ashishsa_hw2_p1

We take the USArrests Data and perform Heirarchical Clustering with complete linkage and eucledian distance to cluster the states.

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.6.3
```

```
data("USArrests")
df1 <- USArrests
head(df1)
```

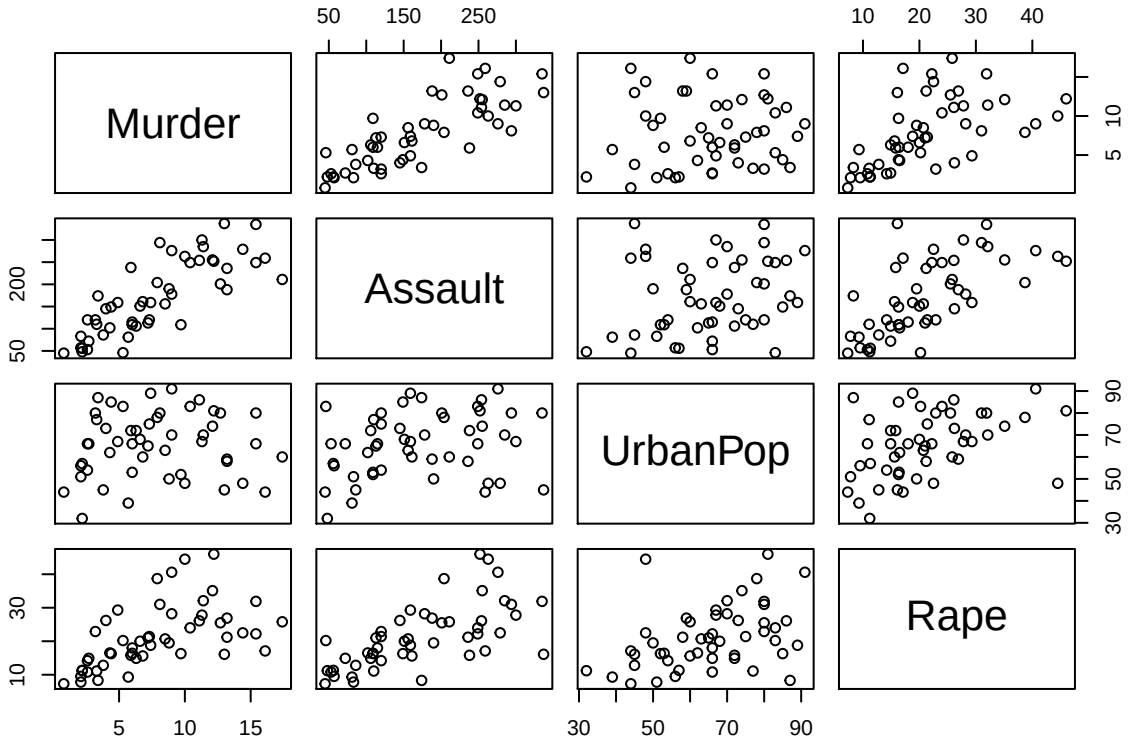
```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
nrow(df1)
```

```
## [1] 50
```

We observe that there are 4 columns(Murder, Assault, UrbanPop and Rape). There are 50 Rows here.

