PROJECT REPORT

Prediction of Forest Cover Type – UCI ML Repository

Submitted by

Student Name	SIS ID
Prashanth G	JGB3CJVL39
Amulya Manne	1ONCHCD7TT
Apoorva	Q5GLNBWRSW
Ganesh Darshan	T4JDHYFWFO

Batch: DSE_BANGALORE_AUG2019

Mentor: Mr. Ankush Bansal

GREAT LAKES
INSTITUTE OF MANAGEMENT, CHENNAI

greatlearning
Learning for Life

Abstract

Natural resource managers responsible for developing ecosystem management strategies require basic descriptive information including inventory data for forested lands to support their decision-making processes. However, managers

generally, do not have this type of data for inholdings or neighboring lands that are outside their immediate jurisdiction. One method of obtaining this information is

using predictive models. The aim of this study is to determine the cover type of the Roosevelt National Forest of Northern Colorado. The obtained data of Cartographic variables only (no remotely sensed data) is fed to machine learning classification methods to build a model. The results show that on raw data Decision tree produces significantly higher accuracy than Logistic Regression.

Acknowledgements

At the outset, we are indebted to our Mentor Mr. Ankush Bansal for his time, valuable inputs, and guidance. His experience, support and structured thought process guided us to be on the right track towards completion of this project.

We also thank all the course faculty of the DSE program for providing us a strong foundation in various concepts of analytics & machine learning.

Last but not the least, we would like to sincerely thank our respective families for giving us the necessary support, space and time to complete this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Prashanth G
Amulya Manne
Apoorva
Ganesh
Darshan

Date: 17.01.2020

Place: Bangalore

Table of Contents

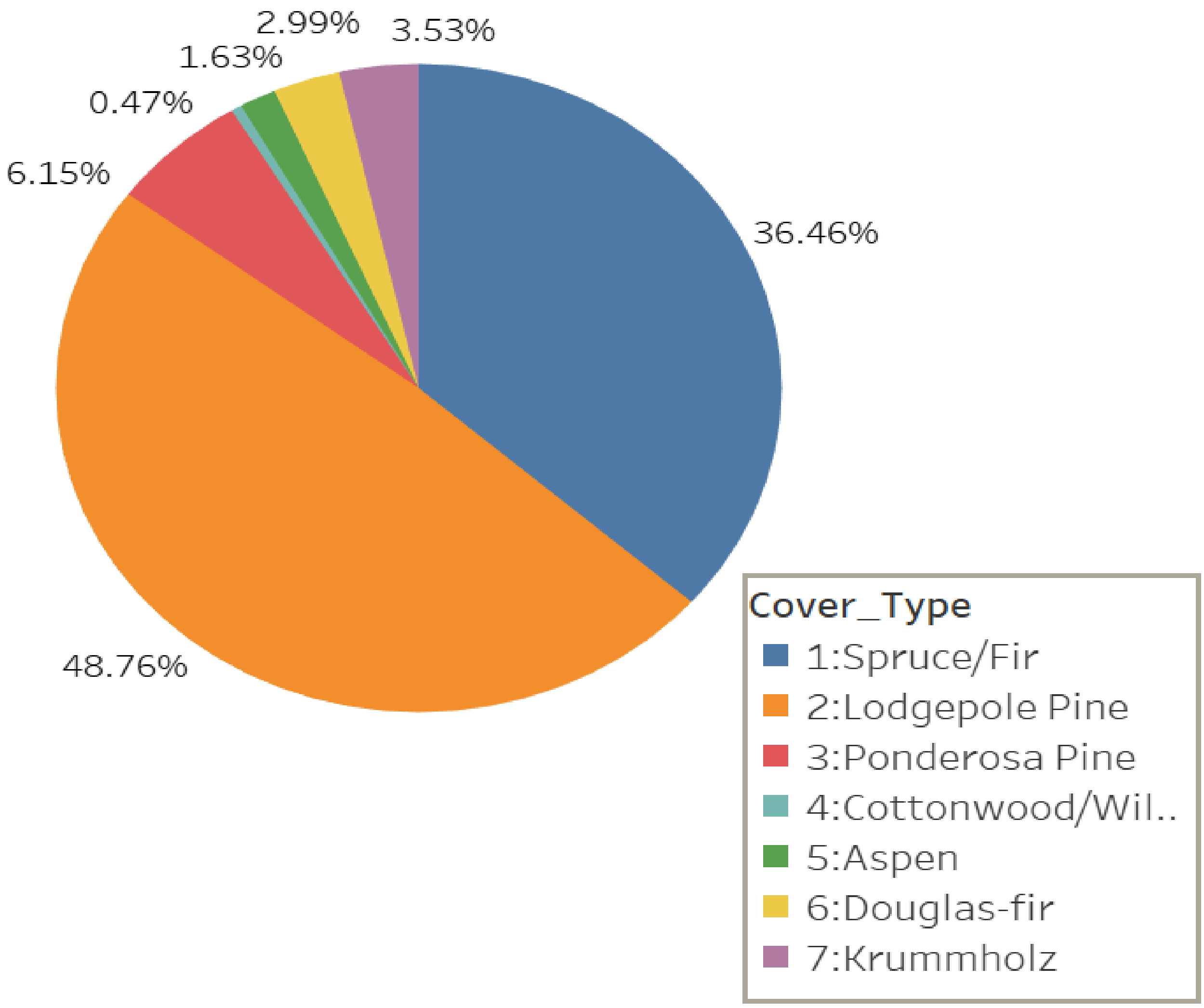
Abbreviations used

Abbreviation	Expansion
LR	Logistic Regression
DT	Decision Tree
RF	Random Forest
NB	Naïve Bayes
KNN	K nearest Neighbors

Chapter 1 - Project overview

Predicting forest cover type from cartographic variables only (no remotely sensed data). The actual forest cover type for a given observation (30 x 30 meter cell) was determined from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. Independent variables derived from data originally obtained from US Geological Survey (USGS) and USFS data. Data is in raw form (not scaled) and contains binary (0 or 1) columns of data for qualitative independent variables (wilderness areas and soil types).

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.



Executive summary

Background & need for study: The forest region in the globe is occupying a wide area and it is important to have a proper data of these regions to know the where abouts and to look into the safety of the animals and people residing in these area. It is practically not possible to investigate every nomenclature in order to determine the cover type, which would take a large amount of time. Hence, we are trying to collect the cartographic variables and we are trying to predict the cover type by using machine-learning models. The study area included four wilderness areas found in the Roosevelt National Forest of northern Colorado. Twelve cartographic measures considered as independent variables in the predictive models, while seven major forest cover types were used as dependent variables. Several subsets of these variables were examined to determine the best overall predictive model.

Scope & Objectives: The objective of this project is to do a research and develop a methodology by building models for Forest Cover type Detection. By analyzing the hill shade at various time, soil types, distance from various geographical features as the result of these data will be determined. Acceptable actions will be used as labels during pattern definition with supervised learning algorithms. Thus, in future when they get the data of any forest with predefined pattern, they will be tagged with the obtained pattern function and the action to be taken instantaneously will be determined.

Approach & methodology: The data extracted from UCI machine learning Repository platform. After processing the dataset and cleaning the inconsistencies, the numerical and categorical features used in the purchasing intention prediction model is generated. As of now no treatment on data is done Classification algorithms are used to produce emerging customer segments and relevant target marketing activities for each segment.

Key learnings: The forty types of soil already being label encoded, several class have very negligible amount of presence in the survey hence it is better to combine them into classes in order to understand the data in a better way.

Recommendations & actionable insights: The high-level recommendations for the project are developed by predicting forest cover type. The forestry department of the area to understand the whereabouts of the forest in a better way. Analysis on the Fire Points in the forest.

Current baseline & business mission:

For each subset of cartographic variables examined in this study, relative classification accuracies indicate the neural network approach outperformed the traditional discriminant analysis method in predicting forest cover types. The final neural network model had a higher absolute classification accuracy (70.58%) than the final corresponding linear discriminant analysis model(58.38%). In support of these classification results, thirty additional networks with randomly selected initial weights were derived. From these networks, the overall mean absolute classification accuracy for the neural network method was 70.52%, with a 95% confidence interval of 70.26% to 70.80%. Consequently, natural resource managers may utilize an alternative method of predicting forest cover types that is both superior to the traditional statistical methods and adequate to support their decision-making processes for developing ecosystem management strategies.

The Current Business mission is to increase the prediction efficiency and explore the data to gain more insight.

Data sources:

(a) Original owners of database: Remote Sensing and GIS Program Department of Forest Sciences

College of Natural Resources Colorado State University

Fort Collins, CO 80523

(contact Jock A. Blackard, jblackard 'at' fs.fed.us

or Dr. Denis J. Dean, denis.dean 'at' utdallas.edu)

(b) Donors of database:

Jock A. Blackard (jblackard 'at' fs.fed.us) GIS Coordinator

USFS - Forest Inventory & Analysis

Rocky Mountain Research Station

507 25th Street

Ogden, UT 84401

Dr. Denis J. Dean (denis.dean 'at' utdallas.edu) Professor

Program in Geography and Geospatial Sciences

School of Economic, Political and Policy Sciences

800 West Campbell Rd

Richardson, TX 75080-3021

Dr. Charles W. Anderson (anderson 'at' cs.colostate.edu)

Associate Professor

Department of Computer Science

Colorado State University

Fort Collins, CO 80523 USA

Date donated: August 199

Dataset Description

Data is in raw form (not scaled) and contains binary (0 or 1) columns of data for qualitative independent variables (wilderness areas and soil types). 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.

Table 1 – Numerical Description of the features used in the analysis model

Feature Name	Feature Description	Min Value	Max Value	Std Dev.
Elevation	Height of the tree above the ground level. Elevation in meters	1859	3858	279.99
Aspect	The aspect identifies the compass direction that the downhill slope faces for each location. Aspect in degrees azimuth	0	360	111.91
Slope	Identifies the slope (gradient or steepness) from each cell of a raster. Angle of the slope of the region Slope in degrees	0	66	1.26
Horizontal Distance To Hydrology	Horizontal Distance to nearest surface water features	0	1397	212.55
Vertical Distance To Hydrology	Vertical Distance to nearest surface water features	-173	601	58.295
Horizontal Distance To Roadways	Horizontal Distance to nearest roadway	0	7117	1559.25
Hill-shade 9am	Hill-shade index at 9am, summer solstice	0	254	26.76
Hill-shade Noon	Hill-shade index at noon, summer solstice	0	254	19.76
Hill-shade 3pm	Hill-shade index at 3pm, summer solstice	0	254	38.27
Horizontal Distance To Fire Points	Horizontal Distance to nearest wildfire ignition points	0	7173	1324.19

Table 1 shows the numerical features along with their statistical parameters. Among these features 'Elevation', 'Aspect', 'Slope', 'Horizontal Distance To Hydrology', 'Vertical Distance To Hydrology', 'Horizontal Distance To Roadways', 'Hill-shade 9am', 'Hill-shade Noon', 'Hill-shade 3pm', 'Horizontal Distance To Fire Points' represent the various numerical features present in the data .

Table 2 – Categorical Features used in the Model

Feature Name	Feature Description	Number of Categorical Values
Wilderness Area	an area of undeveloped Federal land retaining its primeval character and influence, without permanent improvements or human habitation	4
Soil Type	Browser of the visitor	40
Cover Type	Geographic region from which the session has been started by the visitor	7

Table 2 shows the categorical features along with their categorical values.

Wilderness Area:

1. Rawah Wilderness Area (44.8864%)
2. Neota Wilderness Area (5.1434%)
3. Comanche Peak Wilderness Area (43.6074%)
4. Cache la Poudre Wilderness Area (6.3627%)

Extra Points :

- Neota (area 2) probably has the highest mean elevation value of the 4 wilderness areas. It would have spruce/fir (type 1),
- Rawah (area 1) and Comanche Peak (area 3) would have a lower mean elevational value.
- Cache la Poudre (area 4) would have the lowest mean elevation value.
- Rawah and Comanche Peak would probably have lodgepole pine (type 2) as their primary species, followed by spruce/fir and aspen (type 5).
- Cache la Poudre would tend to have Ponderosa pine (type 3), Douglas-fir (type 6), and cottonwood/willow (type 4).

The Rawah and Comanche Peak areas would tend to be more typical of the overall dataset than either the Neota or Cache la Poudre, due to their assortment of tree species and range of predictive variable values (elevation, etc.) Cache la Poudre would probably be unique than the others, due to its relatively low elevation range and species composition.

Soil Type:

Climatic zone	geologic zones
1. lower montane dry	1. alluvium
2. lower montane	2. glacial
3. montane dry	3. shale
4. montane	4. sandstone
5. montane dry and montane	5. mixed sedimentary
6. montane and subalpine	6. unspecified in the USFS ELU Survey
7. subalpine	7. igneous and metamorphic
8. alpine	8. volcanic

Note: For the First 2 Digits gives the Climatic and Geologic zones as shown above. The third and fourth ELU digits are unique to the mapping unit and have no special meaning to the climatic or geologic zones

1 to 40: based on the USFS Ecological Land type Units (ELUs) for this study area

SL	USFS ELU Code	Description	Percentage
1	2702	Cathedral family - Rock outcrop complex, extremely stony.	0.5217
2	2703	Vanet - Ratake families' complex, very stony.	1.2952
3	2704	Haploborolis - Rock outcrop complex, rubbly.	0.8301
4	2705	Ratake family - Rock outcrop complex, rubbly.	2.1335
5	2706	Vanet family - Rock outcrop complex complex, rubbly.	0.2749
6	2717	Vanet - Wetmore families - Rock outcrop complex, stony.	1.1316
7	3501	Gothic family.	0.0181
8	3502	Supervisor - Limber families complex.	0.0308
9	4201	Troutville family, very stony.	0.1974
10	4703	Bullwark - Catamount families - Rock outcrop complex, rubbly.	5.6168
11	4704	Bullwark - Catamount families - Rock land complex, rubbly.	2.1359
12	4744	Legault family - Rock land complex, stony.	5.1584
13	4758	Catamount family - Rock land - Bullwark family complex, rubbly.	3.0001
14	5101	Pachic Argiborolis - Aquolis complex.	0.1031
15	5151	unspecified in the USFS Soil and ELU Survey.	0.0005
16	6101	Cryaquolis - Cryoborolis complex.	0.4897

17	6102	Gateview family - Cryaquolis complex.	0.589
18	6731	Rogert family, very stony.	0.3268
19	7101	Typic Cryaquolis - Borohemists complex.	0.6921
20	7102	Typic Cryaquepts - Typic Cryaquolls complex.	1.5936
21	7103	Typic Cryaquolls - Leighcan family, till substratum complex.	0.1442
22	7201	Leighcan family, till substratum, extremely bouldery.	5.744
23	7202	Leighcan family, till substratum - Typic Cryaquolls complex.	9.9399
24	7700	Leighcan family, extremely stony.	3.6622
25	7701	Leighcan family, warm, extremely stony.	0.0816
26	7702	Granile - Catamount families complex, very stony.	0.4456
27	7709	Leighcan family, warm - Rock outcrop complex, extremely stony.	0.1869
28	7710	Leighcan family - Rock outcrop complex, extremely stony.	0.1628
29	7745	Como - Legault families complex, extremely stony.	19.8354
30	7746	Como family - Rock land - Legault family complex, extremely stony.	5.1927
31	7755	Leighcan - Catamount families complex, extremely stony.	4.4175
32	7756	Catamount family - Rock outcrop - Leighcan family complex, extremely stony.	9.0392
33	7757	Leighcan - Catamount families - Rock outcrop complex, extremely stony.	7.7716
34	7790	Cryorthents - Rock land complex, extremely stony.	0.2773
35	8703	Cryumbrepts - Rock outcrop - Cryaquepts complex.	0.3255
36	8707	Bross family - Rock land - Cryumbrepts complex, extremely stony.	0.0205
37	8708	Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony.	0.0513
38	8771	Leighcan - Moran families - Cryaquolls complex, extremely stony.	2.6803
39	8772	Moran family - Cryorthents - Leighcan family complex, extremely stony.	2.3762
40	8776	Moran family - Cryorthents - Rock land complex, extremely stony.	1.506

Data preparation & clean up

The source dataset received has been prepared to ensure that the fields are cleaned up, the values are suitable for model building and the variable names are self-explanatory. The broad approach for data preparation can be outlined as:

Table 4 – Data pre-processing steps

Usually The various steps used in data pre-processing are:
1.Label Encoding
2.Outlier Treatment
3.Standardization
4.Oversampling

Statistical tools &techniques

We have performed 3 statistical tests to identify the significant variables.

