Cynthia Cho
DSC 465

Homework 3

** All codes are attached to the bottom of the problem when applicable. Full code is attached to the end of the homework as an appendix.

Problem 1: Paper review turned in separately on May 8th.

Problem 2: Portland Water Level for 2003

a.) Level plot/Heat map for water level. For the heat map, both RStudio and Tableau was used (code for RStudio is at the bottom of the problem). The heat map shows the progression of the tides over the days of the year. The hours are on the x-axis and the y-axis shows the days/months in the year 2003. To keep both heat maps comparable and consistent, the full range of hours were left on the graphs.

To effectively communicate the trends of the high tides versus the low tides, the spectral color scheme was used in the RStudio version, and a diverging complementary color scheme was used for the Tableau version. The color schemes were used to meaningfully provide context to the graph and data, as well as maximize space. As it can be seen from the both heat maps, the colors bring to our attention to the pattern that emerges: the tide level in relations to the time of day and the months of the year.

During the months October to April, water levels are low during the early hours of the day till roughly 6am; and water levels are increasingly getting higher starring from the mid-morning till about late in the evening before hitting midnight. And during the months April to October, water levels are low during the afternoon well into the late evening, but generally highest during the early mornings till mid-morning. When looking at the months from mid-March till end of May, there are water level fluctuations taking place through the day, as it can be seen from the heat map. Overall, there appears to be seasonality to highs and lows of tidal flow depending on the time of day and primarily between the winter and summer months, with transitions occurring between spring and fall months. By using the different color schemes that create a diverging/contrasting effect, it allows for patterns to be identified through maximization of space.

RStudio heat map code:
Using Spectral Color"

```
ggplot(data=water, aes(x=new_time, y=day, fill=WL)) +
scale_fill_distiller(palette = "Spectral")  +
  labs(title="Portland Water Level HeatMap for 2003", x="Hour", y
= "Month", fill="Water Level") +geom_tile()
```

RStudio heat map:

