Data Analytics Assignment 7

Shanthni Ravindrababu

Exploratory Data Analysis

Model Development and Analysis

Populating england data with wind and precip means

```r
la <- unique(wind_england$lat)
lo <- unique(wind_england$lon)
colnames(wind_england)[1:2] <- c('lat1', 'lon1')

precip_england$lon1 <- unlist(lapply(precip_england$lon, function(x)
                      lo[which.min(abs(lo-x))]))
precip_england$lat1 <- unlist(lapply(precip_england$lat, function(x)
                      la[which.min(abs(la-x))]))

wind_england$wind_mean <- apply(subset(wind_england, select =
                                         -c(lat1,lon1)),1,mean)
precip_england$precip_mean <- apply(subset(precip_england, select =
                                           -c(lat,lon,lat1,lon1)),1,mean)
```

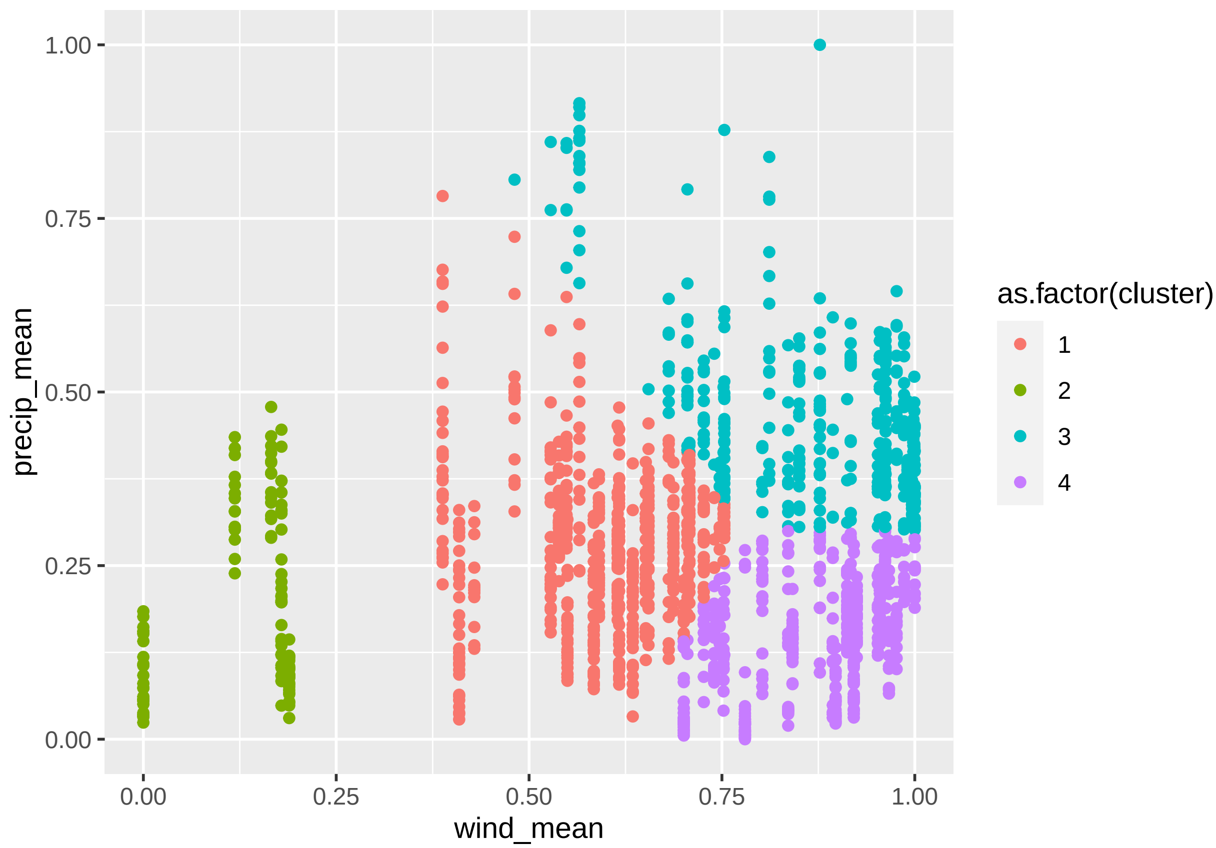
```r
england <- (subset(precip_england, select=c(lat1,lon1,lat,lon,precip_mean))) %>% full_jo
```

```r
la <- unique(wind_Venezuala$lat)
lo <- unique(wind_Venezuala$lon)
colnames(wind_Venezuala)[1:2] <- c('lat1', 'lon1')

precip_Venezuala$lon1 <- unlist(lapply(precip_Venezuala$lon, function(x)
                      lo[which.min(abs(lo-x))]))
precip_Venezuala$lat1 <- unlist(lapply(precip_Venezuala$lat, function(x)
                      la[which.min(abs(la-x))]))

wind_Venezuala$wind_mean <- apply(subset(wind_Venezuala, select =
                                         -c(lat1,lon1)),1,mean)
precip_Venezuala$precip_mean <- apply(subset(precip_Venezuala, select =
                                           -c(lat,lon,lat1,lon1)),1,mean)

venezuala <- (subset(precip_Venezuala, select=c(lat1,lon1,lat,lon,precip_mean))) %>% ful
```

