STAT480_Homework_8

Bin Feng

Setup

```
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

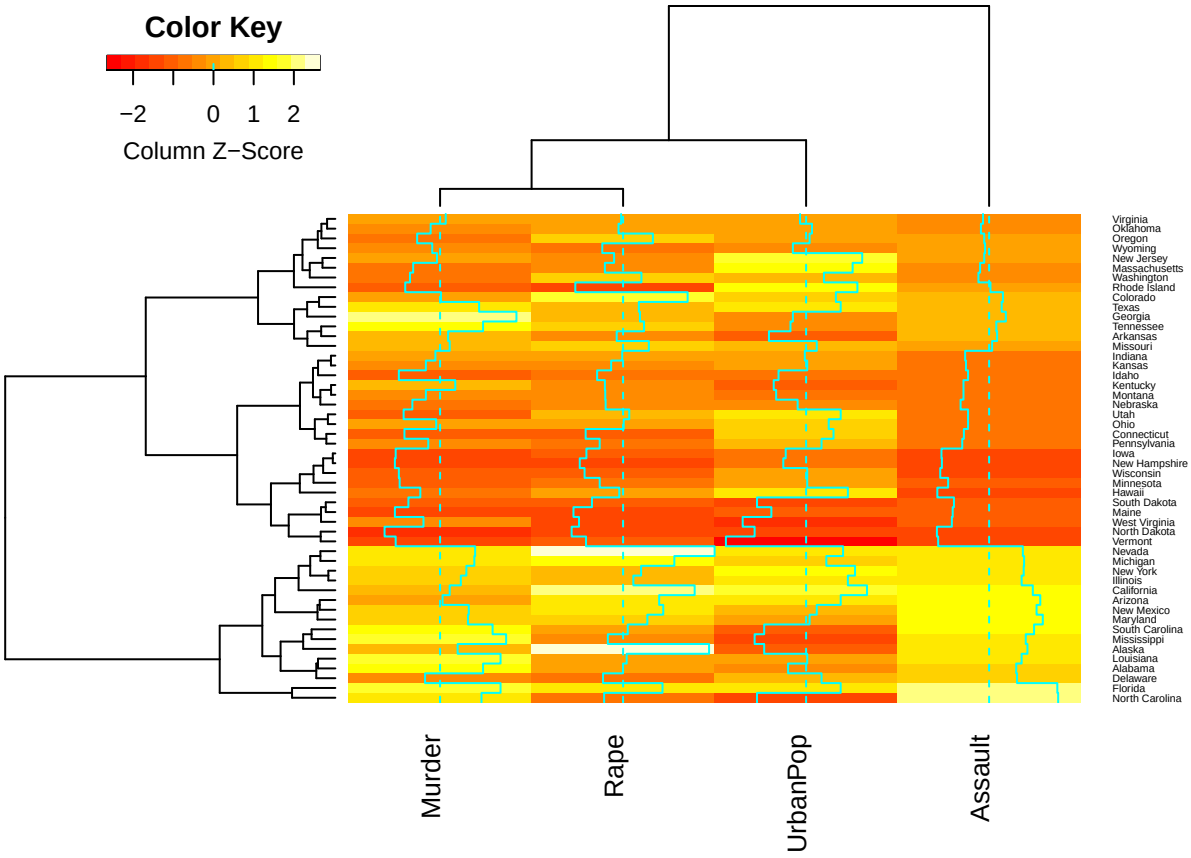
```
library(treemap)
library(knitr)
library(MASS)

x = USArrests
abb = state.abb
region = state.region
```

Exercise 1

Obtain a heatmap based on the USArrests data set, and interpret the plot.
Specifically, comment on which states have higher and lower values of these statistics, comment on any apparent relationships between these crime rates and urban population, and comment on groups of states that are the most similar with respect to these statistics and groups of states that are very different with respect to this statistics.