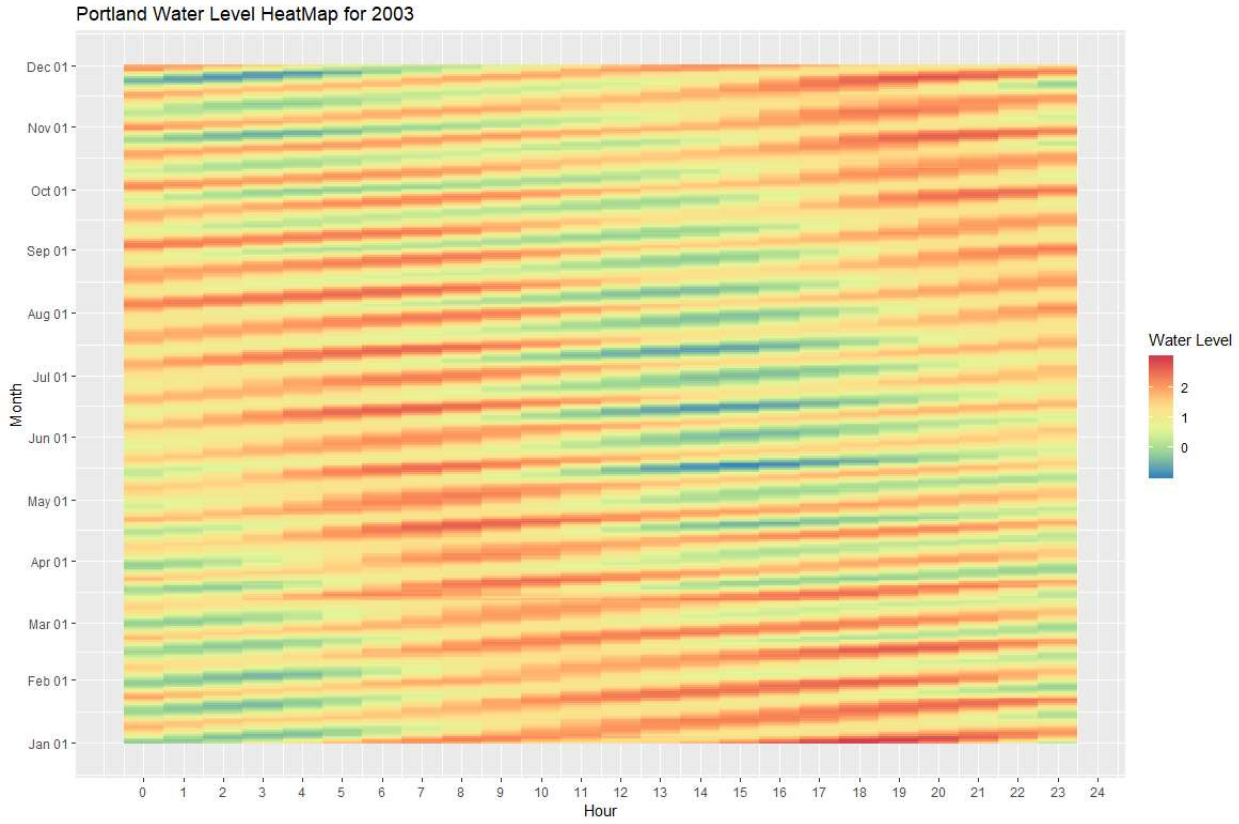
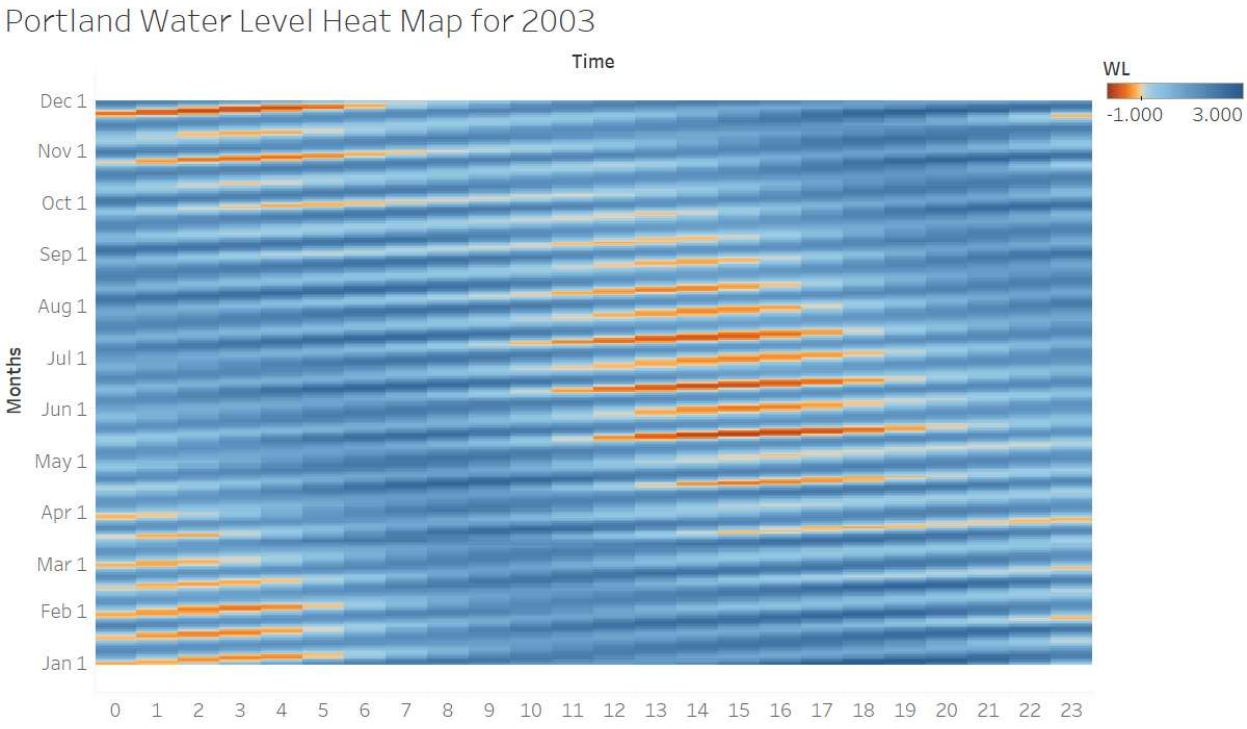
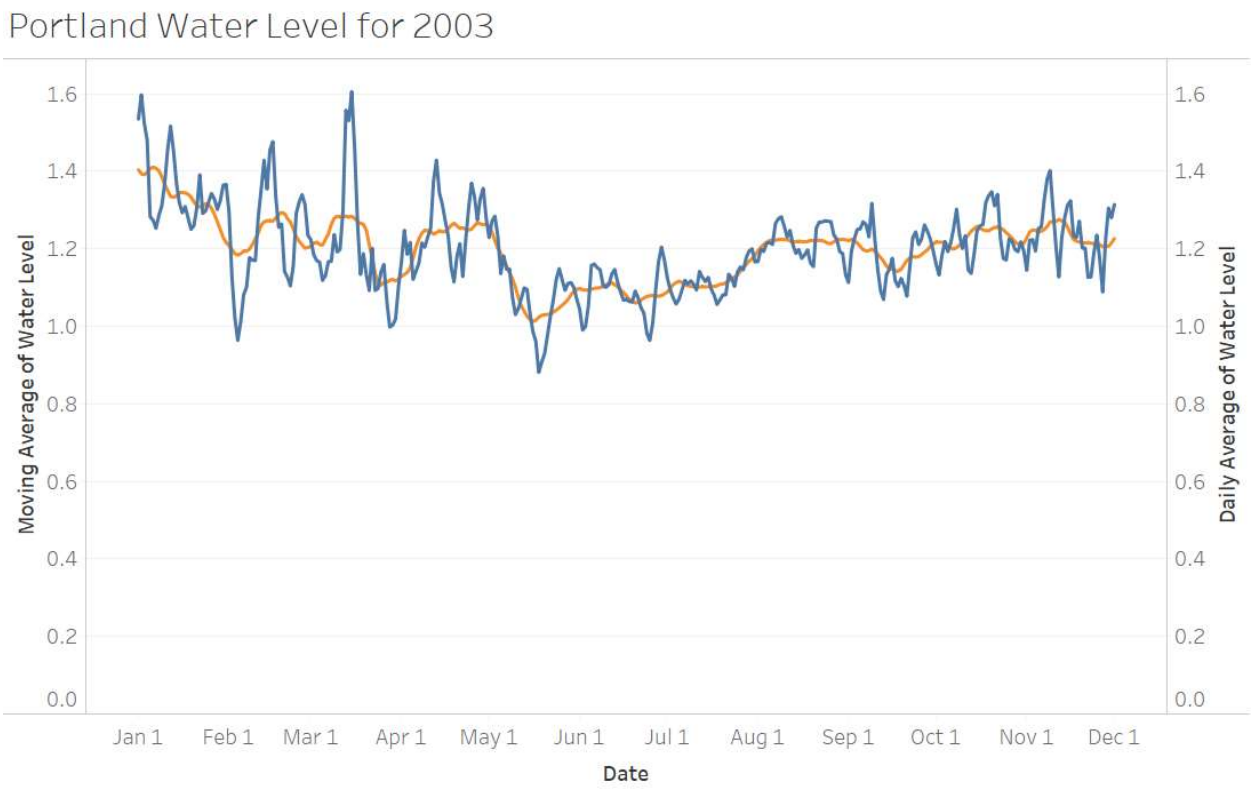


Tableau heat map:



b.) Line graph of Portland Water level: average Water Level with Moving Average Water Level overlay. To smooth out the weekly variations and show the overall trend, a centered moving average of the previous 7 day and next 7 days was calculated for Moving Average water level line graph. Additionally, the axes were synchronized.



c.) The heat map and the line graph with a moving average overlay provides complimentary information about the data. The heat map provides an immediate visual summary of the data using color while maximizing on the space. From the water level heat map, we can immediately see that there is a seasonal pattern to the water level on when the tides are highest, and when they are the lowest. It illustrates the progression of the tides over the days of the year and how they move with time; the heat map conveys granular information in the form of a "big picture" context through space and color. What we also notice on the heat map is that certain months have water level fluctuations throughout the day, which is reinforced by the line graph.

