STAT 351 Final

Charles Hwang

4/30/2020

Charles Hwang

Professor Matthews

STAT 351-001

2 May 2020

Problem 1

Parametric statistics assume that data are derived from a distribution with defined parameters while nonparametric statistics do not. On the same note, parametric statistics also carry the assumption of normality in the data while nonparametric statistics do not which can be helpful for datasets with small sample sizes. It may be easier to conduct a statistical test with nonparametric statistics because there are fewer and more relaxed assumptions which makes it harder for them to be violated.

Problem 2

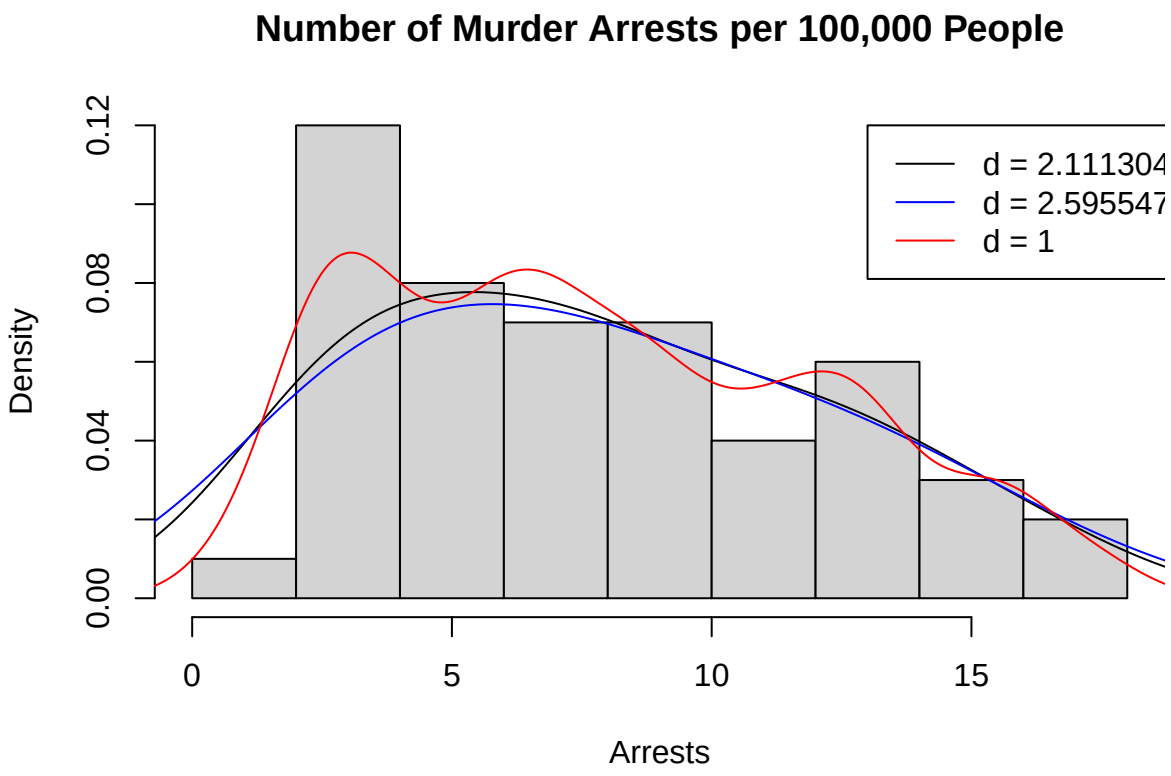
ANOVA can be done with both parametric and nonparametric tests. In parametric statistics, the traditional one-way ANOVA with null hypothesis "H0: m1 = m2 = ... = mk" (where m is the mean of the group and k is the number of groups being compared) and alternative hypothesis "HA: At least one mi is different" can be used. In nonparametric statistics, the equivalent Kruskal–Wallis rank-sum test with null hypothesis "H0: m1 = m2 = ... = mk" (where m is the median of the group and k is the number of groups being compared) and alternative hypothesis "HA: At least one mi is different" are used.

Problem 3

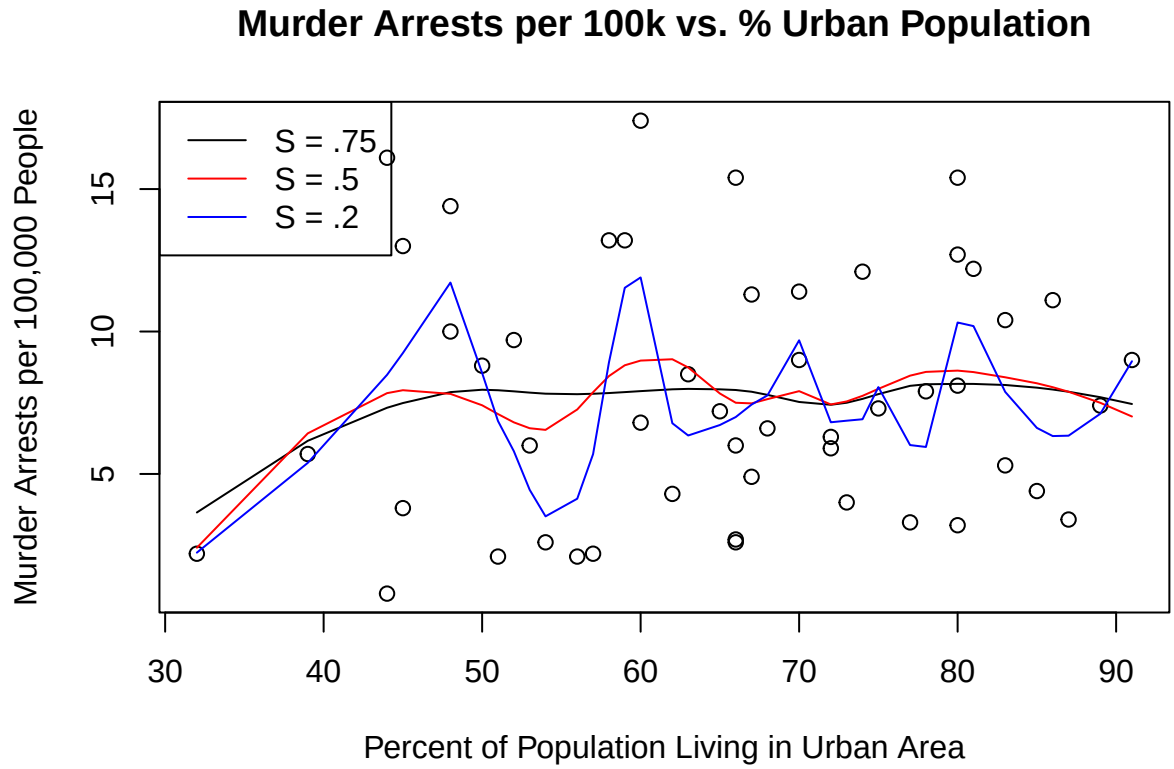
A CART model is a single tree that acts as its own model and is usually pruned to avoid overfitting the data, while a random forest uses several trees, making the model less biased and more predictive. Multiple variables can be used or "tried" at each split of a random forest while each split in a CART model generally only evaluates a single variable at a time.

