Data Wrangling Summary

Tom Thorpe

April 2, 2018

Data Description

The forest coverage data to be cleaned contains physical properties of for a 30 meter by 30 meter square cell including, elevation, slope of the terrain, direction the terrain faces, vertical distance (VD) and horizontal distance (HD) to water, fire ignition points and road features, wilderness area within the study and soil type code. Each cell contains one of seven possible tree codes representing the tree in the cell.

Data Range Check

The first part of the data cleaning is to check data ranges for physical parameters. The results show all the data values have reasonable values and there is no missing data. The ranges were validated checking the min, mean and max of each physical property as shown below.

Property	Min	Mean	Max	Property	Min	Mean	Max
Elev	1859	2959	3858	Aspect	0	155	360
Slope	0	14	66	H2OHD	0	269	1397
H2OVD	-173	46	601	RoadHD	0	2350	7117
FirePtHD	0	1980	7173	Shade9AM	0	212	254
Shade12P	0	223	254	Shade3PM	0	142	254

The elevation ranges from 1859 meters (6099 feet) to 3858 meters (12657 feet) which is valid for elevation in the Colorado wilderness areas. The Aspect which is the compass heading that the terrain faces, ranges from 0 to 360 degrees and is a valid. 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.

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 raod 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).

Binary Data Check

The next phase of the data verification is to ensure multiple columns have not been selected for binary data. Each row was checked to ensure exatcly one of four Wilderness area columns was set to 1 and exactly one of 40 soil types were set to 1. Each row passed. The columns were summed and stored in two new columns to checked to ensure the value was exactly 1. All of the binary data in the forest coverage is clean and no data was missing.

Expansion of Soil Type

I want to investigate expanding the soil type into its aggregated parts to see if the data prediction can be improved. Columns representing the individual properties in the soil description will be added to the data and set based on the soil type code. For example, soil type 33 described as:

Leighcan - Catamount families - Rock outcrop complex, extremely stony in the subalpine climate zone and the igneous and metamorphic geologic zone.

will have binary column values set to 1 for the 'Leighcan' soil family, the 'Catamount' soil family, the 'Rock outcrop complex' rock density, the 'extremely stony' rock density, the 'subalpine' climate and 'igneous and metamorphic' geologic columns.

A transform data set has been created to facilitate this mapping. The columns and values in the transform file (except the first three descriptive columns) will be copied to the forest cover data frame and values set according to the soil type in the cell sample.

Lessons Learned

During the soil type expansion, I initially tried populating the data by setting each cell directly. It took 14.6 hours to update the ~581000 rows in the forest cover data set using this method. Reading and updating individual cells, I learned, is very inefficient. I was able to code simpler functions using column selects and updates.

Results

The full report of the data conversion is located here https://github.com/tomthorpe2000/datasciencefoundation/blob/master/CleanForestCover.pdf .

The full 5810012 row forest cover data set was too big to be loaded to Github. A Sample of the first 500 rows can be seen here https://github.com/tomthorpe2000/datasciencefoundation/blob/master/forestsmall.csv .

And the cleand data data for the 500 rows can be found here https://github.com/tomthorpe2000/datasciencefoundation/blob/master/forestcoversmall_clean.csv .