TeamCCRA: Chenyue Lu, Linda Nieman

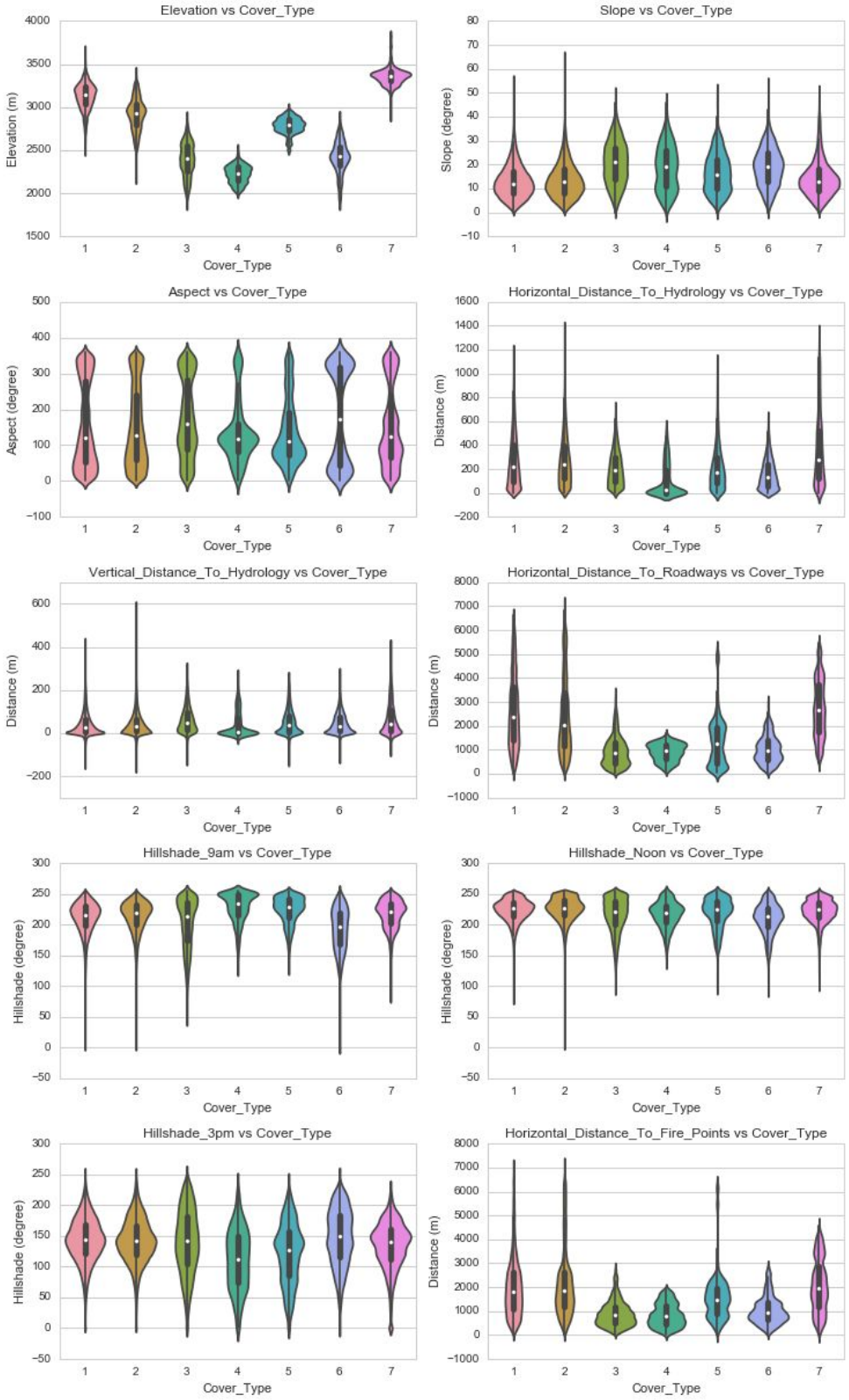

Figure 1. Violinplots of quantitative variables vs Cover_Type

For the **quantitative variables**, we generated violinplots using seaborn, shown in Figure 1 above. From the violinplots, we can see that across different types of tree covers, elevation seems to be distinct. In other words, elevation may be a good determining variable to predict type of tree covers.

Examples: (feel free to skip the examples if we made ourselves clear)

From the Aspect vs Cover_Type plot, we see that each cover type also has distinct aspect profile. For example, tree cover type 4 has a cluster around aspect = 120 degrees, whereas tree cover type 6 are more prevalent when aspect is near either 0 or 360 degrees. These violinplots also show us the spread of the data. For example, from the Horizontal_Distance_To_Roadways vs Cover_Type and Horizontal_Distance_To_Fire_Points vs Cover_Type plots, we see that tree cover types 3 and 4 tend to be cluster near roadways and fire points. Please note that these violinplots are scaled by area for better visualization, because the sample sizes for cover type 1 and 2 are one order of magnitude greater than those of cover type 3 through 7.

