Coffee Classification Project

Classifying Coffee Species

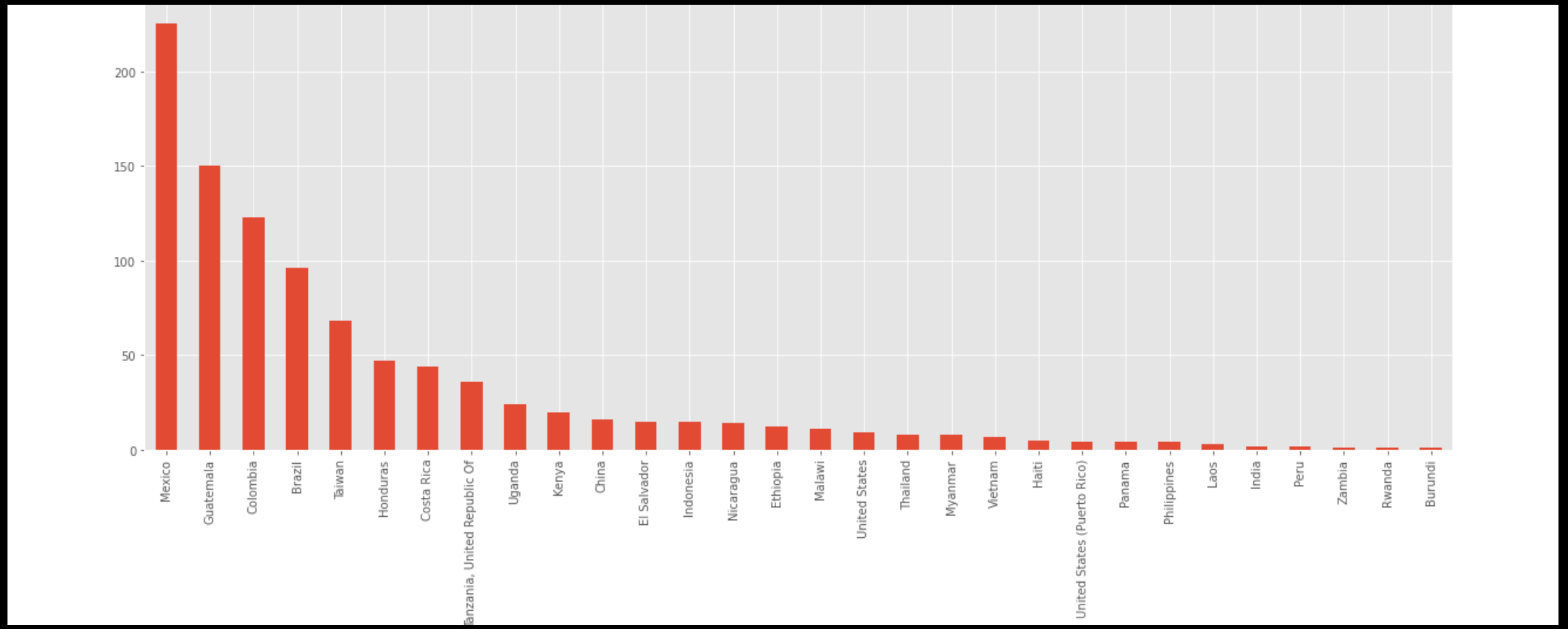
RICARDO BAUTISTA HUERTA

Variables

After filtering the dataset, we have the following variables:

