Bias in Facial Classification ML Models

Patrick Connelly Grace Cooper Bhavana Jonnalagadda Carl Klein
Piya (Leo) Ngamkam Dhairya Veera

2023-12-04

Table of contents

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Report PDF and Code Location

A link to download the PDF version of this report, and a link to the Github source code for this report, are both available as icons in the top right corner of this website.

Features of Quarto:

How we should write this report

- See Karkkainen and Joo (2021) , that is an example on how to cite a bibliography.
- Sections/title headings are automatically numbered.
- Any changes you make, make sure to make a comment of your initials at the top of your work (INCLUDING written text) like so:

```
<!-- BJ !-->
Blah blah etc ....

OR
#BJ
r_var <- ...
```

- Make sure to add a unique name to all code cells, and to also enable the following (the quarto way) (In order for a figure to be cross-referenceable, its label must start with the fig- prefix):

- You can then refer to figures like this `@fig-sec1-unique-name` Figure 1
- Format tables doing the following Link here
- Do all your r work initially in your own custom `.rmd` file in this directory, so that it can be copy-pasted over later into the appropriate section (written descriptions/words can go straight into the `.qmd` files though). For example, Bhav's work is in `5000-final/BJ_work.rmd`.

From the report requirements

A 3-5 summary of the paper. It should address the research question, the methods, and the conclusions of your analysis.
"A good recipe for an abstract is: first sentence: specify the general area of the paper and encourage the reader;

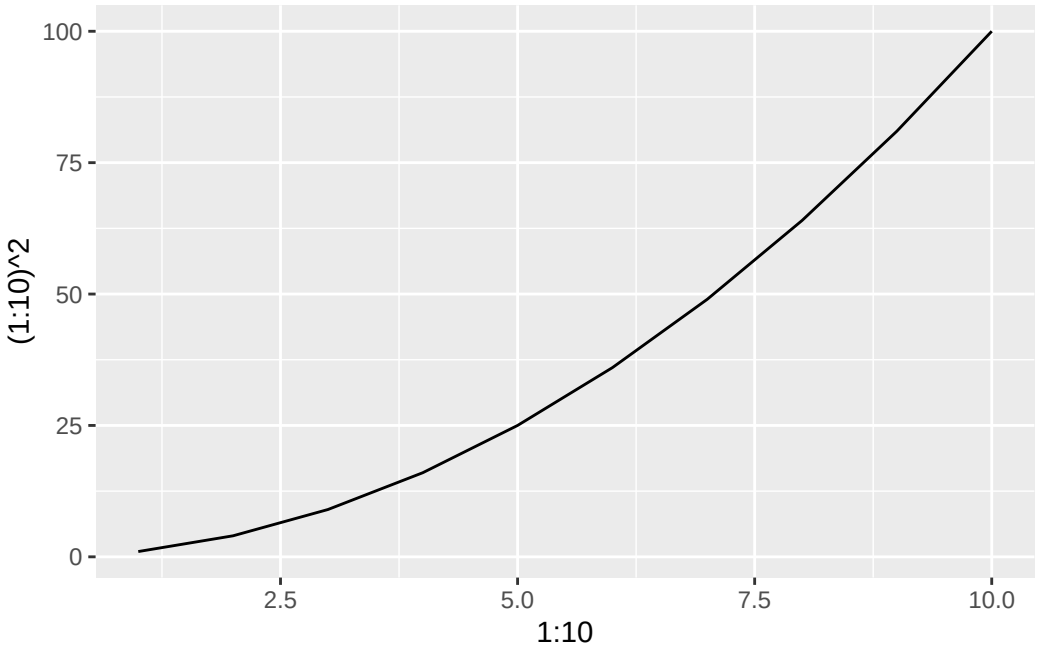


Figure 1: A caption for generated figure

second sentence: specify the dataset and methods at a general level; third sentence: specify the headline result; and a fourth sentence about implications."

1 Introduction

From the report requirements

This section introduces your problem to a **non-expert** audience, describes the context and history of the problem.
For example, if your overall project topic is on Diabetes Prevention and Prediction, then you would use the Introduction to introduce what diabetes is, who it affects, why prevention is important, history on diabetes prevention, etc.
Some questions that you could answer in the introduction:

- What is the "research question"? why is it interesting or worth answering?
- What is the relevant background information for readers to understand your project? Assume that your audience
  is not an expert in the application field.
- Is there any prior research on your topic that might be helpful for the audience?

The goal of the introduction it to capture the audience's interest in your paper. An introduction that starts with "Diabetes kills over 87 thousand people each year and in many cases may be preventable" is more engaging than "This paper is about diabetes prevention".
The introduction should be 2-4 paragraphs long.

2 Data

Pursuant to the study, the team sought out multiple datasets on which we could evaluate the performance of two selected recognition models (Facebook's DeepFace and K. Karkkainen & J. Joo's Fair Face models) to generate performance data and perform statistical analysis on their ability to accurately identify race, age, and gender of a subject in a photograph.

Collectively, we landed on the UTK dataset to perform our evaluation: https://susanqq.github.io/UTKFace/

The dataset has three main sets available for download from the main page: A set of "in-the-wild" faces, which are the raw unprocessed images. The second set is the Aligned & Cropped Faces, which have been cut down to allow facial algorithms to read them more easily. The final file is the Landmarks (68 points) dataset, which contains the major facial landmark points that algorithms use and process to examine the images.

2.1 Exploration of Source Data

(a) Age=6, Gender=F, Race=Indian

(b) Age=38, Gender=M, Race=White

(c) Age=80, Gender=M, Race=Asian

Figure 2.1: Example face images from the UTK dataset ("UTKFace" 2021) with their associated given labels.

> From the report requirements
>
> This section should describe the data you'll be using. Answer **at least all** of the following questions:
>
> - How was the data collected?

The dataset used in this research is a publicly non-commercial available dataset on Github called "UTKFace". The data was collected by the The University of Tennessee, Knoxville. It is specified on its Github page that the images were gathered from the internet. They are likely to be obtained through technique such as web scrapping. The dataset contains more than 20,000 images, representing a highly diversed demographic. However, face images are vary in pose, facial expression, lighting, and resolution.

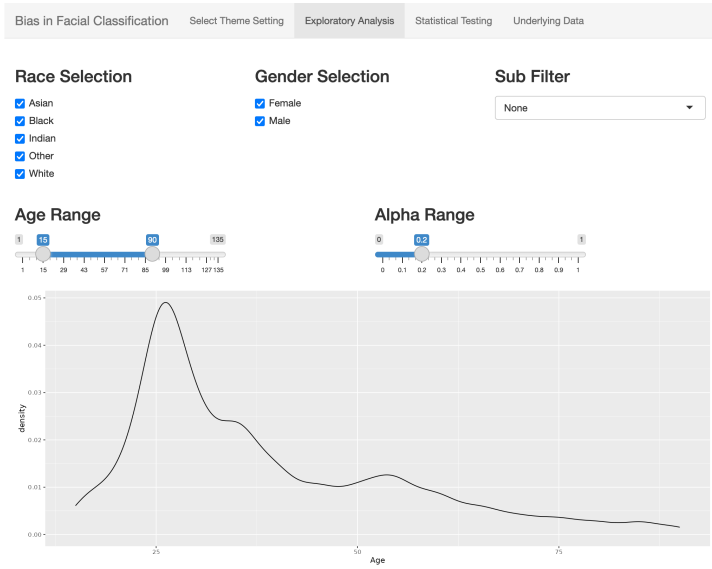


Figure 2.2: Image data EDA

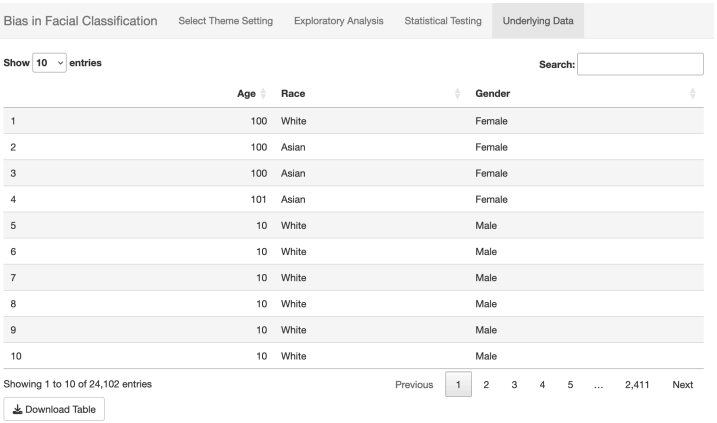


Figure 2.3: Image dataset visualization

Figure 2.4: Screenshots of the interactive figure showcasing the distributions of various data factors in the image dataset, and showcasing the underlying data. To see and interact with this figure, go to the website link

- What are the sources and influences of bias in the data?

