Colorado Forest Cover Types

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.

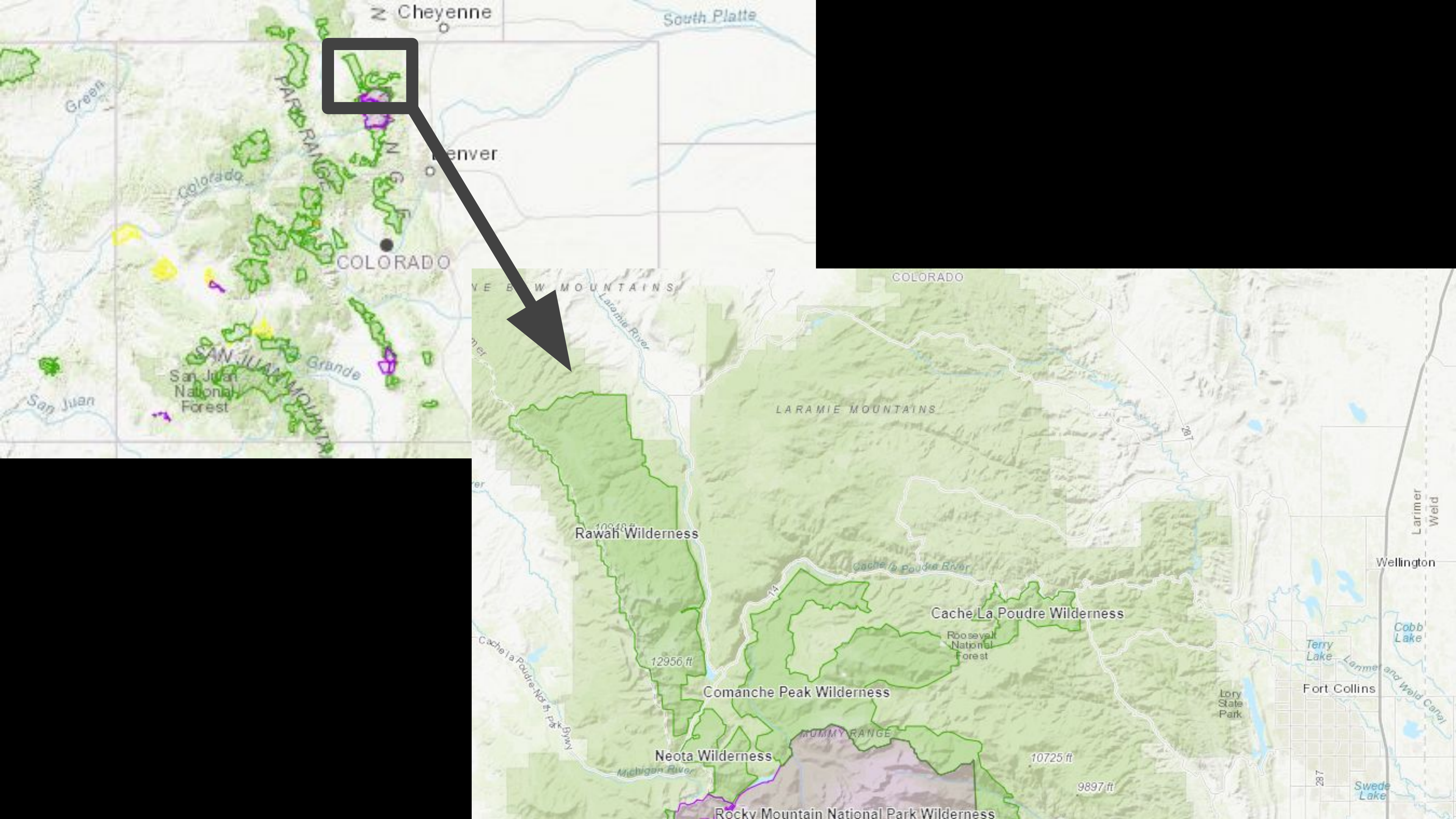
Kaggle: https://www.kaggle.com/c/forest-cover-type-prediction