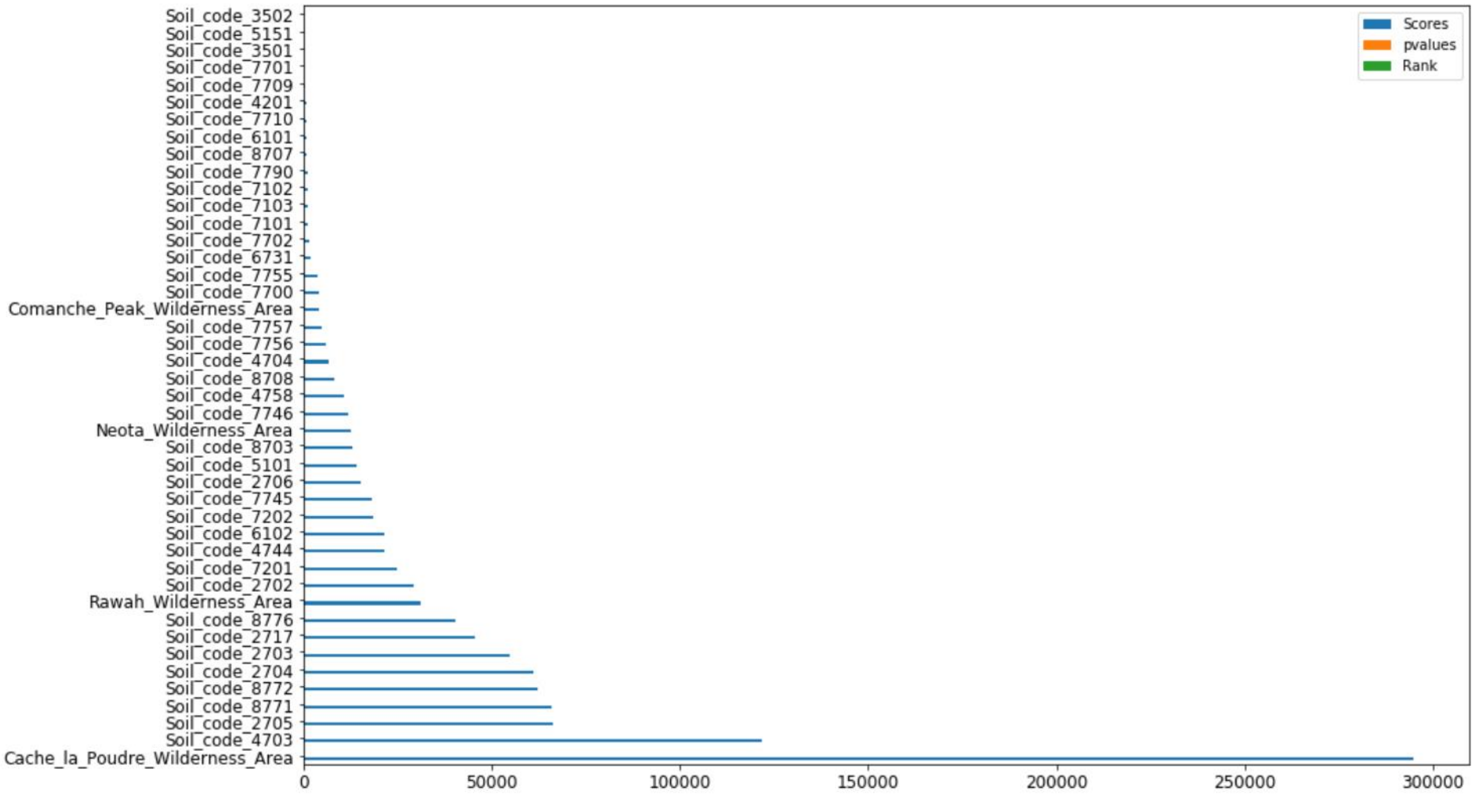
- ANOVA – For continuous variables with Target (Categorical)
- Chi2 test– For categorical variables with Target (Categorical)
- Mutual – For both categorical and continuous with Target (Categorical)

Results of Chi2 test:

Features	Scores	pvalues	Rank
Cache_la_Poudre_Wilderness_Area	0.26696	0.00E+00	1
Soil_code_4703	0.110353	0.00E+00	2
Soil_code_2705	0.060088	0.00E+00	3
Soil_code_8771	0.059469	0.00E+00	4
Soil_code_8772	0.056235	0.00E+00	5
Soil_code_2704	0.055308	0.00E+00	6
Soil_code_2703	0.049745	0.00E+00	7
Soil_code_2717	0.041331	0.00E+00	8
Soil_code_8776	0.036524	0.00E+00	9
Rawah_Wilderness_Area	0.028061	0.00E+00	10
Soil_code_2702	0.026331	0.00E+00	11
Soil_code_7201	0.022342	0.00E+00	12
Soil_code_4744	0.019561	0.00E+00	13
Soil_code_6102	0.019447	0.00E+00	14
Soil_code_7202	0.01662	0.00E+00	15
Soil_code_7745	0.016397	0.00E+00	16
Soil_code_2706	0.013867	0.00E+00	17
Soil_code_5101	0.012907	0.00E+00	18
Soil_code_8703	0.011614	0.00E+00	19
Neota_Wilderness_Area	0.011245	0.00E+00	20
Soil_code_7746	0.010868	0.00E+00	21
Soil_code_4758	0.009783	0.00E+00	22

Soil_code_8708	0.007372	0.00E+00	23
Soil_code_4704	0.006117	0.00E+00	24
Soil_code_7756	0.005495	0.00E+00	25
Soil_code_7757	0.004322	0.00E+00	26
Comanche_Peak_Wilderness_Area	0.003723	0.00E+00	27
Soil_code_7700	0.003584	0.00E+00	28
Soil_code_7755	0.00326	0.00E+00	29
Soil_code_6731	0.001821	0.00E+00	30
Soil_code_7702	0.001495	0.00E+00	31
Soil_code_7101	0.001187	4.90E-280	32
Soil_code_7103	0.001163	2.52E-274	33
Soil_code_7102	0.001033	2.93E-243	34
Soil_code_7790	0.000975	2.50E-229	35
Soil_code_8707	0.000779	1.15E-182	36
Soil_code_6101	0.000716	1.66E-167	37
Soil_code_7710	0.000715	2.52E-167	38
Soil_code_4201	0.000591	7.41E-138	39
Soil_code_7709	0.000221	6.07E-50	40
Soil_code_7701	0.00013	1.91E-28	41
Soil_code_3501	0.0001	1.73E-21	42
Soil_code_5151	0.000088	8.89E-19	43
Soil_code_3502	0.000055	2.51E-11	44

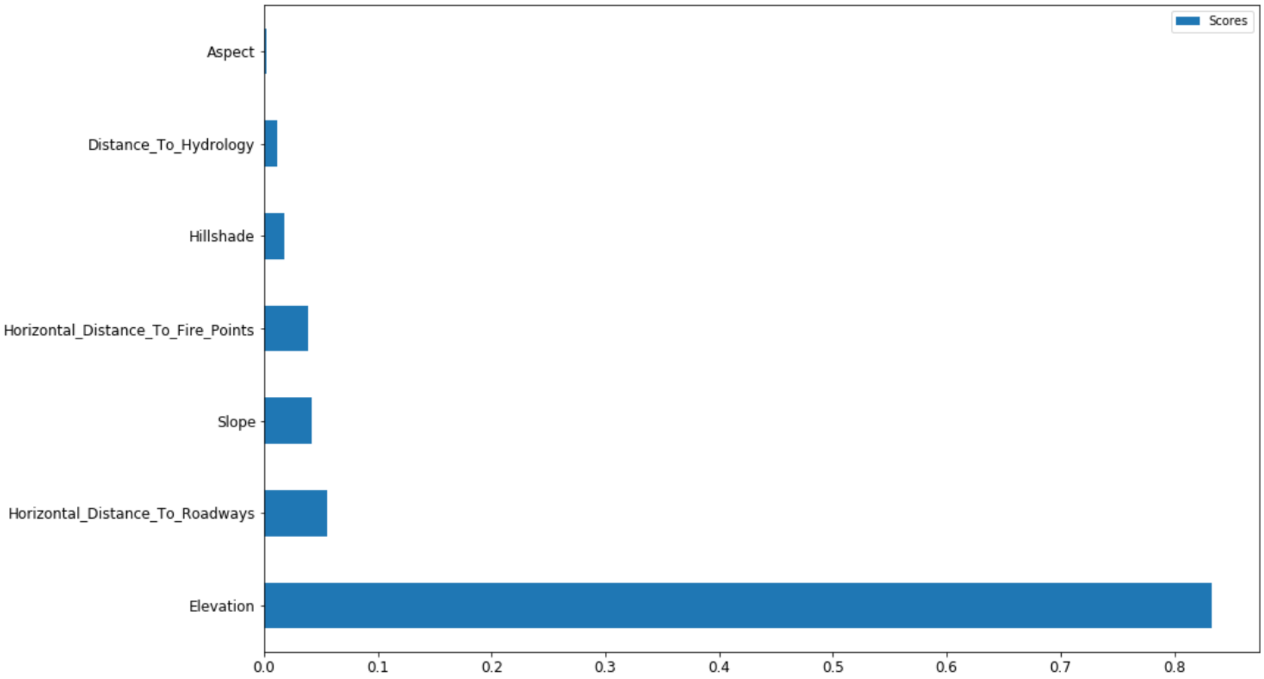
Below is the Image for Chi2 Test



Results of Anova:

Features	Scores	pvalues	Rank
Elevation	0.832921	0	1
Horizontal_Distance_To_Roadways	0.055109	0	2
Slope	0.041632	0	3
Horizontal_Distance_To_Fire_Points	0.03895	0	4
Hillshade	0.0174	0	5
Distance_To_Hydrology	0.011593	0	6
Aspect	0.002396	0	7

Below is the Image for Anova

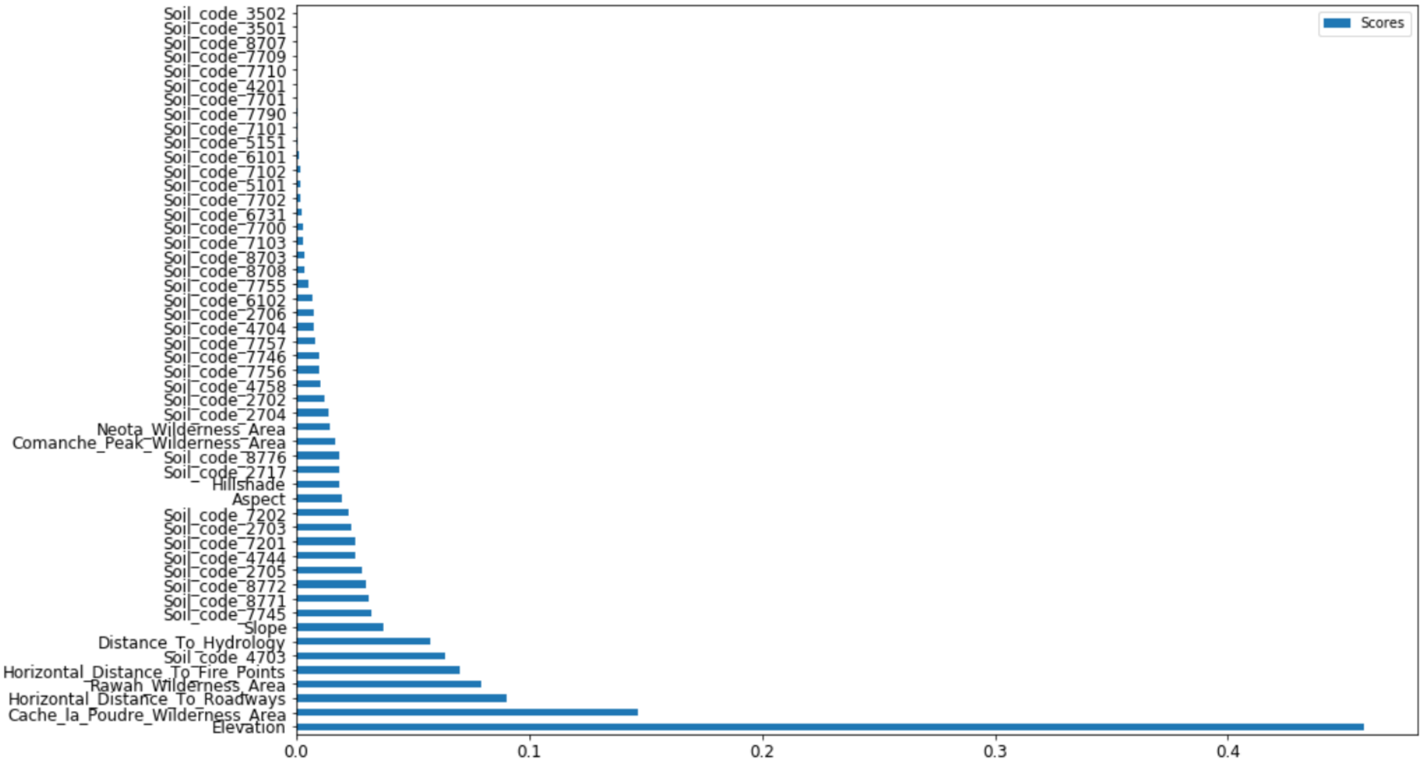


Results of Mutual(SellectKbest):

Features	Scores	Rank
Elevation	0.458464	1
Cache_la_Poudre_Wilderness_Area	0.146662	2
Horizontal_Distance_To_Roadways	0.090144	3
Rawah_Wilderness_Area	0.079381	4
Horizontal_Distance_To_Fire_Points	0.070182	5
Soil_code_4703	0.064018	6
Distance_To_Hydrology	0.057511	7
Slope	0.037712	8
Soil_code_7745	0.032467	9
Soil_code_8771	0.031126	10
Soil_code_8772	0.029958	11
Soil_code_2705	0.028234	12
Soil_code_4744	0.025561	13
Soil_code_7201	0.025195	14
Soil_code_2703	0.023562	15

Soil_code_7202	0.022547	16
Aspect	0.019413	17
Hillshade	0.01872	18
Soil_code_2717	0.018625	19
Soil_code_8776	0.018203	20
Comanche_Peak_Wilderness_Area	0.016845	21
Neota_Wilderness_Area	0.014138	22
Soil_code_2704	0.014094	23
Soil_code_2702	0.012115	24
Soil_code_4758	0.01039	25
Soil_code_7756	0.009981	26
Soil_code_7746	0.009667	27
Soil_code_7757	0.008376	28
Soil_code_4704	0.007625	29
Soil_code_2706	0.007296	30
Soil_code_6102	0.006811	31
Soil_code_7755	0.005419	32
Soil_code_8708	0.003543	33
Soil_code_8703	0.003492	34
Soil_code_7103	0.002928	35
Soil_code_7700	0.002666	36
Soil_code_6731	0.002131	37
Soil_code_7702	0.002055	38
Soil_code_5101	0.001902	39
Soil_code_7102	0.001584	40
Soil_code_6101	0.001012	41
Soil_code_5151	0.000892	42
Soil_code_7101	0.000811	43
Soil_code_7790	0.000386	44
Soil_code_7701	0.000181	45
Soil_code_4201	0.000025	46
Soil_code_7710	0.000011	47

Soil_code_7709	0	48
Soil_code_8707	0	48
Soil_code_3501	0	48
Soil_code_3502	0	48



After performing the above tests, we can see that all the features have p value less than 0 which indicates that they are significant variables.

Model performance measures used for evaluating models

The various models built, must be evaluated based on certain model performance measures to identify the most robust models. The choice of the right model performance measures is highly critical since the dataset is a highly imbalanced dataset and the conversion rate is 15.47%. Model accuracy alone may not be enough to evaluate a model. Hence the following model performance measures have been used to evaluate the models, based on the confusion matrix built for the predictions on the training and test datasets:

	Negative (Predicted)	**Positive (Predicted)**
Negative (Observed)	True Negative (TN)	False positive (FP)
Positive (Observed)	False negative (FN)	True positive (TP)

Accuracy

Accuracy is the number of correct predictions made by the model by the total number of records. The best accuracy is 100% indicating that all the predictions are correct.

Considering the response rate (conversion rate) of our dataset which is ~16%, accuracy is not a valid measure of model performance. Even if all the records are predicted as 0, the model will still have an accuracy of 84%. Hence other model performance measures need to be evaluated.

Sensitivity or recall

Sensitivity (Recall or True positive rate) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall or true positive rate (TPR).

For our dataset, it gives the ratio of actual customers who generated revenue by the total number of customers predicted who will generate the revenue.

Specificity

Specificity (true negative rate) is calculated as the number of correct negative predictions divided by the total number of negatives.

For our dataset, specificity gives the ratio of actual customers who will not generate revenue by the number of customers who are predicted who will not generate revenue.

Precision

Precision (Positive predictive value) is calculated as the number of correct positive predictions divided by the total number of positive predictions.Precision tells us, what proportion of customers who generated revenue as customers actually generated revenue. If precision is low, it implies that the model has lot of false positives.

F1-Score

F1 is an overall measure of a model's accuracy that combines precision and recall A good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0.

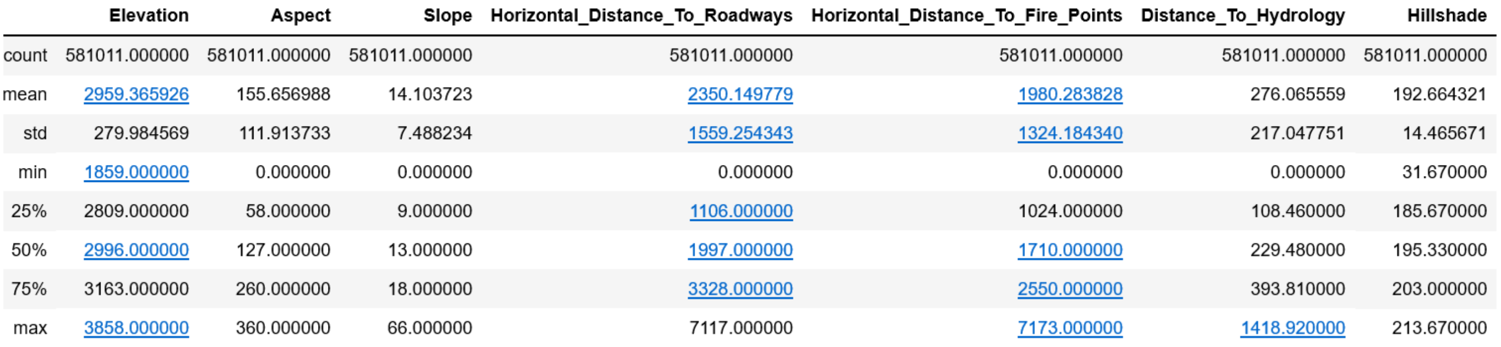
**For this model we are only using Recall to evaluate the performance.