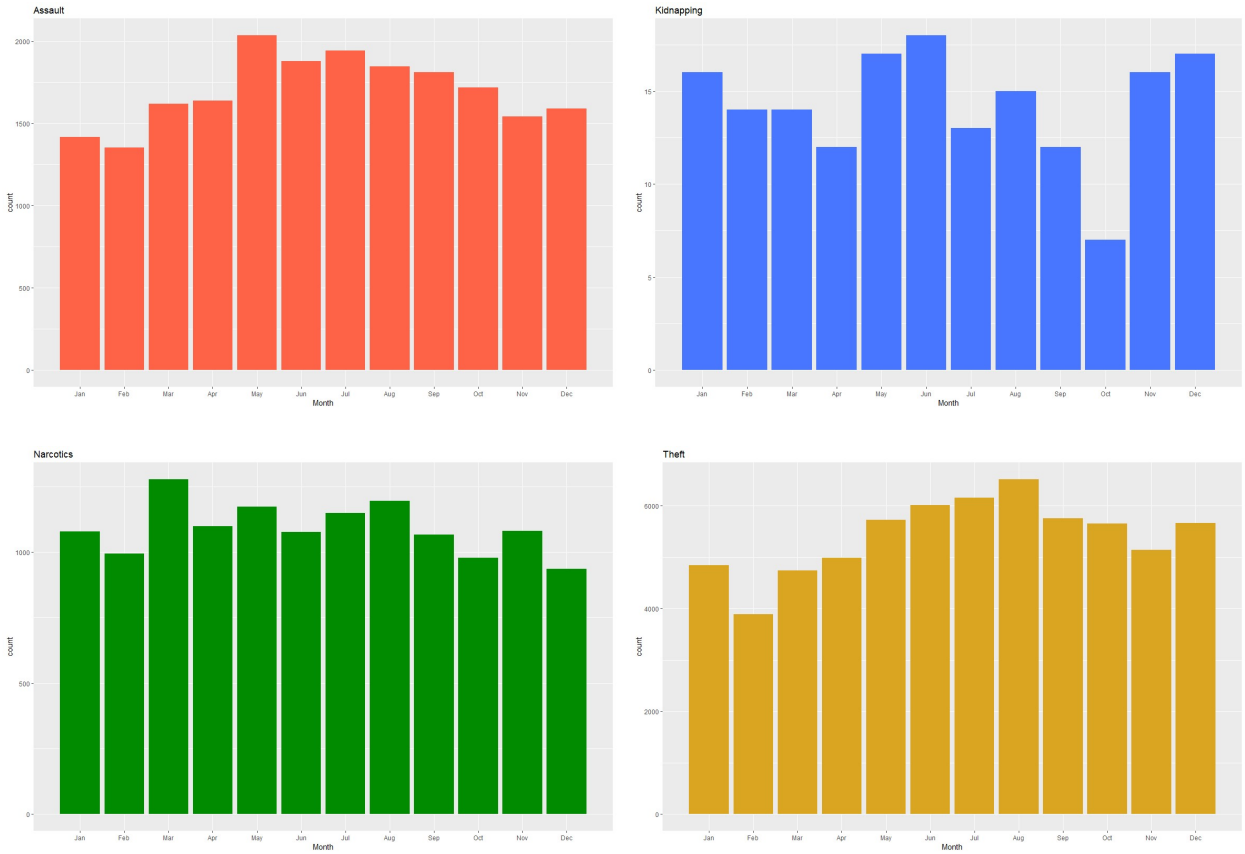
As seen on the line graph, there is a fair amount of fluctuations that take place between the months of January and June. By adding a moving average line, it smooths out the daily fluctuations in water level and identifies a trend; the trend being that the overall water levels

begin to decrease in the beginning of May before increasing in August. While on the heat map, we see that there are seasonal patterns, we are unable to point out the monthly trends as easily that can be identified with line graphs using numbers. The advantage of using the line graph for problem 2b, is the ability to identify overall trends with specific numbers.

Both graphs have their strengths and provide a great deal of detail. By providing both graphs, they are commentary in that they provide additional content together. Both graphics use time, in the heat map, time is broken up into two pieces and used spatially, and in the time-series line graph, we are looking time temporally. Provides two different ways of looking at the data to gain better insight.

Problem 3: Chicago Crime 2018

a.) "small multiples" bar chart for the crimes: Assault, Kidnapping, Narcotics, and Theft.
To create a 2x2 display of the small multiples plot, they had to be composed in Word; due to the difference in the y-axis scale, specifically for Kidnapping, which would have caused some graphs to not show up. The code for the graphics is at the bottom of the problem.



RStudio bar chart code:

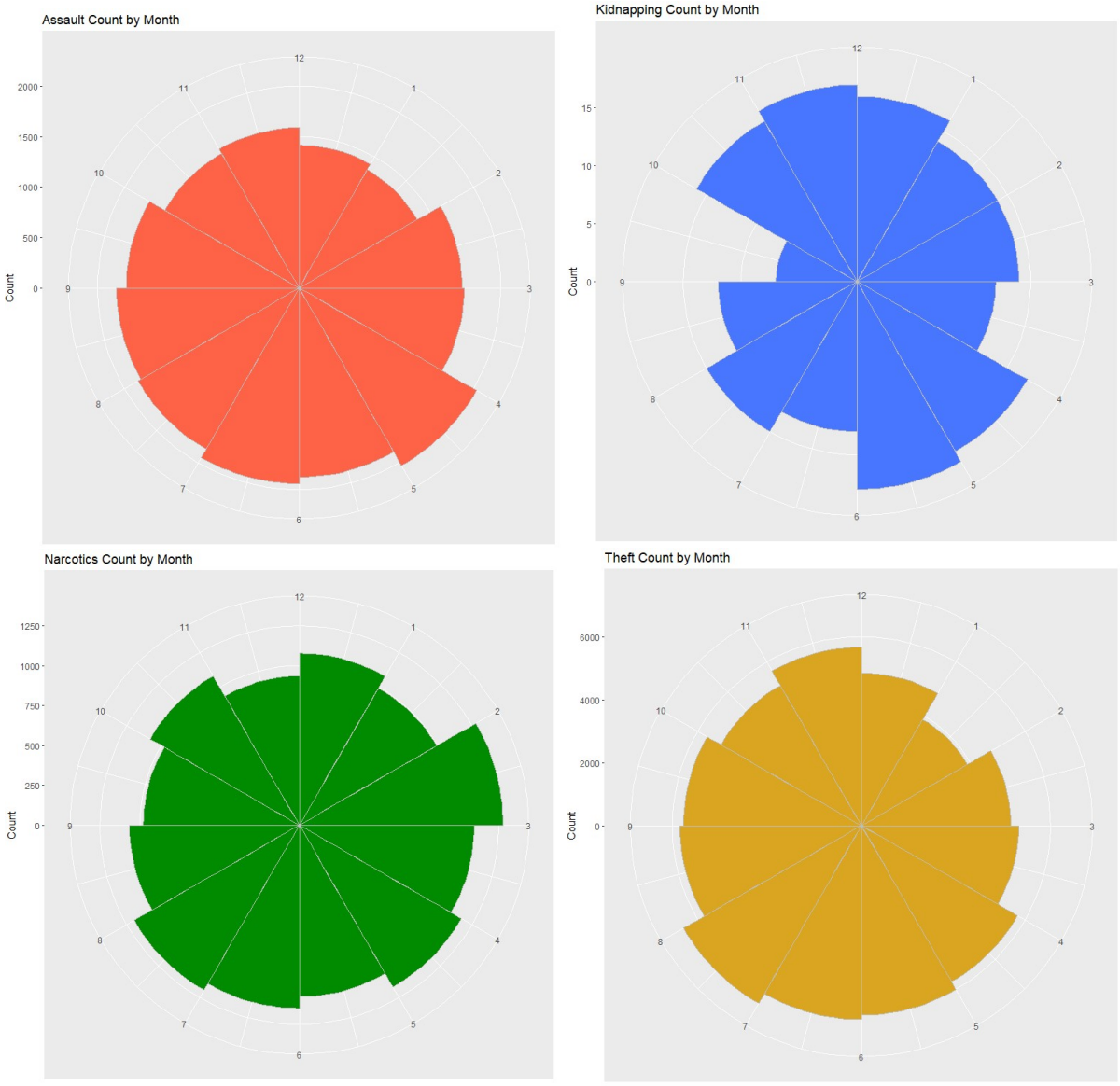
Assault plot:
```
ggplot(Assault, aes(x=Month)) + geom_bar(fill='tomato') +
labs(title="Assault") + scale_x_discrete(limits = month.abb)
```

