HUDM6122 Homework_02

Chenguang Pan

2023-02-06

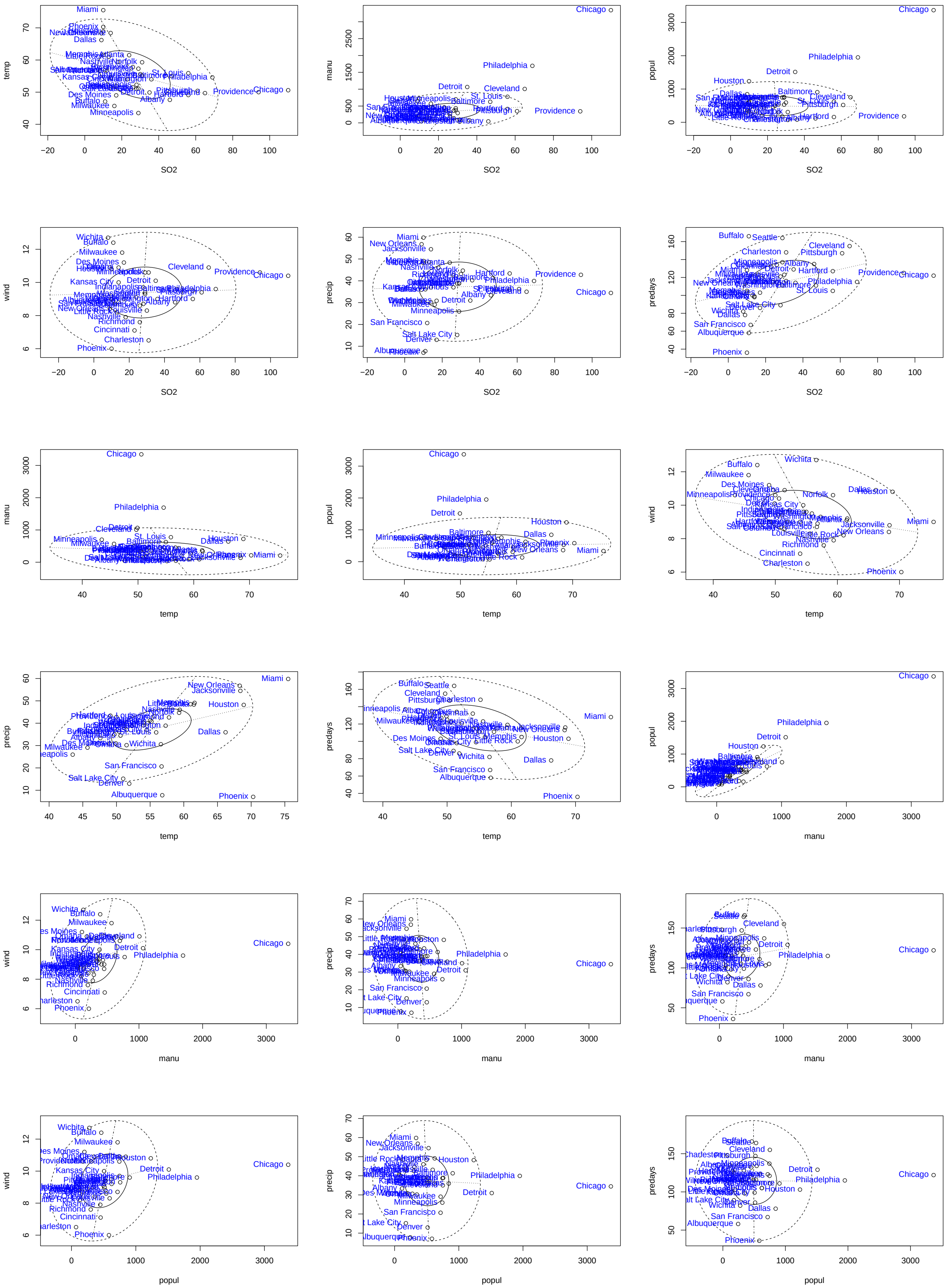
0.1 Ex. 2.1

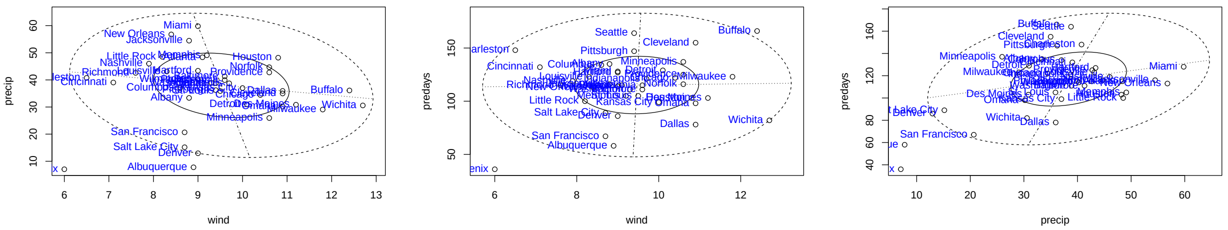
Use the bivariate boxplot on the scatterplot of each pair of variables in the air pollution data to identify any outliers. Calculate the correlation between each pair of variables using all the data and the data with any identified outliers removed. Comment on the results.

