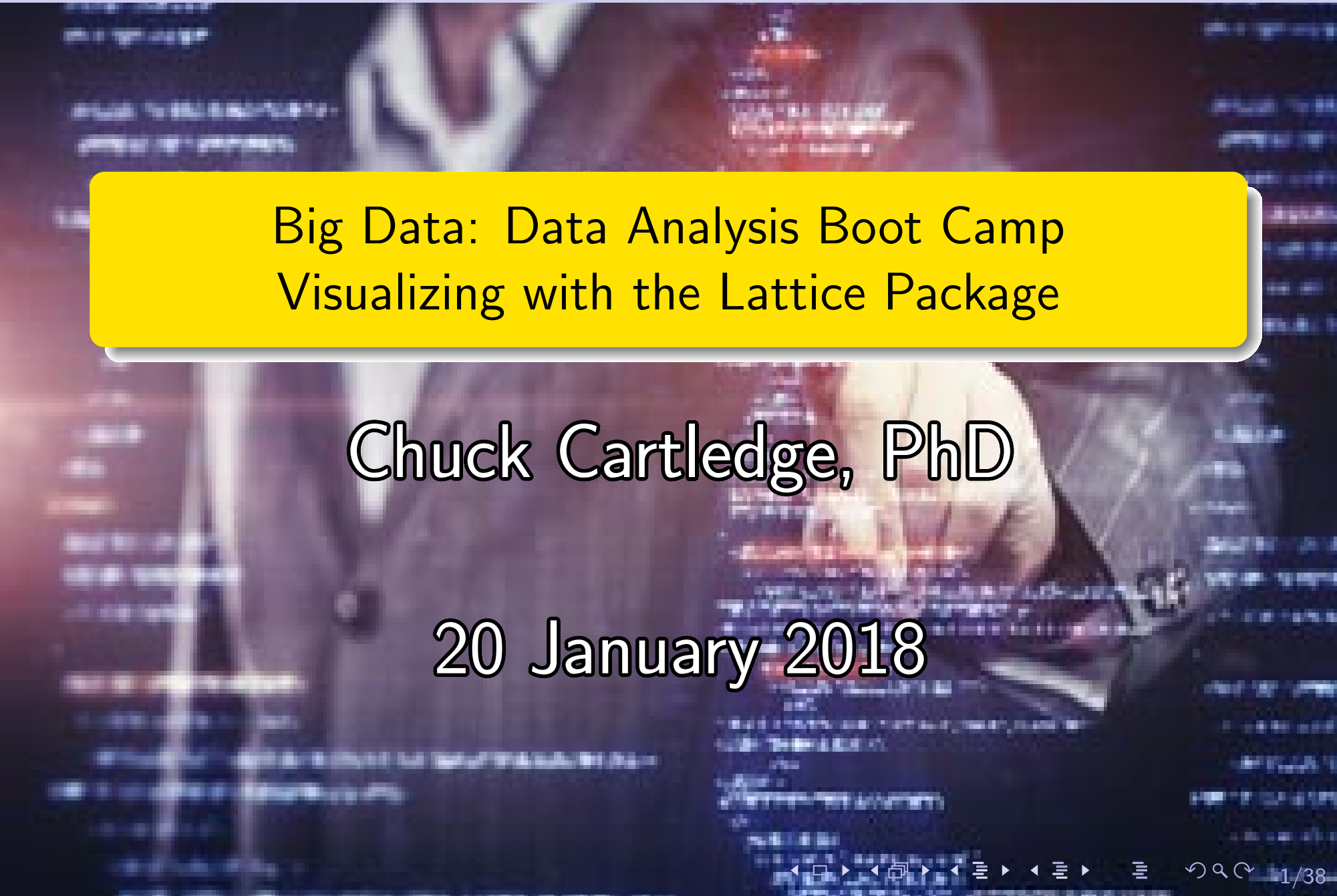


Big Data: Data Analysis Boot Camp
Visualizing with the Lattice Package

Chuck Cartledge, PhD

20 January 2018

Table of contents (1 of 1)

What are we going to cover?

We're going to talk about:

- The `lattice` package functions and capabilities.
- Choroplethrmaps (what they are and how to construct them.)

A description

"A powerful and elegant high-level data visualization system inspired by Trellis graphics, with an emphasis on multivariate data. Lattice is sufficient for typical graphics needs, and is also flexible enough to handle most nonstandard requirements."

D. Sarkar [2]

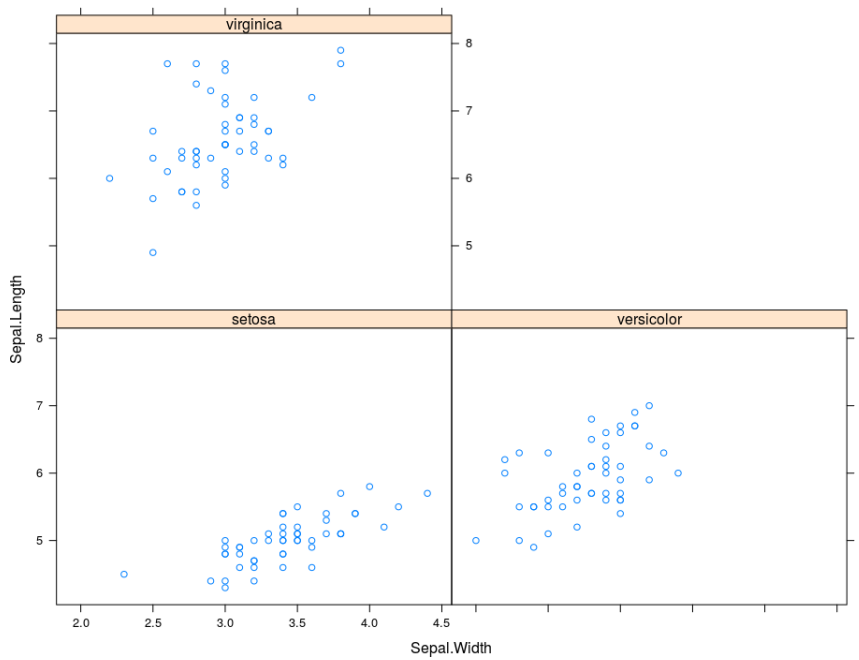
An explanation

"Trellis displays are plots which contain one or more panels, arranged in a regular grid-like structure (a trellis). Each panel graphs a subset of the data. All panels in a Trellis display contain the same type of graph but these graphs are general enough to encompass a wide variety of 2-D and 3-D displays: histogram, scatter plot, dot plot, contour plot, wireframe, 3-D point cloud and more."

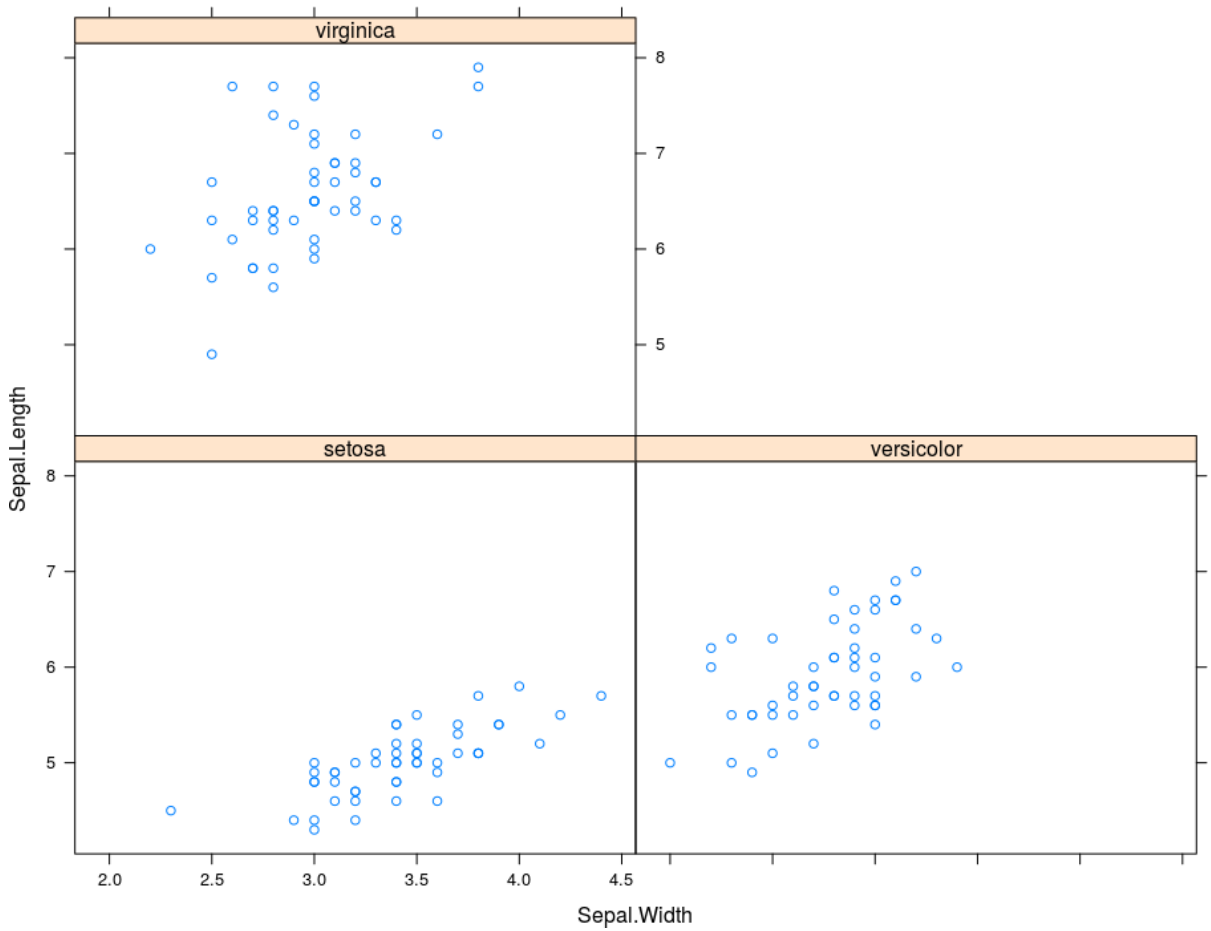
Becker, et al. [1]

A lattice iris example

```
library(lattice)
xyplot(Sepal.Length
~Sepal.Width | Species,
data = iris)
```



Same image.