Problem 4

```
hist(USArrests$Murder,freq=FALSE,main="Number of Murder Arrests per 100,000 People",xlab="Arrests")  # A
legend(13,.12,c("d = 2.111304", "d = 2.595547","d = 1"),lwd="1",col=c("black","blue","red"))
points(density(USArrests$Murder,bw=1.06*sd(USArrests$Murder)/length(USArrests$Murder)^.2)$x,density(USAr
points(density(USArrests$Murder,bw=1.06*IQR(USArrests$Murder)/(1.34*length(USArrests$Murder)^.2))$x,dens
points(density(USArrests$Murder,bw=1)$x,density(USArrests$Murder,bw=1)$y,type="l",col="red") # In my op
```




```
plot(USArrests$Murder~USArrests$UrbanPop,main="Murder Arrests per 100k vs. % Urban Population",xlab="Pe
legend("topleft",c("S = .75","S = .5","S = .2"),lwd="1",col=c("black","red","blue"))
lines(sort(USArrests$UrbanPop),fitted(loess(USArrests$Murder~USArrests$UrbanPop))[order(USArrests$Urban
lines(sort(USArrests$UrbanPop),fitted(loess(USArrests$Murder~USArrests$UrbanPop,span=.5))[order(USArres
lines(sort(USArrests$UrbanPop),fitted(loess(USArrests$Murder~USArrests$UrbanPop,span=.2))[order(USArres
```