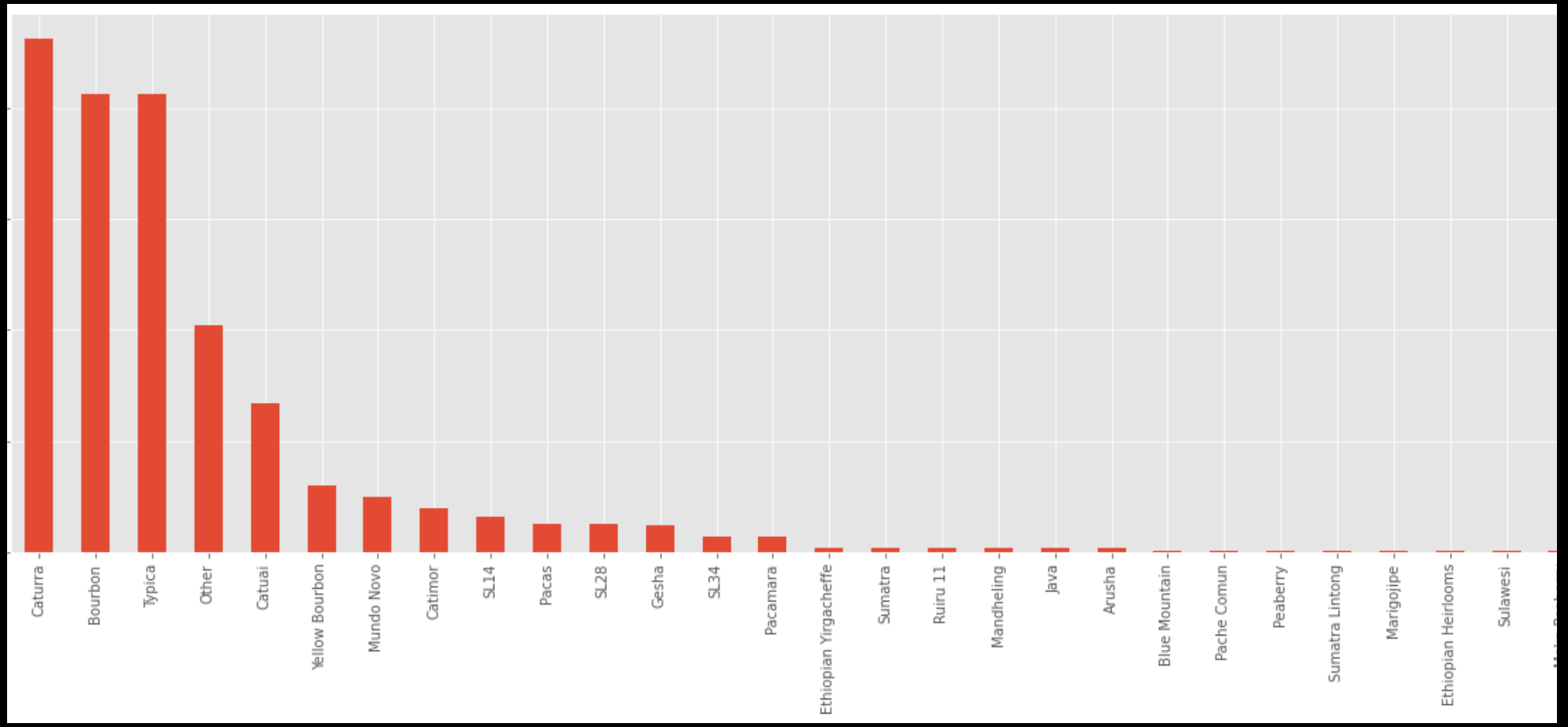
- Species [Arábica or Robusta] (Target Variable)
- Country of Origin
- Harvest Year
- Variety
- Processing Method
- Category One Defects
- Category Two Defects
- Quakers
- Altitude Mean Meters
- Total Cup Points