MY SOLUTION:

Several techniques worth to be noted: - use a for-loop within a for-loop to map all pairs - use `text()` to give each point a name. - the bivariate boxplot function `bvbox` is inclueded in the package `MVA`

```
> # import the data
> library(HSAUR2)
> library(MVA)
> attach(USairpollution)
> head(USairpollution)
            SO2 temp manu popul wind precip predays
Albany       46 47.6   44   116  8.8  33.36     135
Albuquerque  11 56.8   46   244  8.9   7.77      58
Atlanta      24 61.5  368   497  9.1  48.34     115
Baltimore    47 55.0  625   905  9.6  41.31     111
Buffalo      11 47.1  391   463 12.4  36.11     166
Charleston   31 55.2   35    71  6.5  40.75     148
> dim(USairpollution)
[1] 41  7
> # draw the bivariate boxplot of each pair of variables with a for-loop
> for (i in 1:7) {
+   for (j in i:7) {
+     if (i != j) {
+       var_pair <- USairpollution[, c(i, j)]
+       bvbox(var_pair,
+             xlab = names(USairpollution)[i],
+             ylab = names(USairpollution)[j])
+       text(USairpollution[,i],USairpollution[,j],
+           row.names(USairpollution), # to add the point name
+           pos=2,col = "blue")
+     }
+   }
+ }
```


SO2
temp
manu
popul
wind
precip
predays
Chicago
Philadelphia
Detroit
Cleveland
Providence
Houston
Miami
Phoenix
Albuquerque
Denver
Salt Lake City
San Francisco
Wichita
Buffalo
Seattle
Milwaukee
Minneapolis
Des Moines
New Orleans
Jacksonville
Dallas
Charleston
Cincinnati
Richmond
Nashville
Pittsburgh
St. Louis
Hartford
Baltimore
Albany
Kansas City
Memphis
Atlanta
Norfolk
Little Rock
Louisville
Omaha
Indianapolis
Washington
Wilmington
Columbus
Wichita
-20
0
20
40
60
80
100
1000
2000
3000



From the graphs, we can easily find that the outliers among the observations are “Chicago”, “Detroit”, “Cleveland”,“Philadelphia”, “Miami”, “Phoenix”,“Albuquerque”, “Providence”. I run the correlation matrix on all observations first.

```
> # create the correaltion matrix on all observations
> round(cor(USairpollution),2)
          SO2  temp  manu popul  wind precip predays
SO2      1.00 -0.43  0.64  0.49  0.09   0.05    0.37
temp    -0.43  1.00 -0.19 -0.06 -0.35   0.39   -0.43
manu     0.64 -0.19  1.00  0.96  0.24  -0.03    0.13
popul    0.49 -0.06  0.96  1.00  0.21  -0.03    0.04
wind     0.09 -0.35  0.24  0.21  1.00  -0.01    0.16
precip   0.05  0.39 -0.03 -0.03 -0.01   1.00    0.50
predays  0.37 -0.43  0.13  0.04  0.16   0.50    1.00
> # remove all identified outliers
> drop_city <- match(c("Chicago", "Detroit","Cleveland",
+                      "Philadelphia", "Miami","Phoenix",
+                      "Albuquerque", "Providence"), rownames(USairpollution))
>
> round(cor(USairpollution[-drop_city,]),2)
          SO2  temp  manu popul  wind precip predays
SO2      1.00 -0.40  0.10 -0.16 -0.26  -0.03    0.39
temp    -0.40  1.00 -0.02  0.25 -0.21   0.64   -0.41
manu     0.10 -0.02  1.00  0.82  0.26  -0.15   -0.06
popul   -0.16  0.25  0.82  1.00  0.29   0.03   -0.13
wind    -0.26 -0.21  0.26  0.29  1.00  -0.26   -0.13
precip  -0.03  0.64 -0.15  0.03 -0.26   1.00    0.28
predays  0.39 -0.41 -0.06 -0.13 -0.13   0.28    1.00
```

After dropping all the identified outliers, some of the correlation coefficients has changed to the opposite direction, like from positive to negative, others shrink or increase. It is reasonable since some outliers are with high leverage.