```
heatmap.2(as.matrix(x), scale = "column", density.info = "none",
          lwid = c(0.5,1.25), lhei = c(0.5,1.5), cexRow=0.45, cexCol=1)
```



Based on the heatmap as shown above, following observations can be drawn:

1.1) states have higher murder rate: Georgia, Mississippi, Florida & Louisiana (tied)
2.1) states have higher rape rate: Nevada, Alaska, California
3.1) states have higher assault rate: North Carolina, Florida, Maryland
4.1) states have higher urbanpop: California, New Jersey, Rhode Island
1.2) states have lower muder rate: North Dakota, Maine, New Hampshire
2.2) states have lower rape rate: North Dakota, Maine, Rhode Island
3.2) states have lower assault rate: North Dakota, Hawaii, Vermont
4.2) states have lower urbanpop: Vermont, West Virginia, North Dakota

Based on the above observations as well as the trace lines in the heatmap. We note that states with lower urbanpop usually have lower rates for all three crimes and vice versa, e.g. North Dabota has the lowest rates for all three crimes and also has the 3rd lowest urbanpop.

Baed on the dendrograms (shorter the dendrogram line length, more similar the two states based on these statistics; and vice versa), such observations can be drawn:

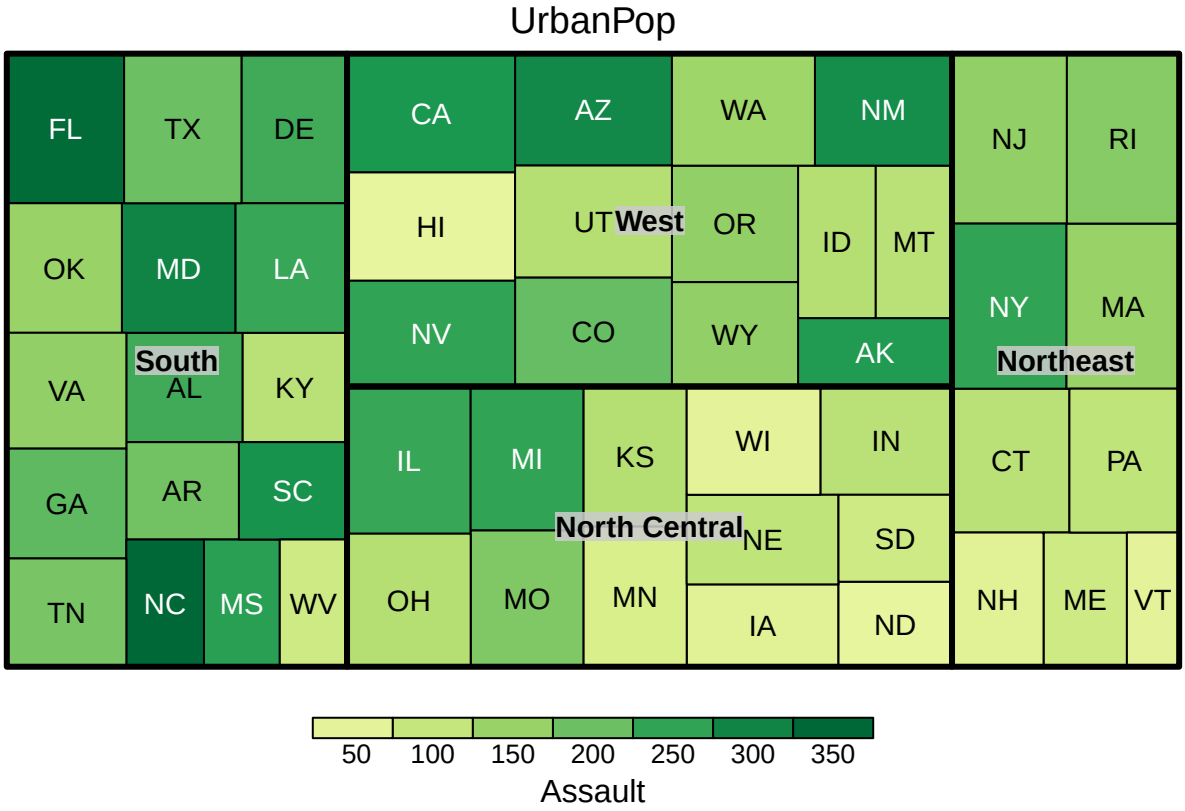
1) groups of states that are most similar: Iowa and New Hampshire.
