```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
import sys
sys.path.append( '/content/drive/MyDrive/CS 480 Kaggle Competition/PlantTraits2024/src')
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os
```

```
data_path = os.getcwd()
df_train = pd.read_csv('/content/drive/MyDrive/CS 480 Kaggle Competition/data/train.csv')
df_train['path'] = '/content/drive/MyDrive/CS 480 Kaggle Competition/data/train.csv' + df_train['id'].astype(str) + '.jpeg'
df_test = pd.read_csv('/content/drive/MyDrive/CS 480 Kaggle Competition/data/test.csv')
df_test['path'] = '/content/drive/MyDrive/CS 480 Kaggle Competition/data/test.csv' + df_test['id'].astype(str) + '.jpeg'
```

```
df_train
```

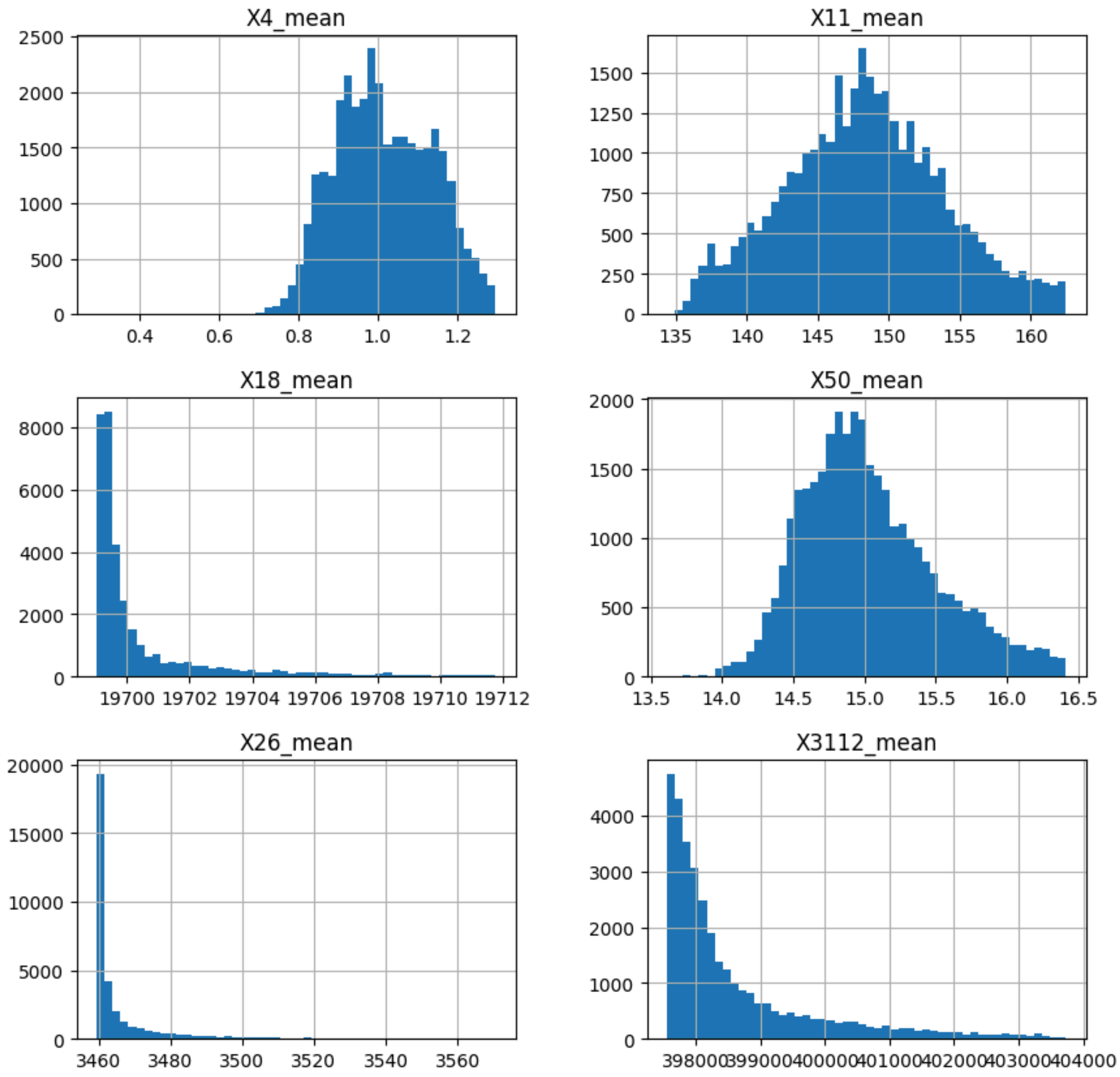
nality	WORLDCLIM_BIO4_temperature_seasonality	WORLDCLIM_BIO7_temperature_annual_range
210766	161.457764	13.886666
906487	178.745422	19.846668
545128	292.781219	23.486668
563957	211.065521	16.768000
409706	36.499138	10.257143
...	...	...
970898	75.369301	12.087244
597702	120.009247	14.226222
789906	473.979675	26.604889
718536	182.917358	22.998470
389532	1090.754761	41.099998

```
#all columns must be identical to be consider the same species
trait_columns = ['X4_mean', 'X11_mean', 'X18_mean', 'X50_mean', 'X26_mean', 'X3112_mean']
aux_columns = list(
            map(lambda x: x.replace("mean", "sd"), trait_columns)
        )
metadata_cols = df_train.drop(
                columns=["id", "path"] + trait_columns
            ).columns
```

```
for col in trait_columns:
    upper_quantile = df_train[col].quantile(0.98)
    df_train = df_train[(df_train[col] < upper_quantile)]
    df_train = df_train[(df_train[col] > 0)]
```

