CSSS 589 Homework 1

Cameron Marsden

October 15, 2016

Problem 1

To cluster schools reported on the Dreamschool Finder website, we could compare a pair of schools using Euclidean distance:

$$\delta_{i,j} = \sqrt{(x_{i,1} - x_{j,1})^2 + (x_{i,2} - x_{j,2})^2 + (x_{i,3} - x_{j,3})^2 + (x_{i,4} - x_{j,4})^2 + (x_{i,5} - x_{j,5})^2 + (x_{i,6} - x_{j,6})^2 + (x_{i,7} - x_{j,7})^2}$$

for schools i and j.

We can also place weights w_k on a set of differences $(x_{i,k} - x_{j,k})$ depending on which variables we feel best indicate a school's level of success.

For example, I believe a school's math, English, AP scores, and expenditure are most important. I also wouldn't pay much attention to a school's racial composition. So, I could use the weight vector $w = [0.2, 0.2, 0.1, 0.1, 0.2, 0.2, 0]$ to obtain a more individualized comparison of two schools. The resulting distance would be

$$\delta_{i,j} = \sqrt{0.2(x_{i,1} - x_{j,1})^2 + 0.2(x_{i,2} - x_{j,2})^2 + 0.1(x_{i,3} - x_{j,3})^2 + 0.1(x_{i,4} - x_{j,4})^2 + 0.2(x_{i,5} - x_{j,5})^2 + 0.2(x_{i,6} - x_{j,6})^2}$$

