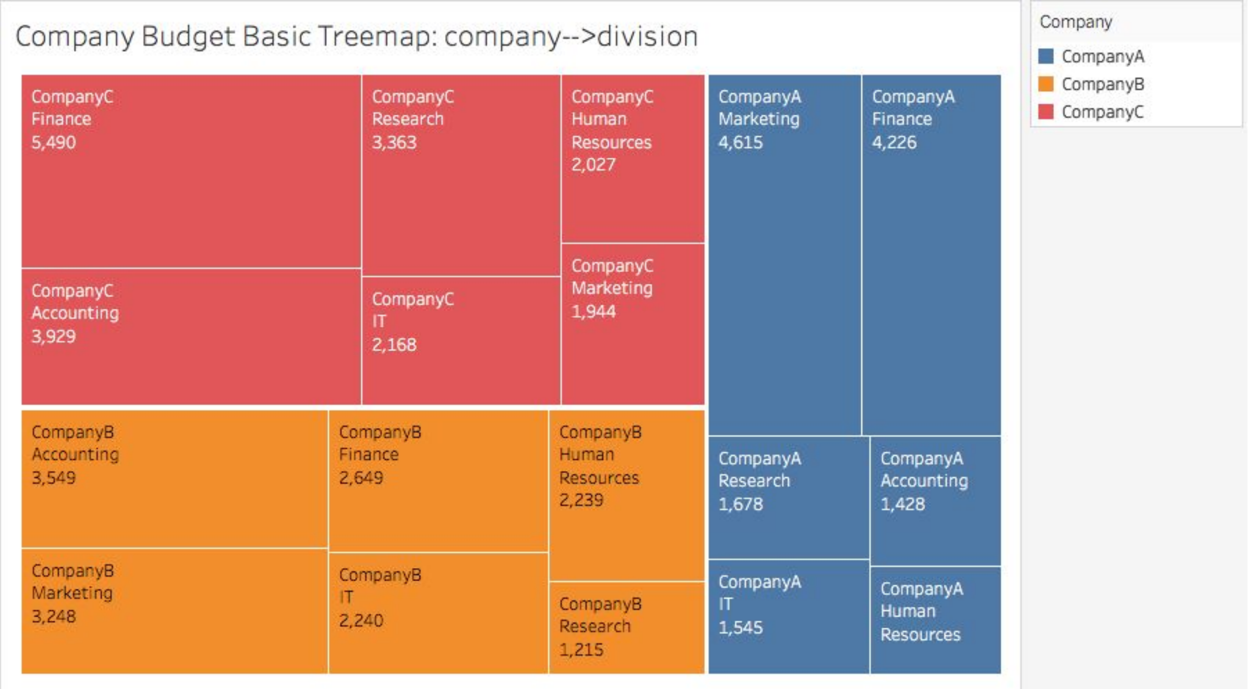


```
#b - hourly
#ASSAULT
ggplot(crimeAssault, aes(x=hour(crimeAssault$Date)+0.5)) +
geom_bar(bins=24,colour="black",fill="blue", width=0.9) + labs(title="Number of Assaults by Hour") +
scale_x_continuous(breaks=seq(0,24,1)) + coord_polar(start=0)

#KIDNAPPING
ggplot(crimeKidnapping, aes(x=hour(crimeKidnapping$Date)+0.5)) +
geom_bar(bins=24,colour="black",fill="green", width=0.9) + labs(title="Number of Kidnappings by
Hour") + scale_x_continuous(breaks=seq(0,24,1)) + coord_polar(start=0)

#NARCOTICS
ggplot(crimeNarcotics, aes(x=hour(crimeNarcotics$Date)+0.5)) +
geom_bar(bins=24,colour="black",fill="red", width=0.9) + labs(title="Number of Narcotics Incidents
by Hour") + scale_x_continuous(breaks=seq(0,24,1))+ coord_polar(start=0)

#THEFT
ggplot(crimeTheft, aes(x=hour(crimeTheft$Date)+0.5)) +
geom_bar(bins=24,colour="black",fill="orange", width=0.9) + labs(title="Number of Thefts by Hour") +
scale_x_continuous(breaks=seq(0,24,1)) + coord_polar(start=0)
```
```
```