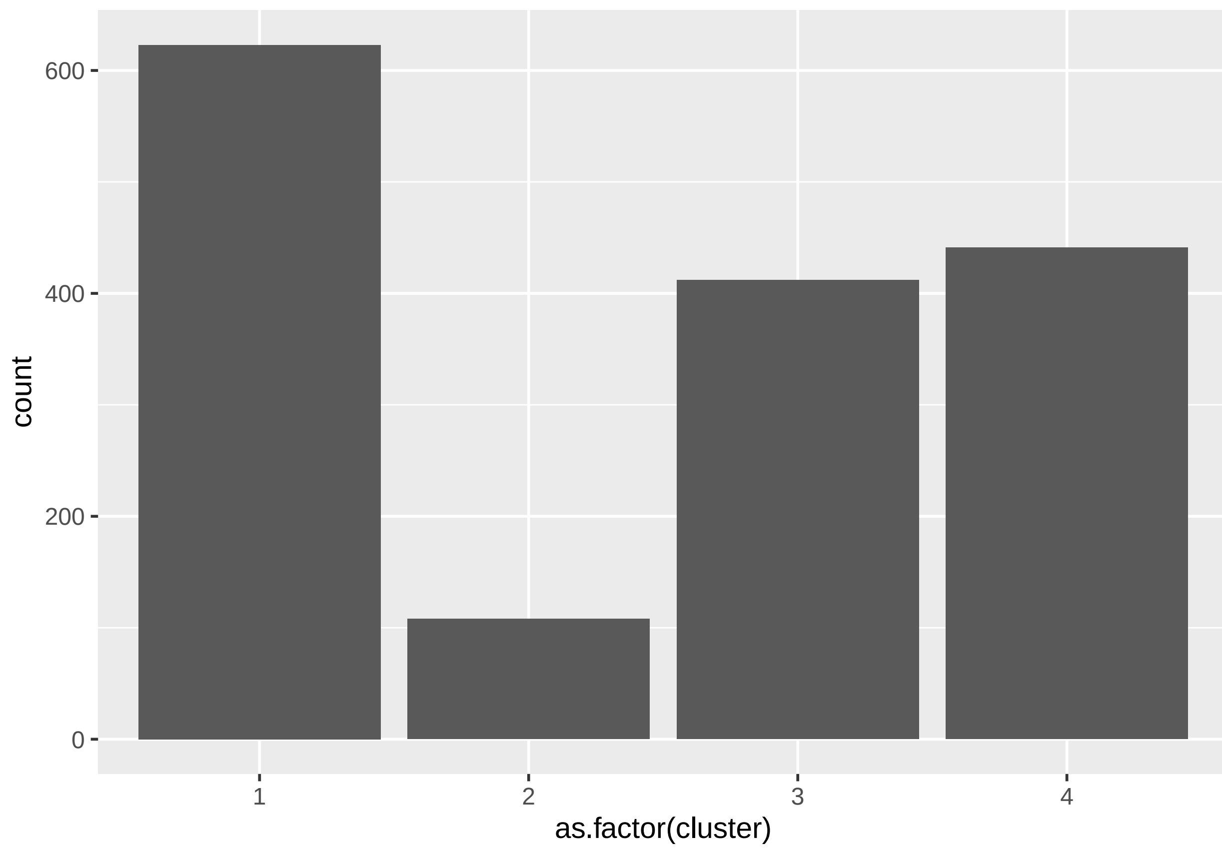
K Means Clustering

```r
england.k <- england

normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}

england.k[5:6] <- as.data.frame(lapply(england.k[5:6], normalize))
```

```r
clustering.kmeans <- kmeans(subset(england.k, select=c(wind_mean,precip_mean)), 2)

england.k$cluster <- as.factor(clustering.kmeans$cluster)

# write.csv(england.k, 'england-kcluster.csv')

graph.england <- read.csv('england-kcluster.csv', header=T)

ggplot(data=graph.england, aes(x=wind_mean, y=precip_mean,color=as.factor(cluster))) +
geom_point() +
scale_fill_manual(values=c('1' = 'blue',
                           '2' = 'green',
                           '3' = 'purple',
                           '4' = 'red'))
```

