20.1 Capstone Exploratory Data Analysis

This project contains the exploratory data analysis (EDA) for my capstone project in the Professional Certificate in Machine Learning and Artificial Intelligence cource from the University of California, Berkley. The goal of the project is to properly identify a person from their image using ML techniques. It uses the celebrity-face-image-dataset data set from kaggle

The Python Image Library (PIL) was used to clean the data, such that all images are the same size with each having 3 channels of data (RGB). The greyscale images have their channels duplicated. All aspect ratios are corrected to the same size through cropping or blurring. Finally, after reading in the images equalize the data set by dropping any oversampled celebrities.

In conclusion, pruning the data set into something usable was fairly straightforward; the numpy libraries do a good job of sizing and handling images for training. The first model trained on logistic regression is probably not the correct approach given how poorly it performed. A better approach would be a Convolutional Neural Network CNN or transfer learning.