Forest Cover Type EDA

We preliminary explored the data to uncover patterns and test hypotheses

Topic	Finding
Data Background	Each data observation represents a 30m x 30m patch of land in Roosevelt National Forest of northern Colorado. Our model will predict a classification for the forest cover type. There are 7 forest cover types.
Data Shape	• 581,012 Rows • 55 Columns • 0 Null Values
Column Types	• 10 variables are Continuous • 44 variables are Binary
Test Variable Distribution	• Class imbalance is present • Lodgepole Pine: accounts for 49% of data observations • Spruce/Fir: 36% • Ponderosa Pine: 6% • Douglas Fir, Aspen, Cottonwood/Willow: 5%
Data Challenges	• 40 different types of Soil variables (all binary) • 4 different types of Wilderness variables (all binary)
Variables Correlation	• Elevation, Aspect, Horizontal Distance to Fires, and Horizontal Distance to Roads variables have high correlation with Forest Cover Types variable • The 4 Wilderness variables have low correlation with Forest Cover Types variable