# Adaptive Mobile: coverType Data Set - Exploratory Analysis

*Brian Carter*

*24 March, 2015*

## Contents



## 1 Excutive Summary

NOTCOMPLETE

This document present an exploration of the data contained Forest CoverType dataset, College of Natural Resources, Colorado State University. An introduction to the information contained in the dataset is presented. The original dataset is described and additional metadata is added for the purposes of more intuitive understanding of the data.

The features of the dataset are analyised. This is performed in the R language. Through examine the data a number of research questions are presented to be answered for the purposes of. EXPAND HERE.

The research questions are answered in Python.

## 2 Introduction

The dataset in its original form contains 581012 rows across 55 columns. The dataset doesn't contain column headers. These are contained in the meta data file **covtype.info** file. These are added to the data. (An excel file with the meta data was created)

XXXXDelete Kaggle Names in metdata.xlsxXXXX XXXXSave with column names?XXXX

Columns 1-10 are quantiative data. Columns 11-14 are a binary resentation of the four possible types of **Wilderness_Area** with a 1 representing the presence of that wilderness type and 0 otherwise. These columns are mutually exclusive, in that, for each row only only of the four columns can contain a 1. For the purpose of exploratory analysis the four columns are reduced to their **original** representation of 1 column.

Similarly columns 15-54 are a binary representation of the 40 possible **Soil_Types**. For the purposes of exploratory analysis, these 40 columns are reduced to their original state with one column REWRITE.

After reducing the binary columns there are `r ncol(coverData)` columns in the dataset. In the **covtype.info** file there is extra meta data associated with the **soilType** column that is not in the original dataset. Each **soilType** has an unique associated 4-digit *ELU.Code*. The first digit of the *ELU.Code* represents its' *Climatic Zone* and the second digit its' *Geological Zone*. This extra information is introduced for exploratory analysis.

The final column represents the **coverType** with seven possible values. **coverType** has been the main predictive target in much analysis concerning this dataset. The lables for this column are used during exploratory analysis.

## 2.1 Data Summary

Having reduced the binary columns representations of **Wilderness_Area, Soil_Type** to their original form and added two additional columns there are 15 columns in the dataset.

A summary of the 15 is presented in Table 1.

| name | dataType | missing | unique | quantity | min |
|---|---|---|---|---|---|
| Elevation | integer | 0 | 1,978 | meters | 1,859 |
| Aspect | integer | 0 | 361 | azimuth | 0 |
| Slope | integer | 0 | 67 | degrees | 0 |
| HD.Hydro | integer | 0 | 551 | meters | 0 |
| VD.Hydro | integer | 0 | 700 | meters | -173 |
| HD.Road | integer | 0 | 5,785 | meters | 0 |
| HD.Fire | integer | 0 | 207 | meters | 0 |
| HS.9am | integer | 0 | 185 | 0-255 | 0 |
| HS.noon | integer | 0 | 255 | 0-255 | 0 |
| HS.3pm | integer | 0 | 5,827 | 0-255 | 0 |
| wildernessArea | character | 0 | 4 | | |
| soilType | character | 0 | 40 | | |
| climaticZone | factor | 0 | 7 | | |
| geologicZone | factor | 0 | 4 | | |
| coverType | factor | 0 | 7 | | |

Table 1: Summary of CoverTye Dataset Featuers (continued below)

| max | mean | std |
|---|---|---|
| 3,858 | 2,959.37 | 279.98 |
| 360 | 155.66 | 111.91 |
| 66 | 14.10 | 7.49 |
| 1,397 | 269.43 | 212.55 |
| 601 | 46.42 | 58.30 |
| 7,117 | 2,350.15 | 1,559.25 |
| 254 | 212.15 | 26.77 |
| 254 | 223.32 | 19.77 |
| 254 | 142.53 | 38.27 |
| 7,173 | 1,980.29 | 1,324.20 |

