Deliverable – 2

Predicting tree types found in the Roosevelt National Forest in Colorado

Team members:

Sai Anish Chowdary Bezawada (801312402)

Pranav Sundaresan Babu (801254287)

Sushma Narne (801269072)

Sai Rachana Reddy Vanipenta (801266847)

Abhishekji Dumala (801307903)

Communication plan to include project artifact repository:

A Word file will be uploaded to Canvas as a way to complete Deliverable 1. Code, screenshots of estimated accuracy and column impact will all be stored in a repository that is established once all other deliverables have been finished. Everyone will have access to the repository as it is made public.

Project repository can be accessed on Github using the following Link given below:

https://github.com/anish1999-13/Big_data_Project

Data Set Selection :

We have selected our dataset from Kaggle. Below are the link to selected data set:

https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset

Business Problem:

Understanding of the various forest types and their features is a business opportunity or challenge that must be kept in mind in order to make decisions about land management and conservation initiations. We will be creating a model that would give an overview of the forest's cover type depending on various characteristics such as, elevation type, slope type and the most important is the type of soil. We are interested in determining the features which are the most important for the overview of the type of forest cover.

Keeping all these factors in mind this dataset's business problem is to create a Machine Learning Model which predicts the forest's cover type with the various features/attributes and to determine the most important features that would help us achieve the goal. This model can assist us in making well-informed decisions regarding land management and conservation activities, such as identifying regions that may need various management

techniques depending on their kind of cover or regions that are especially vulnerable to specific disturbances or threats.

Input Data Set:

We have selected our dataset from Kaggle. The Roosevelt National Forest in northern Colorado's Roosevelt National Forest is included in the Forest Cover Type dataset from the UC Irvine Machine Learning Repository. The target variable is a categorical variable that represents the cover_type of the forest, which has seven potential classes, in the dataset, which also contains continuous and categorical variables.

Here's an explanation of each attribute in the dataset:

Elevation: Elevation in meters

Aspect: Aspect in degrees azimuth

Slope: Slope in degrees

Horizontal_Distance_To_Hydrology: Horizontal distance to nearest surface water features in meters

Vertical_Distance_To_Hydrology: Vertical distance to nearest surface water features in meters

Horizontal_Distance_To_Roadways: Horizontal distance to nearest roadway in meters

Hillshade_9am: Hillshade index at 9am, summer solstice (0 to 255 index)

Hillshade_Noon: Hillshade index at noon, summer solstice (0 to 255 index)

Hillshade_3pm: Hillshade index at 3pm, summer solstice (0 to 255 index)

Horizontal_Distance_To_Fire_Points: Horizontal distance to nearest wildfire ignition points, in meters

Wilderness_Area (4 binary columns): Wilderness area designation

Soil_Type (40 binary columns): Soil Type designation

The "Wilderness_Area" and "Soil_Type" attributes are binary columns indicating whether or not the forest observation falls into a particular wilderness area or soil type, respectively.

Research Objectives:

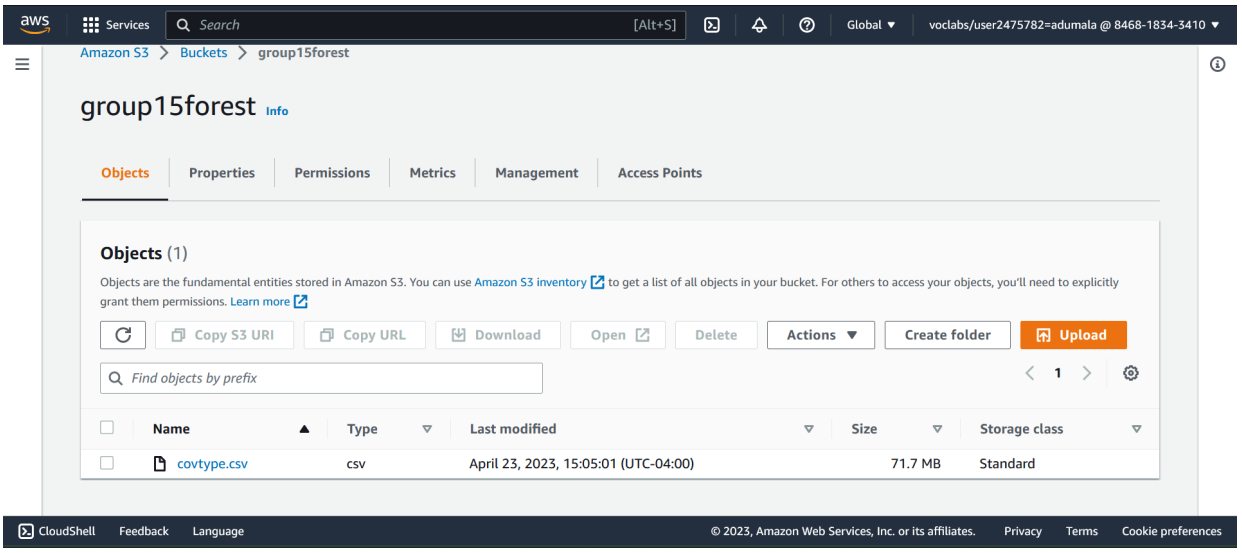
- Evaluating the performance of different machine learning algorithms for the prediction the forest cover type, and identifying the algorithm which performs the best for this model.
- Identify any patterns or trends in the forest cover types over time, and understand how changes in land management and environmental factors may be contributing to these trends.

- Investigate the potential impacts of climate change on the forest cover types and associated vegetation, and understand how the forest cover types may shift over time in response to changing environmental conditions.
- Identify any areas of high conservation value based on the forest cover type and its associated features, and prioritize conservation efforts in these areas.