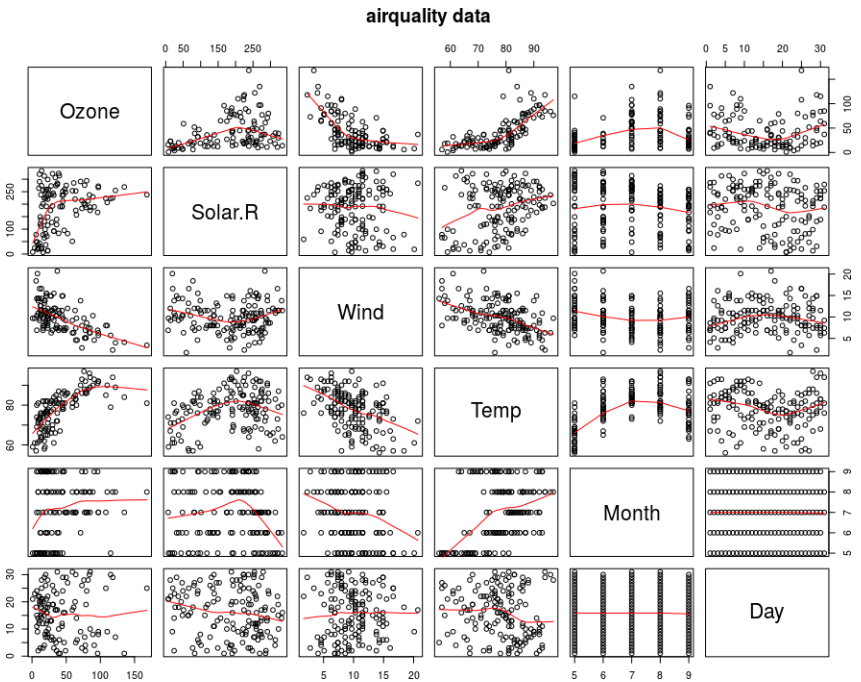

virginica
setosa
versicolor
Sepal.Length
Sepal.Width
8
7
6
5
2.0
2.5
3.0
3.5
4.0
4.5

Traditional pairs plot of air quality

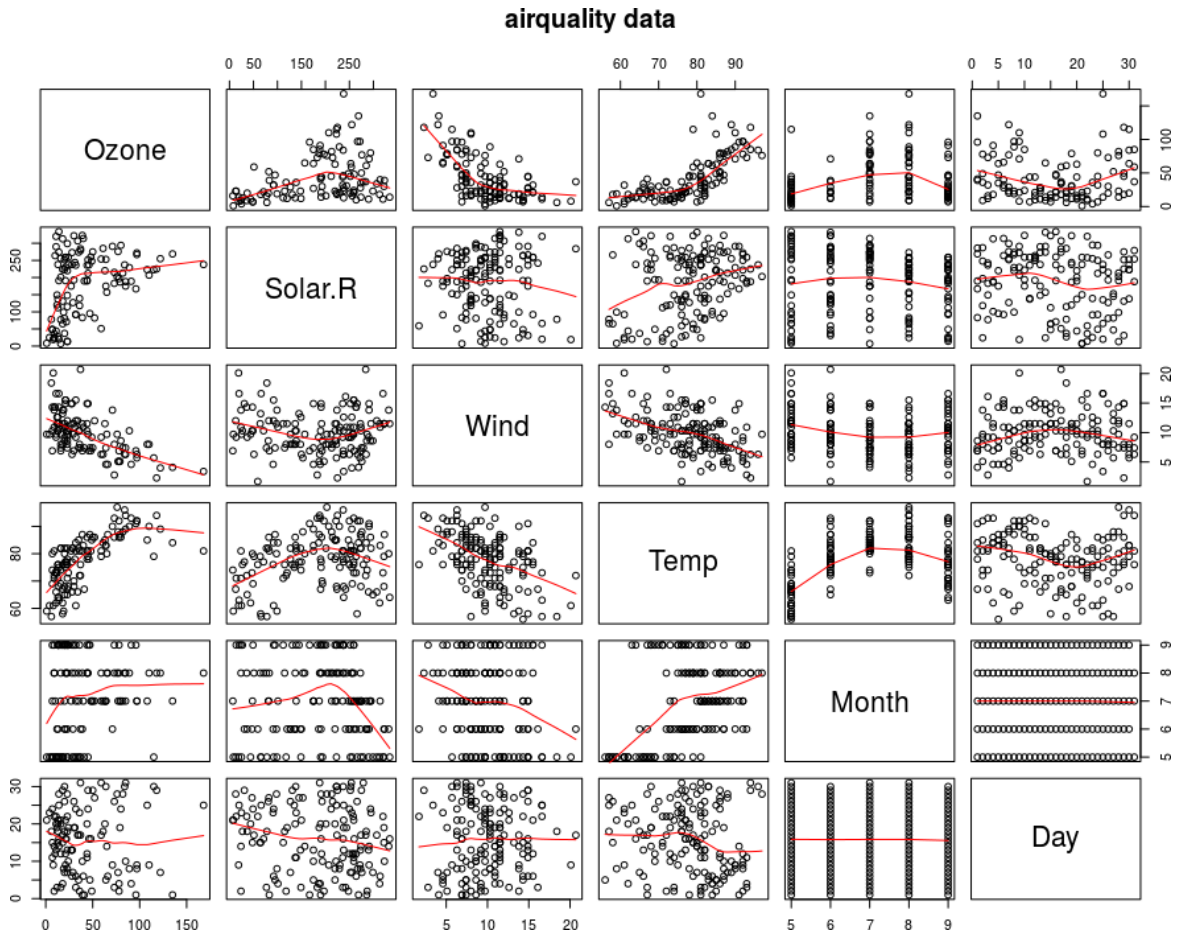
Air quality data is part of the datasets package.

```
pairs(airquality, panel =

panel.smooth, main =
``airquality data'')
```

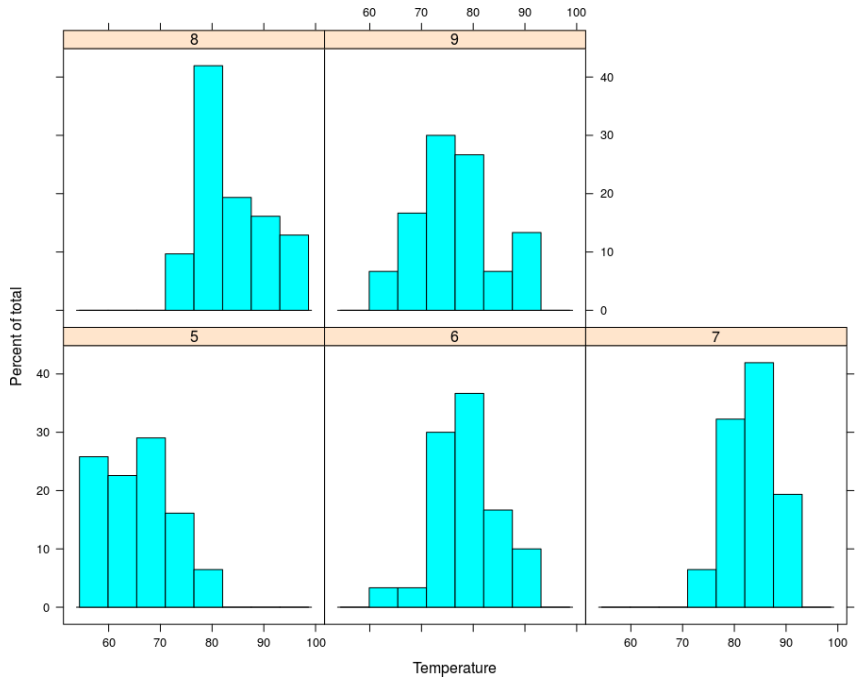


Same image.