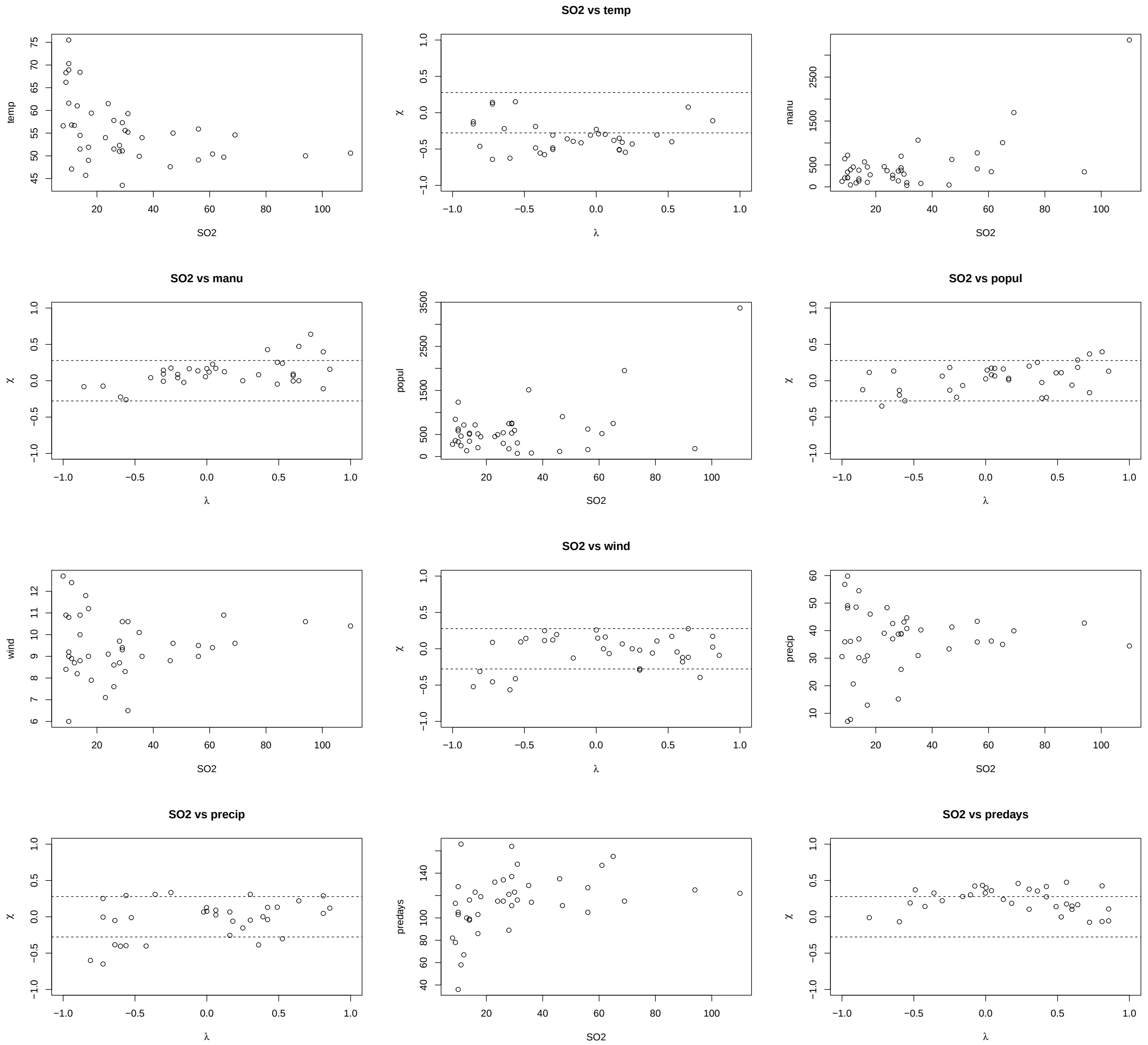
0.2 Ex. 2.2

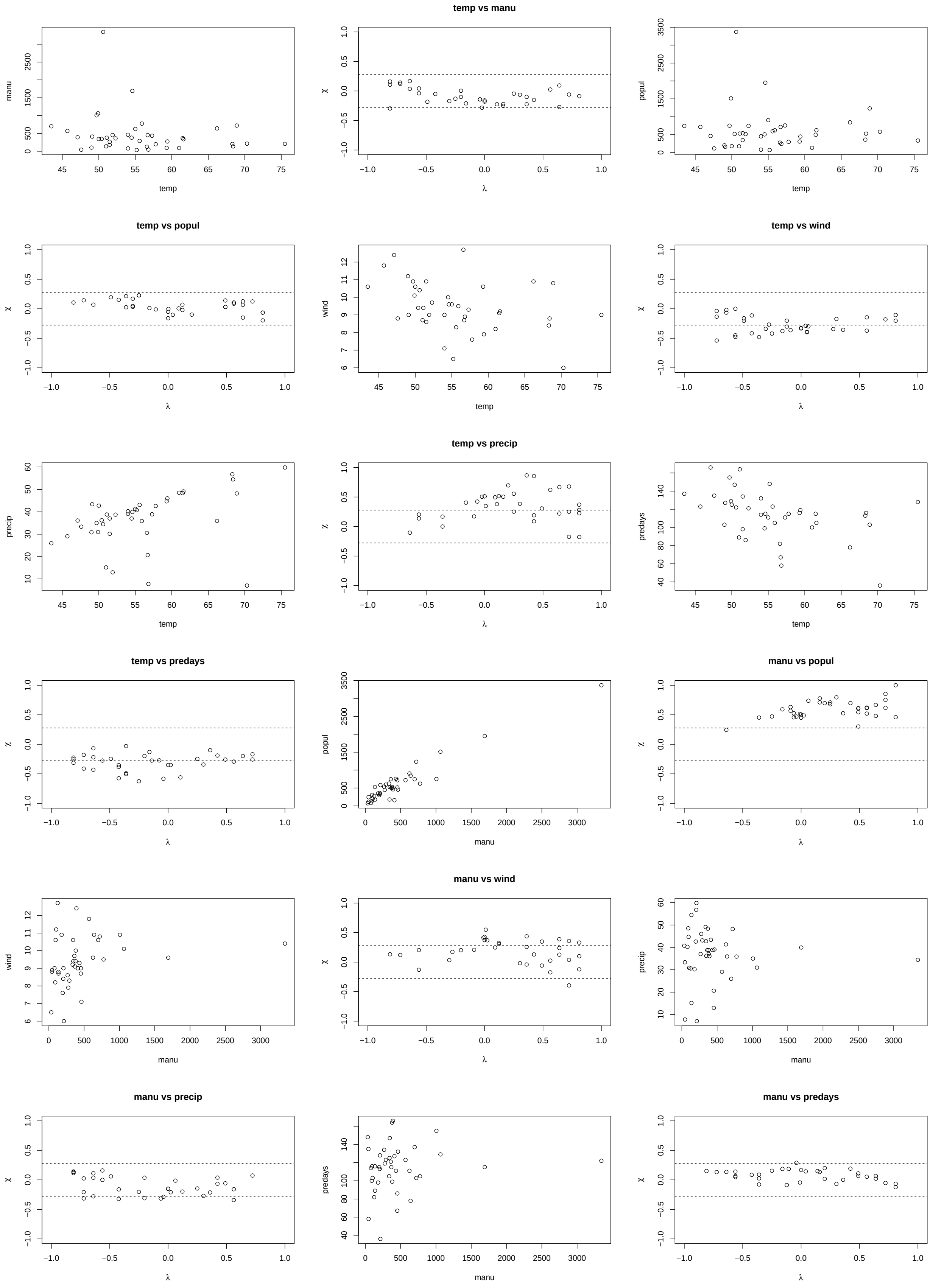
Compare the chi-plots with the corresponding scatterplots for each pair of variables in the air pollution data. Do you think that there is any advantage in the former?