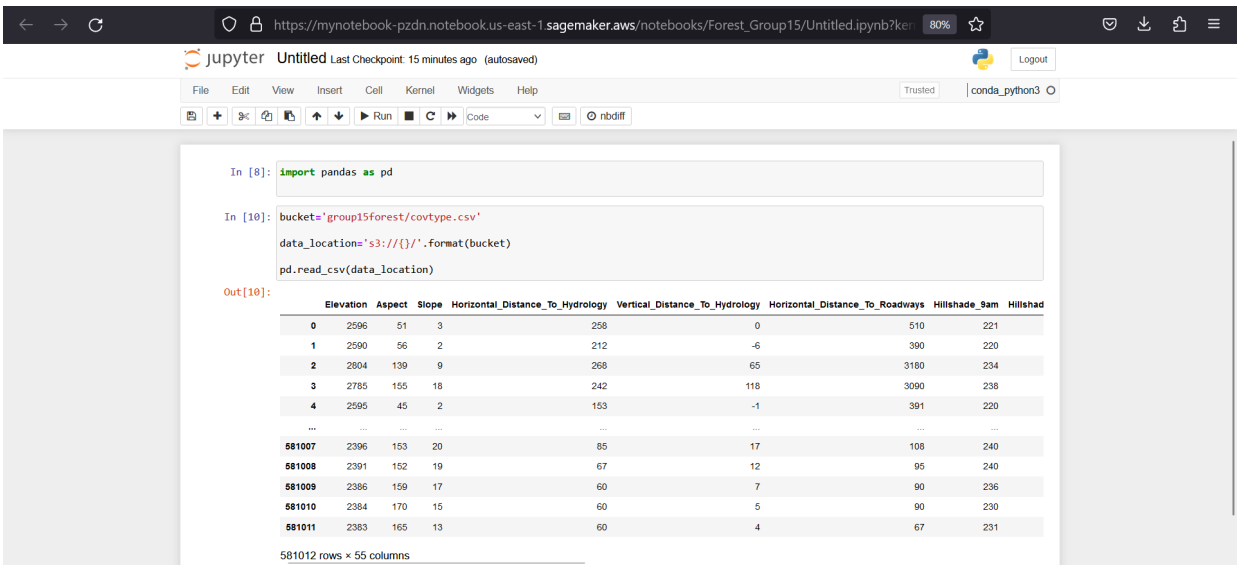
5) Data Understanding

a) Exploratory Data Analysis

We uploaded the data to an S3 bucket named **group15forest** and used the AWS S3 copy command to load it into a Jupyter notebook instance. Afterwards, we combined the data into a single dataframe.

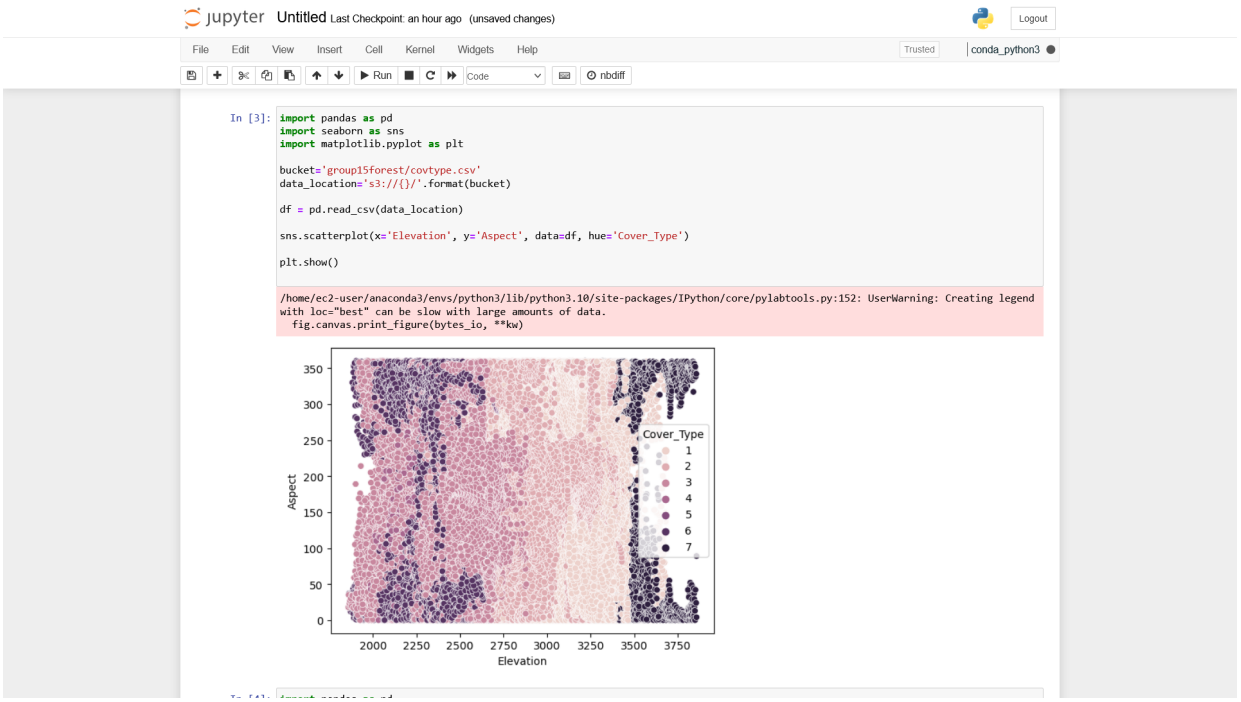


To connect the dataset with Jupyter Notebook, we used the following code to read the dataset:



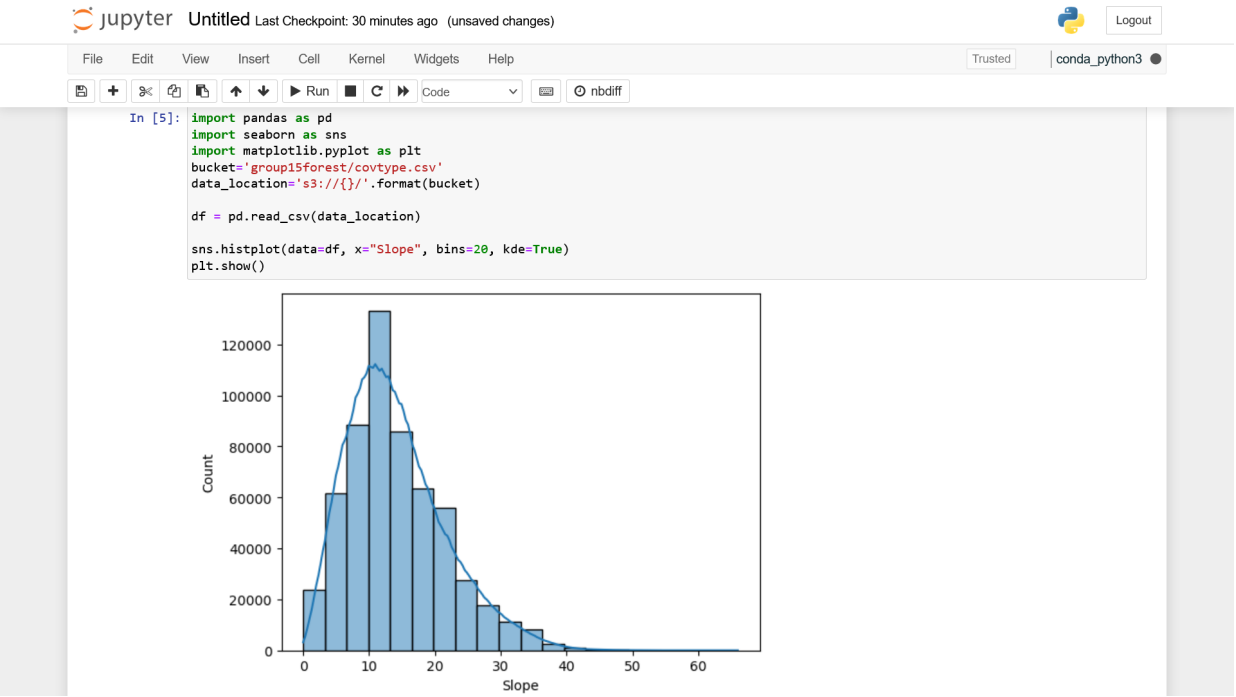
b) Dashboard

Scatter plot analysis of Forest Cover Types based on Elevation and Aspect



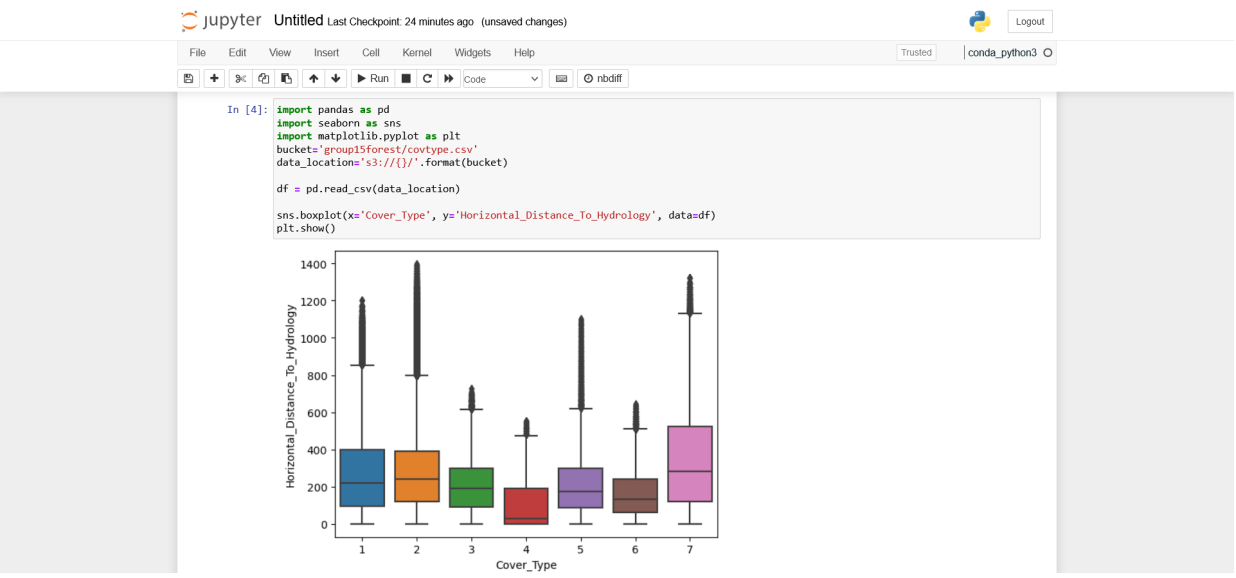
The scatter plot shows how the elevation and aspect of distinct forests correlate with one another. Depending on the type of cover, the plot is color-coded. It is clear that the aspect variation reduces as elevation rises. The plot also shows that different forms of forest cover are concentrated in particular areas, suggesting that elevation and aspect may be important factors in determining the type of forest cover.

Histogram analysis of Slope Values in Forest Areas



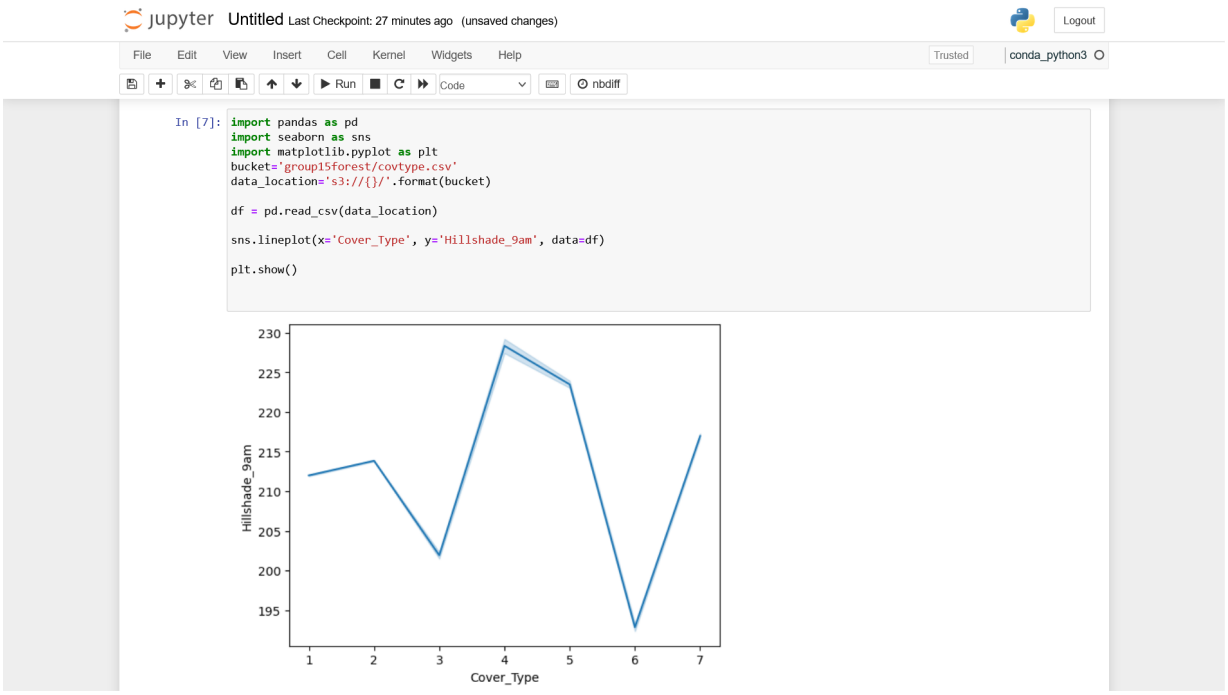
The frequency distribution of slope values across various forest areas is represented by the histogram. The bulk of slopes fall between 0 and 20, with a progressive decline in frequency as slope values increase. The histogram's line is a smoothed estimation of the slope distribution.