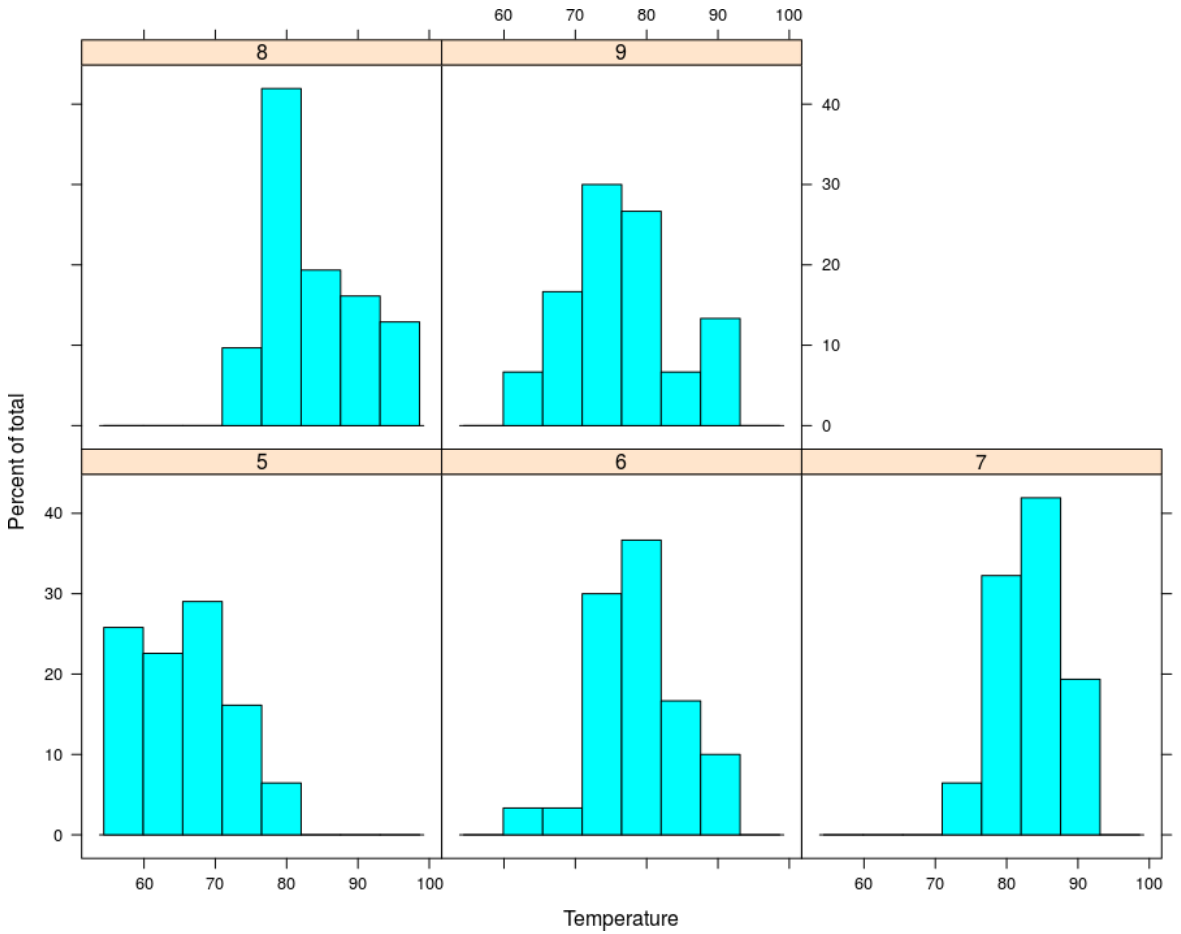


A lattice histogram of air quality

```
histogram(~Temp |
factor(Month),
data = airquality,
xlab = ``Temperature``,
ylab = ``Percent of
total``)
```



Same image.

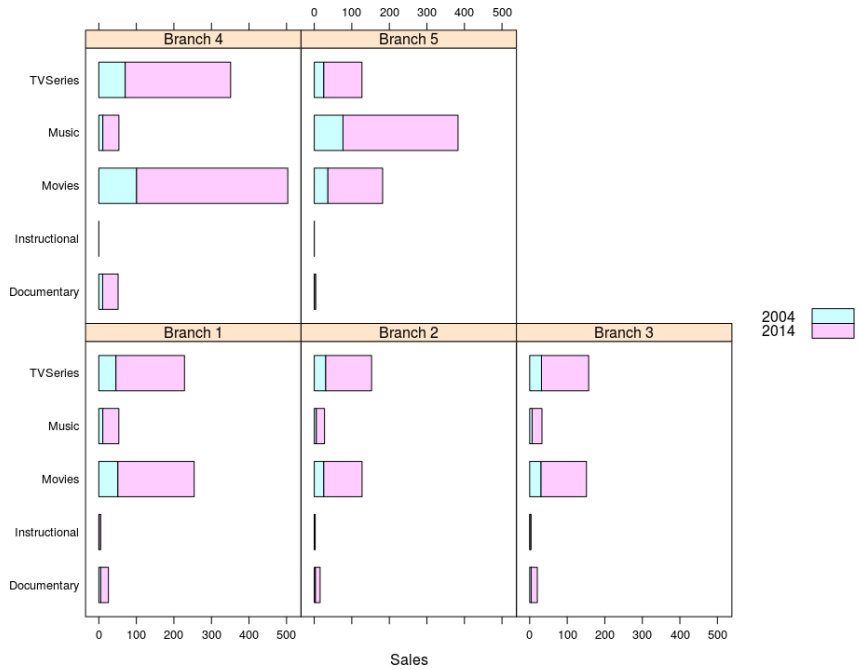

8
9
5
6
7
Percent of total
Temperature
60
70
80
90
100
0
10
20
30
40

A few words about R's formula

Symbol	Example	Meaning
+	+X	include this variable
-	-X	delete this variable
:	X:Z	include the interaction between these variables
*	X*Y	include these variables and the interactions between them
|	X|Z	conditioning: include x given z
^	$(X + Z + W)$^3	include these variables and all interactions up to three way
I	I(X*Z)	as is: include a new variable consisting of these variables multiplied
1	X - 1	intercept: delete the intercept (regress through the origin)

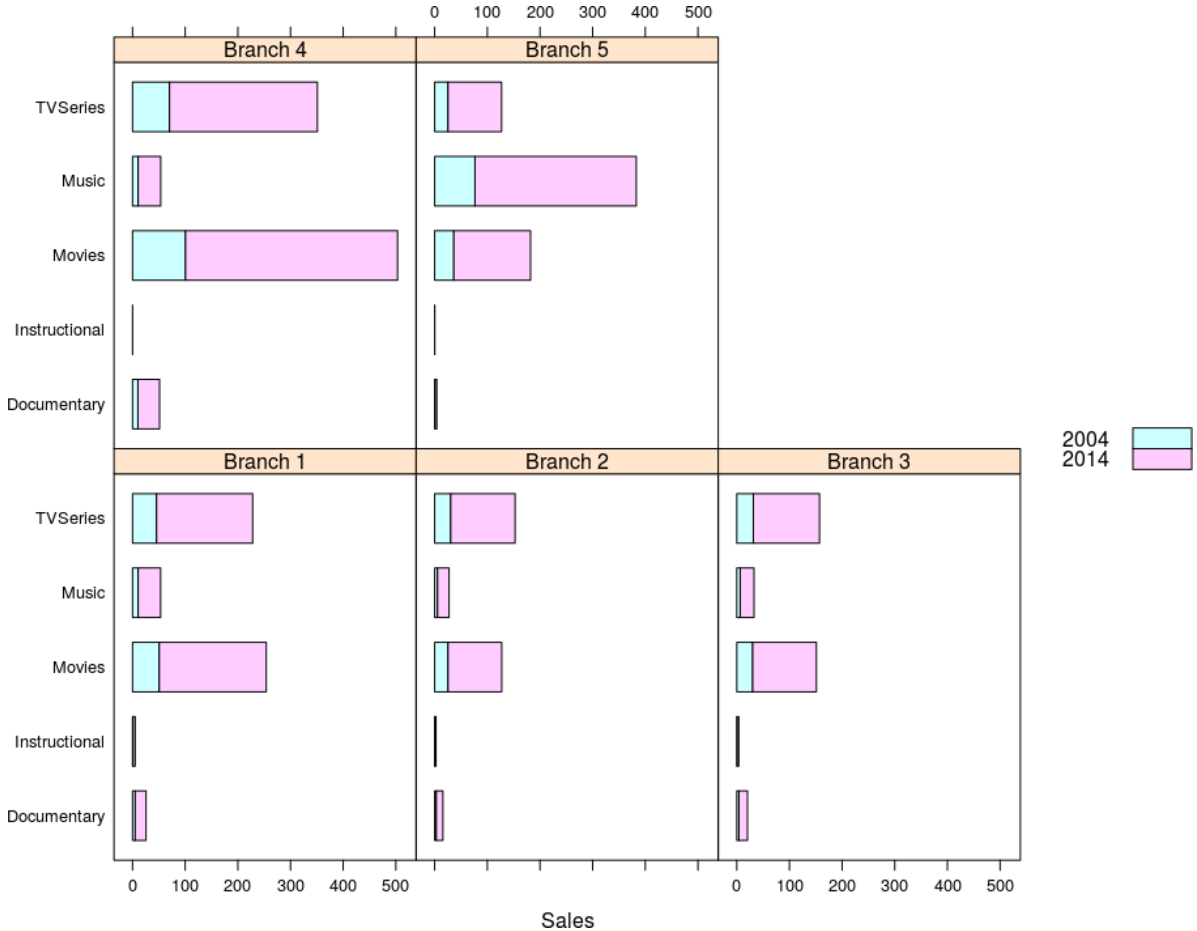
Stacked bar plots from text

(Load attached file "latticeStackedBars.R" into editor)



(Attached file.)

Same image.



(Attached file.)

Functions in package "lattice"

1. To list the functions in any package:

```
ls("package:packageName")
```

2. To see which packages are loaded:

```
search()
```

3. To list the functions in "lattice"

```
ls("package:lattice")
```

There are 149 functions in the `lattice` package.

