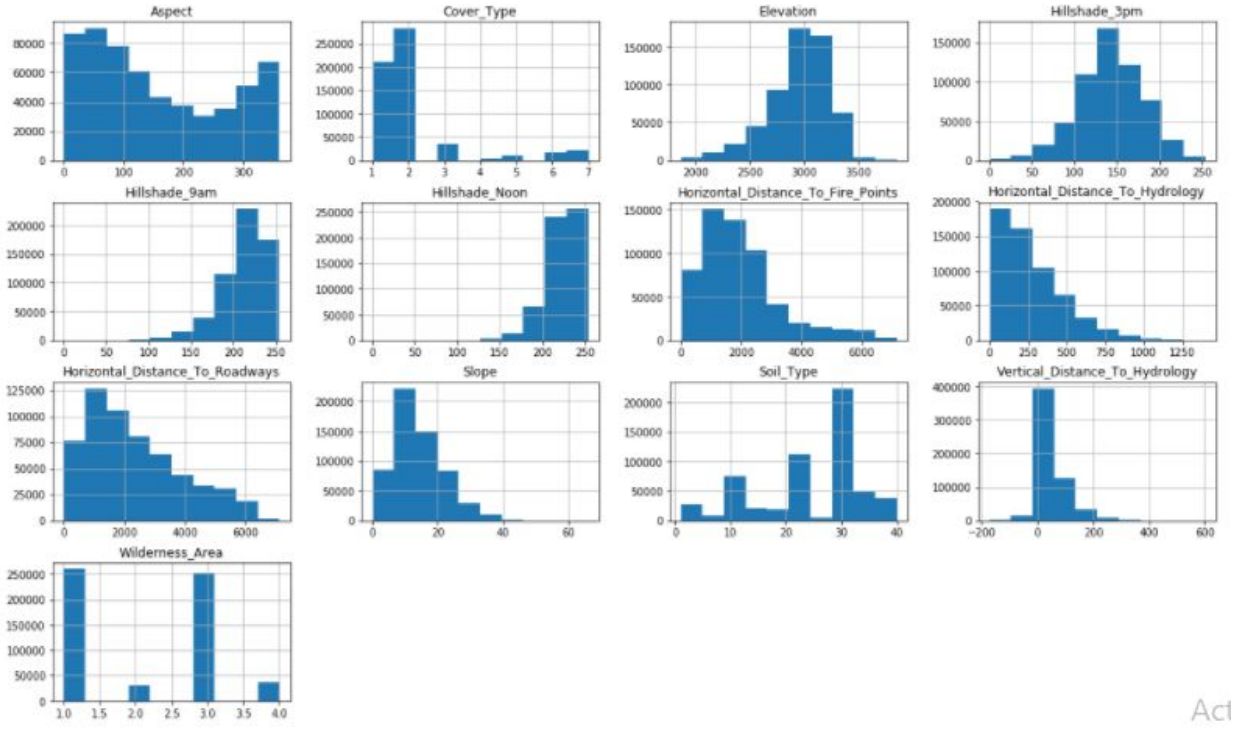
Phase-1

Preprocessing Techniques

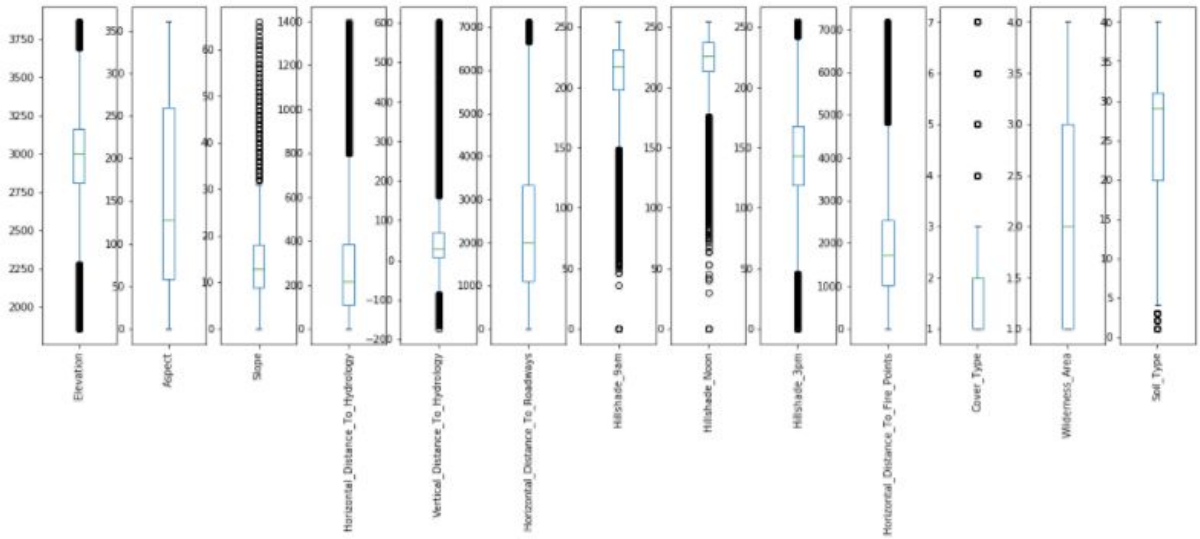
1. **Binary columns to categorical:** There were 4 binary columns for the Wilderness_Area and 40 binary columns for the Soil_Type in our data.We used idmax function of pandas to merge the Widerness_Area and Soil_Type columns separately. This resulted in reducing the number of columns in our data from 54 to 13 columns. This reduced the size of our data.
