Final\_Code

Haley Aldred

04/12/2021

# Introduction

This report contains my code for examining my data set, which will be used in classifying and predicting forest cover types in Roosevelt National Forest in Colorado, USA. My data comes from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/>).

# Preparation Work

## Loading Libraries

library(ggplot2)  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble 3.1.5 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.0.2 v forcats 0.5.1  
## v purrr 0.3.4

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(dplyr)  
library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(scales)

##   
## Attaching package: 'scales'

## The following objects are masked from 'package:psych':  
##   
## alpha, rescale

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

library(texreg)

## Version: 1.37.5  
## Date: 2020-06-17  
## Author: Philip Leifeld (University of Essex)  
##   
## Consider submitting praise using the praise or praise\_interactive functions.  
## Please cite the JSS article in your publications -- see citation("texreg").

##   
## Attaching package: 'texreg'

## The following object is masked from 'package:tidyr':  
##   
## extract

library(corrplot)

## corrplot 0.92 loaded

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(lattice)  
library(e1071)  
library(class)  
library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:psych':  
##   
## outlier

## The following object is masked from 'package:dplyr':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

# Data Exploration

Loading the data.

data <- read.csv("~/School/CIND820/data/covtype.data.gz", header=FALSE, stringsAsFactors=TRUE)  
#View(data)

Looking at the first few rows of data.

head(data)

## V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19  
## 1 2596 51 3 258 0 510 221 232 148 6279 1 0 0 0 0 0 0 0 0  
## 2 2590 56 2 212 -6 390 220 235 151 6225 1 0 0 0 0 0 0 0 0  
## 3 2804 139 9 268 65 3180 234 238 135 6121 1 0 0 0 0 0 0 0 0  
## 4 2785 155 18 242 118 3090 238 238 122 6211 1 0 0 0 0 0 0 0 0  
## 5 2595 45 2 153 -1 391 220 234 150 6172 1 0 0 0 0 0 0 0 0  
## 6 2579 132 6 300 -15 67 230 237 140 6031 1 0 0 0 0 0 0 0 0  
## V20 V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38  
## 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## V39 V40 V41 V42 V43 V44 V45 V46 V47 V48 V49 V50 V51 V52 V53 V54 V55  
## 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 5  
## 2 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 5  
## 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2  
## 4 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 2  
## 5 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 5  
## 6 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 2

Adding column names based off the *covtype.info* document.

colnames(data)<-c("elev", "aspect", "slope", "hHydro", "vHydro", "hRoad", "hs09", "hs12", "hs15", "hFire", "wild1", "wild2", "wild3", "wild4", "s1", "s2", "s3", "s4", "s5", "s6", "s7", "s8", "s9", "s10", "s11", "s12", "s13", "s14", "s15", "s16", "s17", "s18", "s19", "s20", "s21", "s22", "s23", "s24", "s25", "s26", "s27", "s28", "s29", "s30", "s31", "s32", "s33", "s34", "s35", "s36", "s37", "s38", "s39", "s40", "treeType")

Verifying the column names have been updated.

head(data)

## elev aspect slope hHydro vHydro hRoad hs09 hs12 hs15 hFire wild1 wild2 wild3  
## 1 2596 51 3 258 0 510 221 232 148 6279 1 0 0  
## 2 2590 56 2 212 -6 390 220 235 151 6225 1 0 0  
## 3 2804 139 9 268 65 3180 234 238 135 6121 1 0 0  
## 4 2785 155 18 242 118 3090 238 238 122 6211 1 0 0  
## 5 2595 45 2 153 -1 391 220 234 150 6172 1 0 0  
## 6 2579 132 6 300 -15 67 230 237 140 6031 1 0 0  
## wild4 s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20  
## 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## s21 s22 s23 s24 s25 s26 s27 s28 s29 s30 s31 s32 s33 s34 s35 s36 s37 s38 s39  
## 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0  
## s40 treeType  
## 1 0 5  
## 2 0 5  
## 3 0 2  
## 4 0 2  
## 5 0 5  
## 6 0 2

Looking at the last few rows of data.

tail(data)

