STAT480 Homework8

Chenz Zhang, NetID chenziz2

4/10/2019

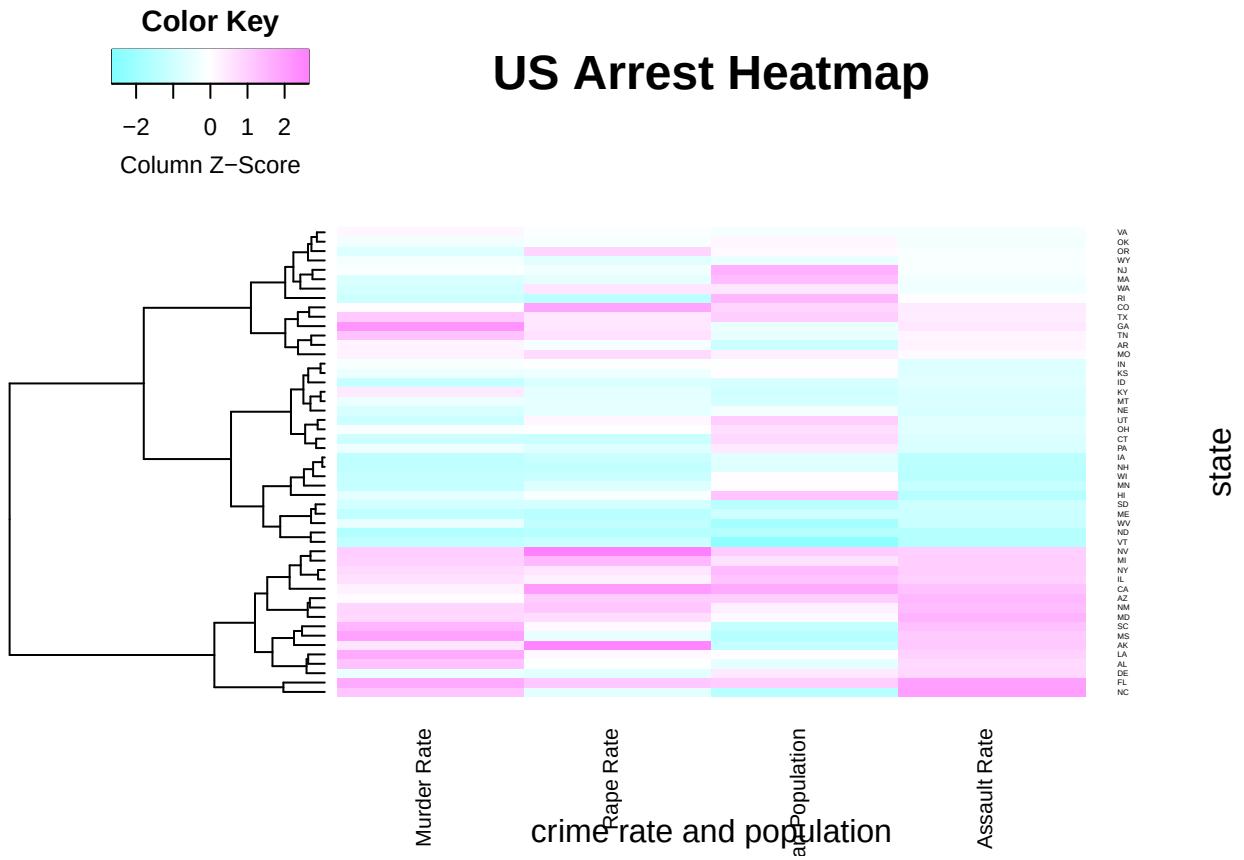
Contents

Command line code in 'ChenziZhangHW8.R' was modified by Chenzi Zhang from R help for heatmap. It is also based on code segments written by James Balamuta for University of Illinois course Stat 430 Big Data Analysis Foundations in Spring 2015 and Darren Glosemeyer's codes 'heatmaptreemapdensityplot.R' for University of Illinois course Stat 480 Data Science Foundations in Spring 2019.

Exercise 1

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

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



I use `heatmap.2` to obtain heatmap from `USArrest` dataset. In this `heatmap.2` function, I set column scaling and remove the trace lines as well as histograms. I use dendrograms clustering rows to find cluster of states. Besides, I change the layout from default setting to new colar and labels.

1. Which states have higher and lower values of these statistics

As a result, I get the heatmap graph as above picture. When the color turns to pink, the magnitude of variables increase. On the contrary (when the color turns to blue), the the magnitude of variables decrease. From the heatmap, I can conclude:

Table 1: States with highest 3 values on statestics

Murder	Assault	Rape	UrbanPop
GA	NC	NV	CA
MS	FL	AK	NJ
LA	MD	CA	RI

Table 2: States with lowest 3 values on statestics

Murder	Assault	Rape	UrbanPop
ND	ND	RI	VT
ME	HI	ME	WV
NH	VT	ND	MS

2. Apparent relationships between these crime rates and urban population

With larger urban population, these crime rates turn to be higher. This relation is more appearant for two crime rates, `Rape` and `Assault`. But unexpectial condition indeed happens. For example, state AR has small urban population with relatively higher three crime rates.