The distribution of each demographic groups are not normally distributed. By plotting a distribution of each demographic group, it is evident that the dataset contains an uneven high volume of older White men. While a smaller porportion of female among …(input race)… is present.

- What are the important features (=columns) that you are using in your analysis? What do they mean?

There are three features in the dataset which are essential to our analysis. They are Race, Gender, and Age. Race is categorized into five groups; Asian, Black, Indian, White, and Other. It should be noted by Asian group in this dataset mostly refers to people from East and Southeast Asia. Whereas, Other includes ethnicities such as Hispanic, Latino, and Middle Eastern.
Gender is divided into two groups, either male or female.
Lastly, Age is represented with an integer. This dataset contains people of all ages ranging from 0 to 116.
Feel free to add anything else that you think is necessary for understanding the paper and the context of the problem.

2.2 Data Selection (UTK Dataset) - LN DV

2.2.1 Motivation for the Selection of UTKFace Dataset

In 2018, Joy Buolamwini, a PhD candidate at MIT Media Lab, published a thesis on gender and racial biases in facial recognition in algorithms. In her paper, she tested facial recongition softwares from multiple large technology companies such as Microsoft, IBM, and Amazom on its effectiveness for different demographic groups. Her research led to a surprising conclusion that most AI algorithms offer a substantially less accurate prediction for feminine/female faces, particularly those with dark skin color.

To determine the degree in which bias is still present in modern facial recognition models, a dataset which comprise of face images with high diversity in regards to ethinicty is required. Upon searching, UTKFace came out as the

largest dataset which fit the preferred qualifications.

2.2.2 Data Collection Method

The dataset utilized for this research is UTKFace dataset. It is a publicly available large scale face dataset non-commercial on Github. The dataset was created by Yang Song and Zhifei Zhang, researchers at Adobe and PhD candidates at The University of Tennessee, Knoxville. On its Github page, it is specified that the images were collected from the internet. They appear to be obtained through the application of technique such as web scrapping. The dataset contains more than 20,000 face images, representing a highly diversed demographic. However, face images are vary in pose, facial expression, lighting, and resolution.

2.2.3 Sources and Influences of Bias in the Dataset

- Facial datasets can be extremely hard to categorize correctly, never mind reducing bias overall. Facial features that are androgynous or defer from the average features of the set can often be misrepresented or reported incorrectly. Those with features that make them look younger or older than their actual age may also be difficult for a computer to accurately guess.
- The datasets used for analysis contain solely male/masculine and female/feminine faces. As stated above, the faces are labelled either 0, for male, or 1, for female. There are no gender non-conforming/non-binary/trans faces or people reported in the datasets, which could introduce potential bias. This absence of an entire category of facial features could also result in inaccurate guesses should these faces be added to the data later.
- The datasets do not report nationality or ethnicity. This can introduce inaccuracy in the part of the identification, and it also may identify the face in a racial group that the person identified would consider inaccurate. This is as much a matter of potentially inaccurate data as it is social labels. There is also a level of erasure associated with simply creating a "multi-racial" category, given that it would bin all multiracial faces together with no further consideration. That is to say, there is no ideal solution to the issue at this time. However, it is always worth pointing out potential biases in data, research, and analysis.
- The data given in the UTK dataset is composed purely of people who have their faces on the internet. This introduces a potential sampling bias. Given the topic, it is also likely to come from populations well-versed in technology. This can often exclude rural populations. Thus, the facial data present can be skewed towards urban residents or other characteristics, which can potentially create "lurking variables" that we aren't aware of within the data. This is a common problem that many Anthropological and Sociological studies face when collecting and analyzing data. Being aware of the possibility is often the first, and most crucial, step towards reducing it.

Overall, all of the given potential biases listed above are simply the largest and most easily identified. It is possible that other sources of bias are present in the data that we haven't noticed. And identifying these biases does not mean that the data is not sound, or that any conclusions drawn from it are invalid. It simply indicates that further research should be done and that this data is far from the most complete picture of human facial features and identification.

2.3 Selected Models (LN DV)

2.3.1 FairFace

Developed by researchers at University of California, Los Angeles, FairFace was specifically designed to mitigate gender and racial biases. The model was trained on 100K+ face images of people of various ethnicities with approximately equal stratification across all groups. Beside facial recognition model, FairFace also provided the

dataset which it was trained on. The dataset is immensely popular among facial recognition algorithm developers. Owing to its reputation in bias mitigation, FairFace appears to be a valuable piece for the objective of this research.

2.3.2 DeepFace

DeepFace is a lightweight open-source model developed and used by Meta (Facebook). Since the model is used by one of the largest social media company, it is widely known among developers. Therefore, its popularity prompts us to evaluate its performance. It should be noted that this model of DeepFace is a free open source version. It is highly likely that this version is less advanced than what Meta is actually utilizing. Thus, we should not view the result of this model as a representative of Meta's algorithm.

2.4 Dataset Features

Understanding and selecting the appropriate features for our data is key to success in analysis. Furthermore, understanding feature differences and planning standardization across datasets is key in making sound comparisons and analyses upon the dataset.

2.4.1 Input data set

The input dataset, being UTKFace, provided feature information natively in each filename without additional external data. The features contained therein include the following items for each image's subject. They are defined as follows in the UTKFace Readme:

- "[race] is an integer from 0 to 4, denoting White, Black, Asian, Indian, and Others (like Hispanic, Latino, Middle Eastern)."
- "[gender] is either 0 (male) or 1 (female)"
- "[age] is an integer from 0 to 116, indicating the age"

We processed each image to extract these features from each image and create a table of source information (link to source data file here).

As our work is focused in potential biases in protected classes such as race, gender, and age, the features of UTKFace are sufficient to meet the needs for an input dataset for category prediction in our selected models.

2.4.2 FairFace Outputs

FairFace outputs provided predictions age and race, and two different predictions for race - one based upon their "Fair4" model, and the other based upon their "Fair7" model. In addition to these predictions, the output included scores for each category. With the nature of our planned analyses, the scores are of less import to us in our evaluation.

To examine more in detail on "Fair" and "Fair4" models, the latter provided predictions of race in the following categories: [White, Black, Asian, Indian]. Of key note, the "Fair4" model omitted "Other" categories as listed in the race category for the UTK dataset. However, the "Fair7" model provides predictions across [White, Black, Latino_Hispanic, East Asian, Southeast Asian, Indian, Middle Eastern]. We elected to use the the Fair7 model, and to refactor the output categories to match those of the UTK dataset. Namely, we refactored instances of Middle Eastern and Latino_Hispanic as "Other" and instances of "East Asian" and "Southeast Asian" as "Asian" to match the categories explicitly listed in UTKFace.

Additionally, FairFace only provides a predicted age range as opposed to a specific, single, predicted age as a string. To enable comparison of actual values to the predicted values, we maintained this column as a categorical variable, and split it into a lower and upper bound of predicted age as an integer in the event we require it for our analyses.

With the above considerations in mind, the following output features are of import to the team:

Column Name	Data Type	Significance	Valid Values
name_face_string	String	The name and path of the file upon which FairFace made predictions	[filepath]
race_preds_string	String	The predicted race of the image subject	[White|Black|Latino_Hispanic|East Asian|Southeast Asian|Middle Eastern|Indian]
gender_preds_string	String	The predicted gender of the image subject	[Male|Female]
age_preds_string	String	The predicted age range of the image subject	['0-2'|'3-9'|'10-19'|'20-29'|'30-39'|'40-49'|'50-59'|'60-69'|'70+']

2.4.3 DeepFace Outputs

Default outputs have a wide-range of information for the user. In addition to providing its predictions, DeepFace also provides scores associated with each evaluation on a per class basis (i.e. 92% for Race #1, 3% Race #2, 1% Race #3, and 4% Race #4). For the purpose of our planned analyses, the score features are of less concern to us.

