Title: Establish Data to Analysis Plan
Group name: City Gorls
Members: Grace Berrien, Elizabeth Breslin, Lilleth Snavely
Group Leader: Elizabeth Breslin
Course: DS4002
Date: 11/09/22

https://github.com/elizabeth-breslin/Coded-Bias

Restated Hypothesis:

If we train a model to classify face images as either male or female with only white faces as opposed to all racial groups, it will be at least 10% less accurate in classification of males and females for all racial groups in comparison to a model trained with images of all racial groups.

Executive Summary:

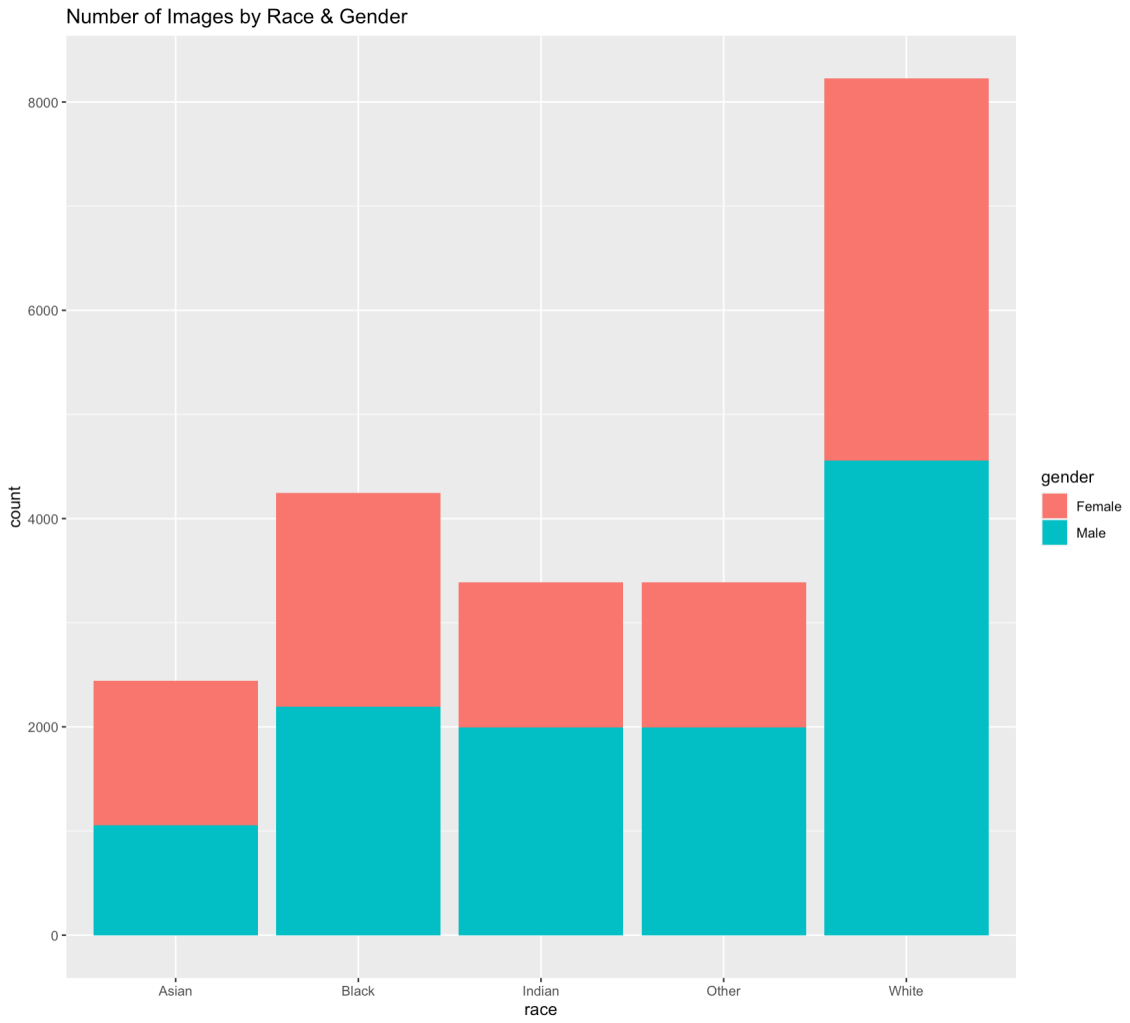
This document explains how we chose to incorporate the UTKFace Large Scale Face Dataset to test our hypothesis. We displayed the proportion of races in each data set to get initial insight into the racial groupings that will be tested in later stages. Our next steps include refining our data into two sets (one with all races and one with only the white race), training our logistic regression model with both sets of photo data, and then testing the models on our test data set.

Data Discovery Findings:

The data set we are using is the UTKFace data set. This contains the faces of over 20,000 people from age 0 to 116. Due to the extremely large size of the data set, we were not able to load all of the data into python. Thus, we decided to use just the population of adults in the data set. So any photo of someone under the age of 18 was excluded from the data. Each photo is labeled in the file name by the age, gender, race, and date and time the photo was taken.

Using a python script we split the data set into ten groups – Asian males, Asian females, Black males, Black females, Indian males, Indian females, White males, White females, Other males, and Other females.