Country of Origin



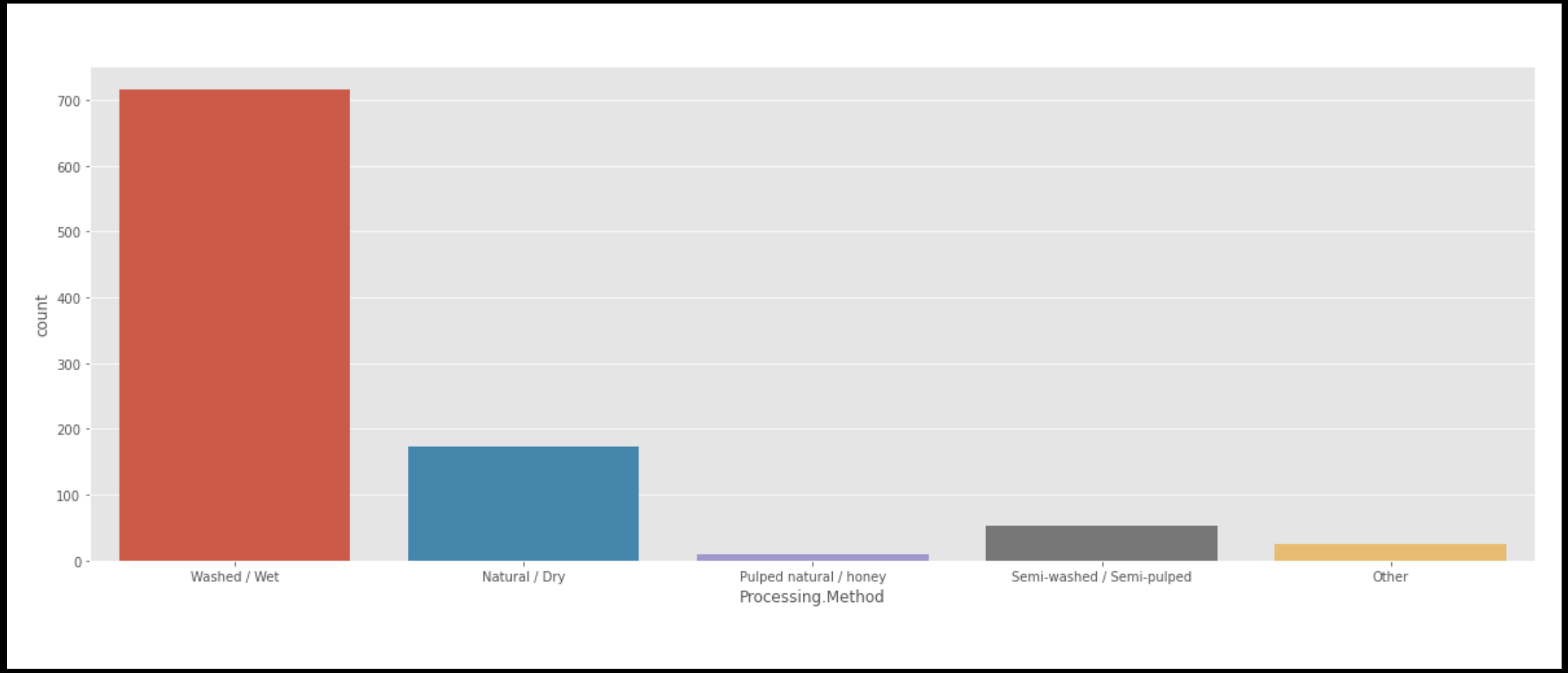
Most of the Coffee in the dataset comes from Mexico, Guatemala, Colombia and Brazil