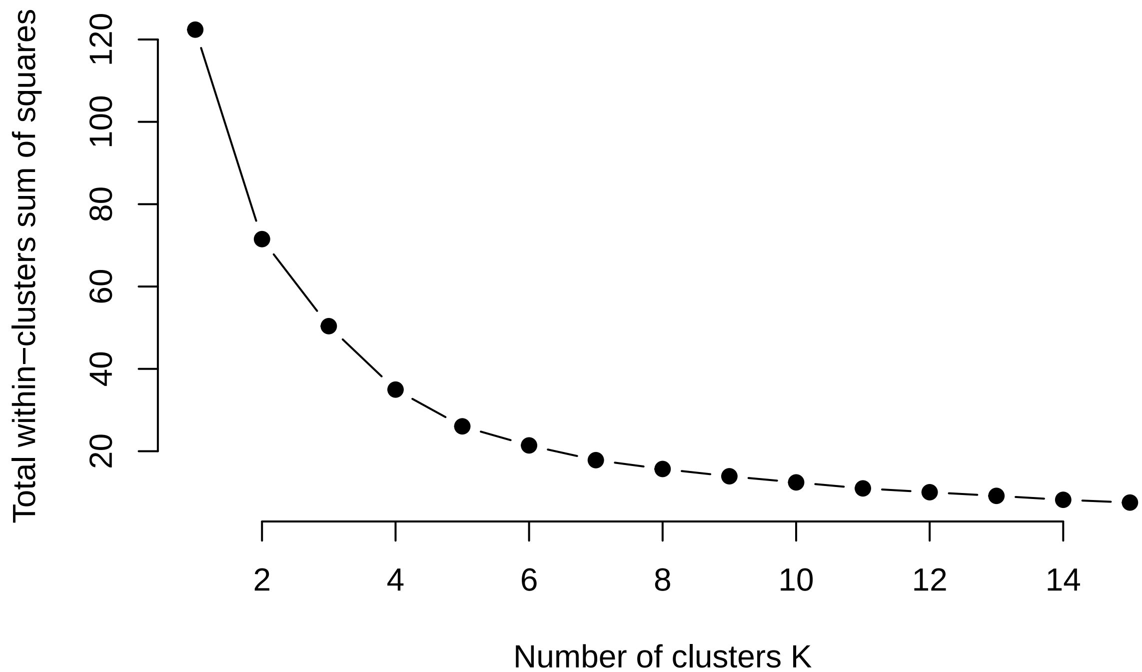



```
ggplot(graph.england, aes(x = as.factor(cluster))) + geom_bar()
```




```
# 1 - blue, 2 - green, 3- purple, 4 - red


#Elbow Method for finding the optimal number of clusters
set.seed(123)
# Compute and plot wss for k = 2 to k = 15.
k.max <- 15
data <- subset(england.k, select=c(wind_mean,precip_mean))
wss <- sapply(1:k.max,
              function(k){kmeans(data, k, nstart=50,iter.max = 15 )$tot.withinss})


plot(1:k.max, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
```

```
    ylab="Total within-clusters sum of squares")
```



