Descriptive Statistics: the Education edition

M P Gururajan, Hina A Gokhale and Dayadeep Monder

Indian Institute of Technology Bombay, Mumbai

In this session, we are going to use a data set downloaded from `data.gov.in` site on education for our analysis. The dataset is called `Education.csv` and is stored in `csv` format.

Of course, we know how to load the data set. Before we do that, we will also remove all data from the previous sessions. And, after reading the data, we will check it once before proceeding further.

```
unlink("~/.RData")
X <- read.csv("../Data/Education.csv",header=TRUE)
str(X)
```

```
## 'data.frame': 36 obs. of  16 variables:
##  $ State                                                      : chr  "Andhra Padesh" "Arunacha
##  $ Population.in..Cr..of.Total.Census.2001                    : num  7.62 0.11 2.67 8.3 2.08 0
##  $ Population.in..Cr..of.6.14.age.2004                        : num  1.3 0.02 0.53 1.88 0.41 0
##  $ Literacy.Rate                                              : num  60.5 54.3 63.3 47 64.7 82
##  $ Gross.Enrollment.Ratio..Classes...I.VIII.                  : num  87 106.7 91.9 65.2 112.6
##  $ Drop.out.Classes..I.X.                                     : num  63.7 70.8 75 83.1 0 ...
##  $ Pupil.Teacher.Ratio                                        : int  33 34 42 104 48 21 35 44
##  $ Pupil.Teacher.Ratio.Upper.Primary                          : int  31 30 16 75 46 17 39 30 3
##  $ Elementary.School.per.lakh.population                      : int  99 163 137 57 203 76 73 6
##  $ Sec.Hr.Sec.Schools.per.lakh.population                     : int  22 19 19 4 12 31 14 23 37
##  $ Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure...Rs.cr.        : num  1807 187 1165 2479 1440 .
##  $ Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure.as.percentage.of.Total: num  5.09 0.53 3.28 6.99 4.06
##  $ Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure...Per.capita.6.14.age : int  1387 8179 2186 1319 3513
##  $ Lakh.of.Population.per.Institution..University.            : num  31.6 11.4 46.5 46.2 43.8
##  $ Lakh.of.Population.per.Institution..College.               : num  0.59 1.14 0.88 1.18 1.03
##  $ Lakh.of.Population.per.Institution..Technical.             : num  1.97 5.7 25.34 43.87 18.2
```

We know that this data, in addition to the statewise information, also contains all India information. In some cases, we want to remove the India information; we can generate a new data frame in which the last line is removed as follows. Note that we are storing one row of data corresponding to India in a different variable.

```
Y <- X[-36,]
Z <- X[36,]
str(Y)
```