2. **Missing Values:** There were no missing values in the data.
3. **Datatypes:** Allocating space ahead of time allows computers to optimize storage and processing efficiency hence we analyzed the datatypes of all the attributes in our dataset and changed them according to the requirements and processing of data.

Plots:

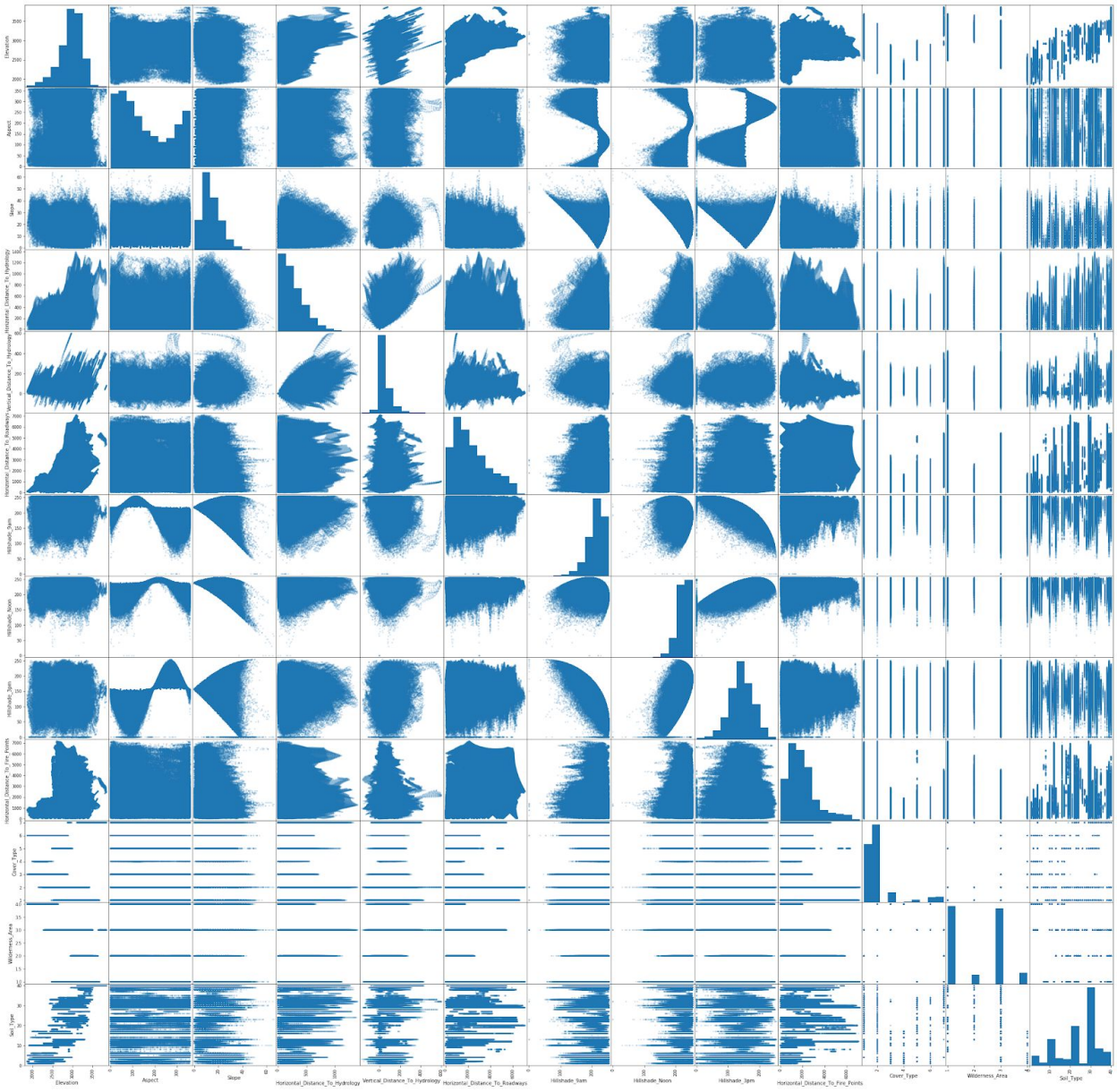
- **Histogram:** We plotted the histograms to observe the distributions of data.