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



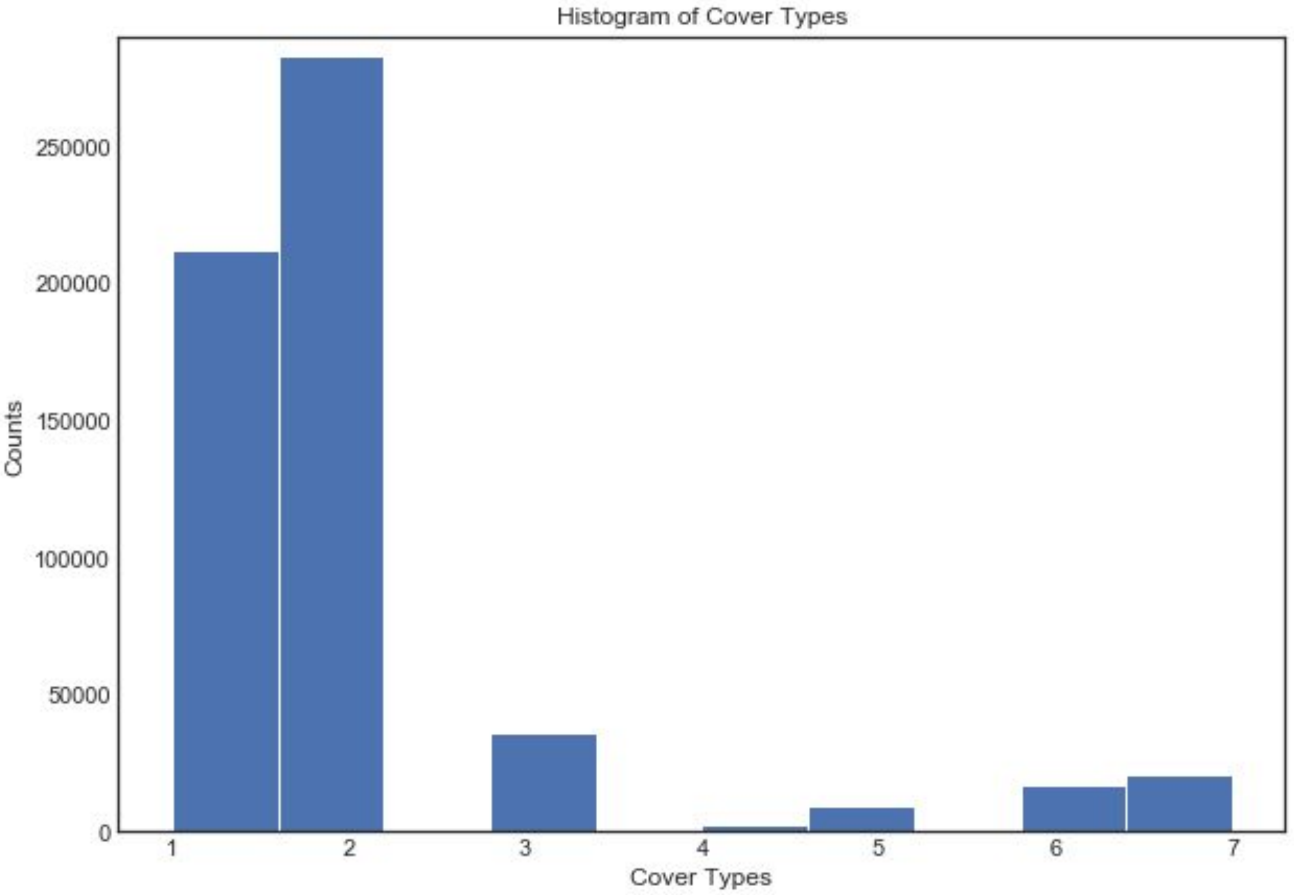
Cheyenne
South Platte
Green
Denver
Colorado
COLORADO
Grande
San Juan National Forest
San Juan
COLORADO
LARAMIE MOUNTAINS
Laramie River
Rawah Wilderness
12956 ft
Cache la Poudre River
Cache La Poudre Wilderness
Roosevelt National Forest
Comanche Peak Wilderness
MUMMY RANGE
Neota Wilderness
Michigan River
Rocky Mountain National Park Wilderness
Cache la Poudre-North Park Byway
10725 ft
9897 ft
287
14
Larimer
Weld
Wellington
Cobb Lake
Terry Lake
Larimer and Weld Canal
Fort Collins
Lory State Park
Swede Lake


Western Colorado
Continental Divide
Eastern Colorado
Alpine
11,500 ft. and higher
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
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.

Histogram of the Cover Types

Uneven balance in the data within the 7 cover type categories.

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




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?