## elev aspect slope hHydro vHydro hRoad hs09 hs12 hs15 hFire wild1 wild2  
## 581007 2401 157 21 90 15 120 238 238 119 830 0 0  
## 581008 2396 153 20 85 17 108 240 237 118 837 0 0  
## 581009 2391 152 19 67 12 95 240 237 119 845 0 0  
## 581010 2386 159 17 60 7 90 236 241 130 854 0 0  
## 581011 2384 170 15 60 5 90 230 245 143 864 0 0  
## 581012 2383 165 13 60 4 67 231 244 141 875 0 0  
## wild3 wild4 s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17  
## 581007 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 581008 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 581009 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 581010 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 581011 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 581012 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## s18 s19 s20 s21 s22 s23 s24 s25 s26 s27 s28 s29 s30 s31 s32 s33 s34 s35  
## 581007 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 581008 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 581009 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 581010 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 581011 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 581012 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## s36 s37 s38 s39 s40 treeType  
## 581007 0 0 0 0 0 3  
## 581008 0 0 0 0 0 3  
## 581009 0 0 0 0 0 3  
## 581010 0 0 0 0 0 3  
## 581011 0 0 0 0 0 3  
## 581012 0 0 0 0 0 3

Verifying that there are no missing values.

colSums(is.na(data))

## elev aspect slope hHydro vHydro hRoad hs09 hs12   
## 0 0 0 0 0 0 0 0   
## hs15 hFire wild1 wild2 wild3 wild4 s1 s2   
## 0 0 0 0 0 0 0 0   
## s3 s4 s5 s6 s7 s8 s9 s10   
## 0 0 0 0 0 0 0 0   
## s11 s12 s13 s14 s15 s16 s17 s18   
## 0 0 0 0 0 0 0 0   
## s19 s20 s21 s22 s23 s24 s25 s26   
## 0 0 0 0 0 0 0 0   
## s27 s28 s29 s30 s31 s32 s33 s34   
## 0 0 0 0 0 0 0 0   
## s35 s36 s37 s38 s39 s40 treeType   
## 0 0 0 0 0 0 0

Getting rid of binary wilderness area type and soil type columns. This will be done by concatenating the binary columns, converting them to numeric, and finding the log of the binary number.

data$wildArea<-paste(data$wild1,data$wild2,data$wild3,data$wild4,sep="")  
data$wildArea<-as.numeric(as.character(data$wildArea))  
data$area<-(4-log(data$wildArea,10))

Next, we will do the same procedure for the 40 different soil types.

data$soilType<-paste(data$s1,data$s2,data$s3,data$s4,data$s5,data$s6,data$s7,data$s8,data$s9,data$s10,data$s11,data$s12,data$s13,data$s14,data$s15,data$s16,data$s17,data$s18,data$s19,data$s20,data$s21,data$s22,data$s23,data$s24,data$s25,data$s26,data$s27,data$s28,data$s29,data$s30,data$s31,data$s32,data$s33,data$s34,data$s35,data$s36,data$s37,data$s38,data$s39,data$s40, sep="")  
data$soilTypes<-as.numeric(as.character(data$soilType))  
data$soil<-(40-log(data$soilTypes,10))

Next, we will calculated the euclidean distance to water.

data$edHydro = (data$hHydro^2+data$vHydro^2)^.5

Now to delete the useless columns. By eliminating a combined 42 columns, it will make the data easier to work with in the future without losing any critical information.

data<-select(data,-c(hHydro,vHydro,wild1,wild2,wild3,wild4,s1,s2,s3,s4,s5,s6,s7,s8,s9,s10,s11,s12,s13,s14,s15,s16,s17,s18,s19,s20,s21,s22,s23,s24,s25,s26,s27,s28,s29,s30,s31,s32,s33,s34,s35,s36,s37,s38,s39,s40,wildArea,soilType,soilTypes))

Verify the process worked.

head(data)

## elev aspect slope hRoad hs09 hs12 hs15 hFire treeType area soil edHydro  
## 1 2596 51 3 510 221 232 148 6279 5 1 29 258.0000  
## 2 2590 56 2 390 220 235 151 6225 5 1 29 212.0849  
## 3 2804 139 9 3180 234 238 135 6121 2 1 12 275.7698  
## 4 2785 155 18 3090 238 238 122 6211 2 1 30 269.2360  
## 5 2595 45 2 391 220 234 150 6172 5 1 29 153.0033  
## 6 2579 132 6 67 230 237 140 6031 2 1 29 300.3748

tail(data)