```
pairs(df1)
```



We compute the Eucledian Distance between the data points and calculate the complete linkage for each of the points.

We use the ggplot package to plot the Dendrogram

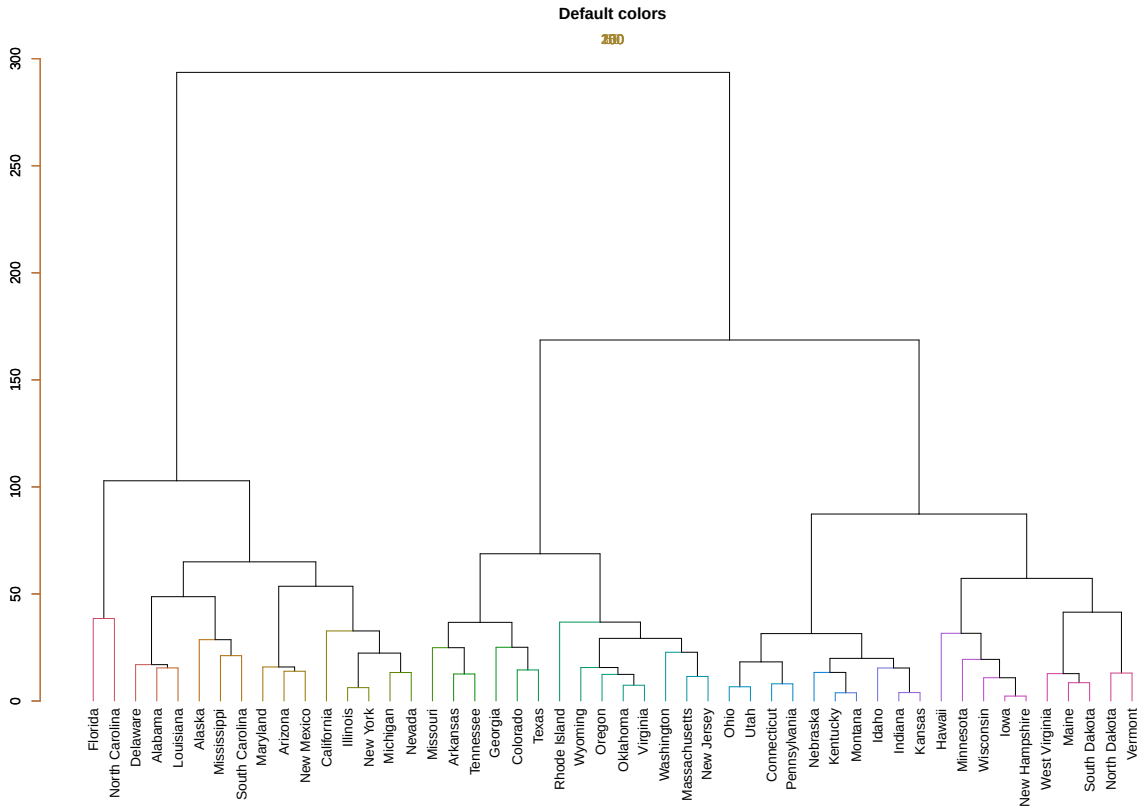
```
library(ggplot2)
library(ggdendro)
```

```
## Warning: package 'ggdendro' was built under R version 3.6.3
```

```
library(dplyr)
library(dendextend)
```

```
## Warning: package 'dendextend' was built under R version 3.6.3
```

```
dend1 <- as.dendrogram(hclust1)
dend1 %>% set("branches_k_color") %>%
  plot(main = "Default colors") %>%
  axis(side = 2,col = "#F38630",labels = TRUE) %>%
  mtext(col = "#A38630")
```



We now cut the Dendrogram at a height which results in three distinct clusters. We also print which cluster each state belongs to.

```
cutree1 <- cutree(dend1,3)
cutree1
```

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

```
clust1_1 <- names(which(cutree1==1))
```

```
clust1_2 <- names(which(cutree1==2))
clust1_3 <- names(which(cutree1==3))
```

```
clust1_1
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "California"
##  [5] "Delaware"       "Florida"        "Illinois"       "Louisiana"
##  [9] "Maryland"       "Michigan"       "Mississippi"    "Nevada"
## [13] "New Mexico"     "New York"       "North Carolina" "South Carolina"
```

```
clust1_2
```

```
##  [1] "Arkansas"      "Colorado"      "Georgia"       "Massachusetts"
##  [5] "Missouri"      "New Jersey"    "Oklahoma"      "Oregon"
##  [9] "Rhode Island"  "Tennessee"     "Texas"         "Virginia"
## [13] "Washington"    "Wyoming"
```