Looking at the data

The text devotes a fair number of pages to analyzing and plotting cancer related data from the `latticeExtra` package. The purpose is to demonstrate plotting and analytical techniques.

There is a problem with the data.

Not all US states and territories are represented.

- Data is in the data frame: `USCancerRates`
- `USCancerRates` column names are:
  rate.male, LCL95.male, UCL95.male, rate.female,
  LCL95.female, UCL95.female, state, county
- Use
  `unique(sort(unlist(as.character(USCancerRates$state))))`
  to find the unique state names

Hawaii and the District of Columbia are not in the data frame.
We'll explore cancer rates with better data.

Which data and where to get it?

Various sources of data:

1. Cancer rates for men and women (use 65 and under): `https://statecancerprofiles.cancer.gov/`
2. Insurance rates: `https://www2.census.gov/programs-surveys/sahie/datasets/time-series/estimates-acs/sahie-2013-csv.zip`
3. Connect state and county names to Federal Information Processing Standards (FIPS) `https://www2.census.gov/geo/docs/reference/codes/files/national_county.txt`
4. Get state and county shapefiles based FIPS `http://www2.census.gov/geo/docs/maps-data/data/gazetteer/Gaz_counties_national.zip`

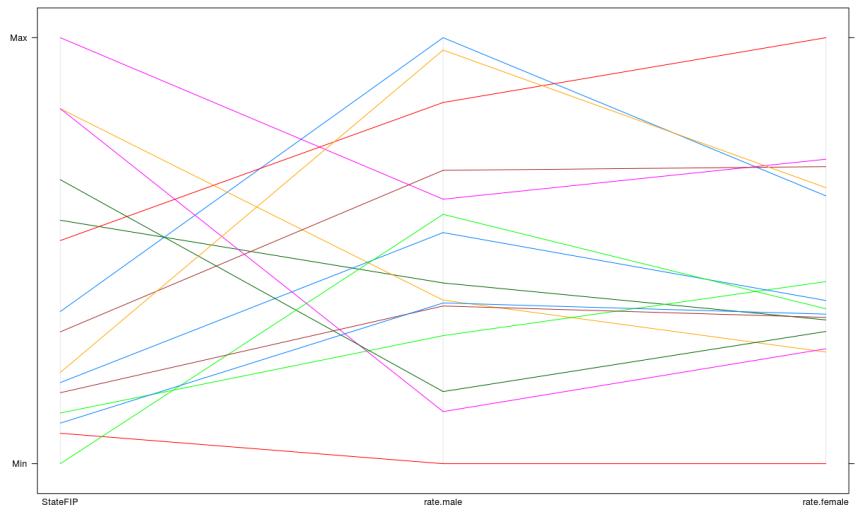
What do we do after we know where the data is?

1. Download the various data files (some will require bringing up a browser, and selecting a data file manually).
2. Build a data frame based on FIPS ID
   1. Combine cancer rates
   2. Expand with insurance rates
   3. Expand with positional (latitude and longitude) data
3. "Cleanse" the data (numerical data wrangling)
4. Analyze the data

All of these steps are in the attached file: ``cancerData.R''

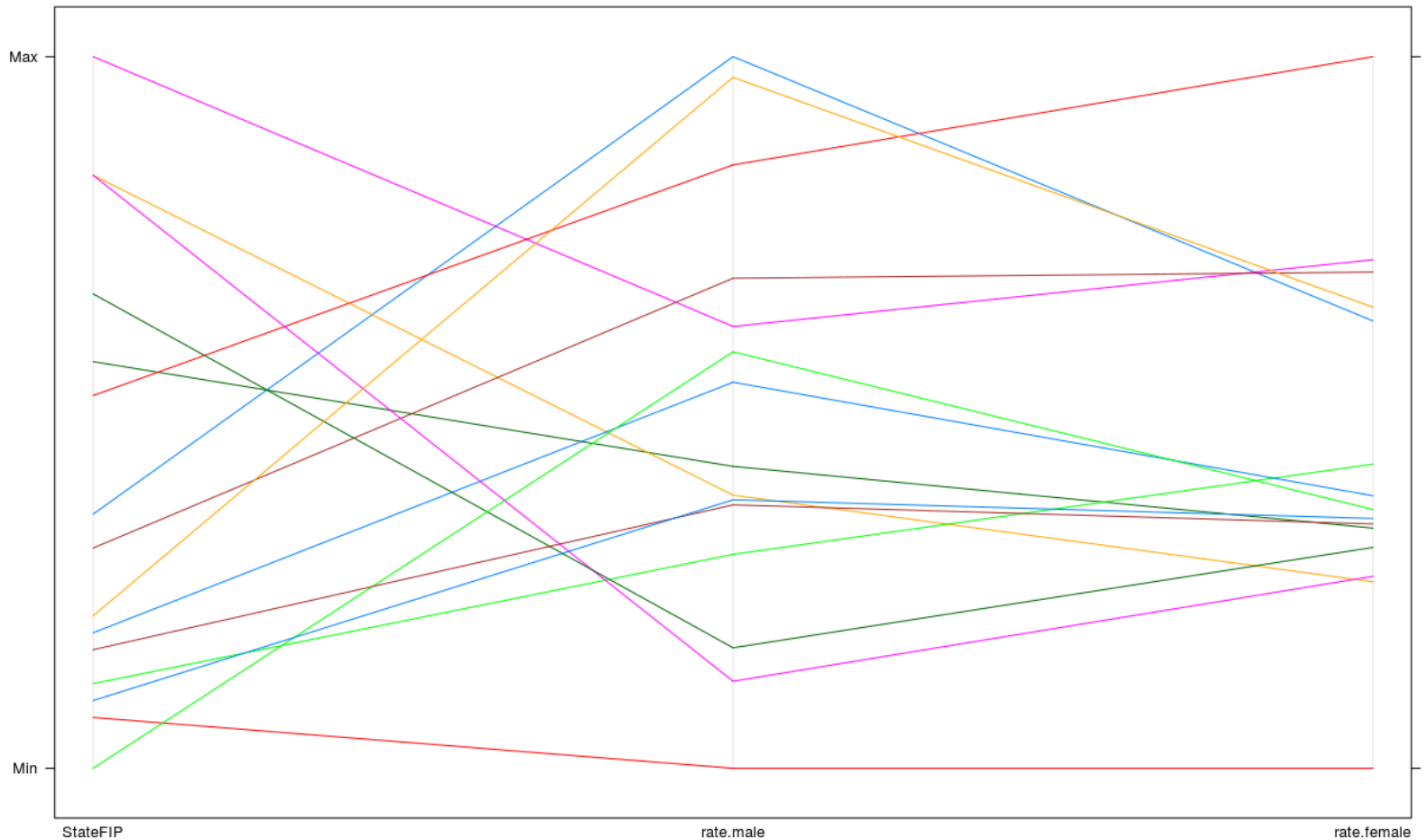
Parallel plot of selected counties and gender based cancer rates.

```
set.seed(987)
subSample <- data[sample(1:nrow(data),
size=15), c("StateFIP", "rate.male",
"rate.female")]
parallelplot(subSample, horizontal.axis=FALSE,
groups=data$StateFIP)
```



Important thing is that cancer rates differ based on gender.

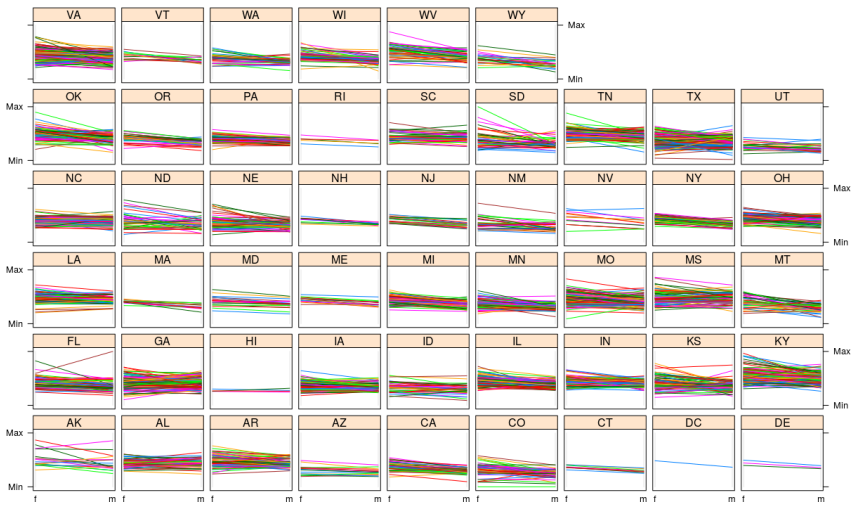
Same image.



Important thing is that cancer rates differ based on gender.

