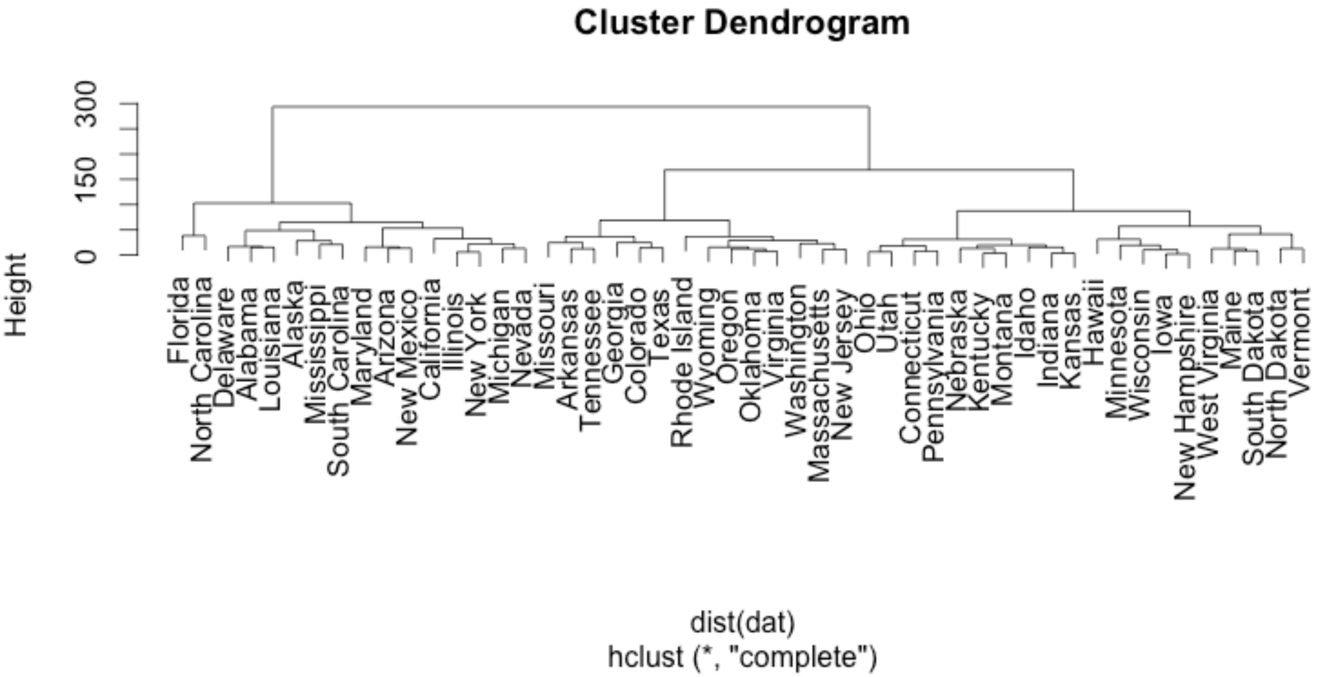
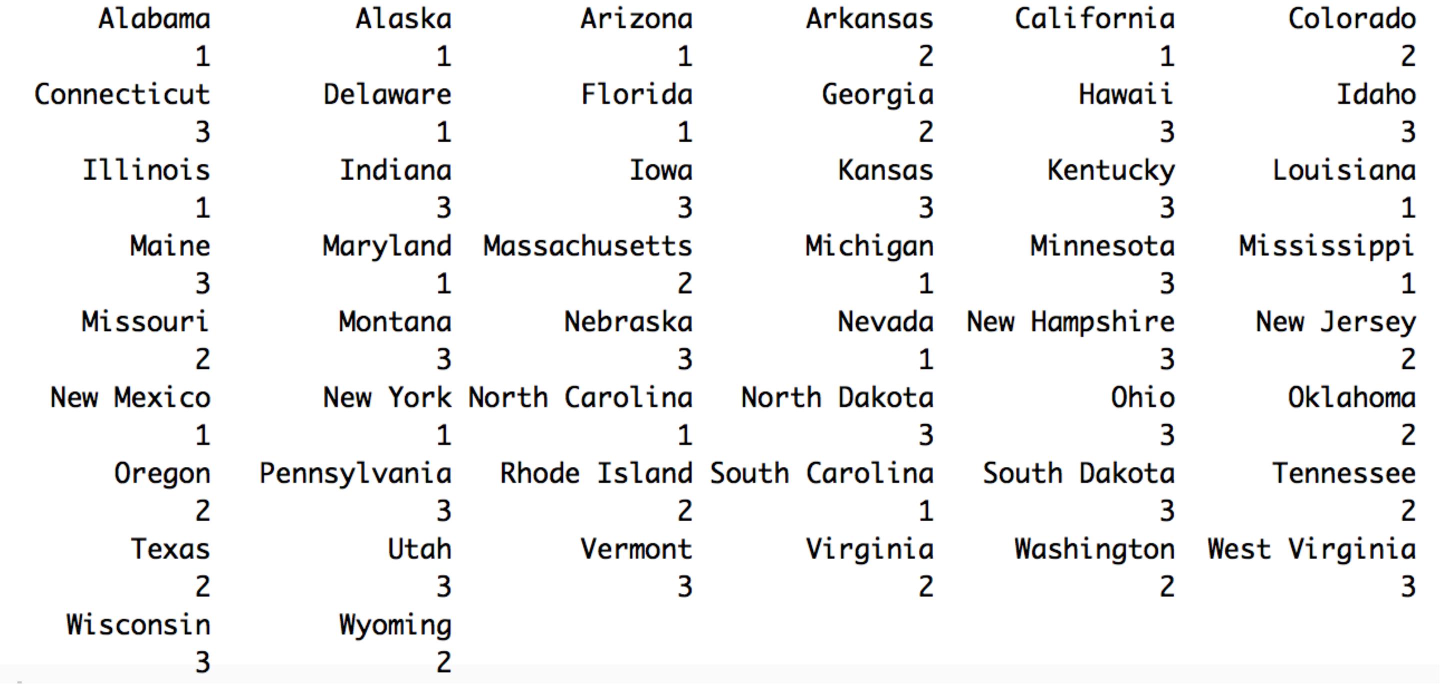
Name: Parth Pareek
UNI: PP2547
Date: 4/6/2016
Assignment: Homework 8