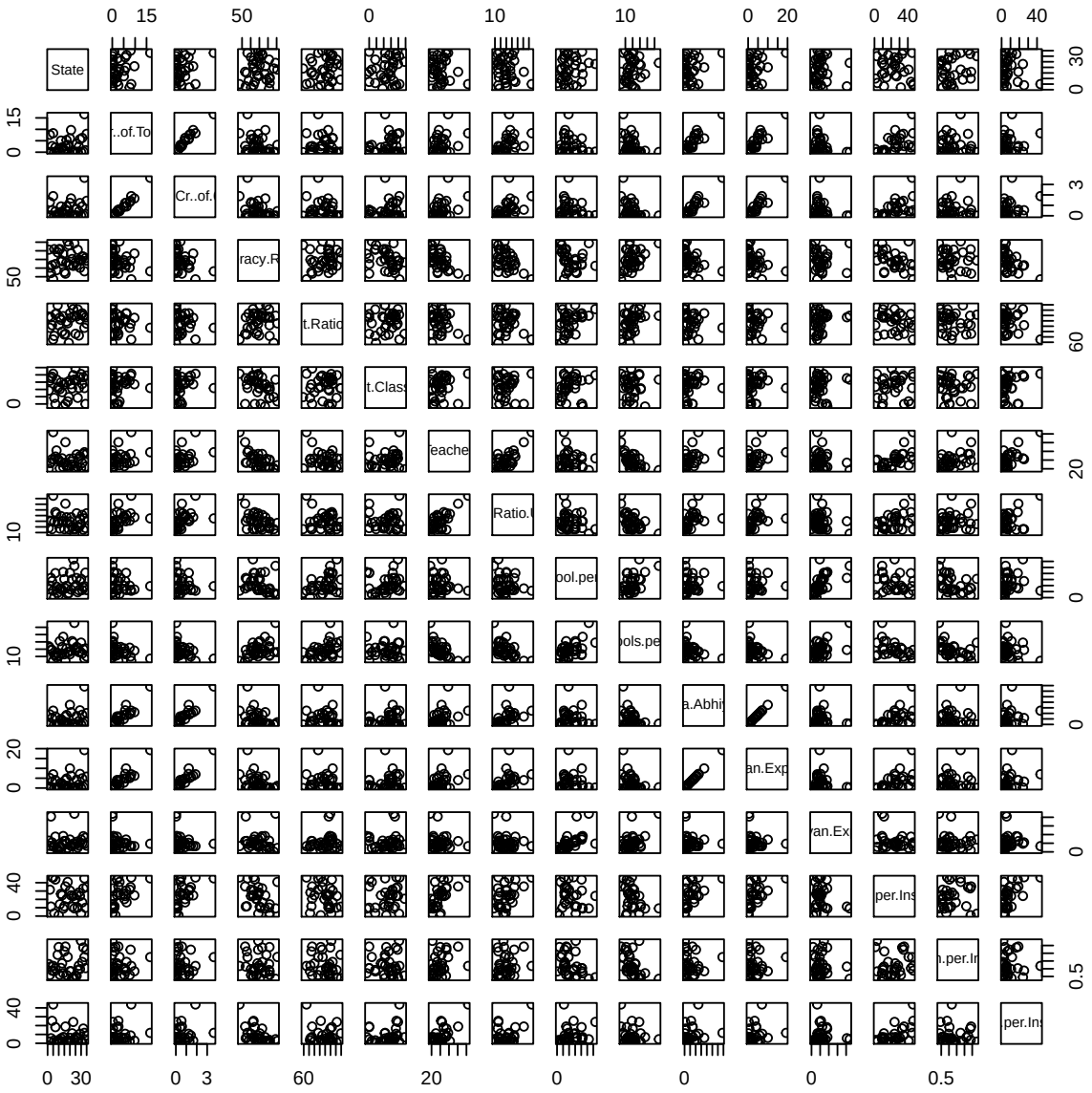
```
## 'data.frame': 35 obs. of  16 variables:
##  $ State                                                      : chr  "Andhra Padesh" "Arunacha
##  $ Population.in..Cr..of.Total.Census.2001                    : num  7.62 0.11 2.67 8.3 2.08 0
##  $ Population.in..Cr..of.6.14.age.2004                        : num  1.3 0.02 0.53 1.88 0.41 0
```

```
##  $ Literacy.Rate                                                  : num  60.5 54.3 63.3 47 64.7 82
##  $ Gross.Enrollment.Ratio..Classes...I.VIII.                      : num  87 106.7 91.9 65.2 112.6
##  $ Drop.out.Classes..I.X.                                         : num  63.7 70.8 75 83.1 0 ...
##  $ Pupil.Teacher.Ratio                                            : int  33 34 42 104 48 21 35 44
##  $ Pupil.Teacher.Ratio.Upper.Primary                              : int  31 30 16 75 46 17 39 30 3
##  $ Elementary.School.per.lakh.population                          : int  99 163 137 57 203 76 73 6
##  $ Sec.Hr.Sec.Schools.per.lakh.population                         : int  22 19 19 4 12 31 14 23 37
##  $ Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure...Rs.cr.           : num  1807 187 1165 2479 1440
##  $ Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure.as.percentage.of.Total: num  5.09 0.53 3.28 6.99 4.06
##  $ Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure...Per.capita.6.14.age : int  1387 8179 2186 1319 3513
##  $ Lakh.of.Population.per.Institution..University.                : num  31.6 11.4 46.5 46.2 43.8
##  $ Lakh.of.Population.per.Institution..College.                   : num  0.59 1.14 0.88 1.18 1.03
##  $ Lakh.of.Population.per.Institution..Technical.                 : num  1.97 5.7 25.34 43.87 18.2
```

1 Correlation

It is possible to plot the complete data – each column against all the other columns. This can be done by simply calling the plot function:

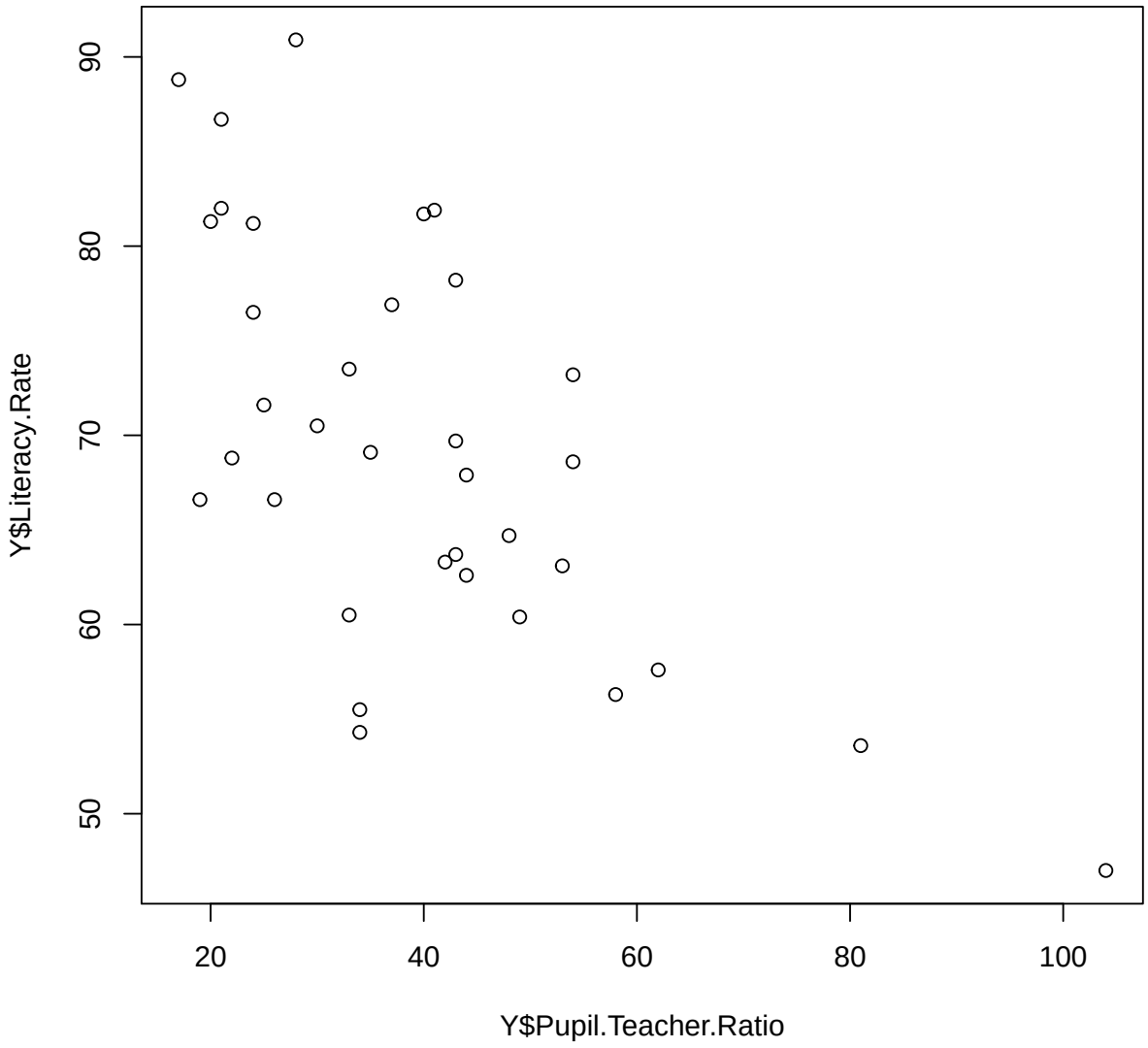
```
plot(Y)
```



From these plots, once you zoom in and see, it is easy to notice the trends and correlations, if any!

Of course, one can also plot specific variables. For example, is literacy rate related to the pupil teacher ratio? Let us check!

```
plot(Y$Pupil.Teacher.Ratio,Y$Literacy.Rate)
```



The overall trend seems to be that the literacy rate decreases as the pupil teacher ratio increases.

How about the correlation coefficient? As you have learnt from the lecture, the correlation coefficient is defined as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

So, let us write an R script that does this. We are going to calculate the coefficient of correlation in two different ways. In one case, we are going to use the `sum` function. In the other, we are going to explicitly run a loop.

```
x <- Y$Pupil.Teacher.Ratio
y <- Y$Literacy.Rate
xbar = mean(x) # Calculate the mean x
ybar = mean(y) # Calculate the mean y
```

```
xp <- x-xbar # Define xprime = x - xbar
yp <- y-ybar # Define yprime = y -ybar
r <- sum(xp*yp)/sqrt(sum(xp*xp)*sum(yp*yp))
r

## [1] -0.6509299

# Method 2
n <- length(x)
sum1 = 0
sum2 = 0
sum3 = 0
for(i in 1:n){
sum1 = sum1 + (x[i]-xbar)*(y[i]-ybar)
sum2 = sum2 + (x[i]-xbar)*(x[i]-xbar)
sum3 = sum3 + (y[i]-ybar)*(y[i]-ybar)
}
r = sum1/(sqrt(sum2*sum3))
r

## [1] -0.6509299
```

As one can see, there is correlation and it is negative.

