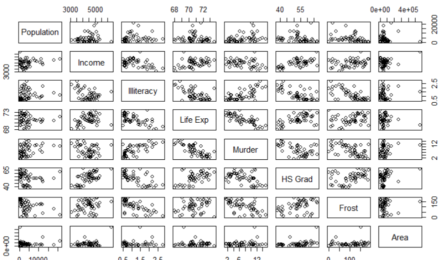
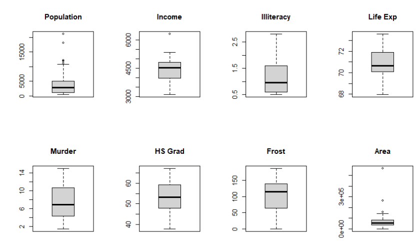
2. PCA on State Statistics:

a. The dataset has 8 columns and 50 rows. I started off by identifying if any correlation exists between the columns using a pair plot. From the pairplot, we can see that there is some correlation or relationship among some columns as we see some linear trends. For example, we see a linear trend between income and HS grad, Illiteracy and life expectancy, etc. After doing the summary() function we can see the unit of measurements of these columns differ a lot, hence scaling will be needed moving forward. From the boxplots, we see that population, area, and income have a few outliers.

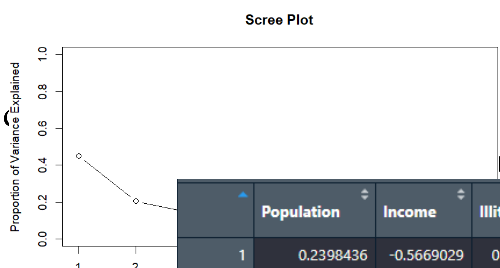
```
> summary(st)
   Population        Income       Illiteracy       Life Exp         Murder          HS Grad
 Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96   Min.   : 1.400   Min.   :37.80
 1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12   1st Qu.: 4.350   1st Qu.:48.05
 Median : 2838   Median :4519   Median :0.950   Median :70.67   Median : 6.850   Median :53.25
 Mean   : 4246   Mean   :4436   Mean   :1.170   Mean   :70.88   Mean   : 7.378   Mean   :53.11
 3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89   3rd Qu.:10.675   3rd Qu.:59.15
 Max.   :21198   Max.   :6315   Max.   :2.800   Max.   :73.60   Max.   :15.100   Max.   :67.30
     Frost             Area
 Min.   :  0.00   Min.   :  1049
 1st Qu.: 66.25   1st Qu.: 36985
 Median :114.50   Median : 54277
 Mean   :104.46   Mean   : 70736
 3rd Qu.:139.75   3rd Qu.: 81163
 Max.   :188.00   Max.   :566432
```

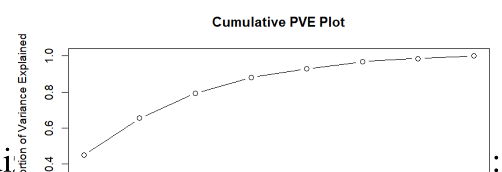




b. Yes, as discussed above standardizing the variables will be a good idea as we can see by results of summary() our columns have different ranges

c. After plotting the scree and cumulative pve plot, we can see that we should use 5 components indicated by the elbow shape, and insignificant gain afterwards. Hence, 5 pc helps us explain around 95% variability.





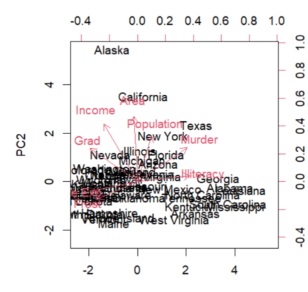
	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
1	0.2398436	-0.5669029	0.88720374	-0.7808560	0.8427885	-0.8056584	-0.6780384	-0.06333314
2	0.5248778	0.6629778	0.06766573	-0.1043129	0.3921173	0.3816642	-0.1961984	0.75067024

ii. Cumulative percentage of the total variability by the two component:

```
> cumsum(pve)[1:2]
[1] 0.4498619 0.6538519
```

iii. Scores on the two components & bi plot:

```
> pca.fit$rotation[,1:2]
                   PC1         PC2
Population  0.12642809  0.41087417
Income     -0.29882991  0.51897884
Illiteracy  0.46766917  0.05296872
Life Exp   -0.41161037 -0.08165611
Murder      0.44425672  0.30694934
HS Grad    -0.42468442  0.29876662
Frost      -0.35741244 -0.15358409
Area       -0.03338461  0.58762446
```



iv. Interpretation of results: we can clearly see that PC1 is more correlated with Illiteracy, life expectancy, murder, HS grad, frost, while PC2 is more correlated with population, income, area. Furthermore, a southern component is illiteracy and murder as we see alot of southern states pointed in that direction on the biplot. Hence, southern states scored better on PC1 than PC2.