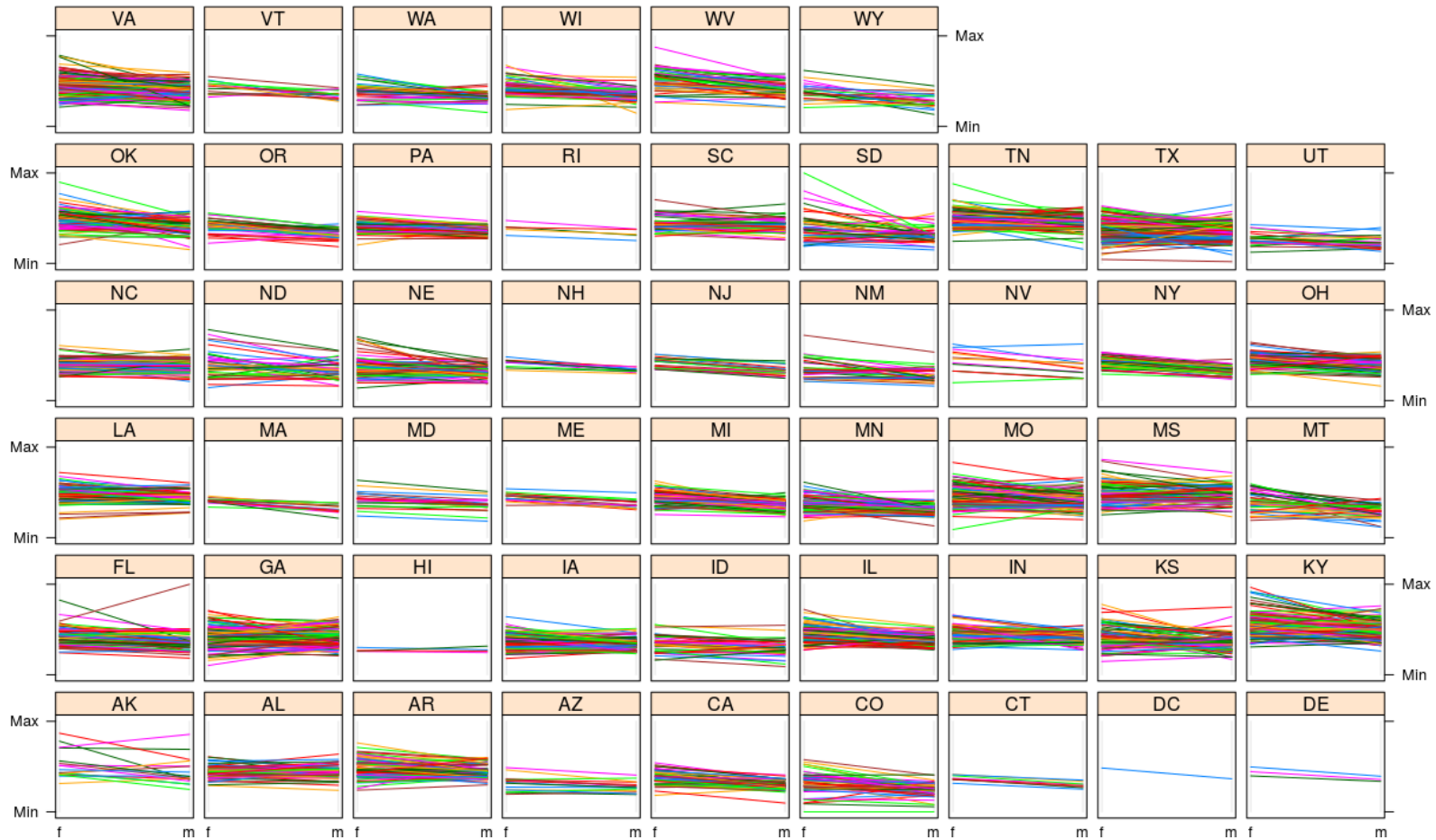
Male and female rates across all states and the DC.

```
parallelplot(~data[,c(‘‘rate.female’’,‘‘rate.male’’)]
| data$State,
horizontal.axis=FALSE, data=data,
varnames=c(‘‘f’’,‘‘m’’))
```



In general, female rate is higher than male.

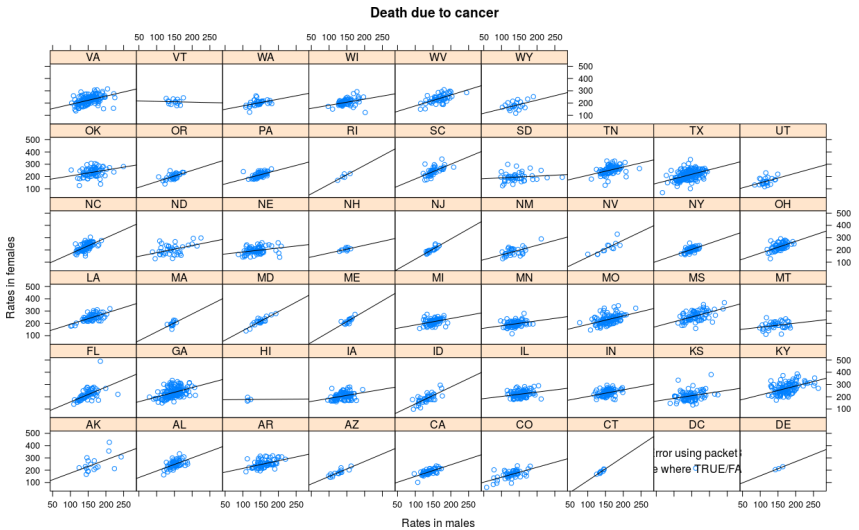
Same image.



In general, female rate is higher than male.

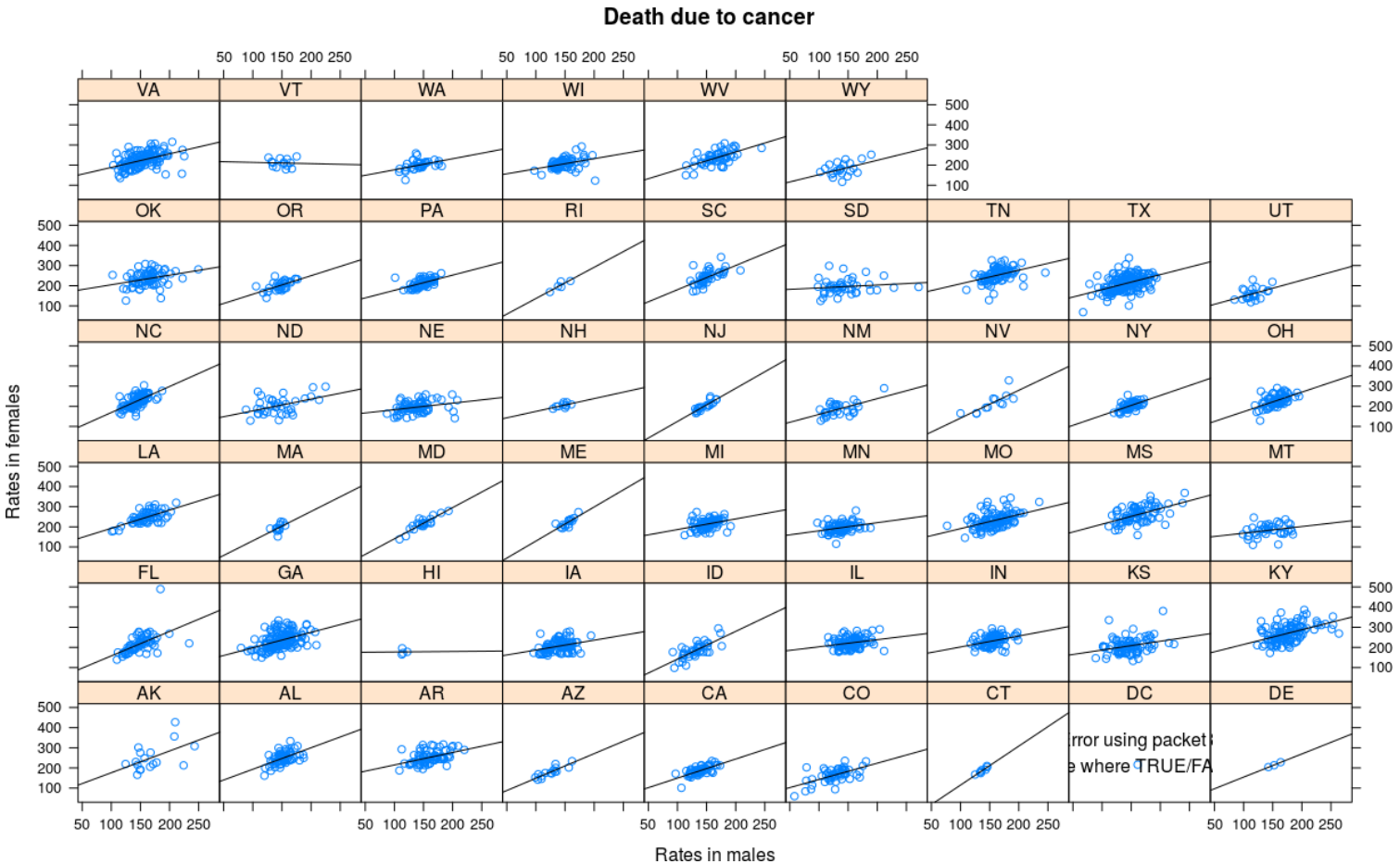
Fitted male and female rates across all states and the DC.

```
xyplot(rate.male ~rate.female | State,
data=data,
main = ‘‘Death due to cancer’’,
xlab=‘‘Rates in males’’,
ylab=‘‘Rates in females’’,
panel=function(x,y,...){
panel.xyplot(x,y)
panel.abline(lm(y~x))
}
)
```



Limited data points for DC.

