Capstone Data Exploration and Statistics

Tom Thorpe
June 12, 2018

Objective

View different plots of the cleaned Forest Cover data set and run some statistics to learn more about the data.

Data Overview

The forest cover data has a row for each sample representing a 30 meter by 30 meter square area of land. Each cell sample is described by elevation, slope and direction the cell faces, distance to water, roads, and fire and binary columns for wilderness area and soil type. One of 4 possible wilderness areas and one of 40 possible aggregated soil types are set in each row. The predicted variable is the coverage type indicating 1 of 7 possible trees found in the cell sample.

As part of data cleaning, the aggregated soil type was split into it's individual components of one possible climate zone, one possible geologic zone, one or more soil families and one or more "rocky-ness" categories. I hope to learn if breaking out the aggregated soil type into its components will improve the effectiveness of predicting the coverage type.

The data is described in detail here: https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/covtype.info. The data names have been abbreviated but can be related to the data descriptions easily.

List Selected Data Ranges

Data ranges are listed to help validate the data is within expected limits.

```
##        Data  min              mean median  max nonzero
## 1      Elev 1859  2959.36530054457   2996 3858  581012
## 2      Elev   18  29.0998344268277     29   38  581012
## 3    Aspect    0  155.656807432549    127  360  576098
## 4     Slope    0  14.1037035379648     13   66  580356
## 5     H2OHD    0  269.428216628916    218 1397  556409
## 6     H2OVD -173   46.418855376481     30  601  542347
## 7    RoadHD    0  2350.14661142971   1997 7117  580888
## 8  FirePtHD    0    1980.291226343   1710 7173  580961
## 9  Shade9AM    0  212.146048618617    218  254  580999
## 10 Shade12P    0  223.318716308785    226  254  581007
## 11 Shade3PM    0   142.52826275533    143  254  579674
## 12   RWwild    0 0.448865083681576      0    1  260796
```

```
## 13   NEwild    0 0.051434393781884      0    1   29884
## 14   CMwild    0 0.436073609495157      0    1  253364
## 15   CPwild    0 0.063626913041383      0    1   36968
```

The results show all the data values have reasonable values and there is no missing data. The elevation ranges from 1859 meters (6099 feet) to 3858 meters (12657 feet). These are valid ranges for elevation in the Colorado wilderness areas being sampled, but the rule of thumb for timberline (the maximum elevation for where trees are found) is 11500 feet. It might be interesting to see how accurate predictions are if samples above 11800 feet are removed.

ElevSlot, "Elevation Slot" is a new column that creates bins for elevation data for use with Chi-square testing. It is calculated by diving the elevation by 100 and truncating the value by saving as an integer. This results in 21 elevation bins.

The Aspect which is the compass heading that the terrain faces, ranges from 0 to 360 degrees and is a valid data range. The Slope is the steepness of the terrain with 0 degrees being flat and 90 degrees being vertical. The maximum Slope was found to be 66 degrees which seems logical since trees are not usually seen on near-vertical cliffs. (It's a different story in New Zealand!)

The horizontal distance to the nearest water features, range from 0 to 1397 meters which seems reasonable. The vertical distance to nearest water features, range from -173 to 601 meters which seems reasonable and can be negative since the nearest water may be below the forest cover data sample.

The horizontal distance to the nearest road ranges from 0 to 7117 meters which is reasonable. The horizontal distance to the nearest fire features range from 0 to 7173 meters which is reasonable. The amount of shade present in a cell sample at 9AM, 12PM and 3PM ranges from 0 (full sun) to 254 (fully shaded).