1.

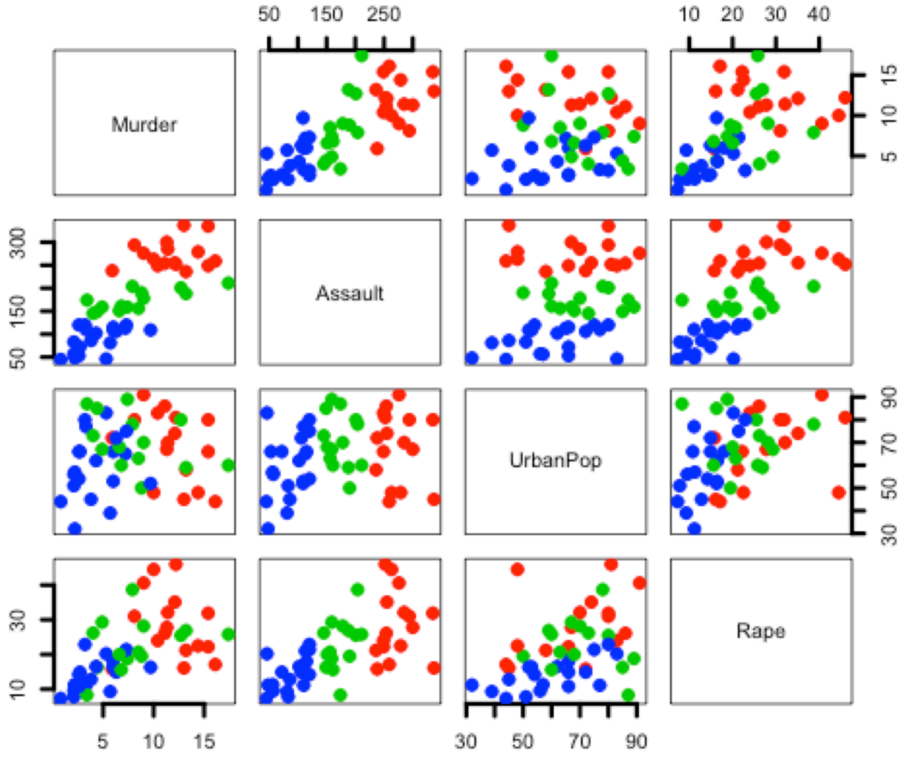
a.