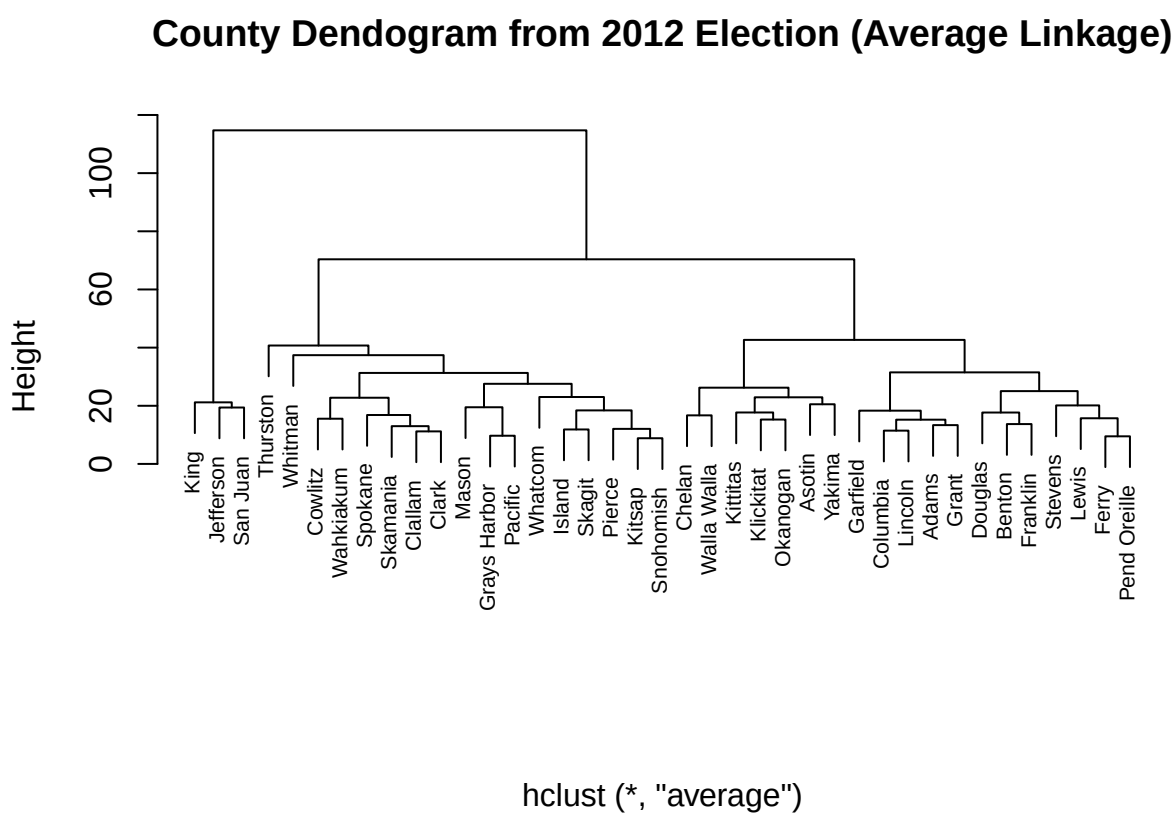
Problem 2

Part A

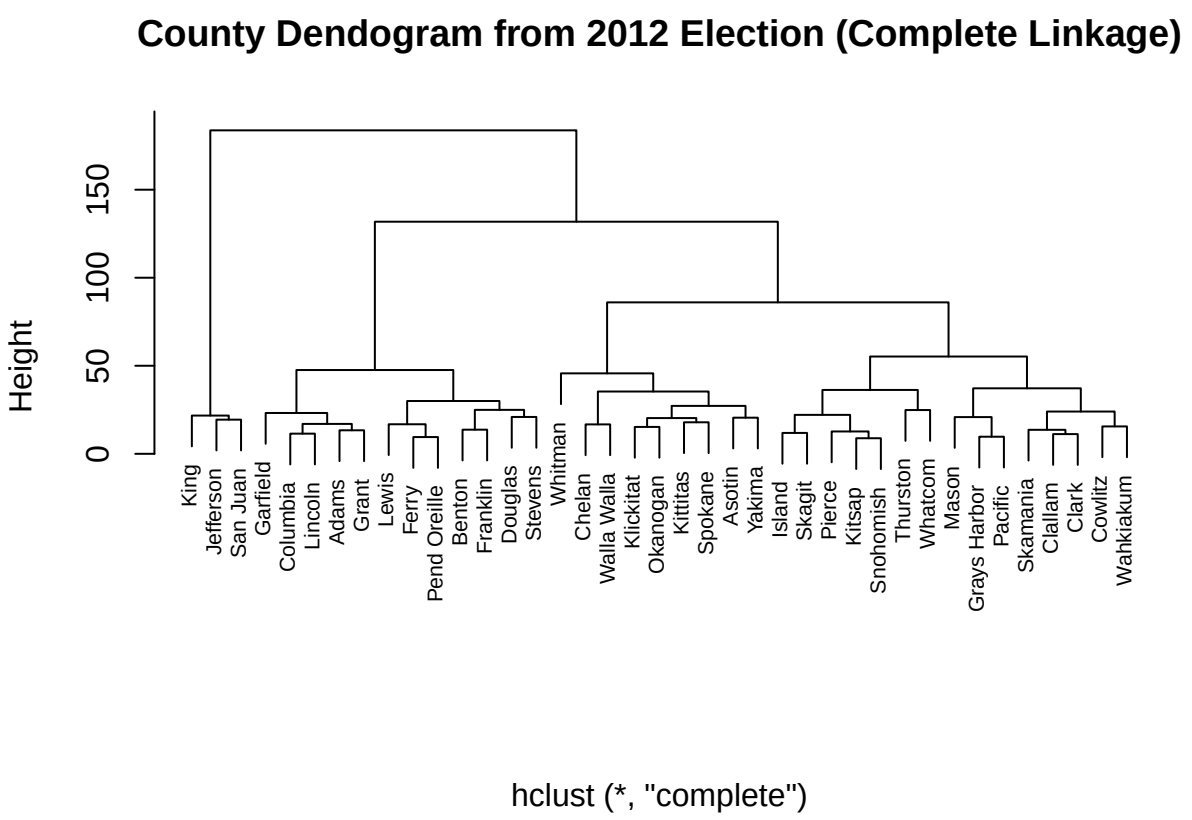
Average vs. Complete Linkage Clusters and Dendrograms

We want to compare the clusters created for the WA 2012 election data using average and complete linkage. Below are the resulting dendrograms:

```
setwd("~/University of Washington Master's Program/2016 Q3/CSSS 589/Homework/Homework 1")
library(knitr)
#Read in the data and create distance values
election2012 <- read.csv("wa_election.csv")
election.dist <- dist(election2012[,-1])
#Recreate the dendrogram/clusters from lab (average linkage)
county.tree.average <- hclust(d = election.dist, method = "average")
#Create the dendrogram/clusters with complete linkage
county.tree.complete <- hclust(d = election.dist, method = "complete")
#Create plots for both linkage types
plot(county.tree.average, labels = election2012[,1], cex = .7, xlab = "",
     main = "County Dendogram from 2012 Election (Average Linkage)")
```




```
plot(county.tree.complete, labels = election2012[,1], cex = .7, xlab = "",
     main = "County Dendogram from 2012 Election (Complete Linkage)")
```



The two dendrograms are similar in ways. For example, both have a small cluster for King, Jerfferson, and San Juan counties. If we use complete linkage, however, we can see four distinct clusters as opposed to the three we see when we use average linkage. To get a better idea of the cluster comparison, we can create a table that lists each county along with its cluster for each linkage method—setting three clusters for average linkage and four for complete linkage:

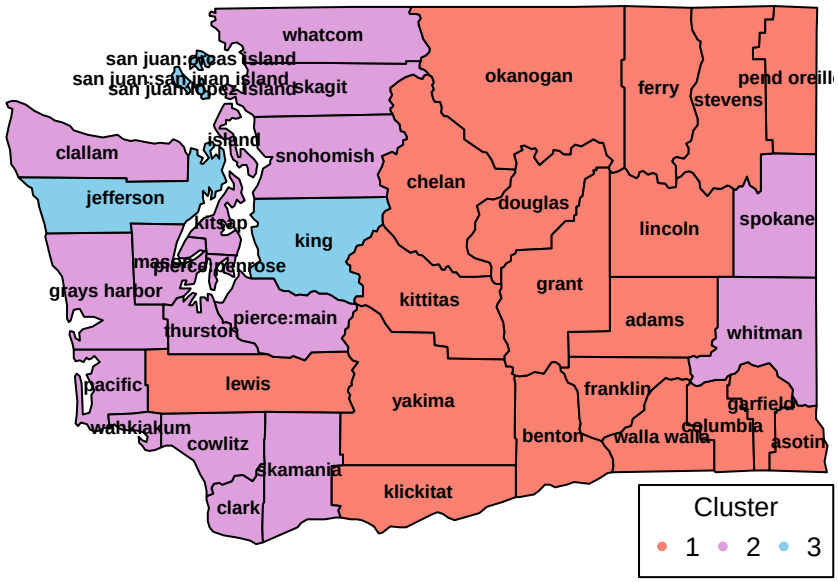
```
#Show list of counties with cluster if we choose k=3 clusters for both linkage types
county.clusters.average <- cutree(tree = county.tree.average, k=3)
county.clusters.complete <- cutree(tree = county.tree.complete, k=4)
cluster.list <- data.frame(election2012$County,
                           county.clusters.average,
                           county.clusters.complete)
colnames(cluster.list) <- c("County","Cluster (Average Linkage)","Cluster (Complete Linkage)")
kable(cluster.list)
```

County	Cluster (Average Linkage)	Cluster (Complete Linkage)
Adams	1	1
Asotin	1	2
Benton	1	1
Chelan	1	2
Clallam	2	3
Clark	2	3
Columbia	1	1
Cowlitz	2	3
Douglas	1	1