2) groups of state that are very different: Navada and Vermont, Michigan and North Dakota, etc.

Exercise 2

Combine the USArrests data with the region and abbreviation (abb) variables from the state data set to create a data set for the treemap exercises.
Use a treemap to compare state-by-state urban population and assault rates, grouping by region within the plot.
Comment on urban populations and assault rates by region and within region. What were the general

differences between regions at the time? What were the differences within regions? Which states had the highest and lowest urban populations and assault rates overall in 1973, and which were highest and lowest within regions.

```
combine = cbind(x, abb, region)
treemap(combine,
        index=c("region", "abb"),
        vSize="UrbanPop",
        vColor="Assault",
        type="value")
```



Based on the treemap shown above, following observations can be drawn:

Between regions:

1) South and West have higher assault rate than North Central and Northest;

2) States in South region usually have smaller urban population than the other three regions.

Within regions:

1.1) In South region, Florida has larger urban population and higher assault rate; North Carolina has small urban population but very high assault rate;

1.2) In South region, Kentucky and West Virginia have smaller urban population and lower assault rate;

1.3) In South region, most states have high assault rate except Kentucky and West Virginia.

2.1) In West region, Arizona has larger urban population and higher assault rate;

2.2) In West region, Hawaii has larger urban population but very low assault rate;

2.3) In West region, state wise assualt rate varies significantly from very high (e.g. Arizona) to very low (e.g. Hawaii).

3.1) In North Central region, Michigan has larger urban population and higher assault rate;

3.2) In North Central region, many states (e.g. Wisconsin, Iowa, etc) have moderate urban population and low assault rate;

3.3) In North Central region, state wise assualt rate varies noticeably from high (e.g. Michigan) to low (e.g. Wisconsin).

4.1) In Northeast region, New York has larger urban population and higher assault rate;

4.2) In Northeast region, Vermont and New Hampshire have small urban population and low assault rate;

4.2) In Northeast region, most states have moderate assault rate except New York.

Overall:

```
# create dataframe for table
table2.1 = data.frame("Assault" = c("North Carolina", "North Dakota"),
                      "UrbanPop" = c("California", "Vermont"))
rownames(table2.1) <- c("Highest","Lowest")
kable(table2.1, caption = "Overall")
```

Table 1: Overall