We hone in on the following select features from DeepFace outputs to have the ability to cross-compare between UTKFace, FairFace, and DeepFace:

Column Name	Data Type	Significance	Valid Values
Age	Integer	The predicted age of the image subject	Any Integer
Dominant Gender	String	The predicted gender of the iamge subject	[Man|Woman]
Dominant Race	String	The predicted race of the image subject	[middle eastern|asian|white|latino hispanic|black|indian]

2.4.4 Standardizing model outputs

As can be seen above, there are some key differences between the outputs of both models as well as the source data that we needed to resolve to enable comparison of each dataset to one another. We'll focus on the primary features of age, gender, and race from each dataset.

2.4.4.1 FairFace Output Modifications

We'll discuss FairFace first, as it introduces a requirement for modification to both our input information as well as the outputs for DeepFace.

Age - FairFace only provides a categorical predicted age range as opposed to a specific numeric age. We retain this age format and modify the last category of "70+" to "70-130" to ensure we can capture the gamut of all input and output ages in all datasets.

Gender - No changes to predicted values; use "Male" and "Female"

Race - the source data from UTKFace has 5 categories "White" "Black" "Asian" "Indian" and "Other". Using the defintions from UTKFace, we collapse the output categories of FairFace's Fair7 model as follows:

["Southeast Asian","East Asian"] => "Asian" ["Middle Eastern" , "Latino_Hispanic] =>"Other"

2.4.5 DeepFace Output Modifications

Age Cut the predicted age into bins based upon the same prediction ranges provided by FairFace. If the DeepFace predicted age falls into a range provided by FairFace, provide that as the predicted age range for DeepFace.

Gender: we adjust the DeepFace gender prediction outputs to match that of the source and FairFace data with the following refactoring: "Man" => "Male" "Woman" => "Female"

Race: we adjust the DeepFace race prediction outputs to match that of the source dataset with the following refactoring:

- "white" => "White"
- "black" => "Black"
- "indian" => "Indian"
- "asian" => "Asian"
- ["middle eastern", "latino hispanic"] => "Other"

2.4.6 Source Data Modifications

Age: We cut the predicted age into bins based upon the same prediction ranges provided by FairFace. If the input / source data age falls into a range provided by FairFace, provide that is the source age range for the image subject.

Gender: No changes.

Race: No changes.

2.5 Evaluating Permutations of Inputs and Models for Equitable Evaluation

Aside from the differences in the outputs of each model in terms of age, race, and gender, there are also substantial differences between FairFace and DeepFace in terms of their available settings when attempting to categorize and predict the features associated with an image.

The need for this permutation evaluation rose from some initial scripting and testing of these models on a small sample of images from another facial dataset - the Asian Face Age Dataset (**need citation here**). We immediately grew concerned with DeepFace's performance using default settings (namely, enforcing requirement to detect a face prior to categorization/prediction, and using OpenCV as the default detection backend). Running these initial scripting tests, we encountered a face detection failure rate, and thus a prediction failure rate, in DeepFace of approximately 70%.

We performed further exploratory analysis on both models in light of these facts, and sought some specific permutations of settings to determine what settings may provide the most fair and equitable comparison of the models prior to proceeding to further analysis.

The ultimate goal for us in performing this exploration was to identify the settings for each model that might best increase the likelihood that the model's output would result in a failure to reject our null hypotheses. In lamens terms, our tests sought out the combination of settings that give each model the benefit of the doubt, and for each to deliver the greatest accuracy in their predictions. For simplicity's sake, we leaned solely on the proportion of true positives across each category when compared with the source information to decide which settings to use.

2.5.1 DeepFace Analysis Options

DeepFace has a robust degree of avaialble settings when performing facial categorization and recognition. These include enforcing facial detection prior to classification of an image, as well as 8 different facial detection models to detect a face prior to categorization. The default of these settings is OpenCV detection with detection enabled. Other detection backends include ssd, dlib, mtcnn, retinaface, mediapipe, yolov8, yunet, and fastmtcnn.

In a Python 3.8 environment, attempting to run detections using dlib, fastmtcnn, retinaface, mediapipe, yolov8, and yunet failed to run, or failed to install the appropriate models directly from source during exeuction. Repairing any challenges or issues with the core functionality of DeepFace and FairFace's code is outside the scope of our work, and as such, we have excluded any of these non-functioning models from our settings permutation evaluation.

2.5.2 FairFace Analysis Options

The default script from FairFace provided no options via its command line script to change runtime settings. It uses dlib/resnet34 models for facial detection and image pre-processing, and uses its own Fair4 and Fair7 models for categorization. There are no other options or flags that can be set by a user when processing a batch of images.

We converted the simple script to a class in Python without addressing any feature bugs or errors in the underlying code. This change provided us some additional options when performing the analysis of an input image using FairFace - namely, the ability to analyze and categorize an image with or without facial detection, similar to the functionality of DeepFace. FairFace remains limited in the fact that is only detection model backend is built in dlib, but this change from a script to a class object gave us more options when considering what type of images to use and what settings to use on both models before generating our final dataset for analysis.

2.5.3 Specific Permutations

With the above options in mind, we designed the following permutations for evaluation on a subset of the UTK dataset:

Detection	Detection Model	Image Source
Enabled	FairFace=Dlib; DeepFace=OpenCV	Pre-cropped
Enabled	FairFace=Dlib; DeepFace=OpenCV	In-The-Wild
Enabled	FairFace=Dlib; DeepFace=mtcnn	Pre-cropped
Enabled	FairFace=Dlib; DeepFace=mtcnn	In-The-Wild
Disabled	FairFace,DeepFace=None	Pre-cropped
Disabled	FairFace,DeepFace=None	In-The-Wild

We processed each of the above setting permutations against approximately 9800 images, consisting of images from part 1 of 3 from the UTK datset. Each of the cropped images (cropped_UTK_dataset.csv) and uncropped images (uncropped_UTK_dataset.csv) came from the same underlying subject in each image; the only difference between each image was whether or not it was pre-processed before evaluation by each model. Having the same underlying source subject enables us to perform a direct comparison of results between cropped vs. in-the-wild images, and better support a conlcusion of which settings to use.

pred_model	detection_enabled	detection_model	image_type	all_rate	age_grp_rate	gender_rate	race_
DeepFace	False	None	cropped	0.0724949	0.1601227	0.6667689	0.695
DeepFace	False	None	uncropped	0.0834356	0.1522495	0.7326176	0.645
DeepFace	True	mtcnn	cropped	0.0889571	0.1534765	0.7249489	0.680
DeepFace	True	mtcnn	uncropped	0.1023517	0.1615542	0.7834356	0.666
DeepFace	True	opencv	cropped	0.0267894	0.0765849	0.1887526	0.198
DeepFace	True	opencv	uncropped	0.0806748	0.1455010	0.6619632	0.585
FairFace	False	None	cropped	0.4015337	0.6101227	0.8921268	0.768
FairFace	False	None	uncropped	0.1031697	0.2671779	0.7599182	0.447
FairFace	True	dlib	cropped	0.4015337	0.6101227	0.8921268	0.768
FairFace	True	dlib	uncropped	0.4353783	0.6230061	0.9155419	0.791

Examining the true positive ratios for each case, our team came to the conclusion that the settings that gave both models the best chance for success in correctly predicting the age, gender, and race of subject images are as follows:

- FairFace: enforce facial detection with dlib, and use uncropped images for evaluation
- DeepFace: enforce facial detection with MTCNN detection backend, and use uncropped images for evaluation.

These settings are equitable and make a degree of sense. Using facial detection, specifically-coded for each model, should give each model the ability to isolate the portions of a face necessary for them to make a prediction, as opposed to using a pre-cropped image that could include unneeded information, or excluded needed information.

Having decided on these settings, our team proceeded to run the entirety of the UTK dataset through both DeepFace and FairFace models using a custom coded script MasterScript.py that allowed us to apply multiprocessing across the list of images and evaluate all items in a reasonable amount of time. (cite FairFace here, may also need to reference that we rebuilt their script into a class format).

Due to the resource-intensive design of FairFace, our script enables multiprocessing of FairFace to allow for multiple simultaneous instances of the FairFace class as a pool of worker threads to iterate over all of the source data.