County	Cluster (Average Linkage)	Cluster (Complete Linkage)
Ferry	1	1
Franklin	1	1
Garfield	1	1
Grant	1	1
Grays Harbor	2	3
Island	2	3
Jefferson	3	4
King	3	4
Kitsap	2	3
Kittitas	1	2
Klickitat	1	2
Lewis	1	1
Lincoln	1	1
Mason	2	3
Okanogan	1	2
Pacific	2	3
Pend Oreille	1	1
Pierce	2	3
San Juan	3	4
Skagit	2	3
Skamania	2	3
Snohomish	2	3
Spokane	2	2
Stevens	1	1
Thurston	2	3
Wahkiakum	2	3
Walla Walla	1	2
Whatcom	2	3
Whitman	2	2
Yakima	1	2

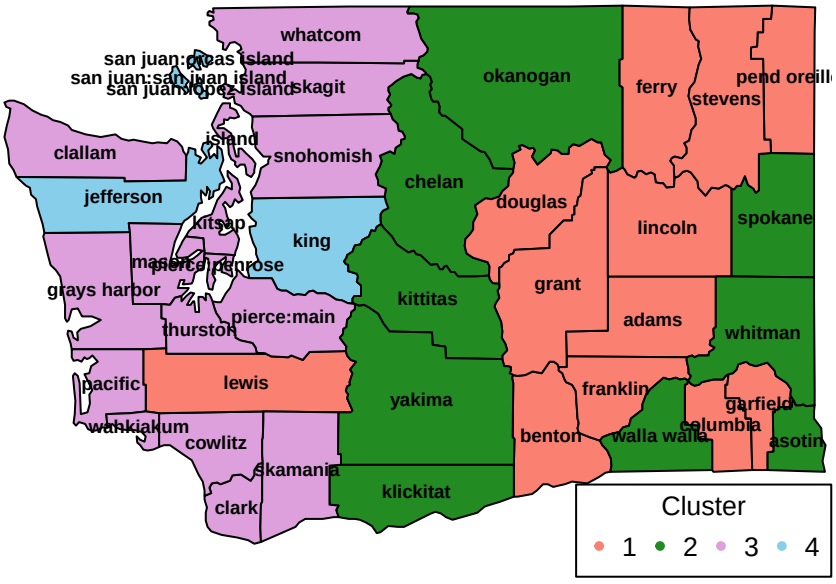
Average Linkage Map

```
library(maps)
# vector which counts pierce county twice and the san juans 3 times
clusters.for.map <- c(county.clusters.average[1:26],
                      rep(county.clusters.average[27],2),
                      rep(county.clusters.average[28],3),
                      county.clusters.average[29:39])
# colors to plot for each cluster
colors <- c("salmon", "plum", "skyblue")
# plot the counties with colors governed by the cluster
map(database = "county", region = "washington",fill = T, col = colors[clusters.for.map])
# add county name
map.text(database = "county", region = "washington", add = T, font =2, cex = .6)
# add a legend for the colors
legend(x = "bottomright", col = colors, pch = 20,
       legend = paste(c(1:3)), title = "Cluster", ncol = 3,
       cex = .8, xpd = T, inset = c(0, -.05))
```



Complete Linkage Map

```
# vector which counts pierce county twice and the san juans 3 times
clusters.for.map <- c(county.clusters.complete[1:26],
                      rep(county.clusters.complete[27],2),
                      rep(county.clusters.complete[28],3),
                      county.clusters.complete[29:39])
# colors to plot for each cluster
colors <- c("salmon", "forest green", "plum", "skyblue")
# plot the counties with colors governed by the cluster
map(database = "county", region = "washington",fill = T, col = colors[clusters.for.map])
# add county name
map.text(database = "county", region = "washington", add = T, font =2, cex = .6)
# add a legend for the colors
legend(x = "bottomright", col = colors, pch = 20,
       legend = paste(c(1:4)), title = "Cluster", ncol = 4,
       cex = .8, xpd = T, inset = c(0, -.05))
```



Cluster Map Analysis

The two maps show similar clusters for all counties in Western Washington. When we use complete linkage, however, a distinct fourth cluster splits counties between Central and Eastern Washington (except Spokane, Whitman, Walla Walla, and Asotin—which align with the Central Washington cluster).

Both maps make intuitive sense when considering the political atmosphere in Washington State. Western counties tend to lean toward liberal ideologies—with densely cities (e.g., Seattle in King County) accounting

for the largest left-leaning pull—while more rural counties in Central and Eastern Washington typically align with conservative thought. Spokane and Whitman Counties cluster together with western counties (average linkage) or central counties (complete linkage) which we expect given that Spokane is the second largest city in Washington by metropolitan population.

Which cluster method is better?

The clusters found using average and complete linkage paint very similar pictures, so either method seems valid. However, the additional cluster that the complete linkage method creates seems unnecessary.

Because complete linkage uses the farthest points between clusters A and B for its measure of distance, it is much more susceptible to outliers when compared to average linkage.

In addition, complete linkage tends to create similar-sized clusters—as evidenced by the creation of a new cluster to split central and eastern counties. We expect a large cluster of central and eastern counties to explain political ideology patterns in Washington.

Overall, average linkage provides very similar solutions to complete linkage while maintaining a simpler interpretation using three clusters instead of four.

Part B

Some political races are more important than others. It is reasonable to assume that the presidential race garners much more attention and holds the most weight when analyzing an individual's political views. During the presidential election, the majority of ad time, campaigning, rallies, and funding go toward presidential candidates rather than state positions and initiatives.