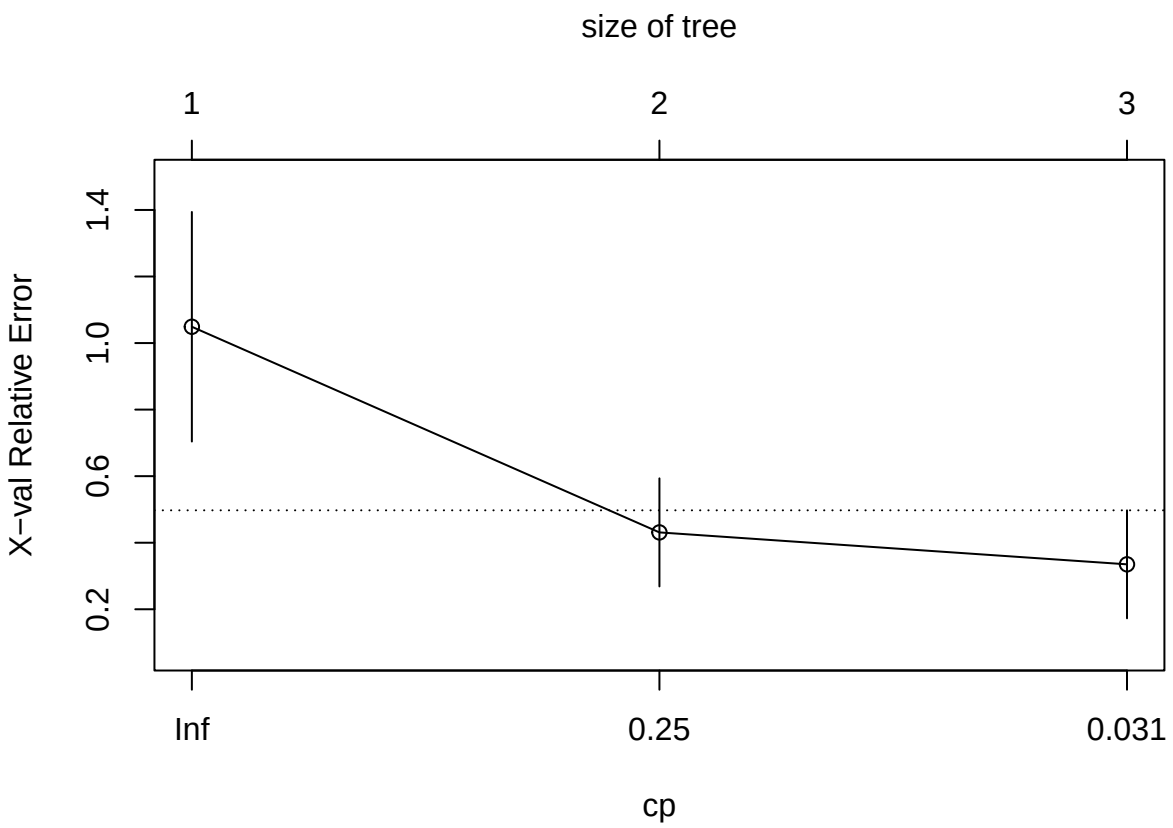


Problem 5

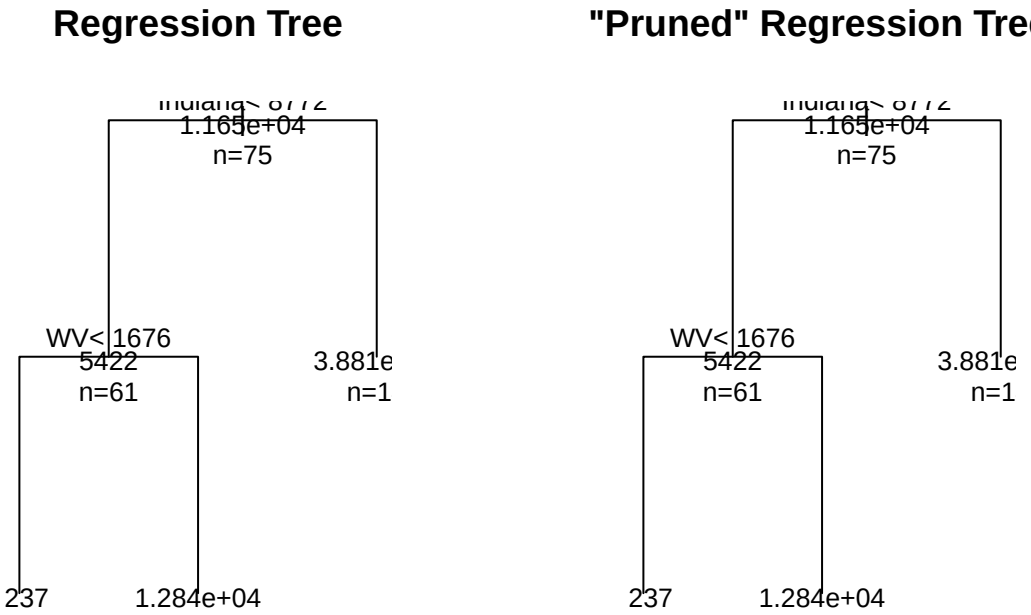
```
library(dslabs)
library(randomForest)
library(reshape2)
library(rpart)
library(tree)
measles <- subset(us_contagious_diseases,disease=="Measles")
measles <- melt(measles,id=c("disease","state","year"),measure="count")
measles <- dcast(measles,formula=year~state)
names(measles)[10] <- "DC"
names(measles)[31] <- "NH"
names(measles)[32] <- "NJ"
names(measles)[33] <- "NM"
names(measles)[34] <- "NY"
names(measles)[35] <- "NC"
names(measles)[36] <- "ND"
names(measles)[41] <- "RI"
names(measles)[42] <- "SC"
names(measles)[43] <- "SD"
names(measles)[50] <- "WV"
rownames(measles) <- measles$year # Assigning "year" column as index of dataset
measles$year <- NULL
set.seed(2025)                                                    # Problem 5a
tree <- rpart(Illinois~.,data=measles)
printcp(tree)
```

```
##
## Regression tree:
## rpart(formula = Illinois ~ ., data = measles)
##
## Variables actually used in tree construction:
## [1] Indiana WV
##
## Root node error: 1.9881e+10/75 = 265086642
##
## n= 75
##
##         CP nsplit rel error  xerror    xstd
## 1 0.638597      0   1.00000 1.04896 0.34497
## 2 0.095253      1   0.36140 0.43101 0.16241
## 3 0.010000      2   0.26615 0.33506 0.16227
```

```
plotcp(tree)
```




```
par(mfrow=c(1,2))
plot(tree,uniform=TRUE,main="Regression Tree")
text(tree,use.n=TRUE,all=TRUE,cex=.8)
ptree <- prune(tree,cp=tree$cptable[which.min(tree$cptable[,"xerror"]),"CP"])
plot(ptree,uniform=TRUE,main="''Pruned'' Regression Tree") # There appears to be no difference between
text(ptree,use.n=TRUE,all=TRUE,cex=.8)
```



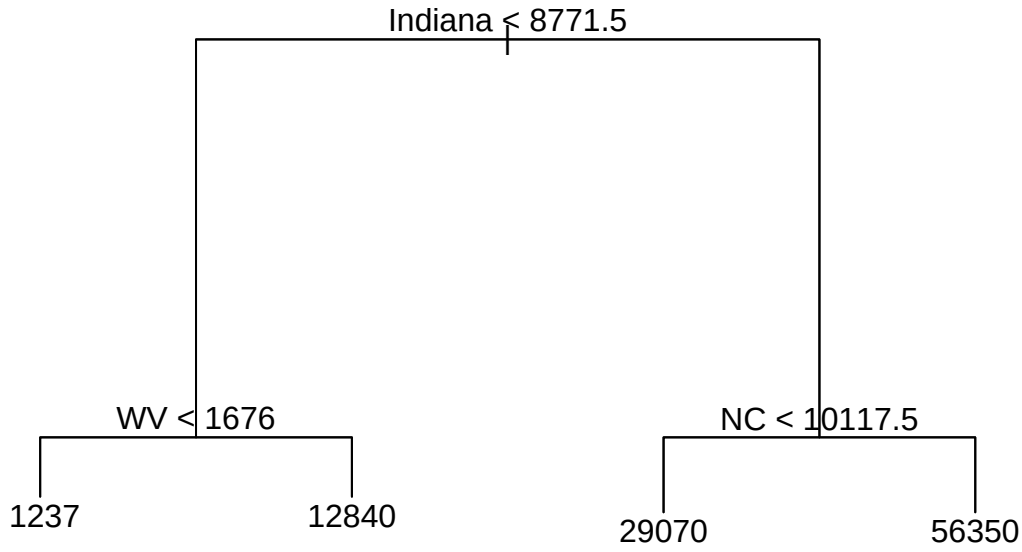

```
par(mfrow=c(1,1))
cat("Cross validation error:",min(tree$cptable[,"xerror"]))
```

```
## Cross validation error: 0.3350563
```

```
T <- tree(Illinois~.,data=measles) # Growing different tree
T
```

```
## node), split, n, deviance, yval
##       * denotes terminal node
##
## 1) root 75 1.988e+10 11650
##   2) Indiana < 8771.5 61 2.566e+09  5422
##     4) WV < 1676 39 1.258e+08  1237 *
##     5) WV > 1676 22 5.460e+08 12840 *
##   3) Indiana > 8771.5 14 4.620e+09 38810
##     6) NC < 10117.5 9 6.207e+08 29070 *
##     7) NC > 10117.5 5 1.607e+09 56350 *
```

```
plot(T) # This tree has four terminal nodes instead of three which is slightly better.
text(T) # Both CART models make their first split on the number of Indiana measles cases. This makes se
```

