Forest Cover EDA

wilderness\_areas <- tbl\_df(data.frame(  
 wa = stringr::str\_c("wilderness\_area", 1:4),  
 wilderness\_area = c("Rawah Wilderness Area",  
 "Neota Wilderness Area",  
 "Comanche Peak Wilderness Area",  
 "Cache la Poudre Wilderness Area"),  
 stringsAsFactors = FALSE  
))  
  
soil\_types <- tbl\_df(data.frame(  
 st = stringr::str\_c("soil\_type", 1:40),  
 soil\_type = c("Cathedral family - Rock outcrop complex, extremely stony",  
 "Vanet - Ratake families complex, very stony",  
 "Haploborolis - Rock outcrop complex, rubbly",  
 "Ratake family - Rock outcrop complex, rubbly",  
 "Vanet family - Rock outcrop complex complex, rubbly",  
 "Vanet - Wetmore families - Rock outcrop complex, stony",  
 "Gothic family",  
 "Supervisor - Limber families complex",  
 "Troutville family, very stony",  
 "Bullwark - Catamount families - Rock outcrop complex, rubbly",  
 "Bullwark - Catamount families - Rock land complex, rubbly.",  
 "Legault family - Rock land complex, stony",  
 "Catamount family - Rock land - Bullwark family complex, rubbly",  
 "Pachic Argiborolis - Aquolis complex",  
 "unspecified in the USFS Soil and ELU Survey",  
 "Cryaquolis - Cryoborolis complex",  
 "Gateview family - Cryaquolis complex",  
 "Rogert family, very stony",  
 "Typic Cryaquolis - Borohemists complex",  
 "Typic Cryaquepts - Typic Cryaquolls complex",  
 "Typic Cryaquolls - Leighcan family, till substratum complex",  
 "Leighcan family, till substratum, extremely bouldery",  
 "Leighcan family, till substratum - Typic Cryaquolls complex",  
 "Leighcan family, extremely stony",  
 "Leighcan family, warm, extremely stony",  
 "Granile - Catamount families complex, very stony",  
 "Leighcan family, warm - Rock outcrop complex, extremely stony",  
 "Leighcan family - Rock outcrop complex, extremely stony",  
 "Como - Legault families complex, extremely stony",  
 "Como family - Rock land - Legault family complex, extremely stony",  
 "Leighcan - Catamount families complex, extremely stony",  
 "Catamount family - Rock outcrop - Leighcan family complex, extremely stony",  
 "Leighcan - Catamount families - Rock outcrop complex, extremely stony",  
 "Cryorthents - Rock land complex, extremely stony",  
 "Cryumbrepts - Rock outcrop - Cryaquepts complex",  
 "Bross family - Rock land - Cryumbrepts complex, extremely stony",  
 "Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony",  
 "Leighcan - Moran families - Cryaquolls complex, extremely stony",  
 "Moran family - Cryorthents - Leighcan family complex, extremely stony",  
 "Moran family - Cryorthents - Rock land complex, extremely stony"),  
 stringsAsFactors = FALSE))

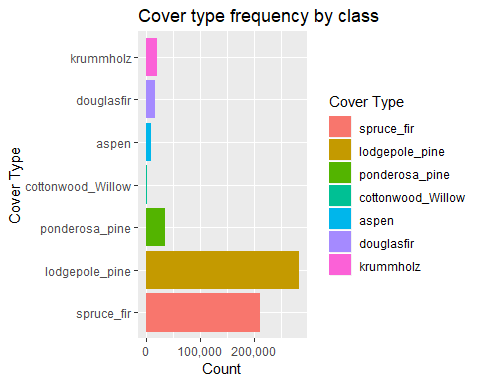
# reshape soil types and convert to labels  
raw\_data <- raw\_data %>%  
 tidyr::gather(st, st\_flag, soil\_type1:soil\_type40) %>%  
 filter(st\_flag != 0) %>%  
 left\_join(soil\_types, by = "st") %>%  
 dplyr:: select(-st, -st\_flag) %>%  
 mutate(soil\_type = as.factor(soil\_type))  
  