On the contrary, with smaller urban population, all the three crime rates tend to be lower. This relation is more appearant for states from SD to VT. But unexpectial condition happens, too. For example, state HI has large urban population with relatively lower three crime rates.
Thus, I should say that the real relation (negatively related or positively related) depends on which group or cluster the state is in.

In the heatmap, state are automatically divided into groups. Then, I analyze these groups by two parts:

3. Groups of states that are the most similar with respect to these statistics

Group1: Similarly big value.
NV, MI, NY, IL, CA, AZ, NM, MD

Group2: Similarly small value.
SD, ME, WV, ND, VT

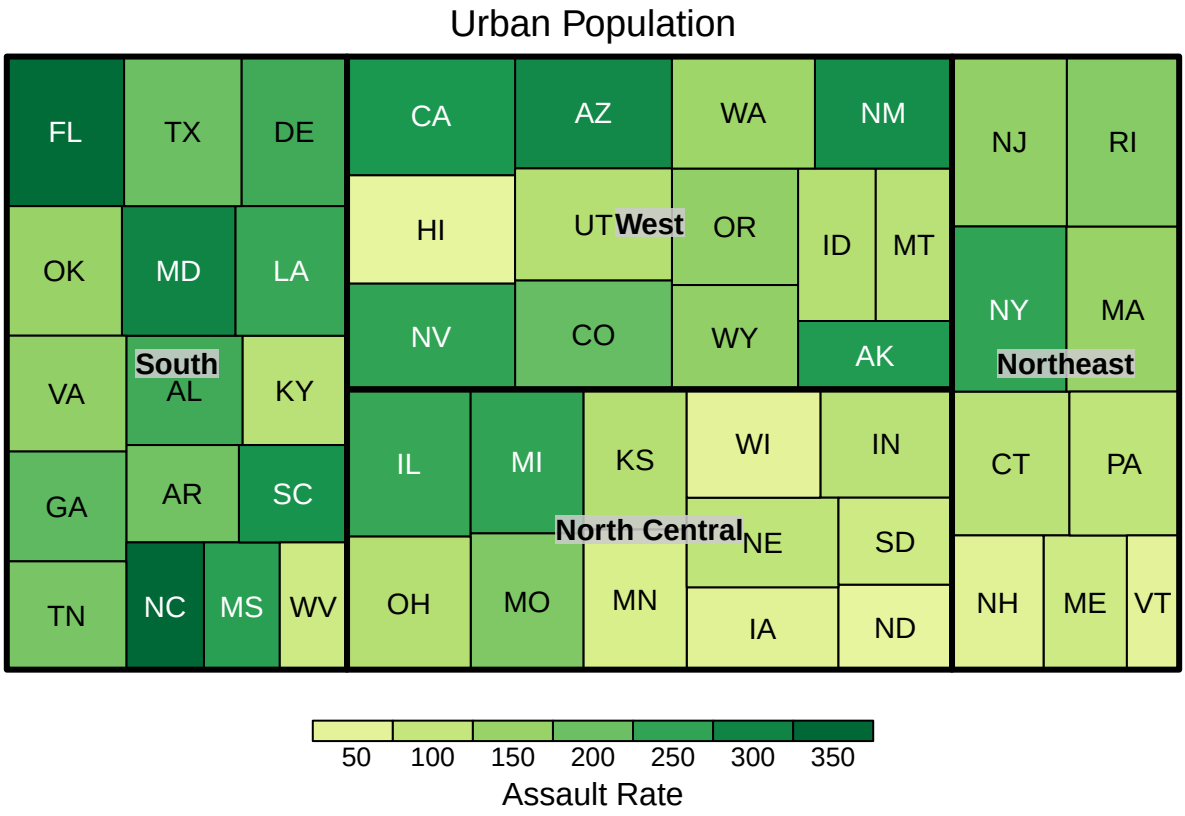
4. Groups of states that are very different with respect to this statistics

