Research Questions

1. How effectively can Principal Component Analysis (PCA) extract meaningful facial features that distinguish individuals?
2. How many principal components are required to achieve accurate classification performance?
3. What challenges arise when using real-world facial images, and how do they affect model accuracy?

Data Description

The dataset used for this project comes from Kaggle's Labeled Faces in the Wild dataset, which includes over 13,000 facial images belonging to more than 5,000 individuals. The dataset is organized such that each person has a folder with multiple facial images. Since many individuals have only a single image available, I will filter the dataset to include only individuals with between 20 and 30 images. This is still a work in progress as I am getting to know how to work with github through google colab. Ultimately, this helps ensure that the classifier has enough examples per person to learn meaningful patterns.

Each image is represented as pixel intensity values, though they will be converted into grayscale for PCA application. The main variable of interest is the identity label, which corresponds to the folder name. A limitation of the dataset is that variations of lighting, pose, and expression introduce noise that PCA may not fully remove. Additionally, restricting the dataset to individuals with sufficient samples reduces coverage of the full dataset but improves model reliability.

Methods

To begin, I am in the process of cleaning the data by isolating only those individuals with more than 20 but less than 30 images. The next step is converting each image into grayscale and flattening the pixel matrix into a single numerical vector. PCA will be performed to reduce dimensionality and extract principal components, which will represent the eigenfaces that highlight key visual features shared among images.

Once the PCA-transfomed features are generated, I will train a classifier using these reduced-dimension features. For evaluation, I will assess classification accuracy, confusion matrices, and visual reconstruction of sample faces using PCA inverse transforms to examine interpretability.

Expected visualizations include:

- Sample eigenfaces
- Comparison between original and reconstructed faces
- Accuracy comparisons across different numbers of principal components

Through these visual and quantitative approaches, I will evaluate how well PCA preserves identity-specific structure and how much information is lost.