## elev aspect slope hRoad hs09 hs12 hs15 hFire treeType area soil edHydro  
## 581007 2401 157 21 120 238 238 119 830 3 3 2 91.24144  
## 581008 2396 153 20 108 240 237 118 837 3 3 2 86.68333  
## 581009 2391 152 19 95 240 237 119 845 3 3 2 68.06614  
## 581010 2386 159 17 90 236 241 130 854 3 3 2 60.40695  
## 581011 2384 170 15 90 230 245 143 864 3 3 2 60.20797  
## 581012 2383 165 13 67 231 244 141 875 3 3 2 60.13319

Identifying tree types, adding a column for tree name

data$treeName = 'a'  
data$treeName[data$treeType==1] = 'Spruce-fir'  
data$treeName[data$treeType==2] = 'Lodgepole Pine'  
data$treeName[data$treeType==3] = 'Ponderosa Pine'  
data$treeName[data$treeType==4] = 'Cottonwood-Willow'  
data$treeName[data$treeType==5] = 'Aspen'  
data$treeName[data$treeType==6] = 'Douglas-fir'  
data$treeName[data$treeType==7] = 'Krummholz'

Identifying areas, adding a column for area name

data$areaName = 'a'  
data$areaName[data$area==1] = 'Rawah'  
data$areaName[data$area==2] = 'Neota'  
data$areaName[data$area==3] = 'Comanche Peak'  
data$areaName[data$area==4] = 'Cache La Poudre'

Looking at the structure of the data. We will first create a dataset without the names.

forest<-select(data,-c(treeName,areaName))  
forest<-forest %>% relocate(treeType,.after=edHydro)  
  
str(forest)

## 'data.frame': 581012 obs. of 12 variables:  
## $ elev : int 2596 2590 2804 2785 2595 2579 2606 2605 2617 2612 ...  
## $ aspect : int 51 56 139 155 45 132 45 49 45 59 ...  
## $ slope : int 3 2 9 18 2 6 7 4 9 10 ...  
## $ hRoad : int 510 390 3180 3090 391 67 633 573 666 636 ...  
## $ hs09 : int 221 220 234 238 220 230 222 222 223 228 ...  
## $ hs12 : int 232 235 238 238 234 237 225 230 221 219 ...  
## $ hs15 : int 148 151 135 122 150 140 138 144 133 124 ...  
## $ hFire : int 6279 6225 6121 6211 6172 6031 6256 6228 6244 6230 ...  
## $ area : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ soil : num 29 29 12 30 29 29 29 29 29 29 ...  
## $ edHydro : num 258 212 276 269 153 ...  
## $ treeType: int 5 5 2 2 5 2 5 5 5 5 ...

Looking at the basic data statistics.

summary(forest)

## elev aspect slope hRoad hs09   
## Min. :1859 Min. : 0.0 Min. : 0.0 Min. : 0 Min. : 0.0   
## 1st Qu.:2809 1st Qu.: 58.0 1st Qu.: 9.0 1st Qu.:1106 1st Qu.:198.0   
## Median :2996 Median :127.0 Median :13.0 Median :1997 Median :218.0   
## Mean :2959 Mean :155.7 Mean :14.1 Mean :2350 Mean :212.1   
## 3rd Qu.:3163 3rd Qu.:260.0 3rd Qu.:18.0 3rd Qu.:3328 3rd Qu.:231.0   
## Max. :3858 Max. :360.0 Max. :66.0 Max. :7117 Max. :254.0   
## hs12 hs15 hFire area soil   
## Min. : 0.0 Min. : 0.0 Min. : 0 Min. :1.000 Min. : 1.00   
## 1st Qu.:213.0 1st Qu.:119.0 1st Qu.:1024 1st Qu.:1.000 1st Qu.:20.00   
## Median :226.0 Median :143.0 Median :1710 Median :2.000 Median :29.00   
## Mean :223.3 Mean :142.5 Mean :1980 Mean :2.114 Mean :24.36   
## 3rd Qu.:237.0 3rd Qu.:168.0 3rd Qu.:2550 3rd Qu.:3.000 3rd Qu.:31.00   
## Max. :254.0 Max. :254.0 Max. :7173 Max. :4.000 Max. :40.00   
## edHydro treeType   
## Min. : 0.0 Min. :1.000   
## 1st Qu.: 108.5 1st Qu.:1.000   
## Median : 229.5 Median :2.000   
## Mean : 276.1 Mean :2.051   
## 3rd Qu.: 393.8 3rd Qu.:2.000   
## Max. :1418.9 Max. :7.000

