Preprocessing steps and rationale

- First Seen The Dataset It seems like a image of 448 columns that is 28*16 or vice versa
- Extracting the features in separate frame and the target value in separate dataframe
- Deleting hsi_id column as its no longer required in model building or visualization
- Tried to plot image to understand the image and seems like a normalized image with a cnn layer already been imposed on it

Insights from dimensionality reduction

- The First Component of pca includes all the important and major information
- Second component of pca gives only 0.04% of information

Model selection, training, and evaluation details

- Model Selected - CNN
- Why - Because The dataset contains the values of images so cnn is always the best for image data
- Training - epochs = 100, batch size= 32
- Evaluation - CNN MAE = 3009.46
- Not calculated mse and r factor since any of the three matrix can tell the performance of model

Key findings and suggestions for improvement

- CNN model helps to predict the values but the dataset provided for training is very less.
- The first component of pca gives all the information but not applied pca on training data because deep learning models are data hungry.
- While plotting all the point we can see that the data is concentrated across a certain range.
- Very few outliers present