	Assault	UrbanPop
Highest	North Carolina	California
Lowest	North Dakota	Vermont

Regions Wise:

```
# create dataframe for table
table2.2 = data.frame("Assault(highest)" = c("North Carolina", "Arizona", "Michigan", "New York"),
                      "Assault(lowest)" = c("West Virginia", "Hawaii", "North Dakota", "Vermont"),
                      "UrbanPop(highest)" = c("Texas", "California", "Illinois", "New Jersey"),
                      "UrbanPop(lowest)" = c("West Virginia", "Alaska", "North Dakota", "Vermont"))
rownames(table2.2) <- c("South","West", "North Central", "Northest")
kable(table2.2, caption = "State Wise")
```

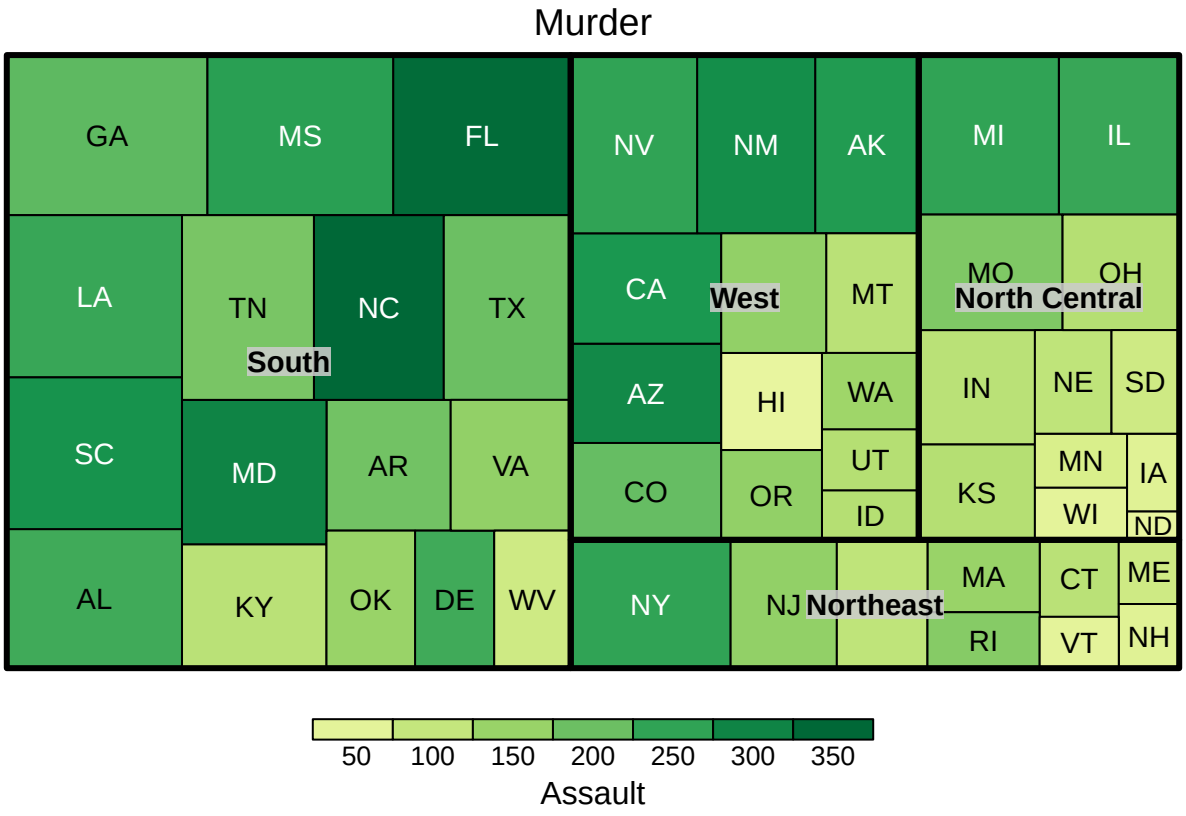
Table 2: State Wise

	Assault.highest.	Assault.lowest.	UrbanPop.highest.	UrbanPop.lowest.
South	North Carolina	West Virginia	Texas	West Virginia
West	Arizona	Hawaii	California	Alaska
North Central	Michigan	North Dakota	Illinois	North Dakota
Northest	New York	Vermont	New Jersey	Vermont

Exercise 3

Repeat Exercise 2 with murder and assault rates instead of urban population and assault rates.

```
treemap(combine,
        index=c("region", "abb"),
        vSize="Murder",
        vColor="Assault",
        type="value")
```



Based on the treemap shown above, following observations can be drawn:

Between regions:

1) South region has the highest murder rate; Northeast region has the lowest murder rate;
2) States in South region also usually have higher assault rate than the other three regions;
3) state with high murder rate usually also have high assault rate and vise versa.

Within regions:

1.1) In South region, Florida and North Carolina have high murder and assault rate;
1.2) In South region, West Virginia have relative small murder and assault rate;
1.3) In South region, most states have high assault rate except Kentucky and West Virginia, all states have moderate to high murder rate;
2.1) In West region, five states (NV, NM, AK, CA, AZ) have high murder and assault rate;
2.2) In West region, four states (OR, WA, UT, ID) have very low murder and assault rate;
2.3) In West region, murder and assault rate differ largely among different states;
3.1) In North Central region, Michigan and Illinois have both high assault and murder rate;
3.2) In North Central region, North Dakota has both very low assault and murder rate;
3.3) In North Central region, Most states have moderate to low assault and murder rate;
4.1) In Northeast region, New York has both high murder and assault rate;
4.2) In Northeast region, Vermont and New Hampshire have low murder and assault rate;
4.3) In Northeast region, most states have moderate to low assault and murder rate except New York.