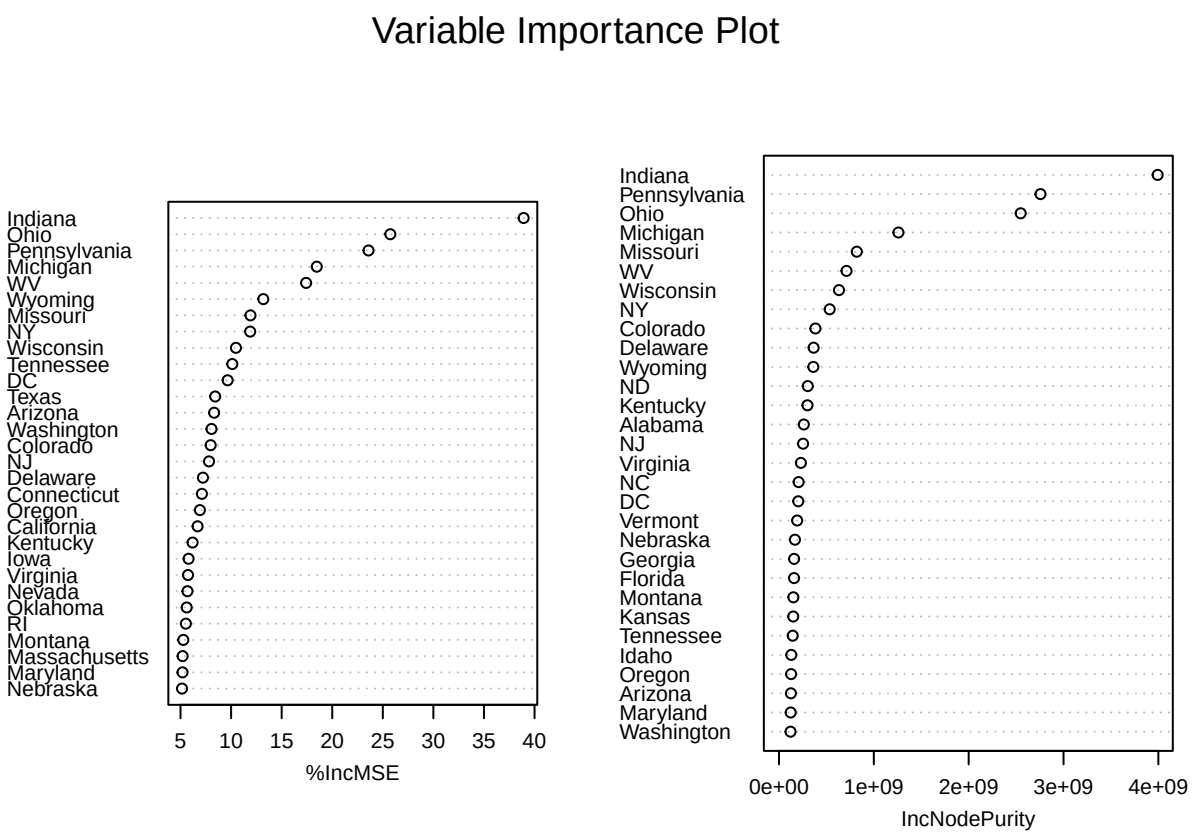



```
set.seed(2025)                                                    # Problem 5b
rf <- randomForest(Illinois~.,data=measles,ntree=5000,importance=TRUE)
predict(rf,measles)
```

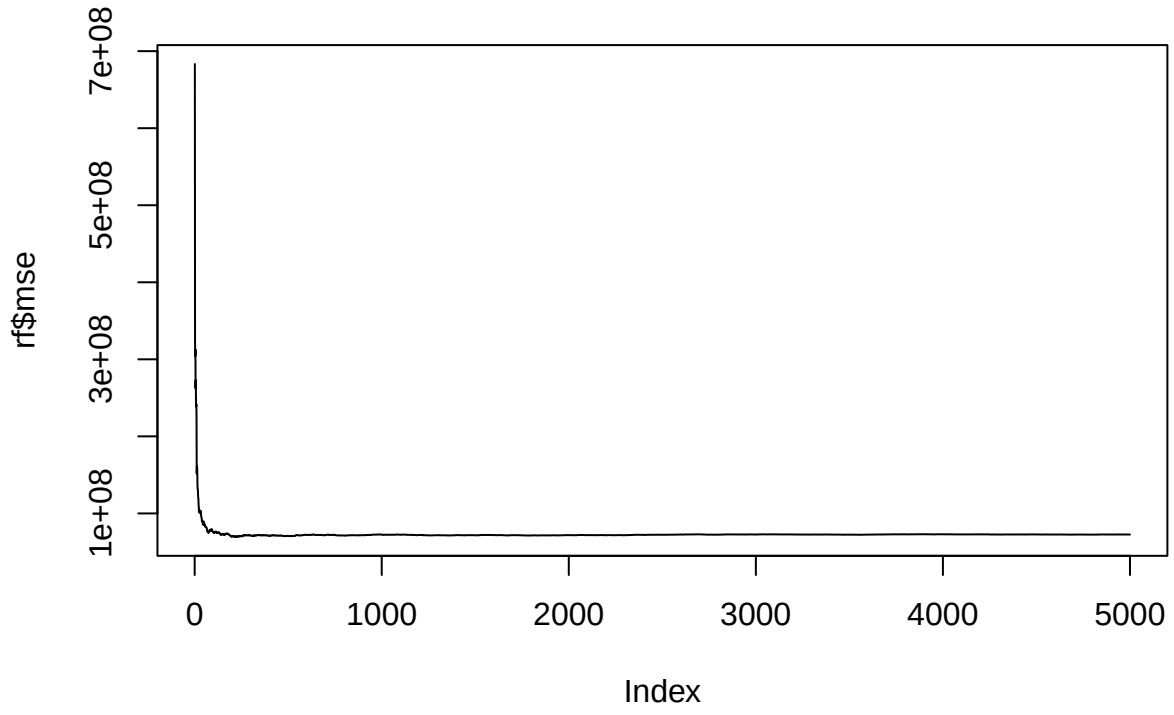
```
##         1928         1929         1930         1931         1932         1933
## 14280.928230 33360.962947 16355.041657 36745.317703 17095.106333 12977.414810
##         1934         1935         1936         1937         1938         1939
## 44104.534417 50627.773887  5071.884557 12566.610690 64321.049957  9921.166370
##         1940         1941         1942         1943         1944         1945
##  8148.318740 48880.648240 12155.135713 23783.728980 19353.390993  6949.133533
##         1946         1947         1948         1949         1950         1951
## 25631.573313  9133.747133 32256.259020 13473.500610 16171.772863 16633.009490
##         1952         1953         1954         1955         1956         1957
## 22128.796183 16695.842630 31330.606097 14159.745380 38363.944737 12095.077337
##         1958         1959         1960         1961         1962         1963
## 27105.585413 12011.947667 22199.250600 16391.660607 16082.620010 12755.587887
##         1964         1965         1966         1967         1968         1969
## 20264.941087  8153.943270 10530.921017  2964.220827  1543.968150  1432.587790
```

```
##         1970         1971         1972         1973         1974         1975
##  3362.591827  4043.602557  3342.277660  2220.970320  2069.683720  2007.767813
##         1976         1977         1978         1979         1980         1981
##  2711.438553  2785.367463  1541.137823  1382.965163   700.505703    74.799307
##         1982         1983         1984         1985         1986         1987
##    85.793067   234.869283   199.376320   322.677173   456.938427   298.034363
##         1988         1989         1990         1991         1992         1993
##   271.360657  1887.937633  1227.922793   277.989740   170.895247     8.210823
##         1994         1995         1996         1997         1998         1999
##    47.929030    20.284947    26.461080   104.280607     4.333717     5.835357
##         2000         2001         2002
##     6.882827     2.948040     6.995873
```