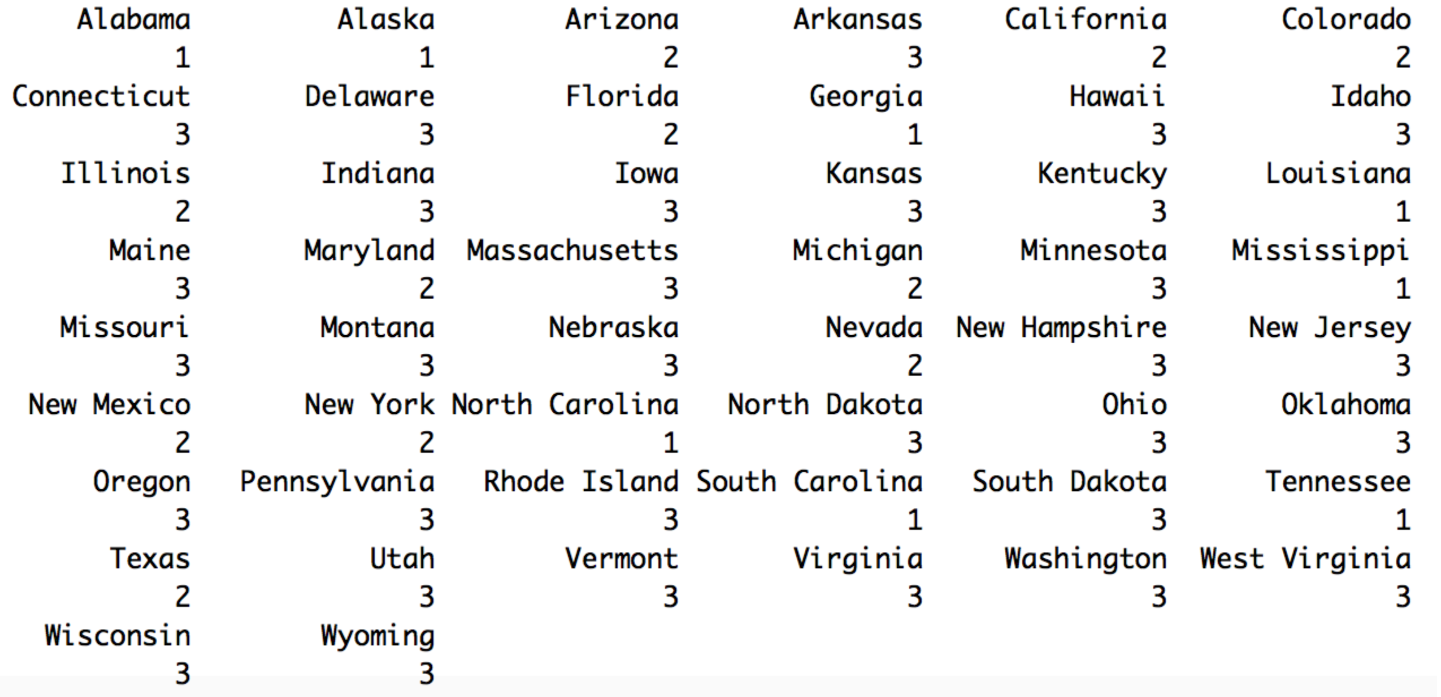
b.