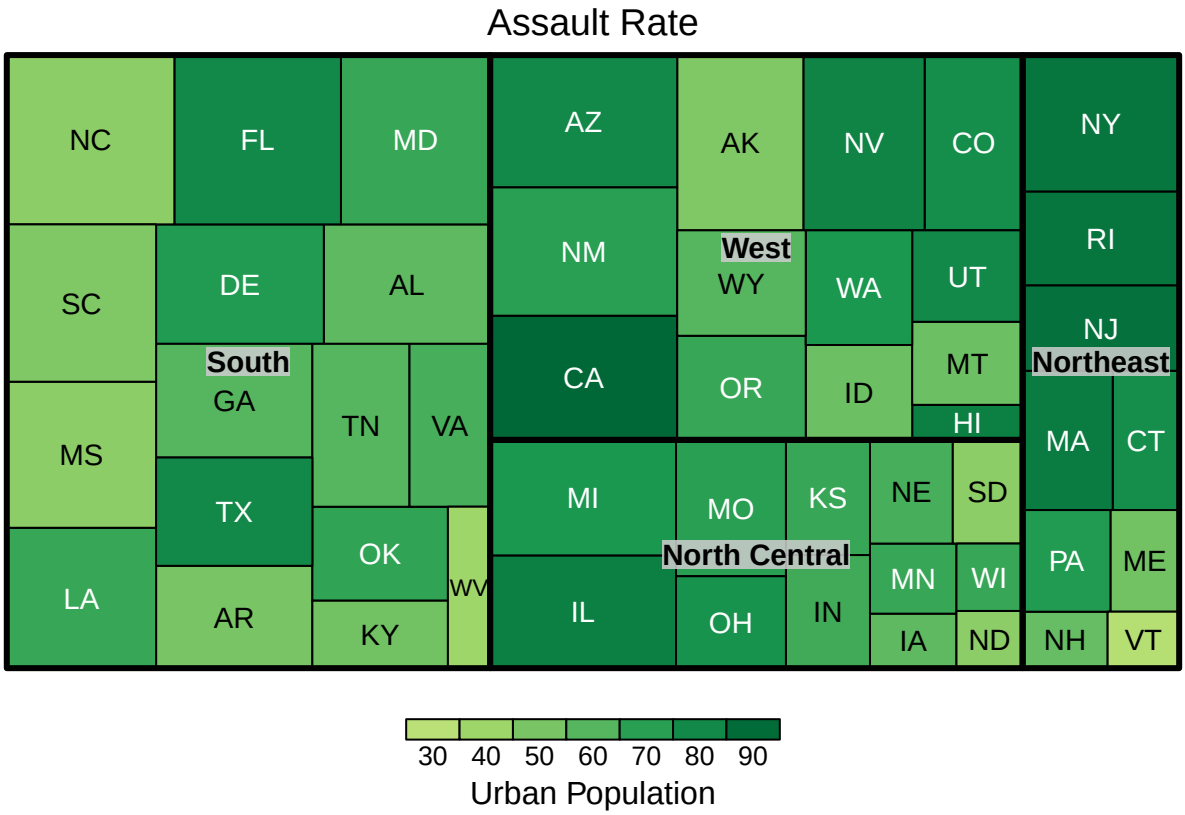
Group1: Large urban population with low crime rate.
NJ, MA, WA; UT, OH, CT, PA

Group2: Small urban population with high crime rate.
SC, MS, AK

Other groups do not have very appearant relation between crime rates and population.

Exercise 2


Urban Population
FL
TX
DE
CA
AZ
WA
NM
NJ
RI
HI
UT
West
OR
ID
MT
OK
MD
LA
NY
MA
NV
CO
WY
AK
VA
South
AL
KY
Northeast
GA
AR
SC
IL
MI
KS
WI
IN
CT
PA
North Central
NE
SD
TN
NC
MS
WV
OH
MO
MN
IA
ND
NH
ME
VT
50
100
150
200
250
300
350
Assault Rate



I plot two treemaps. One shows Urban Population by size and Assault Rate by color compared state-by-state. One shows Assault Rate by size and Urban Poplution by color compared state-by-state.

General differences between regions

In the term of `Urban Population`, `North Central` has the largest area in the first treemap and and the darkest green color in the second treemap. From the area size in the first treemap and darkness of green color in the second treemap, I can sort `Urban Population` by regions in descent order: South, West, North Central, Northeast.

In the term of `Assault Rate`, `South` has the darkest green color in the first treemap and the largest area in the second treemap. From the area size in the second treemap and darkness of green color in the first treemap, I can sort `Assault Rate` by regions in descent order: South, West, Northeast, North Central.

General differences within regions

Let's see the first treemap.

In South region, FL is the second largest rectangular with the darkest green color. This means that FL has the second largest Urban Popluation and highest Assault Rate. WV is the smallest rectangular with the lightest green color. This means that WV has the smallest Urban Population and lowest Assualt Rate.

In West region, AZ is a relatively large rectangular with the darkest gree color. This mean taht AZ has relatively large population and highest Assault Rate in the West. HI is a large rectangular with the lightest green color. This means that HI has large Urban Population and lowest Assualt Rate.

In North Central region, IL is the largest rectangular with the darkest green color. This means that IL has the largest Urban Popluation and highest Assault Rate. ND is the smallest rectangular with the lightest green color. This means that ND has the smallest Urban Population and lowest Assualt Rate.

In Northeast region, NY is the largest rectangular with the darkest green color. This means that NY has the largest Urban Popluation and highest Assault Rate. VT is the smallest rectangular with the lightest green color. This means that VT has the smallest Urban Population and lowest Assualt Rate.