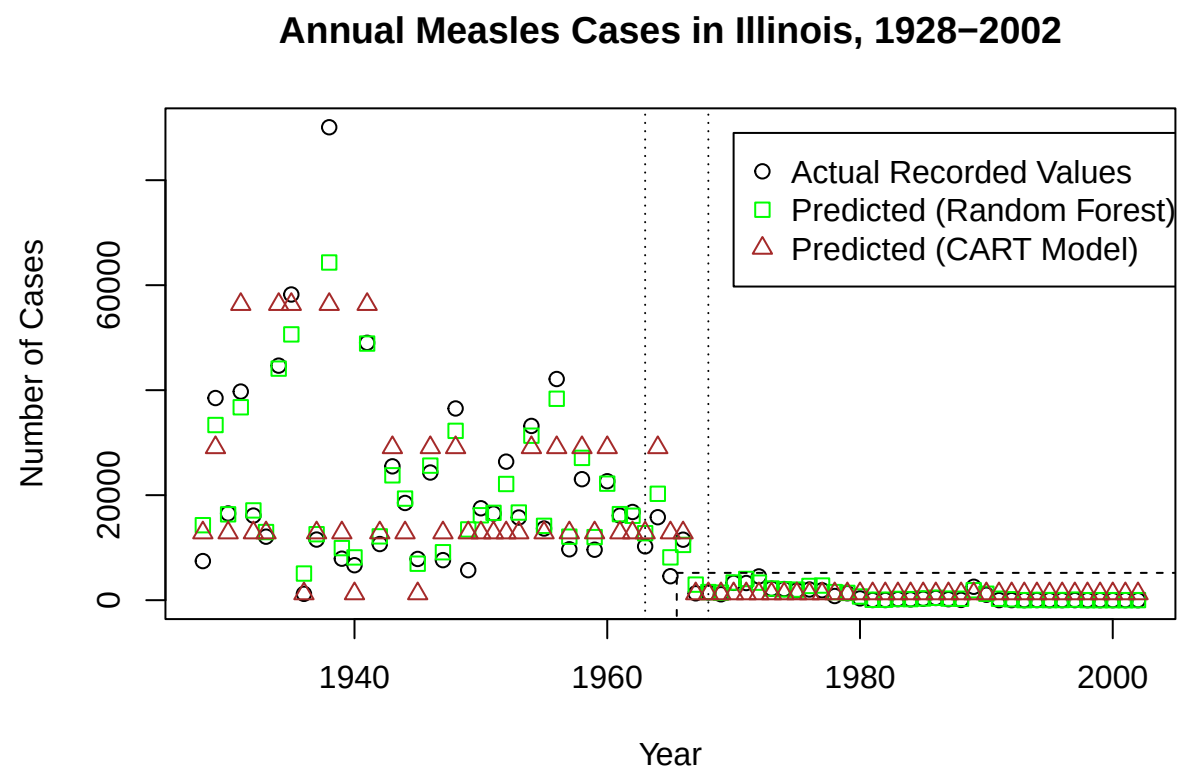
```
varImpPlot(rf,main="Variable Importance Plot",cex=.7) # Adding some graphs to visualize the random fore
```



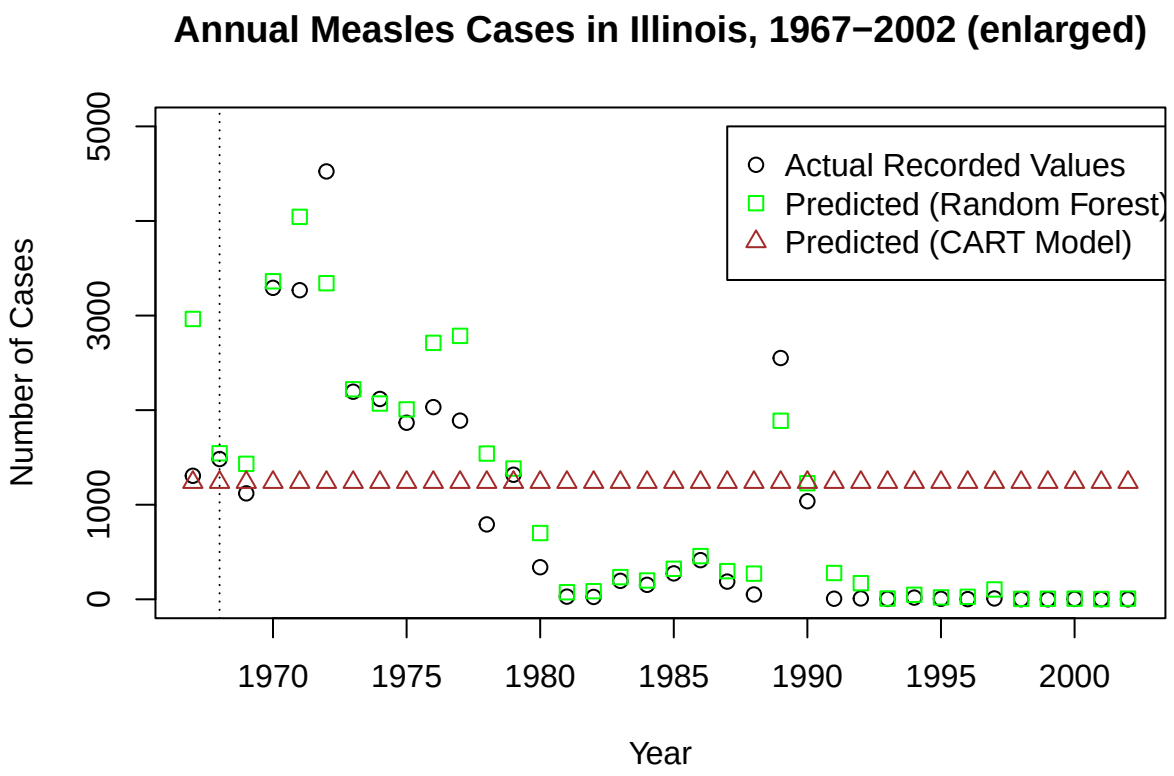

```
plot(rf$mse,type="l")
```




```
plot(rownames(measles),measles$Illinois,main="Annual Measles Cases in Illinois, 1928-2002",xlab="Year",y
points(rownames(measles),predict(rf,measles),pch=0,col="green")
points(rownames(measles),predict(T),pch=2,col="brown") # Comparing random forest to CART model and actu
abline(v=c(1963,1968),lty=3) # The measles vaccine was released in the United States in 1963 and furthe
segments(1965.5,c(-3000,5200),c(1965.5,2009),5200,lty=2) # Window of enlarged plot
legend(1970,89000,c("Actual Recorded Values","Predicted (Random Forest)","Predicted (CART Model)"),col=
```



```
plot(rownames(measles)[40:nrow(measles)],measles$Illinois[40:nrow(measles)],ylim=c(0,5000),main="Annual
points(rownames(measles)[40:nrow(measles)],predict(rf,measles[40:nrow(measles),]),pch=0,col="green")
points(rownames(measles)[40:nrow(measles)],predict(T)[40:nrow(measles)],pch=2,col="brown") # Comparing
abline(v=1968,lty=3)
legend(1987,5000,c("Actual Recorded Values","Predicted (Random Forest)","Predicted (CART Model)"),col=c
```




```
# The random forest appears to be better than the CART model at predicting annual measles cases in Illi
cor.test(measles$Illinois,measles$Massachusetts,method="kendall") # Problem 5c
```

```
##
##  Kendall's rank correlation tau
##
## data:  measles$Illinois and measles$Massachusetts
## z = 8.3992, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.6626963
```

```
cat("Kendall's t =",cor(measles$Illinois,measles$Massachusetts,method="kendall"))
```

```
## Kendall's t = 0.6626963
```