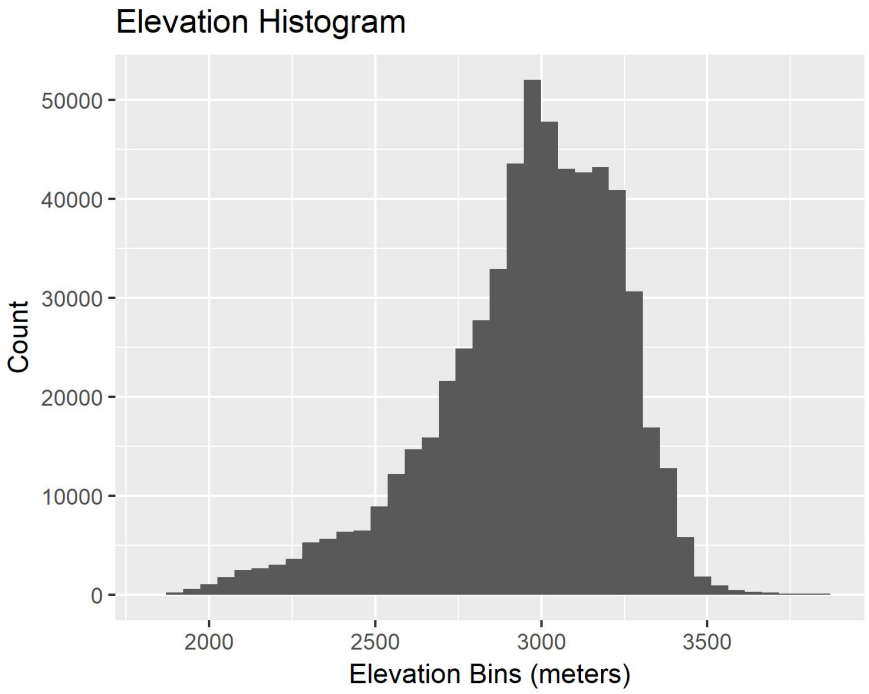
A table showing the number of occurrences for each tree type is shown below.

```
##            Var1   Freq Percent
## 1         Aspen   9493     1.6
## 2 Cotton&Willow   2747     0.4
## 3    DouglasFir  17367     2.9
## 4     Krummholz  20510     3.5
## 5     Lodgepole 283301    48.7
## 6     Ponderosa  35754     6.1
## 7    Spruce&Fir 211840    36.4
```

Lodge pole Pine represents 48.7 percent of the sample. So always guessing "Lodge pole" would provide success rate of 48.7 percent and can be used as a baseline for comparing our predictions. Spruce and Fir represent the next largest number of trees. The two together represent 85.1 percent.

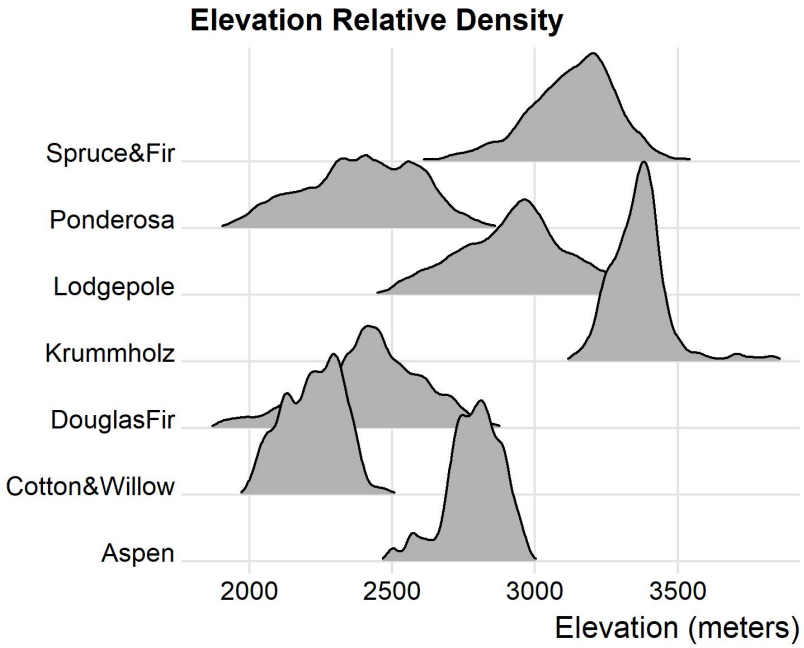
Elevation Histogram

A good histogram for elevation is generated Using 40 bins. There are two humps in the histogram and there may be a more complicated distribution. The elevation may be related to other variables. Next the elevation is grouped by coverage type and wilderness to see how elevation relates to coverage type.