- **Box plot:** Graphic display of five-number summary



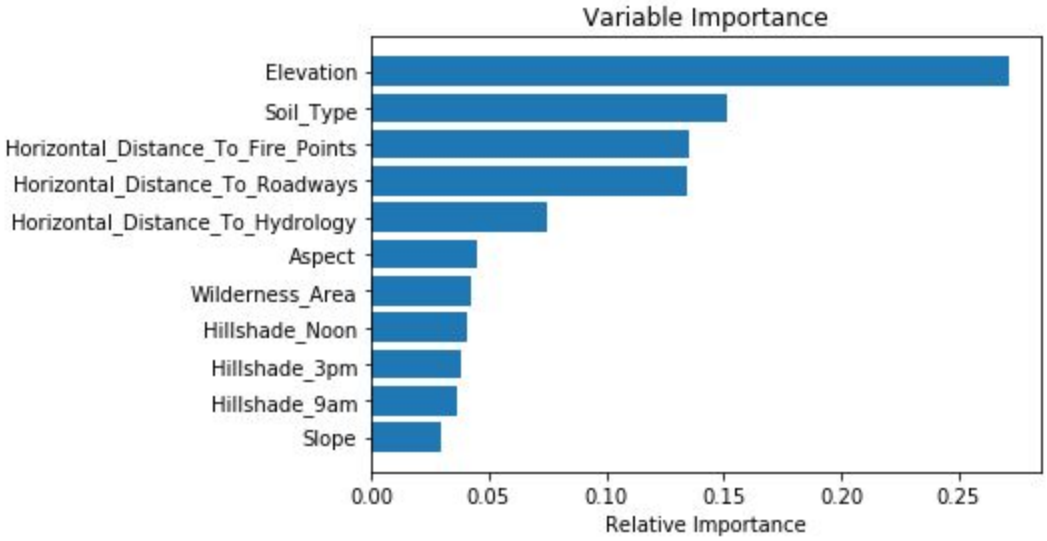
- **Scatter:**


Elevation
Aspect
Slope
Horizontal_Distance_To_Hydrology
Vertical_Distance_To_Hydrology
Horizontal_Distance_To_Roadways
Hillshade_9am
Hillshade_Noon
Hillshade_3pm
Horizontal_Distance_To_Fire_Points
Cover_Type
Wilderness_Area
Soil_Type

Feature Engineering

Identifying important features:

We used Random Forest to generate a variety of decision trees to help us determine which variables are the best at predicting the cover type. It helps us in finding the best features of our data set.The following result shows that the best variable is Elevation while the worst is Slope to predict Cover Type. We can now determine which variables will be the most important for phase 2. This will further help us in identifying that how many variables we should include in our model.



Generating new features

We modified our data so that now it represents the original features along with their interaction features. The top 3 important features in our data were Elevation, Soil_Type and Horizontal_Distance_To_Fire_Points so we generated new features based on their interactions.A total of three new columns were added to our data.

We then later calculated the correlation of the new features with Cover_Type which is the target variable. The correlation of the new features with the Cover_Type did not increase or change significantly.

Correlation Analysis

Categorical Variables:
The chi-square test was used to determine whether an association between 2 categorical variables in a sample is likely to reflect a real association between these 2 variables in the population.