Variety of Coffee



The principal variety of coffee are Caturra, Bourbon and Typica

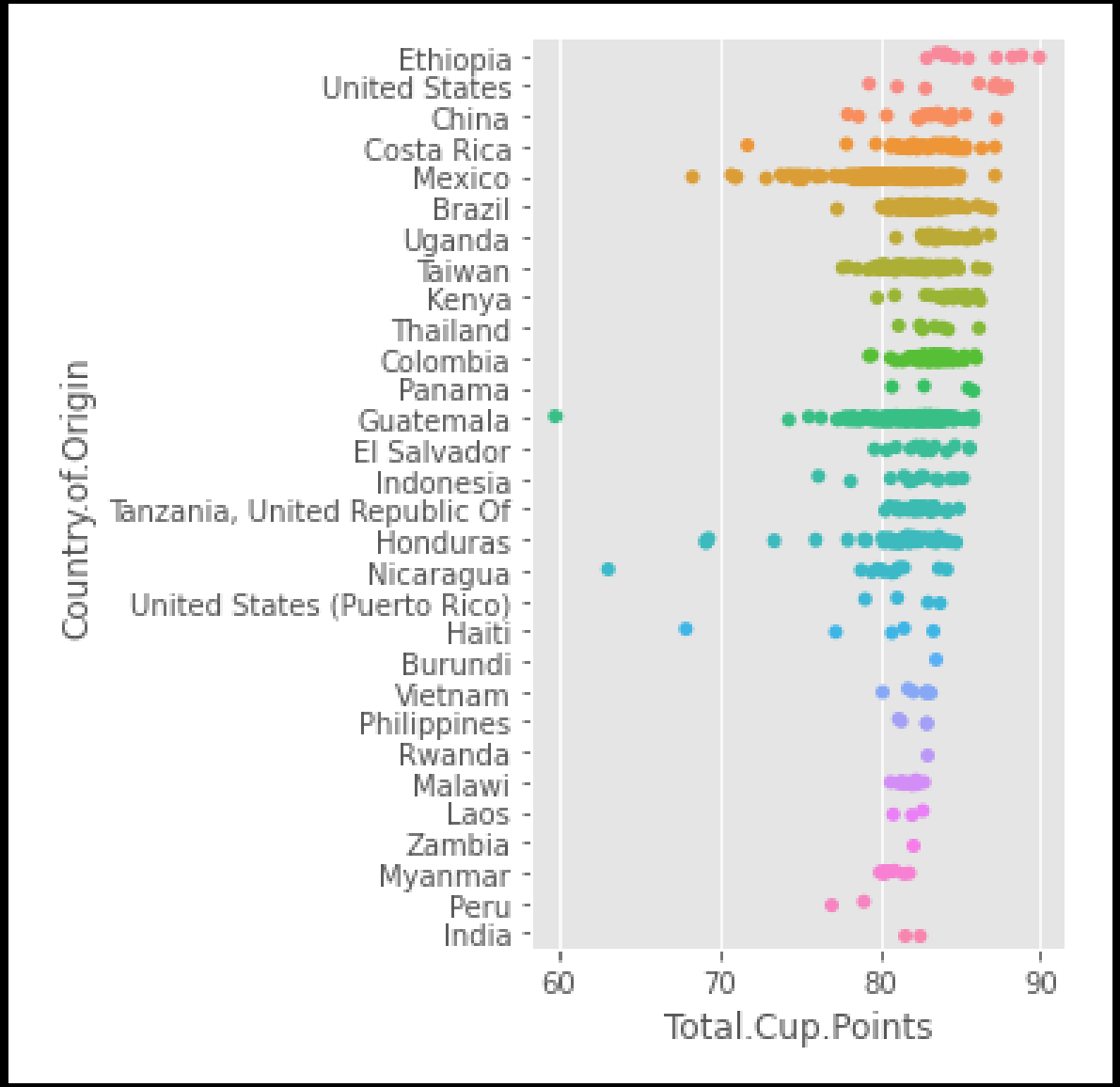
Processing Method



Referring the processing method, the most used is Washed/Wet following Natural/Dry