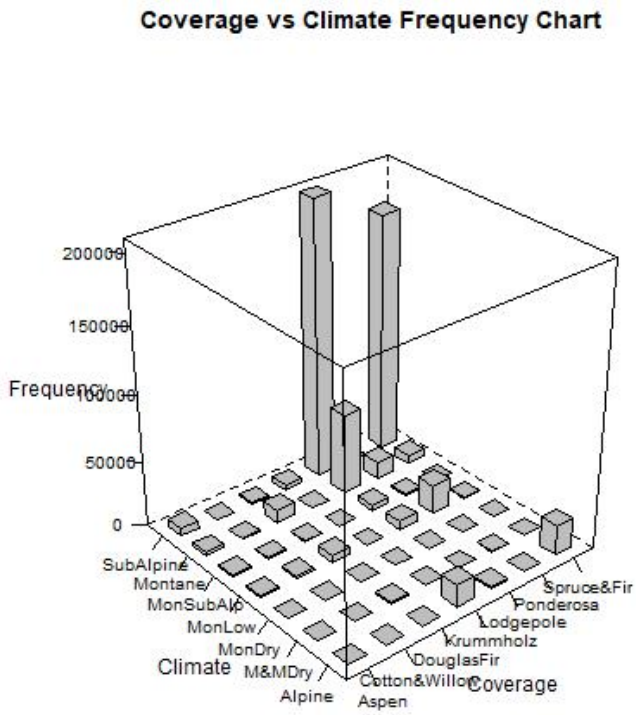
Elevation Density

The density ridges geom gives a good feel for the ranges of elevation for each coverage type. It looks like the elevation is a significant factor in helping determine coverage type.



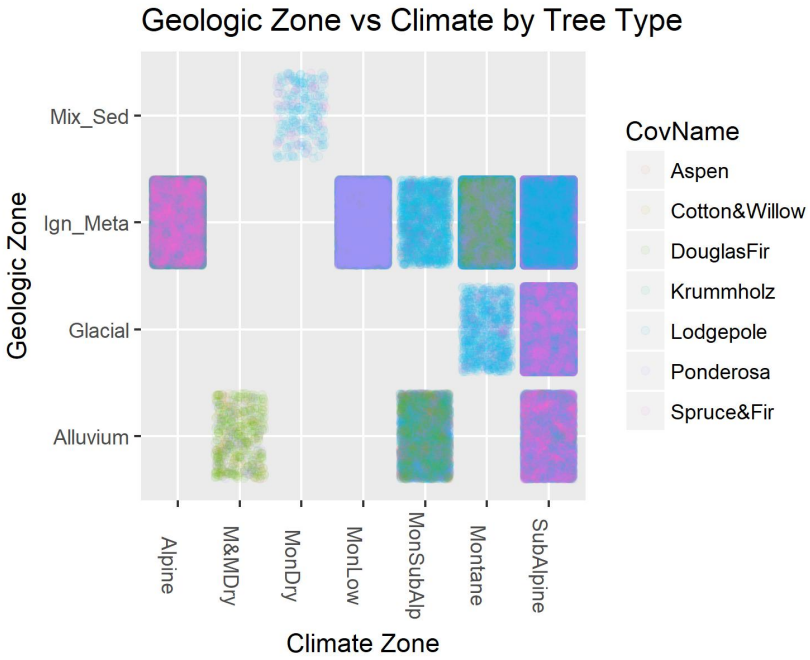
Coverage vs Climate Frequency

This gives a good view of the potential challenge to determine the various coverage types. The Lodge pole and Spruce & Fir trees make up the largest portion of the tree types. Determining the other tree types looks like they are in the “noise” of the data and might be more difficult to determine.