Overall:

```
# create dataframe for table
table3.1 = data.frame("Assault" = c("North Carolina", "North Dakota"),
                      "Murder" = c("Georgia", "North Dakota"))
rownames(table3.1) <- c("Highest","Lowest")
kable(table3.1, caption = "Overall")
```

Table 3: Overall

	Assault	Murder
Highest	North Carolina	Georgia
Lowest	North Dakota	North Dakota
Regions	Wise:**	

```
# create dataframe for table
table3.2 = data.frame("Assault(highest)" = c("North Carolina", "Arizona", "Michigan", "New York"),
                      "Assault(lowest)" = c("West Virginia", "Hawaii", "North Dakota", "Vermont"),
                      "Murder(highest)" = c("Georgia", "Nevada", "Michigan", "New York"),
                      "Murder(lowest)" = c("West Virginia", "Idaho", "North Dakota", "Maine"))
rownames(table3.2) <- c("South","West", "North Central", "Northest")
kable(table3.2, caption = "State Wise")
```

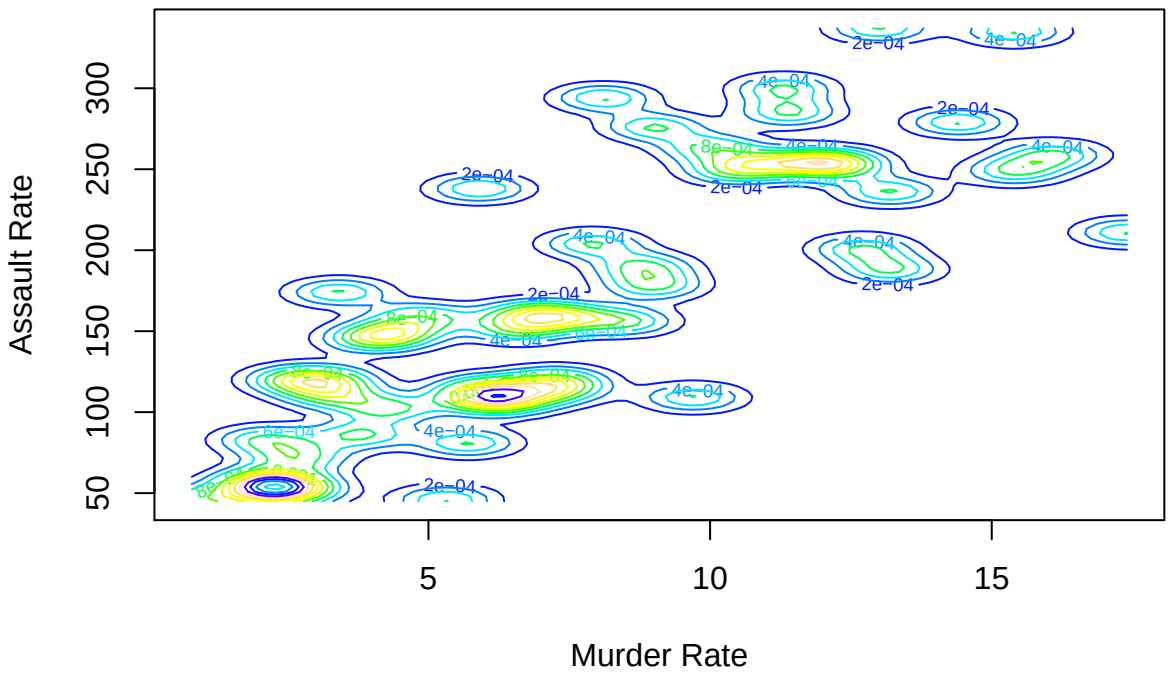
Table 4: State Wise

	Assault.highest.	Assault.lowest.	Murder.highest.	Murder.lowest.
