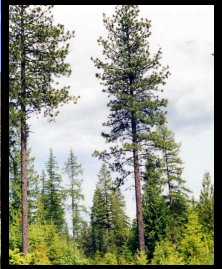
Colorado Forest Cover Types



Background on the data set

This study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices.

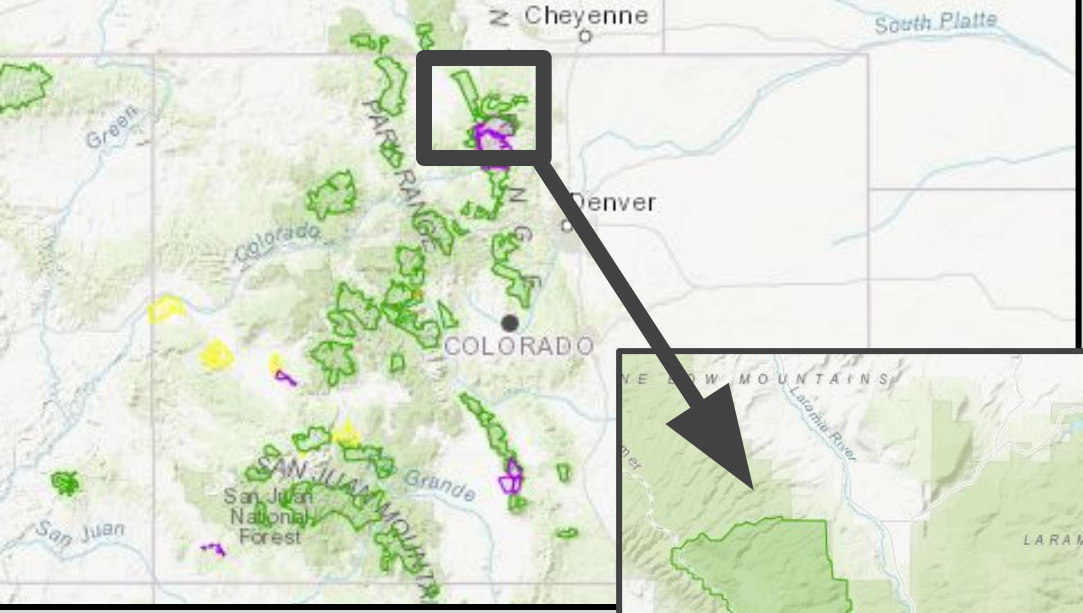
Kaggle: https://www.kaggle.com/c/forest-cover-type-prediction

UCI: http://archive.ics.uci.edu/ml/datasets/covertype

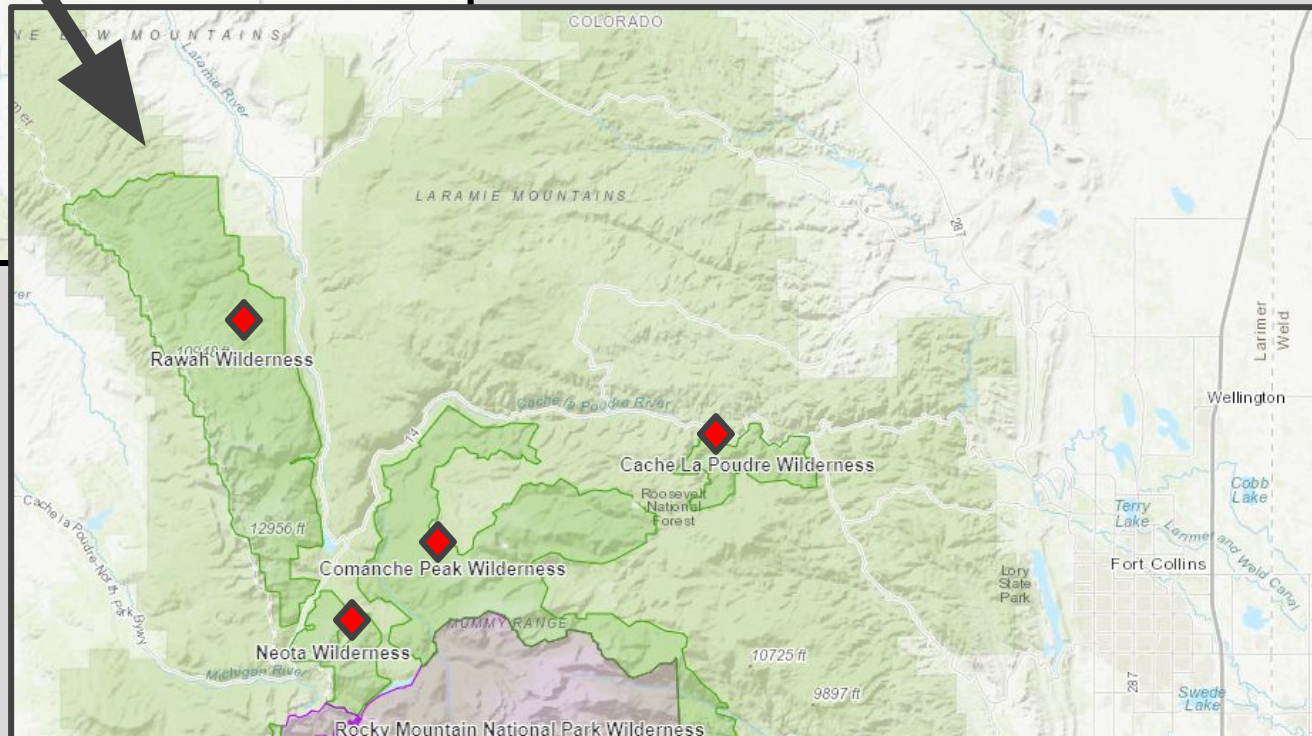
The dataset does not have any missing values.