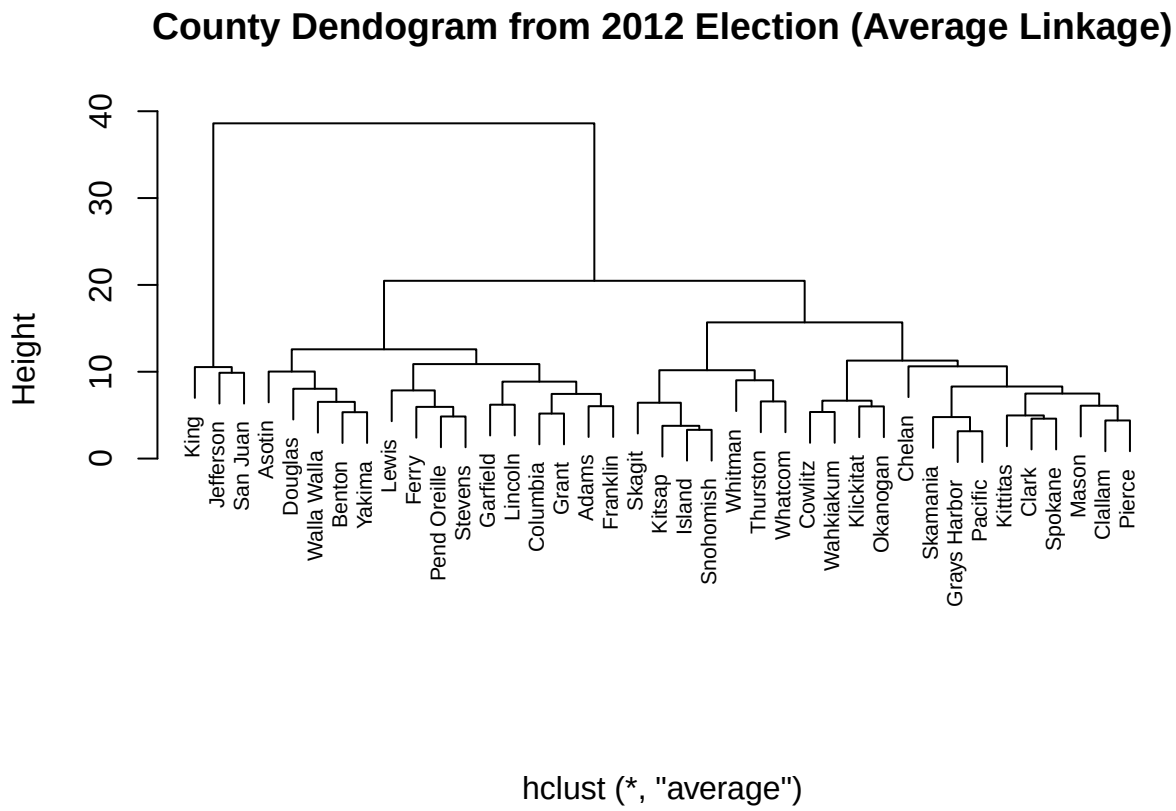
So it may be interesting to run cluster analysis on Washington counties for the 2012 election using weights. The following results replicate the analysis methods used in Part A with the following weights:

Presidential Election: 70% Other Office Elections: 20% Initiatives/Referendums/Non-Office Elections: 10%

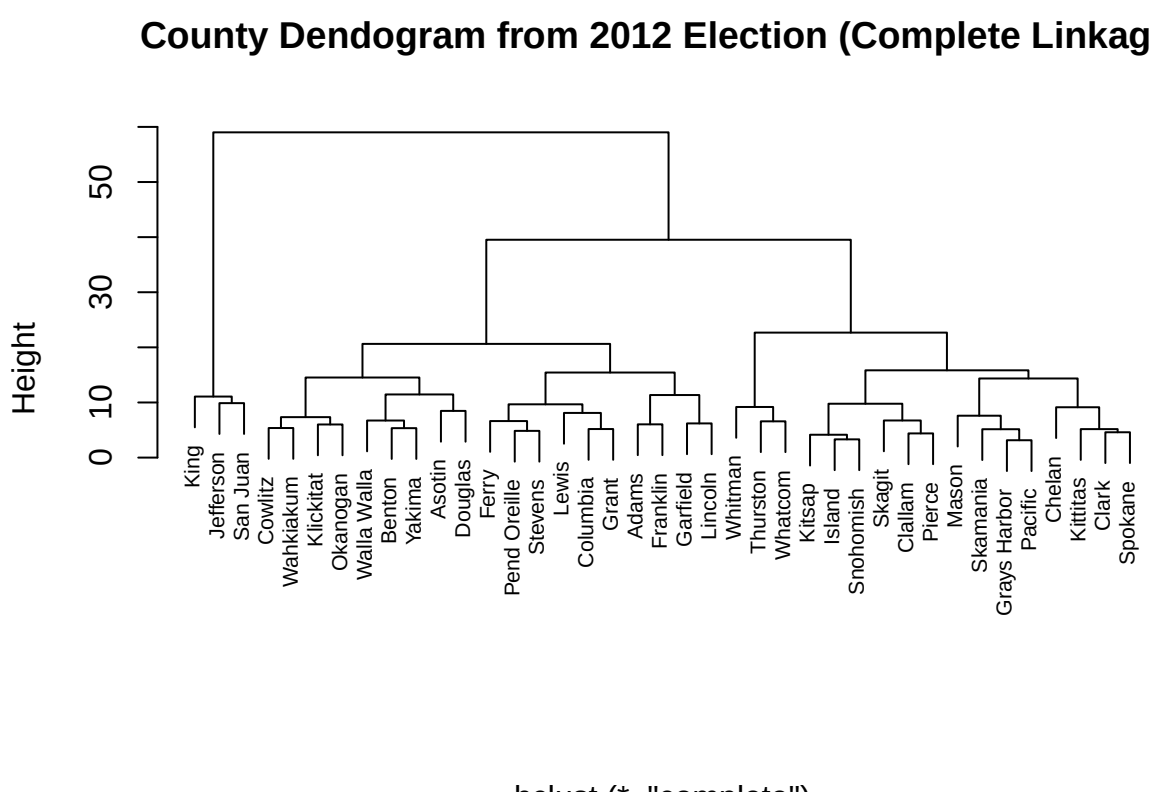
Weighted Data Clusters and Dendrograms

```
#Create weight vector
#1: Presidential race
#2: All government positions except presidential
#3: Initiatives, referendums, etc.
n.ballots <- length(election2012[1,])-1
n.1 <- 2
n.2 <- 18
n.3 <- 16
w1 <- .7/n.1
w2 <- .2/n.2
w3 <- .1/n.3
w <- c(w2,w2,w2,w2,w2,w2,w2,w2,w1,
       w1,w1,w1,w1,w1,w1,w1,w2,w2,
       w1,w1,w1,w1,w2,w2,w3,w3,w2,
       w2,w2,w2,w2,w2,w1,w1,w1,w1)
weighted.dist <- dist(t(t(election2012[,-1])*sqrt(w)))
#Perform the same analysis as part a.
#Recreate the dendrogram/clusters from lab (average linkage)
county.tree.average <- hclust(d = weighted.dist, method = "average")
#Create the dendrogram/clusters with complete linkage
county.tree.complete <- hclust(d = weighted.dist, method = "complete")
```

```r
#Create plots for both linkage types
plot(county.tree.average, labels = election2012[,1], cex = .7, xlab = "",
     main = "County Dendogram from 2012 Election (Average Linkage)")
```




```r
plot(county.tree.complete, labels = election2012[,1], cex = .7, xlab = "",
     main = "County Dendogram from 2012 Election (Complete Linkage)")
```