Box Plot Analysis of Horizontal Distance to Nearest Water Source for Forest Cover Types



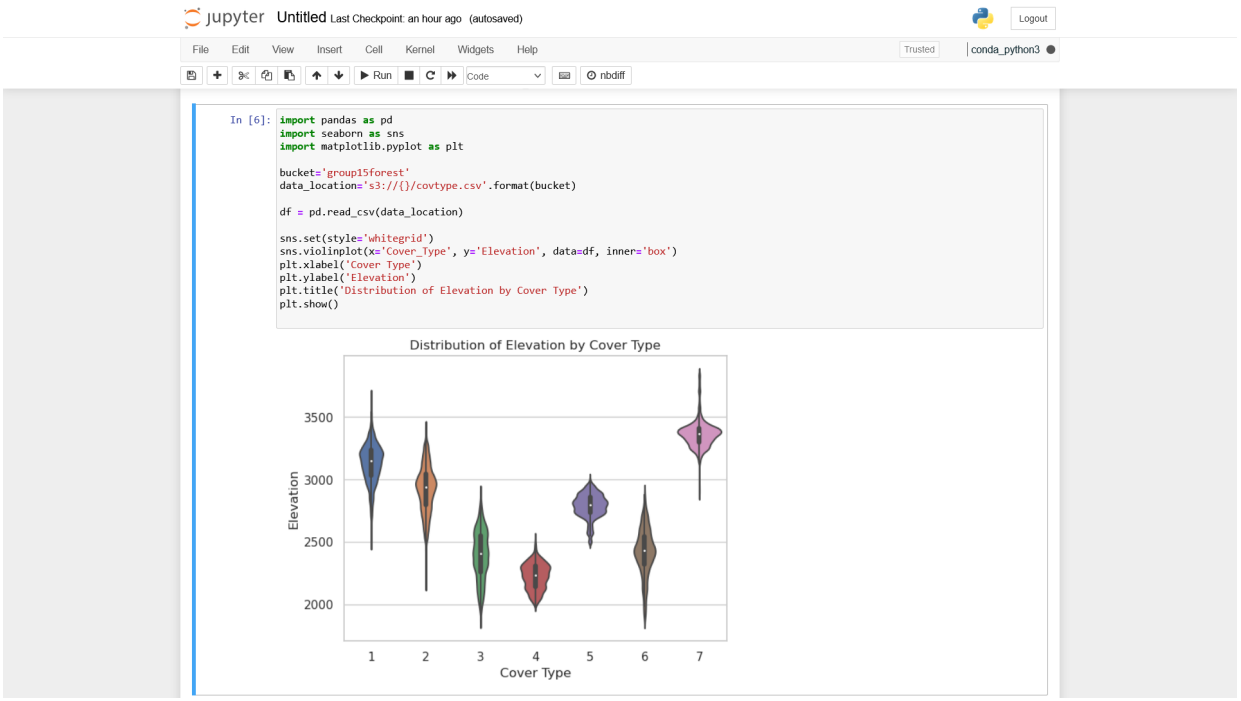
The box plot depicts the distribution of the horizontal distance (hydrology) to the closest water source for each of the seven types of forest cover. Each box's horizontal center line indicates the median value, while the box's upper and lower margins indicate the interquartile range. Individual outliers are additionally shown in the graphic by dots above or below the whiskers.

Hillshade Comparison of Forest Cover Types at 9am



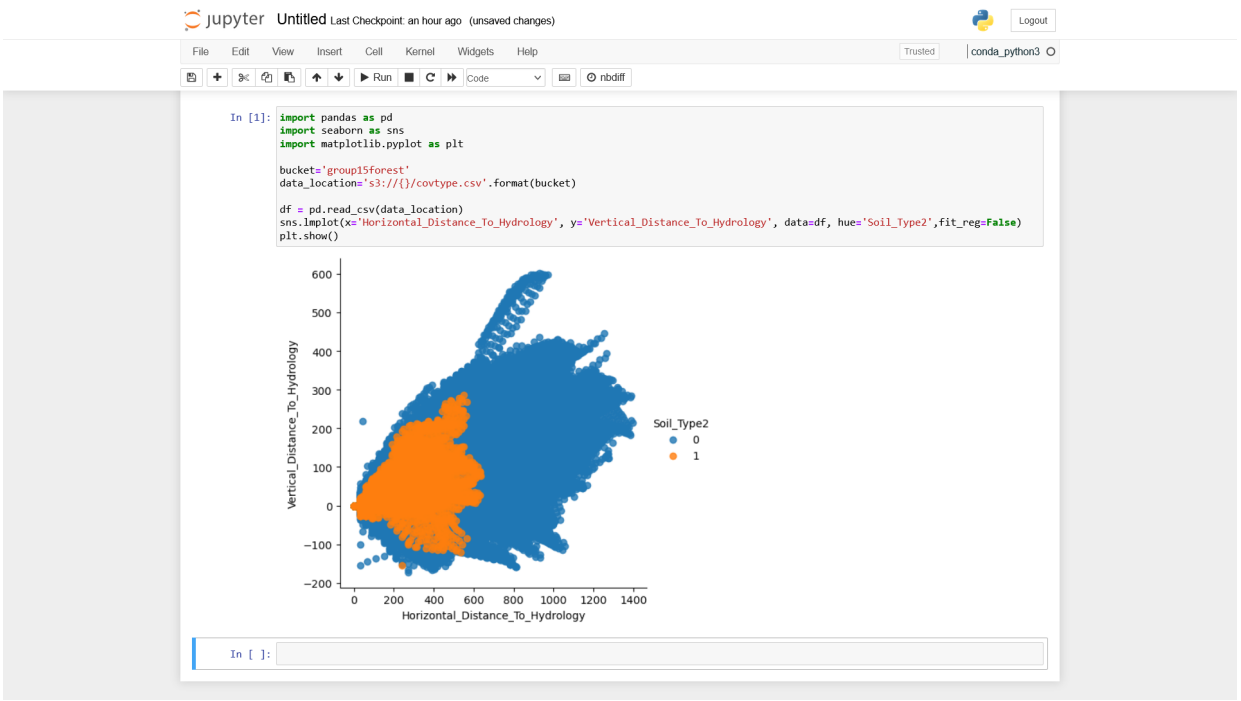
The graph shows the relationship between the forest's type of cover and the hillshade value at 9 am. The y-axis indicates the equivalent hillshade value, and each line in the graph represents a particular type of forest cover. We may compare and contrast the hillshade values for different types of cover by looking at the graph.