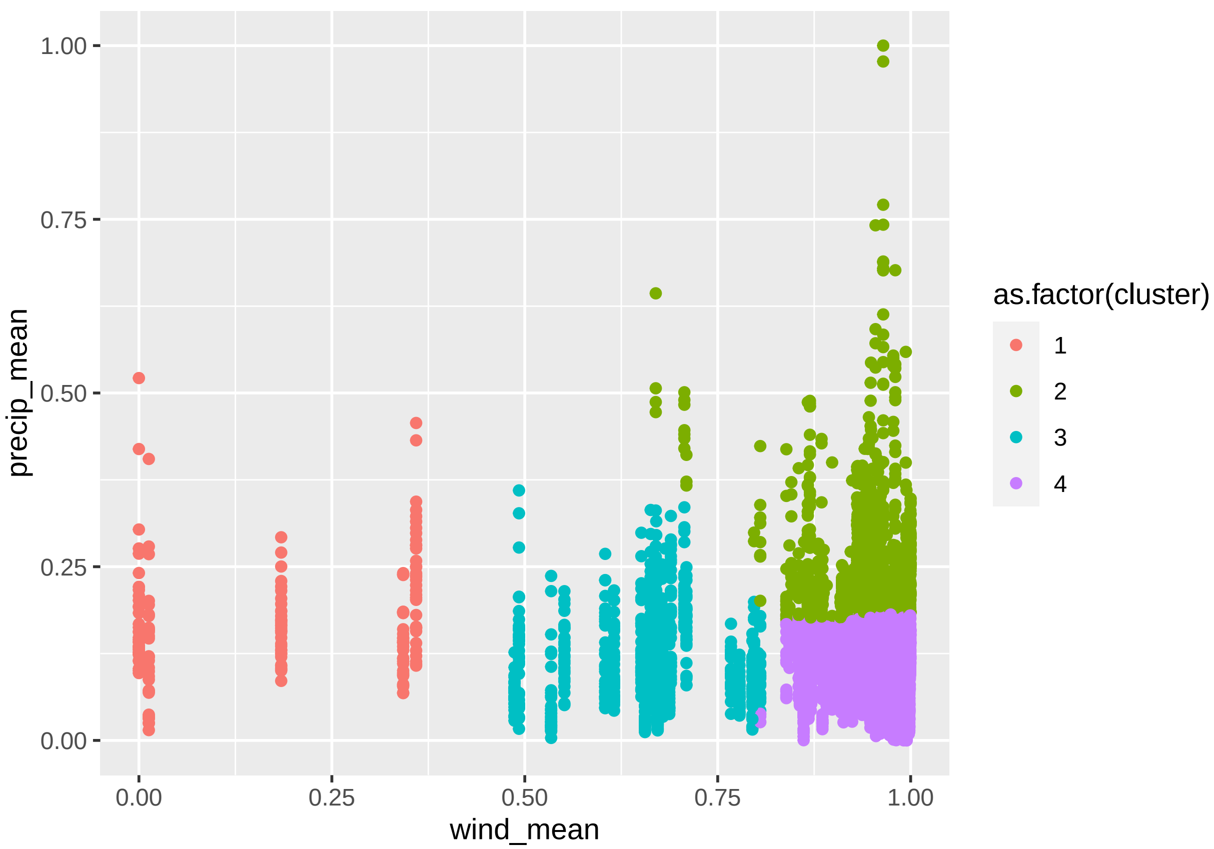

```
venezuala.k <- venezuala
```

```
venezuala.k[5:6] <- as.data.frame(lapply(venezuala.k[5:6], normalize))
```