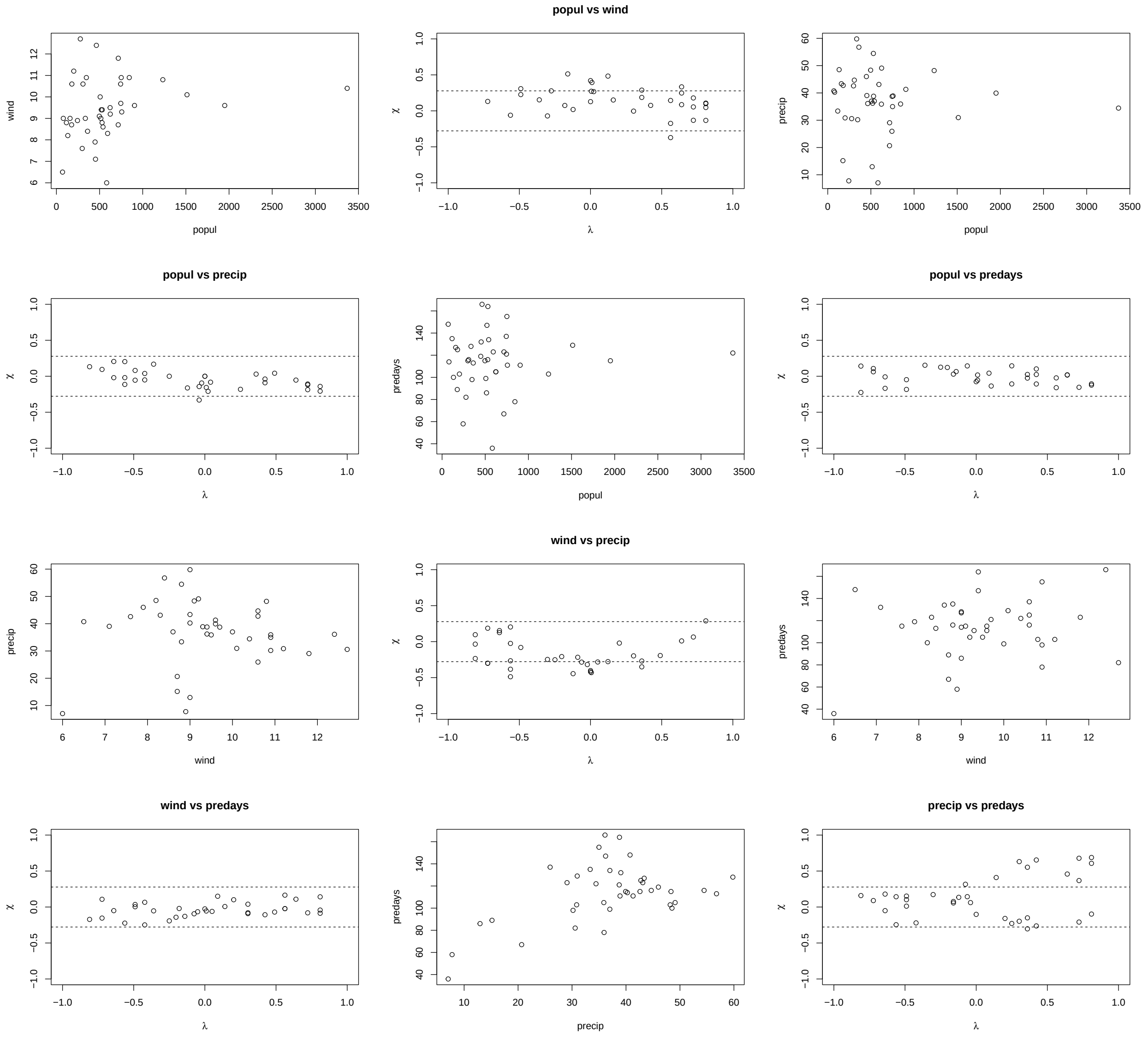
MY SOLUTION:
Several details should be noted.
- For drawing many graphs in `Rmd` file with `knit`, it always reports error or `no such file or directory`. One can clear all the cache in R and cache file and Tex file in the file folder, and do not use the layout function `par()`.
- The `chiplot` function is included in the package `MVA`. If two variables are independent, these value are