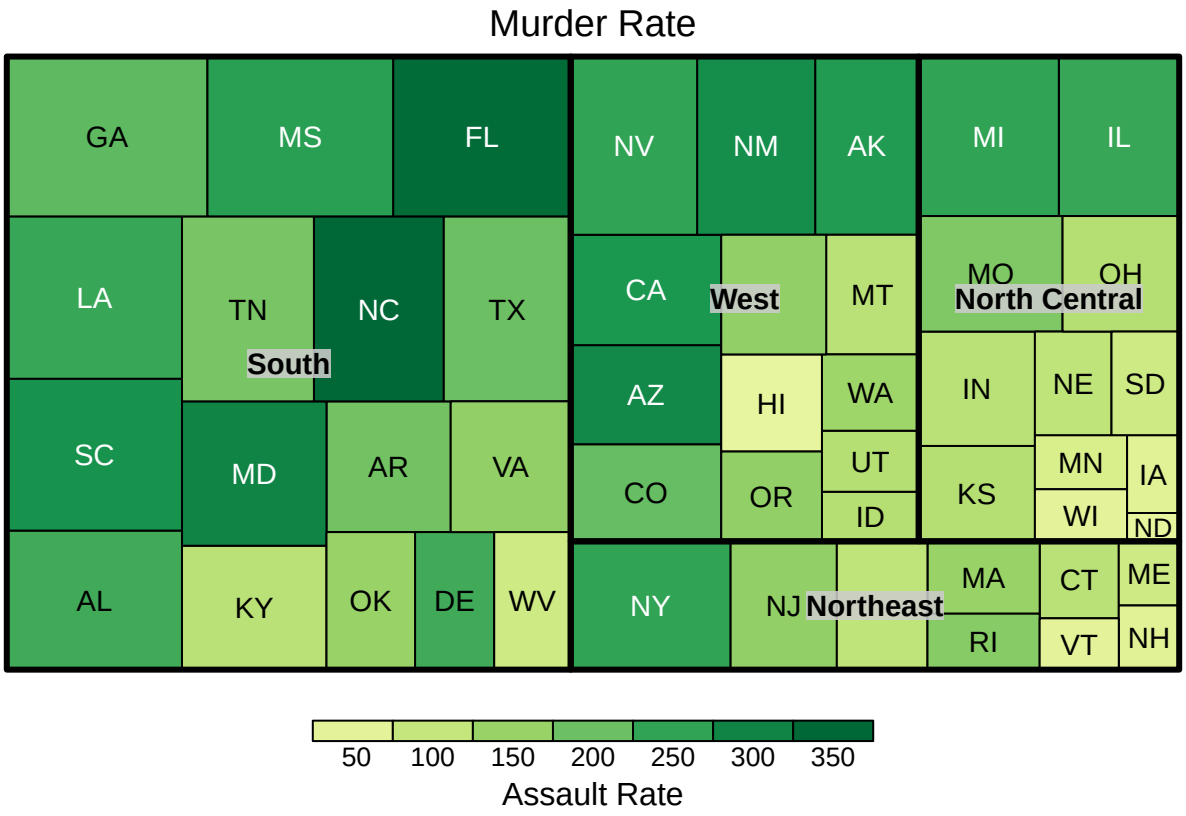
Other comments

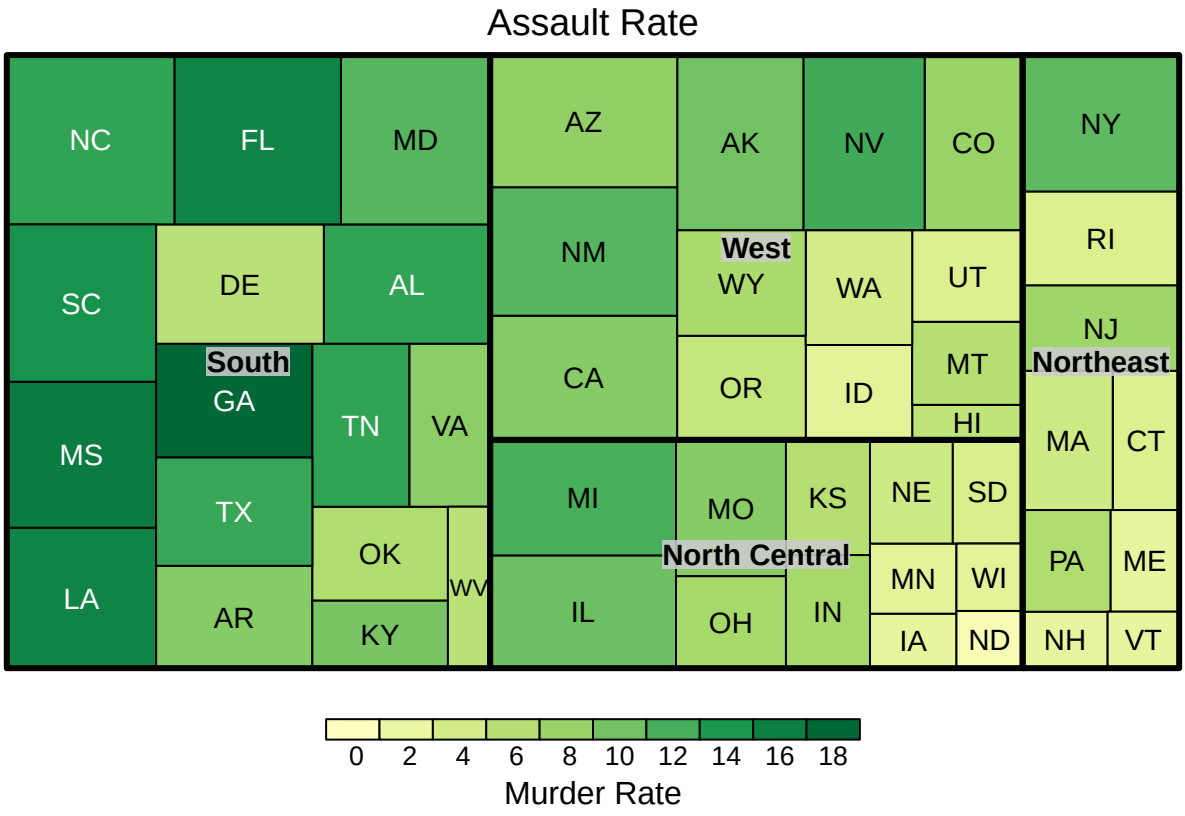
Overall in 1973:
State has the largest Urban Popluation: CA
State has the smallest Urban Popluation: VT
State has the highest assault rate: NC State has the lowest assault rate: ND