We attempted the same multiprocessing methodology for DeepFace, but encountered issues with silent errors and halting program execution when iterating over all images using DeepFace. To alleviate this challenge, we processed DeepFace in a single-threaded manner, and with smaller portions of the dataset vs. pursuing an all-in-one go execution. We proceeded to store the data for each of these smaller runs in multiple output files to combine once we completed all processing requirements.

2.6 Result Output Format

The following table outlines, after the input and output data modifications, the final format of our data for use in our analyses. This file is stored in the /data folder as MasterDataFrame.csv

Column Name	Definition
img_path	Relative path location of the file within the UTK dataset
file	The filename of each file within the UTK dataset
src_age	The age of the subject in each image from the UTK dataset
src_gender	The gender of the subject in each image from the UTK dataset
src_race	The race of the subject in each image from the UTK datset
src_timestamp	The time at which the image was submitted to the UTK dataset
src_age_grp	The age group (matching the predicted age ranges from the FairFace outputs) for each image in
pred_model	The model used to produce the predicted output (FairFace or DeepFace)
pred_race	The race of the subject in the image, predicted by the given prediction model under the pred_m
pred_gender	The gender of the subject in the image, predicted by the given prediction model under the pred_
pred_age_DF_only	The integer-predicted age by DeepFace of the subject in the image
pred_age_grp	The age group of the subject in the image, predicted by the given prediction model under the pr
pred_age_lower	The integer lower bound of the predicted age group
pred_age_upper	The integer upper bound of the predicted age group

3 Methods

Karkkainen and Joo (2021)

3.1 The Big Picture

- Is bias prevalent in facial recognition machine learning models?
- Can one model be shown to have statistically significant less bias than the other?
- Does one model outperform the other in a statistically significant manner, in all aspects?
- Does one model outperform the other in a statistically significant manner, in certain aspects?
  - This is where we can dive into "conventional" bias
- Are there disparate outcomes (i.e. lower chances of correct classification) for one racial group vs. another?

> **Thoughts on Bias**
>
> We need to be careful how we define and use bias. Statistical bias is essentially error, and we could be crossing our definitions between statistical bias and conventional bias.

3.2 Measuring Performance

> **Note**
>
> This performance section is important in choosing the correct models to ensure data integrity, however for the actual statistical tests, we'll focused on more common statistics like mean and proportion.
> These are my recommendations on how we can look at the additional values:
> For cases in which we reject the statistical null hypothesis, we plan to evaluate the below metrics for the same output category. Should we reject a null hypothesis for, let's say, the proportion of females, given that they are asian, we will examine the accuracy and F-1 scores for the same category. We will consider the following range of values:
>
> - Accuracy < 70: poor/unacceptable performance
> - $70 <$ Accuracy < 79: marginally acceptable performance
> - $80 <$ Accuracy < 89: acceptable performance
> - $90 <$ Accuracy < 99: excellent performance
> - **Do we need anything else regarding precision or recall? and does this all need to be a table?**

There are four main measures of performance when evaluating a model:

- **Accuracy**
- **Precision**

- **Recall**
- **F1-Score**

Each of these performance measures has their own place in evaluating models, however, to begin to explain the differences between these models we should start with concepts of positive and negative outcomes.

- **True Positive:** predicted positive, was actually positive (correct)
- **False Positive:** predicted positive, was actually negative (incorrect)
- **True Negative:** predicted negative, was actually negative (correct)
- **False Negative:** predicted negative, was actually positive (incorrect)

These outcomes can be visualized on a confusion matrix. In the image below, green are correct predictions while red are incorrect predictions.

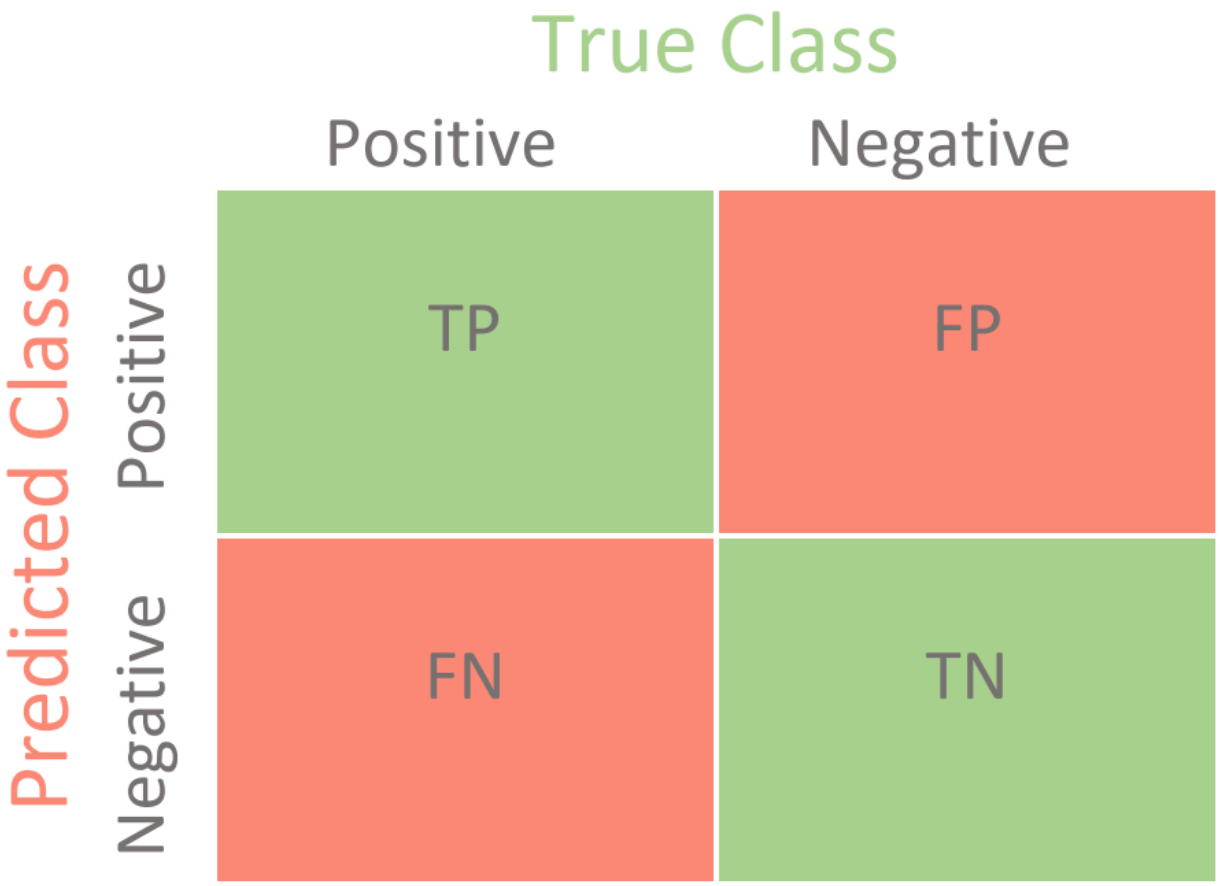


Figure 3.1: confusion_matrix

3.2.1 Accuracy

Accuracy is the ratio of correct predictions to all predictions. In other words, the total of the green squares divided by the entire matrix. This is arguably the most common concept of measuring performance.

$Acccuracy = \frac{TP+TN}{TP+TN+FN}$

3.2.2 Precision

Precision is the ratio of true positives to the total number of positives (true positive + true negative).

$Precision = \frac{TP}{TP+FP}$

3.2.3 Recall

Recall is the ratio of true positives to the number of total correct predictions (true positive + false negative).

$Recall = \frac{TP}{TP+FN}$

3.2.4 F1-Score

F1-Score* is known as the harmonic mean between precision and recall. **Precision** and **Recall** are useful in their own rights, but the f1-Score is useful in the fact it's a balanced combination of both precision and recall.

F1-Score $= \frac{2*Precision*Recall}{Precision+Recall}$

3.3 Hypothesis Testing

Our data consists of three main sets, the source input data, the Fairface output data, and the Deepface output data.

We'll be creating our hypothesis tests by treating the source data as the basis for the original assumptions (our *null hypotheses*), and then using the output from Fairface and Deepface to test for statistically significant differences. Gaining a statistically significant result would allow us to reject our *null hypothesis* in favor of the *alternative hypothesis*. In other words, rejecting the original assumption means there is a statistically large enough difference between the source data and output data, and could indicate a bias in model.