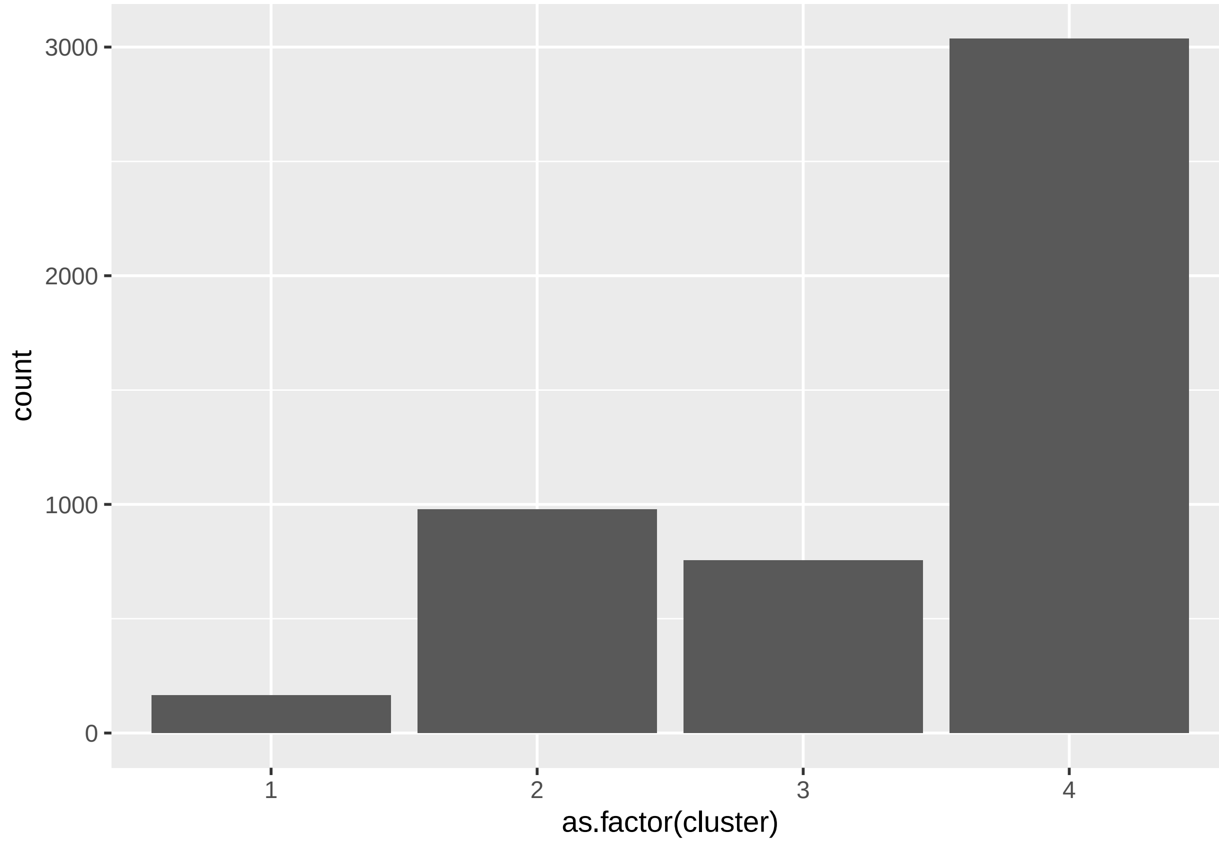
```
clustering.kmeans <- kmeans(subset(venezuala.k, select=c(wind_mean,precip_mean)), 4)
```

```
venezuala.k$cluster <- as.factor(clustering.kmeans$cluster)
```

```r
#write.csv(venezuala.k, 'venezuala-kcluster.csv')



graph.venezuala <- read.csv('venezuala-kcluster.csv', header=T)


ggplot(data=graph.venezuala, aes(x=wind_mean, y=precip_mean,color=as.factor(cluster))) +
geom_point() +
scale_fill_manual(values=c('1' = 'blue',
                           '2' = 'green',
                           '3' = 'purple',
                           '4' = 'red'))
```

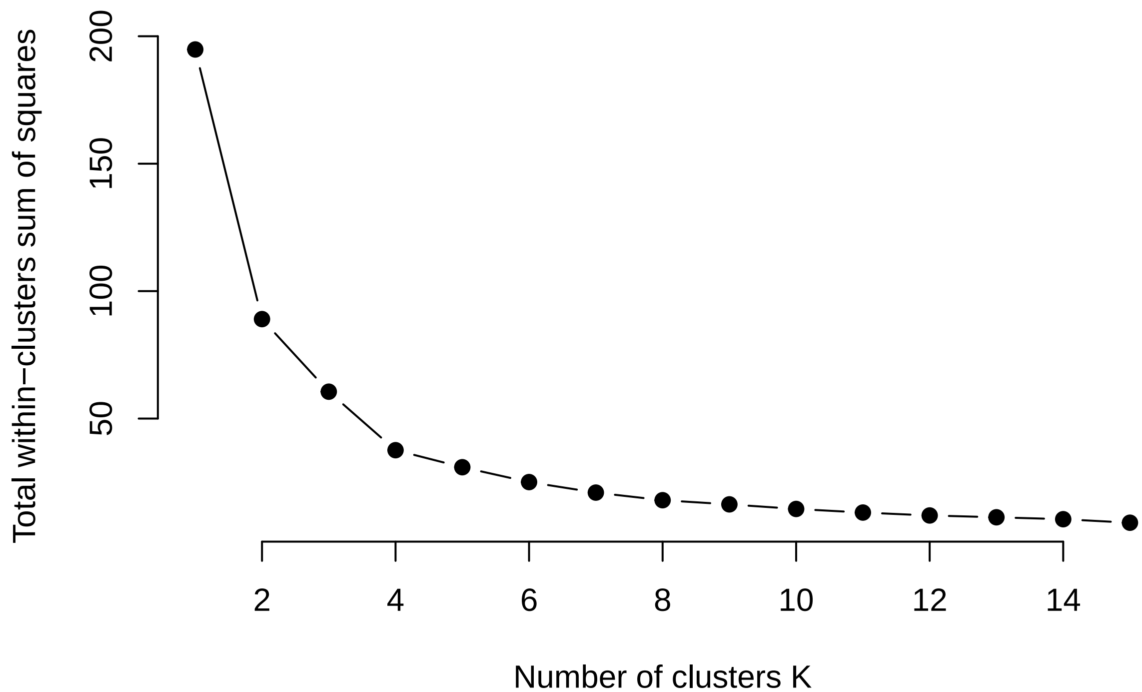


```
ggplot(graph.venezuala, aes(x = as.factor(cluster))) + geom_bar()
```




```
# 1 - blue, 2 - green, 3- purple, 4 - red


#Elbow Method for finding the optimal number of clusters
set.seed(123)
# Compute and plot wss for k = 2 to k = 15.
k.max <- 15
data <- subset(venezuala.k, select=c(wind_mean,precip_mean))
wss <- sapply(1:k.max,
              function(k){kmeans(data, k, nstart=50,iter.max = 15 )$tot.withinss})


plot(1:k.max, wss,
```

```
    type="b", pch = 19, frame = FALSE,
    xlab="Number of clusters K",
    ylab="Total within-clusters sum of squares")
```