# reshape wilderness area and convert to labels  
raw\_data <- raw\_data %>%  
 tidyr::gather(wa, wa\_flag, wilderness\_area1:wilderness\_area4) %>%  
 filter(wa\_flag != 0) %>%  
 left\_join(wilderness\_areas, by = "wa") %>%  
 dplyr ::select(-wa, -wa\_flag) %>%  
 mutate(wilderness\_area = as.factor(wilderness\_area))  
  
dim(raw\_data)

## [1] 581012 13

str(raw\_data)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 581012 obs. of 13 variables:  
## $ elevation : int 2903 2906 2906 2906 2906 2909 2906 2906 2909 2909 ...  
## $ aspect : int 315 45 11 0 333 18 72 0 45 135 ...  
## $ slope : int 2 1 4 3 3 2 2 0 1 5 ...  
## $ horizontal\_distance\_to\_hydrology : int 330 371 360 330 379 360 330 300 390 360 ...  
## $ vertical\_distance\_to\_hydrology : int 0 6 3 3 6 6 3 3 5 6 ...  
## $ horizontal\_distance\_to\_roadways : int 4734 4668 4696 4725 4658 4687 4715 4743 4649 4677 ...  
## $ hillshade\_9am : int 214 219 216 215 212 217 222 218 219 228 ...  
## $ hillshade\_noon : int 237 236 232 234 235 234 234 238 236 238 ...  
## $ hillshade\_3pm : int 161 154 153 156 161 154 149 156 154 144 ...  
## $ horizontal\_distance\_to\_fire\_points: int 4987 5007 4986 4966 4986 4965 4944 4923 4964 4943 ...  
## $ cover\_type : Factor w/ 7 levels "spruce\_fir","lodgepole\_pine",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ soil\_type : Factor w/ 40 levels "Bross family - Rock land - Cryumbrepts complex, extremely stony",..: 13 13 13 13 13 13 13 13 13 13 ...  
## $ wilderness\_area : Factor w/ 4 levels "Cache la Poudre Wilderness Area",..: 4 4 4 4 4 4 4 4 4 4 ...

# Create frequency plot of cover types (Over 80% belong to two classes)  
raw\_data %>%  
 ggplot(aes\_string(x = raw\_data$cover\_type, fill=as.factor(raw\_data$cover\_type))) +  
 geom\_bar() + coord\_flip() +  
 scale\_y\_continuous(labels = scales::comma) +  
 labs(x = "Cover Type",  
 y = "Count",  
 title = "Cover type frequency by class") +   
 scale\_fill\_discrete(name="Cover Type")



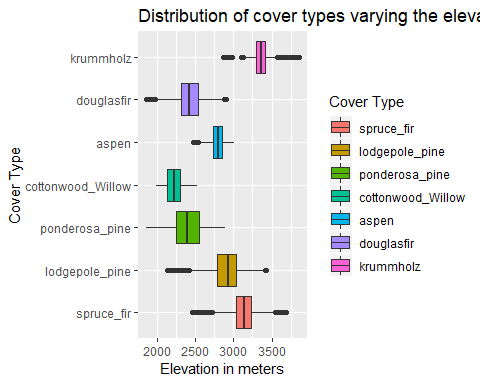
## Including Plots

You can also embed plots, for example:

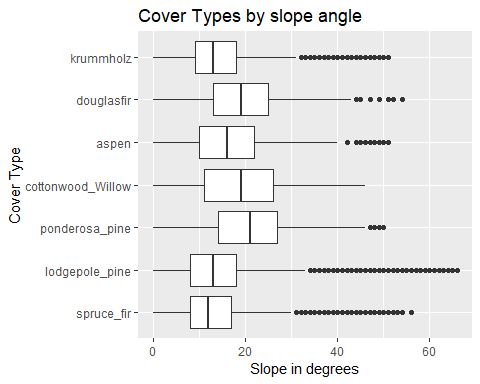
# Check Correlations   
mc <- cor(raw\_data[which(sapply(raw\_data, is.numeric))], use = "complete.obs")  
mc[upper.tri(mc, diag = TRUE)] <- NA  
mc %>%  
 abs() %>%  
 data.frame %>%  
 tbl\_df() %>%  
 mutate(var1 = row.names(mc)) %>%  
 gather(var2, cor, -var1) %>%  
 na.omit() %>%  
 arrange(desc(cor)) %>%  
 filter(cor > 0.5) %>%  
 knitr::kable(digits = 3)

|  |  |  |
| --- | --- | --- |
| var1 | var2 | cor |
| hillshade\_3pm | hillshade\_9am | 0.780 |
| hillshade\_3pm | aspect | 0.647 |
| vertical\_distance\_to\_hydrology | horizontal\_distance\_to\_hydrology | 0.606 |
| hillshade\_3pm | hillshade\_noon | 0.594 |
| hillshade\_9am | aspect | 0.579 |
| hillshade\_noon | slope | 0.527 |