The mean elevation of the area is 2959 m. The mean distance to water is 269.4 m. The mean distance to roads is 2350 m.The median tree type is number 2, also known as the lodgepole pine. The median wilderness area is number 2, also know as Neota, and the median soil type is type 29, also know as Como which contains Legault families complex, extremely stony.

The psych package will be used to take a closer look at the data.

describe(forest)

## vars n mean sd median trimmed mad min max  
## elev 1 581012 2959.37 279.98 2996.00 2983.39 259.46 1859 3858.00  
## aspect 2 581012 155.66 111.91 127.00 150.10 126.02 0 360.00  
## slope 3 581012 14.10 7.49 13.00 13.49 7.41 0 66.00  
## hRoad 4 581012 2350.15 1559.25 1997.00 2203.24 1541.90 0 7117.00  
## hs09 5 581012 212.15 26.77 218.00 215.09 23.72 0 254.00  
## hs12 6 581012 223.32 19.77 226.00 225.11 17.79 0 254.00  
## hs15 7 581012 142.53 38.27 143.00 143.36 37.06 0 254.00  
## hFire 8 581012 1980.29 1324.20 1710.00 1797.61 1111.95 0 7173.00  
## area 9 581012 2.11 1.06 2.00 2.06 1.48 1 4.00  
## soil 10 581012 24.36 9.49 29.00 25.08 7.41 1 40.00  
## edHydro 11 581012 276.07 217.05 229.48 250.21 204.43 0 1418.92  
## treeType 12 581012 2.05 1.40 2.00 1.73 1.48 1 7.00  
## range skew kurtosis se  
## elev 1999.00 -0.82 0.75 0.37  
## aspect 360.00 0.40 -1.22 0.15  
## slope 66.00 0.79 0.58 0.01  
## hRoad 7117.00 0.71 -0.38 2.05  
## hs09 254.00 -1.18 1.88 0.04  
## hs12 254.00 -1.06 2.07 0.03  
## hs15 254.00 -0.28 0.40 0.05  
## hFire 7173.00 1.29 1.65 1.74  
## area 3.00 0.09 -1.61 0.00  
## soil 39.00 -0.70 -0.43 0.01  
## edHydro 1418.92 1.13 1.36 0.28  
## treeType 6.00 2.28 4.95 0.00

This indicates that there are 260796 instances for wilderness area 1 (Rawah), 29884 instances for wilderness area 2 (Neota), 253364 instances for wilderness area 3 (Comanche Peak), and 36968 instances for wilderness area 4 (Cache la Poudre). The Rawah wilderness area has the most instances in the collected data, next is Comanche Peak.

Lets create a table of the different tree types.

table(data$treeName)

##   
## Aspen Cottonwood-Willow Douglas-fir Krummholz   
## 9493 2747 17367 20510   
## Lodgepole Pine Ponderosa Pine Spruce-fir   
## 283301 35754 211840

This means there are 211840 instances of tree type 1 (spruce/fir), 283301 instances of tree type 2 (lodgepole pine), 35754 instances of tree type 3 (ponderosa pine), 2747 instances of tree type 4 (cottonwood/willow), 9493 instances of tree type 5 (aspen), 17367 instances of tree type 6 (douglas-fir), and 20510 instances of tree type 7 (krummholz). Tree type 2, or lodgepole pine, is the most frequent tree type found in the national forest. Cottonwood/willow trees are the most rare tree type in the forest.

