Programming Code:

Training:

1st cell:

```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read_csv('covtype.csv')

df.head()
```

2nd cell:

```
#Inspecting the data for missing values

df.isnull().sum()
```

3rd cell:

```
# checking the data types

df.info()
```

4th cell:

```
#checking our target variable

df['Cover_Type'].value_counts()

##looks like a very balanced data set
```

5th cell:

```
df.columns
```

6th cell: No output

```
continous_variables = ['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']

categorical_variables = ['Wilderness_Area1',

    'Wilderness_Area2', 'Wilderness_Area3', 'Wilderness_Area4',

    'Soil_Type1', 'Soil_Type2', 'Soil_Type3', 'Soil_Type4', 'Soil_Type5',
    'Soil_Type6', 'Soil_Type7', 'Soil_Type8', 'Soil_Type9', 'Soil_Type10',

    'Soil_Type11', 'Soil_Type12', 'Soil_Type13', 'Soil_Type14',

    'Soil_Type15', 'Soil_Type16', 'Soil_Type17', 'Soil_Type18',

    'Soil_Type19', 'Soil_Type20', 'Soil_Type21', 'Soil_Type22',

    'Soil_Type23', 'Soil_Type24', 'Soil_Type25', 'Soil_Type26',

    'Soil_Type27', 'Soil_Type28', 'Soil_Type29', 'Soil_Type30',

    'Soil_Type31', 'Soil_Type32', 'Soil_Type33', 'Soil_Type34',

    'Soil_Type35', 'Soil_Type36', 'Soil_Type37', 'Soil_Type38',

    'Soil_Type39', 'Soil_Type40', 'Cover_Type']
```

7th cell:

```
wilderness = df[['Cover_Type',  'Wilderness_Area1',

    'Wilderness_Area2', 'Wilderness_Area3', 'Wilderness_Area4']]


wilderness_long = pd.melt(wilderness, id_vars = "Cover_Type", var_name = "Wilderness_Area", value_name = "Area")

wilderness_pivot = pd.pivot_table(wilderness_long, index = 'Cover_Type', columns = 'Wilderness_Area', values = 'Area', aggfunc= 'sum')

wilderness_pivot
```

8th cell:

```
wilderness_long
```

9th cell: No output

```
## same analysis for soil types


soil_types = df[[

    'Soil_Type1', 'Soil_Type2', 'Soil_Type3', 'Soil_Type4', 'Soil_Type5',

    'Soil_Type6', 'Soil_Type7', 'Soil_Type8', 'Soil_Type9', 'Soil_Type10',

    'Soil_Type11', 'Soil_Type12', 'Soil_Type13', 'Soil_Type14',

    'Soil_Type15', 'Soil_Type16', 'Soil_Type17', 'Soil_Type18',

    'Soil_Type19', 'Soil_Type20', 'Soil_Type21', 'Soil_Type22',

    'Soil_Type23', 'Soil_Type24', 'Soil_Type25', 'Soil_Type26',

    'Soil_Type27', 'Soil_Type28', 'Soil_Type29', 'Soil_Type30',
    'Soil_Type31', 'Soil_Type32', 'Soil_Type33', 'Soil_Type34',

    'Soil_Type35', 'Soil_Type36', 'Soil_Type37', 'Soil_Type38',

    'Soil_Type39', 'Soil_Type40', 'Cover_Type']]
```

10th cell:

```
soil_types
```

11th cell:

```
soil_long = pd.melt(soil_types, id_vars = "Cover_Type", var_name = "Soil Types", value_name = "Soil_Types")

soil_long

soil_long['Soil Type Number']= soil_long['Soil Types'].str.replace('Soil_Type','')

soil_long['Soil Type Number']= pd.to_numeric(soil_long['Soil Type Number'])

soil_long
```

12th cell: No output

```
soil_types_pivot = pd.pivot_table(soil_long, index = 'Cover_Type', columns = 'Soil Type Number', values = 'Soil_Types', aggfunc= 'sum')
```

13th cell:

```
soil_types_pivot
```

14th cell :

```
##filter the names of the cover types

list(enumerate(soil_types_pivot.index))
```

15th cell:

```
df[['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']]
```

16th cell:

```
##filter the names of the cover types

list(enumerate(continous_variables))
```

Testing code:

1st cell:

```
plt.figure(figsize=[8,5])

sns.barplot(x= 'Cover_Type', y = 'Area', hue= 'Wilderness_Area', data = wilderness_long,ci= None)

plt.title('Widerness Area for different Forest Cover Types')
```

2nd cell:

```
plt.figure(figsize = (15,13))

for i in enumerate(soil_types_pivot.index):

    plt.subplot(4,2,i[0]+1)

    soil_types_pivot.loc[i[1]].plot(kind= 'bar', color='green')

    plt.title(f'Bar Plot of Forest Cover {i[1]} by Soil Types')

plt.tight_layout()
```

3rd cell:

```
plt.figure(figsize = (15,15))

for i in enumerate(continous_variables):

    plt.subplot(5,2,i[0]+1)

    sns.boxplot(x= df['Cover_Type'], y = df[i[1]], palette = 'turbo')

    plt.title(f'Box Plot of {i[1]} by Forest Covers')

plt.tight_layout()
```

4th cell: No output

```
# correlation and headtmap

corr = df[continous_variables].corr()
```

5th cell:

```
##corelation between continous variables

plt.figure(figsize = (10,5))

sns.heatmap(corr, annot=True,cmap='Reds', fmt = '.2f')
```