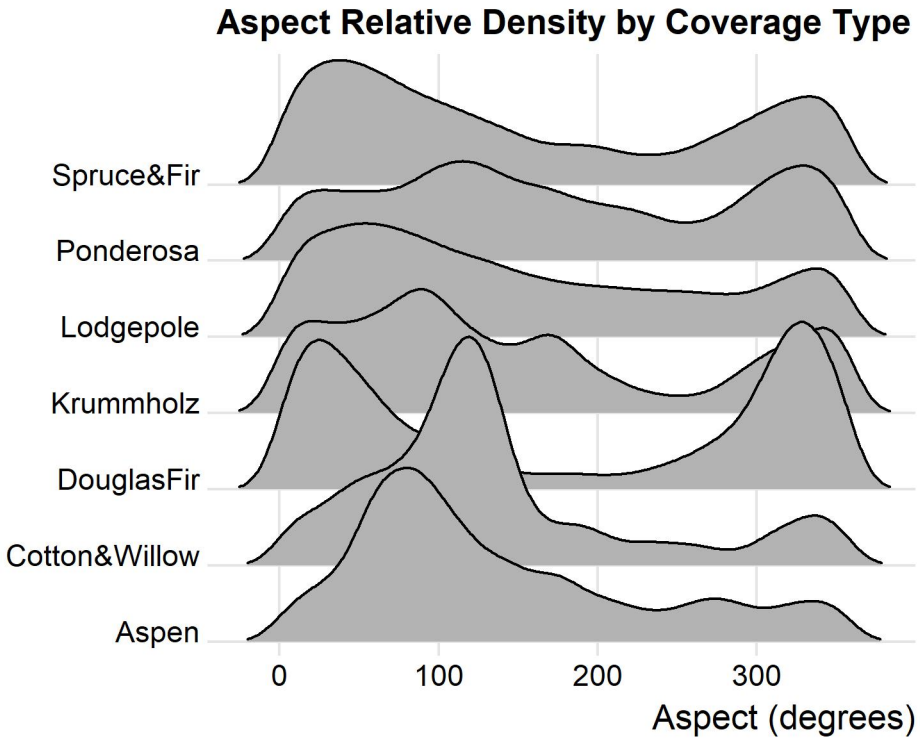


Geologic Zone vs Climate with Tree Type

Looking at the coverage type vs Climate and Geologic zones shows the two combinations may be helpful in determine coverage type but it is difficult to determine from this graph. The jitter geom was used to try to show the density, but the color coding is not distinct enough to get a feeling of the relative density of the tree coverage.



Aspect Relative Density vs Tree Type



Many other data were examined but did not suggest as clear a relationship to coverage type as the previous graphs. For example, the aspect of the slope (direction the slope of the cell faces) looks similar for each tree type. There are concentrations of tree types near aspects of 100 and 360 degrees. This occurs for all tree types and shows that the aspect will probably not be a significant factor in determining coverage type.

Statistics Analysis

After looking at some of the data relationships graphically, some statistical tests are applied to the data to test if variables follow a normal distribution or are independent.

Shaprio Test - Elevation

The Elevation histogram looks like it possibly has a normal distribution. It is not perfect but might be close enough statistically.

The Shapiro test is used to determine if data is normally distributed. The maximum number of data points for this Shaprio test is 5000. A sample of the forest cover data set was extracted for the Shapiro test. The Shapiro test result is shown below.

```
shapiro.test(altforestcover$Elev)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  altforestcover$Elev
## W = 0.95909, p-value < 2.2e-16
```

The null hypothesis for the Shapiro test is that the data follows a normal distribution. If the P-value is less than the 0.05 significance level, the null hypothesis is rejected and the data is not considered to be normally distributed otherwise the data is normally distributed.

The P-value for elevation data is 2e-16 which is nearly zero and much less than 0.05, therefore the null hypothesis is rejected and the data is not normally distributed. The previous histogram shows this visually: The graph has a long left tail and a short right tail.

Chi Square Test - Elevation & Coverage Name

It looks like elevation can be used to help identify coverage type. A chi-square test will be used to see if the coverage type and elevation variables are independent.

```
## Warning in chisq.test(table(forestcover$CovName,
forestcover$ElevSlot), :
## Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  table(forestcover$CovName, forestcover$ElevSlot)
## X-squared = 763760, df = 120, p-value < 2.2e-16
```

If the P-value is less than significance factor of 0.05, the null hypothesis is rejected and the variables are not independent. The P-value is 2e-16 which is nearly zero. This shows that the coverage type and elevation are dependent, if the chi-square test is valid.

Coverage Type vs Soil Type Independence check

The original paper used the Soil Type categories to predict coverage type. Let's try a chi-square test on them.

```
## Warning in chisq.test(table(forestcover$CovName,
forestcover$SoilType), :
## Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  table(forestcover$CovName, forestcover$SoilType)
## X-squared = 762250, df = 234, p-value < 2.2e-16
```

If the P-value is less than significance factor of 0.05, the null hypothesis is rejected and the variables are not independent. The P-value is 2e-16 which is nearly zero. This shows that the coverage type and soil type are dependent, if the chi-square test is valid.

Chi Square Test - Climate & Coverage Name Independence Test

The climate vs coverage type frequency graphs looked like there was a relationship between the two. A chi-square test on climate and coverage name is shown below.

```
## Warning in chisq.test(table(forestcover$CovName, forestcover
## $ClimateName), : Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  table(forestcover$CovName, forestcover$ClimateName)
## X-squared = 551740, df = 36, p-value < 2.2e-16
```

If the P-value is less than significance factor of 0.05, the null hypothesis is rejected and the variables are not independent. The P-value is 2e-16 which is nearly zero. This shows that the climate zone and coverage type are dependent, if the chi-square test is valid.

Conclusion

There are many interesting data distributions in the continuous data. The elevation data seems to be the most easy to intuitively see a relationship predicting the outcome of the coverage type.

The other categorical data seems difficult to relate to outcome intuitively.

Statistically, chi-square testing indicates that both elevation and soil type are related to the coverage type.

It will be interesting to see how the machine learning algorithms find relationships with the different data and if splitting out the soil type into individual components improves the accuracy of the predicted tree type.