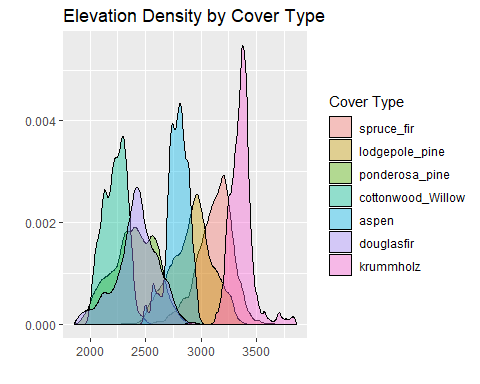
# Boxplots of elevation   
raw\_data %>%  
 ggplot(aes\_string(x = raw\_data$cover\_type, y = "elevation", fill = as.factor(raw\_data$cover\_type))) +  
 geom\_boxplot() +  
 coord\_flip() +  
 labs(x = "Cover Type",  
 y = "Elevation in meters",  
 title = "Distribution of cover types varying the elevation") +  
scale\_fill\_discrete(name = "Cover Type")



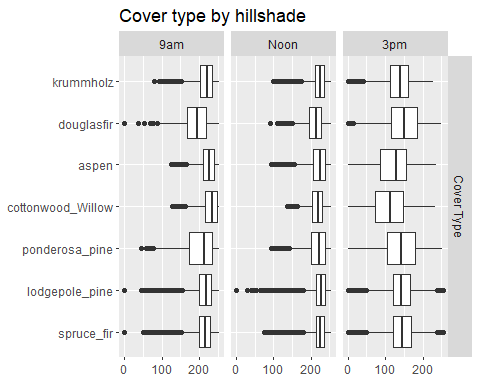
# Boxplots of slope   
raw\_data %>%  
 ggplot(aes\_string(x = raw\_data$cover\_type, y = "slope")) +  
 geom\_boxplot() +  
 coord\_flip() +  
 labs(x = "Cover Type",  
 y = "Slope in degrees",  
 title = "Cover Types by slope angle")



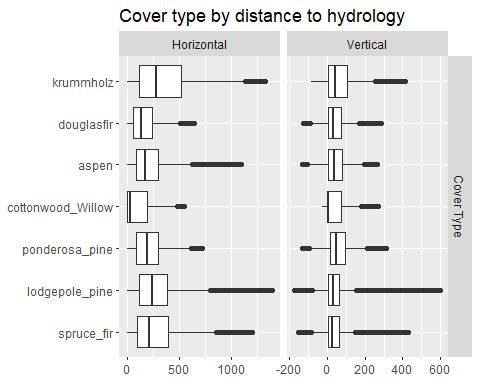
# Density plot of elevation by Cover Type  
ggplot(raw\_data, aes(raw\_data$elevation, fill=as.factor(raw\_data$cover\_type))) +  
 geom\_density(alpha=0.4) +  
 labs(title="Elevation Density by Cover Type", x="", y="") +  
 scale\_fill\_discrete(name="Cover Type")