Chapter 2 - Exploratory data analysis

The purpose of exploratory data analysis is two-fold:

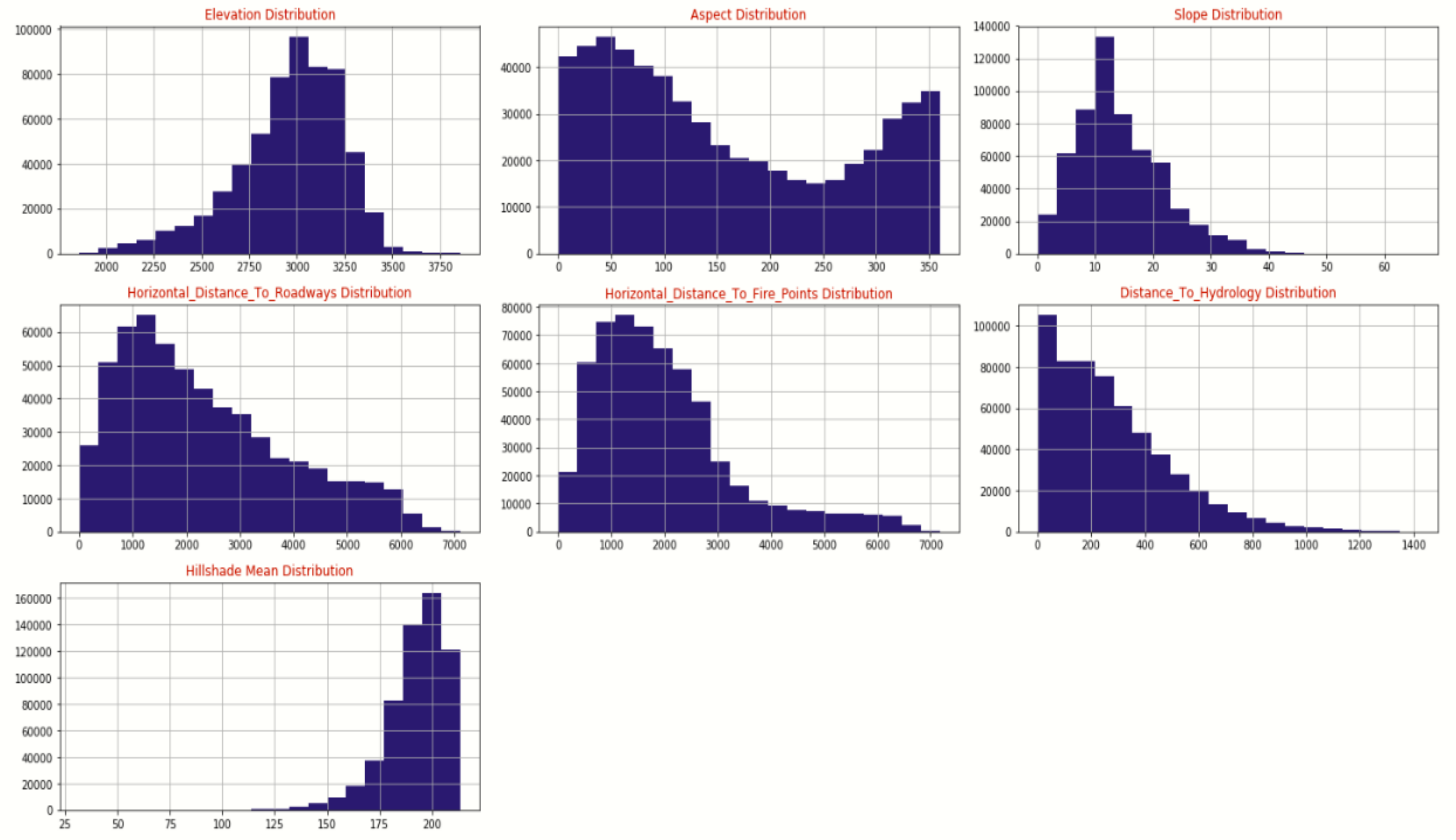
- to understand the features in the given data in a better way and its influence on the target column.
- Get insights on various features.

Data Statistical Description:



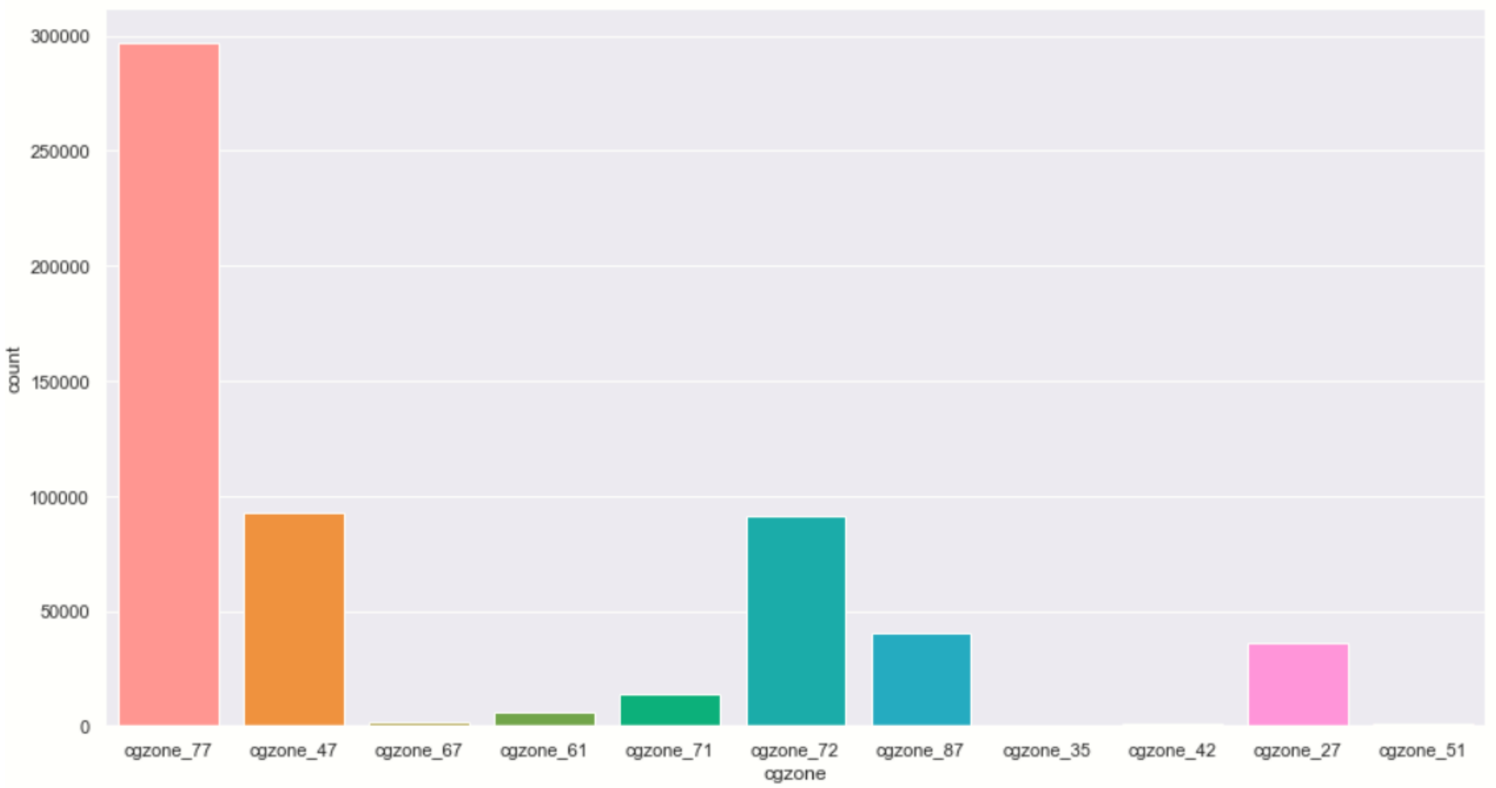
	Elevation	Aspect	Slope	Horizontal_Distance_To_Roadways	Horizontal_Distance_To_Fire_Points	Distance_To_Hydrology	Hillshade
count	581011.000000	581011.000000	581011.000000	581011.000000	581011.000000	581011.000000	581011.000000
mean	2959.365926	155.656988	14.103723	2350.149779	1980.283828	276.065559	192.664321
std	279.984569	111.913733	7.488234	1559.254343	1324.184340	217.047751	14.465671
min	1859.000000	0.000000	0.000000	0.000000	0.000000	0.000000	31.670000
25%	2809.000000	58.000000	9.000000	1106.000000	1024.000000	108.460000	185.670000
50%	2996.000000	127.000000	13.000000	1997.000000	1710.000000	229.480000	195.330000
75%	3163.000000	260.000000	18.000000	3328.000000	2550.000000	393.810000	203.000000
max	3858.000000	360.000000	66.000000	7117.000000	7173.000000	1418.920000	213.670000

EDA Univariate Analysis.



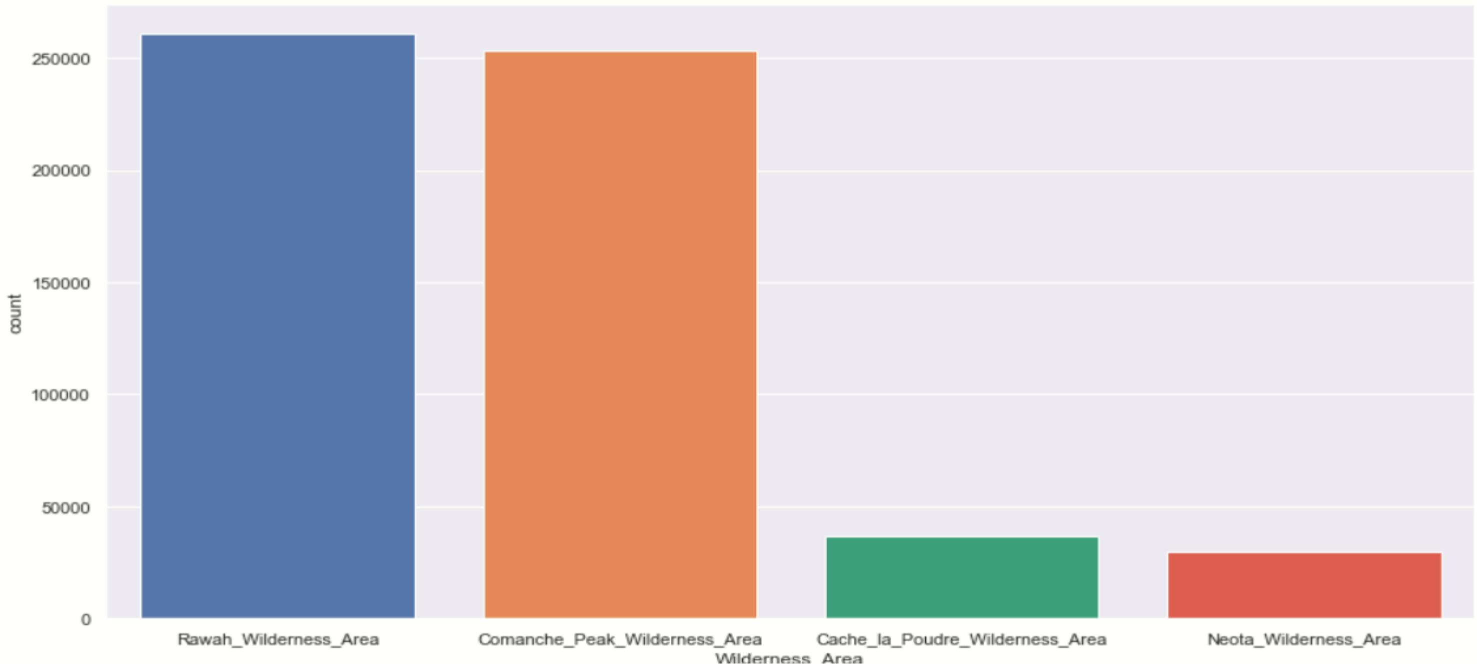
We can see that most of the data max is good except Distance to Hydrology, Distance to Fire points and Roadways were the maximum values are very far from Q3. Also We can see that most of the Distributions are skewed.

CG ZONE



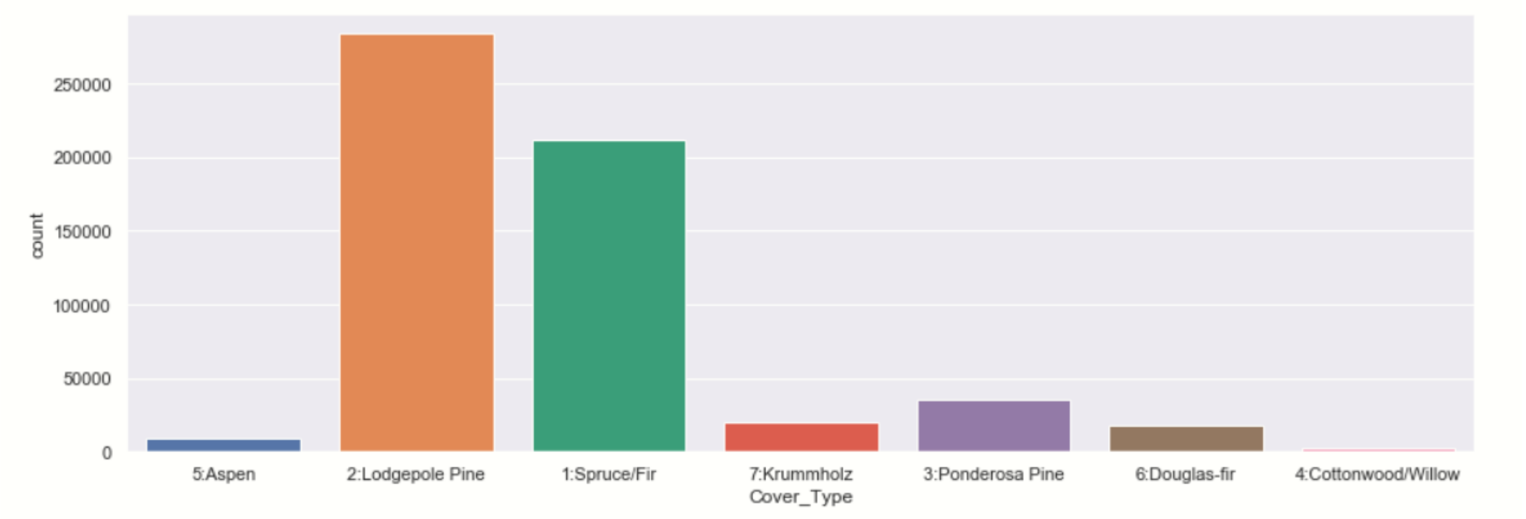
The Soil in the Cgzone_77 has the highest presence and the cgzone_35, cgzone_42, cgzone_51 Has the least presence.