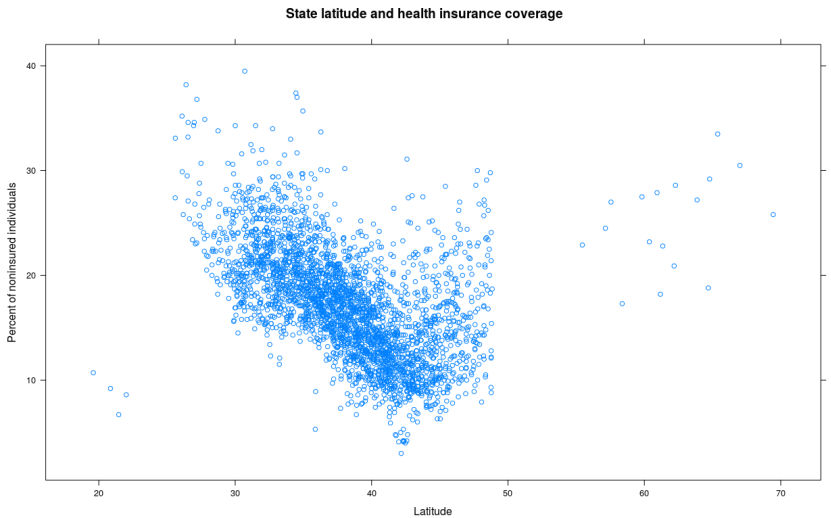
Same image.



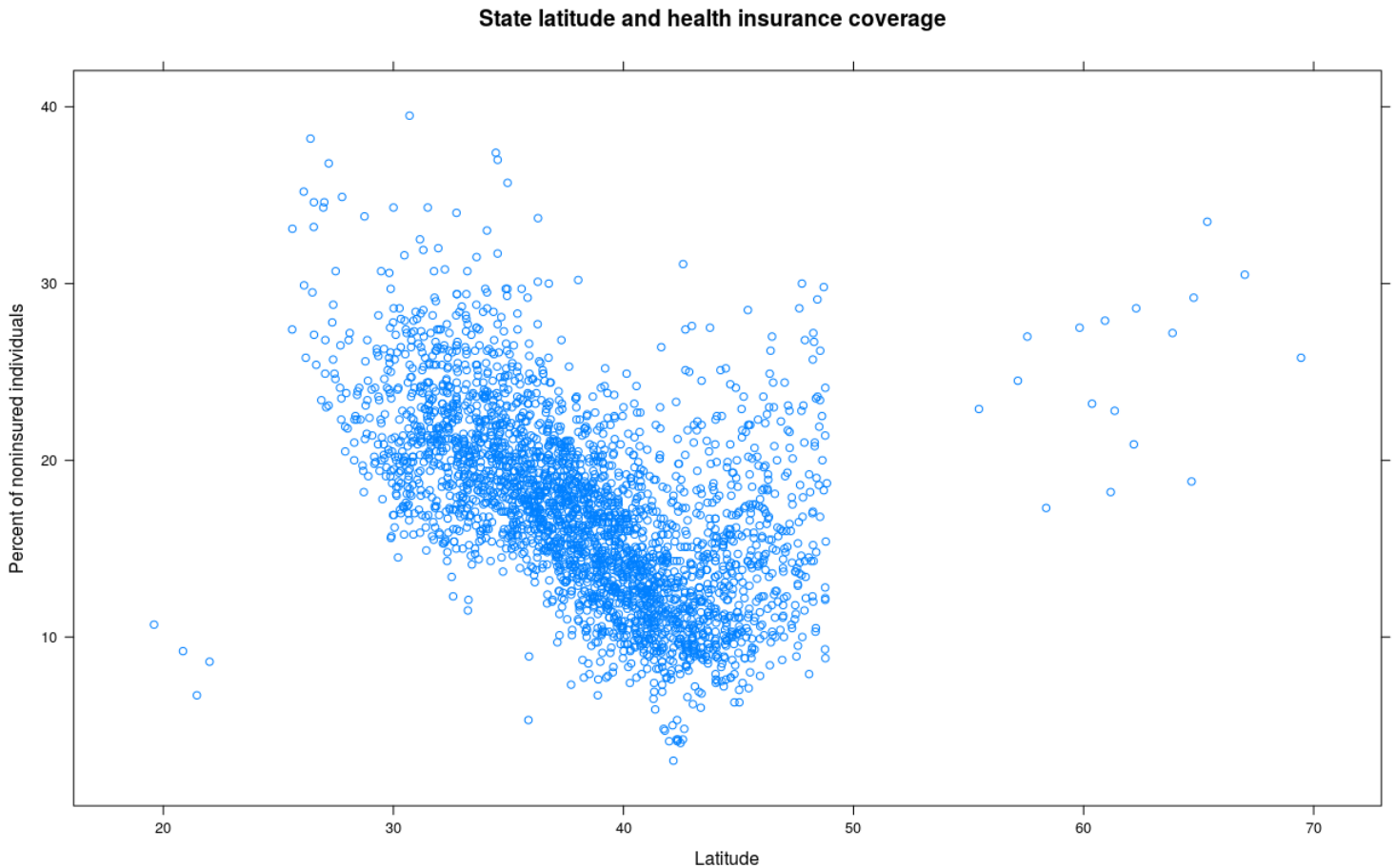
Limited data points for DC.

Insurance based on latitude (north or south)

```
xyplot(data$NoInsur ~data$CentLat,
main = ‘‘State latitude and health insurance
coverage’’,
xlab = ‘‘Latitude’’,
ylab = ‘‘Percent of noninsured individuals’’)
```

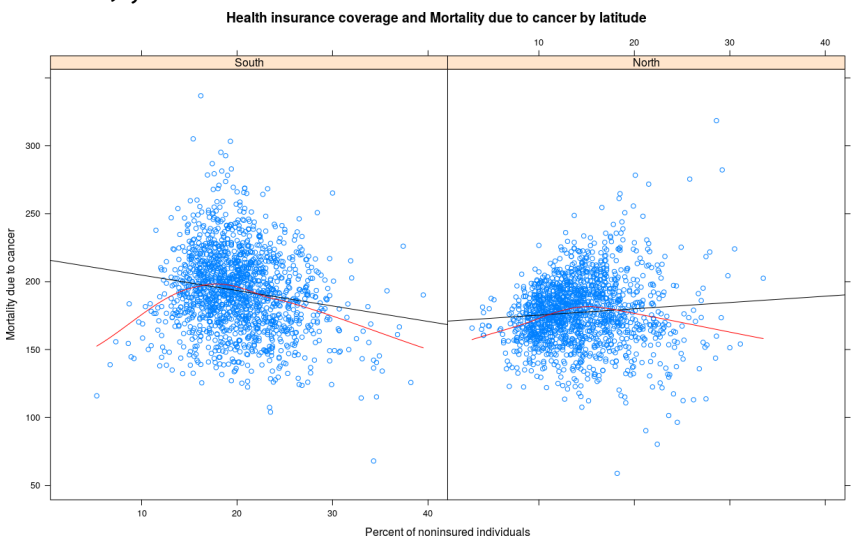


Same image.

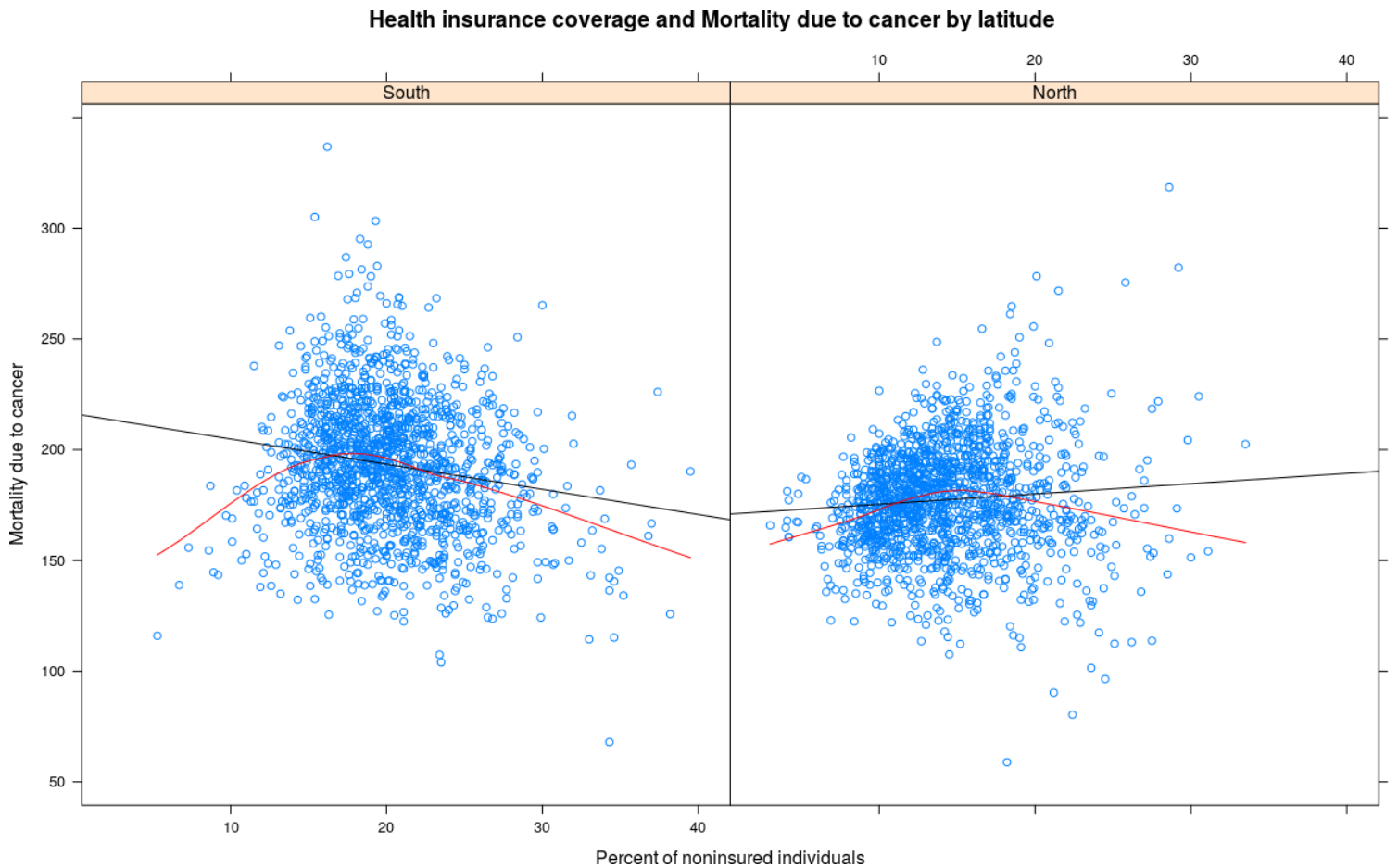

State latitude and health insurance coverage
Percent of noninsured individuals
40
30
20
10
20
30
40
50
60
70
Latitude