Data Visualization of variables to be predicted

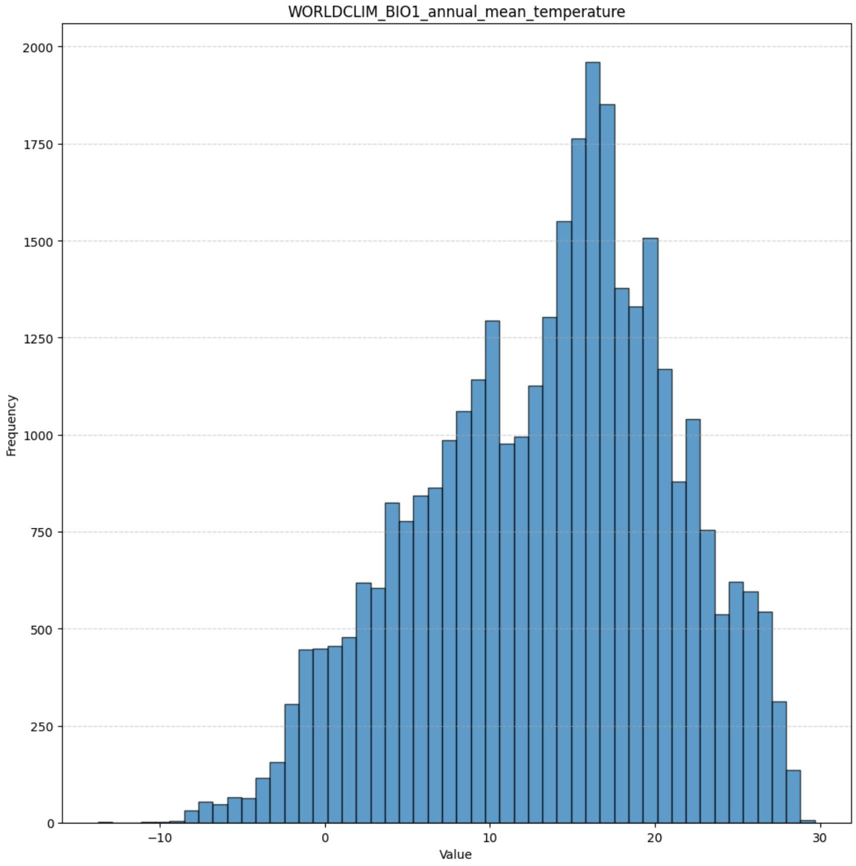
```
df_train[trait_columns].hist(bins=50, figsize=(10, 10))
plt.show()