asymptotically normal with mean zero; the xi values should show a non-systematic random fluctuation around zero.

```
> for (i in 1:7) {
+   for (j in i:7) {
+     if (i != j) {
+       plot(USairpollution[,i],USairpollution[,j],
+            xlab = names(USairpollution)[i], ylab=names(USairpollution)[j])
+       chiplot(USairpollution[,i],USairpollution[,j],
+               main = paste(names(USairpollution)[i],
+                            "vs",
+                            names(USairpollution)[j]))
+     }
+   }
+ }
```




temp vs manu
manu
temp
χ
λ
popul
temp vs popul
wind
temp vs wind
temp vs precip
precip
predays
temp vs predays
manu vs popul
manu vs wind
manu vs precip
manu vs predays
-1.0
-0.5
0.0
0.5
1.0
45
50
55
60
65
70
75
0
500
1000
1500
2000
2500
3000
3500



From the results, one can easily find that the scatter plots are sometimes difficult to identify the independence between two variables. But, comparatively the `chiplot` presents more straightforward way to tell this attribute. For example, it is hard to find the relation from the scattorplot for `manu` and `predays`, but the chiplot clearly demonstrates that these two varriables are independent.

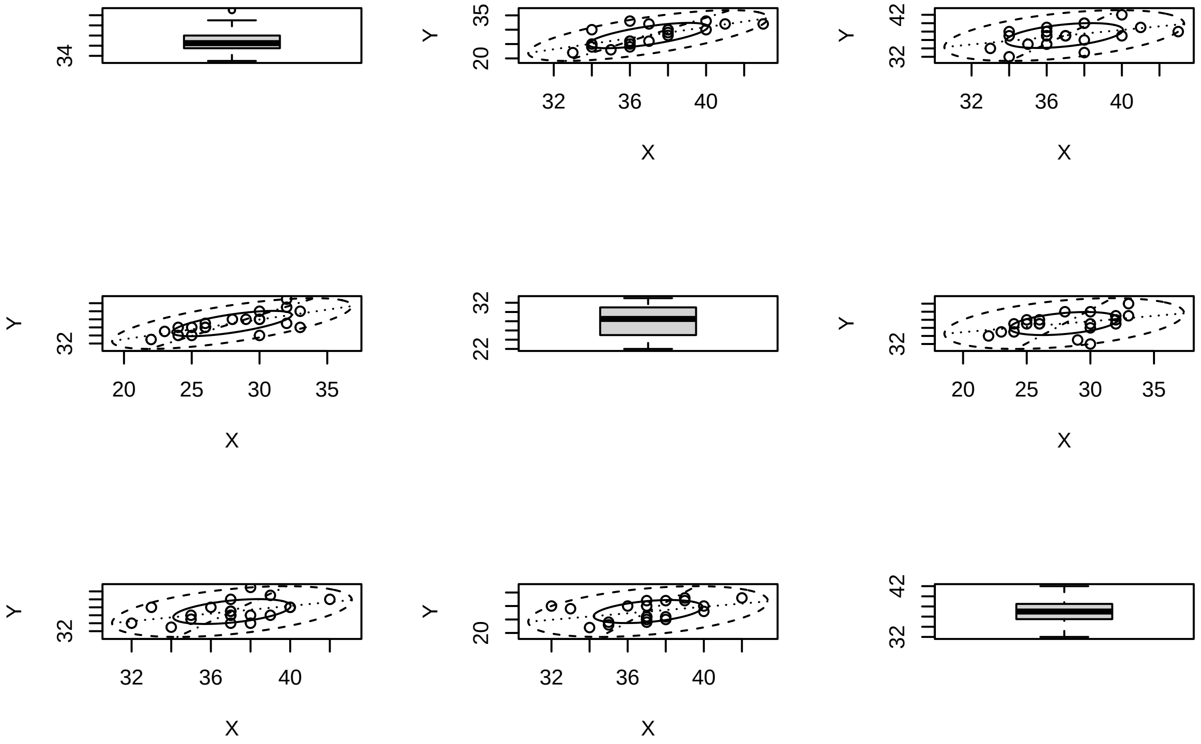
0.3 Ex. 2.3

Construct a scatterplot matrix of the body measurements data that has the appropriate boxplot on the diagonal panels and bivariate boxplots on the other panels. Compare the plot with Figure 2.17, and say which diagram you find more informative about the data.