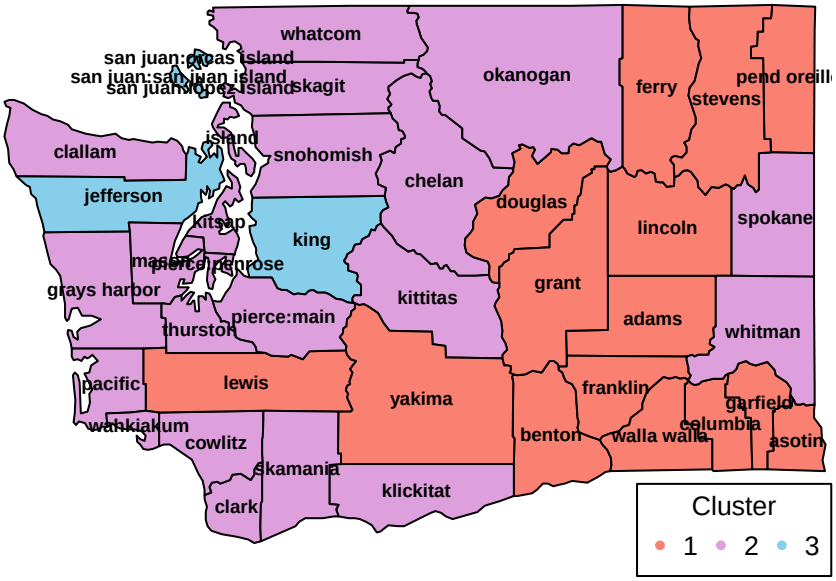

```
#Show list of counties with cluster if we choose k=3 clusters for both linkage types
county.clusters.average <- cutree(tree = county.tree.average, k=3)
county.clusters.complete <- cutree(tree = county.tree.complete, k=3)
cluster.list <- data.frame(election2012$County,county.clusters.average,county.clusters.complete)
kable(cluster.list)
```

election2012.County	county.clusters.average	county.clusters.complete
Adams	1	1
Asotin	1	1
Benton	1	1
Chelan	2	2
Clallam	2	2
Clark	2	2
Columbia	1	1
Cowlitz	2	1
Douglas	1	1
Ferry	1	1
Franklin	1	1
Garfield	1	1
Grant	1	1
Grays Harbor	2	2
Island	2	2
Jefferson	3	3
King	3	3
Kitsap	2	2

election2012.County	county.clusters.average	county.clusters.complete
Kittitas	2	2
Klickitat	2	1
Lewis	1	1
Lincoln	1	1
Mason	2	2
Okanogan	2	1
Pacific	2	2
Pend Oreille	1	1
Pierce	2	2
San Juan	3	3
Skagit	2	2
Skamania	2	2
Snohomish	2	2
Spokane	2	2
Stevens	1	1
Thurston	2	2
Wahkiakum	2	1
Walla Walla	1	1
Whatcom	2	2
Whitman	2	2
Yakima	1	1

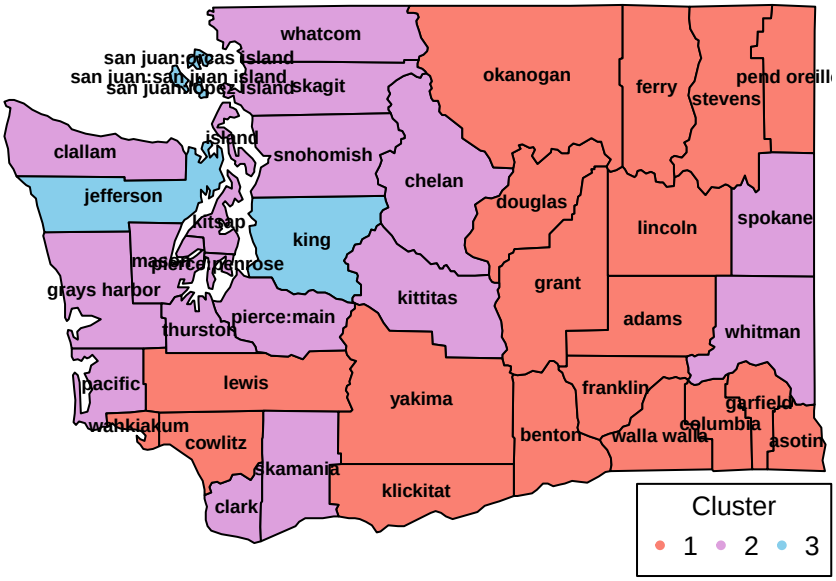
Average Linkage Map (Weighted)

```
#Create maps for both cluster types using weighted dist.
# vector which counts pierce county twice and the san juans 3 times
clusters.for.map <- c(county.clusters.average[1:26],
                      rep(county.clusters.average[27],2),
                      rep(county.clusters.average[28],3),
                      county.clusters.average[29:39])
# colors to plot for each cluster
colors <- c("salmon", "plum", "skyblue")
# plot the counties with colors governed by the cluster
map(database = "county", region = "washington",fill = T, col = colors[clusters.for.map])
# add county name
map.text(database = "county", region = "washington", add = T, font =2, cex = .6)
# add a legend for the colors
legend(x = "bottomright", col = colors, pch = 20,
       legend = paste(c(1:3)), title = "Cluster", ncol = 3,
       cex = .8, xpd = T, inset = c(0, -.05))
```