```

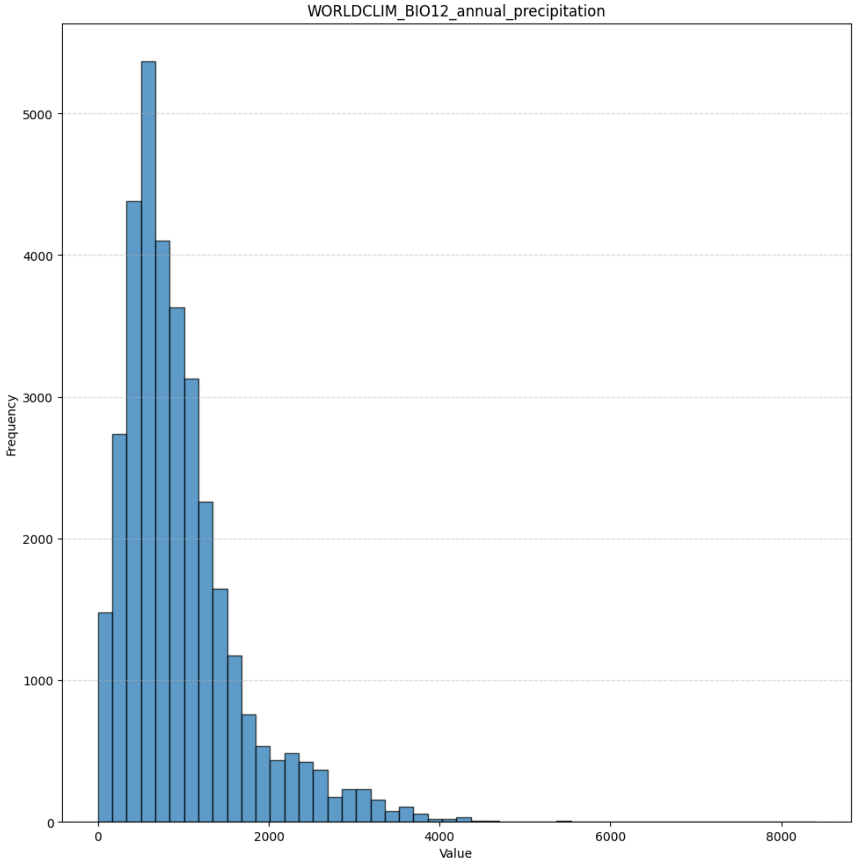


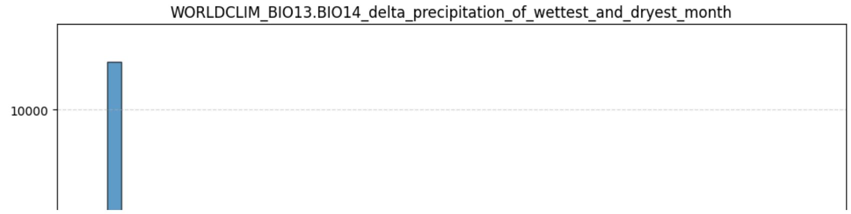
```
def visualize(df_train, colQuery):
  climate_cols = df_train.filter(like=colQuery)
  fig, axs = plt.subplots(nrows=len(climate_cols.columns), ncols=1, figsize=(10, 70))

  # loop through each column and create a histogram
  for i, col in enumerate(climate_cols.columns):
      axs[i].hist(climate_cols[col], bins=50, alpha=0.7, edgecolor = 'black')
      axs[i].set_title(col)
      axs[i].set_xlabel('Value')
      axs[i].set_ylabel('Frequency')
      axs[i].grid(True, axis='y', linestyle='--', alpha=0.5)

  # adjust the layout to avoid overlapping titles
  plt.tight_layout(rect=[0, 0.03, 1, 0.95])


  # show the plot
  plt.show()
```

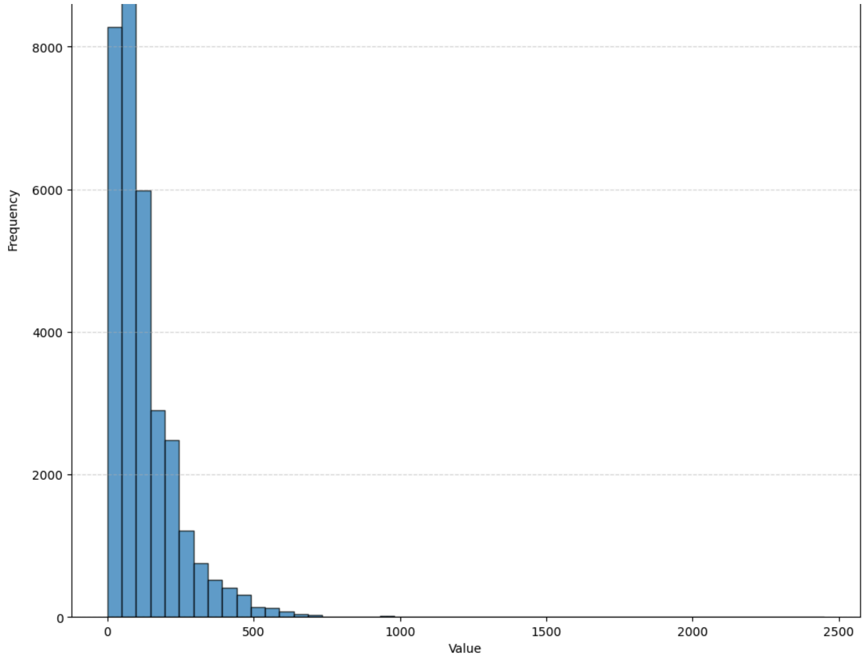
WORLDCLIM_BIO* Data Visualization

```
visualize(df_train, 'WORLDCLIM_BIO')
```

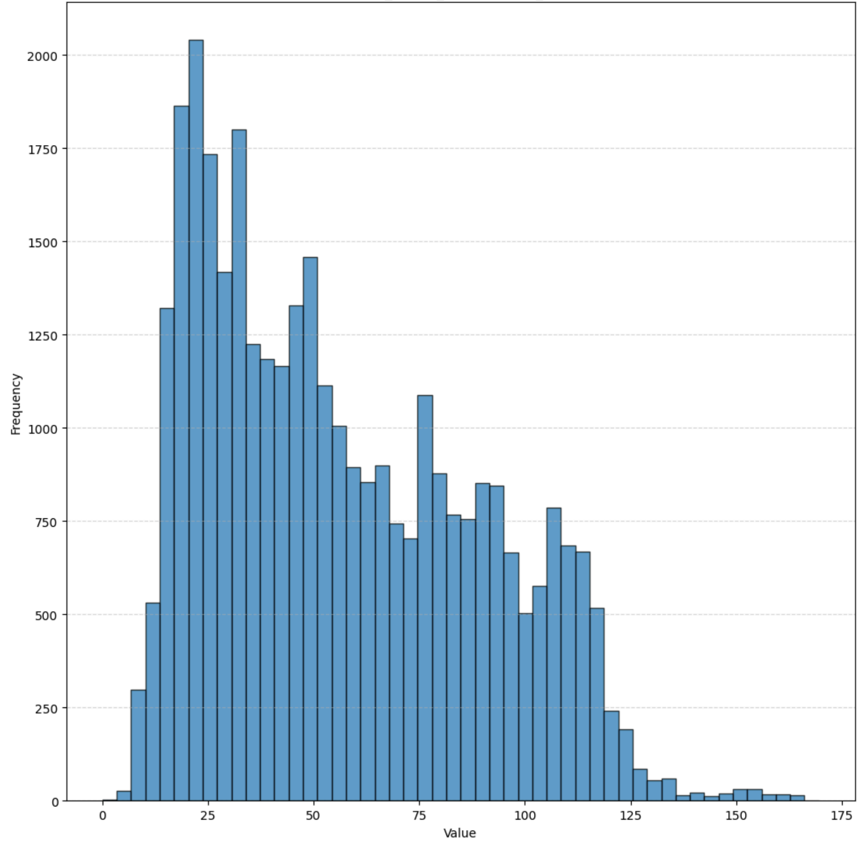

WORLDCLIM_BIO1_annual_mean_temperature
Frequency
0
250
500
750
1000
1250
1500
1750
2000
−10
0
10
20
30
Value