WILDERNESS AREA



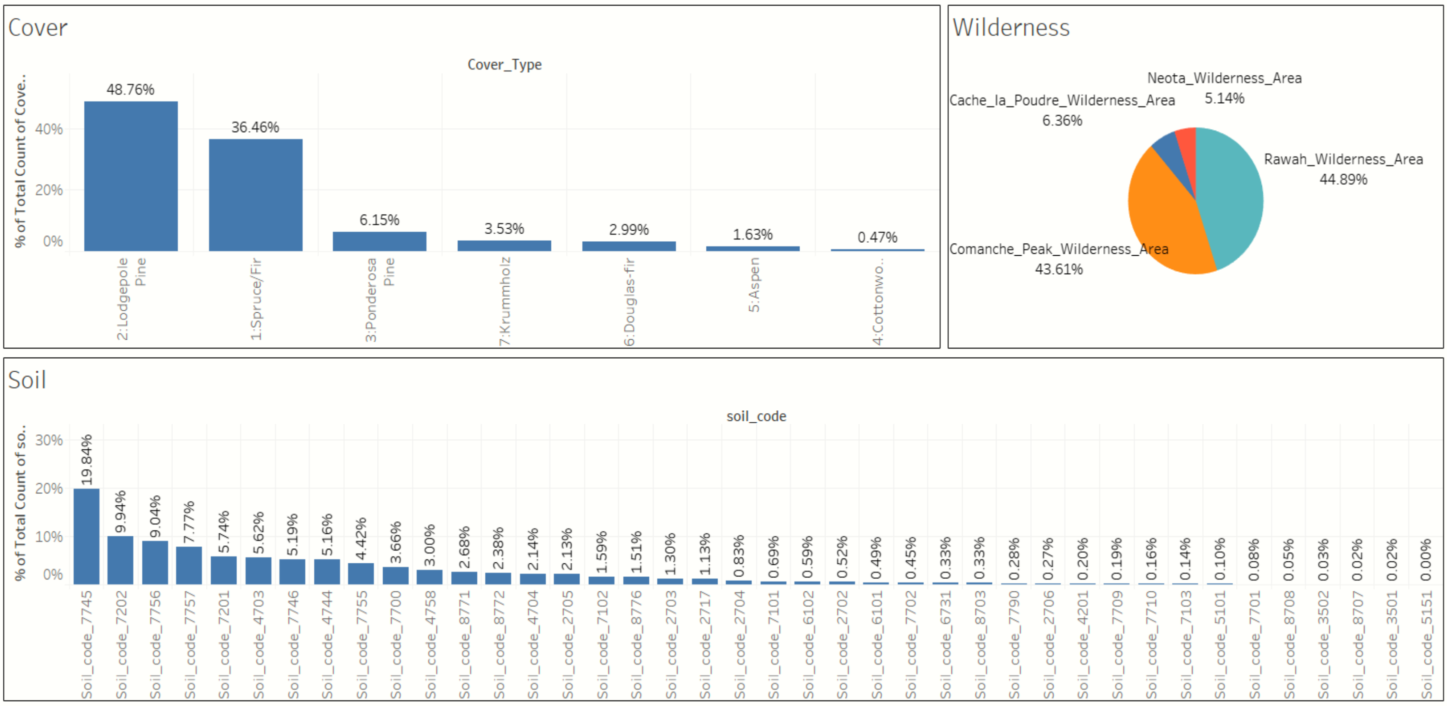
We can see that the Rawah Area has the highest presence and Nota area has the Least presence

COVER TYPE



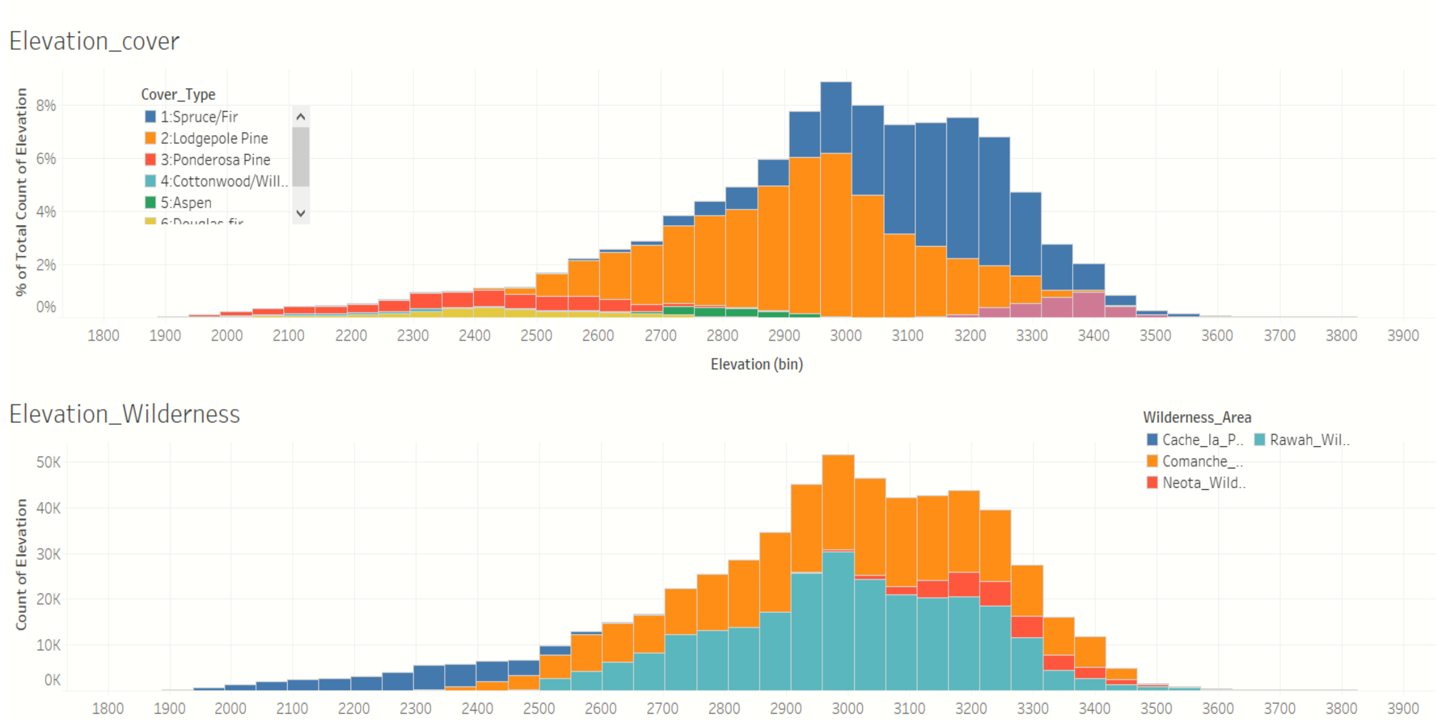
The Lodgepole pine is covering the majority area and cottonwood/willow is covering the least area.

EDA Multivariate Analysis _ Tableau report:



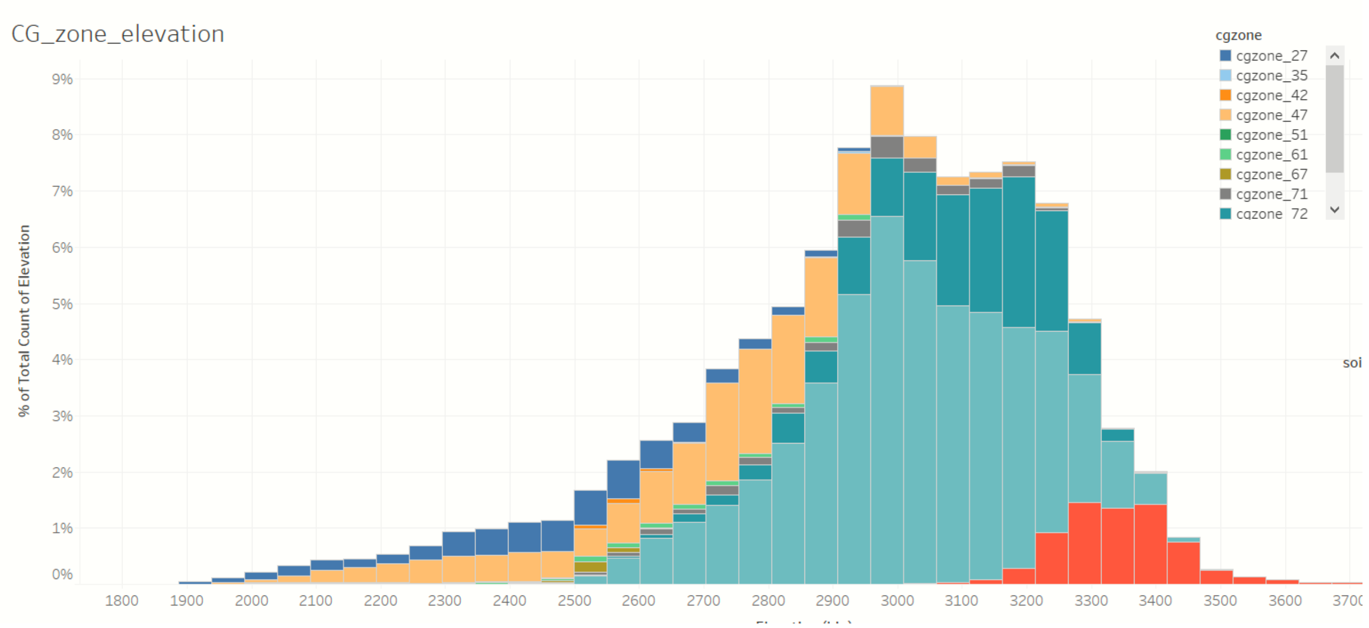
The Data is Highly Imbalanced with both Cover type and Wilderness Area.

Figure 1: Elevation



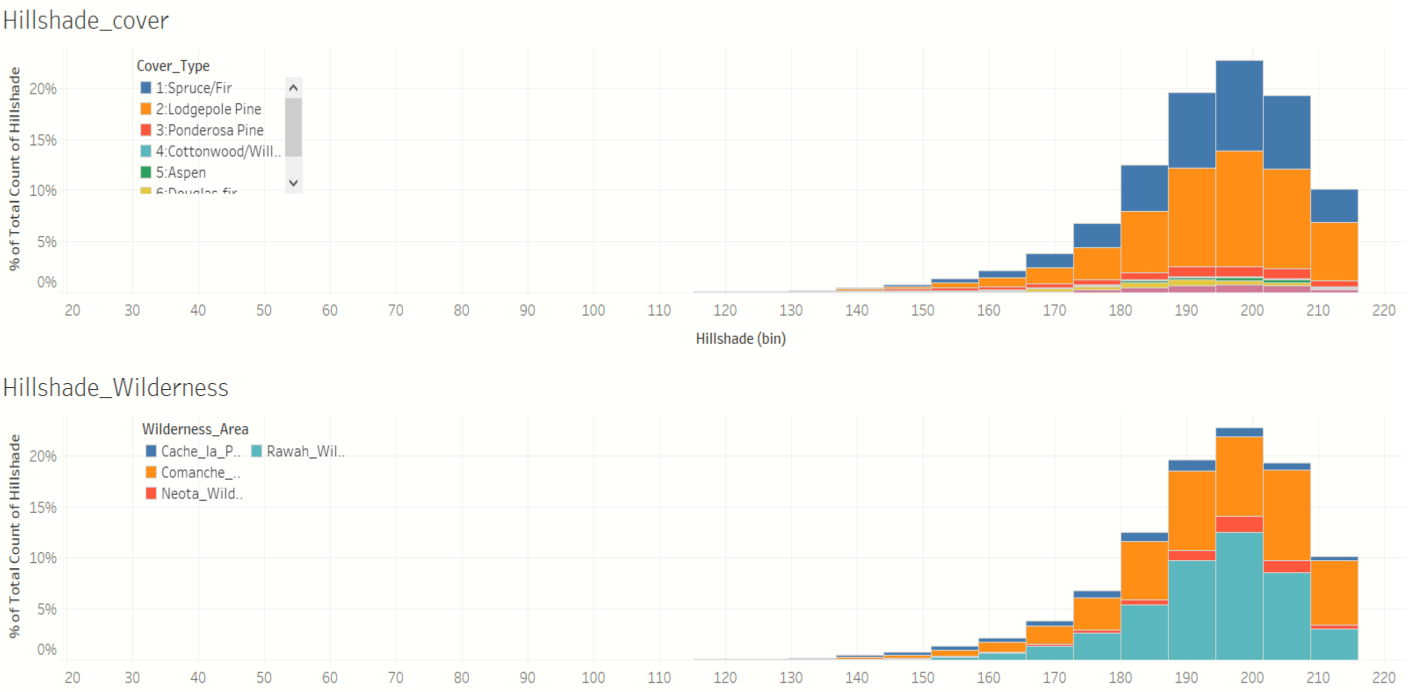
Description 1: Kurmmholz has the highest Mean Elevation and type 3 and 6 has the least Mean Elevation. Aspen is in the Middle of the spectrum and lastly type 1 and 2 has the highest spread of elevation from 2300 to 3500. Cache_la has the least mean elevation and Neota wilderness has the highest mean elevation. Comanche and Rawah have about the same spread across the spectrum

Figure 2: CG_Elevation



Description 2: CG_Zone47 seems to have the highest spread (2400mts). CG_zone35 has the least spread (100mts).CG-zone87 seems to have the Highest Mean Elevation. In zone47, soil_4703 is the highest contributor with maximum Spread. Soil_4758 is found at higher elevations between 2550 &3300. In zone_35 there are 2 soil types (3501,3502) both are equally spread. In Zone_87, soil_8776 has the highest mean spread between 3200, 3850.

Figure 3: Hillshade


Hillshade_cover

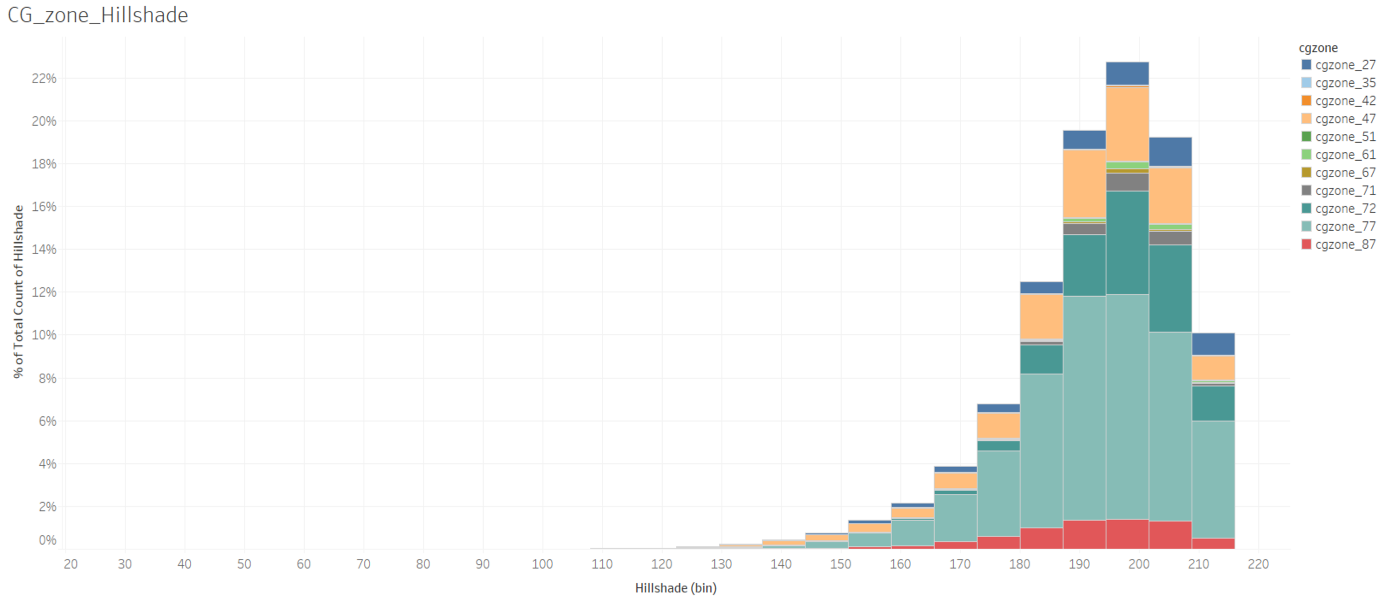
Hillshade_Wilderness

Description 3: Both the Cover Type and Wilderness areas are left skewed which means that mode>median>mean.

The Average sunlight is on the higher side and most of the sunlight index values lies between 115 and 215.

We can also say that most of the cover and Wilderness receive enough sunlight.

Figure 4: CG_Zone Hillshade

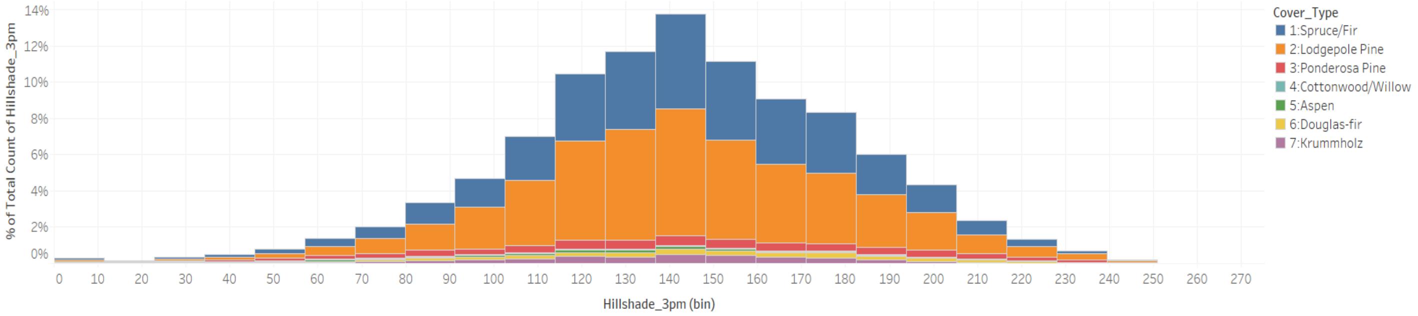

CG_zone_Hillshade

The trend of % of Total Count of Hillshade for Hillshade (bin). Color shows details about cgzone. The view is filtered on cgzone, which keeps 11 of 11 members.

Description 4: Hillshade at 3pm follows Normal Distribution with peek between 140 and 150. The Distribution is spread there is not much separation within the cover types and Wilderness areas.

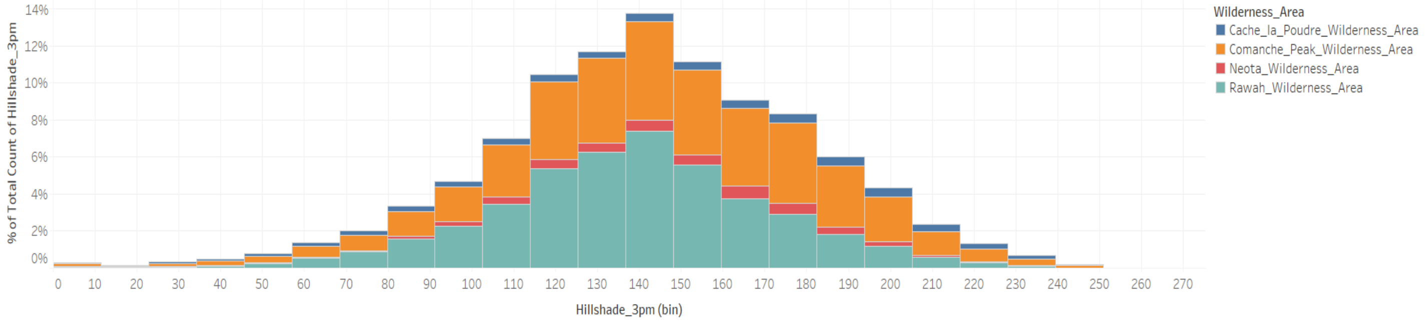
Figure 5: Hiillshade_3pm

Hillshade_3pm_cover



The trend of % of Total Count of Hillshade_3pm for Hillshade_3pm (bin). Color shows details about Cover_Type. The data is filtered on cgzone, which keeps 11 of 11 members. The view is filtered on % of Total Count of Hillshade_3pm, which keeps non-Null values only.