Violin Plot Analysis of Elevation Distribution by Forest Cover Type



The graph uses a violin plot to show the frequency distribution of elevation values for each type of forest cover. Different cover types are represented on the x-axis, while related elevation values are indicated on the y-axis. The "Distribution of Elevation by Cover Type" plot has boxes inside the violin plots.

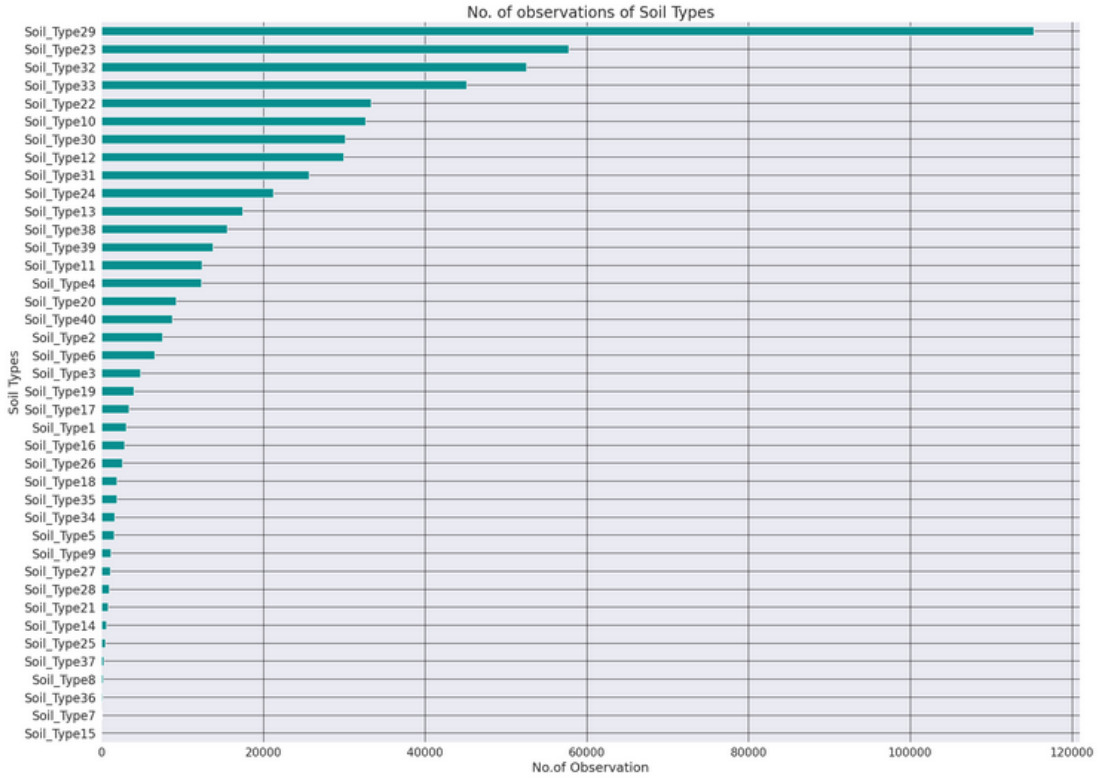
Scatter Plot Analysis of Distances to Hydrology by Soil Type2



The association between the horizontal and vertical distances to hydrology is shown on the scatter plot. Based on whether Soil_Type2 is present or not, the figure is color-coded, and the regression line is ignored (fit_reg=False).

Bar Chart Analysis of Soil Type Frequency in the Dataset

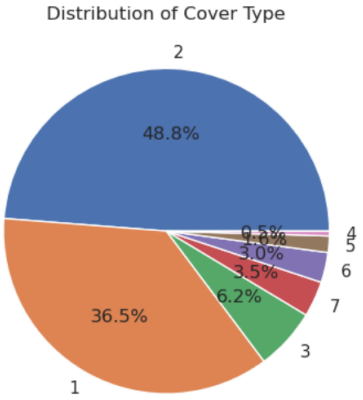
In [8]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

bucket = 'group15forest/covtype.csv'
data_location = 's3://{}/'.format(bucket)

df = pd.read_csv(data_location)

sns.set_style("darkgrid", {'grid.color': '.1'})

Soil_data = df.iloc[:, 14:54]
soil_sum = pd.Series(Soil_data.sum())
soil_sum.sort_values(ascending=False, inplace=True)

soil_sum.plot(kind='barh', figsize=(23, 17), color='#088F8F')

plt.gca().invert_yaxis()

plt.title('No. of observations of Soil Types', size=20)
plt.xlabel('No.of Observation', size=17)
plt.ylabel('Soil Types', size=17)

plt.xticks(rotation='horizontal', size=15)
plt.yticks(size=16)

sns.despine()

plt.show()
```



The graph uses a horizontal bar chart to show the frequency of occurrence for each soil type in the dataset. Based on the quantity of observations, the soil types are arranged in decreasing order, and the plot is designed with a dark grid. The plot also features an inverted y-axis.

Pie Chart Analysis of Cover Type Distribution in the Dataset