Presence 1 = means that it is within the wilderness boundaries Absence 0 = means that it is not within the wilderness boundaries

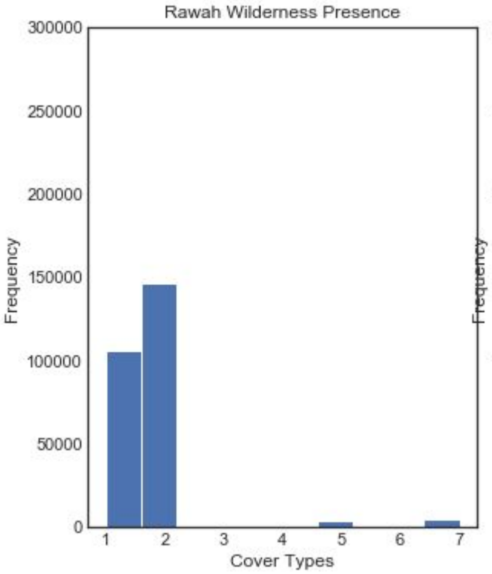
This is evident because Cashe la Poudre and Neota Wilderness Area's majority of surveyed terrain seems to fall outside of the wilderness boundary.

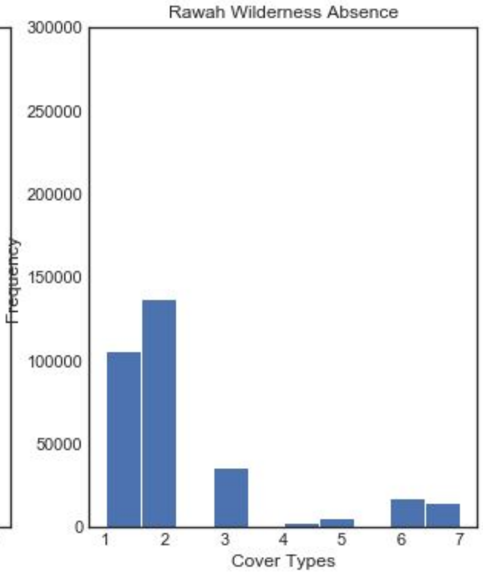
Wilderness_Area1 - Rawah Wilderness Area Total acres = 73,213 acres

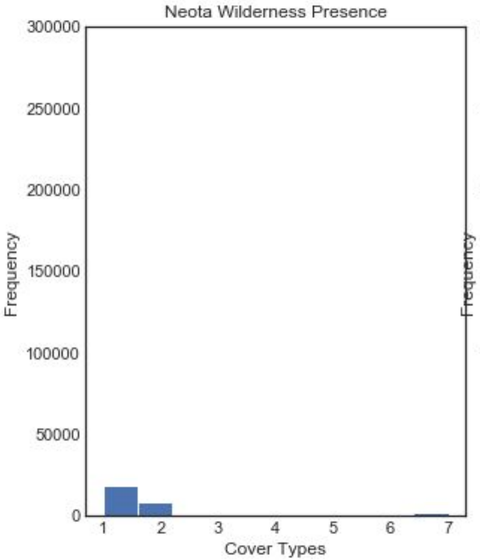
45% of the surveyed area of the dataset. Elevation 8,000 - 13,000

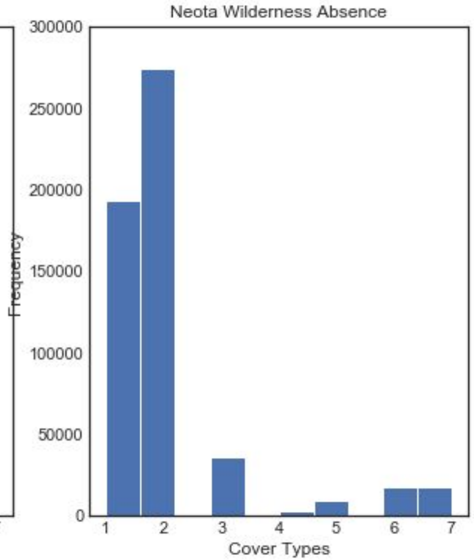
Wilderness_Area2- Neota Wilderness Area Total acres = 9647 acres

Elevation ranges from 10,000 ft (3,000 m) to 11,896 ft (3,626 m)