Our data consisted of 2 categorical variables only. The following are the test results :

Variable-1: Wilderness_Area
Variable-2: Soil_Type

Chi-square value = 819289.809223
Degrees of freedom = 117
P-value = 0.0
Significance level = 0.05

Null Hypothesis: There is no association between the Wilderness_Area and the Soil Type.
Alternative Hypothesis: There is an association between the Wilderness_Area and the Soil Type.

A small p-value indicates that the observations are inconsistent with the Null hypothesis.
The P stands for probability and measures how likely it is that any observed difference between groups is due to chance.
The p-value is less than our significance level hence we conclude that both the variables are dependent and the relationship between both the variables is statistically significant since a low p-value suggests that our sample provides enough evidence that we can reject the null hypothesis for the entire population. We reject the Null Hypothesis. The high chi-squared value supports this observation.

Quantitative Variables:

Both the Spearman and Pearson's Correlation Coefficient were used for investigating the relationship between two quantitative variables. The correlation coefficient (r) is a measure of the strength of the association between the two variables.

Spearman correlation coefficient is a non-parametric test while pearson correlation coefficient is a parametric test. Parametric tests assume underlying statistical distributions in the data whereas

non-parametric do not consider distributions of data. Therefore, several conditions of validity must be met so that the result of a parametric test is reliable such as both the variables should be normally distributed. The distribution of most of the variables in our data was skewed however there are few variables which are normally distributed so we used both tests to conduct our hypothesis.

The following are the correlation between the significant variables in our dataset which we found through randomforest:

Variable-1: Horizontal_Distance_To_Fire_Points
Variable-2: Horizontal_Distance_To_Roadways

Correlation coefficient = 0.322
P-value = 0.0

Analysis: Positively but weakly correlated.

Variable-1: Elevation
Variable-2: Horizontal_Distance_To_Fire_Points

Correlation coefficient = 0.1548
P-value = 0.0
Analysis: Positively but very weakly correlated.

Variable-1: Elevation
Variable-2: Horizontal_Distance_To_Roadways

Correlation coefficient = -0.160
P-value = 0.0

Analysis: Negatively but very weakly correlated.

The following is the only strong correlation found between the feature and the target variable (Cover_Type)

Variable-1: Elevation
Variable-2: Cover_Type

Correlation coefficient = -0.491

P-value = 0.0
Analysis: Negatively correlated.This correlation is statistically significant.

The following are the significant correlations in our dataset :

Variable-1: Hillshade_Noon
Variable-2: Hillshade_9am

Correlation coefficient = -0.82
P-value = 0.0

Variable-1: Aspect
Variable-2: Hillshade_3pm

Correlation coefficient = 0.64
P-value = 0.0

Variable-1: Hillshade_3pm
Variable-2: Hillshade_Noon

Correlation coefficient =0.57
P-value = 0.0

The above correlations are statistically significant because they all have a p value of 0 and they are highly correlated as illustrated by their correlation coefficient value.

Log Transformation of skewed data

We used log transformation on those variables of data whose distribution was skewed.Log transforms are useful when applied to skewed distributions as they tend to expand the values which fall in the range of lower magnitudes and tend to compress or reduce the values which fall in the range of higher magnitudes.This tends to make the skewed distribution as normal-like as possible. We visualized the distributions of our data and after applying the transform we observed that the skewed data distributions were more normal-like as compared to the skewed distribution on the original data. The columns transformed included Slope, Hillshade_9am, Hillshade_noon, Horizontal_Distance_To_Fire_Points and Horizontal_Distance_To_Hydrology.

Adaptive Binning

The technique of fixed-width binning wasn't used since it manually decides the bin ranges due to which the data can end up having irregular bins which are not uniform based on the number of data points or values which fall in each bin.Some of the bins might be densely populated and some of them might be sparsely populated or empty.Adaptive binning is a better strategy since we use the data distribution itself to decide our bin ranges.

We used quantile based binning on our data columns. 10 quantiles were calculated for the continuous columns of the data and then they were used to divide the data into their respective bins.We used histograms to visualize the data after transformation.

The transformed columns included Slope, Elevation , Horizontal_Distance_To_Hydrology, Vertical_Distance_To_Hydrology,Horizontal_Distance_To_Roadways , Horizontal_Distance_To_Fire_Points and Aspect.