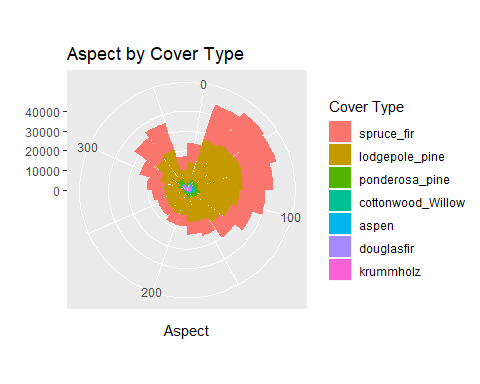
# list of shade plots at all times of the day  
plot\_list <- list()  
plot\_list[[1]] <- raw\_data %>%  
 ggplot(aes\_string(x = raw\_data$cover\_type, y = "hillshade\_9am")) +  
 geom\_boxplot() + coord\_flip()  
plot\_list[[2]] <- raw\_data %>%  
 ggplot(aes\_string(x = raw\_data$cover\_type, y = "hillshade\_noon")) +  
 geom\_boxplot() + coord\_flip()  
plot\_list[[3]] <- raw\_data %>%  
 ggplot(aes\_string(x = raw\_data$cover\_type, y = "hillshade\_3pm")) +  
 geom\_boxplot() + coord\_flip()  
  
# plot all plots in a grid  
GGally::ggmatrix(plot\_list,  
 nrow = 1,  
 ncol = 3,  
 xAxisLabels = c("9am", "Noon", "3pm"),  
 yAxisLabels = "Cover Type",  
 title = "Cover type by hillshade")



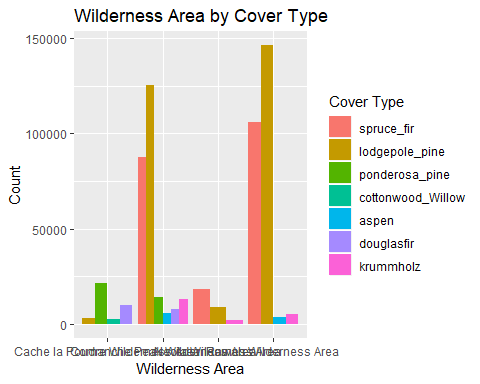
# list of hydrology plots  
plot\_list <- list()  
plot\_list[[1]] <- raw\_data %>%  
 ggplot(aes\_string(x = raw\_data$cover\_type, y = "horizontal\_distance\_to\_hydrology")) +  
 geom\_boxplot() + coord\_flip()  
plot\_list[[2]] <- raw\_data %>%  
 ggplot(aes\_string(x = raw\_data$cover\_type, y = "vertical\_distance\_to\_hydrology")) +  
 geom\_boxplot() + coord\_flip()  
  
# plot all plots in a grid  
GGally::ggmatrix(plot\_list,  
 nrow = 1,  
 ncol = 2,  
 xAxisLabels = c("Horizontal", "Vertical"),  
 yAxisLabels = "Cover Type",  
 title = "Cover type by distance to hydrology")



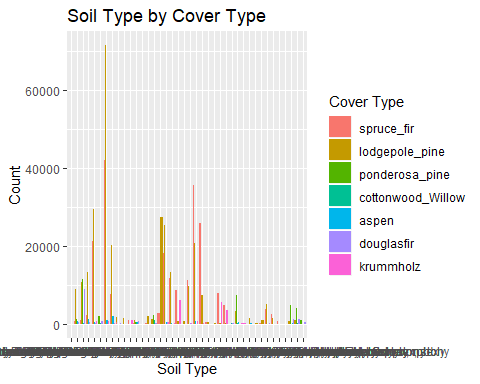
# Aspect by Cover Type  
ggplot(raw\_data, aes(aspect, fill=as.factor(raw\_data$cover\_type))) +  
 geom\_histogram(bins = 20) +  
 coord\_polar() +  
 labs(title="Aspect by Cover Type", x="Aspect", y="") +  
 scale\_fill\_discrete(name="Cover Type")



# Wilderness Area by Cover Type  
ggplot(raw\_data, aes(wilderness\_area, fill=as.factor(cover\_type))) +  
 geom\_bar(position = "dodge") +  
 labs(title="Wilderness Area by Cover Type", x="Wilderness Area", y="Count") +  
 scale\_fill\_discrete(name="Cover Type")



# Soil Types by Cover Type  
ggplot(raw\_data, aes(soil\_type, fill=as.factor(cover\_type))) +  
 geom\_bar(position = "dodge") +  
 labs(title="Soil Type by Cover Type", x="Soil Type", y="Count") +  
 scale\_fill\_discrete(name="Cover Type")



train\_index <- caret::createDataPartition(y = raw\_data$cover\_type,  
 p = 0.7,  
 times = 1,  
 list = F)  
head(train\_index)