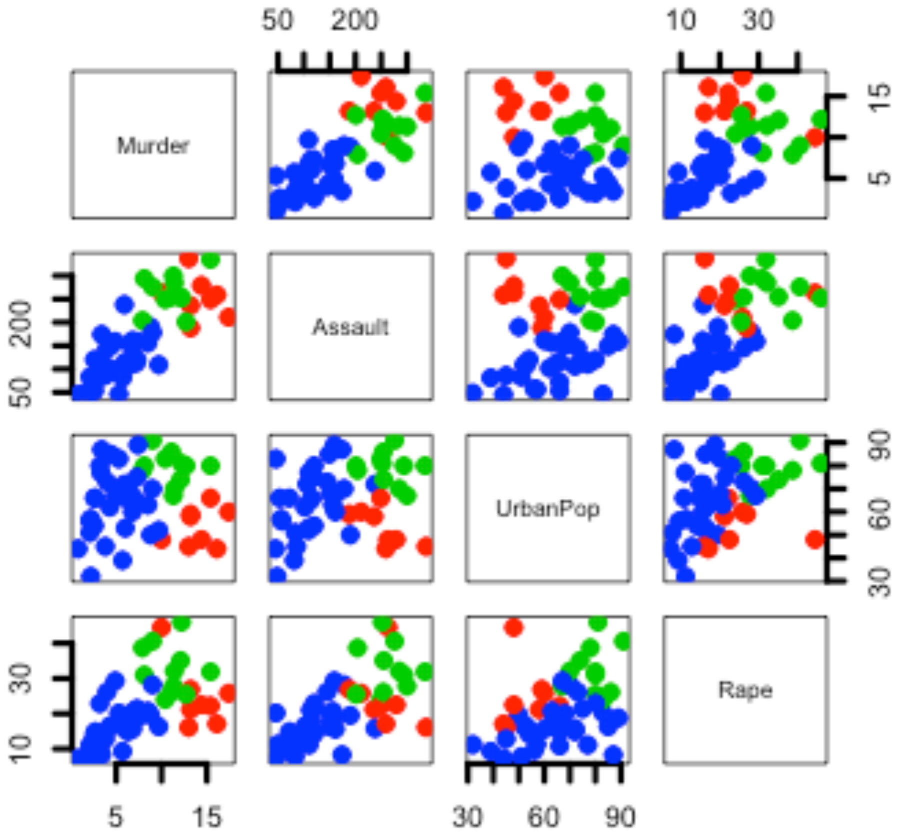


```
       Alabama         Alaska        Arizona       Arkansas     California       Colorado
             1              1              1              2              1              2
   Connecticut       Delaware        Florida        Georgia         Hawaii          Idaho
             3              1              1              2              3              3
      Illinois        Indiana           Iowa         Kansas       Kentucky      Louisiana
             1              3              3              3              3              1
         Maine       Maryland  Massachusetts       Michigan      Minnesota    Mississippi
             3              1              2              1              3              1
      Missouri        Montana       Nebraska         Nevada  New Hampshire     New Jersey
             2              3              3              1              3              2
    New Mexico       New York North Carolina   North Dakota           Ohio       Oklahoma
             1              1              1              3              3              2
        Oregon   Pennsylvania   Rhode Island South Carolina   South Dakota      Tennessee
             2              3              2              1              3              2
         Texas           Utah        Vermont       Virginia     Washington  West Virginia
             2              3              3              2              2              3
     Wisconsin        Wyoming
             3              2
```



c.

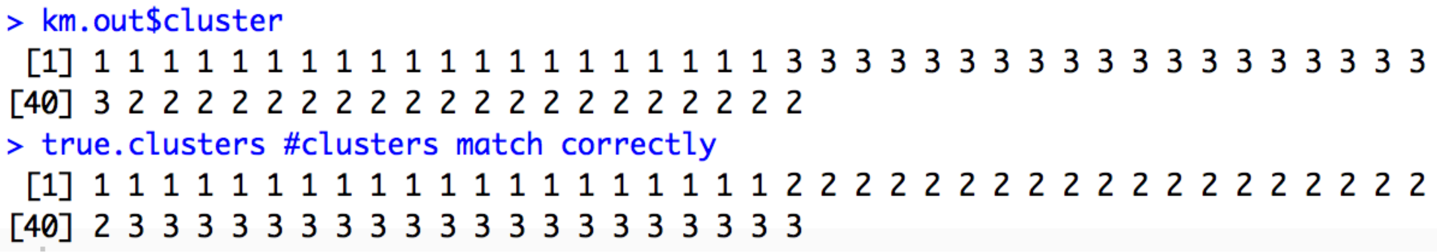

```
       Alabama         Alaska        Arizona       Arkansas     California       Colorado 
             1              1              2              3              2              2 
   Connecticut       Delaware        Florida        Georgia         Hawaii          Idaho 
             3              3              2              1              3              3 
      Illinois        Indiana           Iowa         Kansas       Kentucky      Louisiana 
             2              3              3              3              3              1 
         Maine       Maryland  Massachusetts       Michigan      Minnesota    Mississippi 
             3              2              3              2              3              1 
      Missouri        Montana       Nebraska         Nevada  New Hampshire     New Jersey 
             3              3              3              2              3              3 
    New Mexico       New York North Carolina   North Dakota           Ohio       Oklahoma 
             2              2              1              3              3              3 
        Oregon   Pennsylvania   Rhode Island South Carolina   South Dakota      Tennessee 
             3              3              3              1              3              1 
         Texas           Utah        Vermont       Virginia     Washington  West Virginia 
             2              3              3              3              3              3 
     Wisconsin        Wyoming 
             3              3 
```



d. Scaling changes the results and yes, in this case data should be scaled before clustering since means and variances of the parameters is vastly different in this case. If data is not scaled, parameter with high variance will dominate clustering. Scaling would yield better clustering results since all parameters are now given equal weights.