For the **qualitative variables**, namely Wilderness_Area and Soil_Type, we took different approaches analyzing their correlation with the type of tree covers.

The two stacked bar graphs below visualize the correlation between Wilderness_Area and Cover_Type. We counted the number of occurrences in the dataset that correspond to the separate Wilderness_Area and Cover_Type.

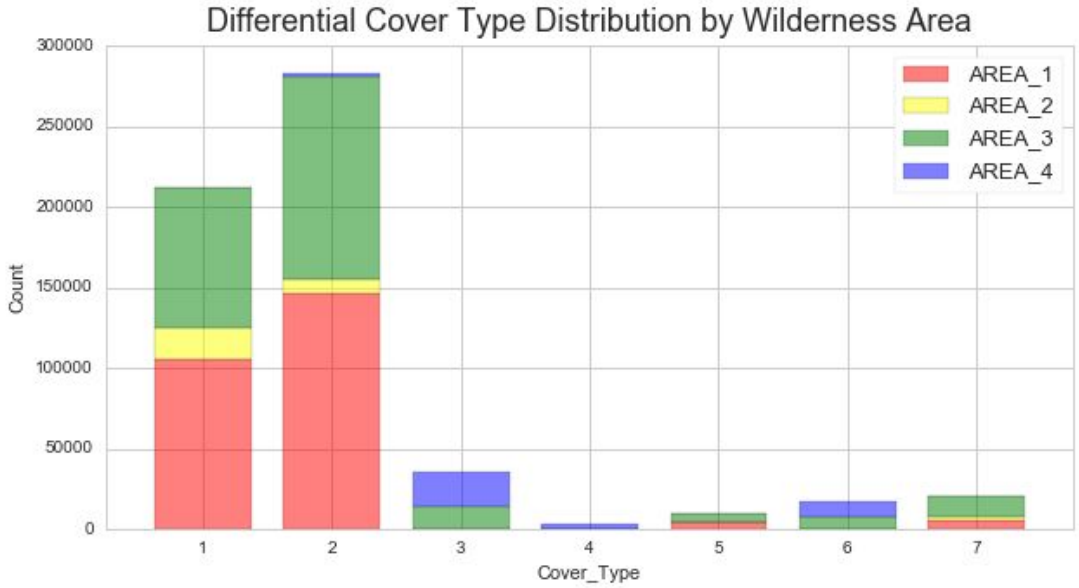

Figure 2. Differential Cover Type Distribution by Wilderness Area

Examples:

First, we can see that we have more data points from Cover_Type 1 and 2. For Cover_Type 1 and 2, we see that many of those types are found in Wilderness_Areas 1 and 3. For Cover_Type 3, most are found in Wilderness_Areas 3 and 4. As for Cover_Type 4 through 7, we can also use the same tool to analyze its correlation with Wilderness_Area by simply plotting them on a separate graph with a smaller y scale.

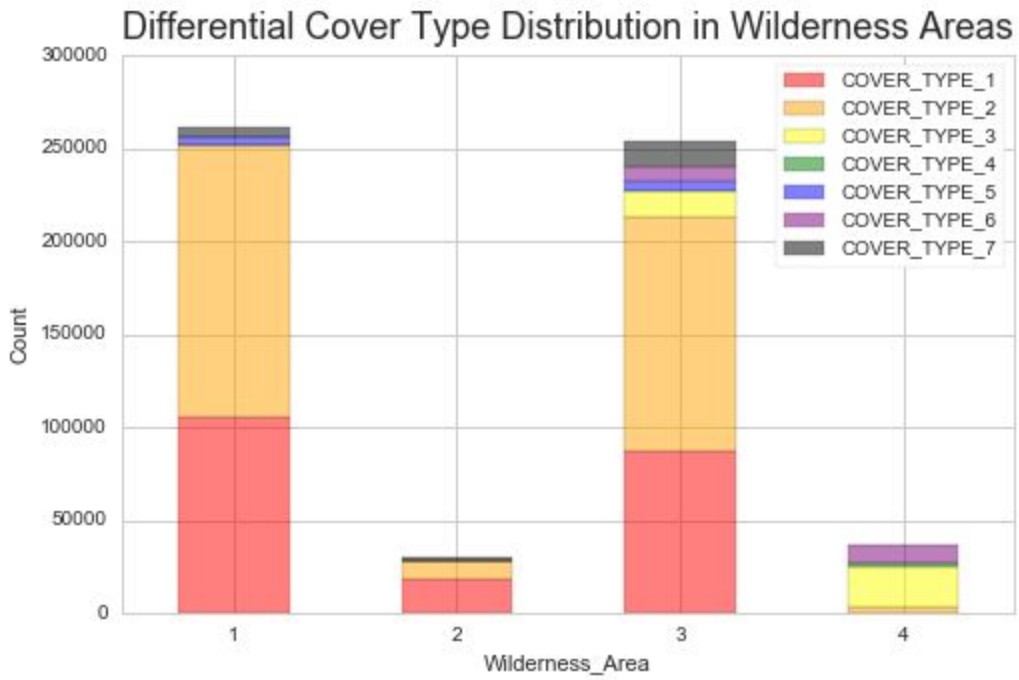

Figure 3. Differential Cover Type Distribution in Wilderness Areas

Examples:

First, we can see that we have more data points from Wilderness_Area 1 and 3. In Wilderness_Area 1, 2 and 3, we found many Cover_Type 1 and 2. Cover_Type 3 is the most prevalent tree cover in Wilderness_Area.