Within regions:

Region	Assault(lowest)	Assault(higest)	UrbanPop(smallest)	UrbanPop(largest)
South	WV	NC	WV	FL
West	HI	AZ	AK	CA
North Central	ND	MI	IL	ND
Northeast	VT	NY	VT	NJ

Exercise 3


Murder Rate
GA
MS
FL
NV
NM
AK
MI
IL
LA
TN
NC
TX
CA
West
MT
MO
OH
South
North Central
AZ
HI
WA
IN
NE
SD
SC
MD
AR
VA
CO
OR
UT
KS
MN
IA
ID
WI
ND
AL
KY
OK
DE
WV
NY
NJ
Northeast
MA
CT
ME
RI
VT
NH
50
100
150
200
250
300
350
Assault Rate



General differences between regions

In the term of `Murder Rate`, `South` has the largest area in the first treemap and and the darkest green color in the second treemap. From the area size in the first treemap and darkness of green color in the second treemap, I can sort `Murder Rate` by regions in descent order: South, West, North Central, Northeast.

In the term of `Assault Rate`, `South` has the darkest green color in the first treemap and the largest area in the second treemap. From the area size in the second treemap and darkness of green color in the first treemap, I can sort `Assault Rate` by regions in descent order: South, West, North Central, Northeast.

The orders `Murder Rate` and `Assault Rate` for are totally the same.

General differences within regions

Let's see the first treemap.

In South region, FL and NC are larges rectangular with the darkest green color. This means that FL and NC have high Murder Rate and highest Assault Rate in the South. GA is the largest rectangular, meaning GA has the highest Murder Rate. WV is the smallest rectangular with the lightest green color. This means that WV has the smallest Murder Rate and lowest Assualt Rate.

In West region, AZ is a relatively large rectangular with the darkest gree color. This mean taht AZ has relatively high Murder Rate and highest Assault Rate in the West. HI is a small rectangular with the lightest green color. This means that HI has low Murder Rate and lowest Assualt Rate. NV is the largest rectangular, meaning NV has the highest Murder Rate. ID is the smallest rectangular, meaning ID has the lowest Murder Rate.

In North Central region, MI and IL are the largest rectangular with the darkest green color. This means that MI and IL have the highest Murder Rate and highest Assault Rate. ND is the smallest rectangular with the lightest green color. This means that ND has the lowest Murder Rate and lowest Assualt Rate.

In Northeast region, NY is the largest rectangular with the darkest green color. This means that NY has the highest Murder Rate and highest Assault Rate. NH is the smallest rectangular with the lightest green color. This means that NH has the lowest Murder Rate and lowest Assualt Rate.