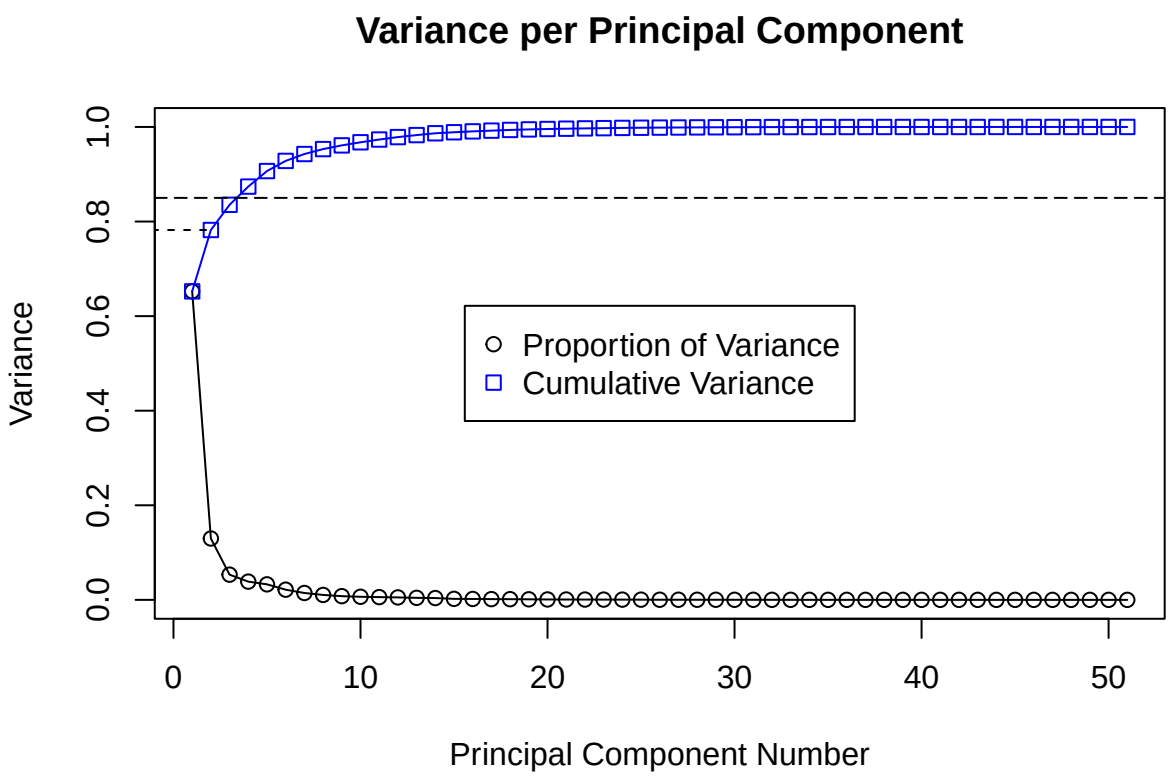
```
set.seed(2025)
BS <- rep(NA,25000)
for (i in 1:25000){
  years <- rownames(measles)[sample(1:nrow(measles),nrow(measles),replace=TRUE)]
  BS[i] <- cor(measles[years,c("Illinois","Massachusetts")]$Illinois,measles[years,c("Illinois","Massach
  }
cat("    Standard Error: ",sd(BS),"\nMean Squared Error: ",mean((BS-cor(measles$Illinois,measles$Massach
```

```
##     Standard Error: 0.053653
## Mean Squared Error: 0.002878849
## Estimated Bias (?): -0.0005657757
```

```
summary(prcomp(measles))$importance["Cumulative Proportion",1:5]  # Problem 5d
```

```
##     PC1     PC2     PC3     PC4     PC5
## 0.65237 0.78212 0.83540 0.87395 0.90675
```

```
plot(summary(prcomp(measles))$importance["Proportion of Variance",],main="Variance per Principal Compone
points(summary(prcomp(measles))$importance["Cumulative Proportion",],col="blue",type="o",pch=0)
abline(h=.85,lty=5) # At least four components are needed to account for 85 percent of the variability.
segments(-2,summary(prcomp(measles))$importance["Cumulative Proportion",2],2,lty=2)
legend("center",c("Proportion of Variance","Cumulative Variance"),col=c("black","blue"),pch=c(1,0))
```



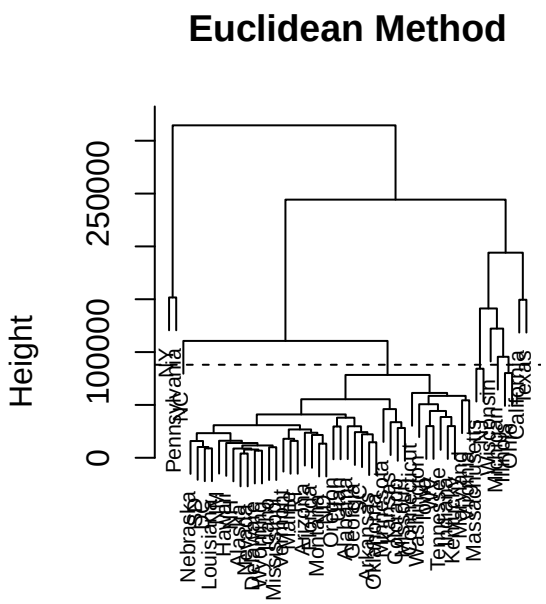

```
cat("The first two components account for approximately",summary(prcomp(measles))$importance["Proportio
```