```
In [9]: import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt

        bucket='group15forest'
        data_location='s3://{}/covtype.csv'.format(bucket)

        df = pd.read_csv(data_location)

        sns.set(style='whitegrid')
        counts = df['Cover_Type'].value_counts()
        plt.pie(counts, labels=counts.index, autopct='%1.1f%%')
        plt.title('Distribution of Cover Type')
        plt.show()
```



The distribution of the different cover types in the dataset is shown in the pie chart. The size of each slice in the graph, which represents a different type of cover, corresponds to the number of observations. The figure also has labels inside each slice that show the proportion of observations for each type of cover.

6) Data Preparation

Exploratory Data Analysis on CSV File using Pandas and NumPy Libraries

1. The code reads a CSV file from an S3 bucket.
2. It uses Pandas and NumPy libraries to perform exploratory data analysis on the dataset.
3. The code prints the number of rows and columns in the data.
4. It also displays the data types, summary statistics, missing values, and unique values for categorical columns in the dataset.
5. The exploratory data analysis helps to identify patterns, trends, and outliers in the data, which is useful in developing insights and making informed decisions.

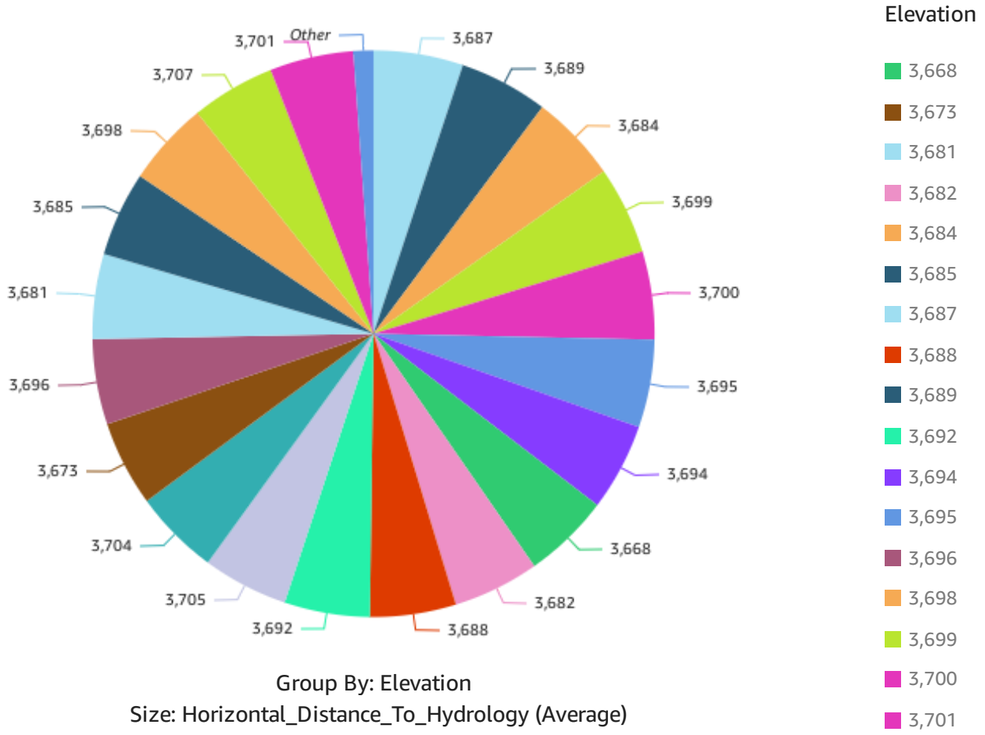
In [11]:

```
import pandas as pd
import numpy as np

bucket='group15forest'
data_location='s3://{}/covtype.csv'.format(bucket)

df = pd.read_csv(data_location)

print('Number of rows:', df.shape[0])
print('Number of columns:', df.shape[1])

print('\nData Types:\n', df.dtypes)

print('\nNumerical Columns:\n', df.describe())
print('\nMissing Values:\n', df.isnull().sum())

print('\nCategorical Columns:\n')
for col in df.select_dtypes(include=['object']):
    print(col, ':', df[col].unique())
```