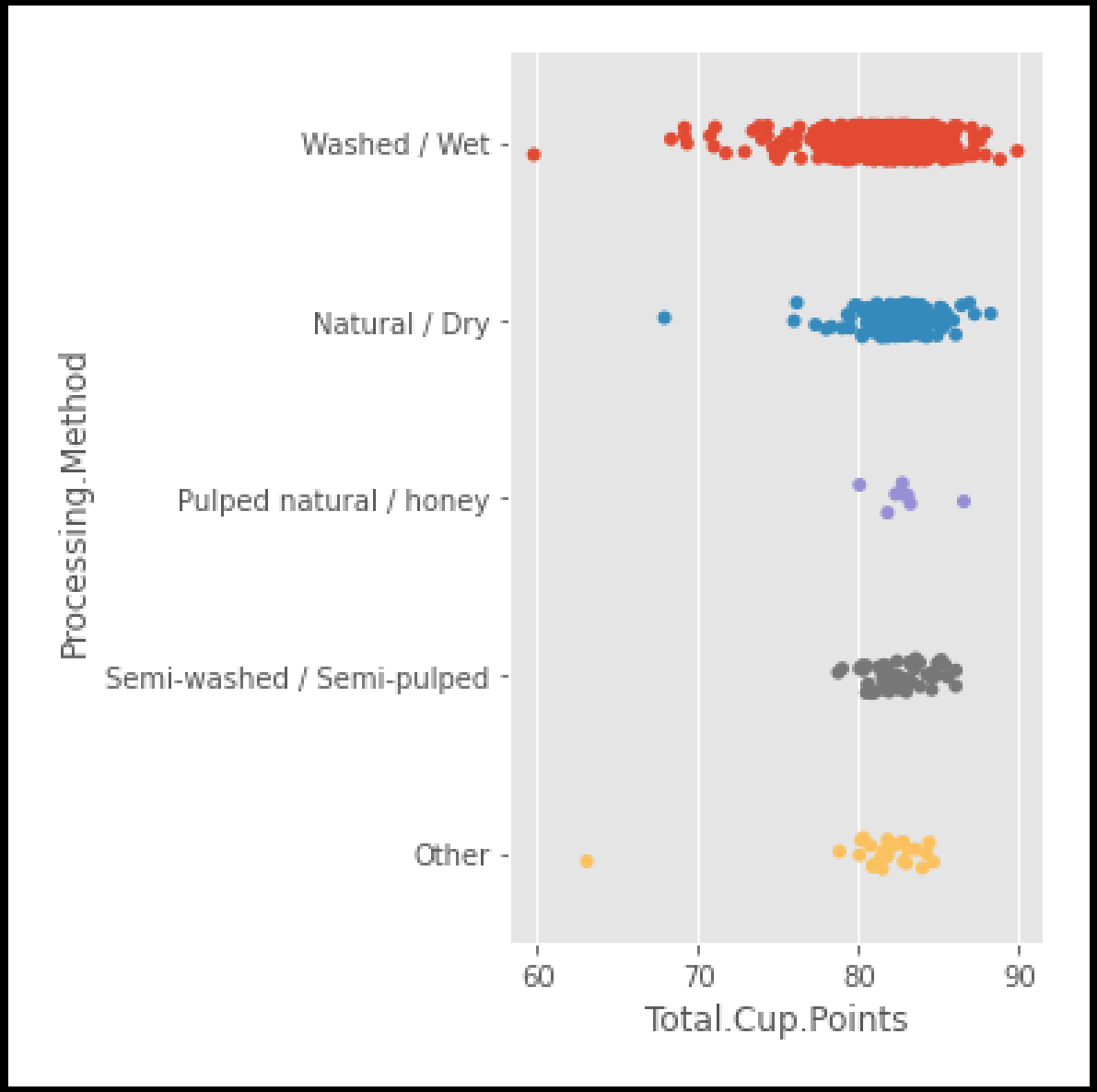
Cup Points by Country

Plotting the Cup Points by Country of Origin is clearly visible that Ethiopia have the best quality coffee in our dataset with a mean of 87 points, followed by United States and then by China