The Cover Type dataset contains 55 features. Of the 55 features in the dataset 10 features are continious and 44 features are binary (wilderness area and soil types). The remaining feature is catergorical Cover Type in 7 forest cover types.

11 MB csv file - 581k X 55



State of Colorado including an inset of the study areas



Western Colorado

Continental Divide

Eastern Colorado

In general as you go up in elevation the more precipitation (rain & snow) falls there and the temperature gets colder.

Elevation: Feet above sea level

Alpine
11,500 ft. and higher

Subalpine:
10,000 to 11,500 ft.

Subalpine:
10,000 to 11,500 ft.

Montane Forests
8,000 to 10,000 ft.

Montane Forests
8,000 to 10,000 ft.

Foothills: Pinyon-Juniper Woodlands & Montane Shrublands
6,000 to 8,000 ft.

Foothills: Open Ponderosa Pine Woodlands
6,000 to 8,000 ft.

Life zones blend together as you go up

Desert Canyonlands & Sage Shrublands
5,000 to 7,000 ft.

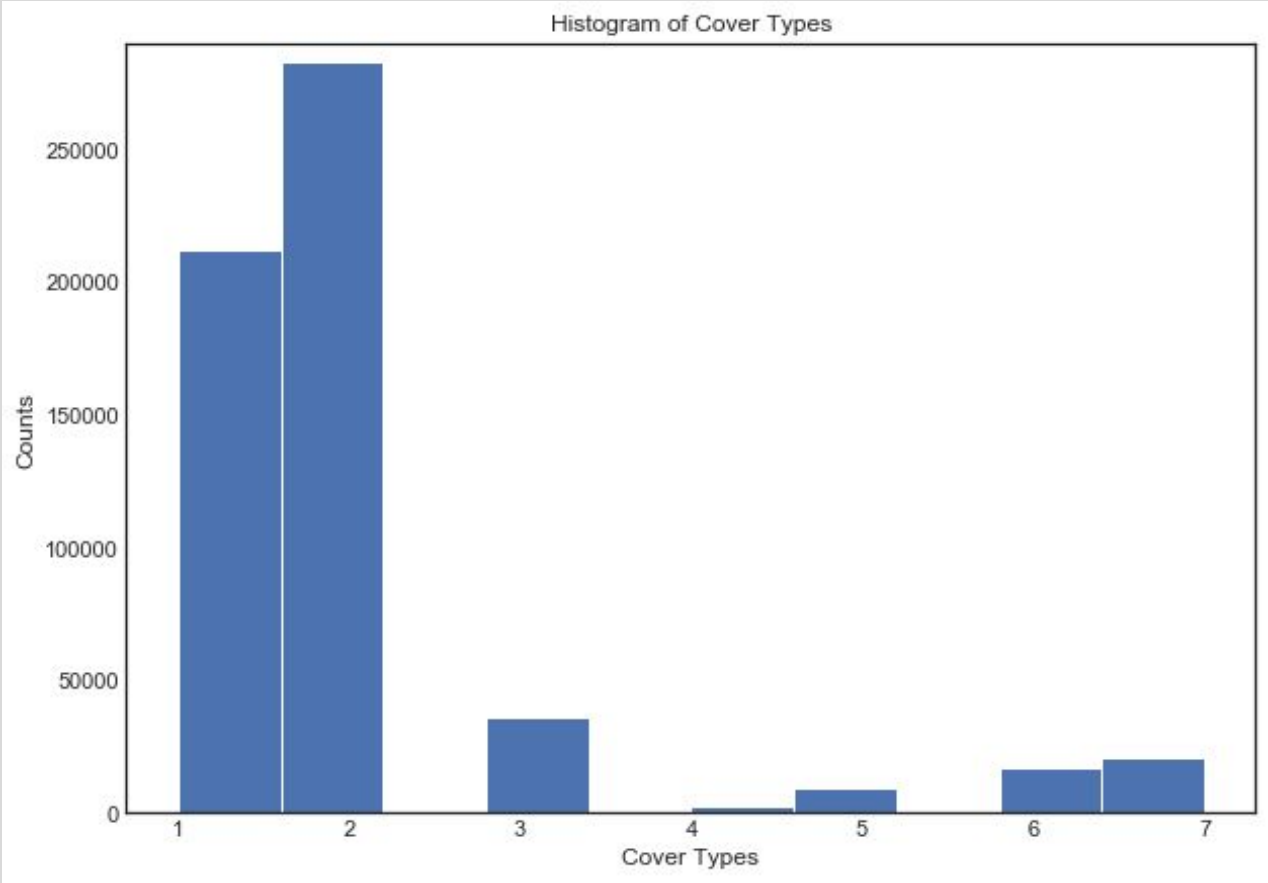
Great Plains: Grasslands
4,000 to 6,000 ft.