The last variable we analyzed is Soil_Type. We used a heat map, shown below in Figure 4, to illustrate the correlation between Soil_Type and Cover_Type (Here Cover_Type goes from 0 to 6 instead of 1 to 7 as we used before. We can find a way to shift the index up by 1 if we had more time). We took the log of the counts for better visualization.

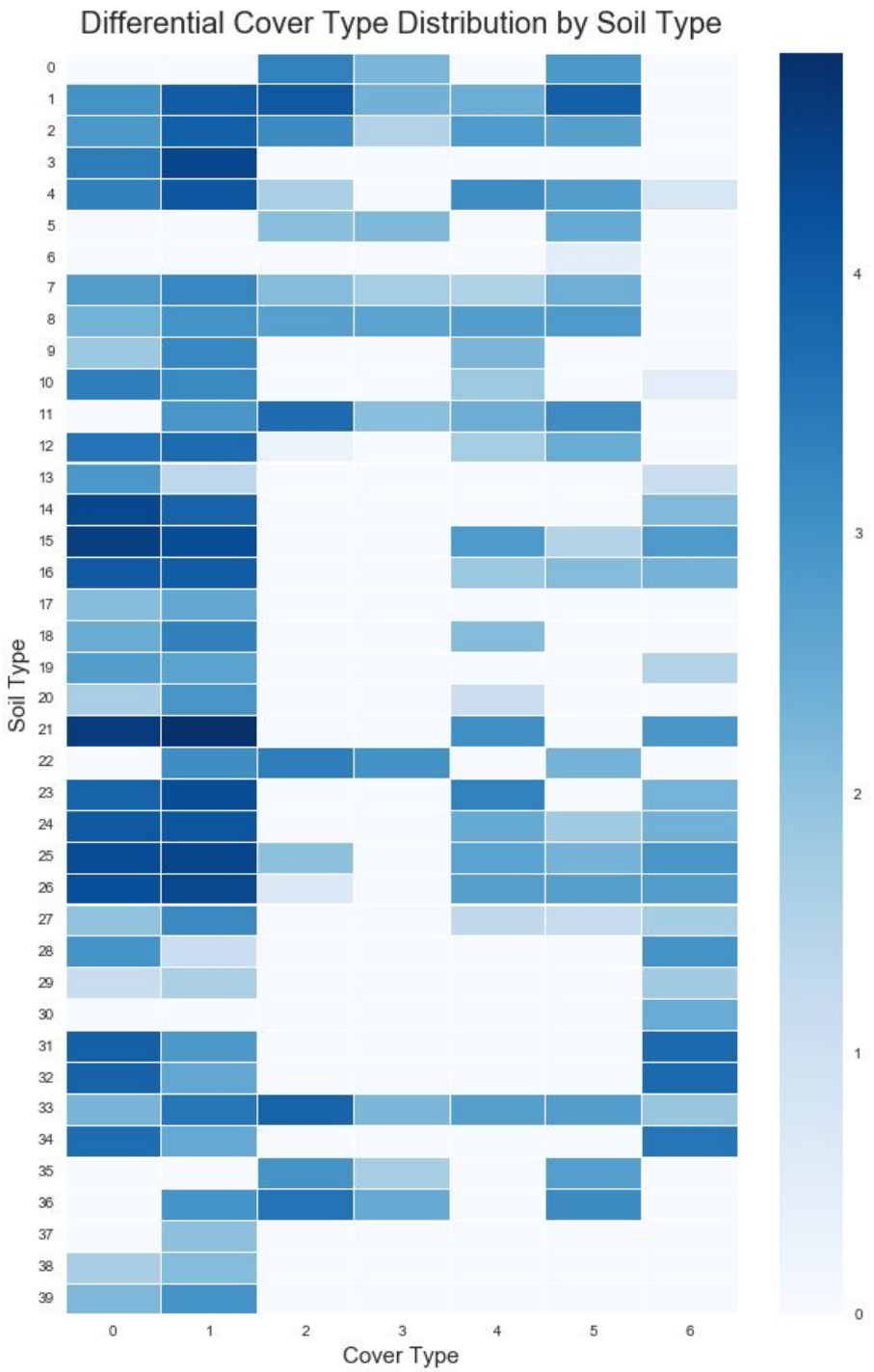

Figure 4. Differential Cover Type Distribution by Soil Type

Examples:

We can see the differential distribution of different Cover_Type in different Soil_Type. For example, Cover_Types 2 are found in most Soil_Types, except Soil_Type 0, 5, 6, 30, 35. If we look horizontally, for example, Soil_Type 22 has Cover_Type 1, 2, 3, and 5, but not many 0, 4, and 6.