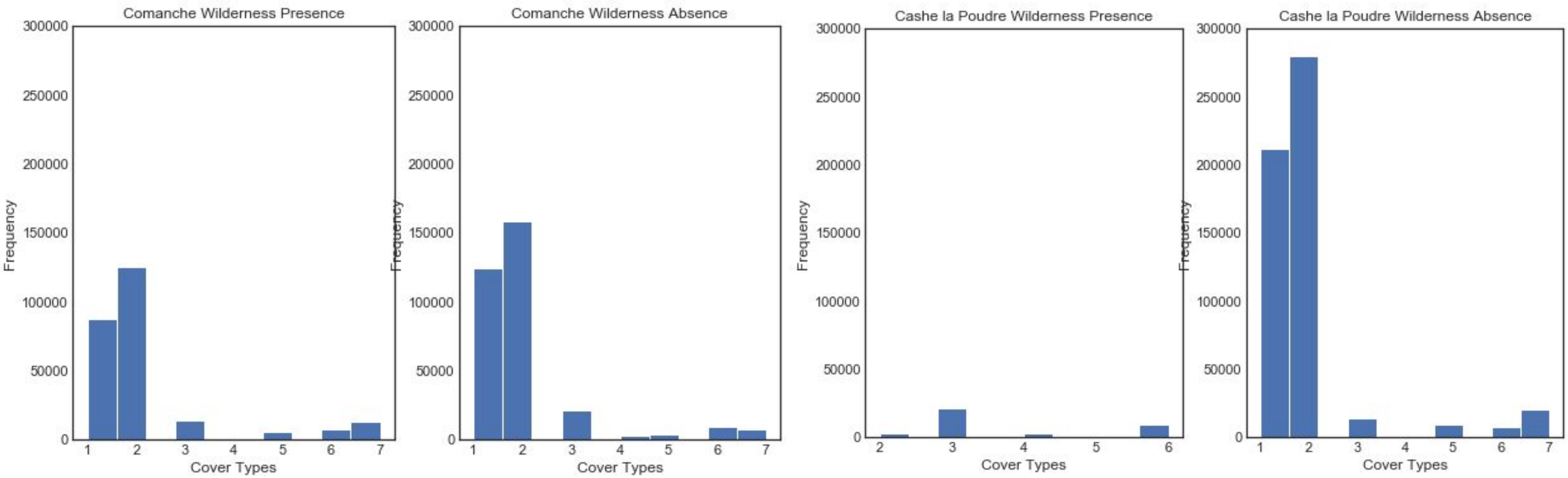




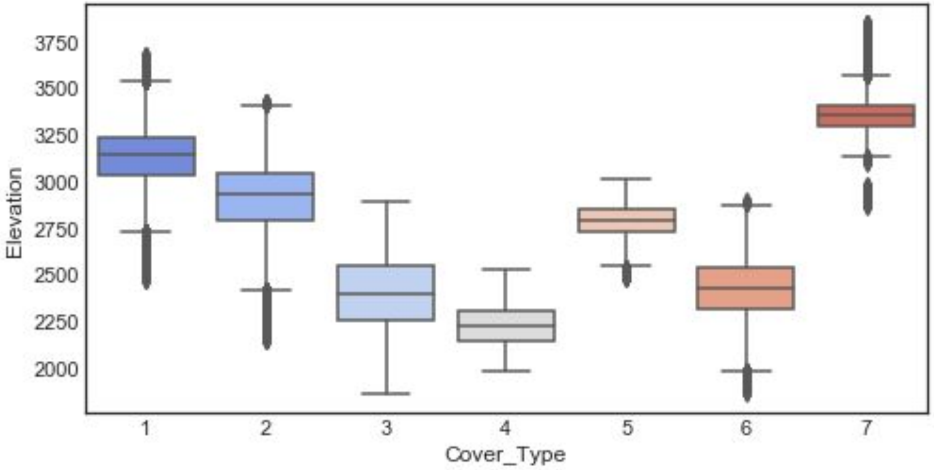
Presence 1 = means that it is within the wilderness boundaries Absence 0 = means that it is not within the wilderness boundaries
This is evident because Cashe la Poudre and Neota Wilderness Area's majority of surveyed terrain seems to fall outside of the wilderness boundary.