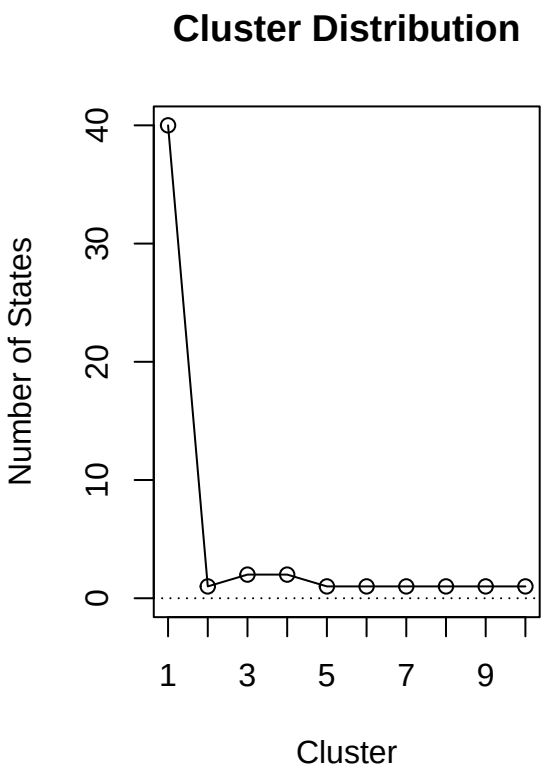
```
## The first two components account for approximately 65.237 and 12.974 percent of the variability, resp
```

```
par(mfrow=c(1,2))                                                    # Problem 5e
plot(hclust(dist(as.data.frame(t(measles)),method="euclidean")),main="Euclidean Method",sub="",xlab="",
abline(h=88000,lty=2) # Choosing arbitrary cutpoints
table(cutree(hclust(dist(as.data.frame(t(measles)),method="euclidean")),h=88000)) # The dataset is trans
```

```
##
##  1  2  3  4  5  6  7  8  9 10
## 40  1  2  2  1  1  1  1  1  1
```

```
plot(table(cutree(hclust(dist(as.data.frame(t(measles)),method="euclidean")),h=88000)),main="Cluster Dis
abline(h=0,lty=3)
```

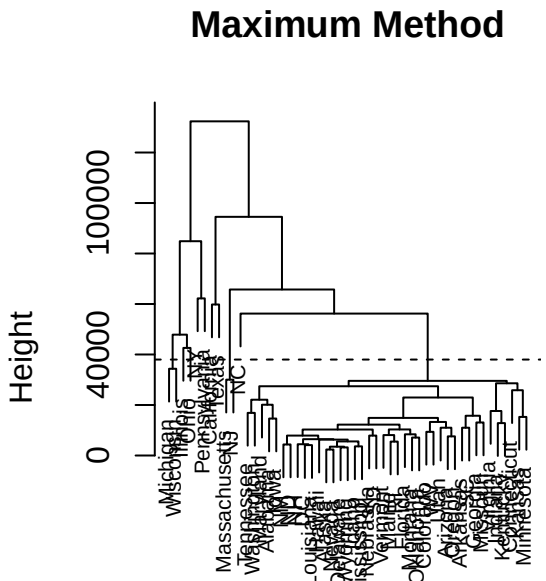


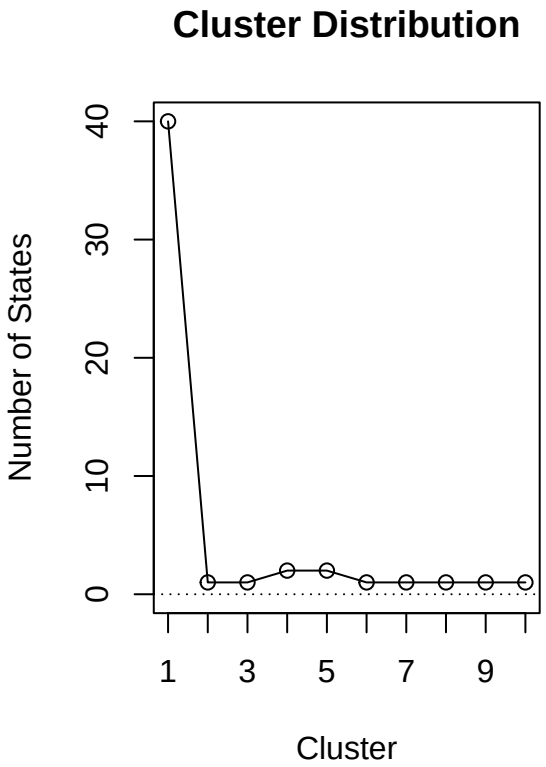



```
plot(hclust(dist(as.data.frame(t(measles)),method="maximum")),main="Maximum Method",sub="",xlab="",cex=
abline(h=38000,lty=2)
table(cutree(hclust(dist(as.data.frame(t(measles)),method="maximum")),h=38000))
```

```
##
##  1  2  3  4  5  6  7  8  9 10
## 40  1  1  2  2  1  1  1  1  1
```

```
plot(table(cutree(hclust(dist(as.data.frame(t(measles)),method="maximum")),h=38000)),main="Cluster Dist
abline(h=0,lty=3)
```

