Facial Aging Report

By : Youssef Kamil Adly

Content

Problem Description:

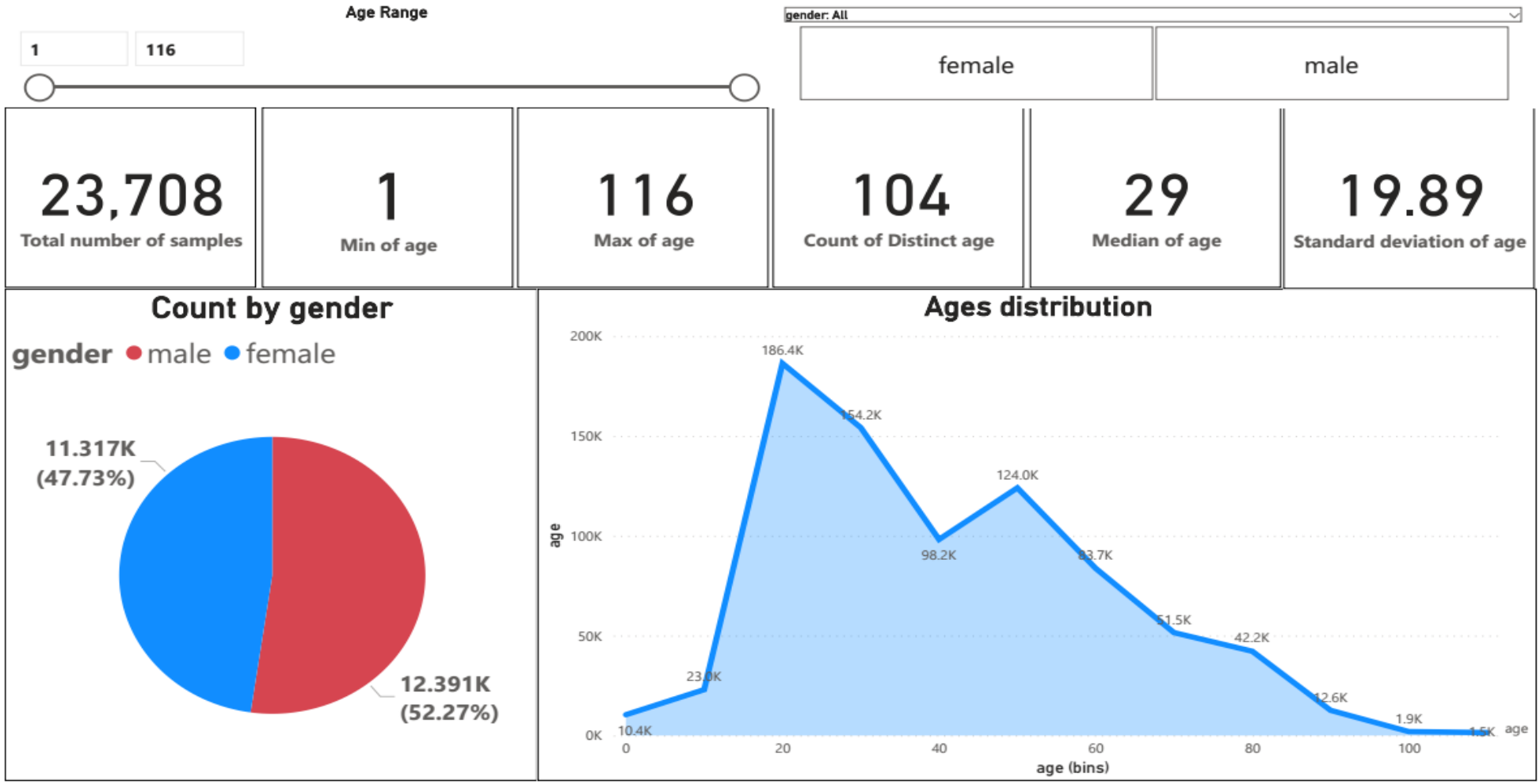
- Facial Aging problem is a problem of predicting the age of a person based on the images of their face.
- The goal of this report is to explore all the datasets available for the facial aging problem and to find the best dataset to use for the analysis.
- The most essential "metrics" to choose the best dataset are (Ages Distribution, Total Number Of Samples, Gender Distribution, Usage constrains, Capturing state and availability).
- I have used python and power BI to preprocess and analyze the Datasets.
- I have used the dataset's Metadata to analyze the dataset, some of them needed to be cleaned by removing missing data or removing the outliers.
- An interactive PowerBi file is attached with the mail.