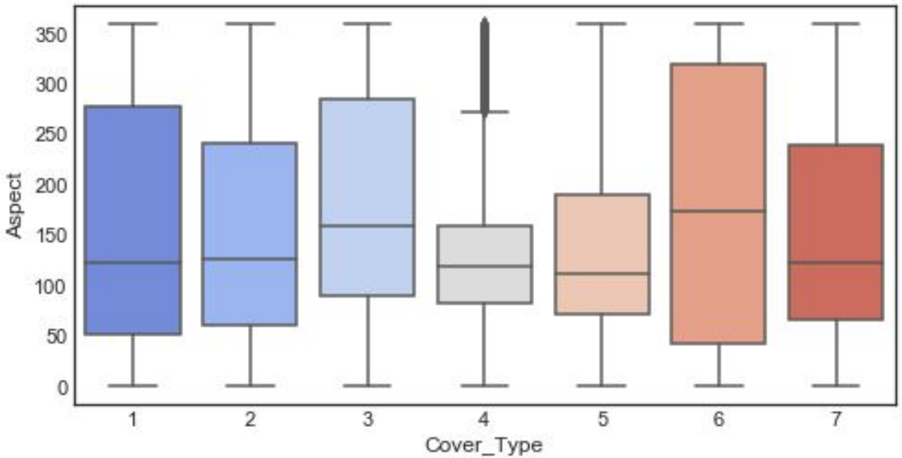
Wilderness_Area3- Comanche Peak Wilderness Area (67,680 acres) **42.8%** of the dataset - Elevation Ranges from 8,000 - to over 12,000 feet

Wilderness_Area4 - Cache la Poudre Wilderness Area (9433 acres) 0.059712736 - 6,200 feet (1,900 m) to 8,600 feet (2,600 m) It follows the Cashe la Pourde River likely Douglas Fir and Cottonwoods