We will do the same for the wilderness areas and soil types.

table(data$areaName)

##   
## Cache La Poudre Comanche Peak Neota Rawah   
## 36968 253364 29884 260796

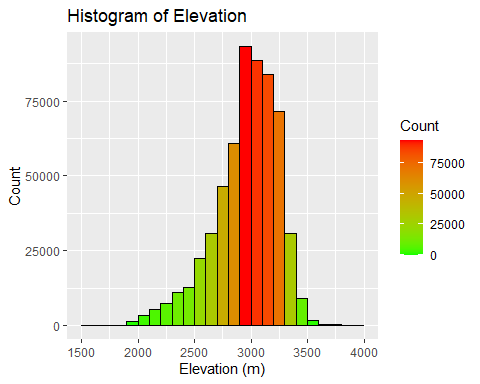
table(data$soil)

##   
## 1 2 3 4 5 6 7 8 9 10 11   
## 3031 7525 4823 12396 1597 6575 105 179 1147 32634 12410   
## 12 13 14 15 16 17 18 19 20 21 22   
## 29971 17431 599 3 2845 3422 1899 4021 9259 838 33373   
## 23 24 25 26 27 28 29 30 31 32 33   
## 57752 21278 474 2589 1086 946 115247 30170 25666 52519 45154   
## 34 35 36 37 38 39 40   
## 1611 1891 119 298 15573 13806 8750

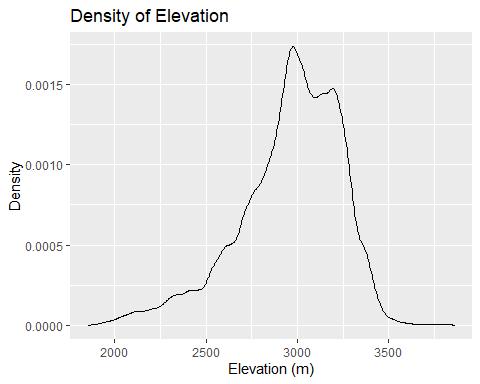
# Data Visualization

Let’s look at the distribution of elevation for the data set.

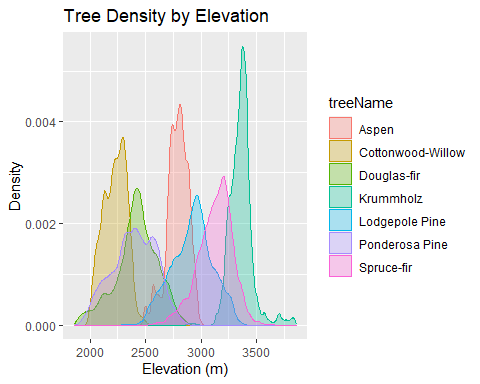
ggplot(data, aes(x=elev))+geom\_histogram(breaks=seq(1500,4000,by=100),col="black",aes(fill=..count..))+scale\_fill\_gradient("Count",low="green",high="red")+labs(title="Histogram of Elevation",x="Elevation (m)",y="Count")



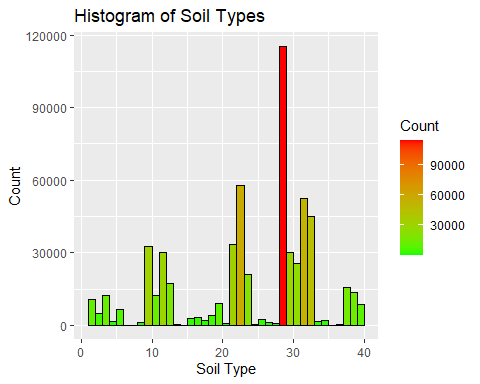
ggplot(data, aes(x=elev)) + geom\_density()+labs(title="Density of Elevation",x="Elevation (m)",y="Density")



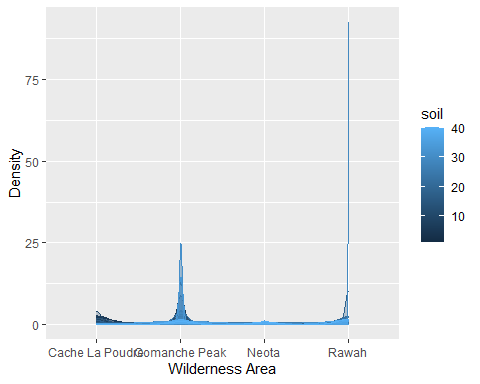
ggplot(data, aes(x=elev)) + geom\_density(aes(group=treeName, colour=treeName, fill=treeName), alpha=0.3)+labs(title="Tree Density by Elevation",x="Elevation (m)",y="Density")