WORLDCLIM_BIO12_annual_precipitation
Frequency
0
1000
2000
3000
4000
5000
0
2000
4000
6000
8000
Value

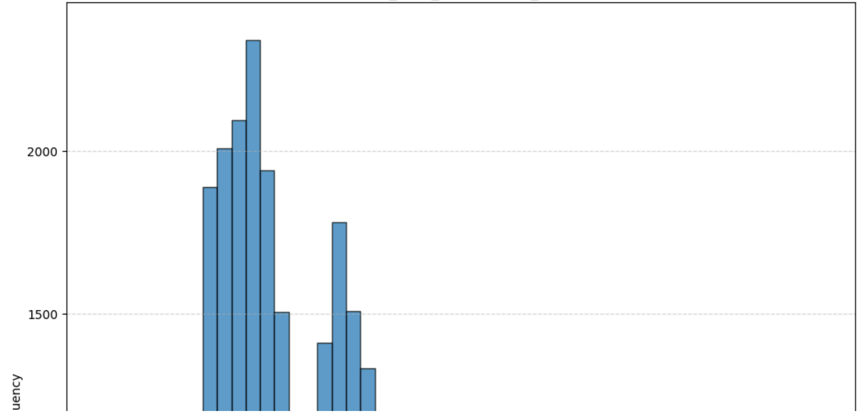

WORLDCLIM_BIO13.BIO14_delta_precipitation_of_wettest_and_dryest_month
10000

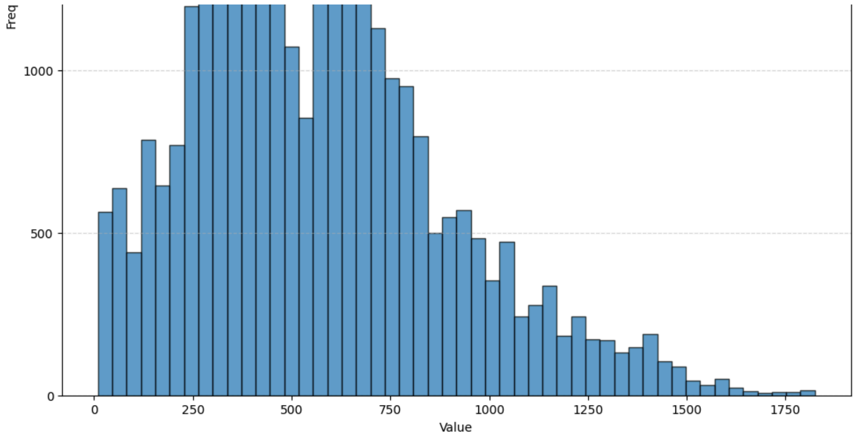


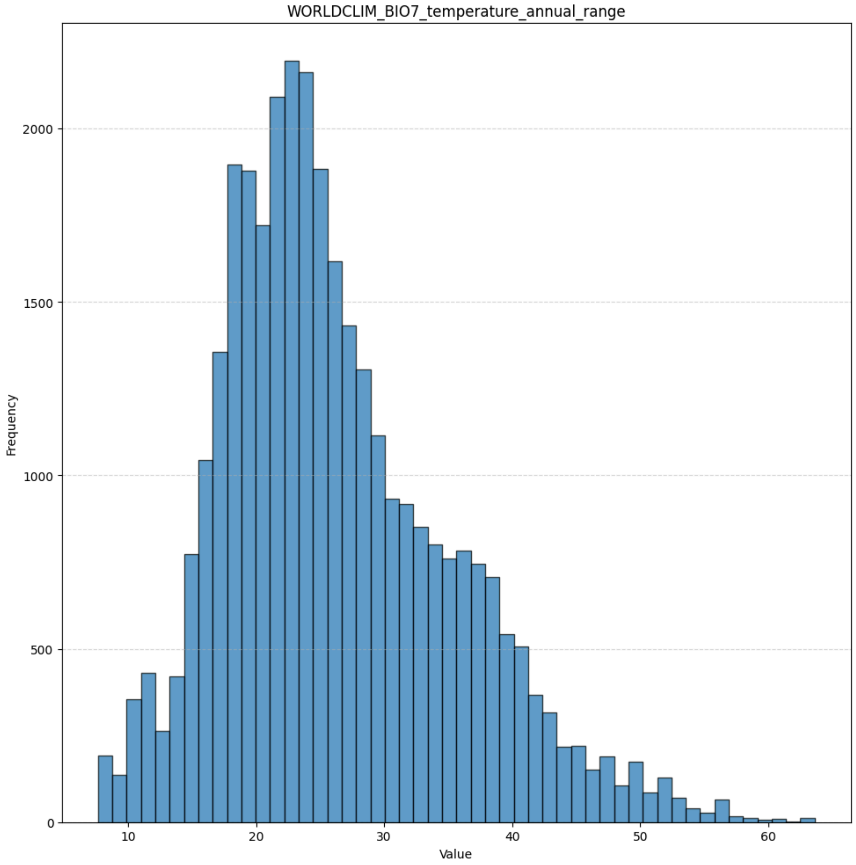
WORLDCLIM_BIO15_precipitation_seasonality