Homework

- Plot the literacy rate against the elementary schools per lakh population. From the plot, can you see if they are correlated? Calculate the correlation coefficient and confirm your answer.
- Plot the literacy rate against the tenth plan Sarva Siksha Abhiyan expenditure (per capita in the 6 to 14 age group). From the plot, can you see if they are correlated? Calculate the correlation coefficient and confirm your answer.
- In terms of spending per capita, number of elementary schools and pupil to teachers ratio, which is the most effective in terms of improving literacy rate? Order them in light of the given data and your analysis.

2 Ordering and plotting data

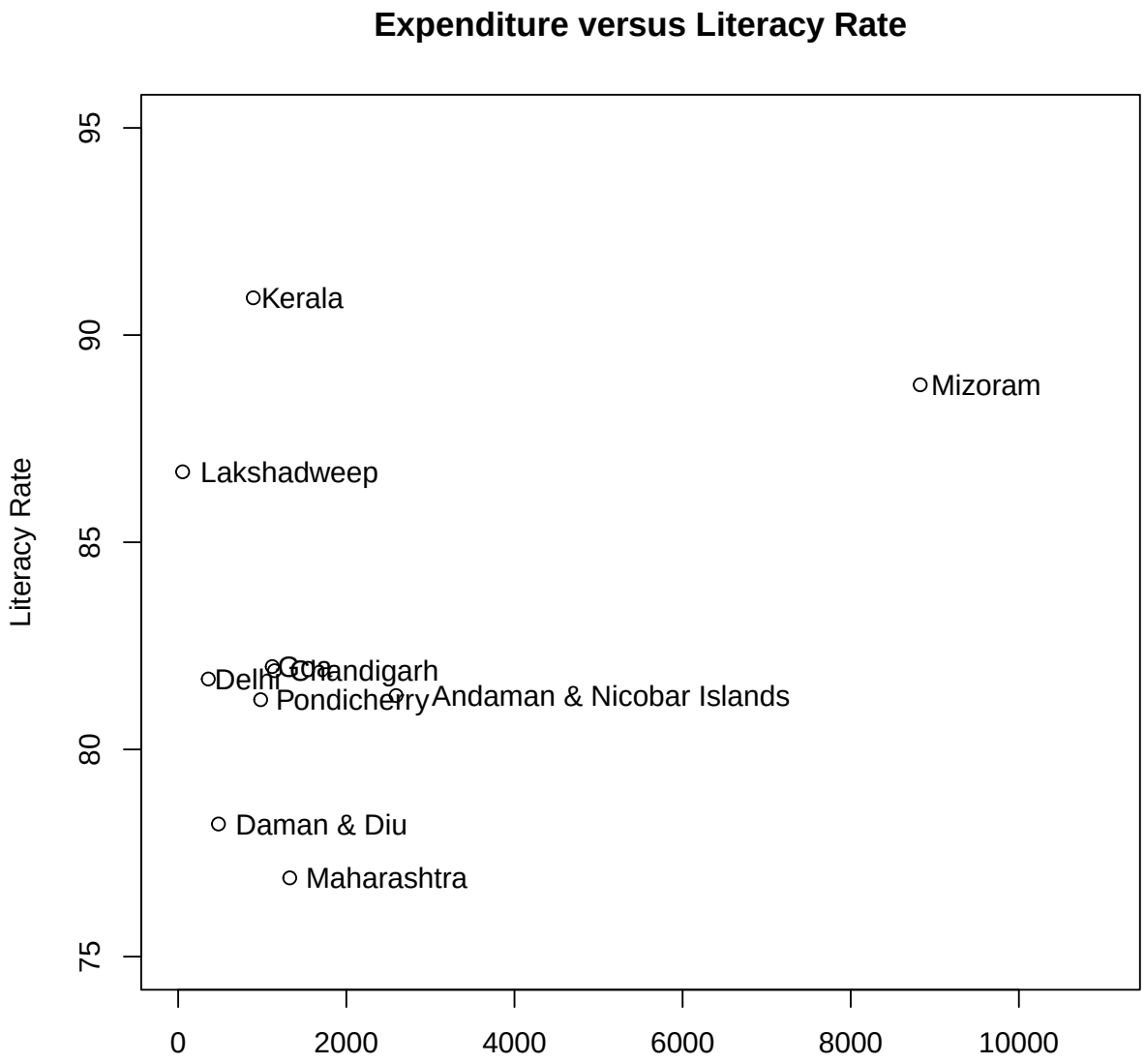
Let us pick the top ten states / union territories in terms of the literacy rate and plot their literacy rate against the spending (per capita in the relevant age group). In this plot, we also want to label the data according to their state. The following script does that.

```
TopTen <- head(Y[order(Y$Literacy.Rate,decreasing=TRUE),],10) # Pick the top ten states
str(TopTen)

## 'data.frame': 10 obs. of  16 variables:
##  $ State                                                    : chr  "Kerala" "Mizoram" "Laksh
##  $ Population.in..Cr..of.Total.Census.2001                  : num  3.18 0.09 0.01 0.13 0.09
##  $ Population.in..Cr..of.6.14.age.2004                      : num  0.44 0.02 0 0.02 0.01 0.2
##  $ Literacy.Rate                                            : num  90.9 88.8 86.7 82 81.9 81
```

```