Modeling approach

Histogram of the Cover Types
1.Spruce/Fir
2.Lodgepole Pine
3.Ponderosa Pine
4.Cottonwood/Willow
5.Aspen
6.Douglas-fir
7.Krummholz

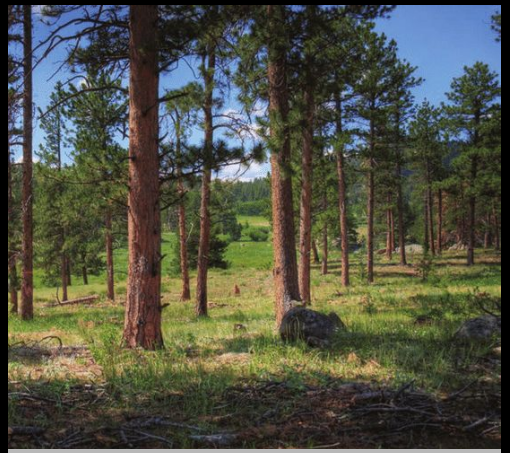
Approaches to modeling
1) Undersampling - Imbalanced data
2) Algorithms

Use the F1-score to find the predictive success multiclass dataset



1: SPRUCE/FIR

2: LODGEPOLE PINE



3: PONDEROSA PINE

4: COTTONWOOD/WILLOW

5: ASPEN

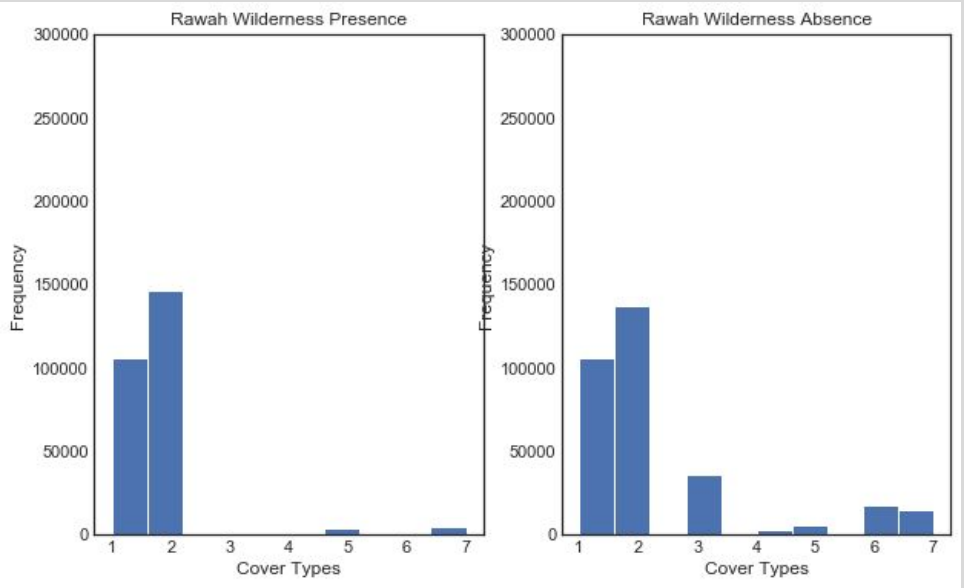
6: DOUGLAS-FIR

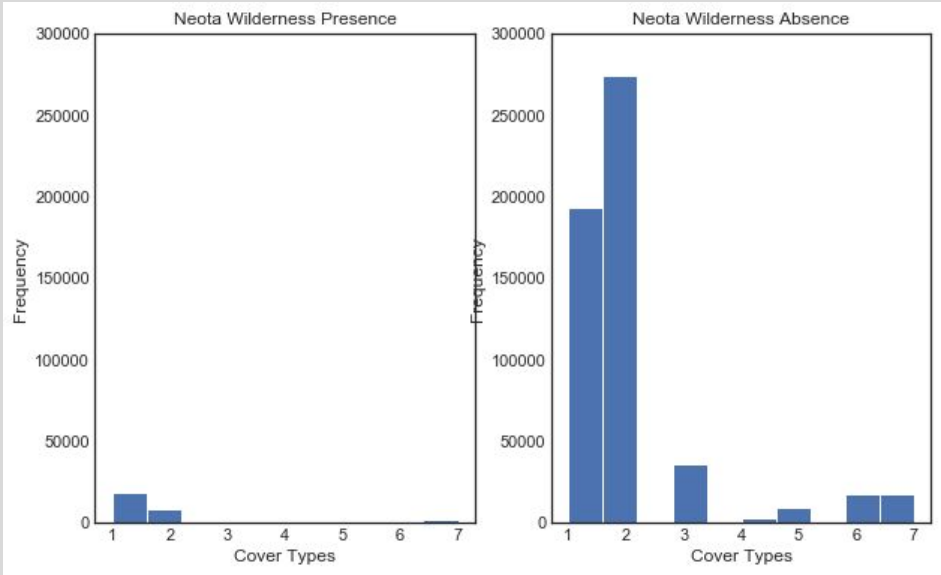
7: KRUMMHOLZ

Why do the Wilderness Areas vary us much?

