Team Checkpoint 1 R Code

Appendix

Annie Condon, Matt Hayden, Matt Robertson, Yvette Gonzalez

```r
# read in the data, create dataframe
gz = gzfile('data/covtype.data.gz','rt')
forest.orig = read.csv(gz,header=F)
forest.orig.colnames = t(read.csv('data/covtyp.colnames.csv',header=F))
colnames(forest.orig) = forest.orig.colnames

# identify continuous variables
forest.var.continuous = c('Elevation','Aspect','Slope', 'HDist.Hydrology', 'VDist.Hydrology',
                          'HDist.Roadway', 'Hillshade.9am', 'Hillshade.12pm', 'Hillshade.3pm',
                          'HDist.FirePoint')

# for speed, will perform eda on subset until ready to do a full run
set.seed(33)
forest = forest.orig[sample(nrow(forest.orig),20000),]
forest.var.discrete.indices = grep("^Area|^SoilType|CoverType", colnames(forest))
forest[,forest.var.discrete.indices] = as.factor(unlist(forest[,forest.var.discrete.indices]))

forest.numeric = as.data.frame(sapply(forest,as.numeric))

covertype.names = c('Spruce-fir','Lodgepole Pine','Ponderosa Pine','Cottonwood-Willow','Aspen','Douglas-

forest$CoverType[forest$CoverType==1] = 'Spruce-fir'
forest$CoverType[forest$CoverType==2] = 'Lodgepole Pine'
forest$CoverType[forest$CoverType==3] = 'Ponderosa Pine'
forest$CoverType[forest$CoverType==4] = 'Cottonwood-Willow'
forest$CoverType[forest$CoverType==5] = 'Aspen'
forest$CoverType[forest$CoverType==6] = 'Douglas-fir'
forest$CoverType[forest$CoverType==7] = 'Krummholz'


# add Area column
# are any in multiple areas? NO. all belong to only one area
idx = grep("Area", colnames(forest))
temp = forest.numeric[,idx]
temp.rows = temp[apply(temp,1,sum) > 1,]
temp$Z.Area = apply(temp,1,function(x) {
  return(which.max(x))
})
forest.numeric$Z.Area = temp$Z.Area

forest.scaled = as.data.frame(scale(forest.numeric))

library(lattice)
library(ggplot2)
library(corrplot)
library(MASS)
```

```r
cover_types <- c("Spruce/Fir", "Lodgepole Pine", "Ponderosa Pine", "Cottonwood/Willow", "Aspen",
                 "Douglas/Fir", "Krummholz")
instances <- c(211840, 283301, 35754, 2747, 9493, 17367, 20510)
ct.df <- data.frame(cover_types, instances)
colnames(ct.df) <- c("Cover Type", "Number of Observations")
```

```
knitr::kable(ct.df, caption = "Cover Type Classes", format="pandoc")
```

```
features <- c("Elevation", "Aspect", "Slope", "HDist.Hydrology", "VDist.Hydrology",
              "HDist.Roadway", "Hillshade.9am", "Hillshade.12pm", "Hillshade.3pm",
              "HHDist.FirePoint", "Area", "SoilType")
descriptions <- c("Elevation in meters", "Aspect in degrees aziumuth", "Slope in degrees",
                 "Horizontal distance to nearest surface water feature in meters",
                 "Vertical distance to nearest surface water feature in meters",
                 "Horizontal distance to nearest roadway in meters", "Hillshade index at 9am, summer sol
                "Hillshade index at noon, summer soltice", "Hillshade index at 3pm, summer solstice",
              "Horizontal Distance to nearest wildfire ignition points", "Wilderness area designation -
              "Soil Type designation - 40 binary values")
features.df <- data.frame(features, descriptions)
colnames(features.df) <- c("Feature", "Descriptions")
```

```
knitr::kable(features.df, caption = "Features", format="pandoc")
```

```
library(scales)

ggplot(as.data.frame(table(forest.orig$CoverType)), aes(x=Var1, y = Freq)) + ggtitle("Forest Cover
Frequency by Class") + geom_bar(stat = "identity", fill="#1f78b4", width=.5,
                                color="black") + xlab("Cover Type") + scale_y_continuous(name="Frequenc
```

```
options(digits=3)
my.summary <- function(x,...){
  c(mean=round(mean(x, ...), digits = 4),
    sd=round(sd(x, ...), digits=4),
    median=median(x, ...),
    min=min(x, ...),
    max=max(x,...),
    nmiss=sum(is.na(x,...)),
    type="Continuous")
}


forest.stats= apply(forest.orig[,1:10], 2, my.summary)

library(knitr)
kable(t(forest.stats), caption= "Summary Statistics for Continuous Data", format="pandoc")
```

```
## area
forest.area= forest.orig[11:14]

summary.area <- data.frame(
  Name = character(),
  Count = numeric(),
  stringsAsFactors = F)

for (i in 1:4){
  summary.area[i,1] <- names(forest.area[i])
  summary.area[i,2] <- sum(forest.area[,i])
}

area<-summary.area[with(summary.area,order(-Count)),]
kable(area,caption= "Area Type Counts", format="pandoc", row.names = F)
```

```r
## soil
forest.soil= forest.orig[15:54]

summary.soil <- data.frame(
  Name = character(),
  Count = numeric(),
  stringsAsFactors = F)

for (i in 1:40){