Hillshade_3pm_Wilderness

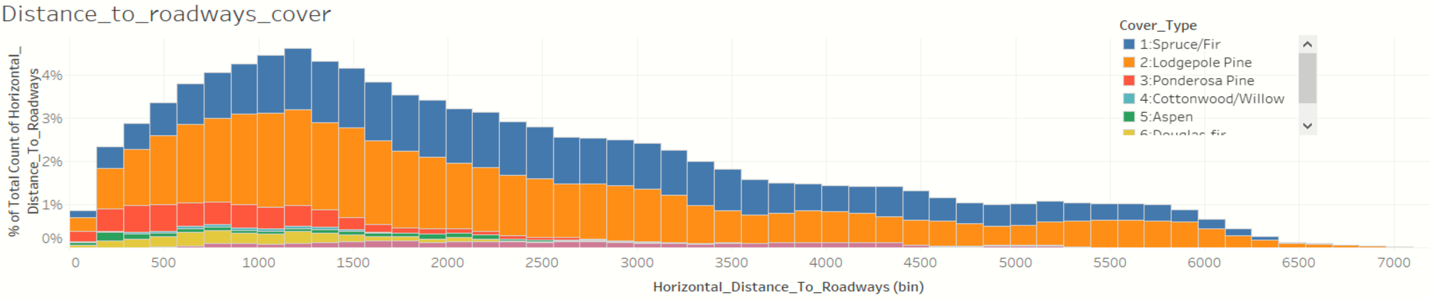


The trend of % of Total Count of Hillshade_3pm for Hillshade_3pm (bin). Color shows details about Wilderness_Area. The data is filtered on cgzone, which keeps 11 of 11 members. The view is filtered on % of Total Count of Hillshade_3pm, which keeps non-Null values only.

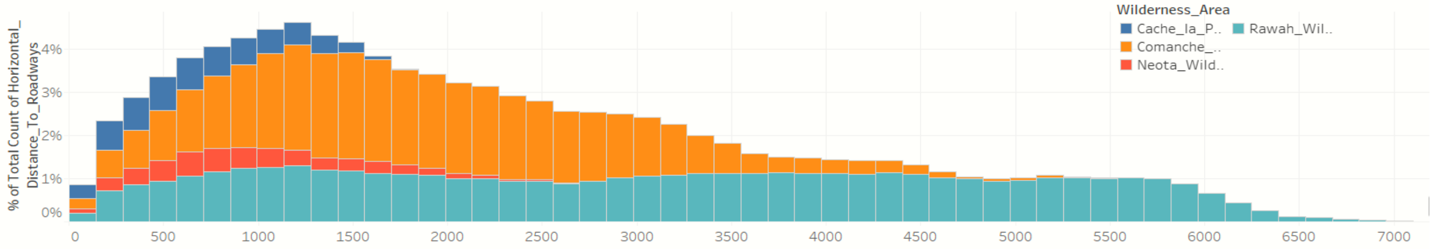
Description 5: Hillshade at 3pm follows Normal Distribution with peek between 140 and 150. The Distribution is spread evenly and not much separation within the cover types and Wilderness areas

Figure 6: Distance to Roadways

Distance_to_roadways_cover



Distance_to_roadways_Wilderness

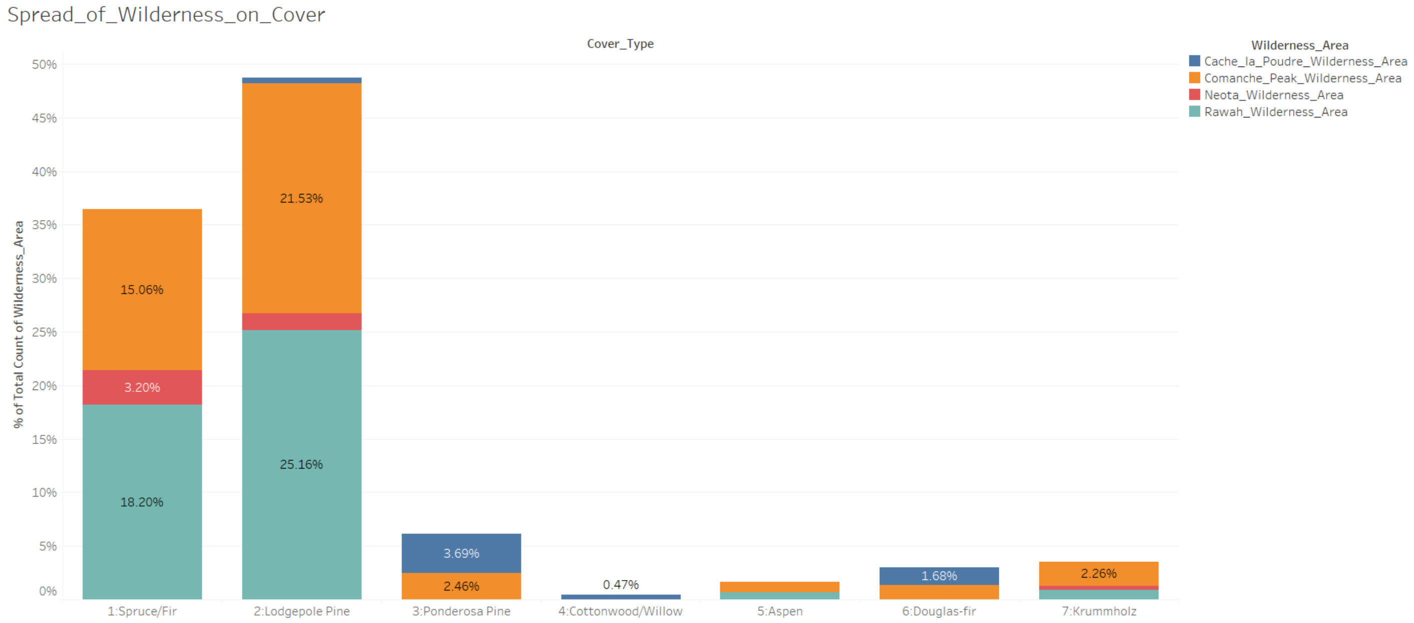


Description 6:

a.Type 4 is dense near the roadways(0-1700mts).Type 3,6 is spread between (0-3000mts). The rest is spread across the spectrum.

b. Cache_la is closest to roadways at (0-1800mts). Neota is the second closest at (0-2400mts). The rest are spread across the spectrum. We can say that the wildlife at Comanche and Rawah is not much affected by humans since this is spread much away from roadways.

Figure 7:Spread of wilderness on cover



% of Total Count of Wilderness_Area for each Cover_Type. Color shows details about Wilderness_Area. The marks are labeled by % of Total Count of Wilderness_Area. The view is filtered on Wilderness_Area, which keeps Cache_la_Poudre_Wilderness_Area, Comanche_Peak_Wilderness_Area, Neota_Wilderness_Area and Rawah_Wilderness_Area.

Description 7: Cache_la area seems to have 3.Ponderosa pine and 6.Douglas-fir as primary species.

Though Comanche area seems to have 1.Spruce/Fir and 2.Lodgepole pine as majority species. Comanche area has the highest Divercity of species.

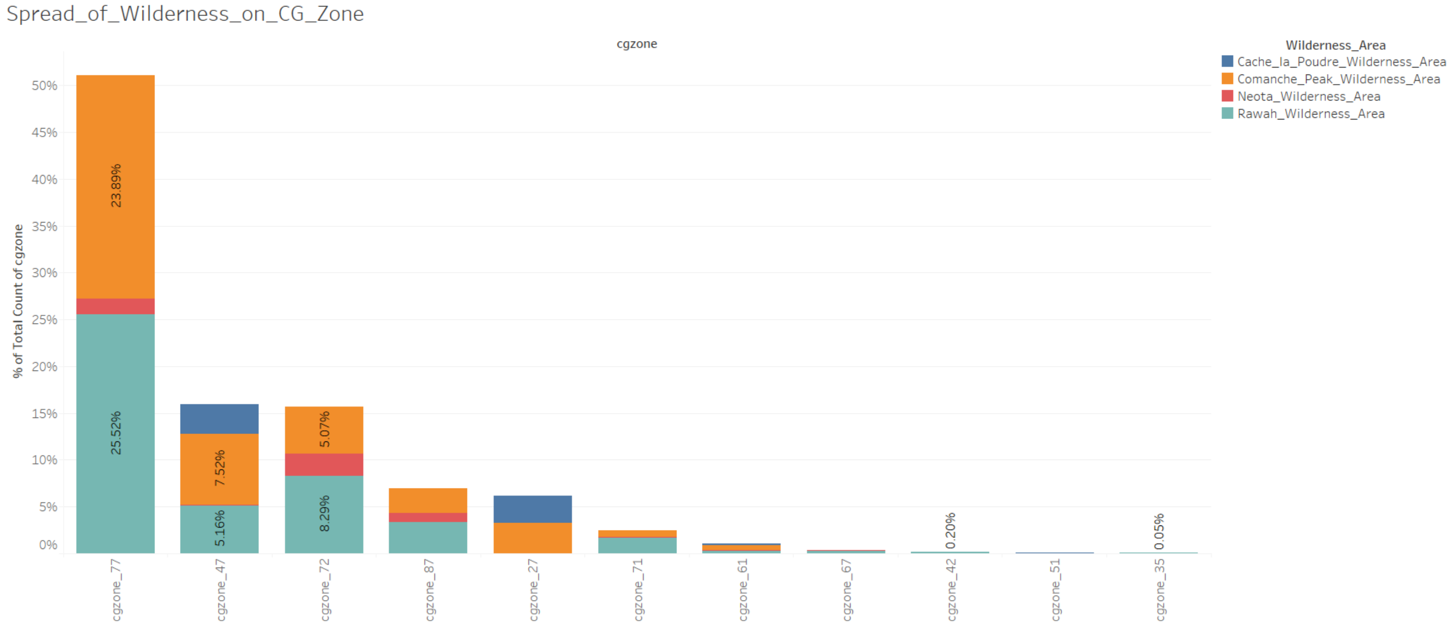
Neota and Rawah areas has 1.Spruce/Fir and 2.Lodgepole pine as mojor species

Cottonwood/willow is specific only to Cache_la area.

1.Spruce/Fir , 2.Lodgepole pine and 7.kurmmholz are found in all the areas except Cache_la

Cache_la area is unique and has species such as 3.Ponderosa pine ,4.Cottonwood/willow, and 6.Douglas-fir.

Figure 8 : Spread of Wilderness on CG Zone



% of Total Count of cgzone for each cgzone. Color shows details about Wilderness_Area. The marks are labeled by % of Total Count of cgzone. The view is filtered on Wilderness_Area, which keeps Cache_la_Poudre_Wilderness_Area, Comanche_Peak_Wilderness_Area, Neota_Wilderness_Area and Rawah_Wilderness_Area.

Description 8:

Cache_la area has predominently CG_zone47 and CG_zone27.

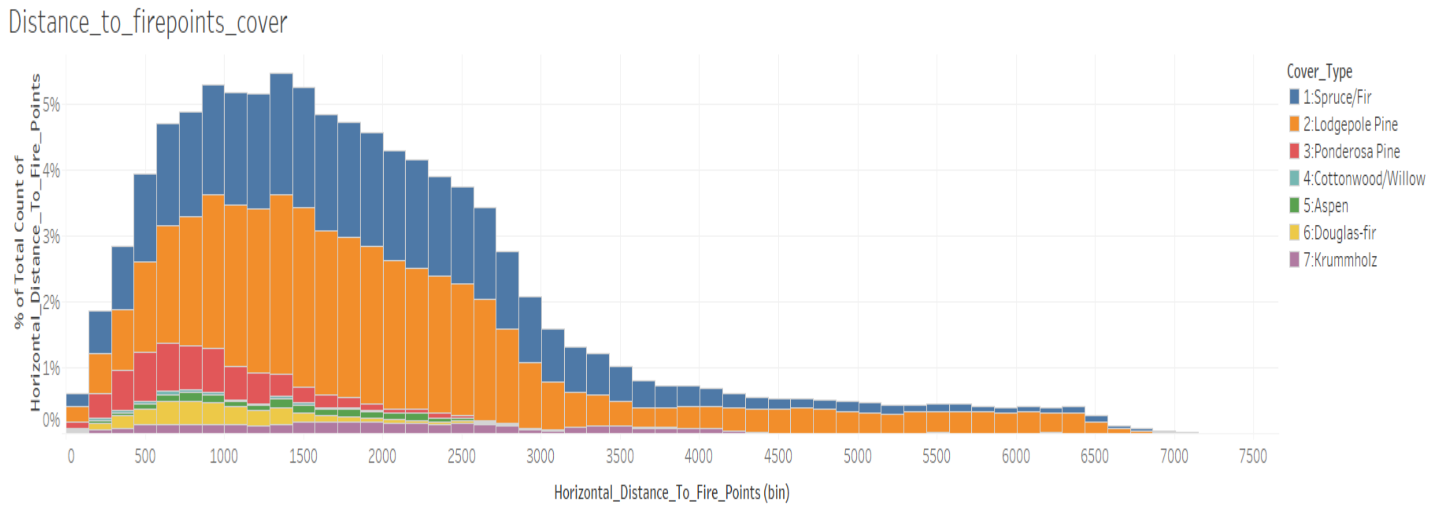
Neota area predominently CG_zone77,CG_zone72 and CG_zone87.

Comanche and Rawah areas has high diversity of CG_zones.

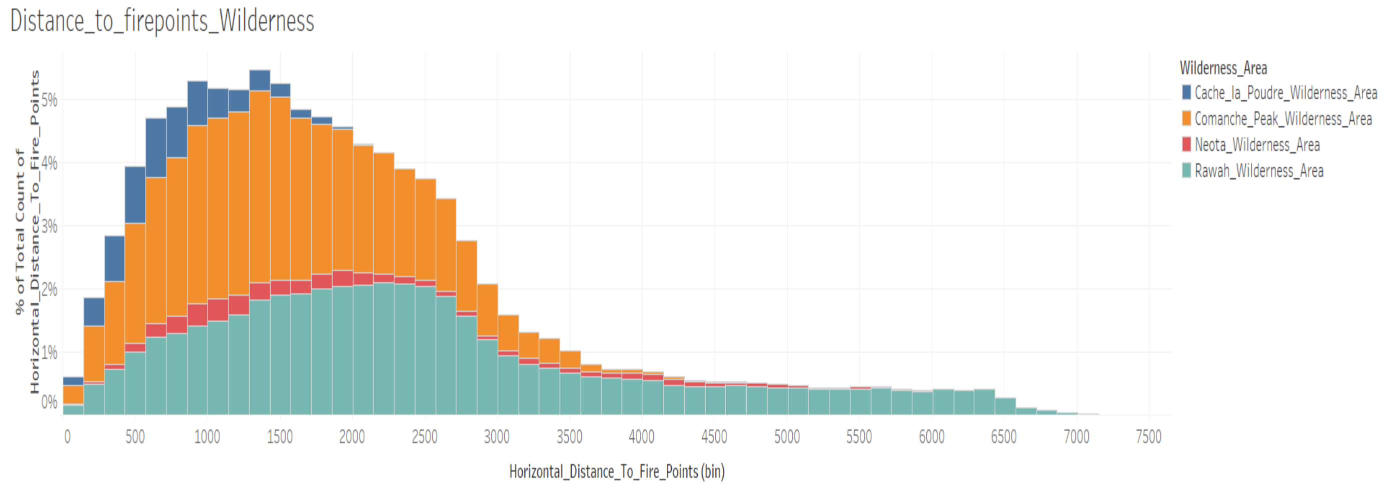
CG_Zone77 alone accounts for 52% of soil across Wilderness areas.

CG_zone47 and CG_zone72 combined accounts for 28% of soil across Wilderness areas

Figure 9: Distance to Firepoints



The trend of % of Total Count of Horizontal_Distance_To_Fire_Points for Horizontal_Distance_To_Fire_Points (bin). Color shows details about Cover_Type. The data is filtered on cgzone, which keeps 11 of 11 members. The view is filtered on % of Total Count of Horizontal_Distance_To_Fire_Points, which keeps non-Null values only.



The trend of % of Total Count of Horizontal_Distance_To_Fire_Points for Horizontal_Distance_To_Fire_Points (bin). Color shows details about Wilderness_Area. The data is filtered on cgzone, which keeps 11 of 11 members.

Description 9:

a.Cottonwood/willow has the least spread(0-2000mts) from the fire points and fire instance may be contained.

Type 1 &2 has the highest spread (0-7000mts) and the fire instances might be difficult to contain.