WORLDCLIM_BIO4_temperature_seasonality

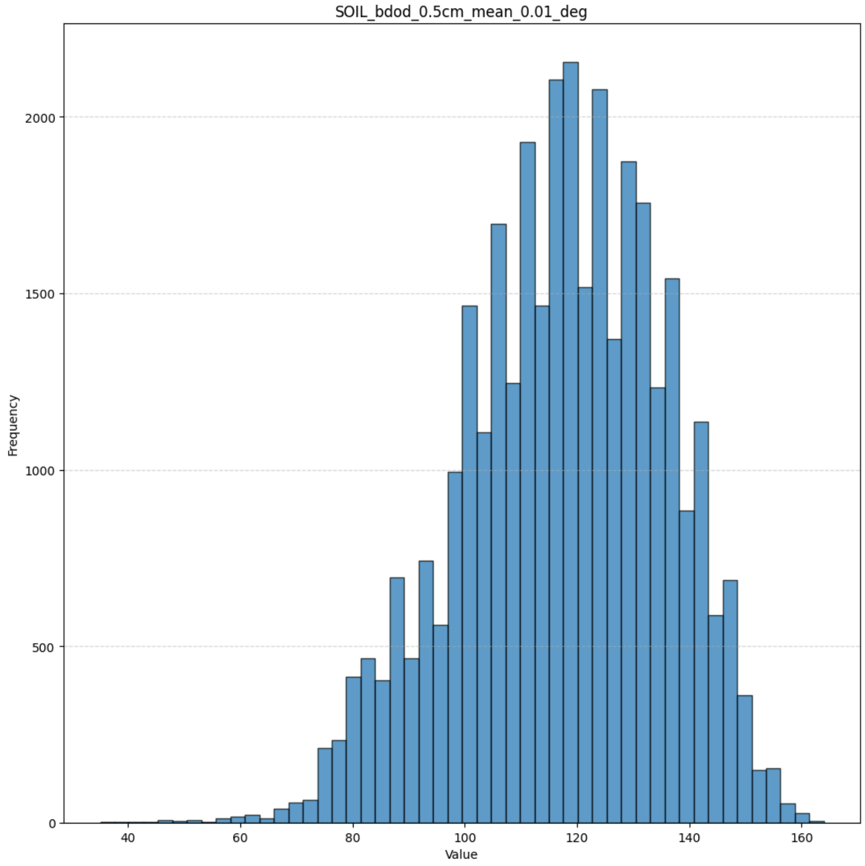


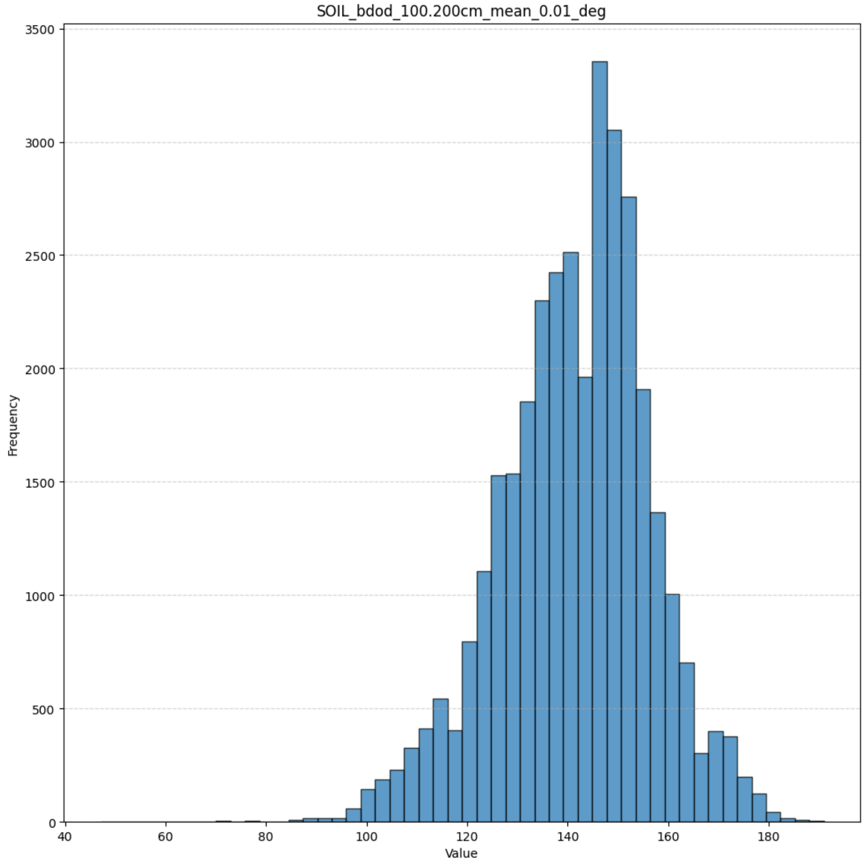

Freq
1000
500
0
0
250
500
750
1000
1250
1500
1750
Value

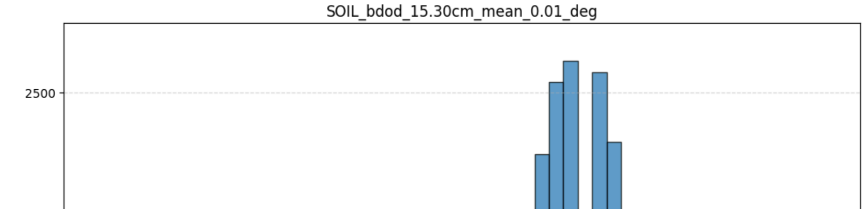

WORLDCLIM_BIO7_temperature_annual_range
Frequency
2000
1500