Kidnapping plot:
```
ggplot(Kidnapping, aes(x=Month)) + geom_bar(fill='royalblue1') +
labs(title="Kidnapping") +scale_x_discrete(limits = month.abb)
```

Narcotics plot:
```
ggplot(Narcotics, aes(x=Month)) + geom_bar(fill='green4') +
labs(title="Narcotics") + scale_x_discrete(limits = month.abb)
```

Theft plot:
```
ggplot(Theft, aes(x=Month)) + geom_bar(fill='goldenrod') +
labs(title="Theft")+ scale_x_discrete(limits = month.abb)
```

b.) “small multiples” rose plot for the crimes: Assault, Kidnapping, Narcotics, and Theft. Similar to the bar chart, to create a 2x2 display of the small multiples plot, they had to be composed in Word; due to the difference in the y-axis scale, specifically for Kidnapping, which would have caused some graphs to not show up. The code for the graphics is at the bottom of the problem.



RStudio rose plot code:

```
## Assault Rose Plot
ggplot(Assault, aes(x = Month)) + geom_histogram(breaks = seq(0,
12),fill = "tomato1", color="grey") + coord_polar(start = 0) +
theme_minimal() + scale_fill_brewer() + ylab("Count") +
ggtitle("Assault Count by Month") + scale_x_continuous("", limits =
c(0, 12), breaks = seq(0, 12), labels = seq(0, 12))

## Kidnapping Rose Plot
ggplot(Kidnapping, aes(x = Month)) +
 geom_histogram(breaks = seq(0, 12),fill = "royalblue1",
color="grey") + coord_polar(start = 0) + theme_minimal() +
scale_fill_brewer() + ylab("Count") + ggtitle("Kidnapping Count by
Month") + scale_x_continuous("", limits = c(0, 12), breaks = seq(0,
12), labels = seq(0, 12))

## Narcotics Rose Plot
ggplot(Narcotics, aes(x = Month)) + geom_histogram(breaks = seq(0,
12),fill = "green4", color="grey") + coord_polar(start = 0) +
theme_minimal() + scale_fill_brewer() + ylab("Count") +
ggtitle("Narcotics Count by Month") + scale_x_continuous("", limits =
c(0,12), breaks = seq(0, 12), labels = seq(0, 12))

## Theft Rose Plot
ggplot(Theft, aes(x = Month)) + geom_histogram(breaks = seq(0, 12),fill
= "goldenrod", color="grey") + coord_polar(start = 0) + theme_minimal()
+ scale_fill_brewer() + ylab("Count") + ggtitle("Theft Count by Month")
+ scale_x_continuous("", limits = c(0, 12), breaks = seq(0, 12), labels
= seq(0, 12))
```

c.) The difference that is most noticeable is the format of the chart and visual summary impact that both small multiples plots provide. Both charts provide a monthly count for the year for each crime on their respective graph. However, what is lacking on the bar chart is continuity. Because months are temporal, there is a continuity aspect that would benefit from using a rose plot. The rose plot provides continuity at the end of a period and shows how the data evolves over time, which enhances the visual summary for the crimes that are committed across the year and can detect a pattern if there is any.

In terms of crime pattern, because of the temporal aspect, the rose plot provides a slightly better visual of the pattern that takes shape over the months. To read the count of instances for each month respective to the crime, the bar chart is much easier. Observations are:

Assault: The lowest assault count took place in February, and the highest count took place in May.
Kidnapping: The lowest kidnapping count took place in October, and the highest count took place in June.

Narcotics: The lowest narcotics count took place in December, and the highest count took place in March.
Theft: The lowest theft count took place in February, and the highest count took place in August.

From the rose plot, it appears most Assault and Theft related crimes take place between the months of April till October. There is a somewhat cyclical nature to narcotics, but it is difficult to tell if this was a one-off for the year 2018, or a repeat occurrence for drug use. And lastly for kidnapping, what the rose plot shows that most of kidnapping crimes take place over January to May and October to December. This is just a conjecture, but based on the plot, it appears that kidnapping is highest during the months when school is in session, and lowest during the summer. An interesting observation is the difference in count for each crime: kidnapping has the lowest (in the tens), and the highest occurrence of crime from the four observed categories is Theft, with a total in the thousands. Assault and Narcotics category sum to the thousands, but not nearly as large as Theft.

Strengths of Rose Plot:

1. provides continuity when used for temporal data
2. can easily make patterns stand out

Weakness of Rose Plot:

1. Similar to a pie chart, the angles and area can cause distortion
2. Comparison between non-adjacent segments are challenging
3. When stacked segments are used (like a stacked bar chart), the angles distort perception and makes area harder to judge and comparing segments challenging
4. Exact value assessment is harder with a rose plot. (Not as easy to read like the bar chart where you just go straight across)

Strengths of Bar Chart:

1. The width for each bar is the same and measuring
2. Makes the comparison process much easier across non-adjacent bars
3. Not ideal when there are many categories in the data
4. Easy to understand