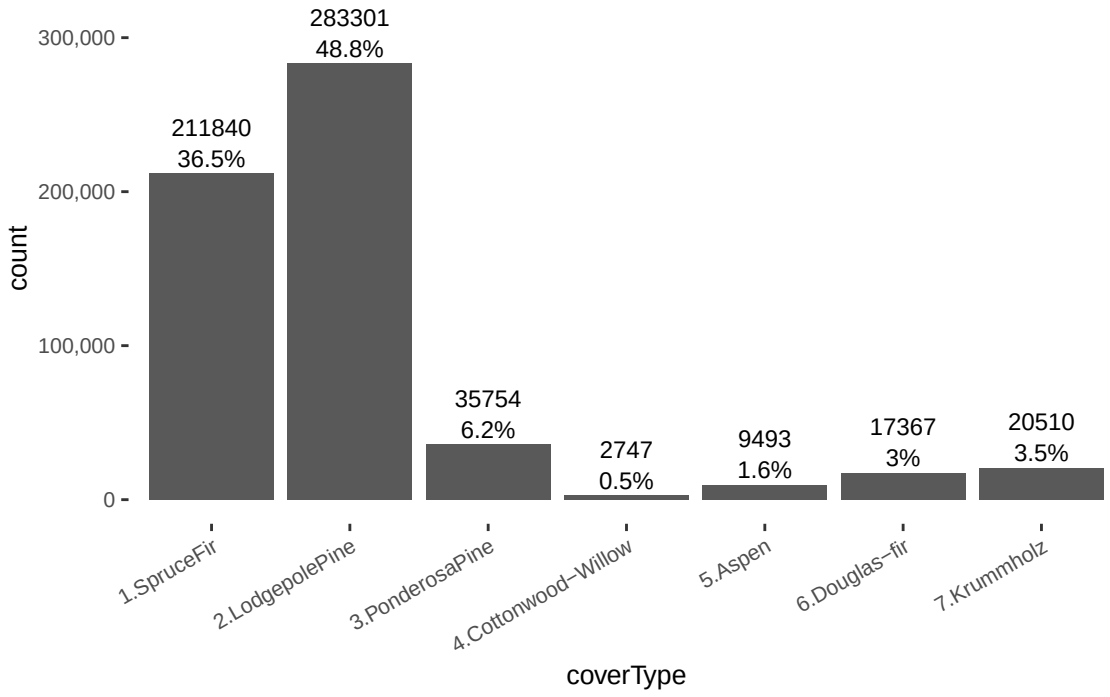
XXXXFix table if have time?XXXX

# 3 Data Exploration

The **covtype.info** file includes refences to a number of papers where predictive models were built with **coverType** as the target variable.

*coverType* is not evenly distributed. Two classes *(SpruceFir, LodgepolePine)* comprimise 85% of all rows. Figure 1 displays a bar graphy of *coverType*.