1000
500
0
10
20
30
40
50
60
Value

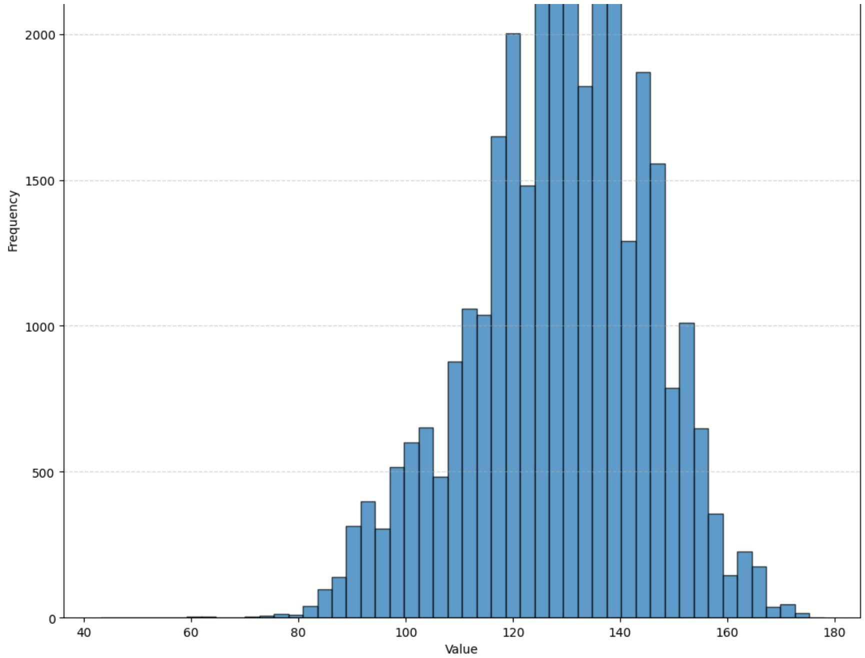
SOIL_bdod* Data Visualization

```
visualize(df_train, 'SOIL_bdod')
```

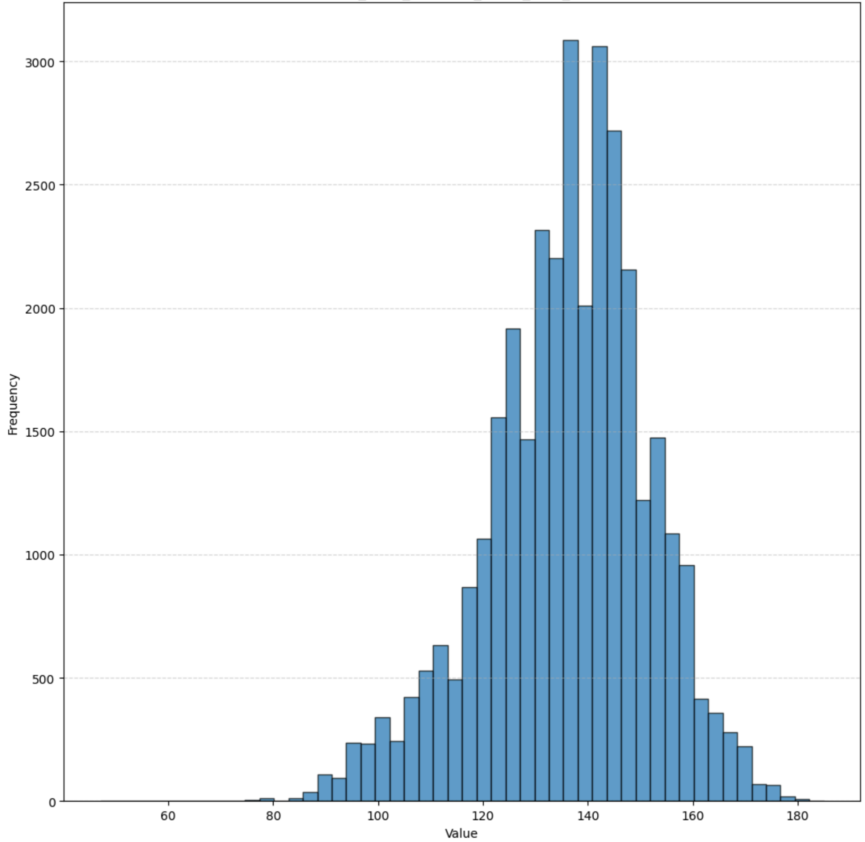

SOIL_bdod_0.5cm_mean_0.01_deg
Frequency
0
500
1000
1500
2000
40
60
80
100
120
140
160
Value