We'll be testing across different subsets contained within the data, as listed below:

3.3.1 Demographics

- Age Group
- Gender
- Race

3.3.2 Demographics' Subgroups

- Age Group (9 groups)
  - 0-2
  - 3-9
  - 10-19
  - 20-29
  - 30-39
  - 40-49
  - 50-59
  - 60-69
  - 70-130

- Gender (2 groups)
  - Female
  - Male
- Race (5 groups)
  - Asian
  - Black
  - Indian
  - Other
  - White

3.3.3 The General Proportion Tests

Our hypothesis tests will be testing different proportions within these subgroups between the source data and the output data.

The general format of our hypothesis tests will be:

$H_0 : p = p_{\text{Source Data Subset}}$

$H_A : p \neq p_{\text{Source Data Subset}}$

With the following test statistic:

$\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}}$

With the p-value being calculated by:

$P(|Z| > \hat{p}|H_0)$

$= P(|Z| > \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}})$,

where

- n: output data subset size
- $\hat{p}$: output data subset proportion
- p: source data subset proportion

NOTE - may be worthwhile to state that we evaluated all cases here, but reduction to specific cases of controlling race and evaluating solely on gender or solely on race is of the most value to us. Performing tight filtering down to, for instance, the proportion of white people given they are 40-49 and female may be too restrictive.

Performing and analyzing calcuations of P(Age|Race) or P(Gender|Race) are of more interest, as I think our research questions are looking at disparate impact to the outcomes for each racial group. ### Notation

Before we list the specific tests, we should introduce some notation.

Let R be race, then $R \in \{Asian, Black, Indian, Other, White\} = \{A, B, I, O, W\}$

Let G be gender, then $G \in \{Female, Male\} = \{F, M\}$

Let A be age, then $A \in \{[0, 2], [3, 9], [10, 19], [20, 29], [30, 39], [40, 49], [50, 59], [60, 69], [70, 130]\} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

Let D be the dataset, then $D \in \{Source, Fairface, Deepface\} = \{D_0, D_f, D_d\}$

3.3.4 More Specific Proportion Tests

Using this notation, we can simplify our nomenclature for testing a certain proportion of an overall demographic.

For example, we can test if the proportion of *Female* in the Fairface output is statistically different than the proportion of *Female* from the source.

Hypothesis Test:

$H_0 : p_F = p_{F|D_0}$

$H_A : p_F \neq p_{F|D_0}$

P-value Calculation:

$P(|Z| > \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}})$,

where

- $p = p_{F|D_0}$: proportion of females from the source data
- $\hat{p} = p_{F|D_f}$: proportion of females from the fairface output
- $n = n_{F \cup M|D_f}$: number of data points in the gender subset form the fairface output

Additionally, we could test for different combinations of subsets within demographics. For instance, if we wanted to test for a statistically significant difference between the proportion of those who *Female*, given that they were *Black*, then we could write a hypothesis test like:

$H_0 : p_{F|B} = p_{F|D_0 \cap B}$

$H_A : p_{F|B} \neq p_{F|D_0 \cap B}$

P-value Calculation:

$P(|Z| > \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}})$,

where

- $p = p_{F|D_0 \cap B}$: proportion of females from the source data, given they were black
- $\hat{p} = p_{F|D_f \cap B}$: proportion of females from the fairface output, given they were black
- $n = n_{F \cup M|D_f \cap B}$: number of data points in the gender subset form the fairface output, given they were black.

These were two specific hypothesis tests, however, we'll be testing many combinations of these parameters and reporting back on any significant findings.

> **i** From the report requirements
>
> Also can be called "Analyses"
> This section might contain several subsections as needed.
>
> - At least one subsection should describe the exploratory data analysis you did.
> - What modifications were necessary to make the dataset ready for analysis? (e.g. dealing with missing values, removing certain rows, replacing/cleaning text values, binning, etc)
> - Describe the analyses you did to answer the question of interest. **Explain why you believe these methods are appropriate.**
> - At least one subsection should describe the exploratory data analysis you did.
> - What modifications were necessary to make the dataset ready for analysis? (e.g. dealing with missing values, removing certain rows, replacing/cleaning text values, binning, etc)

- Describe the analyses you did to answer the question of interest. **Explain why you believe these methods are appropriate.**
- At least one subsection should describe the exploratory data analysis you did.
- What modifications were necessary to make the dataset ready for analysis? (e.g. dealing with missing values, removing certain rows, replacing/cleaning text values, binning, etc)
- Describe the analyses you did to answer the question of interest. **Explain why you believe these methods are appropriate.**

Some methods we learn in this class include distribution comparison, correlation analysis, and hypothesis testing. You are required to include hypothesis tests into the project, but feel free to use additional methods to tell a good story about the data.

4 Results

> **i** From the report requirements
>
> Describe the results of your analysis using visualizations, descriptive statistics, tables and similar.
> Don't focus too much on the implications in this section – that's what the next section is for. Just present the numbers/graphs.

4.1 Model Output

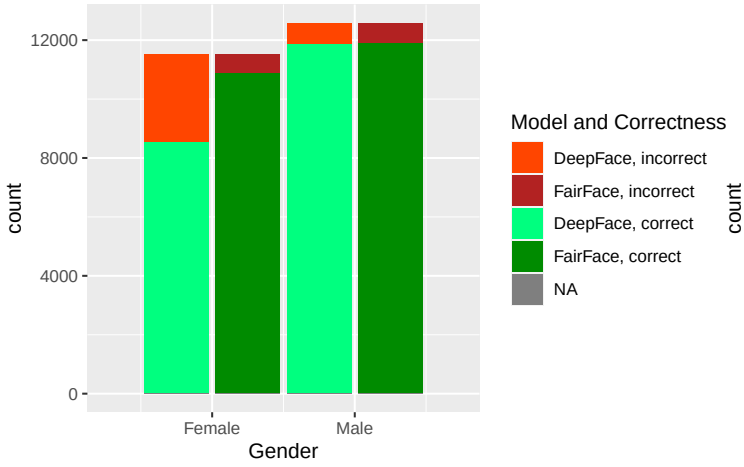
The two models, DeepFace and FairFace, were run on the dataset described previously. In Figure 4.1, one can see the results of the predictions done by each model, by each factor that was considered: age, gender, and race. Note that the total (across correct and incorrect) histogram distributions match the correct (source dataset) distributions of values in each category, so we can see exactly the difference between what was provided and what was predicted, along with how well each model did on each category within each factor.

