This project began with an interest in using data to better understand a climate change related project. The group involved decided that trees would be an interesting topic to go with for the bootcamp project. Trees are related to climate change as a major source for carbon storage. But, human usage of trees can counteract the carbon storage properties by releasing the stored CO_2 into the atmosphere through practices such as clear cutting or burning.

In searching for a dataset, the group settled on the ReforesTree dataset, where the authors explore deep learning techniques for carbon stock estimation. We were interested in the data for the above ground biomass (AGB) prediction, which is a proxy for carbon stock of trees, and whether it could be predicted using deep learning techniques. The main issue is that AGB is highly overestimated in the current estimation techniques despite being very expensive and labor intensive. The paper accompanying the ReforesTree dataset claims that deep learning has promise to make more accurate estimations of the AGB using low-cost drone imagery, which would be of crucial importance in the field of forestry and carbon stock estimation for climate impact. The paper advertised providing a machine learning ready dataset for making AGB predictions with drone images.

We wanted to assess the validity of this claim and explore how well could do with deep learning techniques with drone images. Upon further research, we learned it would be best to utilize convolutional neural networks (CNN) to train, test, and validate the image data to predict AGB as a proxy for carbon storage. The main goal of the project was thus to manipulate the dataset to understand the promise of low-cost drone imagery for AGB prediction.

The coding for this project was done using Python and Jupyter notebooks. There was extensive exploratory data analysis (EDA) done on the provided data. A lot of the field data provided had missing values for tree diameter values, which are important for AGB prediction, so we had to carefully extrapolate the data. Some of the drone images were too big and had to be cropped. We tried out a few different CNN techniques for this work, namely writing a CNN from scratch and using pre-trained models such as EfficientNetB0 and ResNet18. The best results were obtained with ResNet-18. Image pre-processing included cropping, zero padding, image augmentation and resizing to fit the CNN architecture.

The CNN was used to see whether we could classify different tree species that can have high AGB values, such as banana trees. First, binary classification was done to identify banana trees. Then, there was multiclass classification done on banana trees, cacao trees vs all others. It was found that the model is very good at learning to identify banana trees, with over 95% precision and recall values for a dataset split into 70% train, 20% validation and 10% test sets. Multiclass classification results were less encouraging, with a lot of cross talk between cacao and other species, owing to data imbalance, GPS noise, bad image quality for these other tree species, etc. Finally, we tried to predict the AGB and carbon stock values of banana trees using the CNN and a mean squared error loss for this prediction. We found that with the given dataset, the model can only perform as good as learning the average AGB values of the field data. This is contrary to the claims made by the original authors, and it shows that we are still long ways to go before we can only rely on drone imagery to predict AGB and carbon stock.

After extensive work with the data, it was clear that there were limitations to the dataset. More than simple scrubbing and cleaning of the data was necessary to get any good results. Even so, there was a lot of GPS noise and overlap within the data that made it difficult to work on truly classifying based on only the provided photos. Some of the photos being used for learning were not great quality, making it difficult for human and machine alike to classify properly. Also, the DeepForest package version that was initially used to work with the data is no longer available for use, making replication of the original code impossible. By the end of the project, it was clear to us that better drone images and a more accurate drone to field data matching would be needed for classification and prediction purposes.

Due to the time constraints of the project, we could not do any more work with this dataset. Going forward, we would like to apply the work done with this dataset to a different dataset with tree field data as well as accompanying photos. The group is dedicated to staying engaged with this work to further develop prediction models for looking at AGB using the least amount of field data.