SOIL_bdod_100.200cm_mean_0.01_deg
Frequency
0
500
1000
1500
2000
2500
3000
3500
40
60
80
100
120
140
160
180
Value

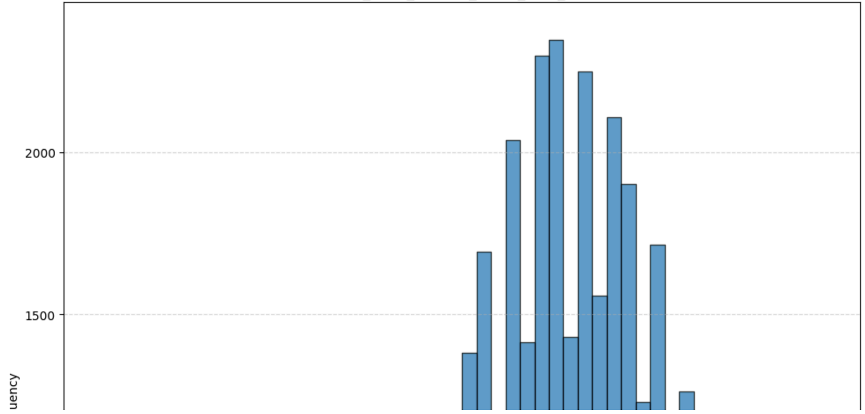

SOIL_bdod_15.30cm_mean_0.01_deg
2500

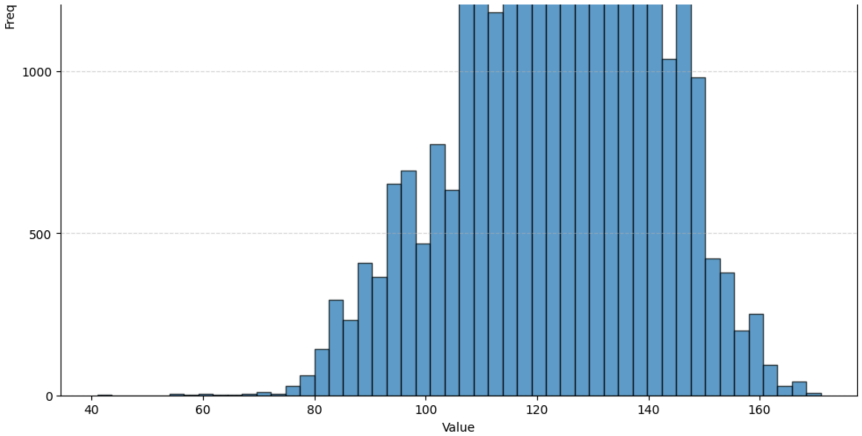


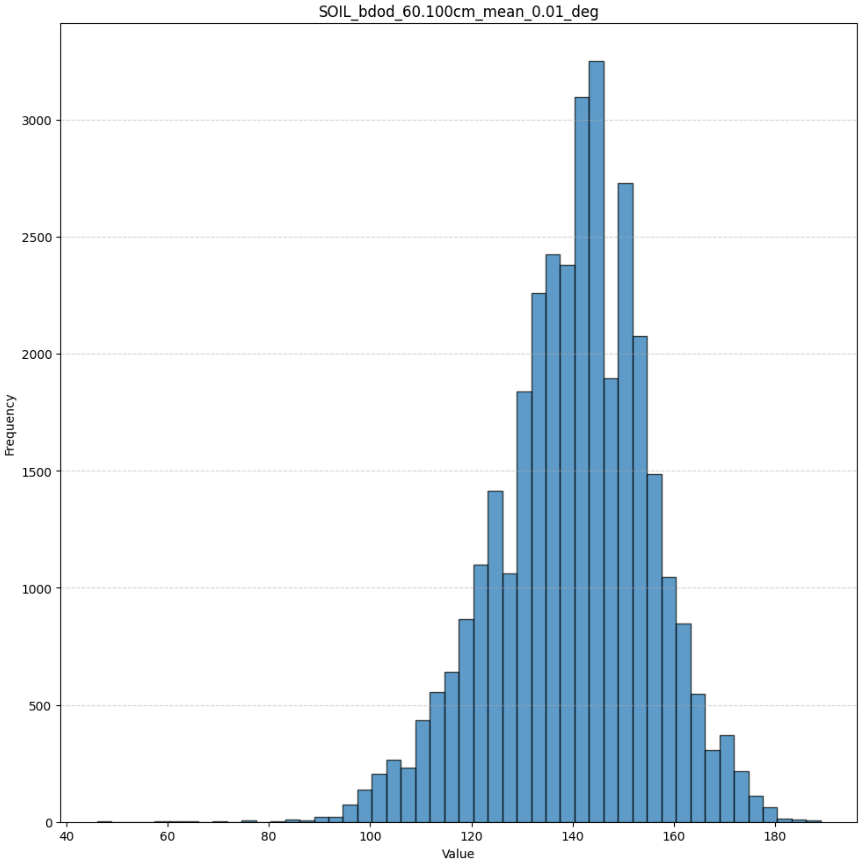
SOIL_bdod_30.60cm_mean_0.01_deg



SOIL_bdod_5.15cm_mean_0.01_deg

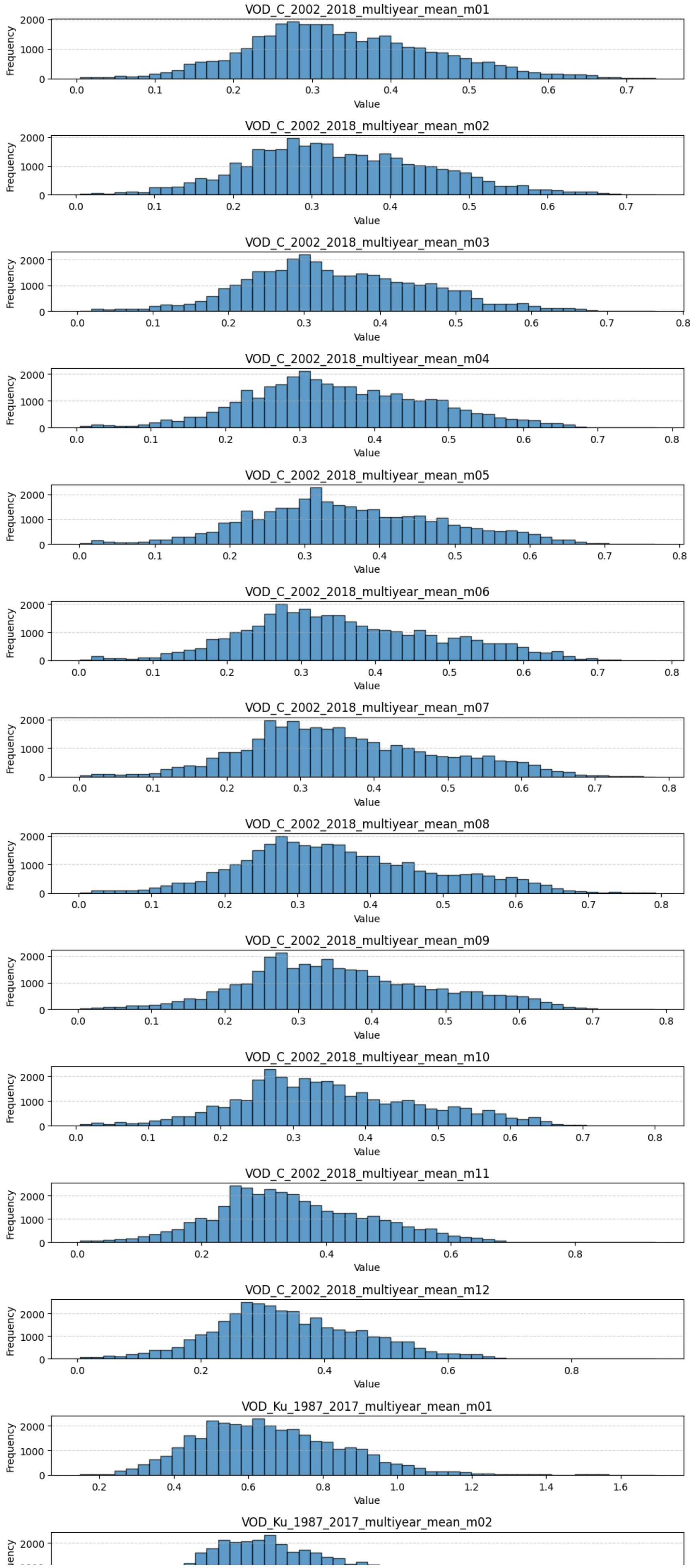





SOIL_bdod_60.100cm_mean_0.01_deg

VOD Data Visualization

```
visualize(df_train, 'VOD')
```


VOD_C_2002_2018_multiyear_mean_m01
Frequency
0
1000
2000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Value
VOD_C_2002_2018_multiyear_mean_m02
Frequency
0
1000
2000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Value
VOD_C_2002_2018_multiyear_mean_m03
Frequency
0
1000
2000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Value
VOD_C_2002_2018_multiyear_mean_m04
Frequency
0
1000
2000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Value
VOD_C_2002_2018_multiyear_mean_m05
Frequency
0
1000
2000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Value
VOD_C_2002_2018_multiyear_mean_m06
Frequency
0
1000
2000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Value
VOD_C_2002_2018_multiyear_mean_m07
Frequency
0
1000
2000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Value
VOD_C_2002_2018_multiyear_mean_m08
Frequency
0
1000
2000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Value
VOD_C_2002_2018_multiyear_mean_m09
Frequency
0
1000
2000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Value
VOD_C_2002_2018_multiyear_mean_m10
Frequency
0
1000
2000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Value
VOD_C_2002_2018_multiyear_mean_m11
Frequency
0
1000
2000
0.0
0.2
0.4
0.6
0.8
Value
VOD_C_2002_2018_multiyear_mean_m12
Frequency
0
1000
2000
0.0
0.2
0.4
0.6
0.8
Value
VOD_Ku_1987_2017_multiyear_mean_m01
Frequency
0
1000
2000
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Value
VOD_Ku_1987_2017_multiyear_mean_m02
ency
2000