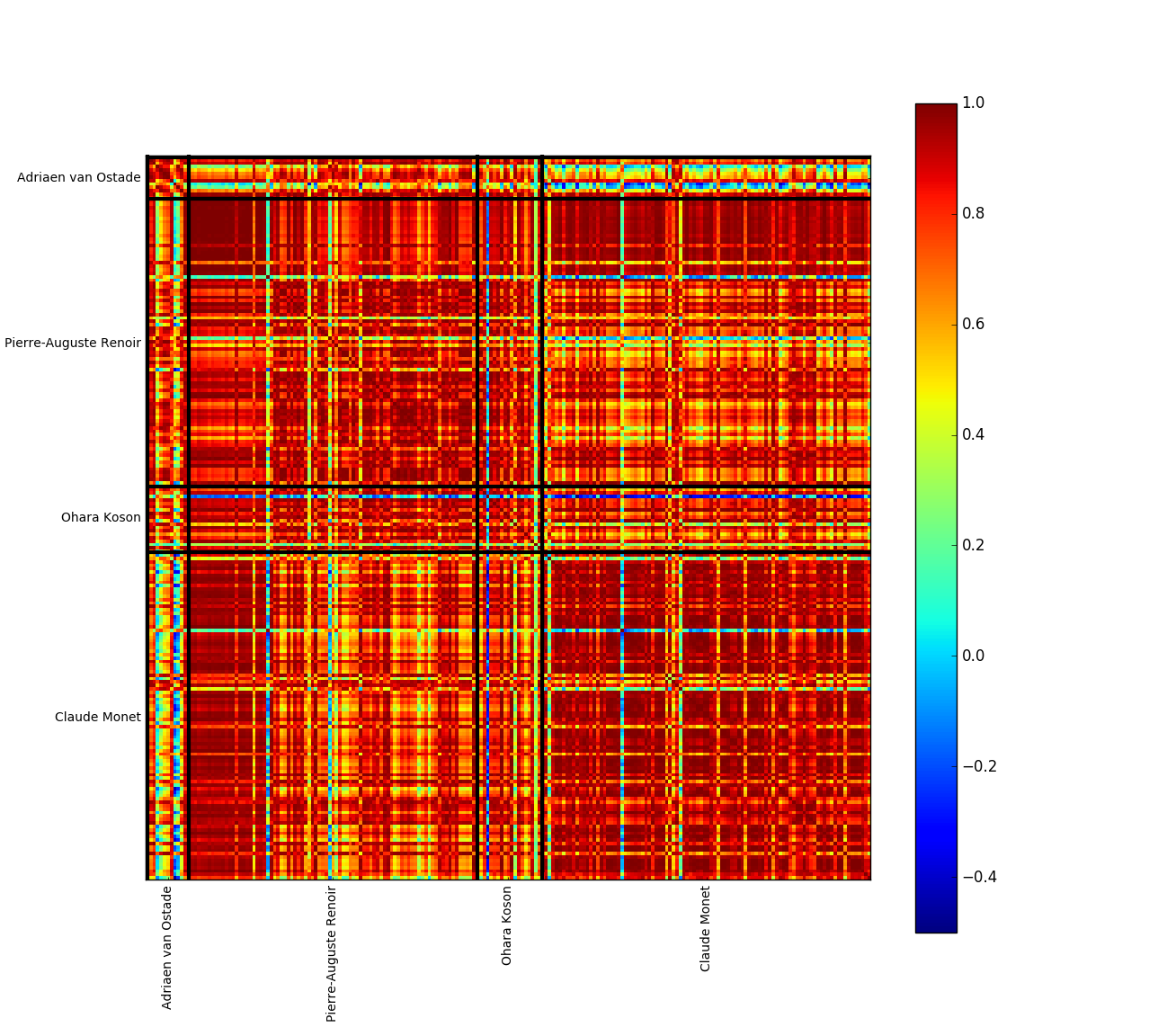
This project is to determine whether a pair of paintings are by the same artists. It is one of Kaggle competitions, online platform where companies and researchers post machine learning problems and data scientists compete for the best models. There are 23,817 images for test dataset and 79,434 images from 1,584 unique painters for training set. The submission includes probabilities of pairwise images having same painters. The evaluation is based on AUC (area under the curve).

Color in a painting is one of key features when we identify its artist. For the initial work, correlation is plotted between painters and their arts’ RGBA: red, green, blue and alpha. ImageStat, a Python Image processing library produces statistics i.e. median pixel level for each RGBA in an image. The four painters which has most paintings in the training sets are highly correlated with their works.



For the beginning of model selection, logistic regression was selected. Each band has become each feature and training data sets were used to train the regression model.