Complete Linkage Map (Weighted)

```
# vector which counts pierce county twice and the san juans 3 times
clusters.for.map <- c(county.clusters.complete[1:26],
                      rep(county.clusters.complete[27],2),
                      rep(county.clusters.complete[28],3),
                      county.clusters.complete[29:39])
# colors to plot for each cluster
colors <- c("salmon", "plum", "skyblue")
# plot the counties with colors governed by the cluster
map(database = "county", region = "washington",fill = T, col = colors[clusters.for.map])
# add county name
map.text(database = "county", region = "washington", add = T, font =2, cex = .6)
# add a legend for the colors
legend(x = "bottomright", col = colors, pch = 20,
       legend = paste(c(1:3)), title = "Cluster", ncol = 3,
       cex = .8, xpd = T, inset = c(0, -.05))
```



Running the same analysis placing heavy weights on the presidential election creates an interesting change in the complete linkage method's clusters. Instead of four distinct clusters we saw in part A, there are only three—matching with the average linkage results for both the weighted and unweighted data:

- 1: Lewis, Cowlitz, Wahkiakum, and most central and eastern counties
- 2: Spokane, Whitman, and most western counties
- 3: King, Jefferson, and San Juan counties

The only counties that diverge between the complete and average linkage methods lie in the southwest part of the state: Klickitat, Cowlitz, and Wahkiakum counties.

Between the weighted and unweighted data, we see a few central counties (e.g., Chelan and Kittitas) align more with western counties when we place greater weight on the presidential race. Otherwise, the clusters produced using weighted and unweighted data are very similar. Once again, we see a strong pattern of coastal counties vs. central/eastern counties with the exception of Spokane and Whitman. In addition, we see King, Jefferson, and San Juan counties creating their own cluster.

Part C

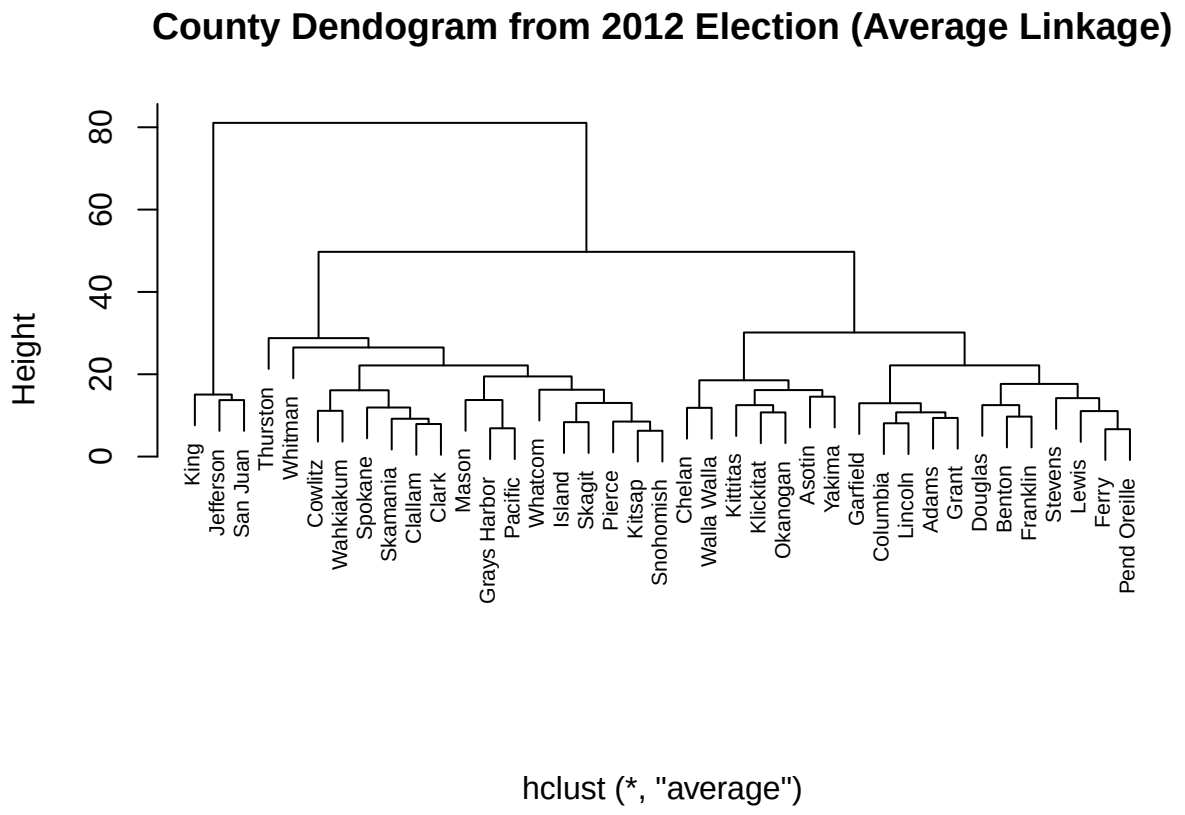
If two variables explaining outcome y are perfectly correlated (i.e., r = -1 or 1), we will not see any change in the resulting clusters if we only used one variable. Intuitively, if one column of our data mimics another, they become redundant.