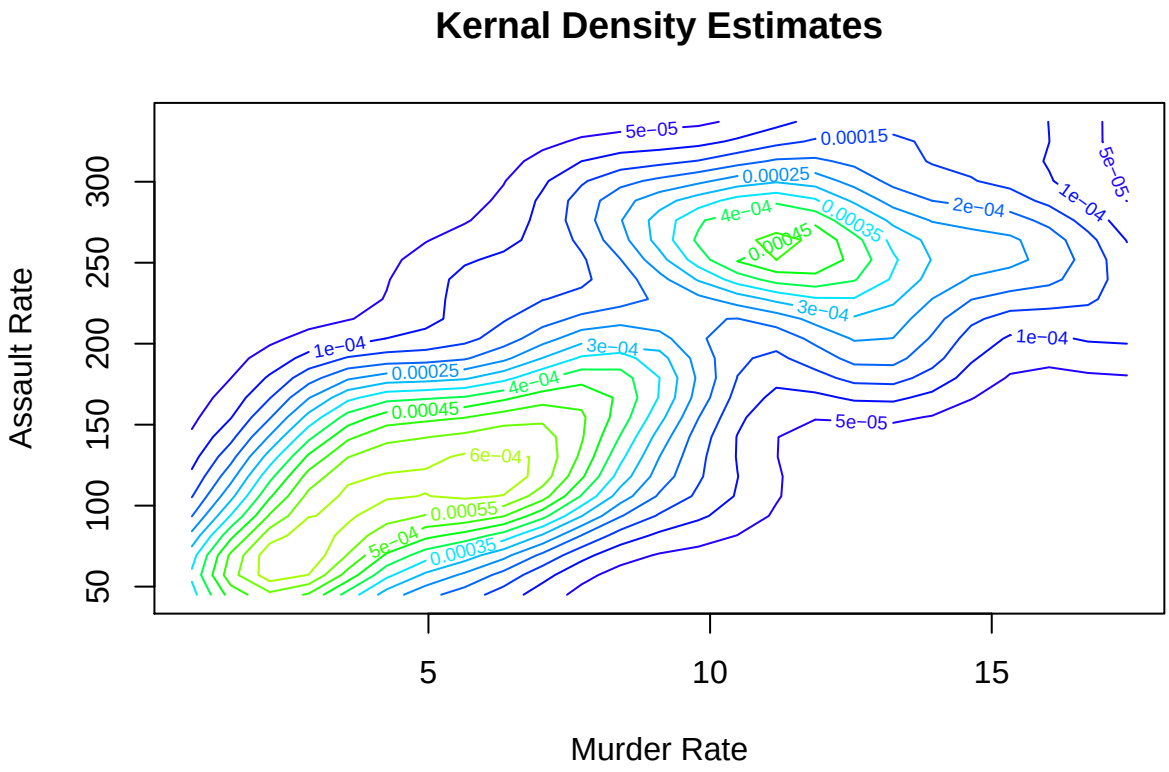
Other commends

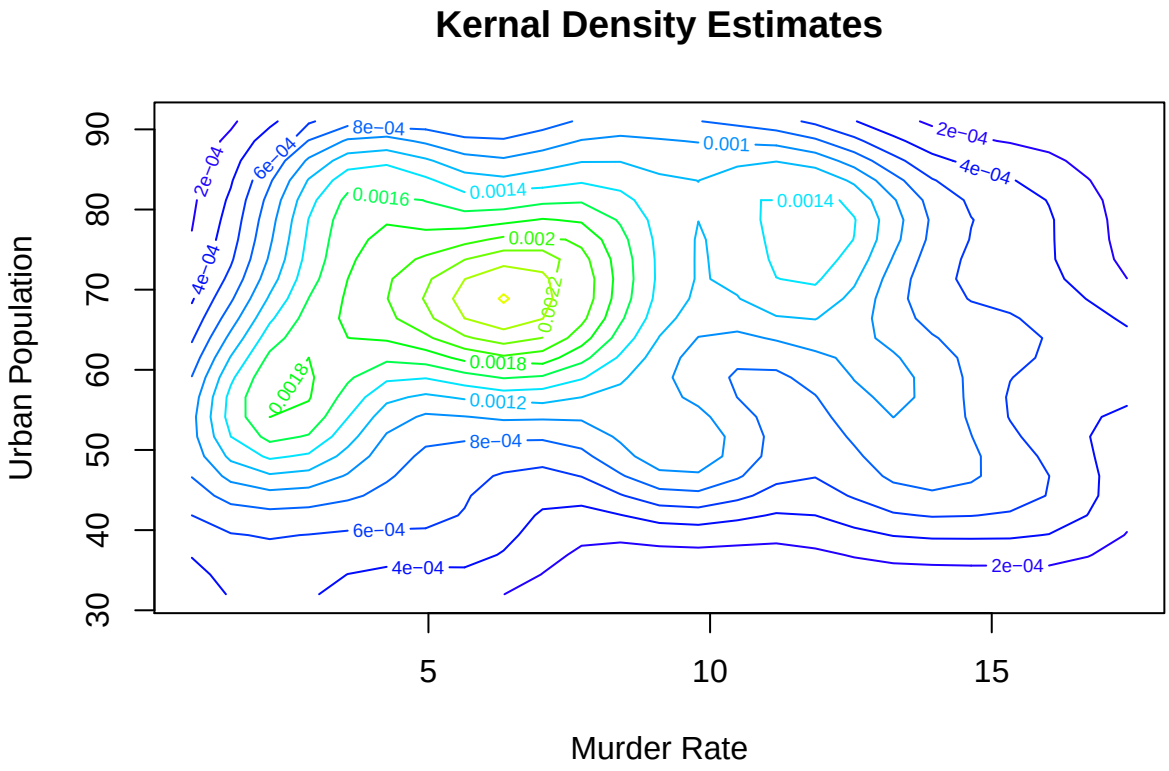
Overall in 1973:
State has the highest Murder Rate: CA
State has the lowest Murder Rate: VT
State has the highest Assault Rate: NC State has the lowest Assault Rate: ND