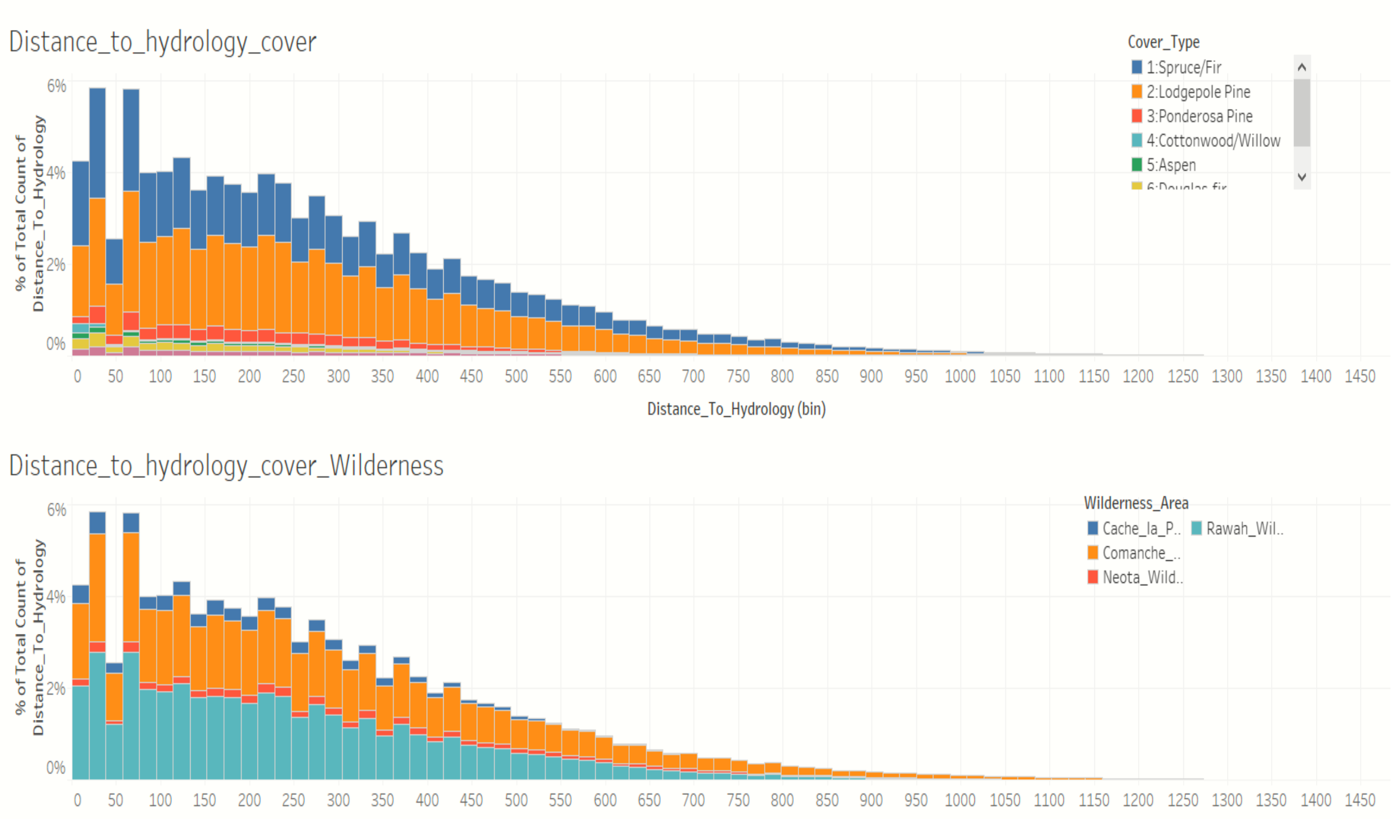
Type 3 & 6 are under 3000mts from fire points.

Type 5 has range (0-6400mts) but has breaks in the spread.

Type 7 extends to 4600mts.

b.Cache_la has least spread and Rawah has the highest spread. Areas with least spread are more prone to fire when Compared to widespread areas.

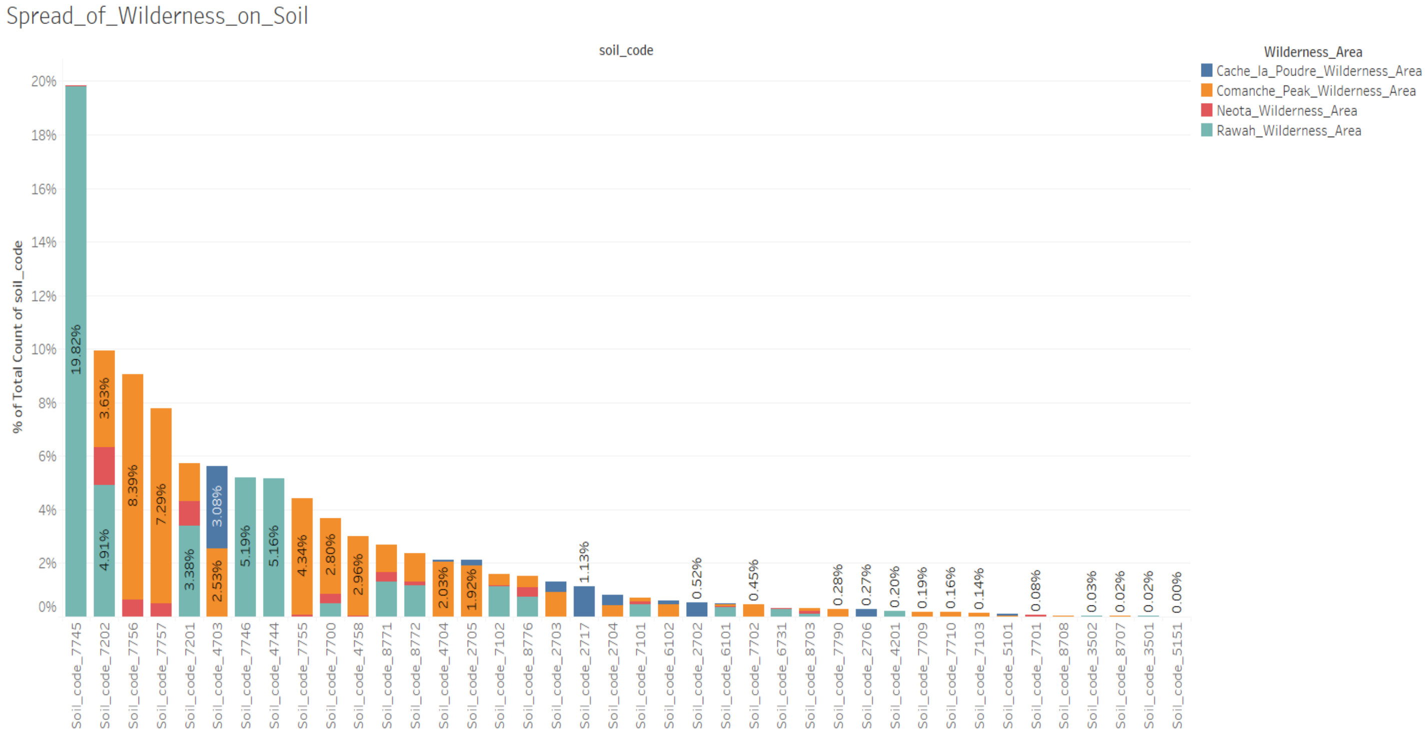
Figure 10: Distance to Hydrology



Description 10:

Cottonwoon/Willow and Douglas-fir seems to be situated near the water(0-600mts), where as the remaining cover types seems to be spread upto 1450mts. Since the scale is only 0-1450mts when compared to firepoints 0-7000mts. we can say that there are multiple hydrology points for these cover types.

Figure 11: Spread of Wilderness on Soil



% of Total Count of soil_code for each soil_code. Color shows details about Wilderness_Area. The marks are labeled by % of Total Count of soil_code. The view is filtered on Wilderness_Area, which keeps Cache_la_Poudre_Wilderness_Area, Comanche_Peak_Wilderness_Area, Neota_Wilderness_Area and Rawah_Wilderness_Area.

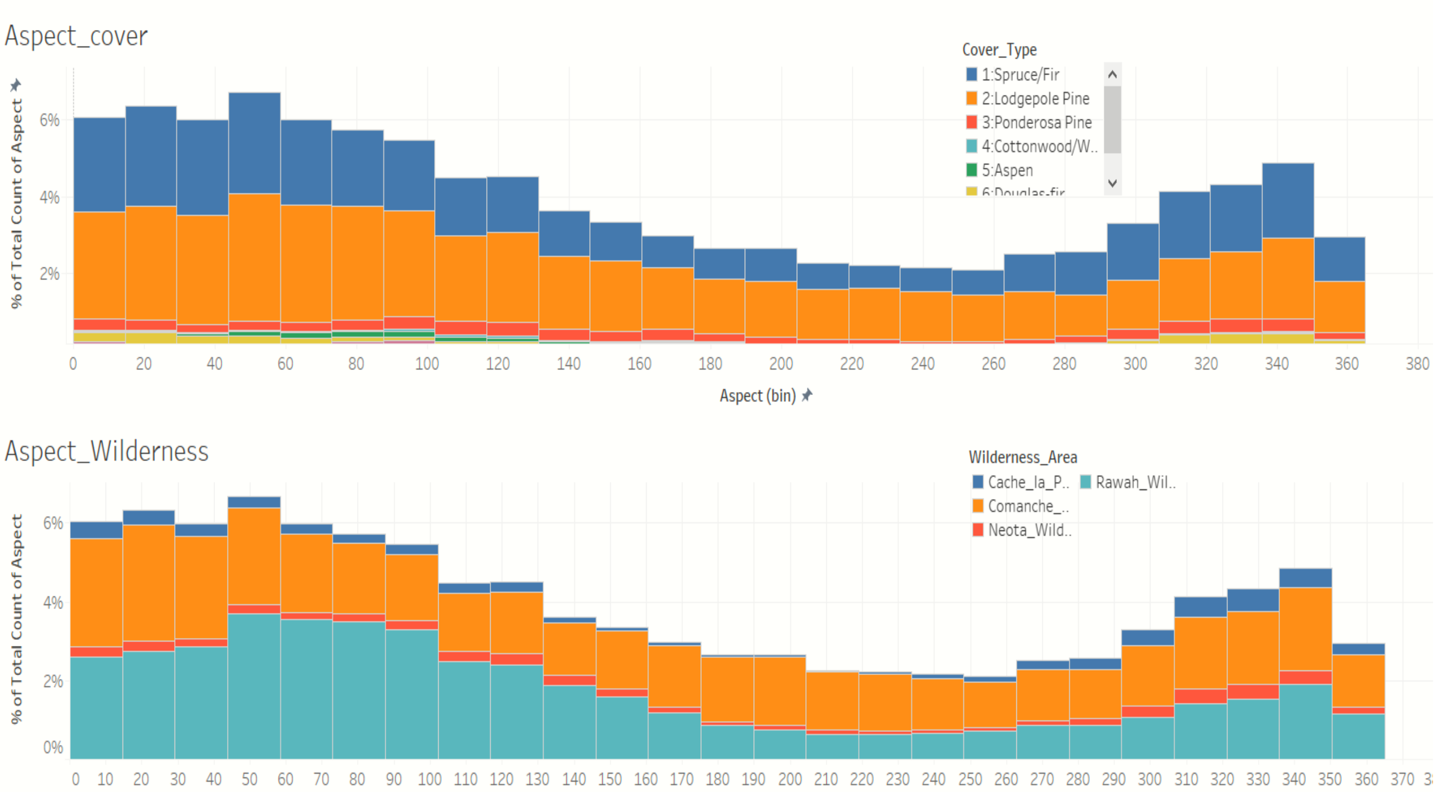
Description 11:

Cache_la area has the least spread and has some of the unique soils such as 5151,2706,2717,2702.

Comanche area has the highest Diversity of soils. Soils 3501, 3502,7746,4744 are unique to Rawah area.

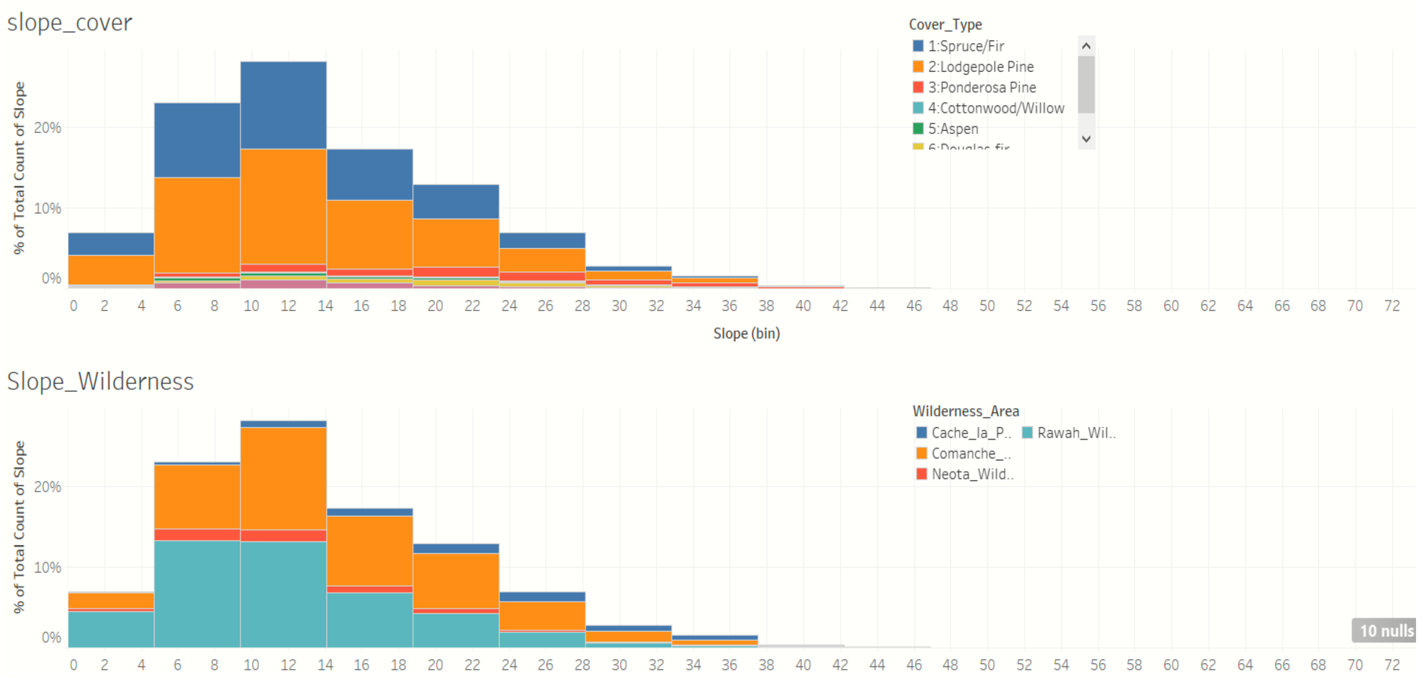
The spread of the remaining soils seems to be uniform across areas.

Figure 12: Aspect



Description 12: We can say that for both covertype and wilderness the Aspect is evenly spread and wilderness area is in form of similar to circle.

Figure 13: Slope



Description 13: We can say that for both cover type and wilderness the Slope is evenly spread between 0-42deg with few outliers. The peek is highest between 10-14 deg.

Chapter 3 - Feature Selection & Model Building

Feature Selection: Feature selection is the process of selecting a subset of relevant attributes to be used in making the model in machine learning. Effective feature selection eliminates redundant variables and keeps only the best subset of predictors in the model which also gives shorter training times. Besides this, it avoids the curse of dimensionality and enhance generalization by reducing overfitting.

Since there was vertical and Horizontal Distance to Hydrology. we have clubbed them both to get the Direct Distance to Hydrology based on Pythagoras theorem (C**2=A**2+B**2).

Also we have taken the Mean of the Hill shade by simple math (a+b+c)/3.

Classification Results:

One of the purposes of this project is to get the prediction for the various forest cover type and to build a model for the future prediction. After looking into the raw data, in order to understand the model building in a better way we are building models from 4 different approaches, and they are:

1. Forest Cover_Soil_Untreated - with outlier
2. Forest Cover_Soil_Treated- without Outliers
3. Forest Cover_CG_Zone_Untreated – With Outliers
4. Forest Cover_CG_Zone_Treated – Without Outliers

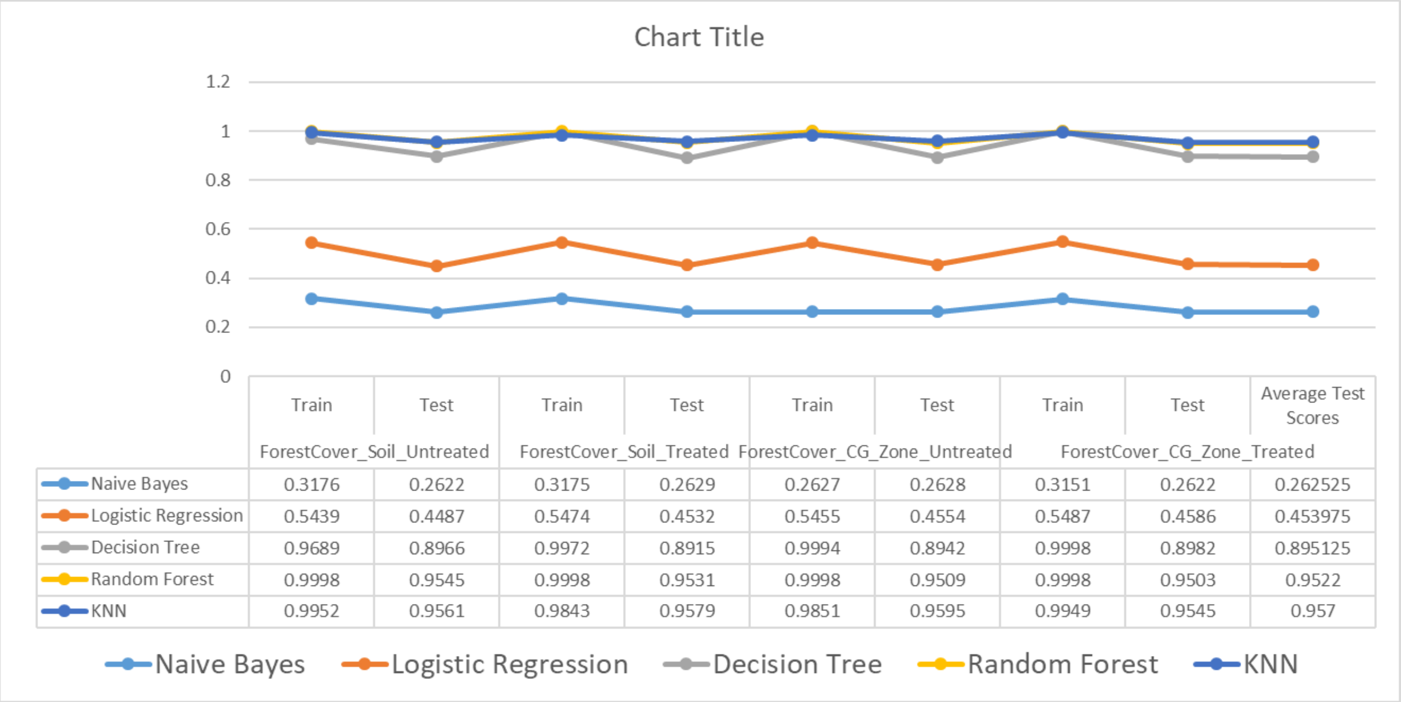
*CG_Zone was extracted from the First 2 Digits of the Soil Code.