South	North Carolina	West Virginia	Georgia	West Virginia
West	Arizona	Hawaii	Nevada	Idaho
North Central	Michigan	North Dakota	Michigan	North Dakota
Northest	New York	Vermont	New York	Maine

Exercise 4

Create visualizations of the following three kernel density estimations: 1. Murder rate and assault rate 2. Murder rate and urban population 3. Assault rate and urban population

```
fit4.1 = kde2d(x$Murder, x$Assault, h=c(2.5,25), n=100)
contour(fit4.1, col = topo.colors(10), xlab = "Murder Rate",
        ylab = "Assault Rate")
```



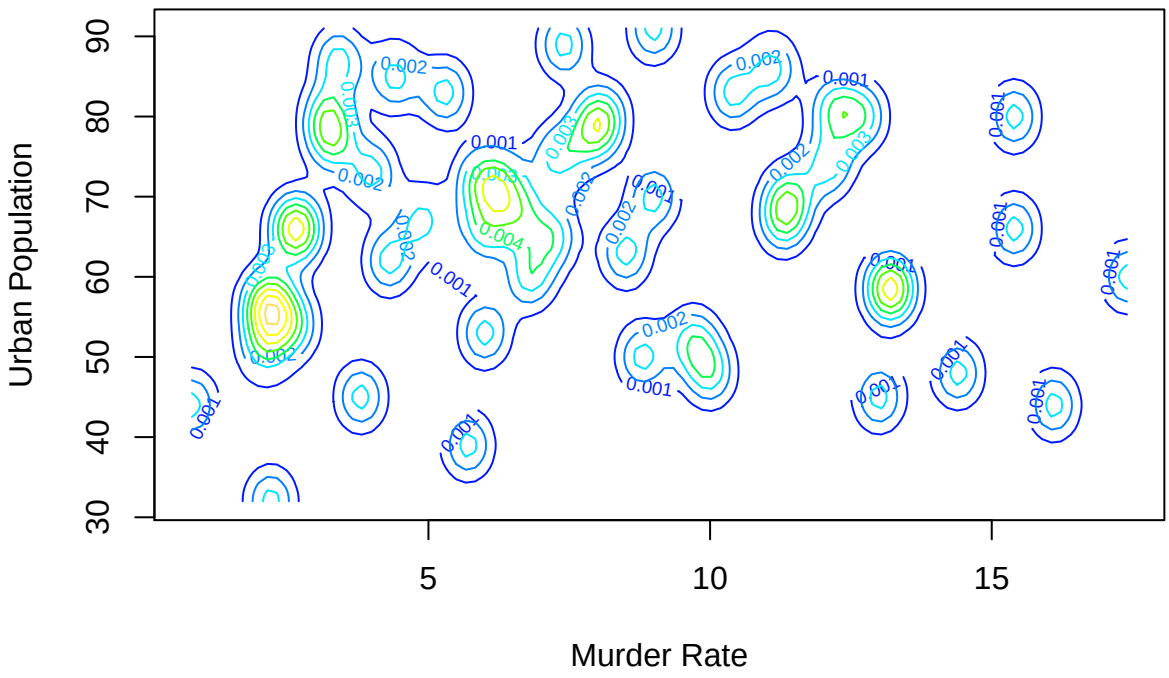
Based on the kernal density plot above, following observations can be drawn:

1) three common magnitude ranges for murder rate are 1-4, 5-8 and 10-13;

2) three common magnitude ranges for assault rate are 50-75, 100-170, and 240-280;

3) muder rate vs. assault rate has positive correlation. If murder rate increases, assault rate will also increase.