4.2 Model Performance

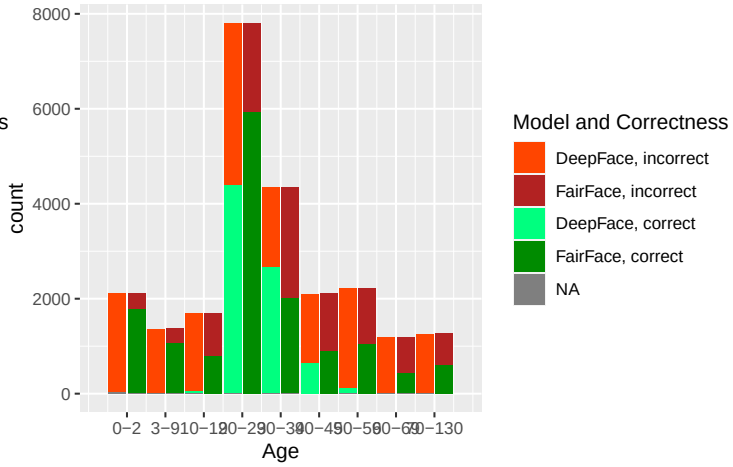
4.2.1 TODO: Remove

Table 4.1: **?(caption)**

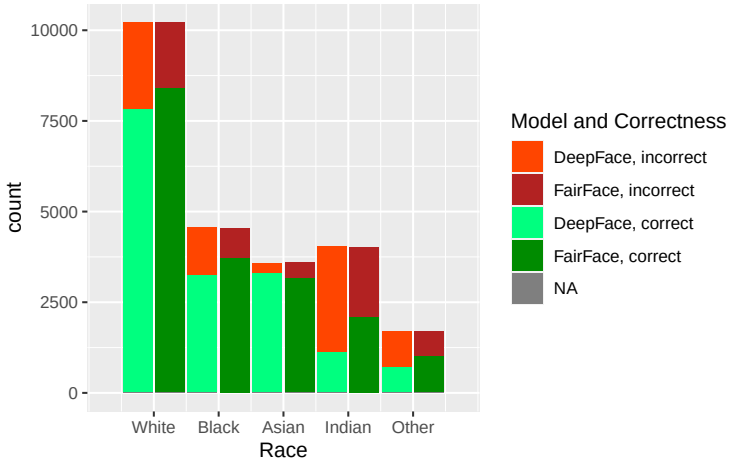
```
# A tibble: 432 x 14
   test_prop test_age test_gender test_race source_n source_prop fairface_n
   <chr>     <chr>    <chr>       <chr>        <dbl>       <dbl>      <dbl>
 1 age_bins  70-130   Female      White         4674      0.105        4539
 2 genders   70-130   Female      White          887      0.556         397
 3 races     70-130   Female      White          638      0.773         408
 4 age_bins  70-130   Female      Asian         1942      0.0340       1895
 5 genders   70-130   Female      Asian          183      0.361         107
 6 races     70-130   Female      Asian          638      0.103         408
 7 age_bins  70-130   Female      Black         2221      0.0221       1851
 8 genders   70-130   Female      Black          136      0.360          30
 9 races     70-130   Female      Black          638      0.0768        408
10 age_bins  70-130   Female      Indian        1742      0.0155       1100
# i 422 more rows
# i 7 more variables: fairface_prop <dbl>, fairface_p_value <dbl>,
#   fairface_power <dbl>, deepface_n <dbl>, deepface_prop <dbl>,
#   deepface_p_value <dbl>, deepface_power <dbl>
```



(a) Gender predictions



(b) Age predictions



(c) Race predictions

Figure 4.1: Histograms of the output from DeepFace and FairFace, with correct vs incorrect values colored. Note that the distributions match the correct (source dataset) distributions.

4.3 Hypothesis Testing

4.3.1 TODO: Remove

Test cateogry	Null Proportion	FairFace Proportion	FairFace P-Value	DeepFace Proportion	DeepFace P-Value
0-2	0.0880010	0.0762941	0.0000029	0.0000000	0
3-9	0.0568832	0.0664564	0.0000125	0.0000000	0
10-19	0.0697867	0.0606451	0.0000482	0.0206211	0
20-29	0.3238735	0.3480138	0.0000000	0.3987029	0
30-39	0.1802755	0.1762899	0.2530466	0.4047312	0
40-49	0.0872542	0.1023619	0.0000000	0.1467177	0
50-59	0.0923160	0.0930638	0.7771406	0.0268158	0
60-69	0.0490831	0.0490640	0.9922575	0.0024113	0
70-130	0.0525268	0.0278112	0.0000000	0.0000000	0

Test cateogry	Null Proportion	FairFace Proportion	FairFace P-Value	DeepFace Proportion	DeepFace P-Value
White	0.4240727	0.3854136	0.0000000	0.3757120	0
Black	0.1891129	0.1652484	0.0000000	0.1479233	0
Asian	0.1487843	0.1448259	0.2195583	0.2408847	0
Indian	0.1670816	0.1030675	0.0000000	0.0611150	0
Other	0.0709485	0.2014445	0.0000000	0.1743649	0

Test cateogry	Null Proportion	FairFace Proportion	FairFace P-Value	DeepFace Proportion	DeepFace P-Value
Female	0.4780101	0.4797642	0.6999259	0.3833202	0
Male	0.5219899	0.5202358	0.6999259	0.6166798	0

Test Category	Test Condition	Null Proportion	FairFace Proportion	FairFace P-Value	DeepFace Proportion	DeepFace P-Value
Female	White	0.4572938	0.4888530	0.0000104	0.4355428	0.0024474
Female	Asian	0.5415505	0.5431356	0.8935647	0.3879876	0.0000000
Female	Black	0.4872751	0.4649586	0.0394280	0.2405846	0.0000000
Female	Indian	0.4325801	0.4430125	0.4097239	0.3496599	0.0000000
Female	Other	0.5508772	0.4477643	0.0000000	0.3972341	0.0000000
Male	White	0.5427062	0.5111470	0.0000104	0.5644572	0.0024474
Male	Asian	0.4584495	0.4568644	0.8935647	0.6120124	0.0000000
Male	Black	0.5127249	0.5350414	0.0394280	0.7594154	0.0000000
Male	Indian	0.5674199	0.5569875	0.4097239	0.6503401	0.0000000
Male	Other	0.4491228	0.5522357	0.0000000	0.6027659	0.0000000

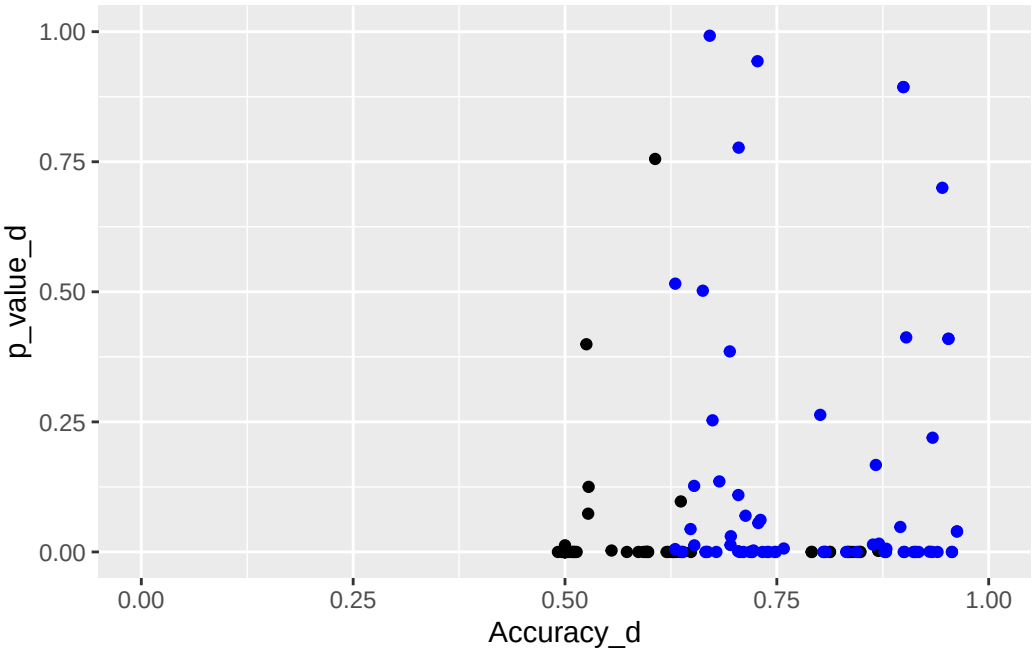
4.3.2 Updated Table Version with Data from Carl, Bhav