Cancer rate by gender and northern or southern

```
xyplot((rate.female + rate.male)/2 ~NoInsur |
ifelse(data$CentLat>median(data$CentLat),''North'',''South''),
data = data,
index.cond=list(c(2,1)),
panel = function(x, y, ...)  {
panel.xyplot(x,y)
panel.abline(lm(y~x))
panel.loess(x,y, col = ''Red'')
} ,
xlab = ''Percent of noninsured individuals'',
ylab = ''Mortality due to cancer'', main =
''Health insurance coverage and Mortality due to
cancer by latitude''
)
```

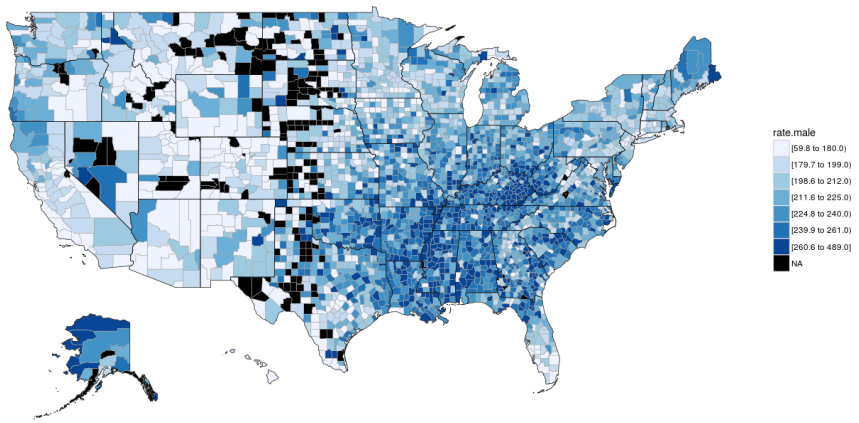


Same image.


Health insurance coverage and Mortality due to cancer by latitude
South
North
Mortality due to cancer
Percent of noninsured individuals
50
100
150
200
250
300
10
20
30
40

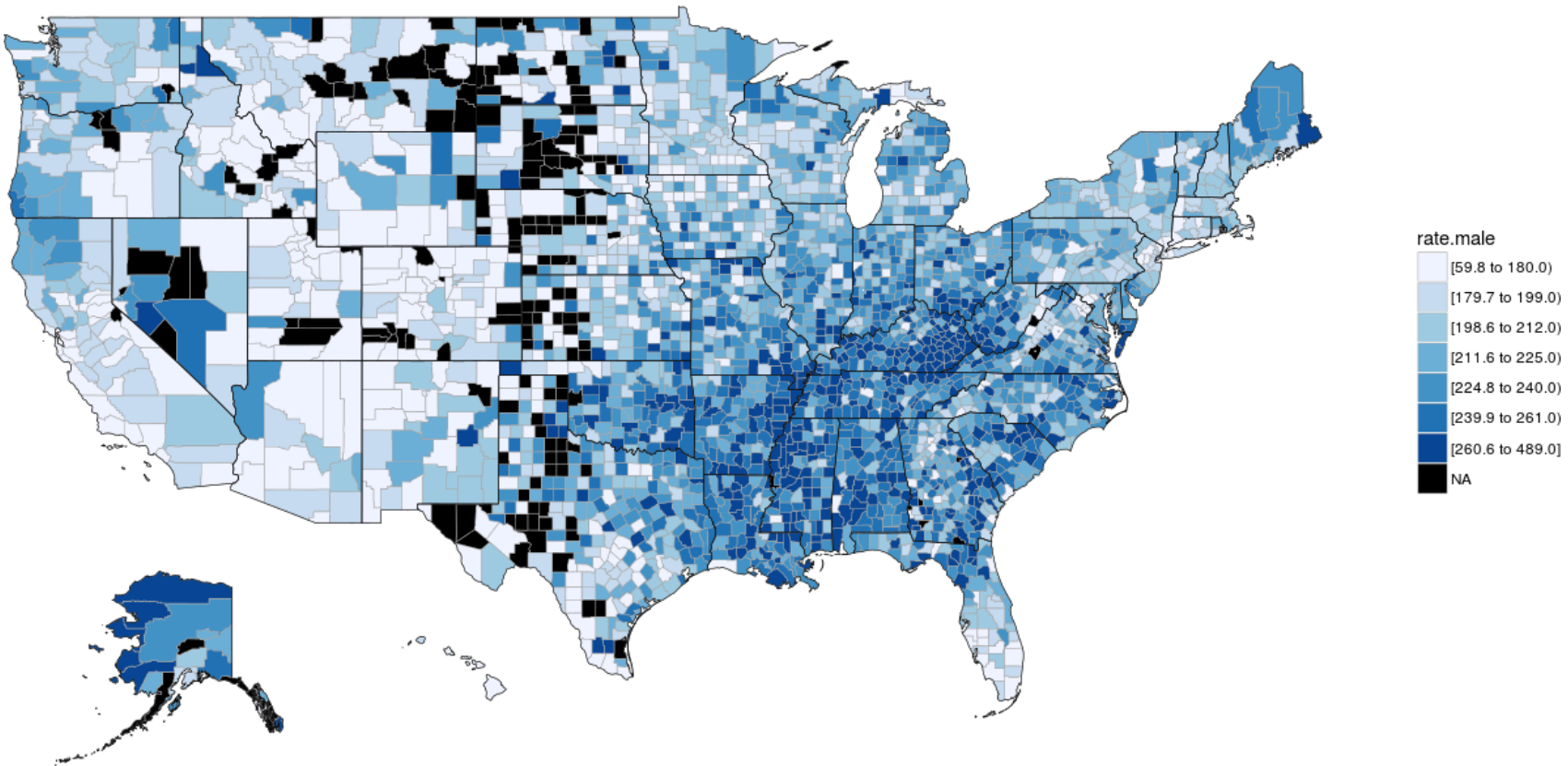
Map of county level cancer rate for males.

```
plotColumn(data,
``rate.male'')
```



(See attached file.)

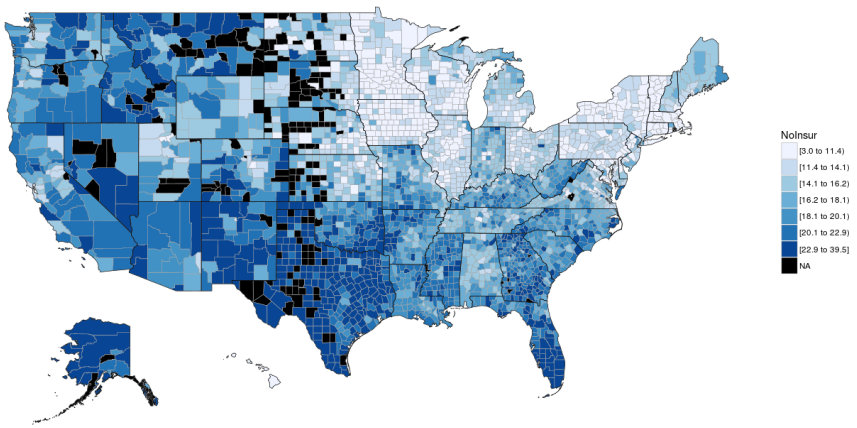
Same image.



(See attached file.)