```
fit4.2 = kde2d(x$Murder, x$UrbanPop, h=c(1.25,12), n=100)
contour(fit4.2, col = topo.colors(10), xlab = "Murder Rate",
        ylab = "Urban Population")
```



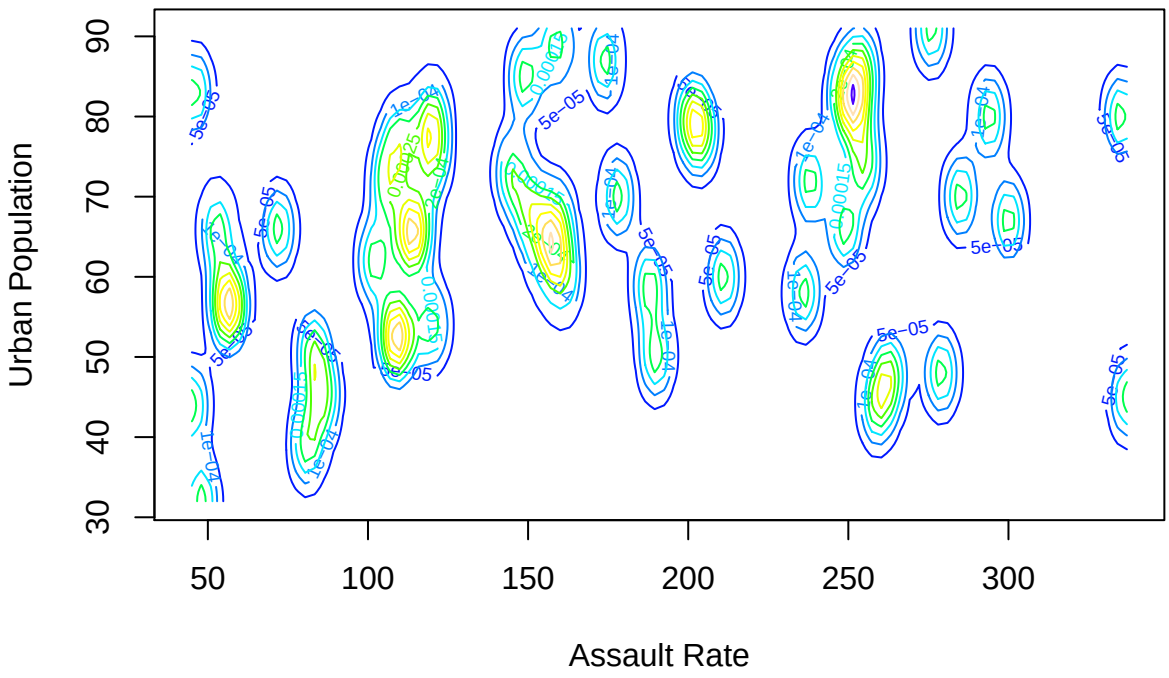
Based on the kernal density plot above, following observations can be drawn:

1) three common magnitude ranges for murder rate are 1-4, 5-8 and 10-13;

2) one common magnitude range for urban population is 50-80;

3) murder rate vs. urban population are quite spreaded. If murder rate increases, urban population won't necessarily increase.