Exploratory Data Analysis

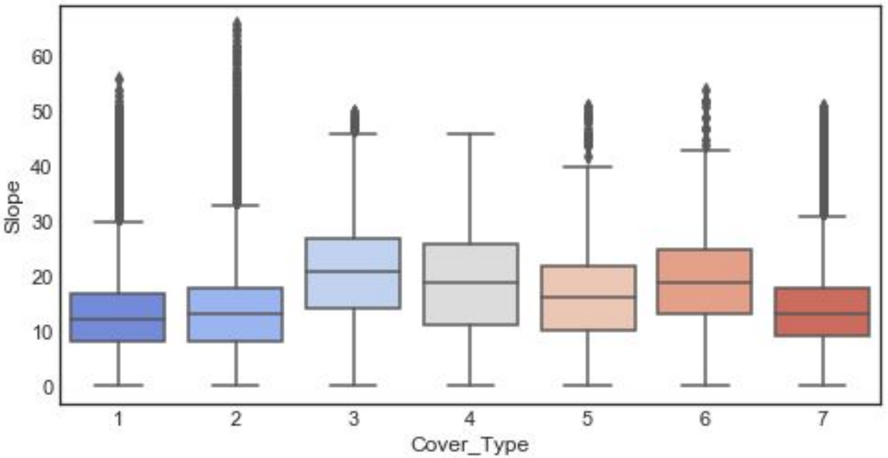


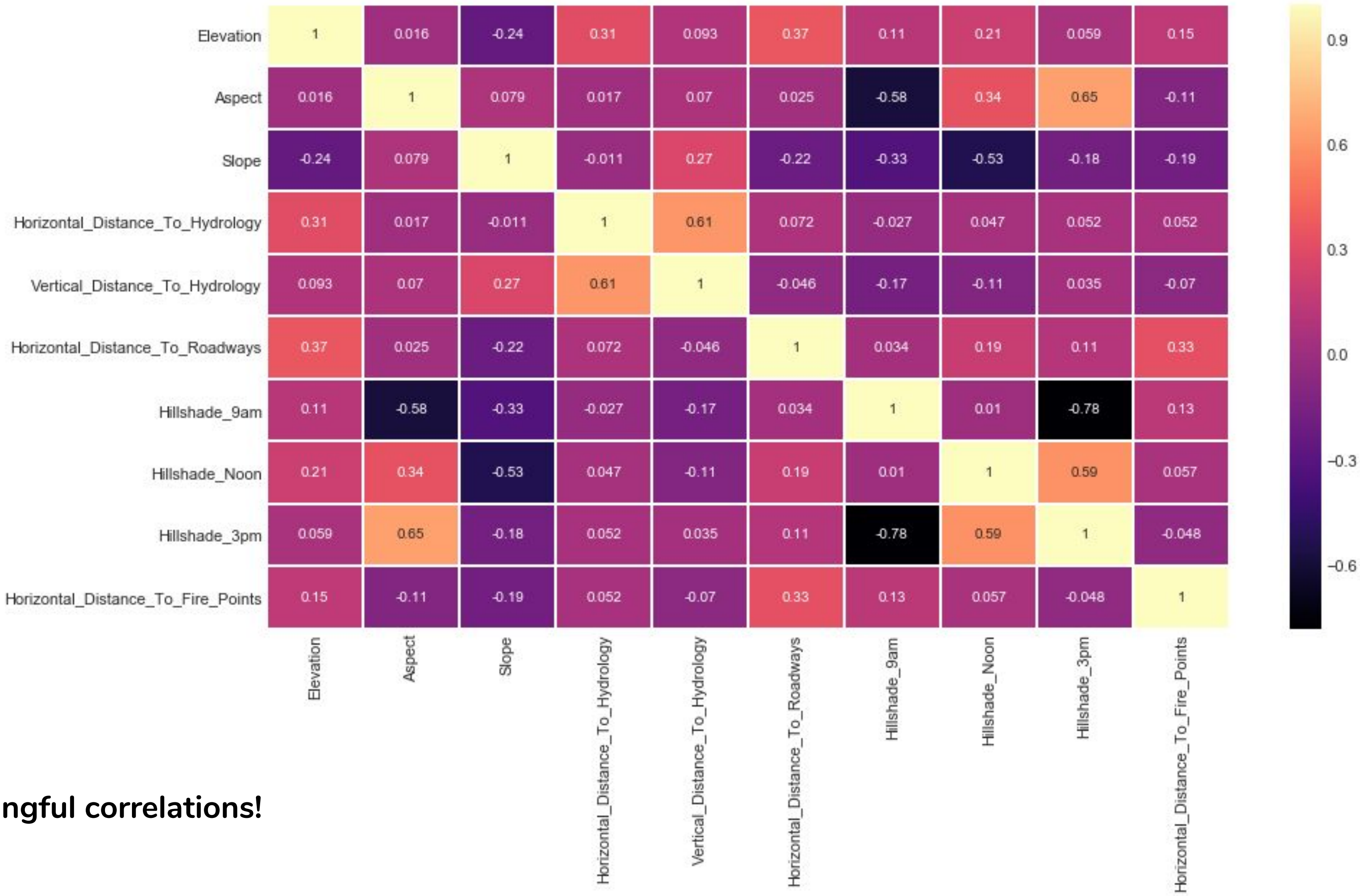


Elevation
- Altitudinal zonation of cover types
- Increase in elevation - different cover types

Aspect - direct the slope faces
- 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?





No meaningful correlations!

Logistic Regression Model

In summary, predicted cover types that were most distinct from others. Cover types (1, 2, 3, and 7) all have a distinct statistical characteristics that define there cover types.

Cover types (4, 5, and 6) are less distinct statistical characteristics that define these cover types. The cover types 4-6 likely have more mixing of other cover types, more restrictive boundaries for area surveyed, and are environmentally restricted to wetted or damper environments.

	precision	recall	f1-score	support
1	0.71	0.68	0.69	69978
2	0.73	0.80	0.76	93523
3	0.61	0.86	0.71	11696
4	0.58	0.23	0.33	875
5	0.00	0.00	0.00	3225
6	0.42	0.06	0.10	5762
7	0.71	0.47	0.57	6675
micro avg	0.71	0.71	0.71191734	
macro avg	0.54	0.44	0.45191734	
weighted avg	0.69	0.71	0.69191734	

Resampling and Undersampling

```python
from imblearn.pipeline import make_pipeline
from imblearn.metrics import classification_report_imbalanced

```

6]:

```python
# fraction of rows
# here you get 75% of the rows
df.sample(frac=0.75, random_state=99)
train = df.sample(frac=0.75, random_state=99)
```

7]:

```python
# you can't simply split 0.75 and 0.25 without overlapping
# this code tries to find that train = 75% and test = 25%
test = df.loc[~df.index.isin(train.index), :]
```

8]:

```python
X = df.ix[:, 'Elevation':'Soil_Type40']
y = df['Cover_Type']
```

9]:

```python


X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)

print('Training target statistics: {}'.format(Counter(y_train)))
print('Testing target statistics: {}'.format(Counter(y_test)))

# Create a pipeline
pipeline = make_pipeline(NearMiss('not majority', version=2),
                         LinearSVC(random_state=42))
pipeline.fit(X_train, y_train)

# Classify and report the results
print(classification_report_imbalanced(y_test, pipeline.predict(X_test)))
```