Wilderness_Area1 - Rawah Wilderness Area Total acres = 73,213 acres - Elevation 8,000 - 13,000 ft

Wilderness_Area2- Neota Wilderness Area Total acres = 9647 acres -Elevation 10,000 to 11,896 ft

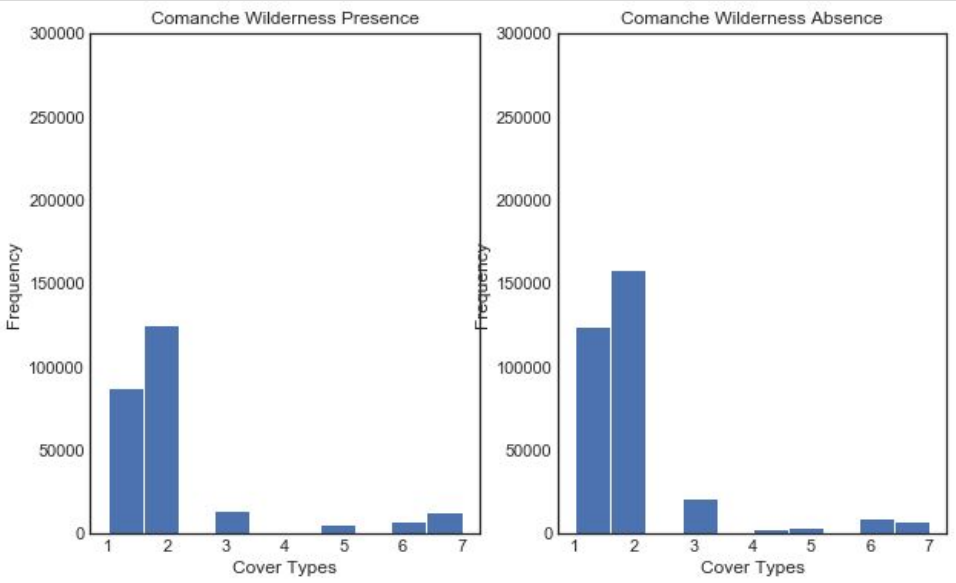


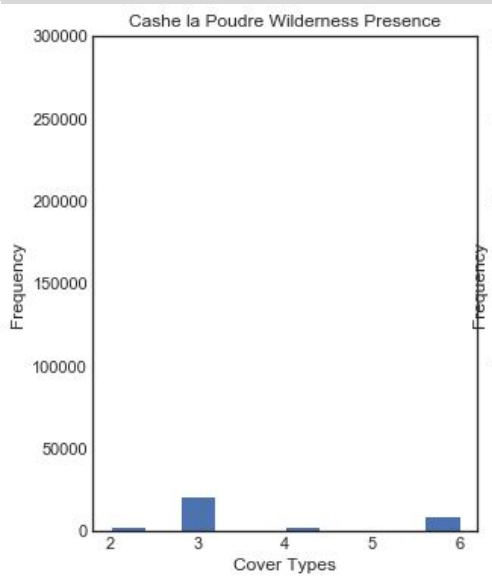


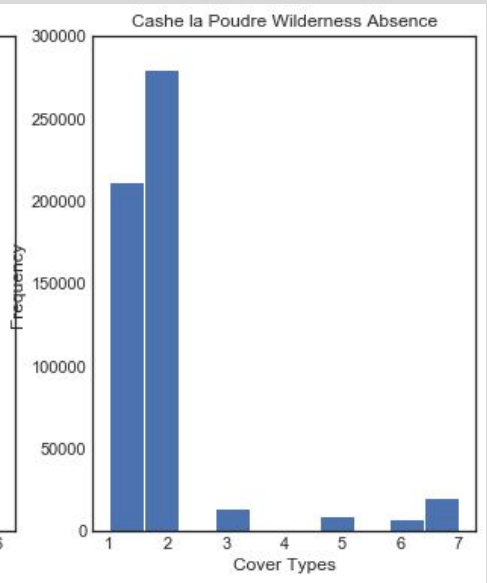
Continued, why do the Wilderness Areas vary us much?

Wilderness_Area3- Comanche Peak Wilderness Area (67,680 acres) Elevation 8,000 - to 12,000 ft

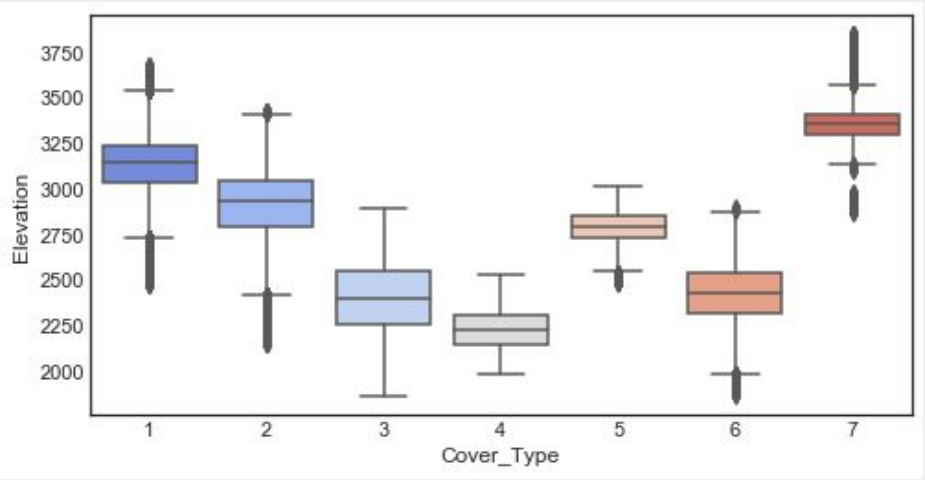
Wilderness_Area4 - Cache la Poudre Wilderness Area (9433 acres) - 6,200 to 8,600 ft