MY SOLUTION:
The body measure dataset is not included in any packages. We need to create the dataset by ourselves. First, based on the codes offered by Prof. Motta, Giovanni, I created measure data

```
> # create the  body measure data. Codes offered by Prof.Motta.
> measure <-
```

```
+   structure(list(V1 = 1:20, V2 = c(34L, 37L, 38L, 36L, 38L, 43L,
+                  40L, 38L, 40L, 41L, 36L, 36L, 34L, 33L, 36L, 37L, 34L, 36L, 38L,
+                  35L), V3 = c(30L, 32L, 30L, 33L, 29L, 32L, 33L, 30L, 30L, 32L,
+                  24L, 25L, 24L, 22L, 26L, 26L, 25L, 26L, 28L, 23L), V4 = c(32L,
+                  37L, 36L, 39L, 33L, 38L, 42L, 40L, 37L, 39L, 35L, 37L, 37L, 34L,
+                  38L, 37L, 38L, 37L, 40L, 35L)), .Names = c("V1", "V2", "V3",
+                  "V4"), class = "data.frame", row.names = c(NA, -20L))
> measure <- measure[,-1]
> names(measure) <- c("chest", "waist", "hips")
> measure$gender <- gl(2, 10)
> levels(measure$gender) <- c("male", "female")
>
> # to only extract the continuous data
> measures <- measure[,c(1:3)]
> par(mfrow=c(3, 3))
> for (i in 1:3) {
+   for (j in 1:ncol(measures)) {
+     if(i != j) {
+       bvbox(measures[, c(i, j)])
+     }
+     else {
+       boxplot(measures[[i]])
+     }
+   }
+ }
```



It is quite challenging to draw the graph in the author's way! First, `par("usr")` will return the coordiante of your current plot in the unit your data in `(xmin,xmax,ymin,ymax)` order.

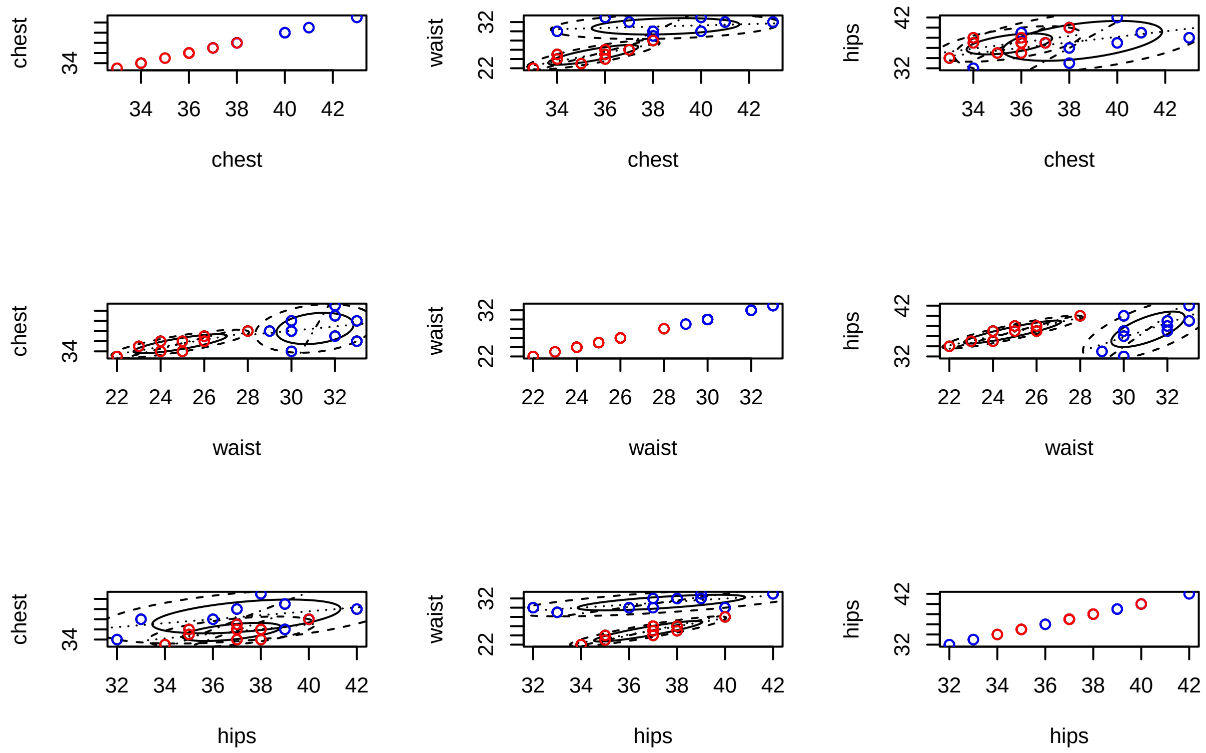
This scatterplot matrix can easily help reader find the outlier and the descriptive information of each variable, like the quantile. In contrasr, the figure 2.17 can help reader to find the joint distribution of each pair and the distribution of the variable itself. I tend to think the plot above is more infomrative.

0.4 Ex. 2.4

Construct a further scatterplot matrix of the body measurements data that labels each point in a panel with the gender of the individual, and plot on each scatterplot the separate estimated bivariate densities for men and women.