```
clust1_3
```

```
##  [1] "Connecticut"   "Hawaii"        "Idaho"         "Indiana"
##  [5] "Iowa"          "Kansas"        "Kentucky"      "Maine"
##  [9] "Minnesota"     "Montana"       "Nebraska"      "New Hampshire"
## [13] "North Dakota"  "Ohio"          "Pennsylvania"  "South Dakota"
## [17] "Utah"          "Vermont"       "West Virginia" "Wisconsin"
```

We can visualize this in the form of a coloured dendrogram as follows:

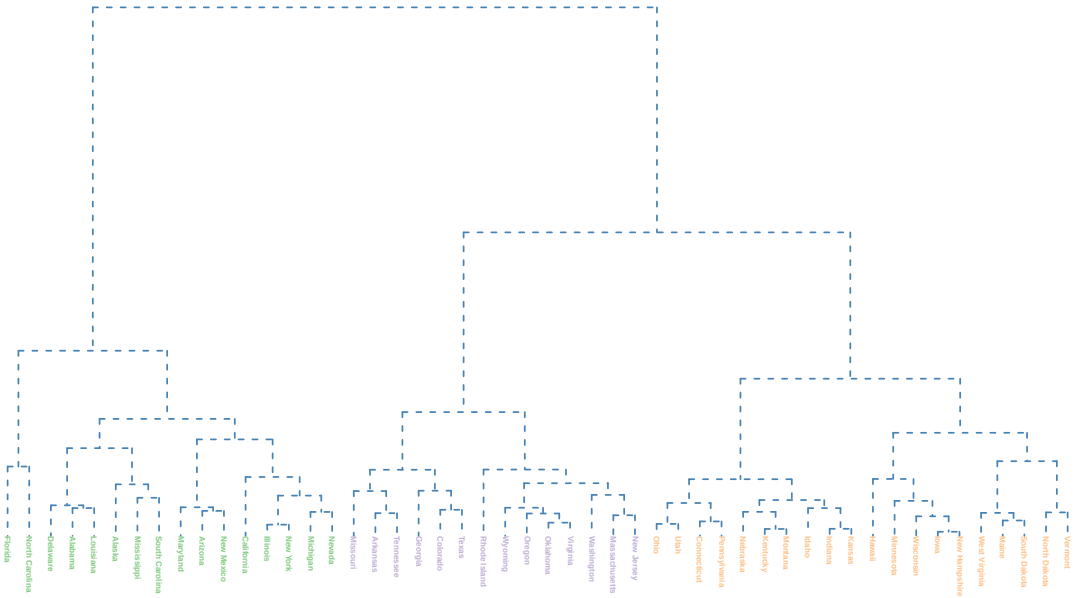
```
library(RColorBrewer)
library(ape)
```

```
## Warning: package 'ape' was built under R version 3.6.3
```

```
##
## Attaching package: 'ape'
```

```
## The following objects are masked from 'package:dendextend':
##
##     ladderize, rotate
```

```
plot(as.phylo(hclust1), type = "phylogram", cex = 0.6,
    tip.color = brewer.pal(3, 'Accent')[cutree1],
      direction = "downwards", font = 2,
     edge.color = 'steelblue', edge.width = 2, edge.lty = 2)
```



Heirarchically cluster states using complete linkage and Eucledian distance after scaling the variables to have a standard deviation of 1.

To scale the dataset we need to find the mean and the standard deviation. After finding the mean and standard deviation we subtract mean from each data point and then divide each data points with standard devaition.

We check if there are any missing values in the dataset.