```
Number of rows: 581012
Number of columns: 55

Data Types:
 Elevation                               int64
Aspect                                  int64
Slope                                   int64
Horizontal_Distance_To_Hydrology        int64
Vertical_Distance_To_Hydrology          int64
Horizontal_Distance_To_Roadways         int64
Hillshade_9am                           int64
Hillshade_Noon                          int64
Hillshade_3pm                           int64
Horizontal_Distance_To_Fire_Points      int64
Wilderness_Area1                        int64
Wilderness_Area2                        int64
Wilderness_Area3                        int64
Wilderness_Area4                        int64
Soil_Type1                              int64
```

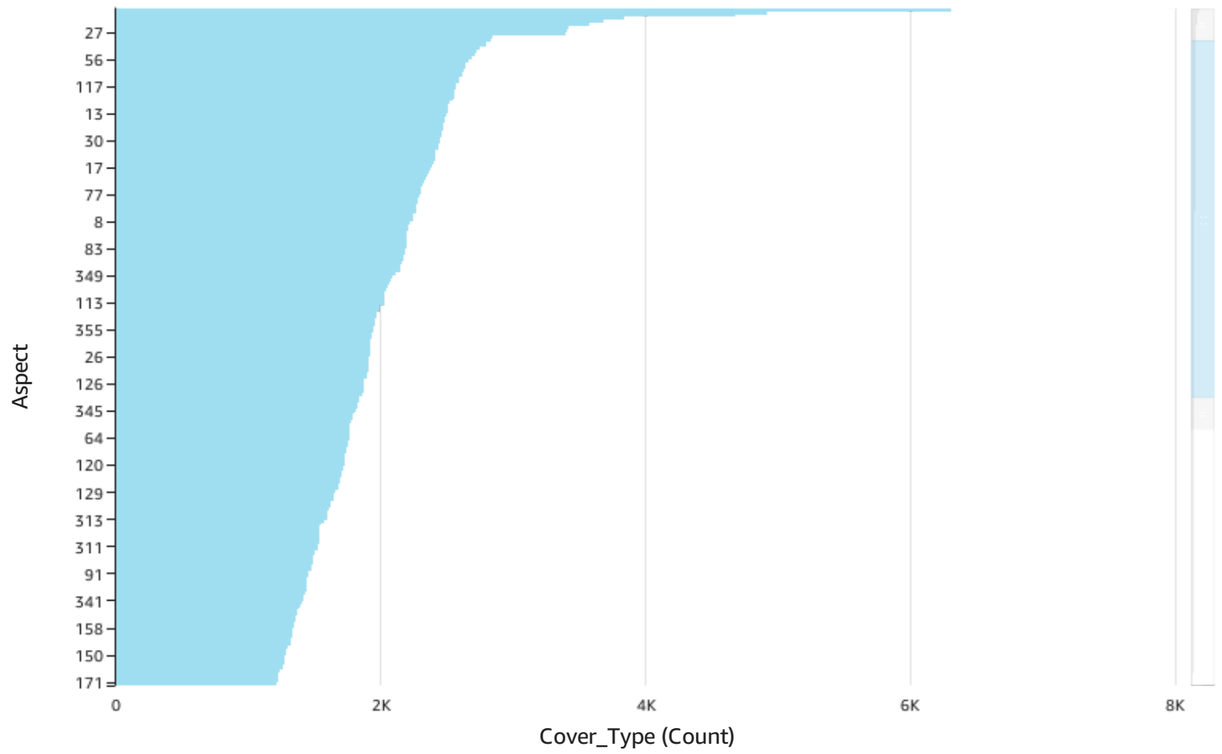
Average of Horizontal_distance_to_hydrology by Elevation

SHOWING TOP 20 IN ELEVATION

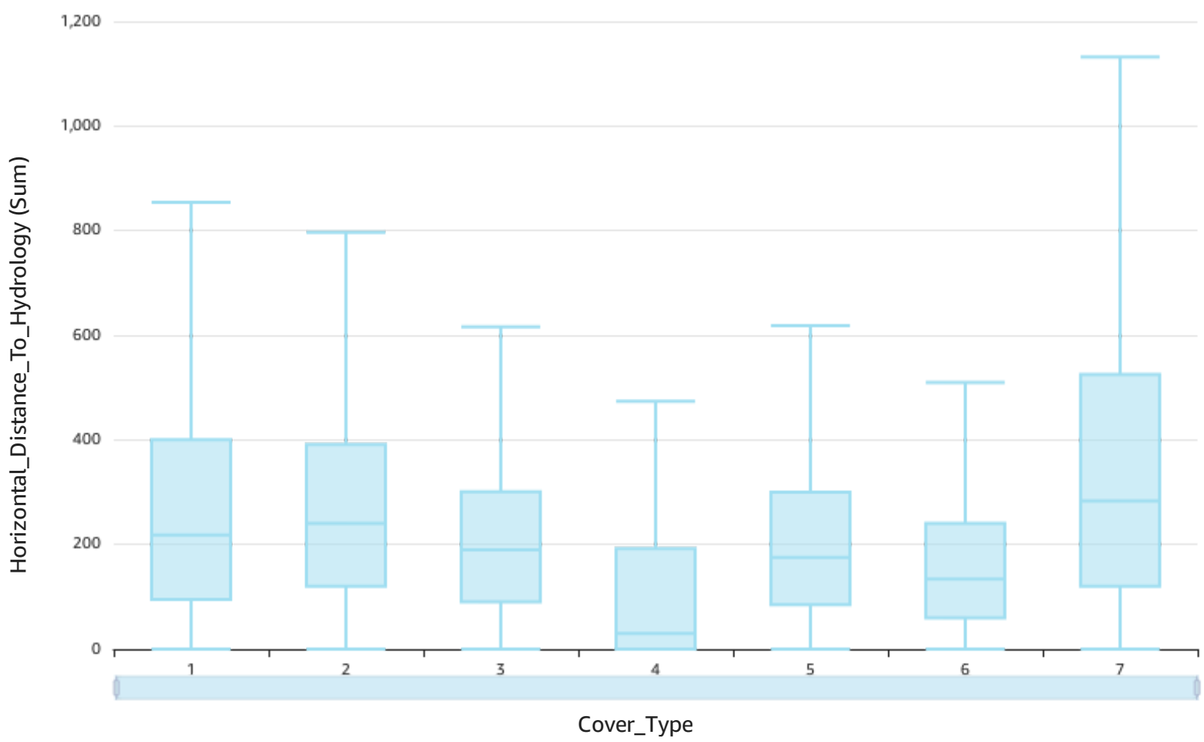


Group By: Elevation
Size: Horizontal_Distance_To_Hydrology (Average)

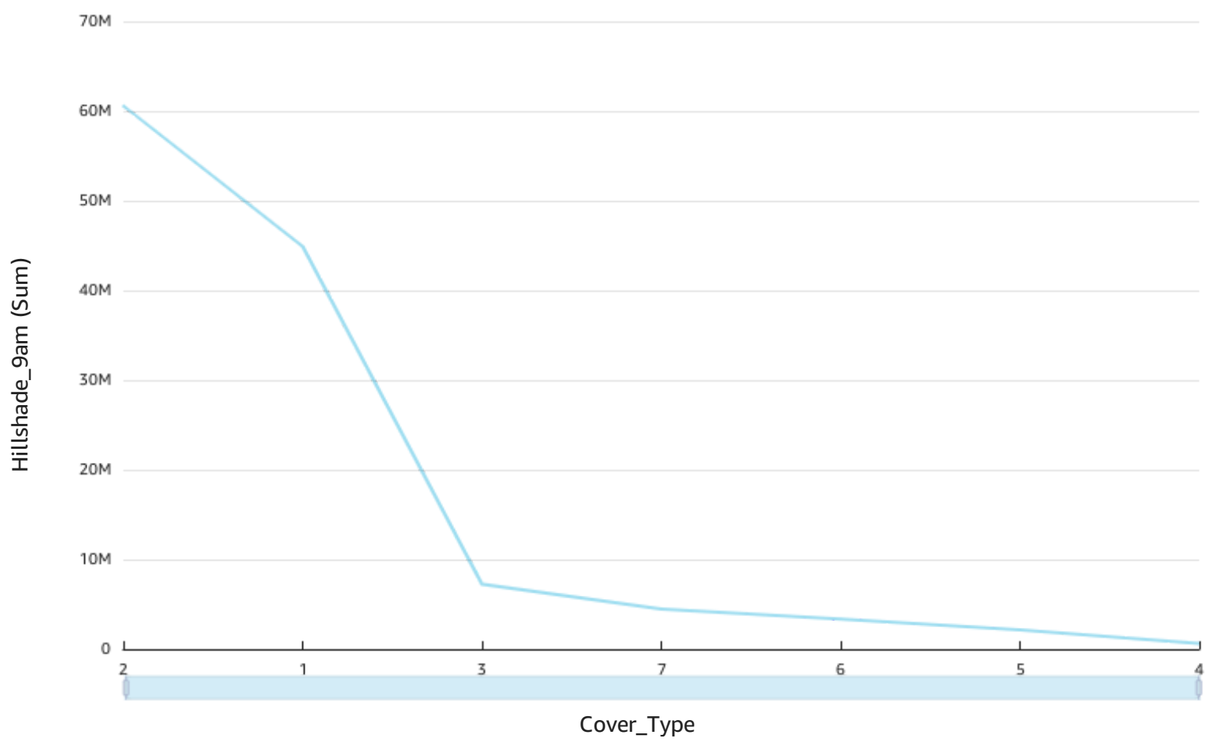
Count of Cover_type by Aspect



Sum of Horizontal_distance_to_hydrology by Cover_type

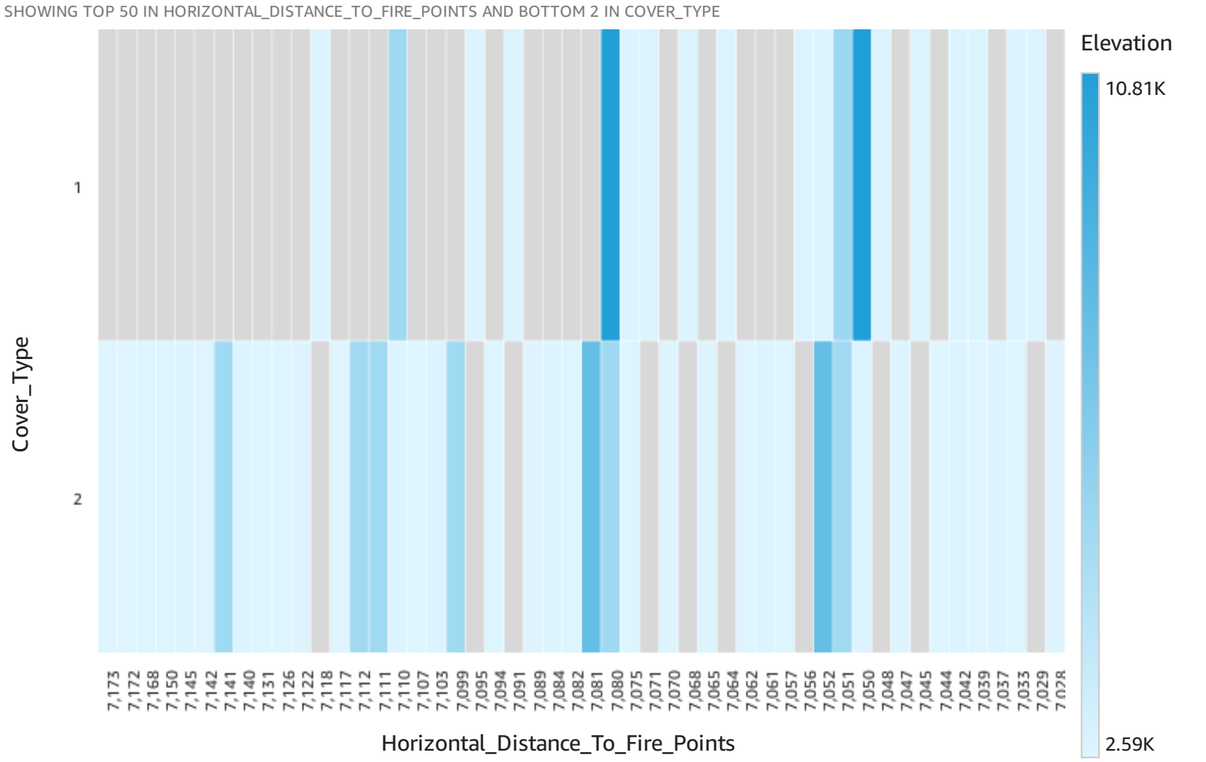


Sum of Hillshade_9am by Cover_type

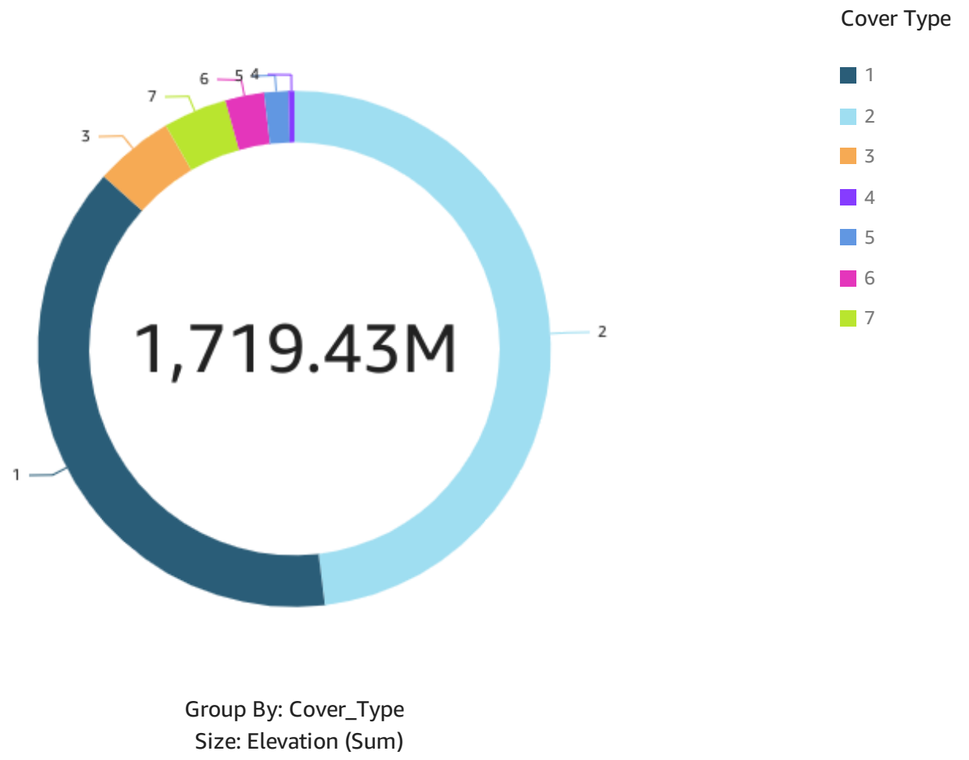


Sum of Elevation by Cover_type and Horizontal_distance_to_fire_points

SHOWING TOP 50 IN HORIZONTAL_DISTANCE_TO_FIRE_POINTS AND BOTTOM 2 IN COVER_TYPE



Sum of Elevation by Cover_type



Group By: Cover_Type
Size: Elevation (Sum)