  summary.soil[i,1] <- names(forest.soil[i])
  summary.soil[i,2] <- sum(forest.soil[,i])
}


soil<-summary.soil[with(summary.soil,order(-Count)),]
kable(soil, caption= "Soil Type Counts", format="pandoc", row.names = F)
```

```r
st = stack(as.data.frame(forest.scaled[,forest.var.continuous]))
ggplot(as.data.frame(st)) +
  geom_boxplot(aes(x = ind, y = values)) +
  theme(axis.text.x = element_text(angle=45, hjust = 1)) +
  scale_x_discrete(name ="") + scale_y_continuous(name ="") +
  ggtitle("Boxplots of Scaled Continuous Variables")
```

```r
density.plots = densityplot(~ Elevation + Aspect + Slope + HDist.Hydrology + VDist.Hydrology +
                              HDist.Roadway + Hillshade.9am + Hillshade.12pm + Hillshade.3pm +
                              HDist.FirePoint,
                            data=forest,
                            groups = CoverType,
                            plot.points = FALSE,
                            auto.key = list(space="right",title="Cover Type",cex=.6),
                            scales= list(x="free",y="free"),
                            xlab = '',
                            ylab=list(cex=.8),
                            aspect="fill",
                            par.strip.text=list(cex=.9))
plot(density.plots)
```

```r
density.plots = densityplot(~ Elevation + Aspect + Slope + HDist.Hydrology + VDist.Hydrology +
                              HDist.Roadway + Hillshade.9am + Hillshade.12pm + Hillshade.3pm +
                              HDist.FirePoint,
                            data=forest.numeric,
                            groups = Z.Area,
                            plot.points = FALSE,
                            auto.key = list(space="right",title="Area",cex=.6),
                            scales= list(x="free",y="free"),
                            xlab = '',
                            ylab=list(cex=.8),
                            aspect="fill",
                            par.strip.text=list(cex=.9))
plot(density.plots)
```

```r
corrplot(cor(forest[, forest.var.continuous]),
         tl.col = "black", tl.cex = 0.8, tl.srt = 45,
         cl.cex = 0.8, pch.cex = 0.8, diag = FALSE,
         type="lower",
         addCoefasPercent = TRUE, addCoef.col = TRUE,number.cex = .6) #Matt added to show correlation a
```

```
idx = grep("SoilType|CoverType", colnames(forest.numeric))
df = as.data.frame(forest.numeric[,idx])
idx.type = grep("CoverType", colnames(df))

df.temp = df[,-idx.type]

soil.sums = apply(df.temp,2,function(x) {
  tbl = table(x,df$CoverType)
  if (dim(tbl)[1] < 2) {
    tbl = rbind('0' = tbl, '1' = rep(0,7), deparse.level = 1)
  }
  return (apply(tbl,1,sum)[2])
})
#soil.sums

soil.sums.byclass = apply(df[,-idx.type],2,function(x) {
  tbl = table(x,df$CoverType)
  tbl = tbl[seq(2,14,by=2)]
  return (tbl)
})
#soil.sums.byclass

soil.ratios = as.data.frame(t(soil.sums.byclass)/soil.sums)
library(RColorBrewer)
soil.ratios.m = na.omit(as.matrix(soil.ratios))
barchart(soil.ratios.m,col=brewer.pal(7, "Pastel2"),xlab='',
         key=list(space="right",
                  lines=list(col=brewer.pal(7, "Pastel2"),lwd=3),
                  text=list(covertype.names)
))
```

```
idx = grep("Area1|Area2|Area3|Area4|CoverType", colnames(forest.numeric))
df = as.data.frame(forest.numeric[,idx])
idx.type = grep("CoverType", colnames(df))

df.temp = df[,-idx.type]

area.sums = apply(df.temp,2,function(x) {
  tbl = table(x,df$CoverType)
  if (dim(tbl)[1] < 2) {
    tbl = rbind('0' = tbl, '1' = rep(0,7), deparse.level = 1)
  }
  return (apply(tbl,1,sum)[2])
})
#area.sums

area.sums.byclass = apply(df[,-idx.type],2,function(x) {
  tbl = table(x,df$CoverType)
  tbl = tbl[seq(2,14,by=2)]
  return (tbl)
```

```r
})
#area.sums.byclass

area.ratios = as.data.frame(t(area.sums.byclass)/area.sums)
library(RColorBrewer)
area.ratios.m = na.omit(as.matrix(area.ratios))
barchart(area.ratios.m,col=brewer.pal(7, "Pastel2"),xlab='',
         key=list(space="right",
                  lines=list(col=brewer.pal(7, "Pastel2"),lwd=3),
                  text=list(covertype.names)
                  )
)
```

```r
ggplot(forest, aes(x=HDist.Roadway)) +
  geom_histogram(aes(group=CoverType, colour=CoverType, fill=CoverType), bins=30, alpha=0.7)+
  ggtitle('')+
  theme(legend.title = element_blank())+
  labs(x="Distance to Roadway",y="Count")
```

```r
ggplot(forest, aes(x=HDist.Hydrology)) +
  geom_histogram(aes(group=CoverType, colour=CoverType, fill=CoverType), bins=30, alpha=0.7)+
  ggtitle('')+
  theme(legend.title = element_blank())+
  labs(x="Distance to Water",y="Count")
```