## Resample1  
## [1,] 2  
## [2,] 5  
## [3,] 6  
## [4,] 7  
## [5,] 10  
## [6,] 11

raw\_train <- raw\_data[train\_index,]  
raw\_test <- raw\_data[-train\_index,]  
  
raw\_train %>% tabyl(cover\_type) %>% adorn\_pct\_formatting()

## cover\_type n percent  
## spruce\_fir 148288 36.5%  
## lodgepole\_pine 198311 48.8%  
## ponderosa\_pine 25028 6.2%  
## cottonwood\_Willow 1923 0.5%  
## aspen 6646 1.6%  
## douglasfir 12157 3.0%  
## krummholz 14357 3.5%

raw\_test %>% tabyl(cover\_type) %>% adorn\_pct\_formatting()

## cover\_type n percent  
## spruce\_fir 63552 36.5%  
## lodgepole\_pine 84990 48.8%  
## ponderosa\_pine 10726 6.2%  
## cottonwood\_Willow 824 0.5%  
## aspen 2847 1.6%  
## douglasfir 5210 3.0%  
## krummholz 6153 3.5%

dim(raw\_train)

## [1] 406710 13

dim(raw\_test)

## [1] 174302 13

# Create a data.table from data.frame  
cover\_data <- as.data.table(raw\_train)  
rm(raw\_train)  
#run model in parallel  
cl <- makeCluster(detectCores())  
registerDoParallel(cl)  
# Run a base forest model for EDA   
base\_forest <- randomForest(cover\_type ~ ., data=cover\_data, mtry=sqrt(ncol(cover\_data)),   
 ntree = 300, importance =T, do.trace=50)

## ntree OOB 1 2 3 4 5 6 7  
## 50: 3.96% 4.49% 2.66% 3.76% 13.78% 17.94% 9.00% 4.69%  
## 100: 3.69% 4.26% 2.40% 3.43% 12.95% 17.48% 8.21% 4.62%  
## 150: 3.59% 4.21% 2.30% 3.28% 13.10% 17.41% 7.95% 4.33%  
## 200: 3.55% 4.13% 2.28% 3.22% 13.00% 17.32% 7.86% 4.36%  
## 250: 3.54% 4.13% 2.26% 3.23% 13.16% 17.32% 7.88% 4.28%  
## 300: 3.51% 4.09% 2.24% 3.24% 12.95% 17.18% 7.69% 4.28%

stopCluster(cl)  
base\_forest

##   
## Call:  
## randomForest(formula = cover\_type ~ ., data = cover\_data, mtry = sqrt(ncol(cover\_data)), ntree = 300, importance = T, do.trace = 50)   
## Type of random forest: classification  
## Number of trees: 300  
## No. of variables tried at each split: 4  
##   
## OOB estimate of error rate: 3.51%  
## Confusion matrix:  
## spruce\_fir lodgepole\_pine ponderosa\_pine  
## spruce\_fir 142223 5731 3  
## lodgepole\_pine 3625 193870 301  
## ponderosa\_pine 1 280 24216  
## cottonwood\_Willow 0 0 196  
## aspen 70 967 77  
## douglasfir 9 240 622  
## krummholz 530 83 0  
## cottonwood\_Willow aspen douglasfir krummholz class.error  
## spruce\_fir 0 53 9 269 0.04090014  
## lodgepole\_pine 1 288 178 48 0.02239412  
## ponderosa\_pine 96 26 409 0 0.03244366  
## cottonwood\_Willow 1674 0 53 0 0.12948518  
## aspen 0 5504 28 0 0.17183268  
## douglasfir 52 12 11222 0 0.07691042  
## krummholz 0 2 0 13742 0.04283625

varImp(base\_forest)