Results - Resampling and Undersampling

```
Training target statistics: Counter({2: 212525, 1: 158834, 3: 26845, 7: 15445, 6: 12994, 5:
7020, 4: 2096})
Testing target statistics: Counter({2: 70776, 1: 53006, 3: 8909, 7: 5065, 6: 4373, 5: 2473,
4: 651})
                    pre       rec       spe        f1       geo       iba       sup


          1      0.25      0.00      1.00      0.01      0.05      0.00     53006
          2      0.51      0.99      0.10      0.68      0.32      0.11     70776
          3      0.81      0.06      1.00      0.11      0.24      0.05      8909
          4      0.22      0.84      0.99      0.35      0.91      0.82       651
          5      0.00      0.00      1.00      0.00      0.00      0.00      2473
          6      0.29      0.23      0.98      0.26      0.48      0.21      4373
          7      0.43      0.08      1.00      0.14      0.29      0.07      5065

avg / total      0.41      0.50      0.56      0.35      0.22      0.07 145253
```

Memory crash on computer multiple times.

The undersampling poorly performed. It did an okay job at predicting cover type 2 (Lodgepole pine). However, the other cover types F1-scores are much lower after undersampling. It appears that the undersampling sample disportionality across the data set. In order to have an effect undersampling all F1-scores should be very similar or the data samples in each cover type class should be similar. The model above shows that undersampling does did effectively work with Logistic regression model.