We have used the five Models such as Logistic Regression, Naive Bayes, Decision Tree, Random Forest, KNN on all the 4 approaches above.

Steps for model building:

1. As per the approach, Train Test split (70:30) with Stratify as a parameter.
2. SMOTE for treating Imbalance
3. Random SearchCV with min 2 hyper parameters for all the models (Random State = 42).
4. Extracted the best Hyper-tuned model and rerun by removing features by using feature imp fn.
5. Extracted the required features for model and Hyper-parameters.

Models output after Hyperparamater tunning:



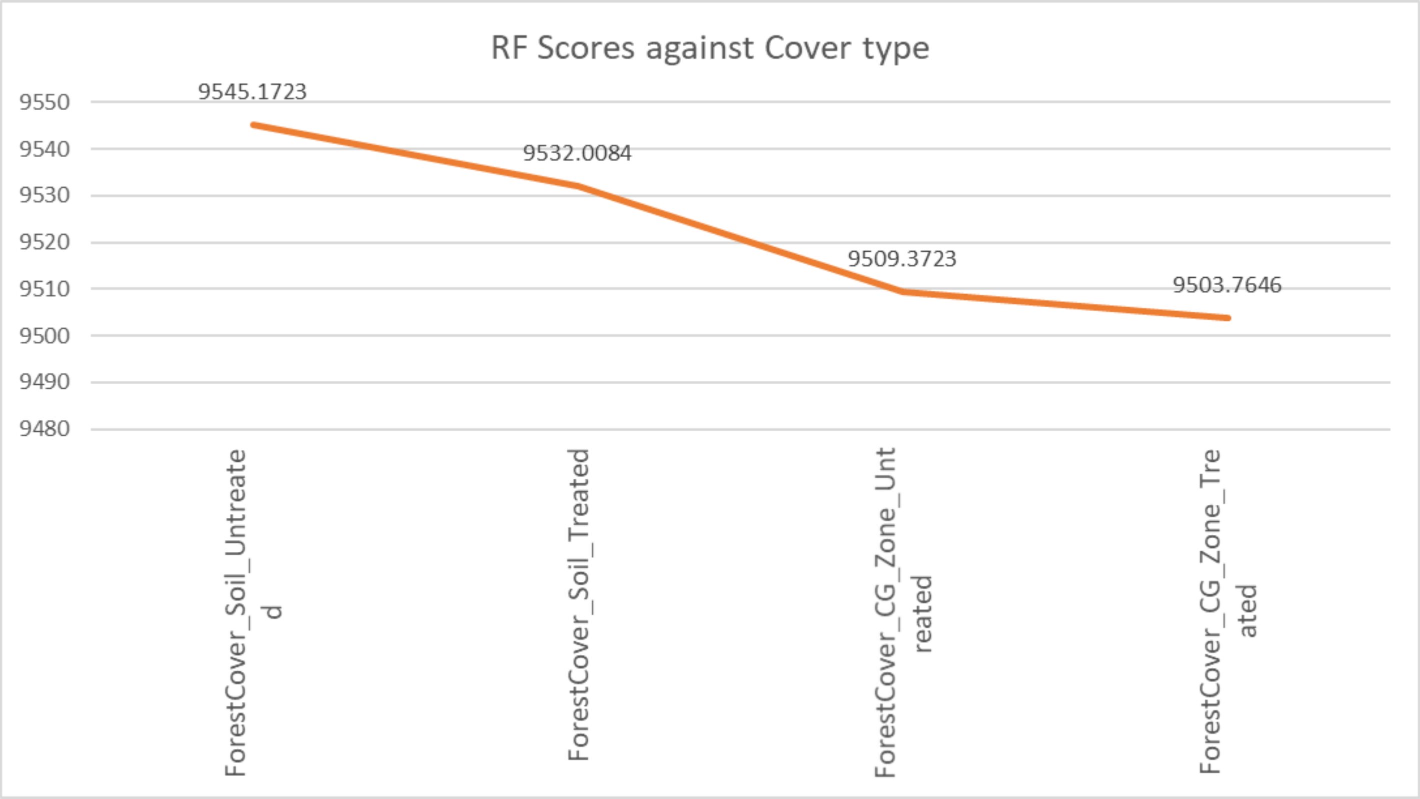
	ForestCover_Soil_Untreated		ForestCover_Soil_Treated		ForestCover_CG_Zone_Untreated		ForestCover_CG_Zone_Treated		
	Train	Test	Train	Test	Train	Test	Train	Test	Average Test Scores
Naive Bayes	0.3176	0.2622	0.3175	0.2629	0.2627	0.2628	0.3151	0.2622	0.262525
Logistic Regression	0.5439	0.4487	0.5474	0.4532	0.5455	0.4554	0.5487	0.4586	0.453975
Decision Tree	0.9689	0.8966	0.9972	0.8915	0.9994	0.8942	0.9998	0.8982	0.895125
Random Forest	0.9998	0.9545	0.9998	0.9531	0.9998	0.9509	0.9998	0.9503	0.9522
KNN	0.9952	0.9561	0.9843	0.9579	0.9851	0.9595	0.9949	0.9545	0.957

		Naive Bayes	Logistic Regression	Decision Tree	Random Forest	KNN
ForestCover_Soil_Untreated	Train	0.3176	0.5439	0.9689	0.9998	0.9952
	Test	0.2622	0.4487	0.8966	0.9545	0.9561
ForestCover_Soil_Treated	Train	0.3175	0.5474	0.9972	0.9998	0.9843
	Test	0.2629	0.4532	0.8915	0.9531	0.9579
ForestCover_CG_Zone_Untreated	Train	0.2627	0.5455	0.9994	0.9998	0.9851
	Test	0.2628	0.4554	0.8942	0.9509	0.9595
ForestCover_CG_Zone_Treated	Train	0.3151	0.5487	0.9998	0.9998	0.9949
	Test	0.2622	0.4586	0.8982	0.9503	0.9545
	Average Test Scores	0.262525	0.453975	0.895125	0.9522	0.957

From the above table and plot we can see that Random Forest have yielded the best results of 0.95% Recall. Now we have decided to use Random Forest. Let's Check which approach is best.

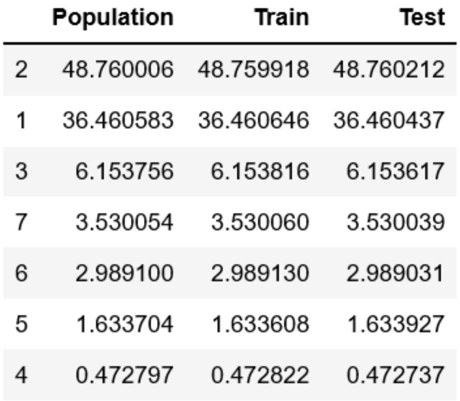
Below is the Scores Obtained for each random Forest Model. We can see that outliers not treated and soil code approach gives the best recall.

ForestCover_Soil_Untreated	ForestCover_Soil_Treated	ForestCover_CG_Zone_Untreated	ForestCover_CG_Zone_Treated
9545.1723	9532.0084	9509.3723	9503.7646



Inference:

The Final Model is Random Forest with Stratified Train test Split and used SMOTE for Imbalance Treatment.



	Population	Train	Test
2	48.760006	48.759918	48.760212
1	36.460583	36.460646	36.460437
3	6.153756	6.153816	6.153617
7	3.530054	3.530060	3.530039
6	2.989100	2.989130	2.989031
5	1.633704	1.633608	1.633927
4	0.472797	0.472822	0.472737

Parameters are: criterion='entropy',n_estimators=50,random_state=42, class_weight='balanced', max_depth=25

The Required Features are 'Elevation', 'Aspect', 'Horizontal_Distance_To_Roadways',

'Horizontal_Distance_To_Fire_Points', 'Distance_To_Hydrology'

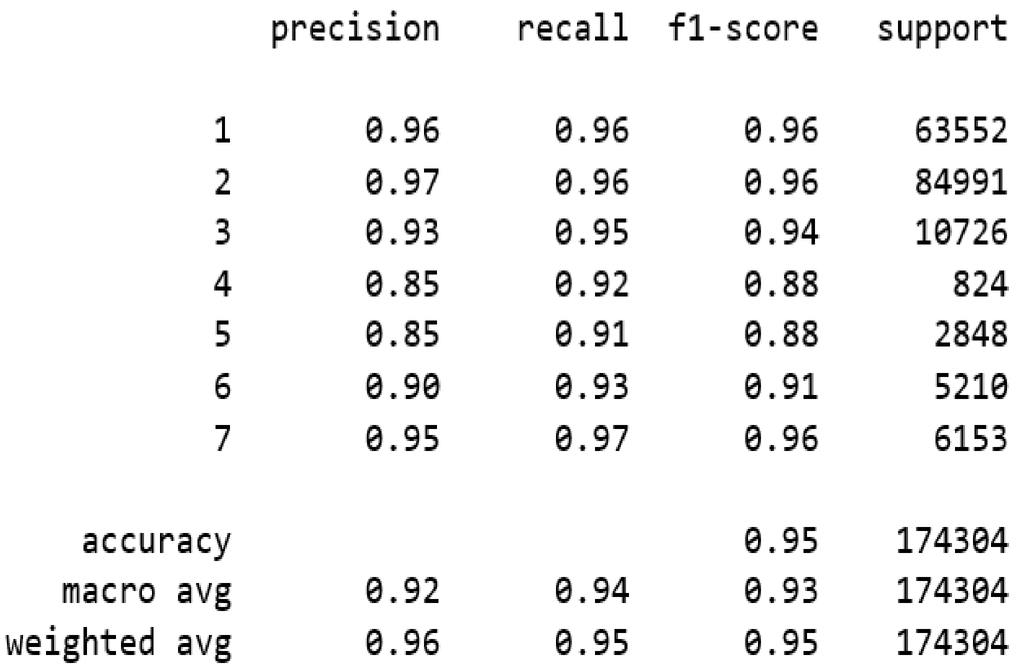
Metrics are:

Train Recall 0.9998364753596461

Test Recall 0.9550670093629521

The f1_weighted of the Train is: 0.9998364581975809

The f1_weighted of the Test is: 0.9551662273823236

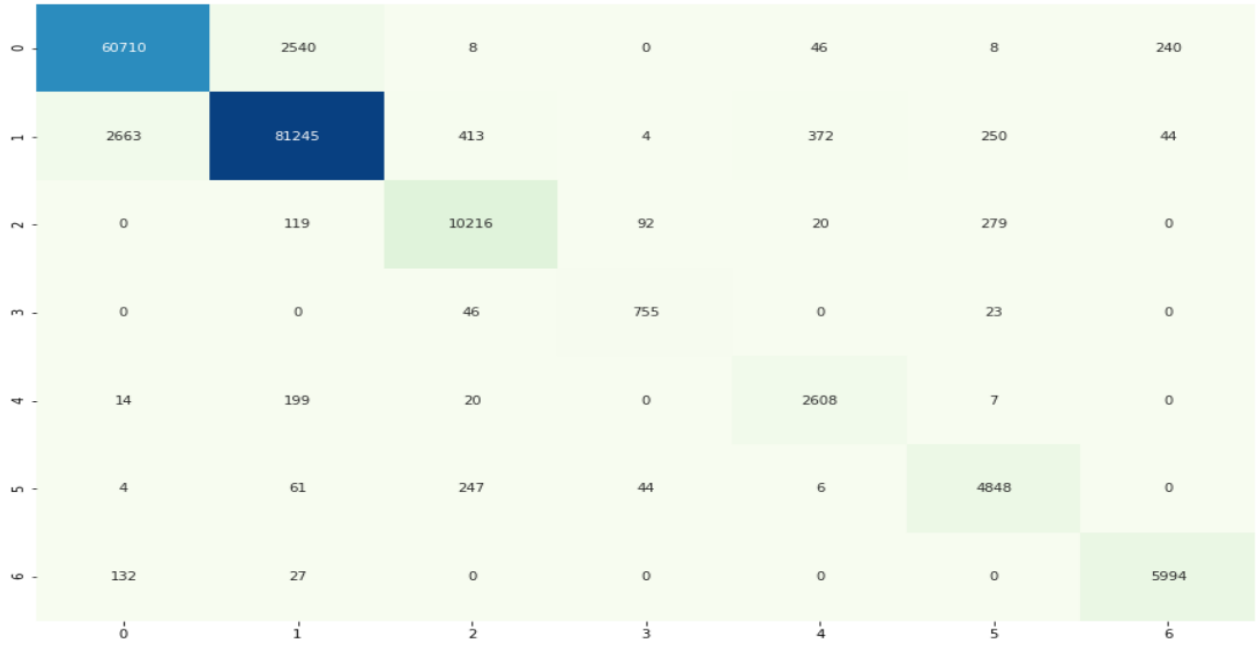


```
              precision    recall  f1-score   support

           1       0.96      0.96      0.96     63552
           2       0.97      0.96      0.96     84991
           3       0.93      0.95      0.94     10726
           4       0.85      0.92      0.88       824
           5       0.85      0.91      0.88      2848
           6       0.90      0.93      0.91      5210
           7       0.95      0.97      0.96      6153

    accuracy                           0.95    174304
   macro avg       0.92      0.94      0.93    174304
weighted avg       0.96      0.95      0.95    174304
```

The Confusion matrix is as below.

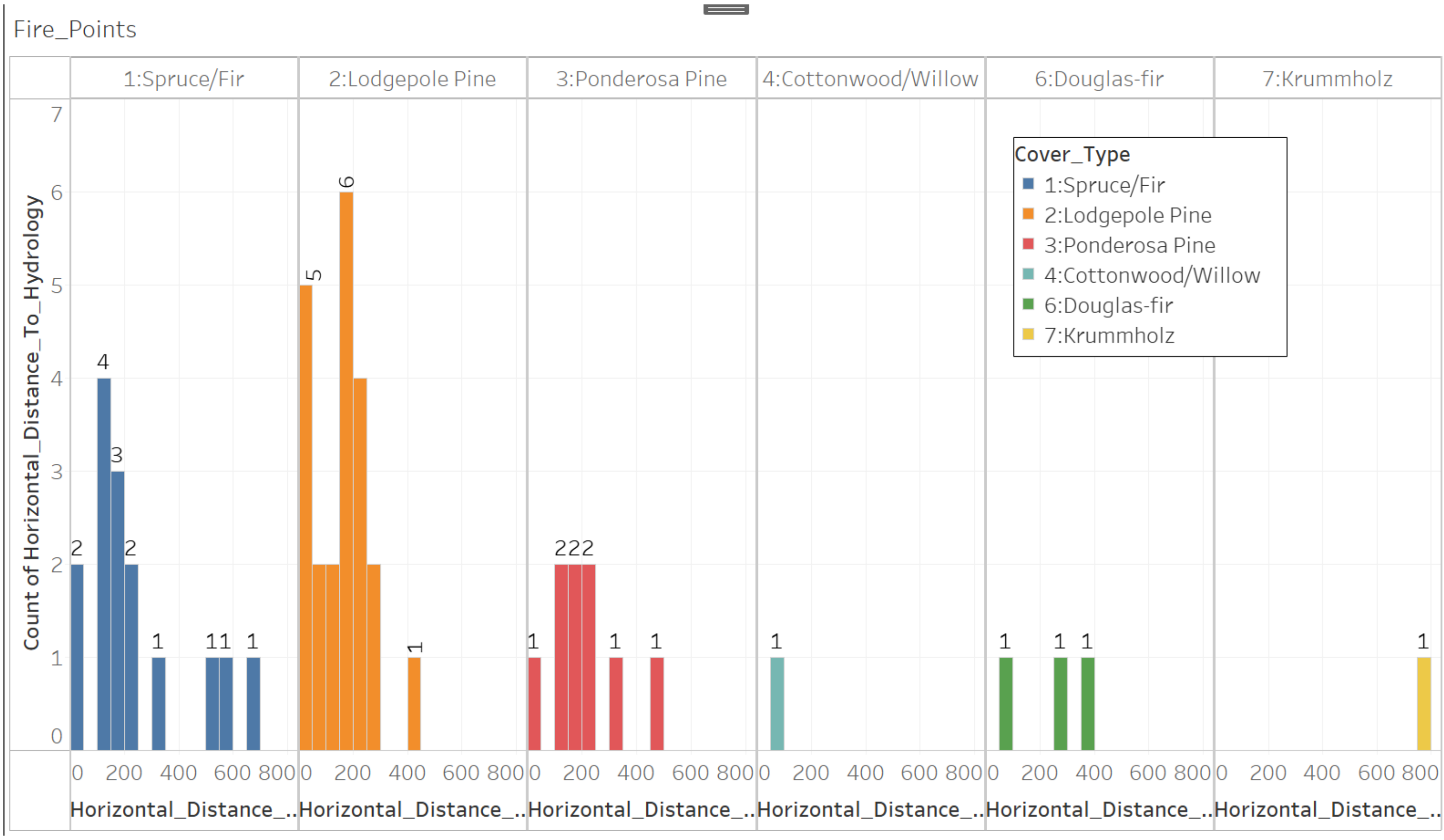


	0	1	2	3	4	5	6
0	60710	2540	8	0	46	8	240
1	2663	81245	413	4	372	250	44
2	0	119	10216	92	20	279	0
3	0	0	46	755	0	23	0
4	14	199	20	0	2608	7	0
5	4	61	247	44	6	4848	0
6	132	27	0	0	0	0	5994