```
is.na(df1)
```

```
##                Murder Assault UrbanPop  Rape
## Alabama         FALSE   FALSE    FALSE FALSE
## Alaska          FALSE   FALSE    FALSE FALSE
## Arizona         FALSE   FALSE    FALSE FALSE
## Arkansas        FALSE   FALSE    FALSE FALSE
## California      FALSE   FALSE    FALSE FALSE
## Colorado        FALSE   FALSE    FALSE FALSE
## Connecticut     FALSE   FALSE    FALSE FALSE
## Delaware        FALSE   FALSE    FALSE FALSE
## Florida         FALSE   FALSE    FALSE FALSE
## Georgia         FALSE   FALSE    FALSE FALSE
## Hawaii          FALSE   FALSE    FALSE FALSE
## Idaho           FALSE   FALSE    FALSE FALSE
## Illinois        FALSE   FALSE    FALSE FALSE
## Indiana         FALSE   FALSE    FALSE FALSE
## Iowa            FALSE   FALSE    FALSE FALSE
## Kansas          FALSE   FALSE    FALSE FALSE
## Kentucky        FALSE   FALSE    FALSE FALSE
```

```
## Louisiana       FALSE   FALSE    FALSE FALSE
## Maine           FALSE   FALSE    FALSE FALSE
## Maryland        FALSE   FALSE    FALSE FALSE
## Massachusetts   FALSE   FALSE    FALSE FALSE
## Michigan        FALSE   FALSE    FALSE FALSE
## Minnesota       FALSE   FALSE    FALSE FALSE
## Mississippi     FALSE   FALSE    FALSE FALSE
## Missouri        FALSE   FALSE    FALSE FALSE
## Montana         FALSE   FALSE    FALSE FALSE
## Nebraska        FALSE   FALSE    FALSE FALSE
## Nevada          FALSE   FALSE    FALSE FALSE
## New Hampshire   FALSE   FALSE    FALSE FALSE
## New