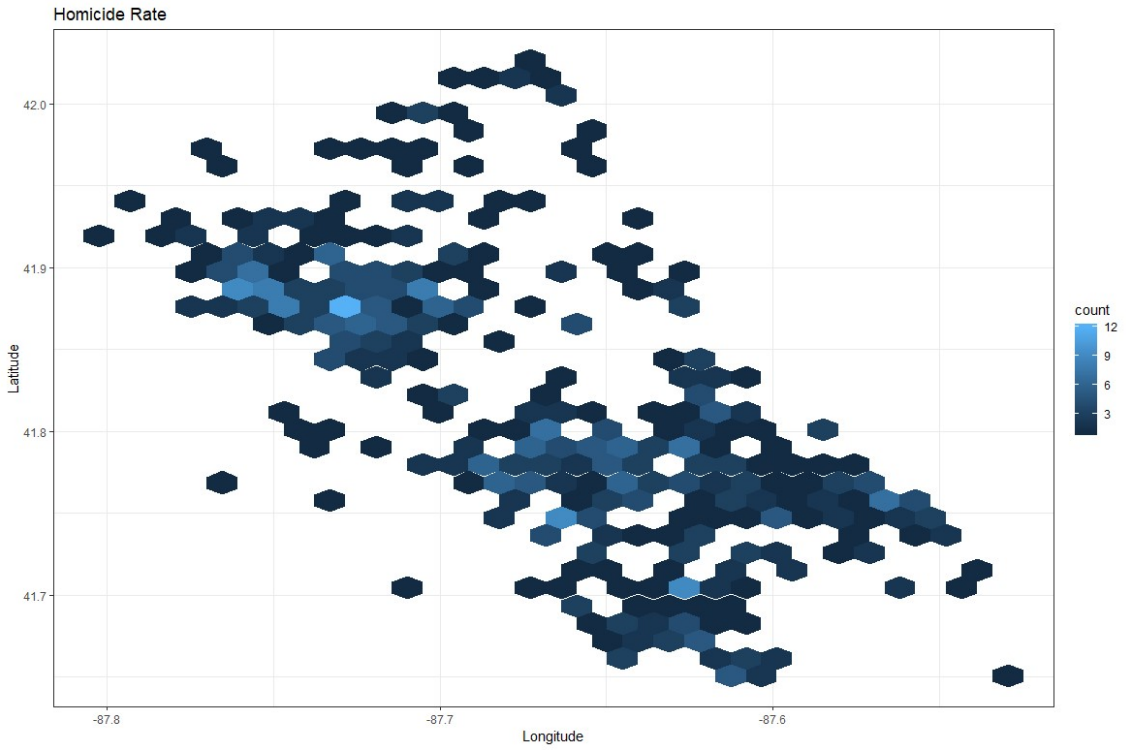
Weakness of Bar Chart:

1. Lack of continuity with temporal data
2. Limited space when the number of categories is large

Both the bar chart and rose plot have their strengths and weaknesses in providing visual summary, however, they both fulfill different visualization needs that can be used to provide context to the data.

d.) Hexbin plot for Chicago Crime 2018 subset for Homicide

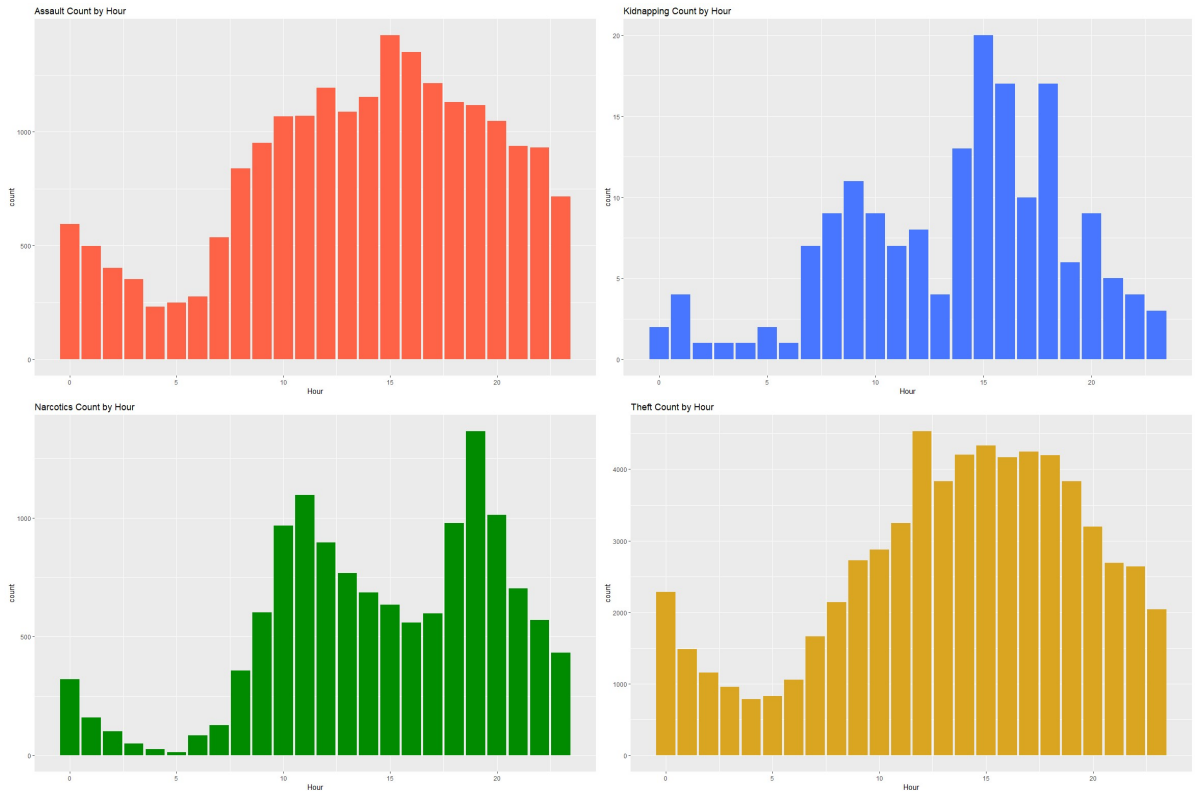
The hexbin plot shows the density of homicide in what should be the Chicago area. However, without an overlay of a map of the city, it's difficult to make out what part of the city where the highest density is located – in terms of the specific subdivisions/towns. What is noticeable is that there are potentially two areas where the density of homicide is high. The first location is longitude: -87.77 to -87.7, latitude: 41.85 to 41.9. The second location is longitude: -87.79 to -87.625, latitude: 41.75 to 41.8.