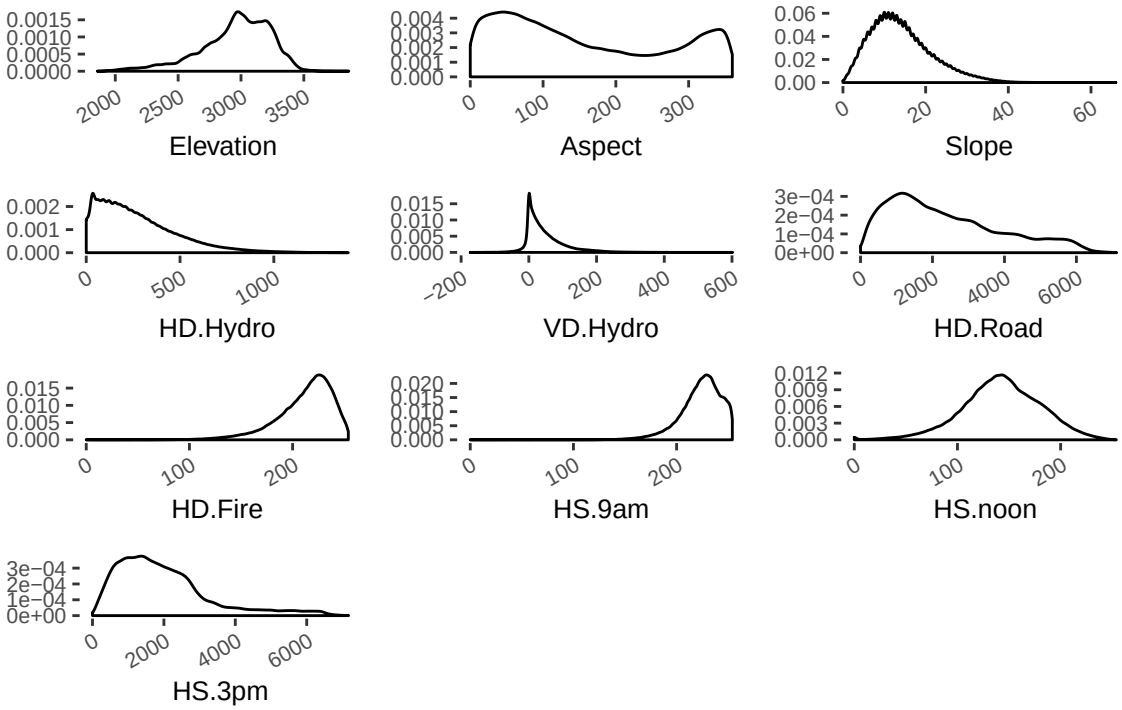
Figure 1: Distribution of coverType




## 3.1 Numerical Data Exploration

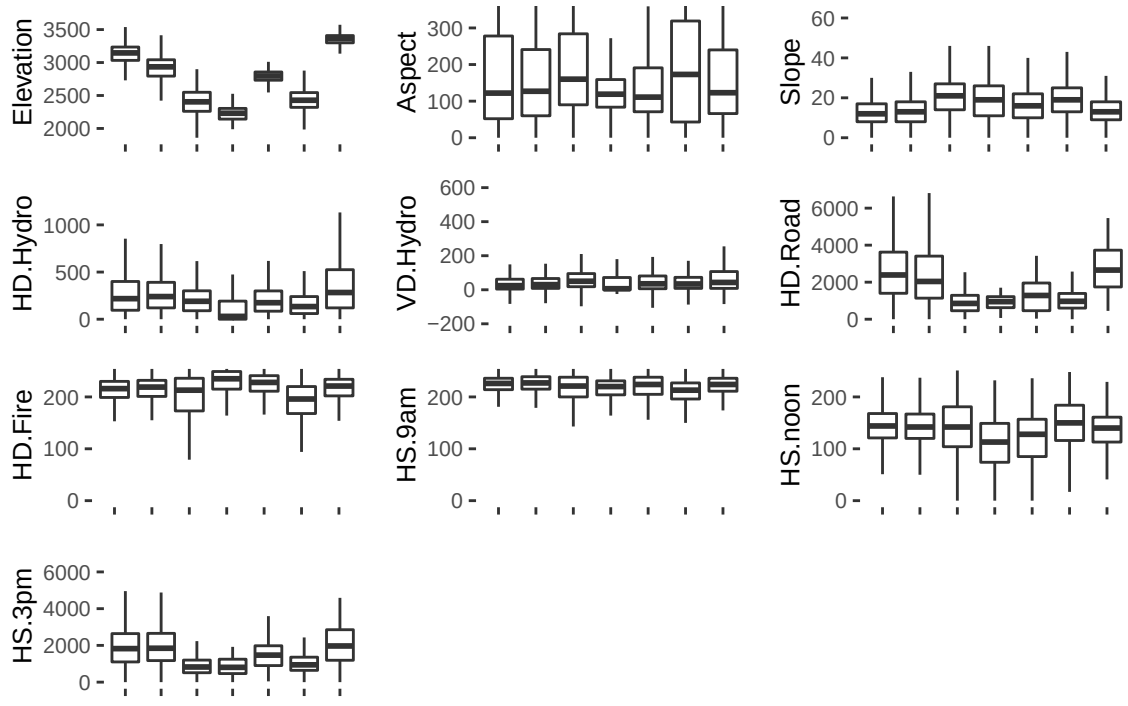
Density Plots

Figure 2: Density Plot of Numeric Data



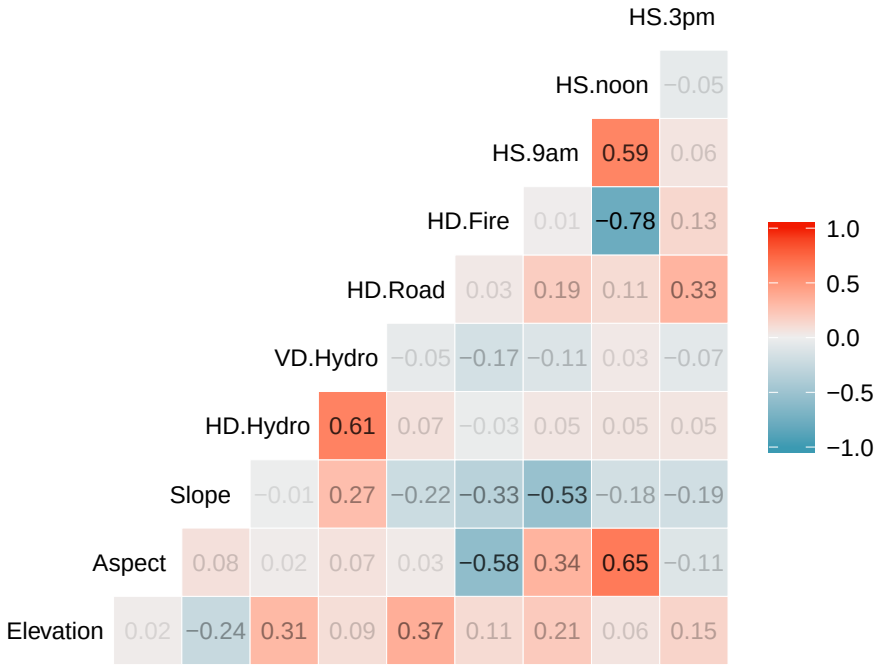

Boxplots

Figure 3: Boxplot of Numeric Data − coverType



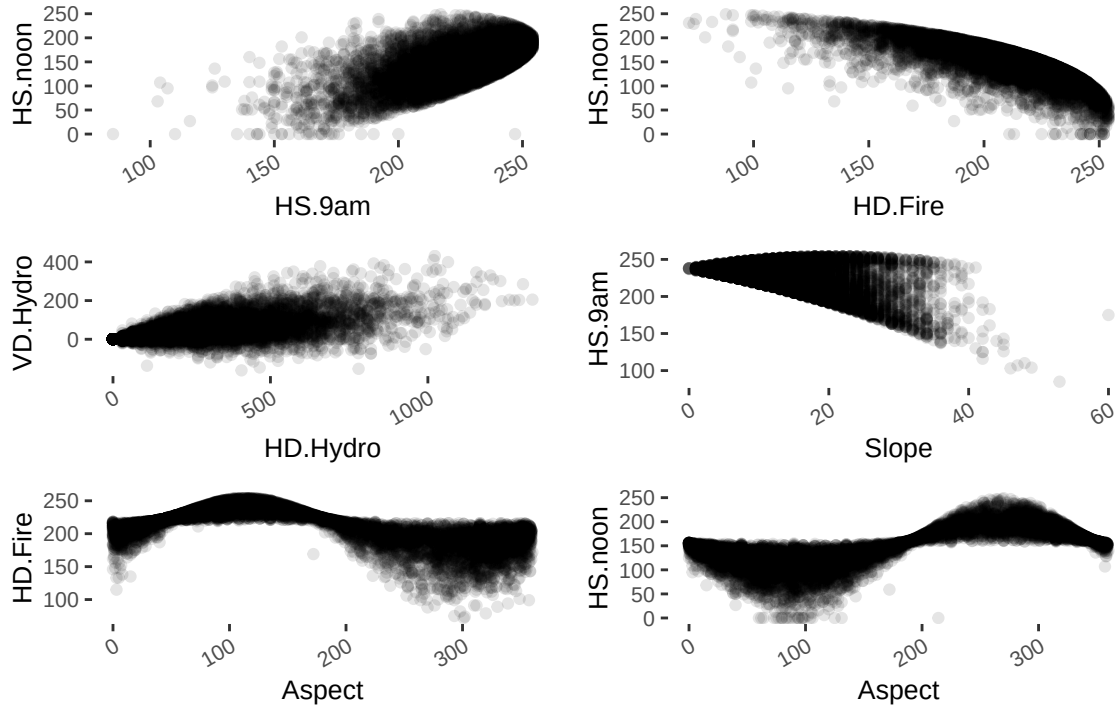

Correlation Matrix

Figure 4: Correlation Matrix of 10 Numerical Features



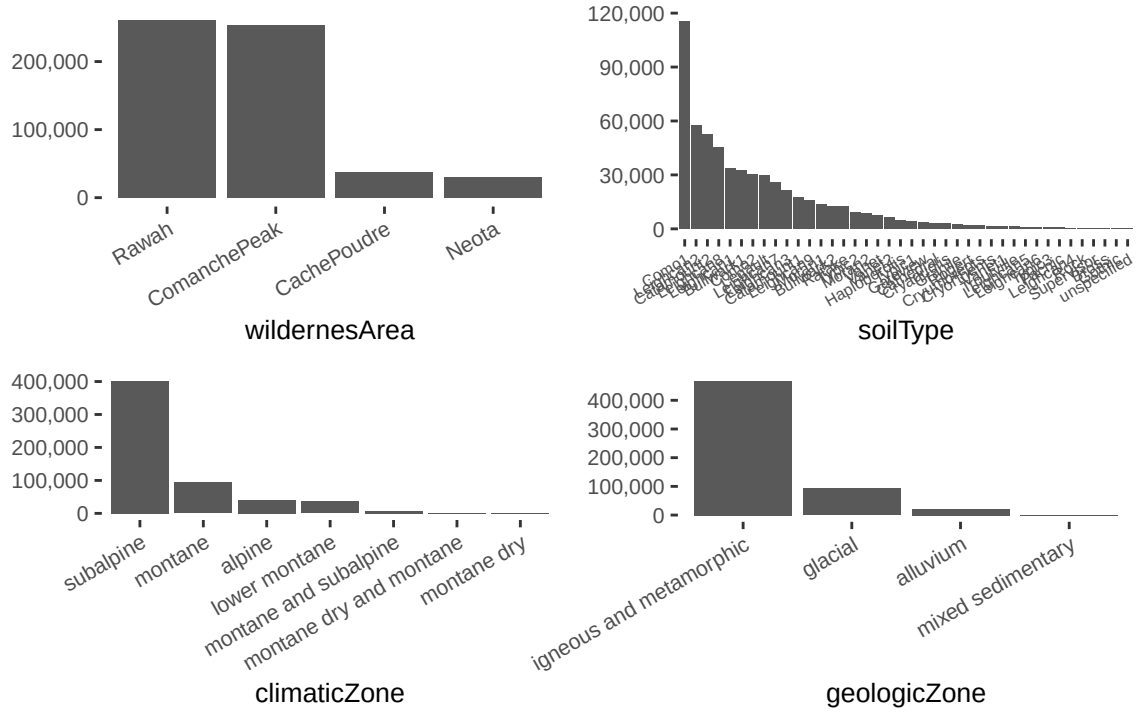