4.3.2.1 TODO: Remove

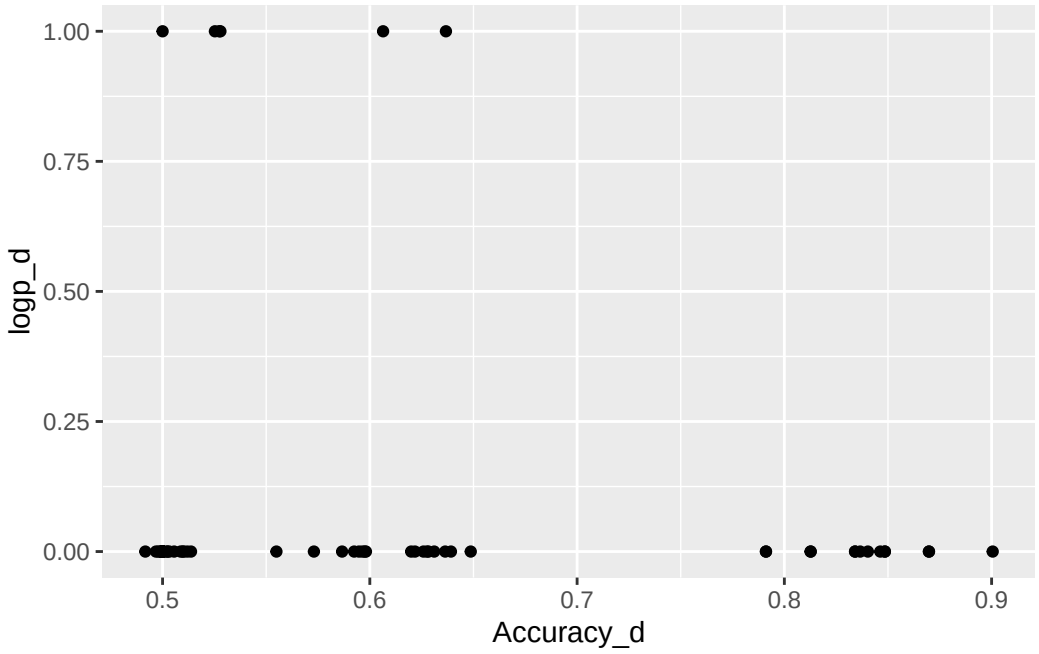
Race	Category	F1_d	F1_f	Accuracy_d	Accuracy_f	prop_d	prop_f	p_value_d	p_valu
All	0-2	NA	0.8959757	0.5000000	0.9172888	0.0000000	0.0762941	0.0000000	0.0000
All	3-9	NA	0.7176035	0.5000000	0.8772778	0.0000000	0.0664564	0.0000000	0.0000
All	10-19	0.0478601	0.5052498	0.5055825	0.7211461	0.0206211	0.0606451	0.0000000	0.0000
All	20-29	0.5054326	0.7332922	0.6217793	0.8050592	0.3987029	0.3480138	0.0000000	0.0000
All	30-39	0.3786318	0.4670003	0.6275447	0.6741504	0.4047312	0.1762899	0.0000000	0.2530
All	40-49	0.2276278	0.3943970	0.5866155	0.6786302	0.1467177	0.1023619	0.0000000	0.0000
All	50-59	0.0801673	0.4633983	0.5137145	0.7049843	0.0268158	0.0930638	0.0000000	0.7771
All	60-69	0.0016129	0.3739425	0.4991769	0.6708204	0.0024113	0.0490640	0.0000000	0.9922
All	70-130	NA	0.6270661	0.5000000	0.7383514	0.0000000	0.0278112	0.0000000	0.0000
All	White	0.8095461	0.8610399	0.8365916	0.8788455	0.3757120	0.3854136	0.0000000	0.0000
All	Black	0.7964994	0.8684858	0.8462797	0.8997692	0.1479233	0.1652484	0.0000000	0.0000
All	Asian	0.7038975	0.8948932	0.9005150	0.9338128	0.2408847	0.1448259	0.0000000	0.2195
All	Indian	0.4092481	0.6402458	0.6310597	0.7488102	0.0611150	0.1030675	0.0000000	0.0000
All	Other	0.2389021	0.3087473	0.6283106	0.7105889	0.1743649	0.2014445	0.0000000	0.0000
All	Female	0.8197702	0.9429153	0.8402892	0.9453080	0.3833202	0.4797642	0.0000000	0.6999
All	Male	NA	NA	NA	NA	0.6166798	0.5202358	0.0000000	0.6999
White	0-2	NA	0.9039010	0.5000000	0.9334307	0.0000000	0.0737749	0.0000000	0.0000
White	3-9	NA	0.7503392	0.5000000	0.8668432	0.0000000	0.0770059	0.0000000	0.1671
White	10-19	0.0634648	0.5638298	0.5102546	0.7330315	0.0163771	0.0693592	0.0000000	0.0000
White	20-29	0.4256326	0.6697460	0.6363584	0.8072440	0.3311940	0.2455574	0.0000000	0.0000
White	30-39	0.3884765	0.4731553	0.6486930	0.6821608	0.4022353	0.1628433	0.0000000	0.1355
White	40-49	0.2224248	0.3847156	0.5730236	0.6683039	0.2100255	0.1186861	0.0000000	0.0001
White	50-59	0.0890599	0.4832502	0.5086819	0.7046059	0.0376231	0.1419494	0.0000000	0.1093
White	60-69	NaN	0.3545817	0.4978778	0.6482054	0.0025451	0.0680668	0.0000000	0.0439
White	70-130	NA	0.6342183	0.5000000	0.7403000	0.0000000	0.0427571	0.0000000	0.0000
White	Male	0.8892356	0.9595281	0.8697687	0.9566327	0.5644572	0.5111470	0.0024474	0.0000
White	Female	0.8556585	0.9526238	0.8697687	0.9566327	0.4355428	0.4888530	0.0024474	0.0000
Asian	0-2	NA	0.9164589	0.5000000	0.9301909	0.0000000	0.2029235	0.0000000	0.0004
Asian	3-9	NA	0.7140255	0.5000000	0.9111790	0.0000000	0.0928633	0.0000000	0.0000
Asian	10-19	0.0395257	0.3798450	0.5023622	0.6944844	0.0457370	0.0366867	0.0022769	0.3854
Asian	20-29	0.5572885	0.8557951	0.5947638	0.8792496	0.5044874	0.4379478	0.0000000	0.0060
Asian	30-39	0.2992611	0.5069357	0.6258649	0.7042053	0.3206766	0.0988822	0.0000000	0.0014
Asian	40-49	0.1520190	0.3320463	0.5924407	0.6626378	0.0995858	0.0369733	0.0000000	0.5021
Asian	50-59	0.0898876	0.4608696	0.5273304	0.7272357	0.0250259	0.0315277	0.0735421	0.9432
Asian	60-69	NaN	0.4141414	0.4988555	0.7581786	0.0044874	0.0315277	0.0000000	0.0066
Asian	70-130	NA	0.7441860	0.5000000	0.8051278	0.0000000	0.0306678	0.0000000	0.0000
Asian	Male	0.7940330	0.8914286	0.7911354	0.8993780	0.6120124	0.4568644	0.0000000	0.8935
Asian	Female	0.7630232	0.9058670	0.7911354	0.8993780	0.3879876	0.5431356	0.0000000	0.8935
Black	0-2	NA	0.8854962	0.5000000	0.9026663	0.0000000	0.0180859	0.0000000	0.4124
Black	3-9	NA	0.7400881	0.5000000	0.8957332	0.0000000	0.0298920	0.0000000	0.0480
Black	10-19	0.0164609	0.3784787	0.5032694	0.6953003	0.0000000	0.0635519	0.0000000	0.0133
Black	20-29	0.5880567	0.6875152	0.6367203	0.7184594	0.4094997	0.4687265	0.0970558	0.0001
Black	30-39	0.4413203	0.4518681	0.5975793	0.6299581	0.4724564	0.2398895	0.0000000	0.0054
Black	40-49	0.1827542	0.3260274	0.5549318	0.6301481	0.1017426	0.0864104	0.0027629	0.5156
Black	50-59	0.0549451	0.3565062	0.5097384	0.6523810	0.0160202	0.0557649	0.0000000	0.1269
Black	60-69	0.0104712	0.3508772	0.5019317	0.6526201	0.0002811	0.0301432	0.0000000	0.0126
Black	70-130	NA	0.4086022	0.5000000	0.6383490	0.0000000	0.0075358	0.0000000	0.0000
Black	Male	0.8471279	0.9637681	0.8126869	0.9625782	0.7594154	0.5350414	0.0000000	0.0394
Black	Female	0.7724665	0.9615732	0.8126869	0.9625782	0.2405846	0.4649586	0.0000000	0.0394

```
# A tibble: 71 x 9
   Race  Category source_prop   n_d  prop_d p_value_d   n_f prop_f p_value_f
   <chr> <chr>          <dbl> <dbl>   <dbl>     <dbl> <dbl>  <dbl>     <dbl>
 1 All   0-2           0.0880 24053 0       0         24091 0.0763  2.87e- 6
 2 All   3-9           0.0569 24053 0       3.29e-293 24091 0.0665  1.25e- 5
 3 All   10-19         0.0698 24053 0.0206  1.43e-148 24091 0.0606  4.82e- 5
 4 All   20-29         0.324  24053 0.399   1.70e- 65 24091 0.348   2.02e- 8
 5 All   30-39         0.180  24053 0.405   0         24091 0.176   2.53e- 1
 6 All   40-49         0.0873 24053 0.147   1.23e- 91 24091 0.102   1.51e- 8
 7 All   50-59         0.0923 24053 0.0268  2.04e-202 24091 0.0931  7.77e- 1
 8 All   60-69         0.0491 24053 0.00241 3.87e-229 24091 0.0491  9.92e- 1
 9 All   70-130        0.0525 24053 0       3.87e-284 24091 0.0278  2.05e-43
10 All   White         0.424  24053 0.376   2.44e- 27 24091 0.385   5.38e-18
# i 61 more rows
```