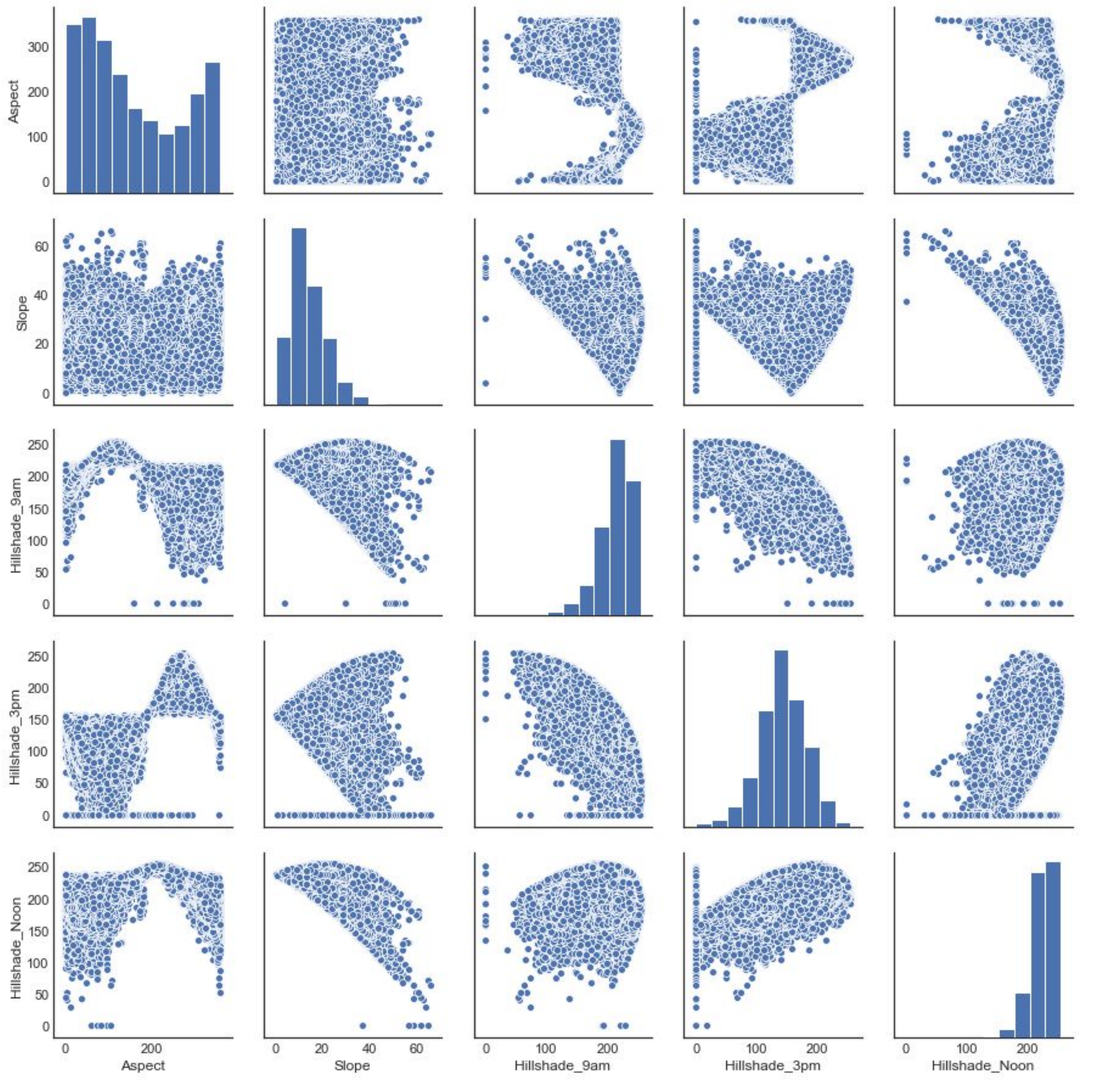
Naive Bayes

The **sns.pairplot** is an indication Classification models such as Naive Bayes.

	precision	recall	f1-score	support
1	0.65	0.48	0.55	211840
2	0.65	0.76	0.70	283301
3	0.60	0.87	0.71	35754
4	0.55	0.43	0.48	2747
5	0.22	0.06	0.10	9493
6	0.24	0.23	0.23	17367
7	0.63	0.61	0.62	20510
micro avg	0.63	0.63	0.63	581012
macro avg	0.51	0.49	0.49	581012
weighted avg	0.63	0.63	0.62	581012

The weighted value of 0.62 is close to the Logistic model of 0.69. Reducing the features in the model may improve the results. It works on conditional probability.

Naives Bayes model does a better job using conditional probabilities to with classes (cover types) with less data.



Xgboost (Boosting)

1 (Spruce/fir) - has a F1-score of 0.73 which is higher then both Naive Bayes (0.55) and Logistic at (0.69).

2 (Lodgepole pine) - has an F1-score of 0.79 which only slightly higher then logistic (0.76) and Naive Bayes at (0.70).

3 (Ponderosa pine) - has an F1-score of 0.76 which both Logistic and Naive Bayes (0.71). The consistency of predicting the ponderosa pine forest at 0.71 is interesting that all three models are predicting similar or same results. That may mean the the data for Ponderosa pine has very few outliers and is consistent throughout the dataset.

Inconsistent predictions:

Cover types 4 (Aspen), 5 (Cottonwood/willow), and 6 (Douglas-fir) all show high variability cross all models for predicting cover type. The inconsistencies goes back to patchy cover types that cover less area overall that are not consistently defined boundaries. These factors lead to the low test accuracy of the F1-scores.

Accuracy: 74.55%

	precision	recall	f1-score	support
1	0.74	0.73	0.73	70052
2	0.76	0.83	0.79	93189
3	0.68	0.85	0.76	11873
4	0.83	0.58	0.68	972
5	0.79	0.10	0.18	3124
6	0.51	0.11	0.18	5687
7	0.84	0.52	0.64	6837
micro avg	0.75	0.75	0.75	191734
macro avg	0.73	0.53	0.57	191734
weighted avg	0.74	0.75	0.73	191734

Random Forest

The bagging method of does a very good job at predicting multiclasses. It was able to use the data to produce a higher test accuracy F1-score for all cover types. With a overall weighted F1-score of 0.95 it is considerably higher than the other models. On average was 0.20 - 0.29 points higher the other models weighted average F1-scores.

The most inconsistent cover types performed better than the best prediction in any of the other models. Random Forest does a very good job at using the entire data set to bag data into classification leading to predictions.

	precision	recall	f1-score	support
1	0.96	0.94	0.95	70052
2	0.95	0.97	0.96	93189
3	0.94	0.96	0.95	11873
4	0.93	0.83	0.88	972
5	0.94	0.75	0.84	3124
6	0.93	0.89	0.91	5687
7	0.97	0.95	0.96	6837
micro avg	0.95	0.95	0.95	191734
macro avg	0.95	0.90	0.92	191734
weighted avg	0.95	0.95	0.95	191734