When testing one column of data vs. two columns of (perfect) negative correlation, the distances used to differentiate clusters all increase linearly—meaning the difference between any two distances remains constant.

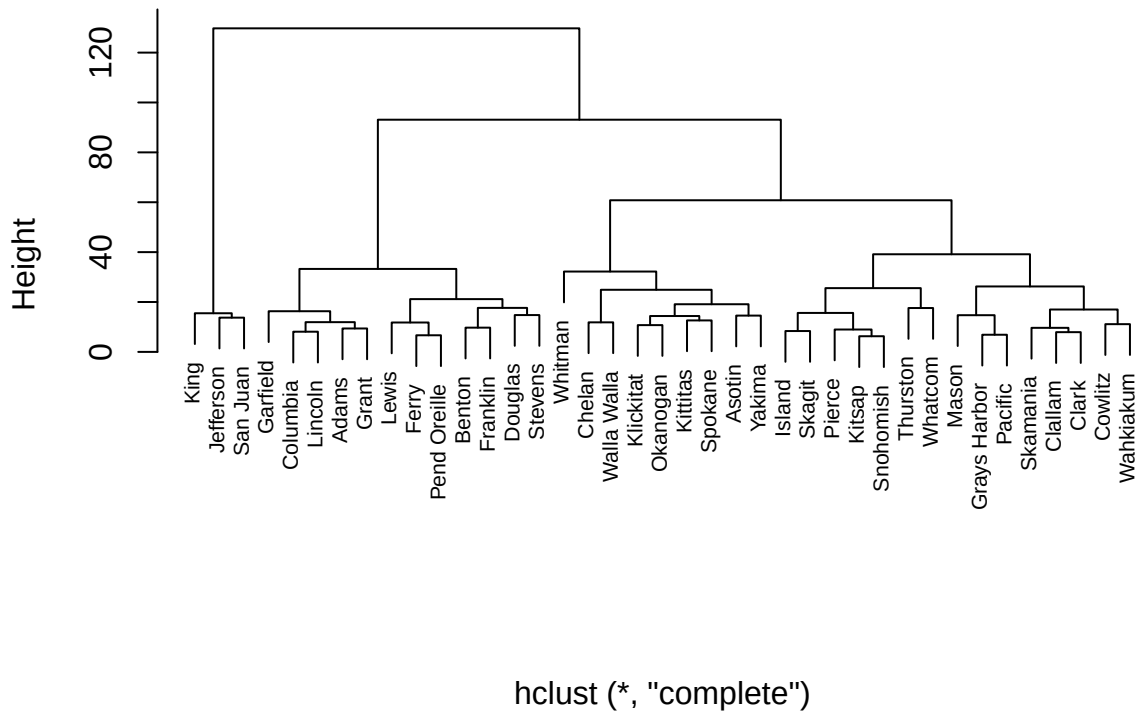
With the election data, we expect that a vote for one candidate likely indicates a vote against the other (assuming there are only two candidates). Since we expect a highly negative correlation between a set of election race results, it is reasonable to exclude every other column in our data for cluster analysis.

Running the same analysis in part A with the reduced dataset produces the following results:

```
election2012.reduced <- election2012[,c(1,seq(from=2,to=37,by=2))]
election.dist <- dist(election2012.reduced[,-1])
#Recreate the dendrogram/clusters from lab (average linkage)
county.tree.average <- hclust(d = election.dist, method = "average")
#Create the dendrogram/clusters with complete linkage
county.tree.complete <- hclust(d = election.dist, method = "complete")
#Create plots for both linkage types
plot(county.tree.average, labels = election2012[,1], cex = .7, xlab = "",
     main = "County Dendogram from 2012 Election (Average Linkage)")
```




```
plot(county.tree.complete, labels = election2012[,1], cex = .7, xlab = "",
     main = "County Dendogram from 2012 Election (Complete Linkage)")
```



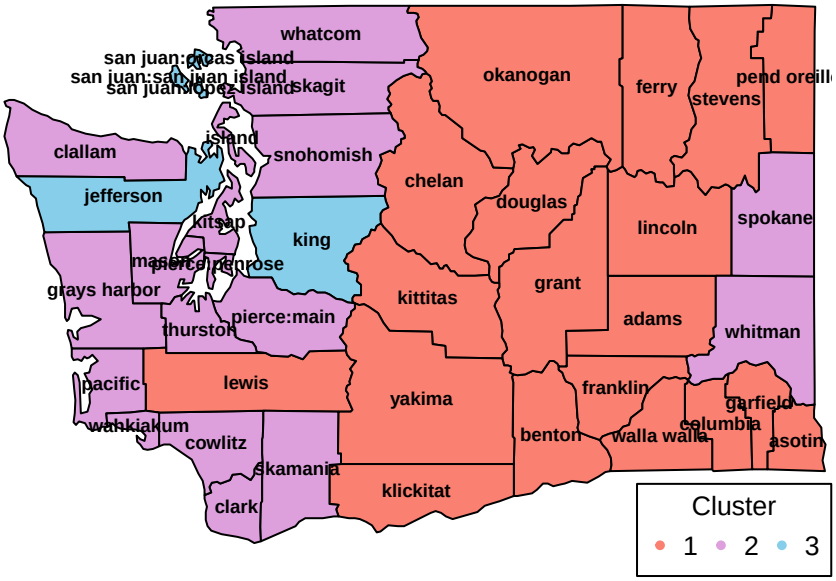

```
#Show list of counties with cluster if we choose k=3 clusters for both linkage types
county.clusters.average <- cutree(tree = county.tree.average, k=3)
county.clusters.complete <- cutree(tree = county.tree.complete, k=4)
cluster.list <- data.frame(election2012$County,
                           county.clusters.average,
                           county.clusters.complete)
colnames(cluster.list) <- c("County","Cluster (Average Linkage)","Cluster (Complete Linkage)")
kable(cluster.list)
```

