Question #2

Perform cluster analysis on 2012-2016 presidential elections database and summarize findings.

Preparing Data

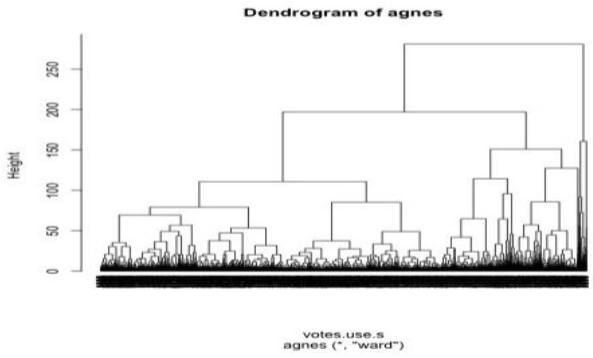
1. Check missing value, there's **no missing value**.
2. Delete fips and non-numeric variables.
3. **Standardizing** each variable since they are in different scales.

Implement Hierarchical Clustering in R

1. Perform **agglomerative HC** with **agnes()** function with four different hierarchical clustering methods. Get the agglomerative coefficient and the agglomerative coefficient of Ward's methods is biggest(ac=0.987767). **Ward's method** has the strongest clustering structure of the four methods used.

```
> hc1<-agnes(votes.use.s,method="complete")
> hc1$ac
[1] 0.974206
> hc2<-agnes(votes.use.s,method="average")
> hc2$ac
[1] 0.9635848
> hc3<-agnes(votes.use.s,method="single")
> hc3$ac
[1] 0.9420727
> hc4<-agnes(votes.use.s,method="ward")
> hc4$ac
[1] 0.987767
```

2. Look at **dendrogram**, each leaf corresponds to one observation(county). **The number of clusters may be 2, 3 and 4**.



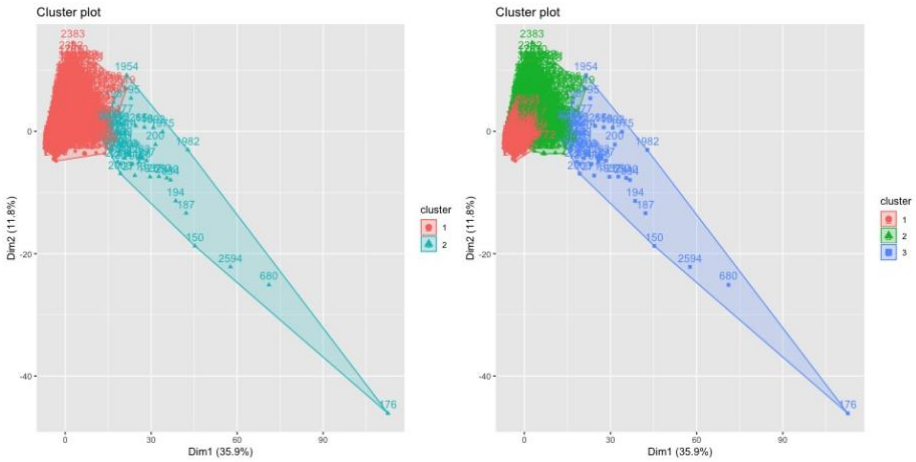
3. Cut tree into 2 to 20 groups and can get the number of observations(counties) in each groups(cluster). **For the two and three cluster solution, the distribution among the clusters looks good** (don't want too many clusters with just a few observations).

2 Cluster	
Cluster 1	Cluster 2
3058	54

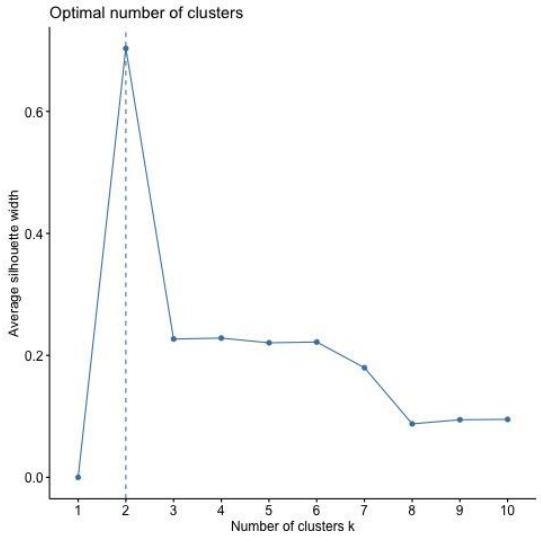
3 Cluster		
Cluster 1	Cluster 2	Cluster3
2178	880	54

4 Cluster			
Cluster 1	Cluster 2	Cluster 3	Cluster 4
2178	880	52	2

4. **Visualize** 2 cluster and 3 cluster in scatter plots. **Scatter plot of 2 cluster looks better** than 3 cluster. There's more overlapping in 3 cluster.



5. In order to **determine optimal cluster**, perform the **average silhouette method**. 2 clusters maximize the average silhouette values and **2 cluster is good** since high average silhouette width indicates a good clustering.



6. **Perform summary statistics with aggregate() function** in order to **see the characteristic of two clusters**. Table below shows some data and more data are in 'median cluster2.output.txt'.

   Cluster 1 contains counties that mostly votes Trump and Romney (GOP). These counties have **higher percent of white people and homeownership rate** than counties in cluster 2. Maybe white people more like Trump and Romney or GOP.

   Cluster 2 contains counties that mostly votes Clinton and Obama (Democratic Party). These counties have **larger and younger population with more racial diversity**. People in these counties mostly are **high-educated and have high income**. There may have some positive relationship between education level and income. Black people will prefer vote Obama. People who have different background and are high-educated with high income may prefer vote Clinton and Obama or Democratic Party. Make sense.

Variable name	Cluster 1	Cluster2
Clinton	-0.2278572	1.9195197
Trump	0.2206224	-1.8679012
Obama	-0.1202091	1.4315307
Romney	0.09965408	-1.41254066
Population2014	-0.2343346	3.8947393
White	0.4370128	-0.9564132
…		
…		