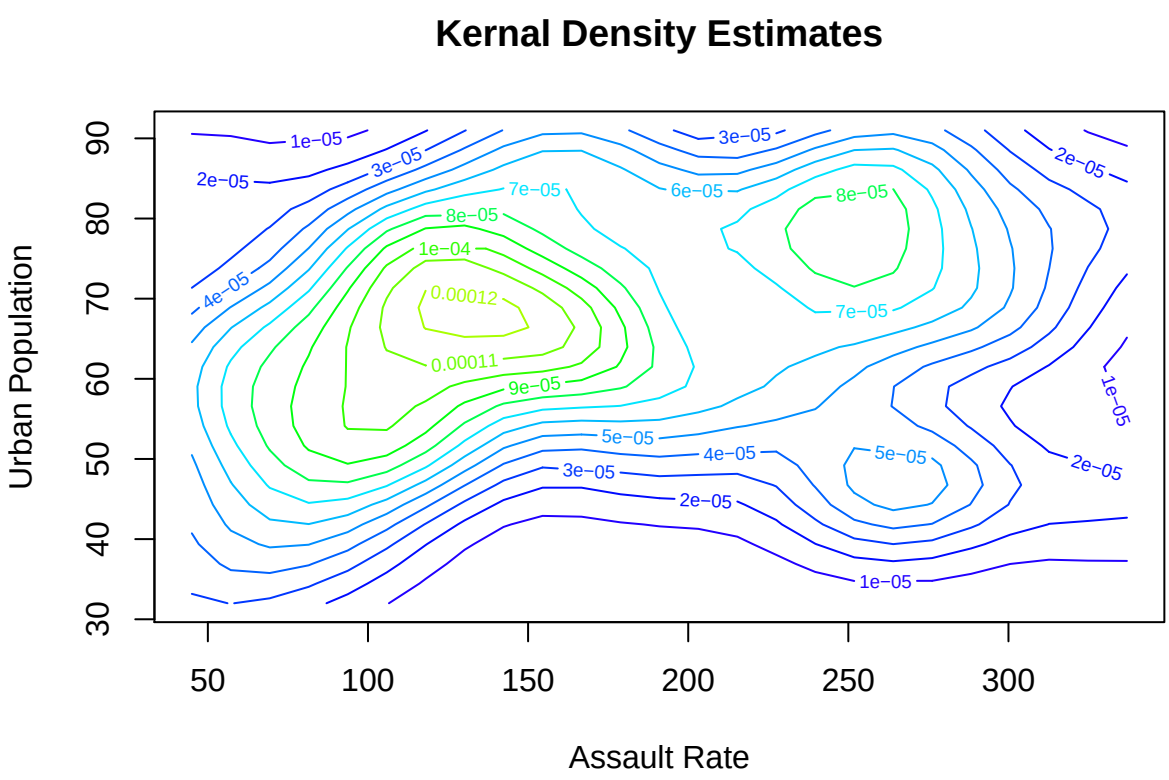
Within regions:

Region	Assault(lowest)	Assault(highest)	Murder(lowest)	Murder(highest)
South	WV	NC	WV	GA
West	HI	AZ	ID	NV
North Central	ND	MI	ND	MI
Northeast	VT	NY	NH	NY

Exercise 4


Kernal Density Estimates
Assault Rate
50
100
150
200
250
300
5
10
15
Murder Rate
5e−05
0.00015
0.00025
4e−04
0.00035
0.00045
3e−04
2e−04
1e−04
5e−05
1e−04
3e−04
0.00025
4e−04
0.00045
6e−04
0.00055
5e−04
0.00035
1e−04
5e−05


Kernal Density Estimates
Urban Population
Murder Rate
30
40
50
60
70
80
90
5
10
15
8e−04
0.001
2e−04
4e−04
6e−04
2e−04
0.0016
0.0014
0.0014
0.002
0.0022
0.0018
0.0018
0.0012
8e−04
6e−04
4e−04
2e−04



I calculated the default bandwiths and refined bandwiths by adjust `h` in `kde2d` function with `c(5,100)`, `c(5,25)` and `c(100,25)`. In this way, I can find the general trend between two variables in every graph.

More common magnitudes

Murder Rate and Assault Rate: (prob $>$ 0.0006) Murder Rate from 2 to 7, Assault Rate from 60 to 140.
Murder Rate and Urban Population: (prob $>$ 0.0024) Murder Rate from 5.5 to 7, Urban Population from 65 to 73.
Assault Rate and Urban Population: (prob $>$ 0.00012) Assault Rate from 118 to 150, Urban Population from 65 to 71.