##  $ Gross.Enrollment.Ratio..Classes...I.VIII.                          : num  95.3 109.5 58.8 106 71.9
##  $ Drop.out.Classes..I.X.                                             : num  7.15 66.95 18.88 40.65 16
##  $ Pupil.Teacher.Ratio                                                : int  28 17 21 21 41 40 20 24 4
##  $ Pupil.Teacher.Ratio.Upper.Primary                                  : int  27 8 16 17 29 26 18 21 29
##  $ Elementary.School.per.lakh.population                              : int  30 263 38 76 3 20 69 45 4
##  $ Sec.Hr.Sec.Schools.per.lakh.population                             : int  16 56 16 31 12 11 24 22 1
##  $ Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure...Rs.cr.                : num  390.7 141.2 0.1 19 16 ...
##  $ Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure.as.percentage.of.Total: num  1.1 0.4 0 0.05 0.05 0.25
##  $ Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure...Per.capita.6.14.age : int  893 8826 53 1119 1146 358
##  $ Lakh.of.Population.per.Institution..University.                    : num  40.9 9.2 NA 14.1 3.33 1 N
##  $ Lakh.of.Population.per.Institution..College.                       : num  1.76 0.35 NA 0.61 0.83 2.
##  $ Lakh.of.Population.per.Institution..Technical.                     : num  2.68 4.6 NA 1.57 3.33 3.6

plot(Literacy.Rate~Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure...Per.capita.6.14.age,data=TopTen,main="L
text(Literacy.Rate~Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure...Per.capita.6.14.age,data=TopTen,labels=
```