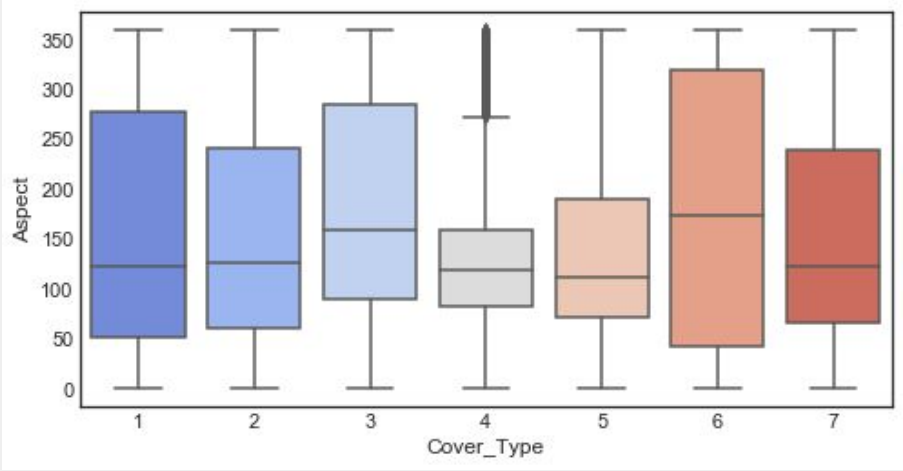






Exploratory Data Analysis

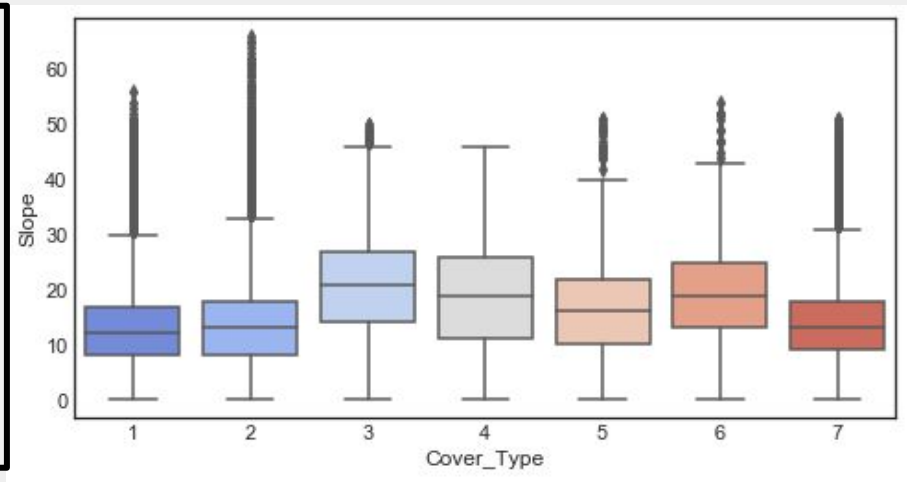




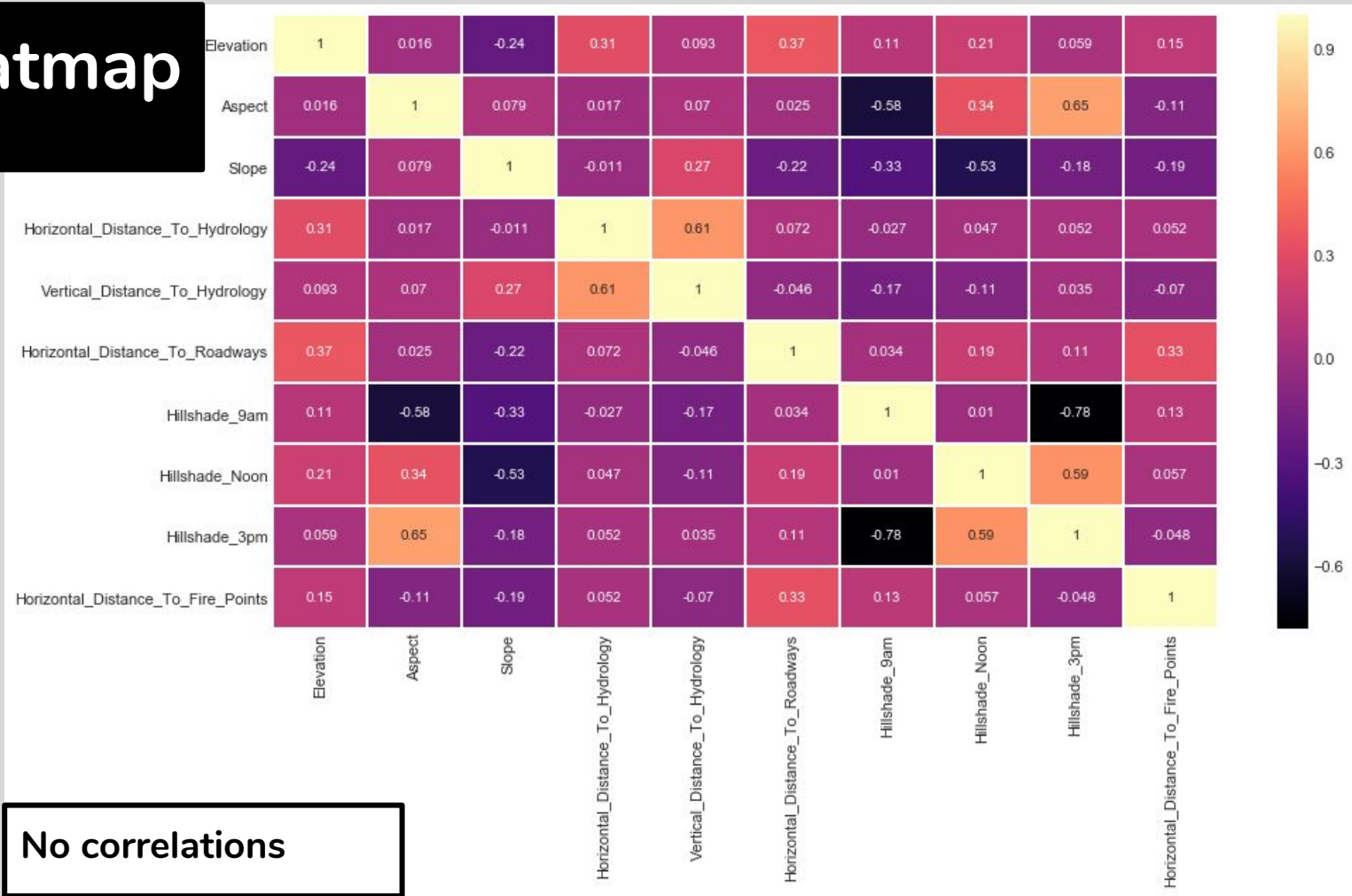
Elevation
- Altitudinal zonation of cover types
- Increase in elevation - different cover types

Aspect
- Pretty difficulty to understand anything
- 4 (cottonwoods/willows) along *Cashe la Poudre River*

Slope
- Cover type 3 and 4 have highest means low elevation species
- Lower elevation canyons perhaps?



EDA - Heatmap



No correlations

Logistic Regression Model

High - Green (strongest predictions)

Low - Orange (weak predictions)

Intermediate - Yellow (average predictions)

Evaluator F1-Score
Weighted avg. F1-score

Cover Types Names	Numbers	Precision	Recall	F1-score	Support
Spruce/Fir	1	0.71	0.68	**0.69**	69978
Lodgepole pine	2	0.73	0.8	**0.76**	93523
Ponderosa pine	3	0.61	0.86	**0.71**	11696
Cottonwood/willow	4	0.58	0.23	**0.33**	875
Aspen	5	0	0	**0**	3225
Douglas-fir	6	0.42	0.06	**0.1**	5762
Krummholz	7	0.71	0.47	**0.57**	6675
	weighted avg	0.69	0.71	**0.69191734**	

Results - Resampling and Undersampling

High - Green (only Lodgepole pine)

Low - Orange (weak predictions 6 of 7 cover types)