 Let’s look at the distribution of soil types for the data set.

ggplot(data=data, aes(soil))+geom\_histogram(breaks=seq(1,40,by=1),col="black",aes(fill=..count..))+scale\_fill\_gradient("Count",low="green",high="red")+labs(title="Histogram of Soil Types",x="Soil Type",y="Count")

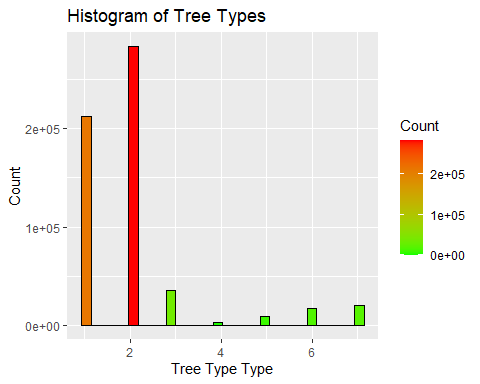


ggplot(data,aes(areaName))+geom\_density(aes(group=soil,colour=soil,fill=soil),alpha=0.3)+labs(main="Soils based on Wilderness Area",x="Wilderness Area",y="Density")

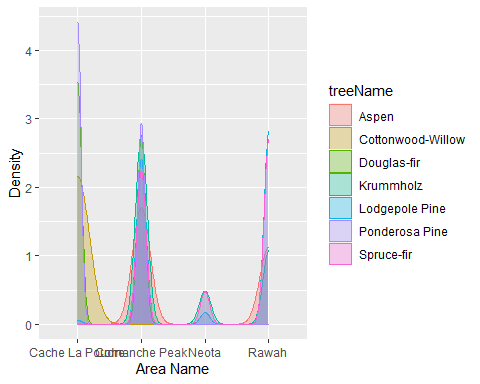
 Let’s look at the distribution of tree types for the data set.

ggplot(data=data, aes(treeType))+geom\_histogram(col="black",aes(fill=..count..))+scale\_fill\_gradient("Count",low="green",high="red")+labs(title="Histogram of Tree Types",x="Tree Type Type",y="Count")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



ggplot(data, aes(x=areaName)) + geom\_density(aes(group=treeName, colour=treeName, fill=treeName), alpha=0.3)+labs(main="Tree Density by Area",x="Area Name",y="Density")

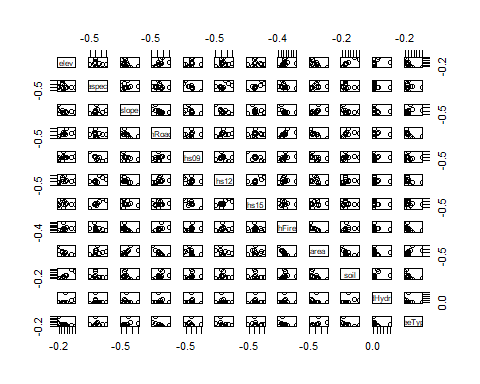
 Next, we will create a correlation matrix of the data.

corForest<-cor(forest)  
head(corForest)

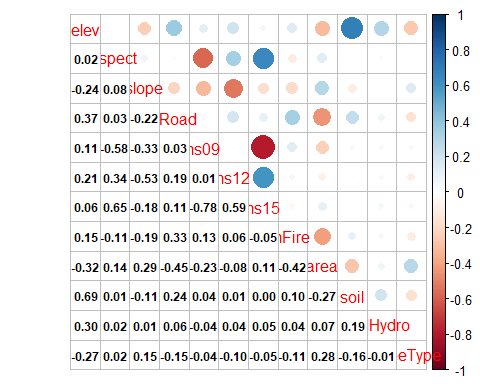
## elev aspect slope hRoad hs09 hs12  
## elev 1.00000000 0.01573494 -0.24269664 0.36555928 0.11217930 0.20588691  
## aspect 0.01573494 1.00000000 0.07872841 0.02512069 -0.57927291 0.33610296  
## slope -0.24269664 0.07872841 1.00000000 -0.21591416 -0.32719899 -0.52691064  
## hRoad 0.36555928 0.02512069 -0.21591416 1.00000000 0.03434912 0.18946096  
## hs09 0.11217930 -0.57927291 -0.32719899 0.03434912 1.00000000 0.01003685  
## hs12 0.20588691 0.33610296 -0.52691064 0.18946096 0.01003685 1.00000000  
## hs15 hFire area soil edHydro treeType  
## elev 0.05914778 0.14802156 -0.31558989 0.689848086 0.298044169 -0.2695538  
## aspect 0.64694395 -0.10917150 0.13870259 0.007148751 0.021169187 0.0170798  
## slope -0.17585362 -0.18566195 0.28617814 -0.105570568 0.009285845 0.1482854  
## hRoad 0.10611918 0.33157958 -0.44592246 0.240829308 0.064298904 -0.1534498  
## hs09 -0.78029595 0.13266890 -0.23345480 0.040876904 -0.037468433 -0.0354150  
## hs12 0.59427365 0.05732865 -0.08076668 0.013661836 0.036901414 -0.0964260