From the figure, one can see that Mizoram is an outlier in terms of spending.

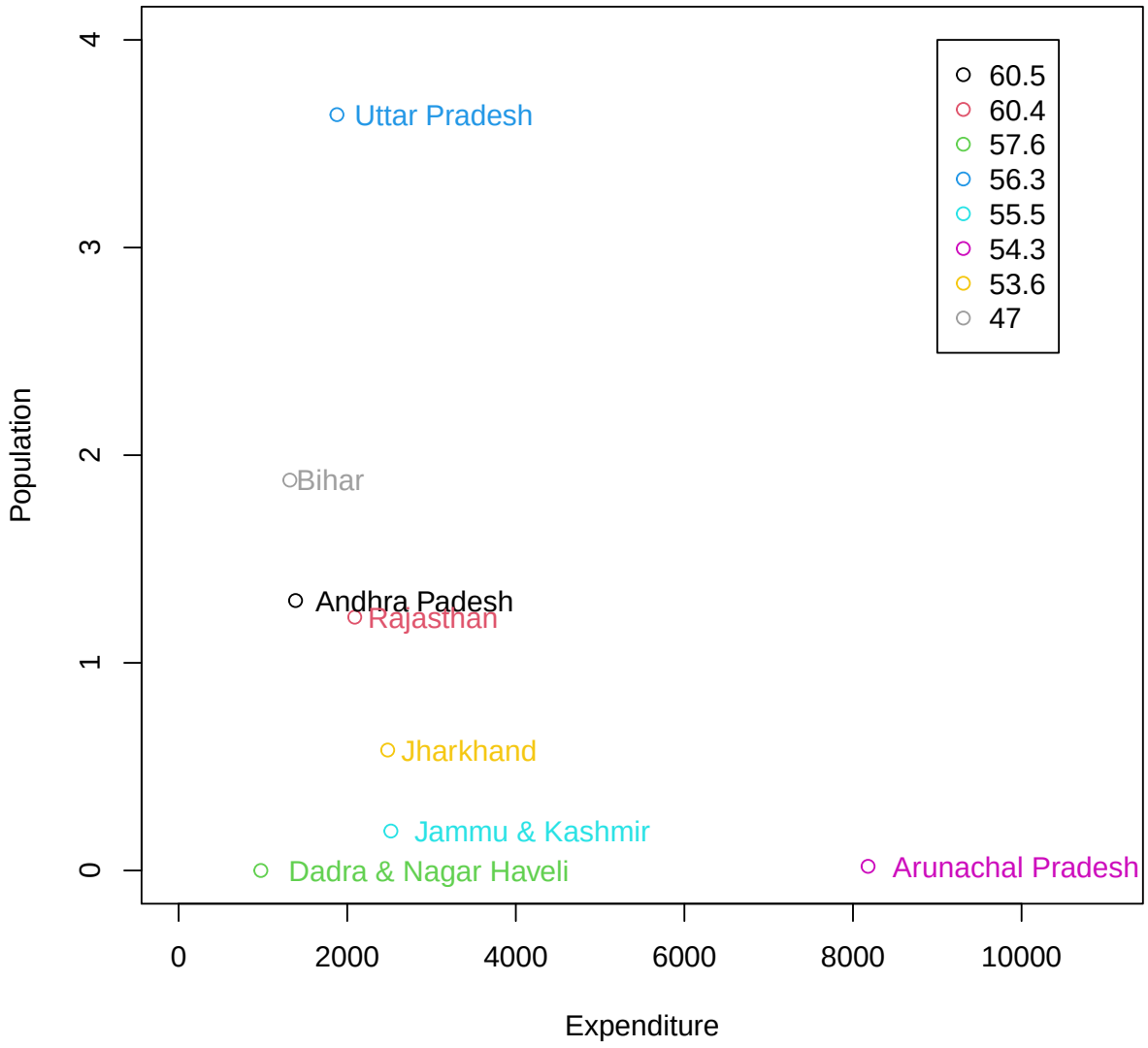
Homework

- In the above script, what happens to the plot if you remove the `xlim` and `ylim` commands?
- Pick the states which are in the bottom ten in terms of literacy and make a plot of spending versus literacy rate. Do you see any outliers?
- Plot two pie-charts next to each other – one of spending and another of literacy rate – for states in the bottom ten. Below them, plot the same two pie-charts for the states in top ten. What do you see?

3 Adding colour!

Adding colour to the plots can help bring in one more dimension to the plots. Here is an example of the expenditure versus population of the bottom eight states plotted with a colour marker which indicated their literacy rate; of course, we have also labeled the data points to indicate the State / UT.

```
BottomEight <- tail(Y[order(Y$Literacy.Rate,decreasing=TRUE),],8)
# Pick the bottom eight states
plot(BottomEight$Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure...Per.capita.6.14.age,
BottomEight$Population.in..Cr..of.6.14.age.2004,col=1:length(BottomEight$Literacy.Rate),
xlab="Expenditure",ylab="Population",
xlim=c(0,11000),ylim=c(0,4))
legend(9000,4,BottomEight$Literacy.Rate,col=1:length(BottomEight$Literacy.Rate),pch=1)
text(BottomEight$Tenth.Plan.Sarva.Siksha.Abhiyan.Expenditure...Per.capita.6.14.age,
BottomEight$Population.in..Cr..of.6.14.age.2004,
labels=(BottomEight[1:length(BottomEight$Literacy.Rate),1]),
adj=c(-0.1,0.5),col=1:length(BottomEight$Literacy.Rate))
```



In the above figure, if I want the names of the states / UTs printed below the point and not next to it, what modification needs to be done?

As a final exercise, plot the histograms of (a) Lakh of Population per University, (b) Tenth plan expenditure, and (c) literacy rate. What can you comment about these distributions? Which data set looks closer to normal, for example? Which data set is skewed? In which direction?

Of course, with the given data set, you can plot many more plots and have much more fun! Happy programming!!