The data files have been loaded into our Github page and separated into different folders by race and gender. The data set identified two genders - male and female - and five race categories - white, black, asian, indian, and other. The plot below shows the number of photos included in the data set for each race and each gender.



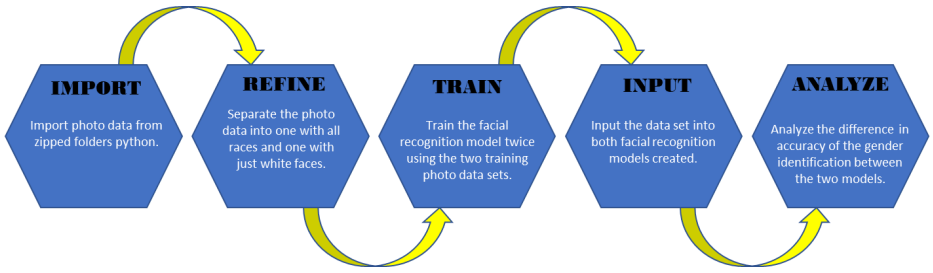
Looking at the plot it is clear that there are significantly more photos of white people than any other race. Therefore, when we train our classifier with photos of people from all racial backgrounds, we will want to ensure we do this with an even amount of every race and not just use all of the images included in the data set.

Analysis Plan:

The first step of our analysis will be refining and organizing our data. Our photo data is from the UTKFace Large Scale Face Dataset. We will extract all the images from the zipped folder so that they can they can be easily imported into Python. Using python, we will take a deeper look into the 20,000+ faces provided in the UTKFace Large Scale Face Dataset. We will be grouping these faces into two groups: one group will contain solely the white race; the other group will contain the white race in addition to every other race included in the dataset. Both groups will need to include equal numbers of images to ensure the classifier is trained and tested with the same amount of images. Furthermore, for the latter group we will want to ensure that all racial groups are included equally. In other words, this group of images will contain the same number of images from each racial group.

Next, for each of the two groups, we will need to split the images into training and testing sets. We will train with ⅔ of each group and test with the other third. Images will be assigned randomly to either testing or training – this can be done with Python. After the training and testing sets are created for each of the two groups, we will examine the sets and make sure they are composed of the same race and gender ratios.

Once the data is organized, we will need to create and train our gender classification Logistic Regression model. We will need to find a logistic regression model that is able to determine the gender of the person in the image. One possible model for this would be, https://www.kaggle.com/code/gulsahdemiryurek/image-classification-with-logistic-regression. First, the training set from the only white image group will be run through the model in order to train the model to recognize gender in the photos. The test set from the second group (the one that includes all races) will then be used to see how well our model that was trained on solely white images classifies images of all races by gender. We will divide the number of correct classifications by total images to get an accuracy percentage. Then, we will retrain the model with the training set from the second group that included all racial image group. The test set from the second group will again be used to see how well our model that was trained on all racial images classifies the images by gender. Again, we will divide the number of correct classifications by total images to get an accuracy percentage. We will then compare the accuracy of the classifier when it was trained on only white images and all racial images and see if performance changed significantly.



In order to answer our initial question about the difference in accuracy between a facial recognition model trained on white faces versus a model trained on various races, we will need to analyze the accuracy score given to each race identification. We will develop a metric that summarizes the accuracy for the entire model, likely the mean or median, using each accuracy score produced. This metric comparison will then be used to determine if the model trained on all races is 10% more accurate than the model trained on all races.

The non-trivial boundaries between each step for the project include ensuring we have adequately created and trained our facial recognition model. Photo data is also much larger than other types of data so the ability to download the hundreds of thousands of images included in our data set onto a computer in order to be input in our model takes a lot of time.

References:

[1] B. Rauenzahn, J. Chung, A. Kaufman, "Facing Bias in Facial Recognition Technology" *The Regulatory Review*, Mar, 20 2021. [Online]. Available: https://www.theregreview.org/2021/03/20/saturday-seminar-facing-bias-in-facial-recognition-technology/#:~:text=According%20to%20the%20researchers%2C%20facial,particularly%20vulnerable%20to%20algorithmic%20bias. [Accessed: Nov. 02 2022]

[2] A. Fawcett, "Understanding racial bias in machine learning algorithms" *DEV,* June, 8 2020. [Online]. Available: https://dev.to/educative/understanding-racial-bias-in-machine-learning-algorithms-4cij [Accessed. Nov, 02 2022]

[3] Darshan M. "How to use logistic regression for image classification?". *Analytics Indian Magazine.* June, 5 2022. [Online]. Available: https://analyticsindiamag.com/how-to-use-logistic-regression-for-image-classification/ [Accessed: Nov, 02 2022]

[4] G. Galario, "Image Classification using Logistic Regression on the AMerican Sign Language MNIST". *Medium*. June, 5 2020. [Online]. Available: https://medium.com/@gryangalario/image-classification-using-logistic-regression-on-the-american-sign-language-mnist-9c6522242ddf [Accessed: Nov, 02 2022]

https://datascience.stackexchange.com/questions/49094/how-to-transform-a-folder-of-images-into-csv-file
https://datatofish.com/move-file-python/

Test Accuracy: 70.18 %
Train Accuracy: 71.27 %