MY SOLUTION:

```
> ncols <- 3 # only the first 3 columns are numeric
> par(mfrow=c(ncols, ncols))
> for (i in 1:ncols) {
+   for (j in 1:ncols) {
+     plot(measure[, i], measure[, j], xlab = names(measure)[i], ylab=names(measure)[j])
+     if(i != j) {
+       bvbox(measure[measure$gender == "male", c(i, j)], add=TRUE)
+       bvbox(measure[measure$gender == "female", c(i, j)], add=TRUE)
+     }
+     points(measure[measure$gender == "male", c(i, j)], col="blue")
+     points(measure[measure$gender == "female", c(i, j)], col="red")
+   }
+ }
```

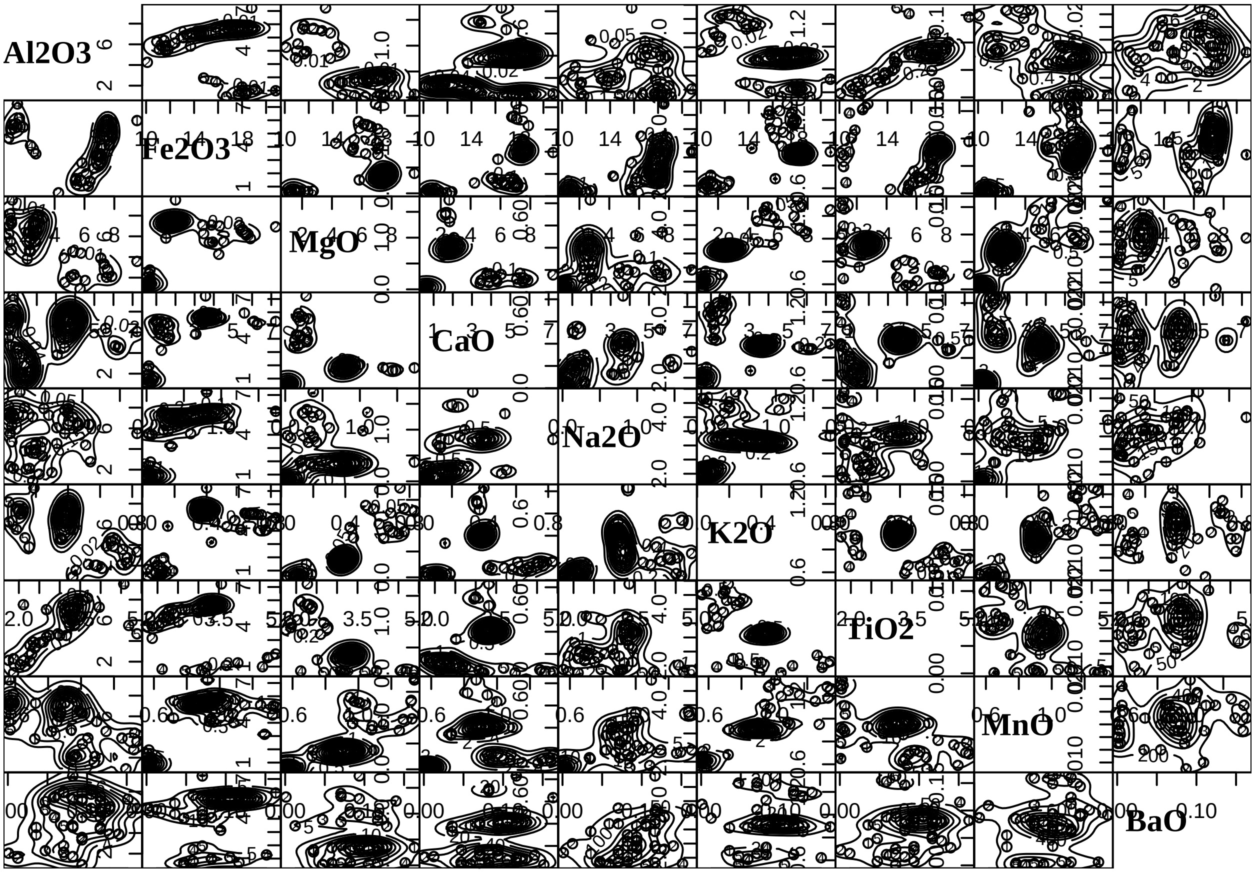


0.5 Ex. 2.5

Construct a scatterplot matrix of the chemical composition of Romano-British pottery given in Chapter 1 (Table 1.3), identifying each unit by its kiln number and showing the estimated bivariate density on each panel. What does the resulting diagram tell you?