Jersey      FALSE   FALSE    FALSE FALSE
## New Mexico      FALSE   FALSE    FALSE FALSE
## New York        FALSE   FALSE    FALSE FALSE
## North Carolina  FALSE   FALSE    FALSE FALSE
## North Dakota    FALSE   FALSE    FALSE FALSE
## Ohio            FALSE   FALSE    FALSE FALSE
## Oklahoma        FALSE   FALSE    FALSE FALSE
## Oregon          FALSE   FALSE    FALSE FALSE
## Pennsylvania    FALSE   FALSE    FALSE FALSE
## Rhode Island    FALSE   FALSE    FALSE FALSE
## South Carolina  FALSE   FALSE    FALSE FALSE
## South Dakota    FALSE   FALSE    FALSE FALSE
## Tennessee       FALSE   FALSE    FALSE FALSE
## Texas           FALSE   FALSE    FALSE FALSE
## Utah            FALSE   FALSE    FALSE FALSE
## Vermont         FALSE   FALSE    FALSE FALSE
## Virginia        FALSE   FALSE    FALSE FALSE
## Washington      FALSE   FALSE    FALSE FALSE
## West Virginia   FALSE   FALSE    FALSE FALSE
## Wisconsin       FALSE   FALSE    FALSE FALSE
## Wyoming         FALSE   FALSE    FALSE FALSE
```

Since there are no missing values in the dataset we donot need to consider missing values.

The main reason to standardize the data is to remove the extreme values in a particular column. The extreme value may not necessarily be an outlier but may be intrinsically a part of the data.

We now scale the data

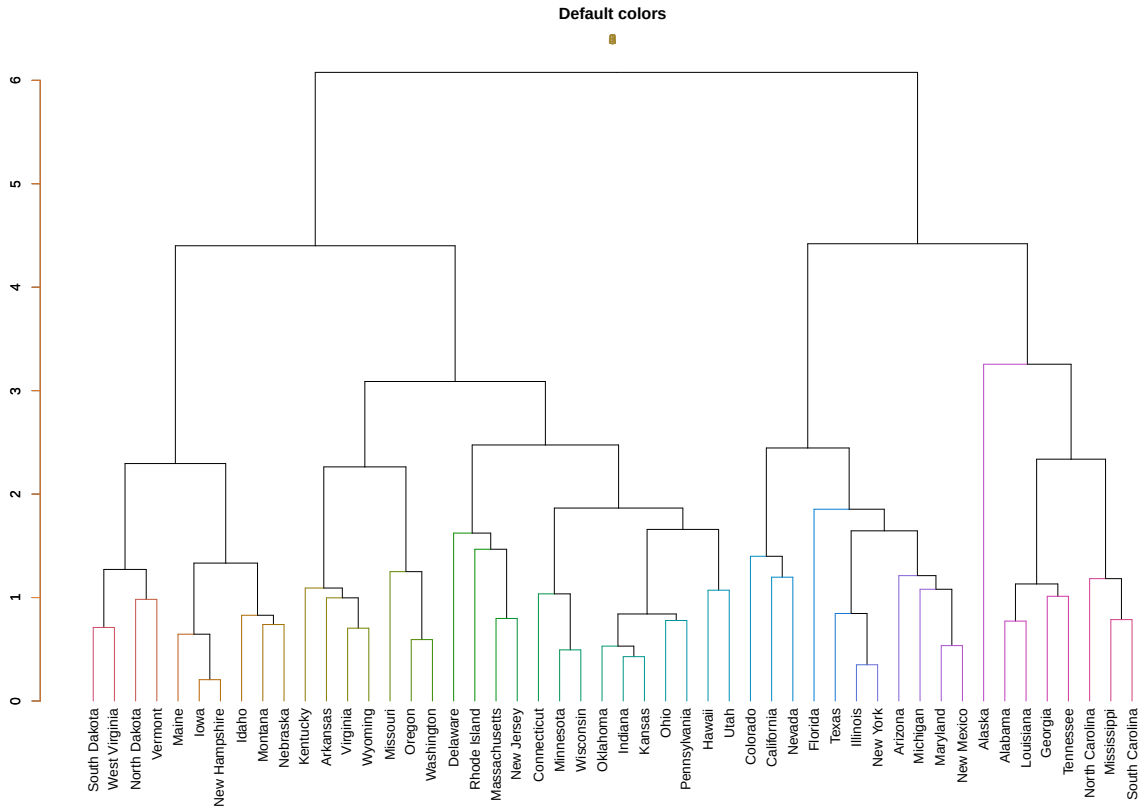
```
m <- apply(df1,2,mean)
s <- apply(df1,2,sd)
z <- scale(df1,m,s)
```

Now we apply the clustering function on this scaled dataset.