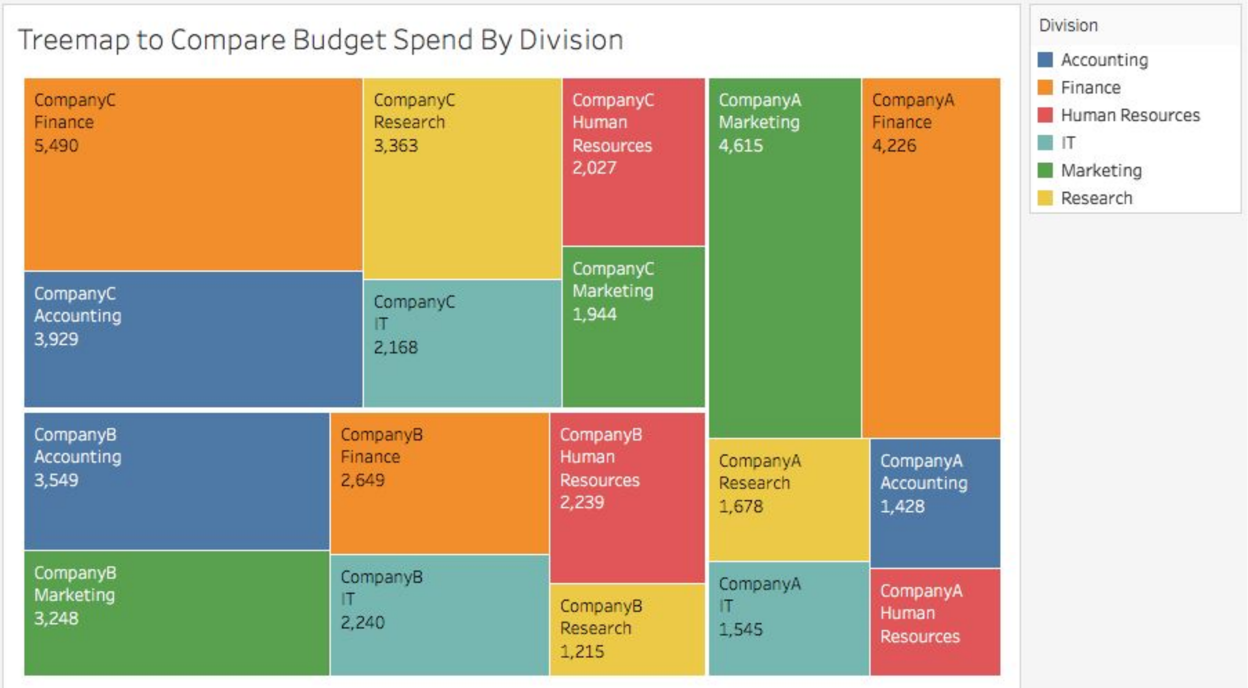
Problem 4 - Tableau

4) Download the Company.csv dataset which contains a list of budgets for offices in a company. Each office is in specific division of a company. So the fields are “company”, “division”, “office” and “budget”. The data has a hierarchy defined by these fields of company → division → office, but you may need to set up the hierarchy in whatever software you are using (Tableau it can be helpful to set up a hierarchical set of dimensions. Right click on a dimension to find the “Hierarchy” submenu). With this dataset create the following visualizations. You may use either Tableau or R to complete this problem.