ScatterPlots

Figure 5: Scatterplot of 6 Correlated Features



### 3.2 Categorical Data Exploration

Figure 6: Row Counts of Categorioal Features




FEATURE IMPORTANCE EXPLAIN

| features1 | chi.square | features2 | random.forest |
|---|---|---|---|
| Elevation | 0.477 | Elevation | 1.108 |
| soilType | 0.468 | HS.3pm | 0.852 |
| wildernessArea | 0.443 | HD.Road | 0.817 |
| climaticZone | 0.398 | HD.Hydro | 0.774 |
| geologicZone | 0.175 | VD.Hydro | 0.751 |
| HD.Road | 0.169 | HS.noon | 0.737 |
| HS.3pm | 0.153 | HS.9am | 0.729 |
| Slope | 0.125 | HD.Fire | 0.714 |
| HD.Fire | 0.122 | Aspect | 0.643 |
| HS.noon | 0.098 | wildernessArea | 0.640 |
| HD.Hydro | 0.097 | Slope | 0.598 |
| HS.9am | 0.096 | climaticZone | 0.497 |
| VD.Hydro | 0.081 | geologicZone | 0.419 |
| Aspect | 0.080 | soilType | 0.047 |

Table 2: Feature Importance (chi,randomForest) of CoverTye Dataset Featuers

# 4 Summary

# 5 Research Questions of Interest