Evaluator F1-Score
Weighted avg. F1-score

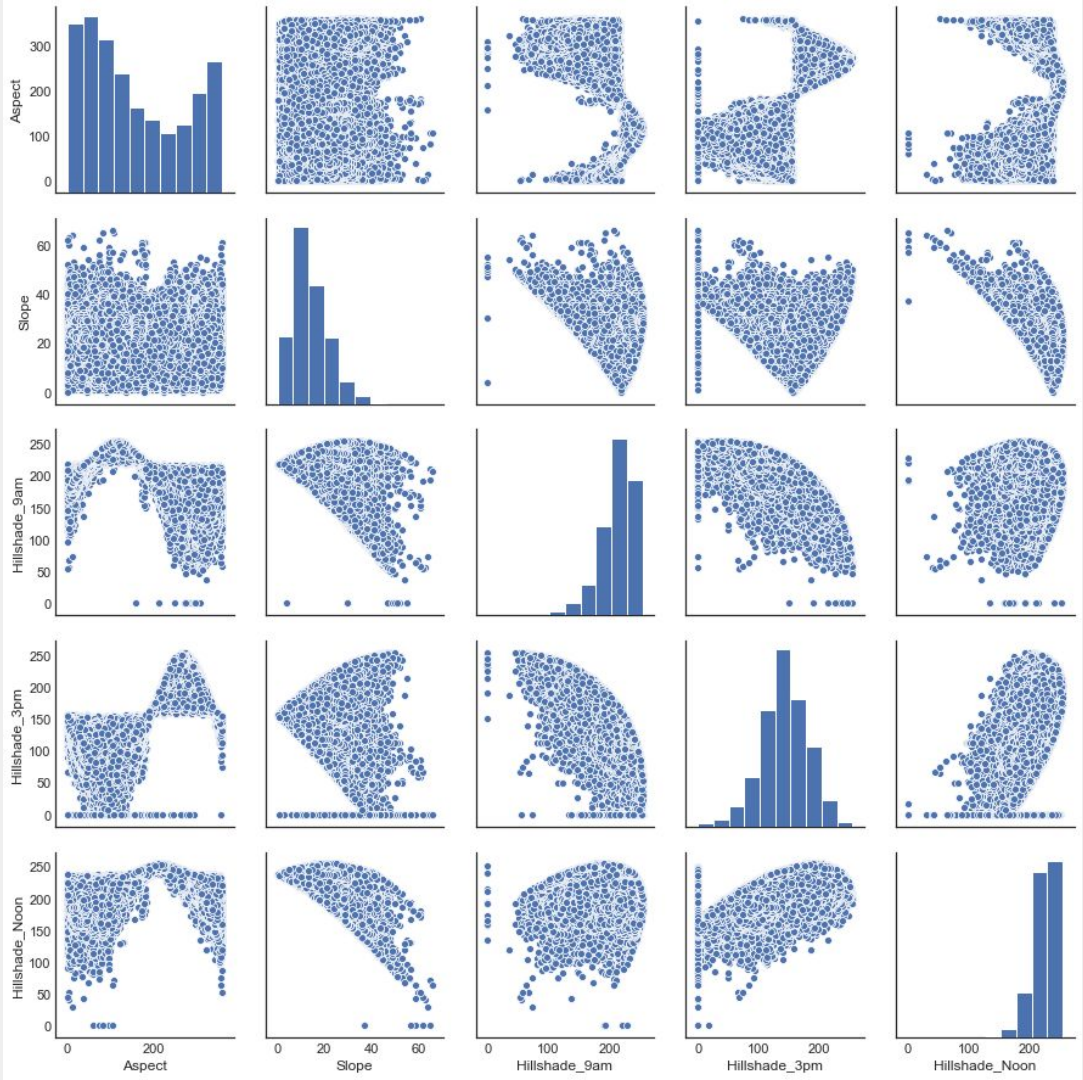
Side note: computer memory crashed multiple times

Cover Types Names	Numbers	Precision	Recall	F1-score	Support
Spruce/Fir	1	0.25	0	**0.01**	53006
Lodgepole pine	2	0.51	0.99	**0.68**	70776
Ponderosa pine	3	0.81	0.06	**0.11**	8909
Cottonwood/willow	4	0.22	0.84	**0.35**	651
Aspen	5	0	0	**0**	2473
Douglas-fir	6	0.29	0.23	**0.26**	4373
Krummholz	7	0.43	0.08	**0.14**	5065
	weighted avg	**0.41**	**0.5**	**0.35**	

Non-linear relationships

The **sns.pairplot** is an indication to use **Naive Bayes.**

Continuous variables show non-linear relationship



Naive Bayes

High - Green (strongest predictions)

Low - Orange (weak predictions)

Intermediate - Yellow (average predictions)

- More average F1-scores
-

Evaluator F1-Score
Weighted avg. F1-score

Cover Types Names	Numbers	Precision	Recall	F1-score	Support
Spruce/Fir	1	0.65	0.48	**0.55**	211840
Lodgepole pine	2	0.65	0.76	**0.7**	283301
Ponderosa pine	3	0.6	0.87	**0.71**	35754
Cottonwood/willow	4	0.55	0.43	**0.48**	2747
Aspen	5	0.22	0.06	**0.1**	9493
Douglas-fir	6	0.24	0.23	**0.23**	17367
Krummholz	7	0.63	0.61	**0.62**	20510
	weighted avg	0.63	0.63	**0.62**	

Xgboost (Boosting)

High - Green (strongest predictions)

Low - Orange (weak predictions)

Intermediate - Yellow (average predictions)

- More higher F1-scores

Evaluator F1-Score
Weighted avg. F1-score

Cover Types Names	Numbers	Precision	Recall	F1-score	Support
Spruce/Fir	1	0.74	0.73	**0.73**	70052
Lodgepole pine	2	0.76	0.83	**0.79**	93189
Ponderosa pine	3	0.68	0.85	**0.76**	11873
Cottonwood/willow	4	0.83	0.58	**0.68**	972
Aspen	5	0.79	0.1	**0.18**	3124
Douglas-fir	6	0.51	0.11	**0.18**	5687
Krummholz	7	0.84	0.52	**0.64**	6837
	weighted avg	0.74	0.75	**0.73**	

Random Forest

All F1- scores are above 0.84

Evaluator F1-Score
Weighted avg. F1-score

Highest performing model

Cover Types Names	Numbers	Precision	Recall	F1-score	Support
Spruce/Fir	1	0.96	0.94	**0.95**	70052
Lodgepole pine	2	0.95	0.97	**0.96**	93189
Ponderosa pine	3	0.94	0.96	**0.95**	11873
Cottonwood/ willow	4	0.93	0.83	**0.88**	972
Aspen	5	0.94	0.75	**0.84**	3124
Douglas-fir	6	0.93	0.89	**0.91**	5687
Krummholz	7	0.97	0.95	**0.96**	6837
	weighted avg	0.95	0.95	**0.95**	