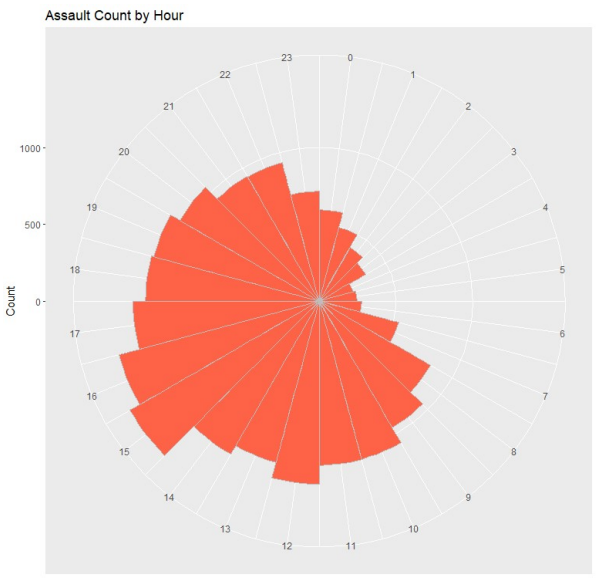


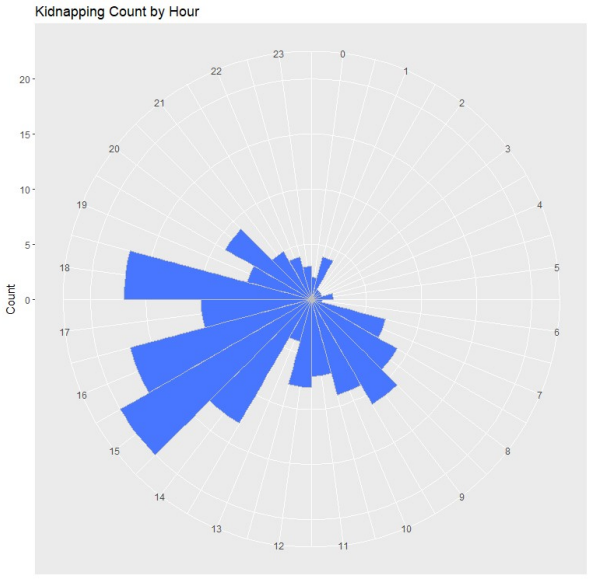
e.) Extra Credit: Small multiples plot for bar chart and rose plot by hour

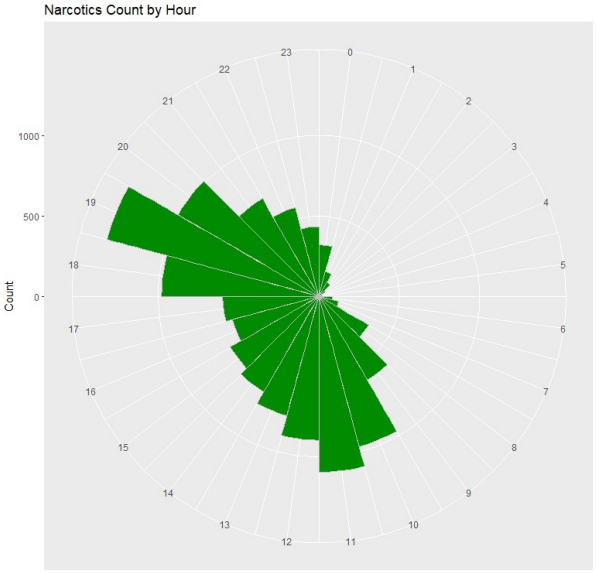
The hourly small multiples plot provides more granular detail and offers the specific time of day the crimes under observation take place. For both Assault and Theft appear to take place during the early afternoon well into the evening; Kidnapping appears to occur around the time students leave school for the day; and lastly, Narcotic arrests occur mostly around the mid-morning and in the evening.

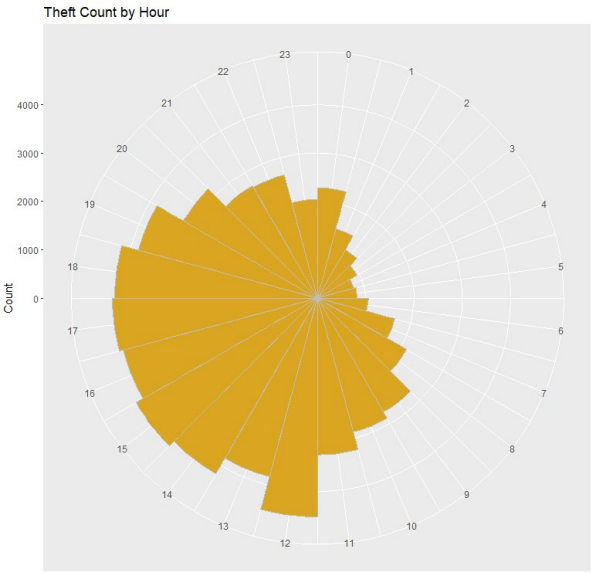
Looking at both monthly and hourly data for the year of 2018 provides insight into not only the cyclical patter of the crime, but also the specific times when crimes are more likely to take place during the day. By being able to change the granularity of the data, the visualization provides different context in a meaningful way.




Assault Count by Hour
Count
1000
500
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