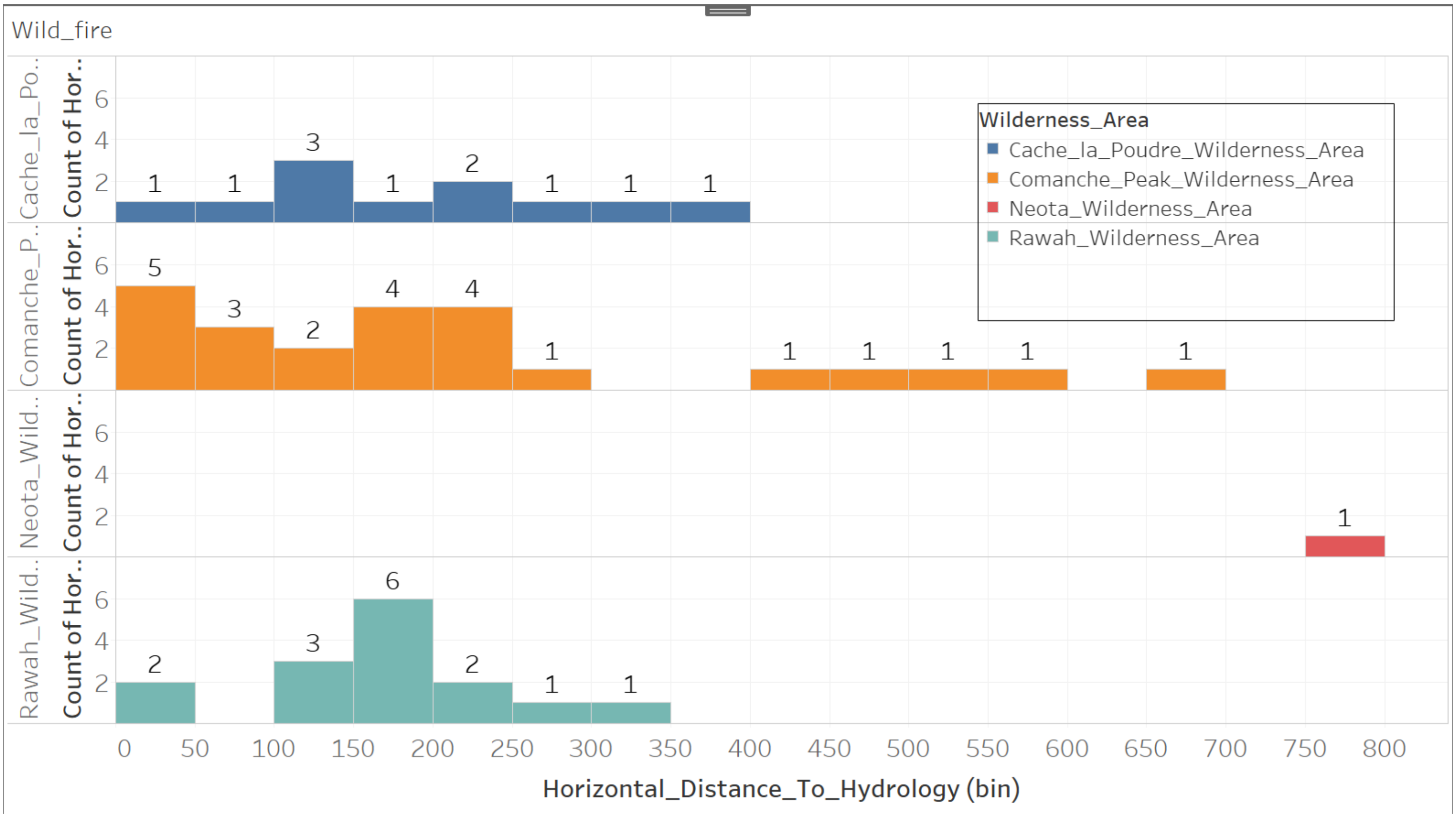
Chapter 4 – Recommendations

Since we have achieved the required business target to increase the efficiency from 70.58 and have achieved 95 %. We have some recommendations from EDA part as shown below

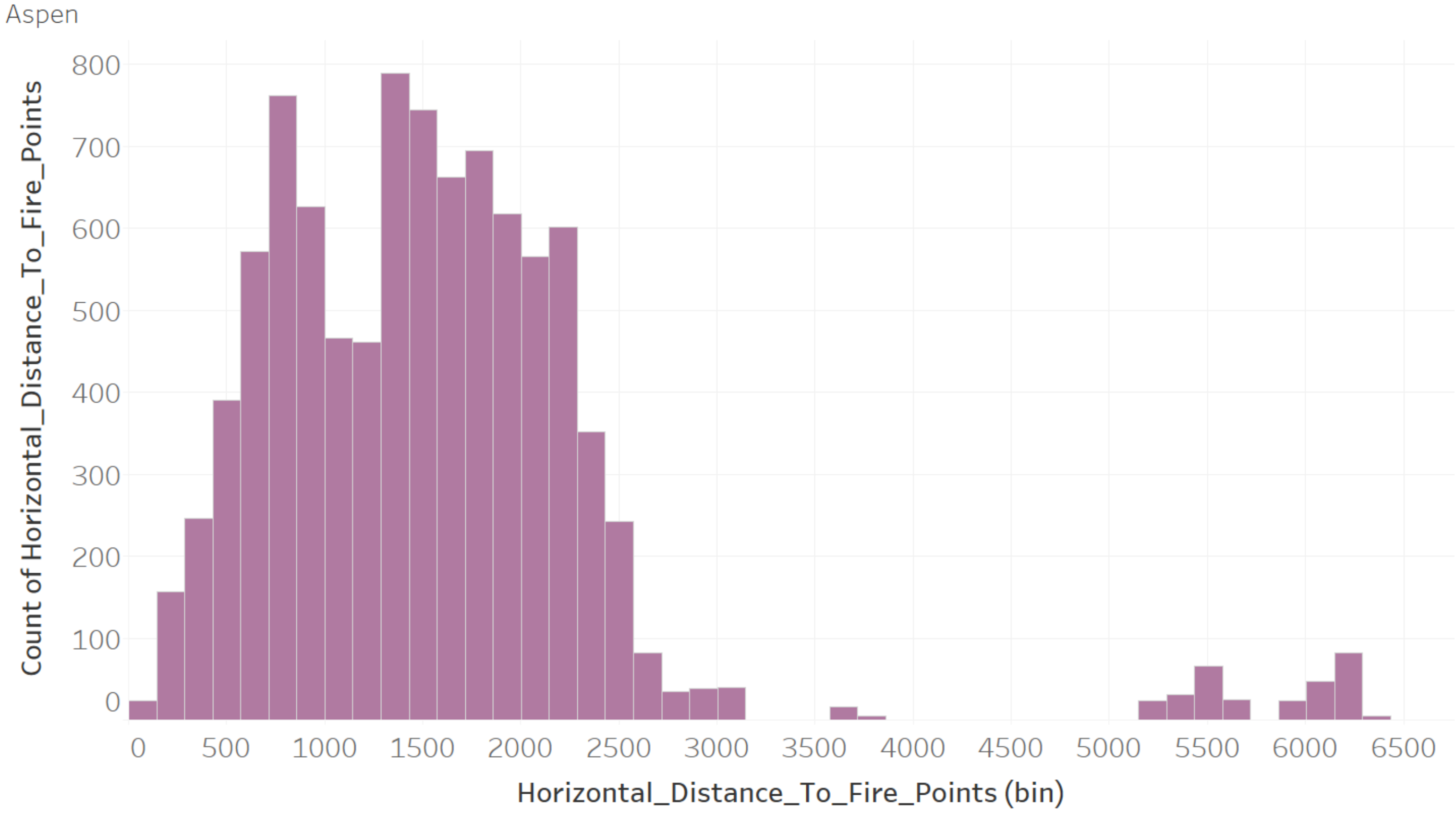
- Only these 5 features are enough to predict the cover type using our model the features are 'Elevation', 'Aspect', 'Horizontal_Distance_To_Roadways', Horizontal_Distance_To_Fire_Points', 'Distance_To_Hydrology'.
- There are about 9 Fire points greater than 300 meters from Hydrology across cover types and Krummholz is the farthest at 800 mts.



- Same as above 9 Fire points which 300mts away from hydrology and the highest 800mts is located in Neota wilderness area.



- There are gaps in Distance to Fire_points between 3900 and 5000 mts in Aspen, which has to be investigated.



Chapter 5 – Conclusion & Further Study

We took up this project as an extension to the existing prediction study to improve the prediction efficiency from 70.58 in neural network. We have done extensive study on features and have created 2 synergy features once just by using the mean of the variables **Hill shade** and the other one by using the Pythagoras theorem to generate the direct **distance to Hydrology**. We have also created CG_zones by extracting the first 2 digits from the Soil ELU Code and considered it to be an separate approach. We have considered 4 approached to the problem,

- Forest_Cover_Soil_(Outliers Untreated)
- Forest_Cover_Soil_(Outliers Treated)
- Forest_Cover_CG_Zone_(Outliers Untreated)
- Forest_Cover_CG_Zone_(Outliers Treated)

Post which we have used the train test split (70:30) with stratify as hyper parameter then have used SMOTE to treat the Imbalance in the Cover Type. Then we have used RandomSearchCV though Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, KNN with min of 2 hyper parameters.

We have selected the Best of all the 20 models build and tuned, Which came to be Random Forest and the best approach was Forest_Cover_Soil_(Outliers Untreated).

The Final Features used was ['Elevation', 'Aspect', 'Horizontal_Distance_To_Roadways',

'Horizontal_Distance_To_Fire_Points', 'Distance_To_Hydrology']

The Final Parameters for RF are: [criterion='entropy',n_estimators=50,random_state=42, class_weight='balanced', max_depth=25]

We still have to investigate the Aspen cover type Missing values in Distance to Fire points and many more insights can be obtained in the Future Study.

Glossary

Azimuth	An azimuth is an angular measurement in a spherical coordinate system. The vector from an observer to a point of interest is projected perpendicularly onto a reference plane; the angle between the projected vector and a reference vector on the reference plane is called the azimuth. link
Rock outcrop	The part of a rock formation that appears above the surface of the surrounding land. Outcrop, outcropping. Belay - something to which a mountain climber's rope will be secured. Outthrust - an outcropping of rock that extends outward.
Cathedral soil	The Cathedral soil consists of very shallow or shallow, well drained or somewhat excessively drained soils that formed in slope alluvium and colluvium from granite, monzonite, schist, or gneiss. These soils are on mountain slopes and hills with slopes of 2 to 100 percent.
Vanet soil	The Vanet soil consists of shallow, well-drained, moderately permeable soils on side slopes of mesas. They formed in colluvium and residuum weathered from sedimentary rocks. Slope ranges from 20 to 40 percent. The mean annual temperature is 44 degrees F. and the mean annual precipitation is 19 inches. link
Ratake soil	The Ratake soil consists of shallow, well drained soils that formed in materials weathered from igneous and metamorphic rocks. Ratake soils are on upland hills and ridges and have slopes of 2 to 60 percent. The mean annual precipitation is about 16 inches and the mean annual temperature is about 45 degrees F. link
Wetmore soil	The Wetmore soil consists of very shallow and shallow, well drained, rapidly permeable soils formed in thin gravelly course textured no calcareous materials weathered from granite. Wetmore soils are on mountain slopes and have slopes of 5 to 80 percent. Mean annual precipitation is about 23 inches and mean annual temperature is about 43 degrees F.link
Gothic soil	The Gothic soil consist of very deep, well drained soils formed in slope alluvium, colluvium or slide deposits. These soils are on mountain slopes, fan remnants, dip slopes and landslides. Slopes are 2 to 60 percent. Mean annual precipitation is about 585 mm, and mean annual air temperature is about 3.3 degrees C. link
Limber soil	Limber soils have light colored A2 horizons, B2t horizons having subangular blocky structure, and continuous Cca horizons. link
Troutville soil	The Troutville soil consists of deep, well drained soils that formed in thick, stony, moderately coarse textured, transported material derived from mixed rock sources. Troutville soils are on alluvial fans, terraces, glacial moraines, eskers, till plains, outwash plains, and mountain sides. Slopes range from 2 to 60 percent. The mean annual precipitation is about 18 inches. The mean annual temperature is about 34 degrees F. link
Bullwark soil	The Bullwark soil consists of moderately deep, well drained soils that formed in colluvium and residuum from schist, gneiss, and granite. Bullwark soils are on mountain slopes. Slopes range from 5 to 50 percent. The mean annual precipitation is about 21 inches and the mean annual temperature is about 40 degrees F.link
Catamount soil	The Catamount soil consists of shallow, excessively or somewhat excessively drained soils that formed in slope alluvium over residuum from granitic rocks, gneiss, and schist. Catamount soils are on mountain side slopes and ridges and have slopes of 5 to 70 percent. The mean annual precipitation is about 21 inches and the mean annual temperature is about 42 degrees F. link
Legault soil	The Legault soil consists of very shallow and shallow, well drained or somewhat excessively drained soils that formed in slope alluvium, colluvium, and residuum from schist, gneiss, and granitic rocks. These soils are on mountain slopes, ridges, structural benches, and spurs that commonly have north aspects. Slopes range from 5 to 80 percent. Average annual precipitation is about 20 inches and mean annual temperature is about 43 degrees F. link
Gateview soil	The Gateview soil consists of very deep, well to somewhat excessively drained soils that formed in alluvial or colluvial material from sedimentary rocks and glacial materials of mixed origin. Gateview soils are on outwash terraces, toeslopes, alluvial fans, till plains, kames, esters, mountains, and low ridges. Slopes are 2 to 40 percent. The mean annual precipitation is about 20 inches and the mean annual temperature is about 36 degrees F.link
Rogert soil	The Rogert soil consists of very shallow and shallow , well drained soils formed in thin, noncalcareous, very gravelly or channery materials weathered residually from granite, sandstone, gneiss or in places from tuff. Rogert soils are on mountain slopes and ridges. Slopes

	are 3 to about 100 percent. The mean annual precipitation is about 18 inches and the mean annual temperature is about 34 degrees F.link
Leighcan soil	The Leighcan soil consists of very deep, well drained soils that formed in till, slope alluvium, or colluvium from acid igneous rocks. Leighcan soils are on mountain slopes and have slopes of 0 to 70 percent. The mean annual precipitation is about 45 inches, and the mean annual temperature is about 32 degrees F. link
Granile soil	The Granile soil consists of very deep, well drained soils on mountain slopes and hills. These soils formed in slope alluvium and colluvium derived principally from gneiss, schist, and granitic rocks. Slopes range from 2 to 60 percent. The mean annual precipitation is about 16 inches and the mean annual air temperature is about 36 degrees F.link
Legault soil	The Legault soil consists of very shallow and shallow, well drained or somewhat excessively drained soils that formed in slope alluvium, colluvium, and residuum from schist, gneiss, and granitic rocks. These soils are on mountain slopes, ridges, structural benches, and spurs that commonly have north aspects. Slopes range from 5 to 80 percent. Average annual precipitation is about 20 inches and mean annual temperature is about 43 degrees F.link
Como soil	The Como soil consists of very deep, somewhat excessively drained soils that formed in colluvium derived from coarse grained igneous and sedimentary rock. These soils occur on mountain slopes. Slopes are 10 to 60 percent. Mean annual precipitation is about 760 mm, and the mean annual air temperature is about 3 degrees C.link
Cryomont soil	The Cryomont soil consists of very deep, somewhat excessively drained soil formed in alluvium. Cryomont soils are on low terraces and fans. Slopes are 0 to 30 percent. The mean annual precipitation is about 55 inches and the mean annual temperature is about 41 degrees F.link
	Loamy-skeletal, mixed, source

Forest Cover Type Classes:

1 -- Spruce/Fir	36.46%	https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb5436769.pdf
2 -- Lodgepole Pine	48.76%	https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/fsbdev7_002604.pdf
3 -- Ponderosa Pine	6.15%	https://plants.usda.gov/plantguide/pdf/pg_pipo.pdf
4 -- Cottonwood/Willow	0.47%	https://plants.usda.gov/plantguide/pdf/cs_poan3.pdf
5 -- Aspen	1.63%	https://www.fs.fed.us/wildflowers/beauty/aspen/ecology.shtml
6 -- Douglas-fir	2.99%	https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/fsbdev7_002606.pdf
7 -- Krummholz	3.53%	https://en.wikipedia.org/wiki/Krummholz