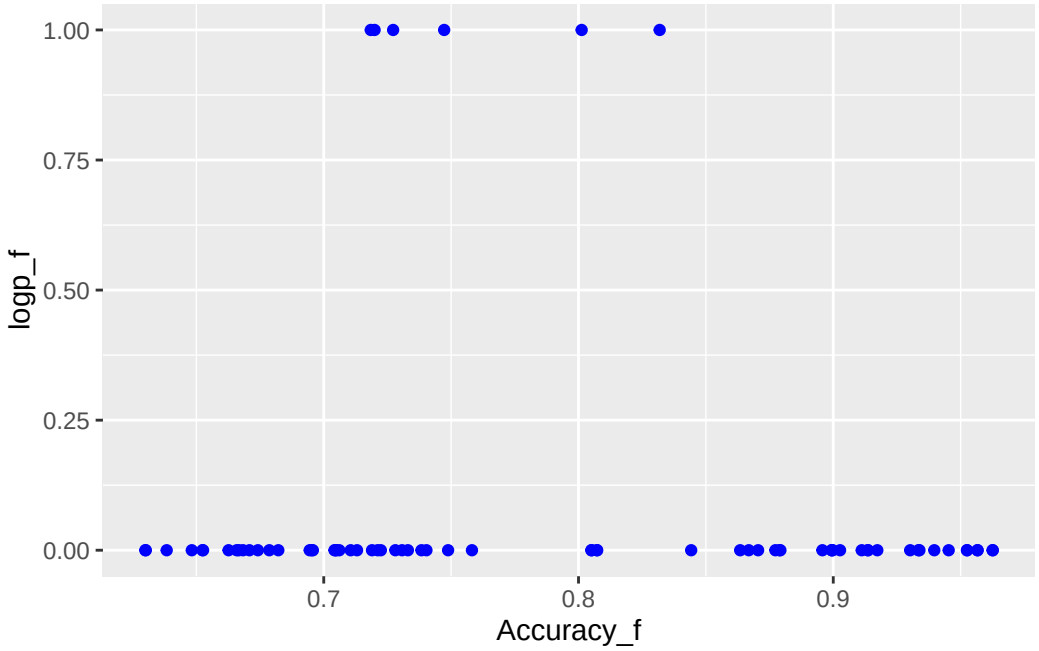
```
Warning: Removed 1 rows containing missing values (`geom_point()`).
Removed 1 rows containing missing values (`geom_point()`).
```




```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```




```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```




```
# A tibble: 71 x 15
   Race  Category     F1_d  F1_f Accuracy_d Accuracy_f  prop_d prop_f p_value_d
   <chr> <chr>       <dbl> <dbl>      <dbl>      <dbl>   <dbl>  <dbl>     <dbl>
 1 All   0-2        NA     0.896      0.5        0.917 0       0.0763 0
 2 All   3-9        NA     0.718      0.5        0.877 0       0.0665 3.29e-293
 3 All   10-19       0.0479 0.505      0.506      0.721 0.0206  0.0606 1.43e-148
 4 All   20-29       0.505  0.733      0.622      0.805 0.399   0.348  1.70e- 65
 5 All   30-39       0.379  0.467      0.628      0.674 0.405   0.176  0
 6 All   40-49       0.228  0.394      0.587      0.679 0.147   0.102  1.23e- 91
 7 All   50-59       0.0802 0.463      0.514      0.705 0.0268  0.0931 2.04e-202
 8 All   60-69       0.00161 0.374     0.499      0.671 0.00241 0.0491 3.87e-229
 9 All   70-130     NA     0.627      0.5        0.738 0       0.0278 3.87e-284
```

```
10 All   White     0.810   0.861      0.837      0.879 0.376   0.385  2.44e- 27
# i 61 more rows
# i 6 more variables: p_value_f <dbl>, power_d <dbl>, power_f <dbl>, n_f <dbl>,
#   source_prop <dbl>, n_d <dbl>
```

4.4 Model Performance, Hypothesis Testing

For each category and model, we calculate the F1 score, accuracy, and p-value, as described in section 3. The results are summarized in **?@tbl-perf-pvalue**. Cell values are colored according to the strength of the metric.

We also specifically looked at the performance metrics of the models, when controlled for specific race groups;

4.4.1 TODO

- Add color key
- Color p-values based on sig level = 99.7%
- Better description of signifiance of numerical values of Accuracy, F1 score
- Add line plot of F1, Accuracy, p-value OR correlation matrix
- Make table caption work?

Static table:

- Values where we FAIL to reject null hypothesis

4.4.2 p-value Critical Values

From the previous table, we extract and highlight key values; namely, the p-values where we fail to reject the null hypothesis at a significance level of 99.7%, displayed in Table 4.2.

Table 4.2: TODO

	p-value	
Test Condition, Category	DeepFace	FairFace
Race: Other		
20-29	0.3992	0.2635
50-59	0.1251	0.0000
Race: Indian		
10-19	0.0000	0.0696
40-49	0.7554	0.0000
50-59	0.0000	0.0302
60-69	0.0000	0.0555
70-130	0.0000	0.0616
Male	0.0000	0.4097
Female	0.0000	0.4097
Race: Asian		
10-19	0.0023	0.3854
40-49	0.0000	0.5021
50-59	0.0735	0.9433
Male	0.0000	0.8936

Female	0.0000	0.8936
Race: Black		
0-2	0.0000	0.4124
3-9	0.0000	0.0480
20-29	0.0971	0.0001
40-49	0.0028	0.5157
50-59	0.0000	0.1270
Male	0.0000	0.0394
Female	0.0000	0.0394
Race: White		
3-9	0.0000	0.1672
30-39	0.0000	0.1356
50-59	0.0000	0.1094
60-69	0.0000	0.0440
No Test Condition		
30-39	0.0000	0.2530
50-59	0.0000	0.7771
60-69	0.0000	0.9923
Asian	0.0000	0.2196
Female	0.0000	0.6999
Male	0.0000	0.6999

4.4.3 Statistical Power

$$\beta = P\left(\left|\frac{\sqrt{n}\cdot\hat{p}-p_a}{\sqrt{p_a(1-p_a)}}\right| \geq \frac{\sqrt{n}\cdot p_0 - p_a}{\sqrt{p_0(1-p_0)}}\right)$$

Our selected level of significance is 99.7% (3-sigma). Type-II error is denoted by β above, and Power will be $1-\beta$

With p_0 being our *assumed* population proportion (from the source dataset and what we used in our tests), p_a being the *actual* population proportion (from one or more of the below methods), n being the number of predicted members of a racial group (i.e. "Indian"),

- For Gender - assume that sex at birth is a bernoulli trial, over time, the proportion for both genders should be 0.5
- For age groups - assume that age has a true normal distribution. Each race may have different means and standard deviations for their distribution of age, but still adhere to a normal distribution. The "population" proportions may be a bit more challenging to calculate, but under this framework, we may be able to get there.
  - May be able to get via bootstrapping the source dataset, average age by race - I think that's what we did in our last project?
  - Could look at external data? May not have time to look through everything.

$$\frac{\sqrt{n_M}\cdot(\bar{p}_M - p_S)}{\sqrt{p_S\cdot(1-P_S)}}$$

5 Conclusions

From the report requirements

- Summarize what the paper has done, and discuss the implications of your Results.
- Explicitly connect the results to the research question.
- Discuss how you would you extend this research

Like the introduction, this section should be written with a **non-expert** in mind. A person should be able to read Introduction+Conclusion and get a rough idea of the meaning and significance of your paper

References

Karkkainen, Kimmo, and Jungseock Joo. 2021. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1548–58.

"UTKFace." 2021. *UTKFace*. https://susanqq.github.io/UTKFace.