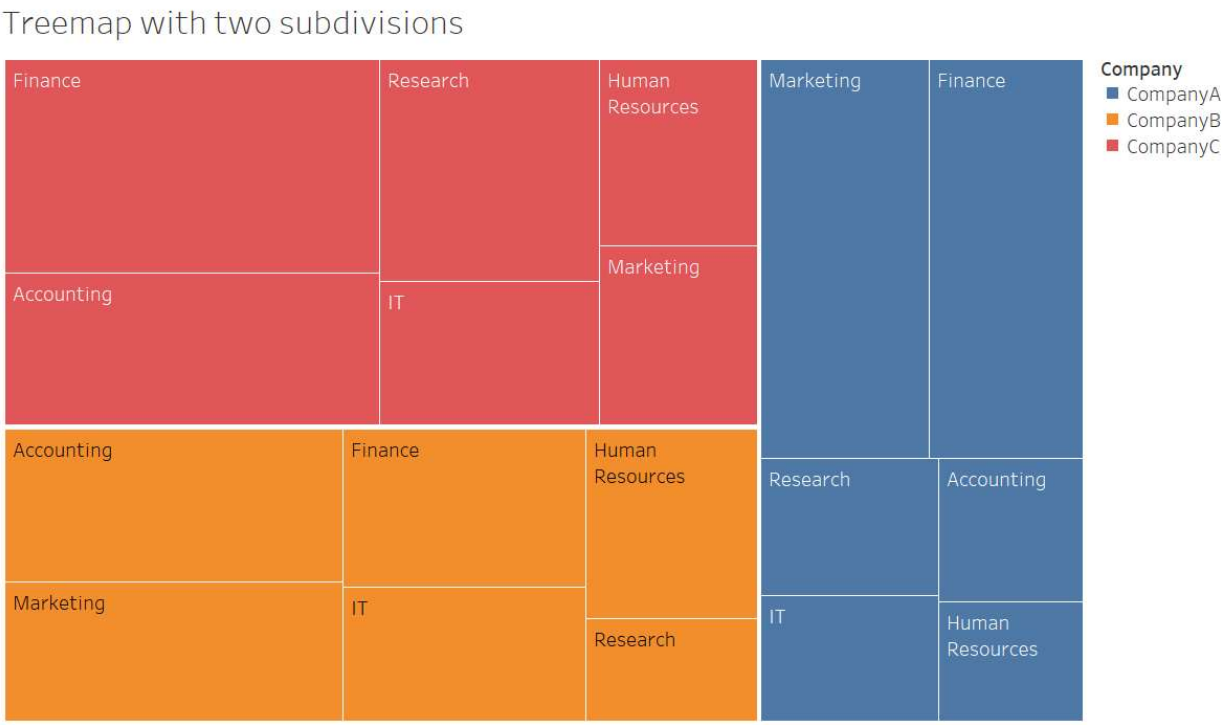

Kidnapping Count by Hour
Count
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23


Narcotics Count by Hour
Count
1000
500
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23


Theft Count by Hour
Count
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

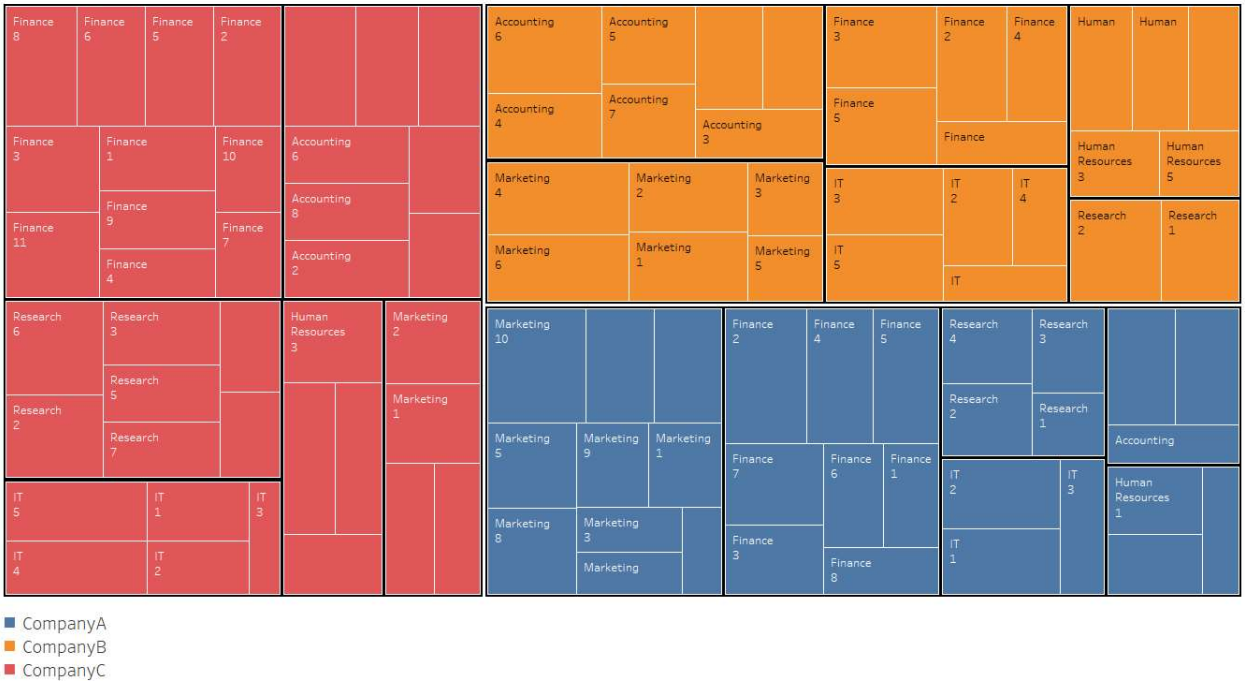
Problem 4: Company dataset

a.) Treemap with two subdivisions: company and division. The division cells are sized by the total budget for the division (aggregated by sum) and Colored by Company



b.) Treemap with three subdivisions: company, division, and office. The Company name and divisions are the labels; the cells are sized by the total budget for each office, and the colors represent the company.

Treemap with three subdivisions



c.) The treemap in 4b.) is a further subdivided than the treemap from 4a). The data hierarchy structures are moving from top to bottom, and then left to right. It seems feasible with this dataset to communicate all three levels with a treemap because it not only shown you the company, the division, and the office numbers, it also shows you the relative budget to the other offices within the same division. The drawback of this Tableau treemap is that unless the graphics are used in the interactive setting within Tableau, the offices with the smaller budgets are difficult to make out due to the granularity of the what the treemap is trying to show with the large number of offices available in the dataset. However, without the interactive interface, the various subdivisions are challenging to read. For instance, while we can see that the Company A's Human Resources office "1" has the larger budget in comparison to the other two offices, it is difficult to tell which cells belong to the other two offices from a static viewpoint.

To color choice was to use the categorical color map because it was used to represent categorical data: company. The choice behind using a categorical color map was because Budget, a continuous variable was already used to size the cells according to office budget; in order to create some distinguishability, color was used to represent the companies. The hues are distinct, and the labels are legible.