Conclusion

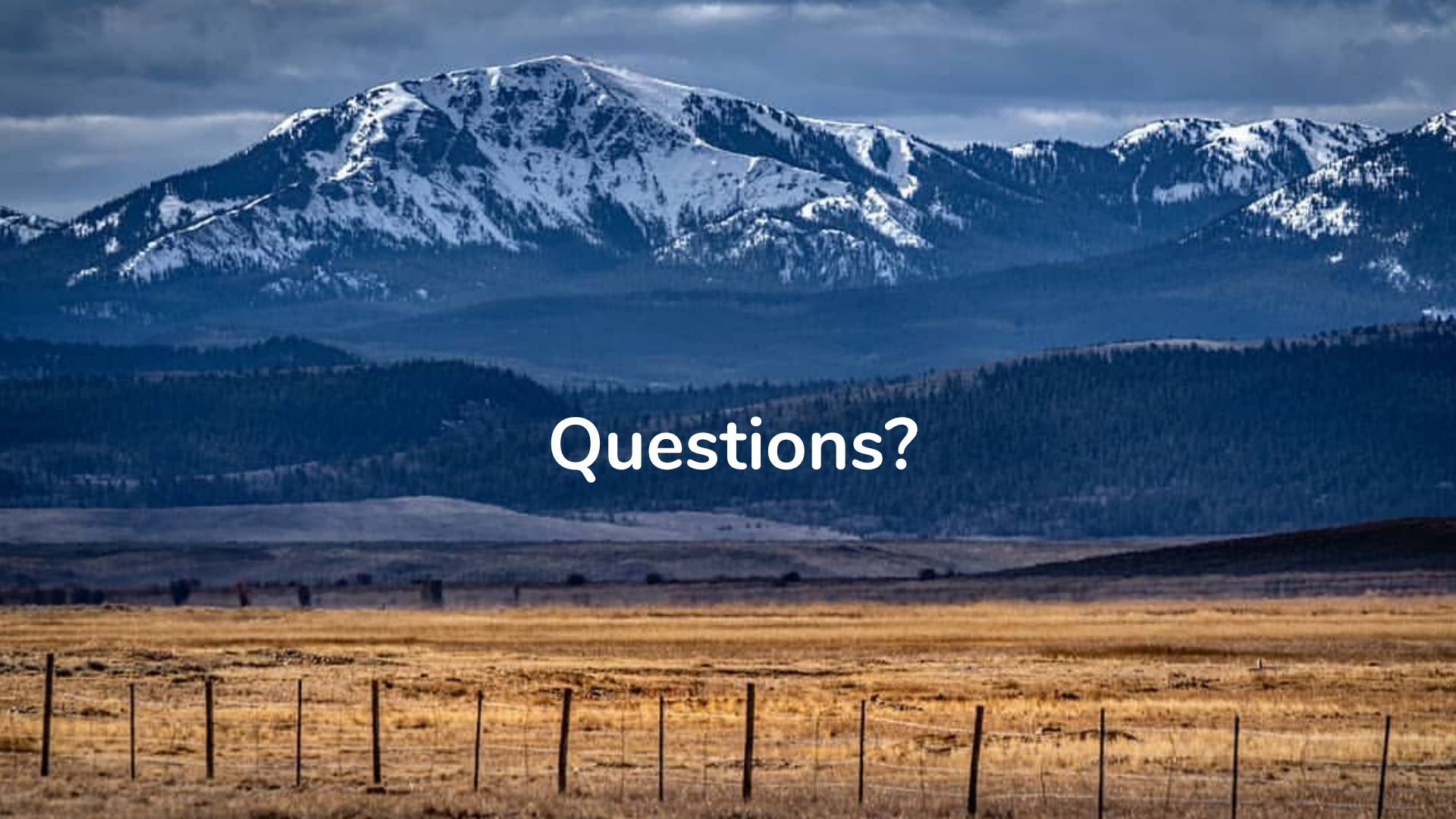
Rank predictions
2 - Lodgepole pine
3 - Ponderosa pine
1 - Spruce/Fir
7 - Krummholz
4 - Cottonwood/willow
6 - Douglas-fir
5 - Aspen

What does it mean?
-Strong sample representations of 2, 3, 1, and 5

-Patchy samples representing 4, 5, and 6

-Bagging technique outperforms all other models

Cover Types Names	Numbers	Logistic Regression F1-score	Undersampling Logistic F1-score	Naive Bayes F1-score	XgBoost F1-score	Random Forest F1-score
Spruce/Fir	1	**0.69**	**0.01**	0.55	0.73	0.95
Lodgepole pine	2	**0.76**	**0.68**	**0.7**	**0.79**	**0.96**
Ponderosa pine	3	**0.71**	**0.11**	0.71	0.76	0.95
Cottonwood/ willow	4	**0.33**	**0.35**	0.48	0.68	0.88
Aspen	5	**0**	**0**	0.1	0.18	0.84
Douglas-fir	6	**0.1**	**0.26**	0.23	0.18	0.91
Krummholz	7	**0.57**	**0.14**	0.62	0.64	0.96
	weighted avg	**0.69191734**	**0.35**	**0.62**	**0.73**	**0.95**


Questions?