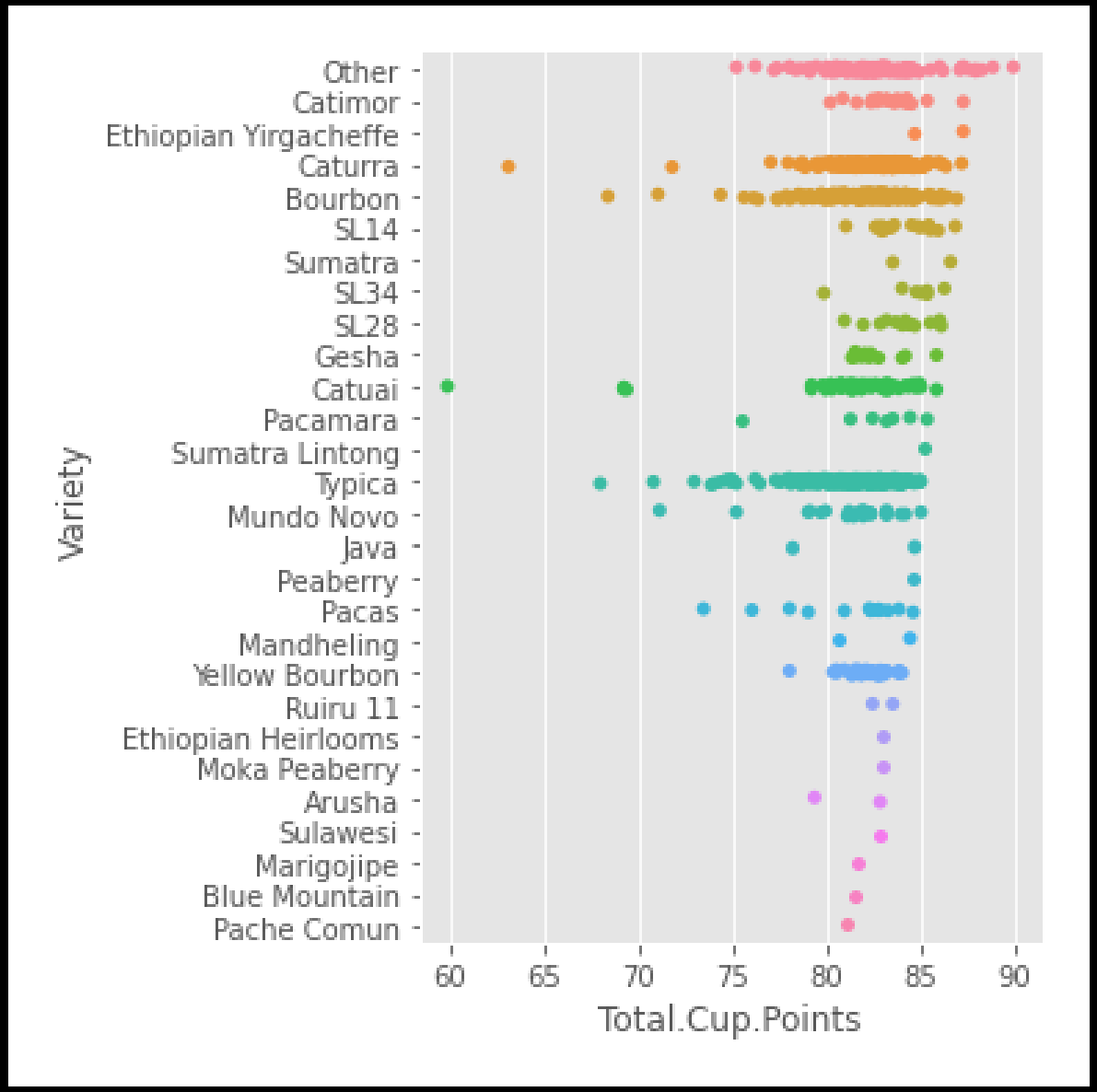
Cup Points by Processing Method

Washed/Wet has the most variety in Cup Points and is the principal Processing Method in the dataset. Also, referring to pulped natural / honey and Semi-washed / Semi-pulped methods, their concentration is around 84 points.

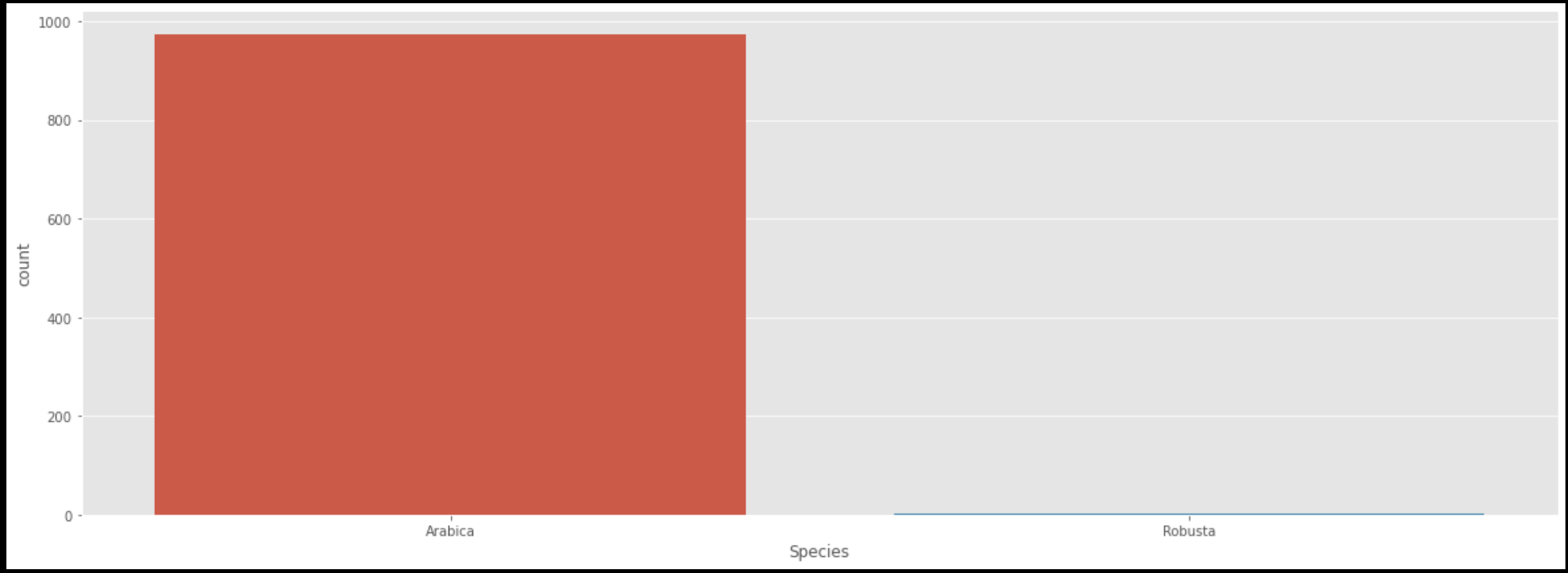


Cup Points by Variety

- Catimor variety have the highest average points.
- Bourbon, Caturra and Typica are the most dispersed in terms of points.
- The 'Others' variety, which we don't have exact information, is in the top with a range of 75 and 90 points