Appendix

```
library(ggplot2)
library(lubridate)
library(stringr)
library(hexbin)

setwd("C:/Users/Cindy/Documents/DSC 465/Homework/Homework 3")

#Problem 2:
water = read.csv("PortlandWaterLevel2003 - for R.csv")

head(water)
water = water[-1,]
head(water)
class(water$Time)

WaterLevel=water$WL
day <- as.Date(water$Date, "%m/%d/%Y")
water$new_time = strptime(water$Time, format="%H:%M")
water$new_time = as.POSIXlt(water$new_time)
water$new_time
class(water$new_time)
water$hour = hour(water$new_time)
head(water)


ggplot(data=water, aes(x=hour, y=day, fill=WL)) + scale_fill_distiller(palette =
"PRGn") +
  labs(title="Portland Water Level HeatMap for 2003", x="Hour", y = "Month",
fill="Water Level") +
  geom_tile()

## Using Spectral Color"
ggplot(data=water, aes(x=hour, y=day, fill=WL)) + scale_fill_distiller(palette =
"Spectral")  +
  labs(title="Portland Water Level HeatMap for 2003", x="Hour", y = "Month",
fill="Water Level") +
  geom_tile() + scale_x_continuous(breaks=seq(0,24)) +
  scale_y_date(date_breaks = "months" , date_labels = "%b %d")


## Problem 3
chicagocrime = read.csv("ChicagoCrime2018.csv")
head(chicagocrime)
class(chicagocrime$Date)
head(chicagocrime$Date)

chicagocrime$Date = strptime(chicagocrime$Date, format="%m/%d/%Y %I:%M:%S %p")
chicagocrime$Date = as.POSIXlt(chicagocrime$Date)
chicagocrime$Date
class(chicagocrime$Date)
chicagocrime$Month = month(chicagocrime$Date)
chicagocrime$Hour = hour(chicagocrime$Date)
#chicagocrime$Date

library(stringr)
#p = str_split_fixed((chicagocrime$Date), " ", 2)
#head(p)
#new_date = p[,1]
#new_date
```

```
#Time = p[,2]
#Time
#chicagocrime$new_date = as.Date(chicagocrime$Date)
#new_date
#class(new_date)
# chicagocrime$Month <- months.Date(new_date)
#chicagocrime$Hour = hour(chicagocrime$Date)
#class(chicagocrime$Hour)
#head(chicagocrime)
#chicagocrime$Hour
#subcrime = subset(chicagocrime, subset = Primary.Type %in% c("ASSAULT","",
"NARCOTICS", "THEFT"))
#subcrime


## Subsetting the data
Assault = subset(chicagocrime, Primary.Type == "ASSAULT")
Kidnapping = subset(chicagocrime, Primary.Type == "KIDNAPPING")
Narcotics = subset(chicagocrime, Primary.Type == "NARCOTICS")
Theft = subset(chicagocrime, Primary.Type == "THEFT")
HOMICIDE = subset(chicagocrime, Primary.Type == "HOMICIDE")

#_____________________________________________________________________________#

## Problem 3a.) ##

##Bar Graphs by Month

ggplot(Assault, aes(x=Month)) + geom_bar(fill='tomato') + labs(title="Assault Count by
Month") + scale_x_discrete(limits = month.abb)
ggplot(Kidnapping, aes(x=Month)) + geom_bar(fill='royalblue1') +
labs(title="Kidnapping Count by Month") +scale_x_discrete(limits = month.abb)
ggplot(Narcotics, aes(x=Month)) + geom_bar(fill='green4') + labs(title="Narcotics
Count by Month") + scale_x_discrete(limits = month.abb)
ggplot(Theft, aes(x=Month)) + geom_bar(fill='goldenrod') + labs(title="Theft Count by
Month")+ scale_x_discrete(limits = month.abb)

#_____________________________________________________________________________#


## Problem 3b.) ##

# Rose Plot by Month

## Assault Rose Plot
ggplot(Assault, aes(x = Month)) +
  geom_histogram(breaks = seq(0, 12),fill = "tomato1", color="grey") +
  coord_polar(start = 0) + ylab("Count") + ggtitle("Assault Count by Month") +
  scale_x_continuous("", limits = c(0, 12), breaks = seq(12), labels = seq(12))

## Kidnapping Rose Plot
ggplot(Kidnapping, aes(x = Month)) +
  geom_histogram(breaks = seq(0, 12),fill = "royalblue1", color="grey") +
  coord_polar(start = 0) +  ylab("Count") + ggtitle("Kidnapping Count by Month") +
  scale_x_continuous("", limits = c(0, 12), breaks = seq(12), labels = seq(12))

## Narcotics Rose Plot
ggplot(Narcotics, aes(x = Month)) +
  geom_histogram(breaks = seq(0, 12),fill = "green4", color="grey") +
  coord_polar(start = 0) +  ylab("Count") + ggtitle("Narcotics Count by Month") +
  scale_x_continuous("", limits = c(0,12), breaks = seq(12), labels = seq(12))

## Theft Rose Plot
```

```
ggplot(Theft, aes(x = Month)) +
  geom_histogram(breaks = seq(0, 12),fill = "goldenrod", color="grey") +
  coord_polar(start = 0) + ylab("Count") + ggtitle("Theft Count by Month") +
  scale_x_continuous("", limits = c(0, 12), breaks = seq(12), labels = seq(12))


#_____________________________________________________________________________#

## Problem 3d.) ##

head(chicagocrime)

ggplot(HOMICIDE, aes(x=Longitude, y=Latitude) ) + geom_hex() + theme_bw() +
  ggtitle("Homicide Rate")


#_____________________________________________________________________________#

## Problem 3e.) ##

##Bar Graphs by Hour

ggplot(Assault, aes(x=Hour)) + geom_histogram(binwidth=3, fill='tomato1') +
labs(title="Assault Count by Hour")
ggplot(Kidnapping, aes(x=Hour)) + geom_bar(fill='royalblue1') + labs(title="Kidnapping
Count by Hour") +scale_x_continuous()
ggplot(Narcotics, aes(x=Hour)) + geom_bar(fill='green4') + labs(title="Narcotics Count
by Hour")  +scale_x_continuous()
ggplot(Theft, aes(x=Hour)) + geom_bar(fill='goldenrod') + labs(title="Theft Count by
Hour") +scale_x_continuous()


# Rose Plot by Hour

## Assault Rose Plot
ggplot(Assault, aes(x = Hour)) +
  geom_bar(width=1, fill = "tomato1", color="grey") +
  coord_polar(start = 0) + ylab("Count") + ggtitle("Assault Count by Hour") +
  scale_x_continuous("", breaks = seq(0,24), labels = seq(0,24))


## Kidnapping Rose Plot
ggplot(Kidnapping, aes(x = Hour)) +
  geom_bar(width=1, fill = "royalblue1", color="grey") +
  coord_polar(start = 0)  + ylab("Count") + ggtitle("Kidnapping Count by Hour") +
  scale_x_continuous("", breaks = seq(0,24), labels = seq(0,24))

## Narcotics Rose Plot
ggplot(Narcotics, aes(x = Hour)) +
  geom_bar(width=1, fill = "green4", color="grey") +
  coord_polar(start = 0) + ylab("Count") + ggtitle("Narcotics Count by Hour") +
  scale_x_continuous("", breaks = seq(0,24), labels = seq(0,24))


## Theft Rose Plot
ggplot(Theft, aes(x = Hour)) +
  geom_bar(width=1, fill = "goldenrod", color="grey") +
  coord_polar(start = 0) +  ylab("Count") + ggtitle("Theft Count by Hour") +
  scale_x_continuous("", breaks = seq(0,24), labels = seq(0,24))
```