Random Forests

```
england.r <- england

england.r[5:6] <- as.data.frame(lapply(england.r[5:6], normalize))

# solar_farms <- as.data.frame(lat=c(53.14, 51.30, 51.21), lon=c(-3.15,-1.58, -1.64))
solar_app_lat <- c(53.14, 51.30, 51.21)
solar_app_lon <- c(-3.15,-1.58, -1.64)
```

```r
# wind_farms <- as.data.frame(lat=c(53.53, 53.00, 52.14), lon=c(-1.47,-0.48,-2.29))
wind_app_lat <- c(53.53, 53.00, 52.14)
wind_app_lon <- c(-1.47,-0.48,-2.29)



england.r$SW <- rep('N', length(england.r))
for(i in 1:nrow(england.r)) {
  lat <- england.r$lat1[i]
  lon <- england.r$lon1[i]
  for (j in 1:length(solar_app_lat)) {
    dist1 <- sqrt((solar_app_lat[j]-lat)^2 + (solar_app_lon[j]-lon)^2)
    dist2 <- sqrt((wind_app_lat[j]-lat)^2 + (wind_app_lon[j]-lon)^2)
    if ((dist1 <= 1 && dist2 <= 1)) {
      england.r$SW[i] <- 'B'
      break
    }
    else if (dist1 <= 1) {
      england.r$SW[i] <- 'S'
      break
    }
    else if (dist2 <= 1) {
      england.r$SW[i] <- 'W'
    }
  }


}
```

```
england.r$SW <- as.factor(england.r$SW)

england.r <- subset(england.r, select=-c(lat1,lon1,lat,lon))

ind <- sample(2, nrow(england.r), replace=TRUE, prob=c(0.7, 0.3))
RFtrain <- england.r[ind==1,]
RFtest <- england.r[ind==2,]

RFmodel <- randomForest(SW ~ precip_mean + wind_mean, data = RFtrain, importance=TRUE)

RFpred <- predict(RFmodel, RFtest, type="class")

table(RFpred)
```

```
## RFpred
##   N   S   W
## 210 109 170
```

```
data.frame(rbind(table(RFtest$SW, RFpred)))
```

```
##     N   S   W
## N 204   2   3
## S   3 103   5
## W   3   4 162
```

```
1 - mean(RFpred != RFtest$SW)
```

```
## [1] 0.9591002
```

```r
# RF predict Venezuala


venezuala.r <- venezuala
venezuala.r[5:6] <- as.data.frame(lapply(venezuala.r[5:6], normalize))


RFpred_venezuala <- predict(RFmodel, venezuala.r, type="class")
venezuala.r$SW <- RFpred_venezuala
write.csv(venezuala.r, 'venezuala-rpredict.csv')
table(RFpred_venezuala)
```

```
## RFpred_venezuala
##    N    S    W
## 3688   88 1159
```