Now we will create a correlogram to visualize the correlation matrix.

pairs(corForest)



corrplot.mixed(corForest,lower.col="black",number.cex=0.75)

 Negative slopes show a negative correlation, while positive slopes show a positive correlation between two attributes. Attributes with the highest correlation to tree type include area.

# Data Analysis

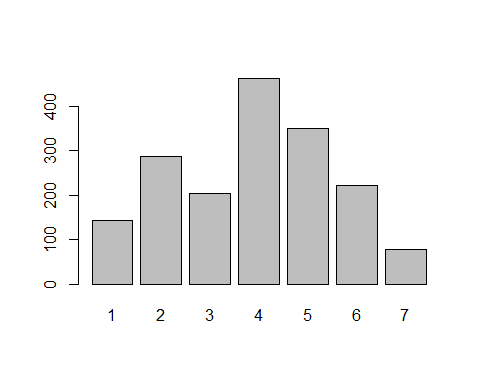
We will split the data into testing and training samples for our detailed analysis. Due to RAM constraints, we will use a small subsection of the large data set.

set.seed(42)  
index1<- sample((1/100)\*nrow(data), replace = FALSE)  
data<- data[index1,]  
forestN<-select(data,-c(treeName,areaName))  
sample\_size=round(nrow(forestN)\*.70)  
index<-sample(seq\_len(nrow(forestN)),size=sample\_size)  
train<-forestN[index, ]  
test<-forestN[-index, ]

## K-Nearest Neighbour Classification

The K-nearest neighbour algorithm will be used to predict the tree type based on other variables.

k.predict<-knn(train = train[,1:12], test = test[,1:12], cl = as.factor(train$treeType), k = 3)  
plot(k.predict)

 The value for k (number of neighbours) was chosen by running the prediction model with different k values each time. Three was chosen as the best k value because it had the highest accuracy out of all the values tested.

Now let’s look at the confusion matrix to determine the accuracy and efficiency of the model.

confMatrixK<-table(k.predict,test$treeType)  
confusionMatrix(confMatrixK)

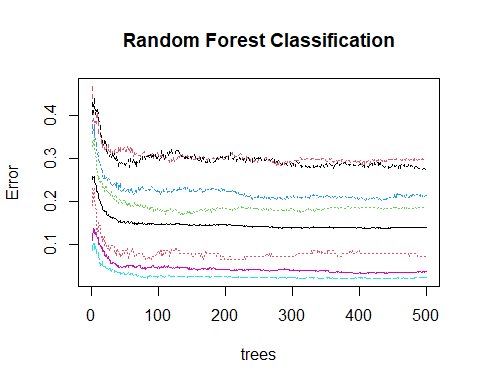
## Confusion Matrix and Statistics  
##   
##   
## k.predict 1 2 3 4 5 6 7  
## 1 100 40 0 0 2 0 2  
## 2 56 219 2 0 5 3 1  
## 3 0 3 165 10 0 26 0  
## 4 0 0 45 399 1 17 0  
## 5 11 42 2 0 289 5 0  
## 6 0 8 27 9 1 176 0  
## 7 4 4 0 0 0 0 69  
##   
## Overall Statistics  
##   
## Accuracy : 0.813   
## 95% CI : (0.7939, 0.831)  
## No Information Rate : 0.2398   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.7746   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6  
## Sensitivity 0.58480 0.6930 0.68465 0.9545 0.9698 0.7753  
## Specificity 0.97201 0.9530 0.97403 0.9525 0.9585 0.9703  
## Pos Pred Value 0.69444 0.7657 0.80882 0.8636 0.8281 0.7964  
## Neg Pred Value 0.95560 0.9334 0.95062 0.9852 0.9935 0.9665  
## Prevalence 0.09811 0.1813 0.13827 0.2398 0.1710 0.1302  
## Detection Rate 0.05737 0.1256 0.09466 0.2289 0.1658 0.1010  
## Detection Prevalence 0.08262 0.1641 0.11704 0.2651 0.2002 0.1268  
## Balanced Accuracy 0.77840 0.8230 0.82934 0.9535 0.9641 0.8728  
## Class: 7  
## Sensitivity 0.95833  
## Specificity 0.99521  
## Pos Pred Value 0.89610  
## Neg Pred Value 0.99820  
## Prevalence 0.04131  
## Detection Rate 0.03959  
## Detection Prevalence 0.04418  
## Balanced Accuracy 0.97677