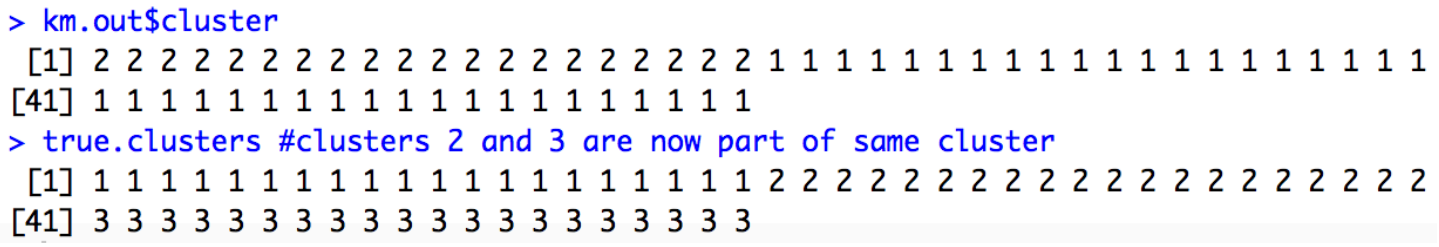
2.

a. K-means clusters match the true clusters



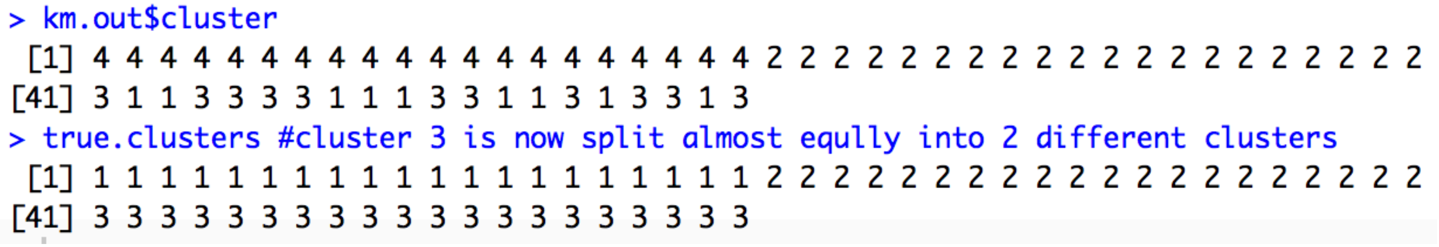
```
> km.out$cluster
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[40] 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
> true.clusters #clusters match correctly
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[40] 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

b. Clusters 2 and 3 (true) are now combined as 1 cluster



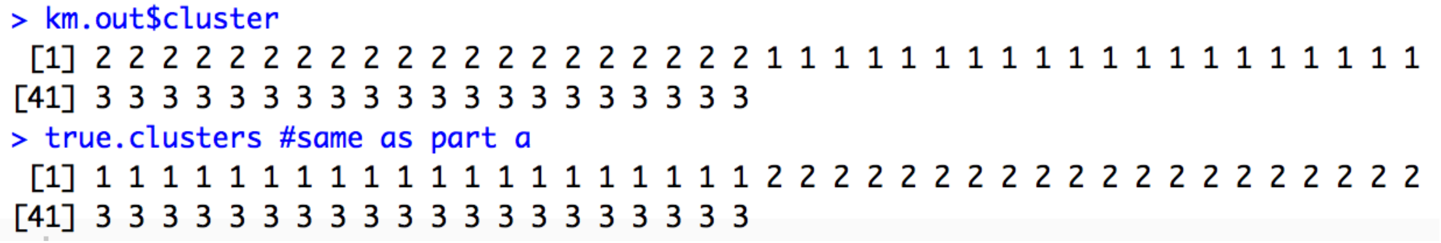
```
> km.out$cluster
 [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[41] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
> true.clusters #clusters 2 and 3 are now part of same cluster
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[41] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

c. Cluster 3 (true) is not split into almost equal clusters



```
> km.out$cluster
 [1] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[41] 3 1 1 3 3 3 3 1 1 1 3 3 1 1 3 1 3 3 1 3
> true.clusters #cluster 3 is now split almost eqully into 2 different clusters
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[41] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

d. Clusters remain same in part (a). Scaling doesn't affect in this case since means and variances of all columns in similar before and after scaling. Scaling is preferable when columns had varied means and variance. However, TSS within columns might reduce after scaling.



```
> km.out$cluster
 [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[41] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
> true.clusters #same as part a
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[41] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

3.

a. Increase in test score = 4.56 (48.42%)
b. Shrinkage coefficient = 0.61066
c. RMSE of predictions = 4.218803
d. Increase in test score (from new model) = 1.52
e. The increase in score reduced in part (d) after taking into account regression effect. When increase in scores is calculated in part (a), it includes regression effect as well as effect of the SIS program, however, after accounting for regression, the increase reduces.