County	Cluster (Average Linkage)	Cluster (Complete Linkage)
Adams	1	1
Asotin	1	2
Benton	1	1
Chelan	1	2
Clallam	2	3
Clark	2	3
Columbia	1	1
Cowlitz	2	3
Douglas	1	1
Ferry	1	1
Franklin	1	1
Garfield	1	1
Grant	1	1
Grays Harbor	2	3
Island	2	3

County	Cluster (Average Linkage)	Cluster (Complete Linkage)
Jefferson	3	4
King	3	4
Kitsap	2	3
Kittitas	1	2
Klickitat	1	2
Lewis	1	1
Lincoln	1	1
Mason	2	3
Okanogan	1	2
Pacific	2	3
Pend Oreille	1	1
Pierce	2	3
San Juan	3	4
Skagit	2	3
Skamania	2	3
Snohomish	2	3
Spokane	2	2
Stevens	1	1
Thurston	2	3
Wahkiakum	2	3
Walla Walla	1	2
Whatcom	2	3
Whitman	2	2
Yakima	1	2

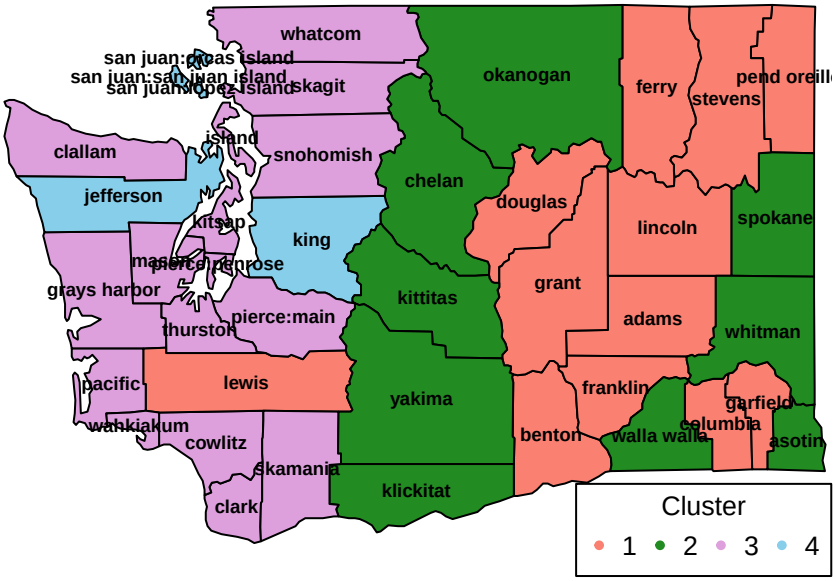
Average Linkage Map (Unweighted Reduced)

```
# vector which counts pierce county twice and the san juans 3 times
clusters.for.map <- c(county.clusters.average[1:26],
                      rep(county.clusters.average[27],2),
                      rep(county.clusters.average[28],3),
                      county.clusters.average[29:39])
# colors to plot for each cluster
colors <- c("salmon", "plum", "skyblue")
# plot the counties with colors governed by the cluster
map(database = "county", region = "washington",fill = T, col = colors[clusters.for.map])
# add county name
map.text(database = "county", region = "washington", add = T, font =2, cex = .6)
# add a legend for the colors
legend(x = "bottomright", col = colors, pch = 20,
       legend = paste(c(1:3)), title = "Cluster", ncol = 3,
       cex = .8, xpd = T, inset = c(0, -.05))
```



Complete Linkage Map (Unweighted Reduced)

```
# vector which counts pierce county twice and the san juans 3 times
clusters.for.map <- c(county.clusters.complete[1:26],
                      rep(county.clusters.complete[27],2),
                      rep(county.clusters.complete[28],3),
                      county.clusters.complete[29:39])
# colors to plot for each cluster
colors <- c("salmon", "forest green", "plum", "skyblue")
# plot the counties with colors governed by the cluster
map(database = "county", region = "washington",fill = T, col = colors[clusters.for.map])
# add county name
map.text(database = "county", region = "washington", add = T, font =2, cex = .6)
# add a legend for the colors
legend(x = "bottomright", col = colors, pch = 20,
       legend = paste(c(1:4)), title = "Cluster", ncol = 4,
       cex = .8, xpd = T, inset = c(0, -.05))
```



As expected, the highly correlated columns used in part A are redundant; excluding them before running the analysis produces the exact same clusters for both average and complete linkage methods.

Problem 3

Part A

Euclidian distance (D) is defined as the square root of the sum of squared distances. In our case,

$$D(x_i, y_i) = \sqrt{\sum_{i=1}^{5}(x_i - y_i)^2}$$

Part B

For a match, $(x_i,y_i) = (0,0)$ or $(1,1)$, so its component for the Euclidian distance is $(x_i - y_i)^2=0$.

Conversely, a mismatch occurs when $(x_i,y_i) = (0,1)$ or $(1,0)$, so its component for the Euclidian distance is $(x_i - y_i)^2=1$.

So the Euclidian distance can be rewritten

$$\begin{aligned} D(x_i, y_i) &= \sqrt{\sum_{i=1}^{5} 1_{[x_i \neq y_i]}} \\ &= \sqrt{b + c} \end{aligned}$$

where b+c is the number of mismatches.

$$\begin{aligned} a+b+c+d &= p \\ \frac{a+b+c+d}{p} &= 1 \\ \frac{a+d}{p} + \frac{b+c}{p} &= 1 \end{aligned}$$

Let $\frac{a+d}{p}$=r so

$$\begin{aligned} r + \frac{b+c}{p} &= 1 \\ \frac{b+c}{p} &= 1-r \\ b+c &= p(1-r) \end{aligned}$$

So we have proven that $D(x_i, y_i) = \sqrt{b+c} = \sqrt{p(1-r)}$.