```
dist3 <- dist(z,method="euclidean")

hclust3 <- hclust(dist3,method="complete")
```

```
dend3 <- as.dendrogram(hclust3)
dend3 %>% set("branches_k_color") %>%
  plot(main = "Default colors") %>%
  axis(side = 2,col = "#F38630",labels = TRUE) %>%
  mtext(col = "#A38630")
```



As we can observe here there is a shift in the position of the states in the dendrogram cluster. Also the Y-Axis now ranges between 0-6 rather than from 0-300. The Major reason behind scaling is as follows:

If you compare states like NewYork and Washington with Alaska(Or Any Other Sparsely populated regions) we observe that even though the rate of crime might be less but it is unfair. As States like NewYork or Washington are much more densely populated and the per person crime rate is much lower than Alaska. And also in certain Datasets this disparity in the data results in Unfair or Unequal Clustering. This Issue is handled by standardizing the datasets.

We can take another Example where we are calculating the income in a particular country we observe that most of the data gets clustered in the top 10% highest earning group of the dataset even though there are very few people earning so high and most of the people earn very less but during clustering the data can get skewered. Although this didnot contain bad data necessarily but the effect of very high or low values effect the clustering mechanism.

We now check the effect of cutting the dendrogram into 3 clusters as follows

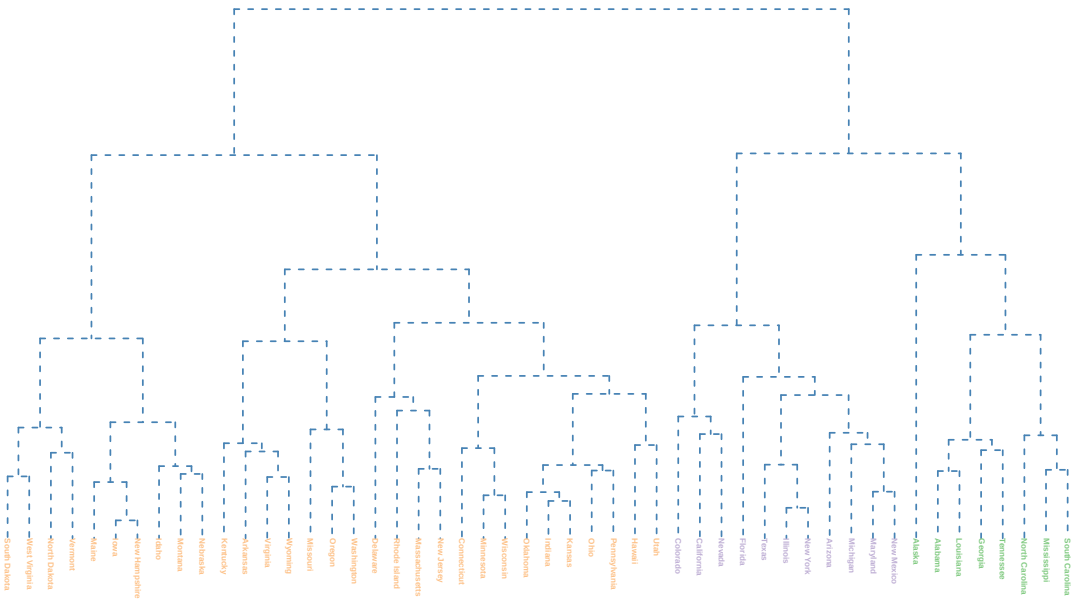
```
cutree2 <- cutree(dend3,3)
cutree2
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              2              3              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              3              2              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              2              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              2              3              1              3
```

```
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              2              3              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              1              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              3              3              3
```

We can visualize this in the form of a coloured dendrogram as follows:

```
plot(as.phylo(hclust3), type = "phylogram", cex = 0.6,
     tip.color = brewer.pal(3, 'Accent')[cutree2],
      direction = "downwards", font = 2,
     edge.color = 'steelblue', edge.width = 2, edge.lty = 2)
```



```
clust2_1 <- names(which(cutree2==1))
clust2_2 <- names(which(cutree2==2))
clust2_3 <- names(which(cutree2==3))
```

```
clust2_1
```

```
## [1] "Alabama"        "Alaska"         "Georgia"        "Louisiana"
## [5] "Mississippi"    "North Carolina" "South Carolina" "Tennessee"
```

```
clust2_2
```

```
##  [1] "Arizona"    "California" "Colorado"   "Florida"    "Illinois"
```

```
##  [6] "Maryland"   "Michigan"   "Nevada"     "New Mexico" "New York"
## [11] "Texas"
```

```
clust2_3
```

```
##  [1] "Arkansas"      "Connecticut"   "Delaware"      "Hawaii"
##  [5] "Idaho"         "Indiana"       "Iowa"          "Kansas"
##  [9] "Kentucky"      "Maine"         "Massachusetts" "Minnesota"
## [13] "Missouri"      "Montana"       "Nebraska"      "New Hampshire"
## [17] "New Jersey"    "North Dakota"  "Ohio"          "Oklahoma"
## [21] "Oregon"        "Pennsylvania"  "Rhode Island"  "South Dakota"
## [25] "Utah"          "Vermont"       "Virginia"      "Washington"
## [29] "West Virginia" "Wisconsin"     "Wyoming"
```

We observe that scaling the dataset results in shift in the number of clusters. Also there is a shift in the classification of states in various dendrograms. This occurs as there might be different units of measurements accross different variables or as the dataset might be distributed assymetrically.