```
fit4.3 = kde2d(x$Assault, x$UrbanPop, h=c(15,15), n=100)
contour(fit4.3, col = topo.colors(10), xlab = "Assault Rate",
        ylab = "Urban Population")
```

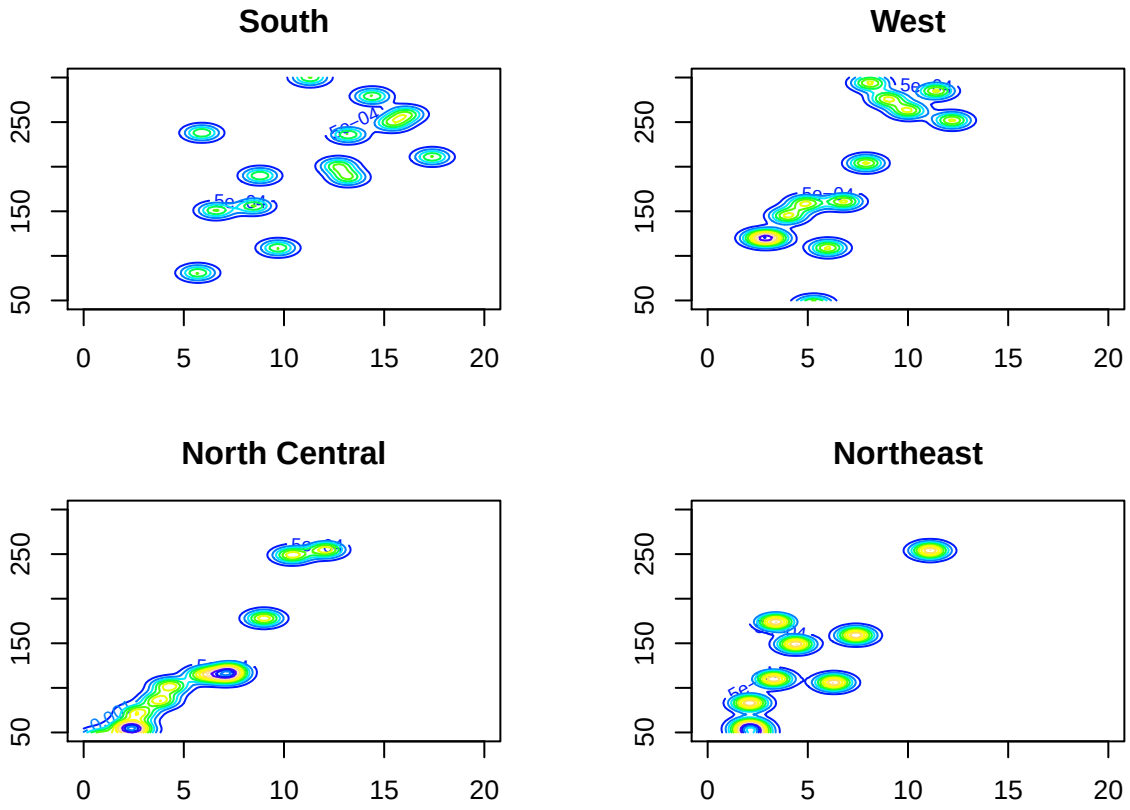


Based on the kernal density plot above, following observations can be drawn:
1) three common magnitude ranges for assault rate are 50-75, 100-170, and 240-280;
2) one common magnitude range for urban population is 50-80;
3) assault rate vs. urban population are quite spreaded. If assault rate increases, urban population won't necessarily increase.

Exercise 5

Obtain and compare visualizations of murder and assault rate kernel density estimations by region using the region variable added for the treemaps. Comment on how the arrest rates differ across regions and how the individual regions compare with the overall density estimation for murder and assault rates from Exercise 4.

```
region = c("South", "West", "North Central", "Northeast")
par(mfrow=c(2,2), mar=c(3,3,3,3))
for(i in region){
  fit5 = kde2d(combine[combine$region==i,]$Murder, combine[combine$region==i,]$Assault,
               h=c(2.5,25), n=100, lims = c(0, 20, 50, 300))
  contour(fit5, col = topo.colors(10), xlab = "Murder Rate",
        ylab = "Assault Rate", xlim = c(0,20),  ylim = c(50, 300), main = i)
}
```



Kernal density plots for four different regions are shown above. For all plots, x axis is murder rate, y axis is assault rate.

Note that, assault rate and murder rate are much higher in South region than the other three. North Central and Northeast are two regions with relative low assault rate and murder rate. Also note that the assault rate and murder rate relationship is speaded in South region while in North Central region, the positive correlation is more significant.

Compare individual region plots with the overall density plot, we note that individual region plots are like one portion of the overall plot (e.g. North Central region plot only covers the low murder and assault rate area). The overall plot contain information from all four regions. Therefore, some trends appear in individual region plots may not be noticeable in the overall region plot.