## spruce\_fir lodgepole\_pine  
## elevation 290.08622 221.56528  
## aspect 67.46829 74.50535  
## slope 68.01035 73.02018  
## horizontal\_distance\_to\_hydrology 144.76455 159.89807  
## vertical\_distance\_to\_hydrology 120.81848 154.00421  
## horizontal\_distance\_to\_roadways 203.95933 191.73733  
## hillshade\_9am 100.37866 113.68761  
## hillshade\_noon 93.57337 96.35059  
## hillshade\_3pm 122.37179 126.12487  
## horizontal\_distance\_to\_fire\_points 206.80752 176.99642  
## soil\_type 137.94989 180.67365  
## wilderness\_area 64.55091 51.54334  
## ponderosa\_pine cottonwood\_Willow  
## elevation 107.47126 119.50038  
## aspect 66.90195 29.98461  
## slope 46.01525 21.53110  
## horizontal\_distance\_to\_hydrology 106.19496 70.92562  
## vertical\_distance\_to\_hydrology 93.50719 27.07603  
## horizontal\_distance\_to\_roadways 74.43851 39.21050  
## hillshade\_9am 57.65276 29.77258  
## hillshade\_noon 67.07044 28.29539  
## hillshade\_3pm 71.01789 23.84498  
## horizontal\_distance\_to\_fire\_points 86.49083 28.91835  
## soil\_type 75.98991 48.31819  
## wilderness\_area 40.31166 38.96306  
## aspen douglasfir krummholz  
## elevation 254.67765 230.16336 223.29823  
## aspect 57.91965 61.15201 39.77305  
## slope 42.11180 41.44453 38.09631  
## horizontal\_distance\_to\_hydrology 105.49071 103.20700 79.81431  
## vertical\_distance\_to\_hydrology 73.09343 88.52472 72.69207  
## horizontal\_distance\_to\_roadways 191.87533 104.90760 104.75031  
## hillshade\_9am 66.28225 58.10633 62.20899  
## hillshade\_noon 48.99390 71.66632 52.22443  
## hillshade\_3pm 56.48221 71.53928 58.83941  
## horizontal\_distance\_to\_fire\_points 138.65155 112.98223 121.74928  
## soil\_type 138.63070 88.82503 106.62086  
## wilderness\_area 37.64571 43.81433 117.10260

varImp(base\_forest)

## spruce\_fir lodgepole\_pine  
## elevation 290.08622 221.56528  
## aspect 67.46829 74.50535  
## slope 68.01035 73.02018  
## horizontal\_distance\_to\_hydrology 144.76455 159.89807  
## vertical\_distance\_to\_hydrology 120.81848 154.00421  
## horizontal\_distance\_to\_roadways 203.95933 191.73733  
## hillshade\_9am 100.37866 113.68761  
## hillshade\_noon 93.57337 96.35059  
## hillshade\_3pm 122.37179 126.12487  
## horizontal\_distance\_to\_fire\_points 206.80752 176.99642  
## soil\_type 137.94989 180.67365  
## wilderness\_area 64.55091 51.54334  
## ponderosa\_pine cottonwood\_Willow  
## elevation 107.47126 119.50038  
## aspect 66.90195 29.98461  
## slope 46.01525 21.53110  
## horizontal\_distance\_to\_hydrology 106.19496 70.92562  
## vertical\_distance\_to\_hydrology 93.50719 27.07603  
## horizontal\_distance\_to\_roadways 74.43851 39.21050  
## hillshade\_9am 57.65276 29.77258  
## hillshade\_noon 67.07044 28.29539  
## hillshade\_3pm 71.01789 23.84498  
## horizontal\_distance\_to\_fire\_points 86.49083 28.91835  
## soil\_type 75.98991 48.31819  
## wilderness\_area 40.31166 38.96306  
## aspen douglasfir krummholz  
## elevation 254.67765 230.16336 223.29823  
## aspect 57.91965 61.15201 39.77305  
## slope 42.11180 41.44453 38.09631  
## horizontal\_distance\_to\_hydrology 105.49071 103.20700 79.81431  
## vertical\_distance\_to\_hydrology 73.09343 88.52472 72.69207  
## horizontal\_distance\_to\_roadways 191.87533 104.90760 104.75031  
## hillshade\_9am 66.28225 58.10633 62.20899  
## hillshade\_noon 48.99390 71.66632 52.22443  
## hillshade\_3pm 56.48221 71.53928 58.83941  
## horizontal\_distance\_to\_fire\_points 138.65155 112.98223 121.74928  
## soil\_type 138.63070 88.82503 106.62086  
## wilderness\_area 37.64571 43.81433 117.10260

varImpPlot(base\_forest)