Unbalanced Dataset



There's a high imbalance in the Species (Target) variable. This will be dealt with the use of SMOTE Oversamplin techinque

XGBClassifier

```
model = xgb.XGBClassifier(colsample_bytree=0.95370991686377759,
                          n_estimators=1000,
                          min_child_weight=8.0,
                          reg_alpha = 58.0,
                          reg_lambda=0.37946710635349 27,
                          max_depth=8,
                          gamma=1.9788276329886632)
```

In this case, we will be using XGBoost Classifier and for tuning, Hyperopt was used resulting in the following hyperparameters

XGBClassifier Results

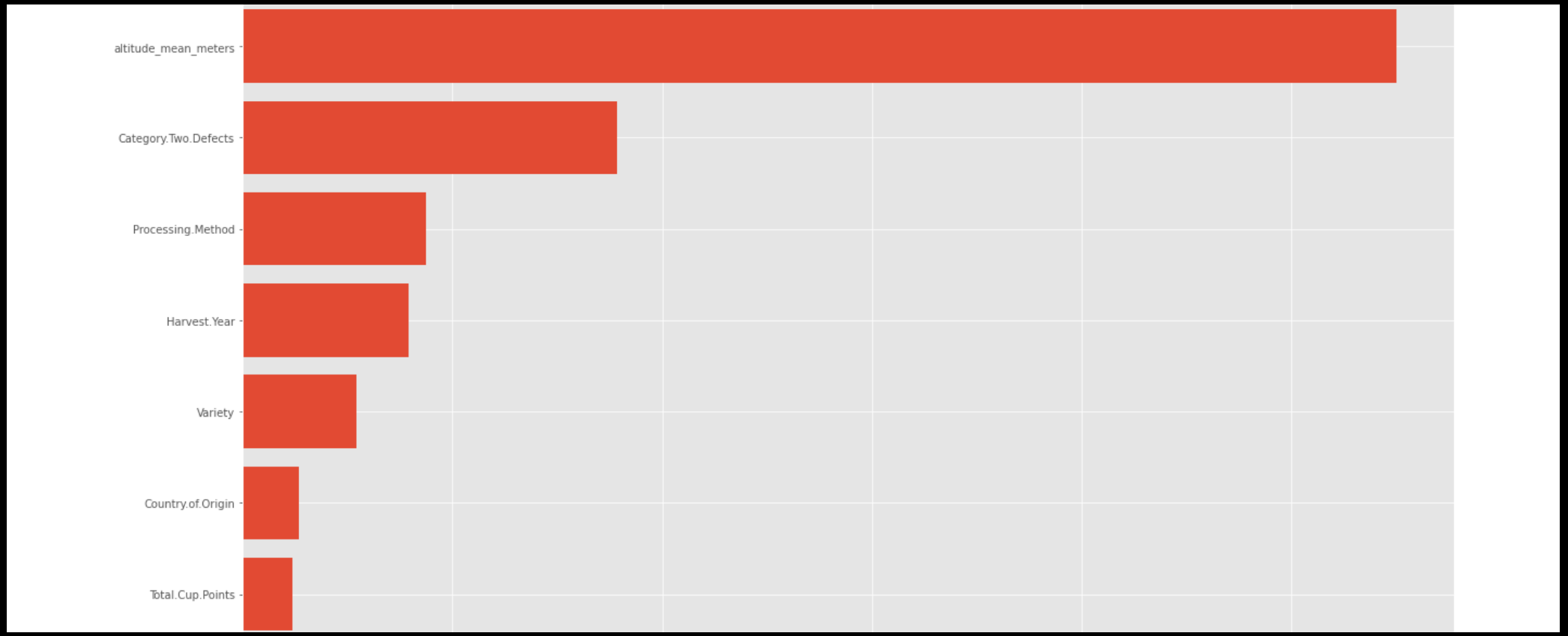
```
1  print('Training accuracy:', model.score(X_train_st, y_train_st))
2  print('Testing accuracy:', model.score(X_test, y_test))
```

✓ 0.2s

```
Training accuracy: 0.9911764705882353
Testing accuracy: 0.9829351535836177
```

We managed to get an accuracy of 98.29% on our validation set, showing that our model performs very good

XGBClassifier Feature Importance



Within the variables with most importance, we can highlight the altitude mean meters, the category two defects and the processing method