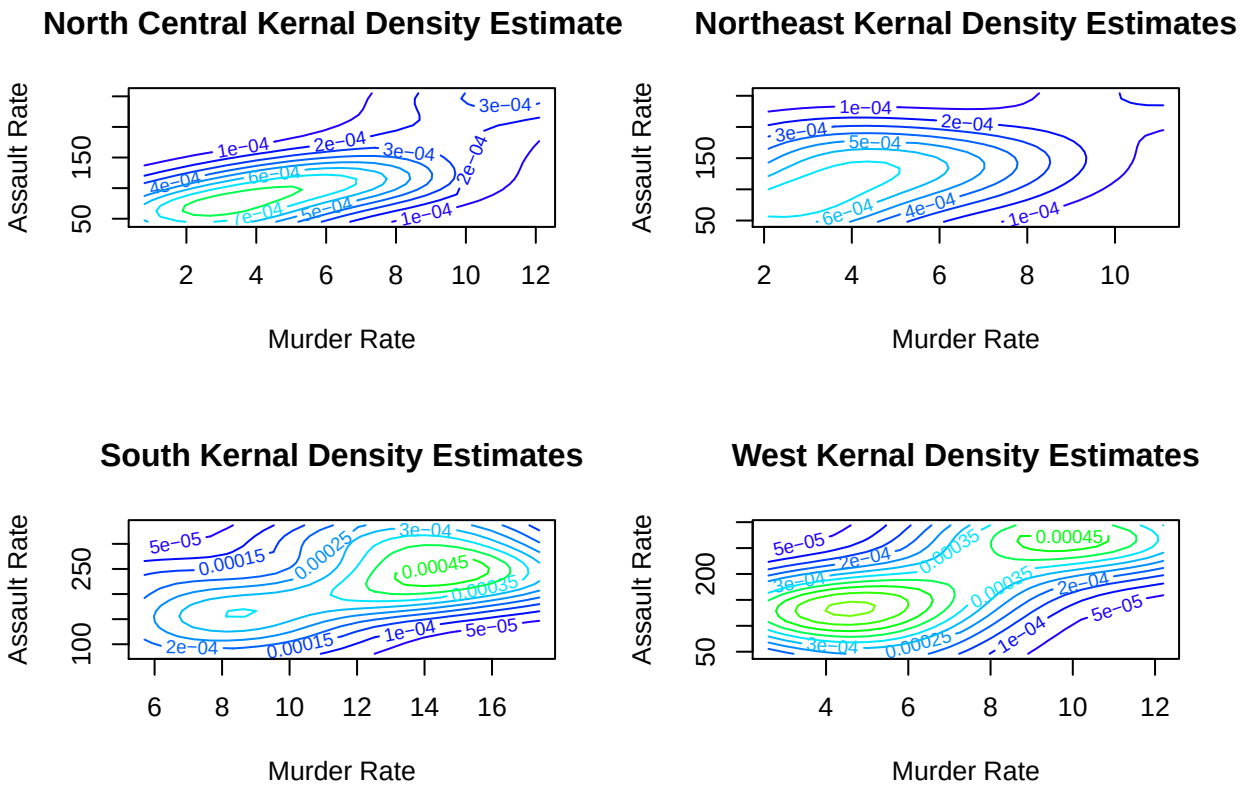
Relation

Murder Rate and Assault Rate have a roughly positive linear relation, meaning that if Murder Rate increases, Assault Rate increases too.

Murder Rate and Urban Population have a positive relation. If population increases, Murder Rate increases too. But Urban Population's increase speed is large than Murder Rate, so the common magnitudes stay at high Urban Population and relatively constant Murder Rate.

Assault Rate and Urban Population have a positive relation, too. If Urban Popualtion increases, Assault Rate has increasing trend too. But comparing to Murder&Pop, the sample points are more sparse, leading to less common magnitudes.

Exercise 5



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.

How the arrest rates differ across regions

North Central and Northeast have relatively low arrest rates. Besides, arrest rates in these two region are sparse, especially Northeast has the most concentrated arrest rate. This means these two regions are relatively safer.

South and West largest estimate prob are 0.00045. This indicates these two regions have relati vely more sparse points. South region has the largest arrest rates magnitude than all other three. West region has two most common magnitude because I can see two centrals in the plot. One is Murder Rate [4,5] with Assult Rate [120,150]. The other is Murder Rate [8.5,11] with Assult Rate [240,290].

How the individual regions compare with the overall density estimation for murder and assault rates from Exercise 4

The overall density estimation for murder and assult rate is:
Murder Rate and Assault Rate: (prob $>$ 0.0006) Murder Rate from 2 to 7, Assault Rate from 60 to 140.

North Central: (prob $>$ 0.0006) Murder Rate from 0 to 8, Assault Rate from 40 to 140. (prob $>$ 0.0008) Murder Rate from 2 to 5, Assault Rate from 50 to 110. It is smaller than the overall estimates. I can conclude arrest rates from North Central are both smaller than the overall arrest rates in the whole picture.

Northeast: (prob > 0.0006) Murder Rate from 0 to 6, Assault Rate from 40 to 160. (prob > 0.0007) Murder Rate from 1.5 to 5, Assault Rate from 50 to 150. Murder Rate is smaller than the overall estimates but Assault Rate is slightly high. I can conclude Murder Rate from Northeast is smaller than the overall arrest rates in the whole picture, but Assault Rate seems the same as overall arrest rates with less certainty.

South: (prob > 0.00045) Murder Rate from 13 to 16, Assault Rate from 210 to 260. It is much larger than the overall estimates. I can conclude arrest rates from South are both larger than the overall arrest rates in the whole picture.

West: (prob > 0.00045) Murder Rate from 2 to 7, from 9 to 11, Assault Rate from 150 to 160, from 250 to 280. It is much larger than the overall estimates. I can conclude arrest rates from West are both larger than the overall arrest rates in the whole picture.