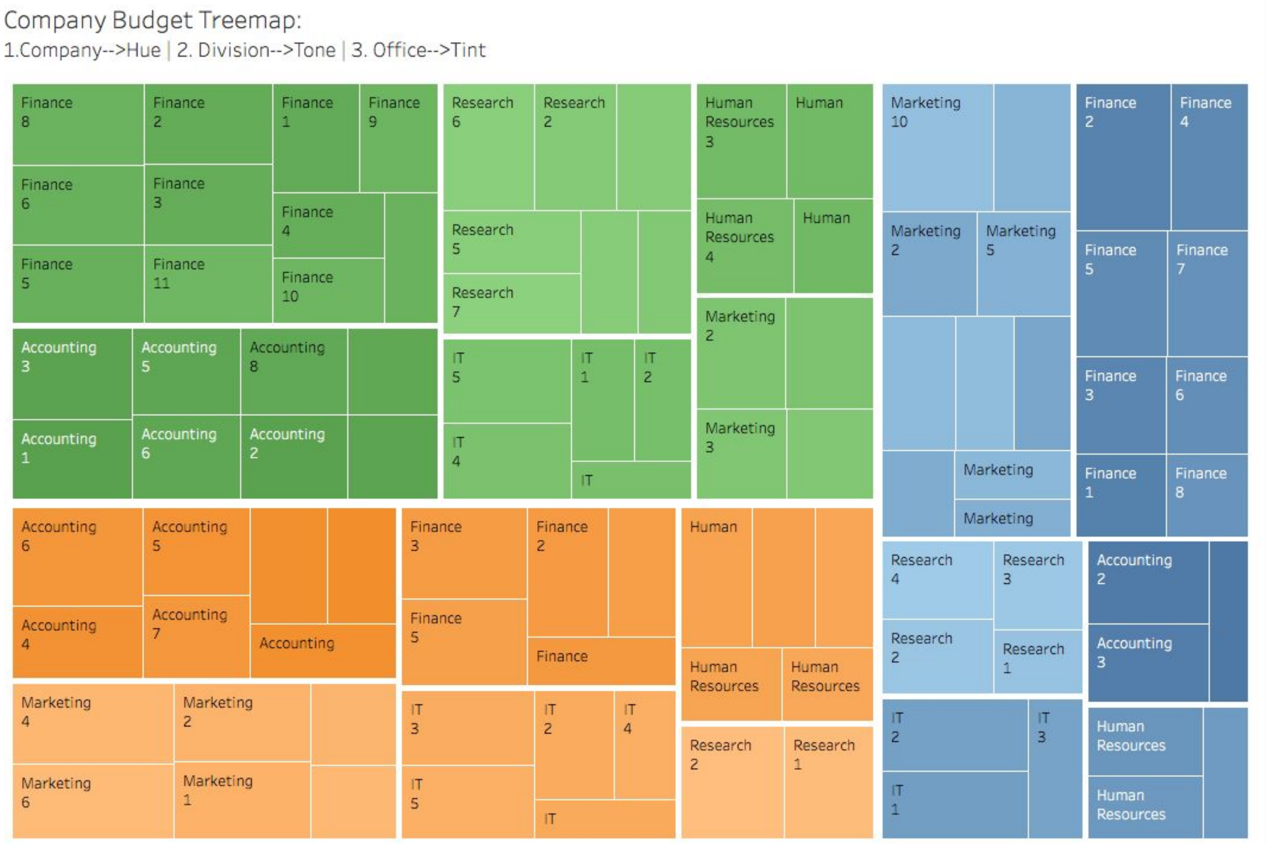
a. Create a basic treemap of this data with two subdivisions “company” and “division”. The division cells should be sized by the total budget for the division (aggregate by sum). Color by the company.



- Above, the basic treemap is shown for the Company dataset. There are 3 Companies, each with divisions for Accounting, Marketing, Finance, IT, HR, and Research. It is interesting to see visually how these different companies are allocating their budgets.
- In this problem we were asked to just color by company but I also think it’s interesting to see the coloring by division. This way we can easily compare say the budget of Company A’s vs Company B’s Finance Division.
- ***This is shown below but not part of the answer to this question.



b. Create a treemap of this data with three subdivisions "company", "division" and "office" with the cells sized by the budget for each office. Experiment with color schemes so that the result communicates as well as possible the three hierarchical subdivisions of the company and also the budget.



- Above shows the 3 subdivision treemap. With the respect to the color scheme, there are three levels. The hue corresponds with the Company (green, orange, and blue). The tone in terms of darkening by tint and shade matches the Division, and the third layer of color is a tint added to lighten based on Office. Layer 2 could also be described as saturation. It's not a great way to tell each division apart, but at least the structure is clear and budget sizes each division.

c. Write a paragraph that compares this visualization to the previous. How well does it communicate compared to a) and is it feasible with this dataset to communicate all three levels with a treemap? Explain your choice of color map based on the criteria we chose in class (categorical, sequential, divergent, etc.). The material from lecture 7 will be helpful.

- Compared with a, the treemap in b does a better job of communicating the 3 different levels of the hierarchy. Specifically, we see the different offices split up in the treemap in b, that we didn't in a. In terms of color, the three different companies get categorically colored with red, orange, and blue. Within each color (hue) a sequential color scheme is applied afterwards. This helps distinguish between the different divisions and offices. Unfortunately with the multiple layers, it becomes difficult to create a color scheme that very effectively matches the hierarchical nature of this dataset.