MY SOLUTION:

```
> # pairs(measure, panel = function(x, y) plot(density(cbind(x, y)), add = TRUE))
> library("KernSmooth")
> par(mar = c(1, 1, 1, 1))
> n = ncol(pottery) - 1
> par(mfrow = c(n, n))
> for (i in 1:n) {
+   for(j in 1:n){
+
+     if(i == j){
+       #d = density(pottery[, i])
+       #plot(d, main = colnames(pottery)[i])
+       #text(d, cex = 0.6,
+            #labels = abbreviate(pottery$kiln))
+       par(mar = c(0,0,0,0))
+       plot(c(0, 1), c(0, 1), ann = F, bty = 'n', type = 'n', xaxt = 'n', yaxt = 'n')
+
+       text(x = 0.3, y = 0.5, paste(colnames(pottery)[i]),
+            cex = 1.5, col = "black", family="serif", font=2, adj=0.5)
+     }
+     else{
+       measured = bkde2D(pottery[, c(i, j)],
+                 bandwidth = sapply(pottery[, c(i, j)], dpik))
+
+       plot(pottery[, c(i, j)],
+            xlab = colnames(pottery)[i],
+            ylab = colnames(pottery)[j])
+       contour(x = measured$x1, y = measured$x2,
+               z = measured$fhat, add = TRUE)
+       text(pottery[,i], pottery[,j], cex = 0.6,
+            labels = abbreviate(pottery$kiln))
+
+     }
+   }
+
+ }
```

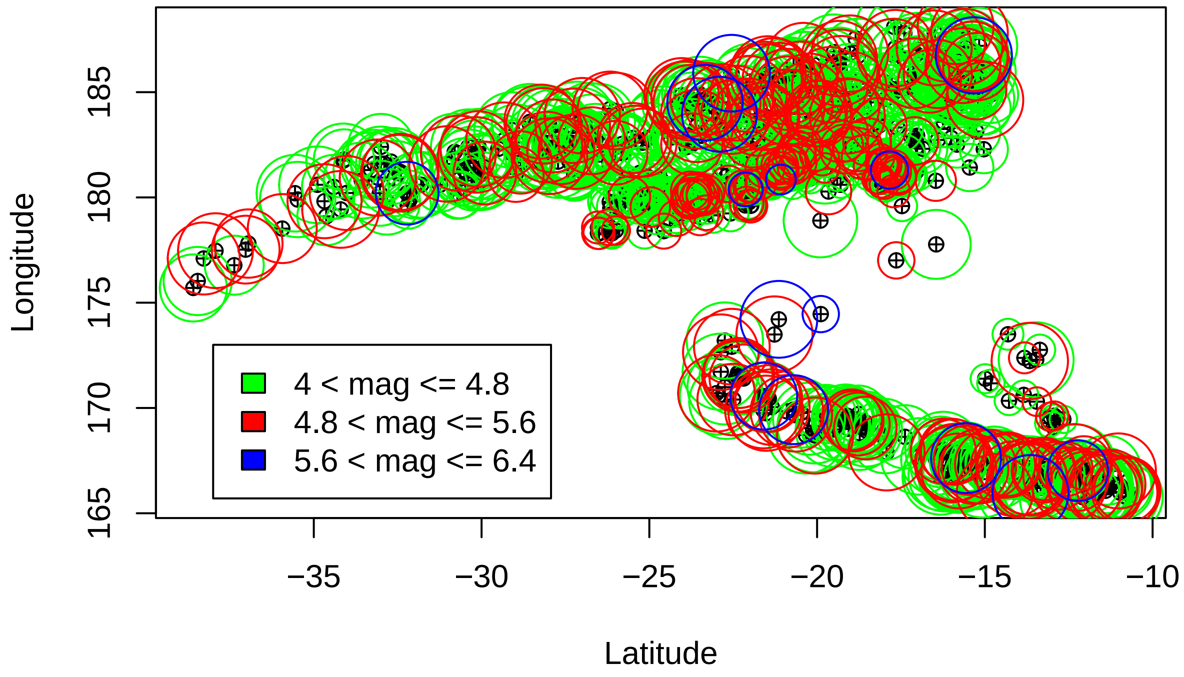


The results is actually not very good-looking, but it still presents the joint distribution of each pair. These joint distributions seems not follow any regular probability distribution.

0.6 Ex. 2.6

Ex. 2.6 Construct a bubble plot of the earthquake data using latitude and longitude as the scatterplot and depth as the circles, with greater depths giving smaller circles. In addition, divide the magnitudes into three *equal ranges and label the points in your bubble plot with a different symbol depending on the magnitude group into which the point falls.*

```
> library(lattice)
> par(mfrow = c(1, 1))
> #ylim <- with(quakes, range(long)) * c(0.95, 1)
> sub1 = subset(quakes, subset = mag > 4 & mag <= 4.8)
> sub2 = subset(quakes, subset = mag > 4.8 & mag <= 5.6)
> sub3 = subset(quakes, subset = mag >5.6 & mag <= 6.4)
>
>
> attach(quakes)
> plot(lat, long,
+      xlab = "Latitude",ylab = "Longitude", pch=10)
>
> with(sub1, symbols(lat, long, circles = 1000- depth,
+                    inches = 0.2, add = T, fg="green"))
> with(sub2, symbols(lat, long, circles = 1000- depth,
+                    inches = 0.2, add = T, fg="red"))
> with(sub3, symbols(lat, long, circles = 1000- depth,
+                    inches = 0.2, add = T, fg="blue"))
> legend(-38,173, legend = c("4 < mag <= 4.8",
+                        "4.8 < mag <= 5.6",
+                        "5.6 < mag <= 6.4"), fill=c("green","red","blue"))
```


Longitude
Latitude
🟩 4 < mag <= 4.8
🟥 4.8 < mag <= 5.6
🟦 5.6 < mag <= 6.4