The accuracy of the model is 81%.

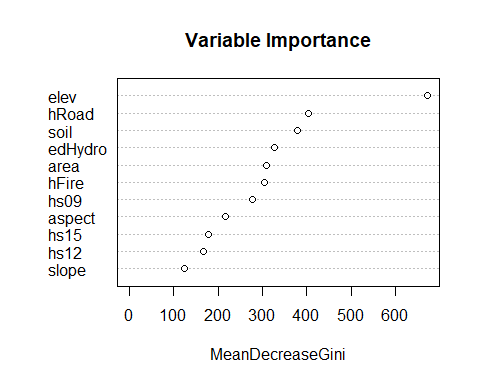
## Random Forest Classification

Now we will look at the Random Forest Classifier.

classifier=randomForest(factor(train$treeType)~.,data=train)  
plot(classifier,main="Random Forest Classification")

 Now we will use the classifier to predict tree type.

rf.predict<-predict(classifier,test)  
varImpPlot(classifier,main="Variable Importance")

 The Variable importance plot shows the mean decrease in node impurity. Therefore elevation is an important indicator of tree type.

Looking at the confusion matrix.

confMatrixTree<-table(rf.predict,test$treeType)  
confusionMatrix(confMatrixTree)

## Confusion Matrix and Statistics  
##   
##   
## rf.predict 1 2 3 4 5 6 7  
## 1 96 37 0 0 1 0 3  
## 2 61 251 0 0 11 1 0  
## 3 0 3 197 6 0 42 0  
## 4 0 0 23 410 0 16 0  
## 5 7 23 1 0 282 4 0  
## 6 0 1 20 2 4 164 0  
## 7 7 1 0 0 0 0 69  
##   
## Overall Statistics  
##   
## Accuracy : 0.8428   
## 95% CI : (0.8249, 0.8596)  
## No Information Rate : 0.2398   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.8106   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6  
## Sensitivity 0.56140 0.7943 0.8174 0.9809 0.9463 0.72247  
## Specificity 0.97392 0.9488 0.9660 0.9706 0.9758 0.98219  
## Pos Pred Value 0.70073 0.7747 0.7944 0.9131 0.8896 0.85864  
## Neg Pred Value 0.95330 0.9542 0.9706 0.9938 0.9888 0.95941  
## Prevalence 0.09811 0.1813 0.1383 0.2398 0.1710 0.13024  
## Detection Rate 0.05508 0.1440 0.1130 0.2352 0.1618 0.09409  
## Detection Prevalence 0.07860 0.1859 0.1423 0.2576 0.1819 0.10958  
## Balanced Accuracy 0.76766 0.8716 0.8917 0.9757 0.9610 0.85233  
## Class: 7  
## Sensitivity 0.95833  
## Specificity 0.99521  
## Pos Pred Value 0.89610  
## Neg Pred Value 0.99820  
## Prevalence 0.04131  
## Detection Rate 0.03959  
## Detection Prevalence 0.04418  
## Balanced Accuracy 0.97677

The accuracy of the Random Forest model is 84%. ```