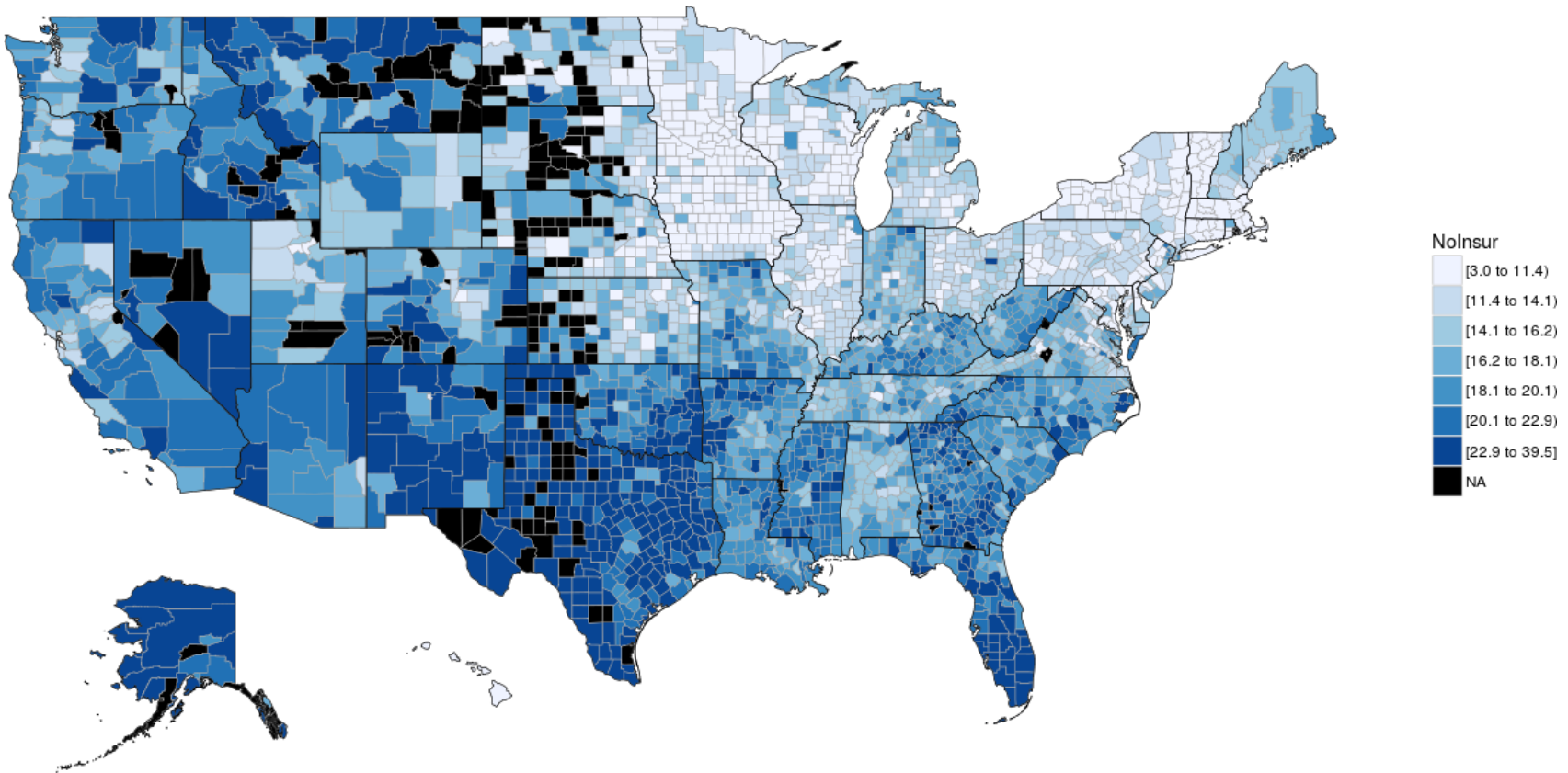
Map of county level insurance rates.

```
plotColumn(data,
''NoInsur'')
```



(See attached file.)

Same image.



(See attached file.)

Some simple exercises to get familiar with data visualization

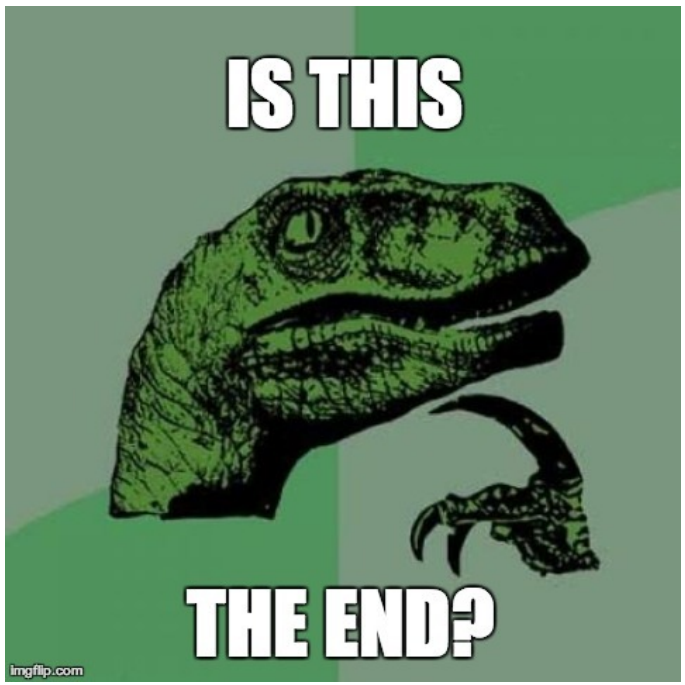
1. Create a choroplethrmap showing the cancer rate for women
2. Create a plot of insurance rate based on longitude
3. Most states have a positive "Death due to cancer" of women to men. Ideas why?

Q & A time.

Q: How does a hacker fix a function which doesn't work for all of the elements in its domain?
A: He changes the domain.

What have we covered?

- Reviewed some of the functions in the `lattice` package
- Reviewed some of the data in the `latticeExtra` package
- Created our own cancer rate data frame
- Analyzed the data using tools from the `lattice` package
- Analyzed the data using tools from the `choroplethr` and `choroplethrmaps` packages



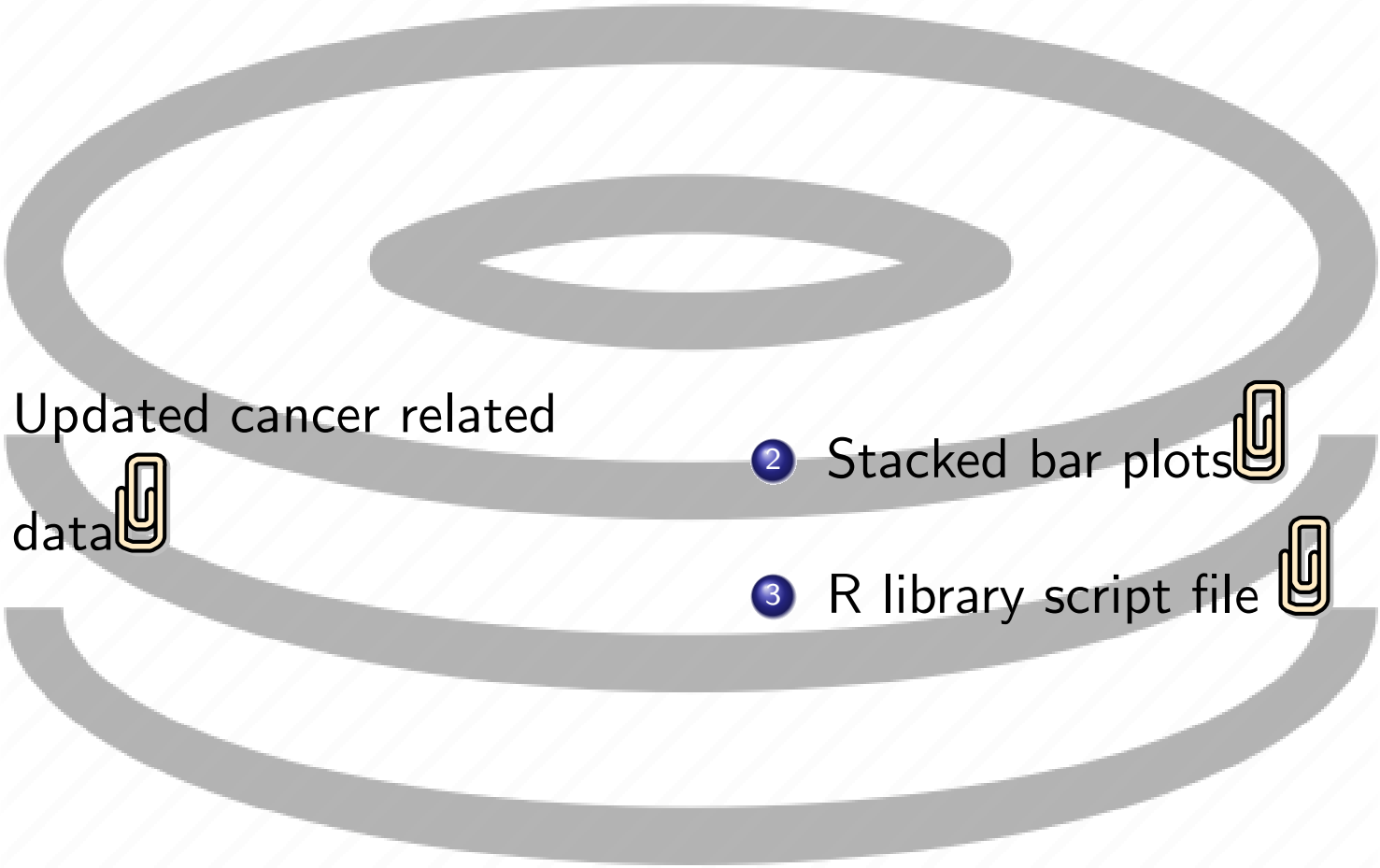
Next: LPAR Chapter 4, cluster analysis

References (1 of 1)

[1] Richard A Becker, William S Cleveland, Ming-Jen Shyu, Stephen P Kaluzny, et al., A Tour of Trellis Graphics, Murray Hill, NJ: AT & T Bell Laboratories **44** (1996).

[2] Deepayan Sarkar, Lattice: Multivariate Data Visualization with R, Springer, New York, 2008, ISBN 978-0-387-75968-5.

Files of interest



1. Updated cancer related data
2. Stacked bar plots
3. R library script file