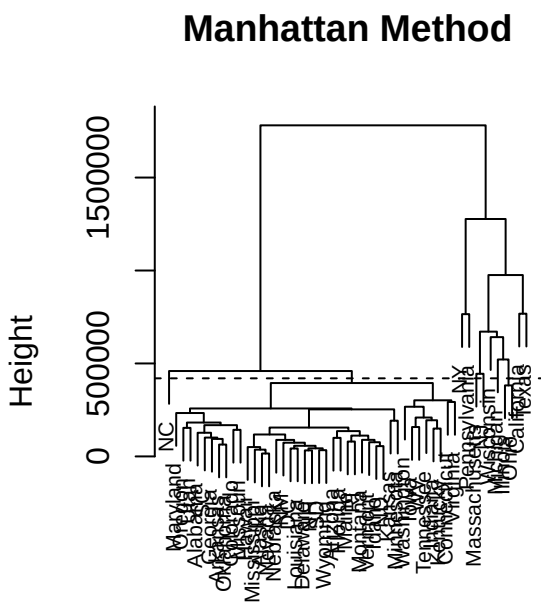


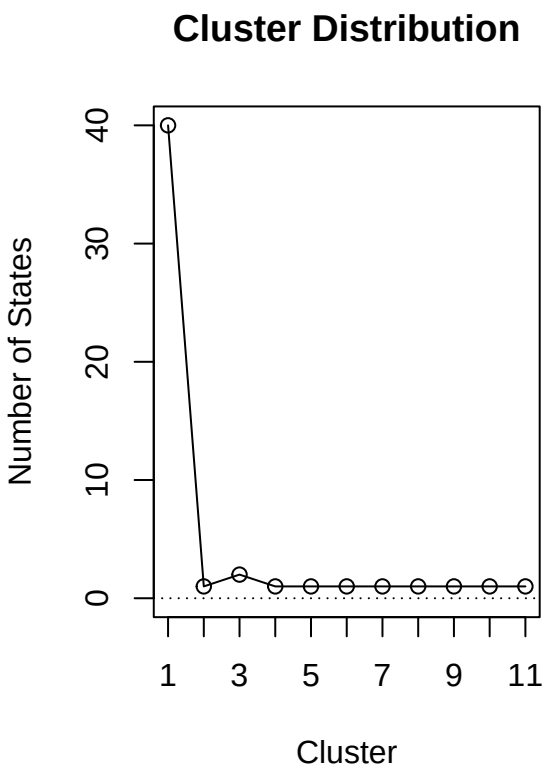



```
plot(hclust(dist(as.data.frame(t(measles)),method="manhattan")),main="Manhattan Method",sub="",xlab="",
abline(h=420000,lty=2)
table(cutree(hclust(dist(as.data.frame(t(measles)),method="manhattan")),h=420000))
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11
## 40  1  2  1  1  1  1  1  1  1  1
```

```
plot(table(cutree(hclust(dist(as.data.frame(t(measles)),method="manhattan")),h=420000)),main="Cluster Di
abline(h=0,lty=3)
```

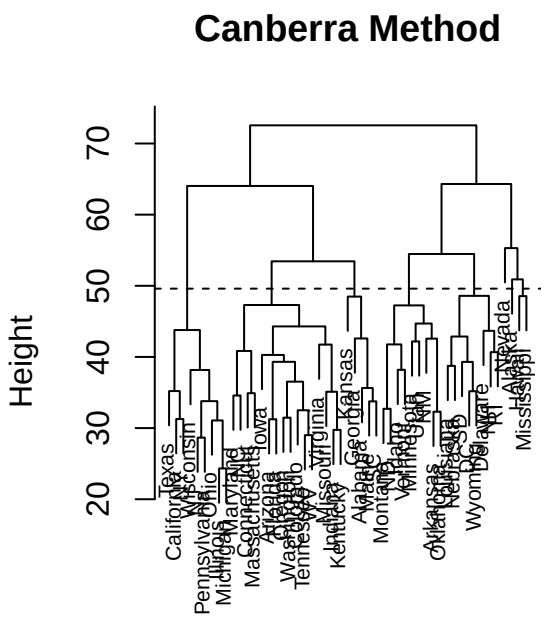


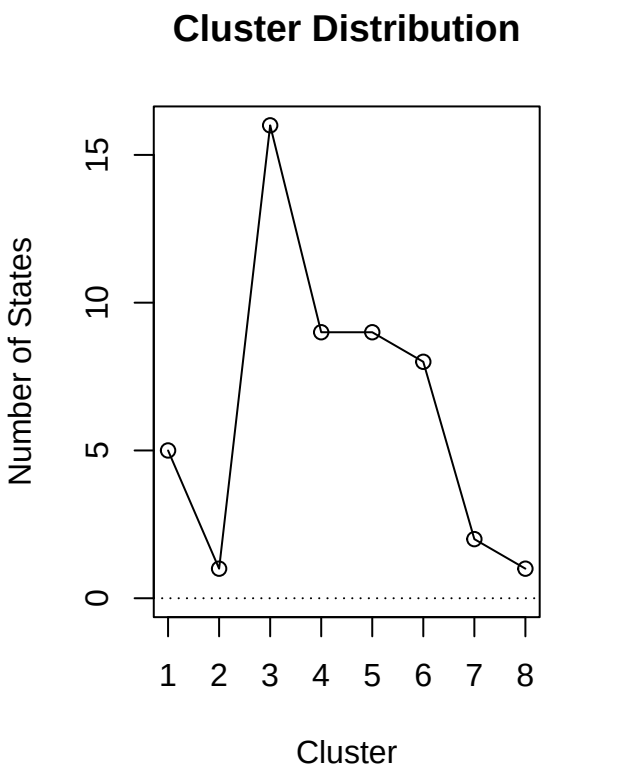



```
plot(hclust(dist(as.data.frame(t(measles)),method="canberra")),main="Canberra Method",sub="",xlab="",ce
abline(h=49.6,lty=2)
table(cutree(hclust(dist(as.data.frame(t(measles)),method="canberra")),h=49.6))
```

```
##
##  1  2  3  4  5  6  7  8
##  5  1 16  9  9  8  2  1
```

```
plot(table(cutree(hclust(dist(as.data.frame(t(measles)),method="canberra")),h=49.6)),main="Cluster Dist
abline(h=0,lty=3)
```






```
par(mfrow=c(1,1))
```

All methods produced nearly the same results except for the Canberra method. The Euclidean, maximum, and Manhattan methods produced clusters for California, Illinois, Massachusetts, Michigan, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Texas, Wisconsin, and the remaining states, with the Euclidean method pairing Illinois and Ohio and Massachusetts and New Jersey together, maximum pairing Massachusetts and New Jersey and Michigan and Wisconsin, and Manhattan pairing Illinois and Ohio, all with cutpoints at heights of thousands of units. The clustered pairs of states makes sense because they belong to the same geographic region and/or are similar (Massachusetts and New Jersey have similar square area and population sizes), and the states in their own cluster were among the ten most populated states in the country (Florida excluded), accounting for a majority of the United States' population. However, the Canberra method produced clusters for Alaska, Hawaii and Mississippi, Nevada, and five other clusters of between five and 16 states each. This is more difficult to accurately interpret because the clusters are states with seemingly no apparent relation with one another. Overall, if the dendrograms produced from the Euclidean, maximum, and Manhattan methods are combined, there are nine to 12 distinct clusters: if the three pairings of states (Illinois and Ohio, Massachusetts and New Jersey, and Michigan and Wisconsin) are all used, there are nine clusters, and if the eleven individual states are all separated, there are twelve clusters, along with the six combinations of pairings resulting in ten or eleven clusters.