Datasets Table:

No.	Dataset	Link
1	UTK	https://susanqq.github.io/UTKFace/
2	IMDB	https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/
3	WIKI	https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/
4	AFAD	https://afad-dataset.github.io/
5	CACD	https://bcsiriuschen.github.io/CARC/
6	MORPH	https://uncw.edu/oic/tech/morph.html
7	FGNET	https://yanweifu.github.io/FG_NET_data/
8	Adience	https://talhassner.github.io/home/projects/Adience/Adience-data.html
9	KANFace	https://sites.google.com/view/kanface-dataset

UTK:

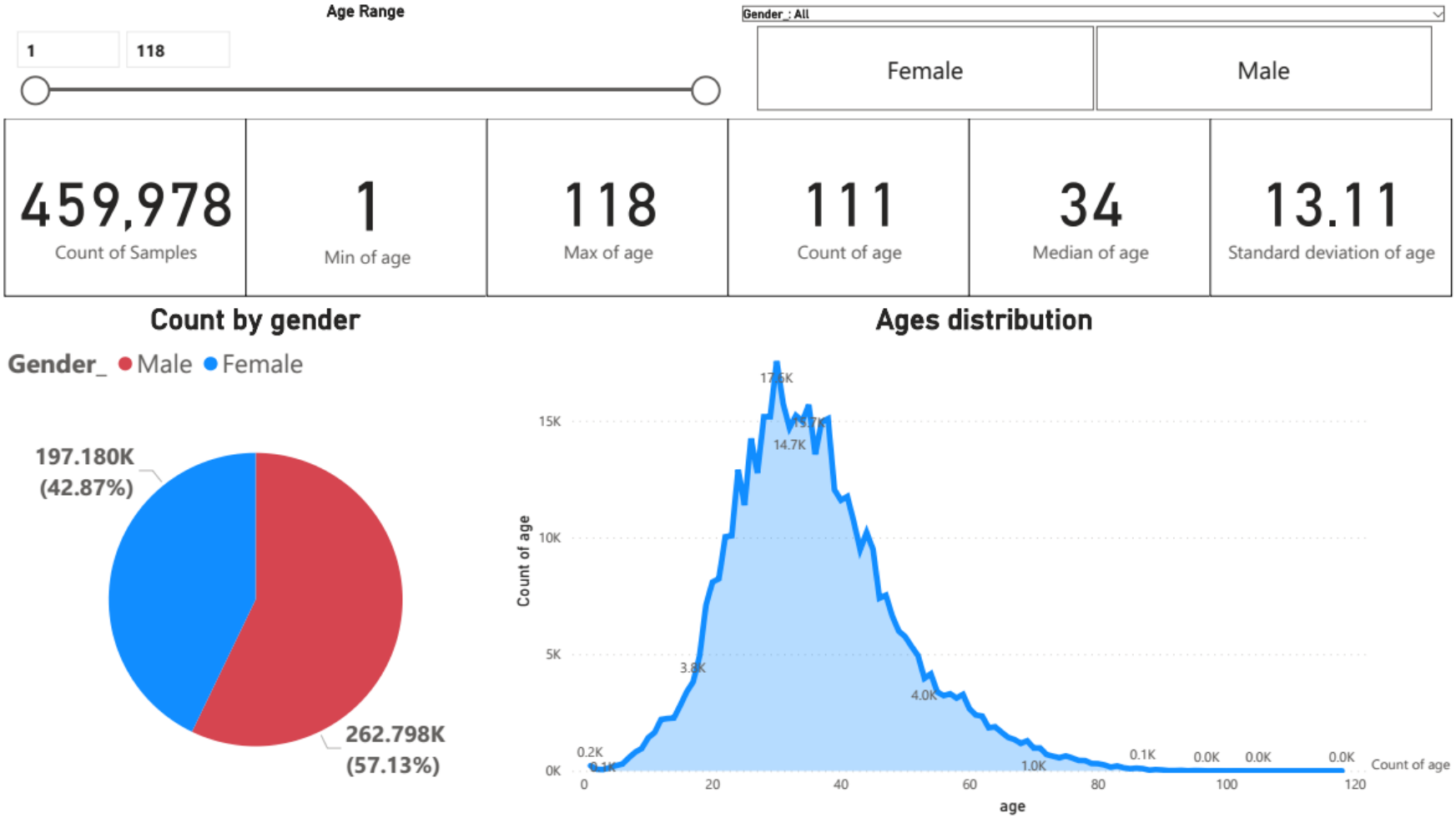
UTKFace dataset is a large-scale face dataset with long age span (range from 0 to 116 years old). The dataset consists of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc.



1.An interactive Page from PowerBI

IMDB:

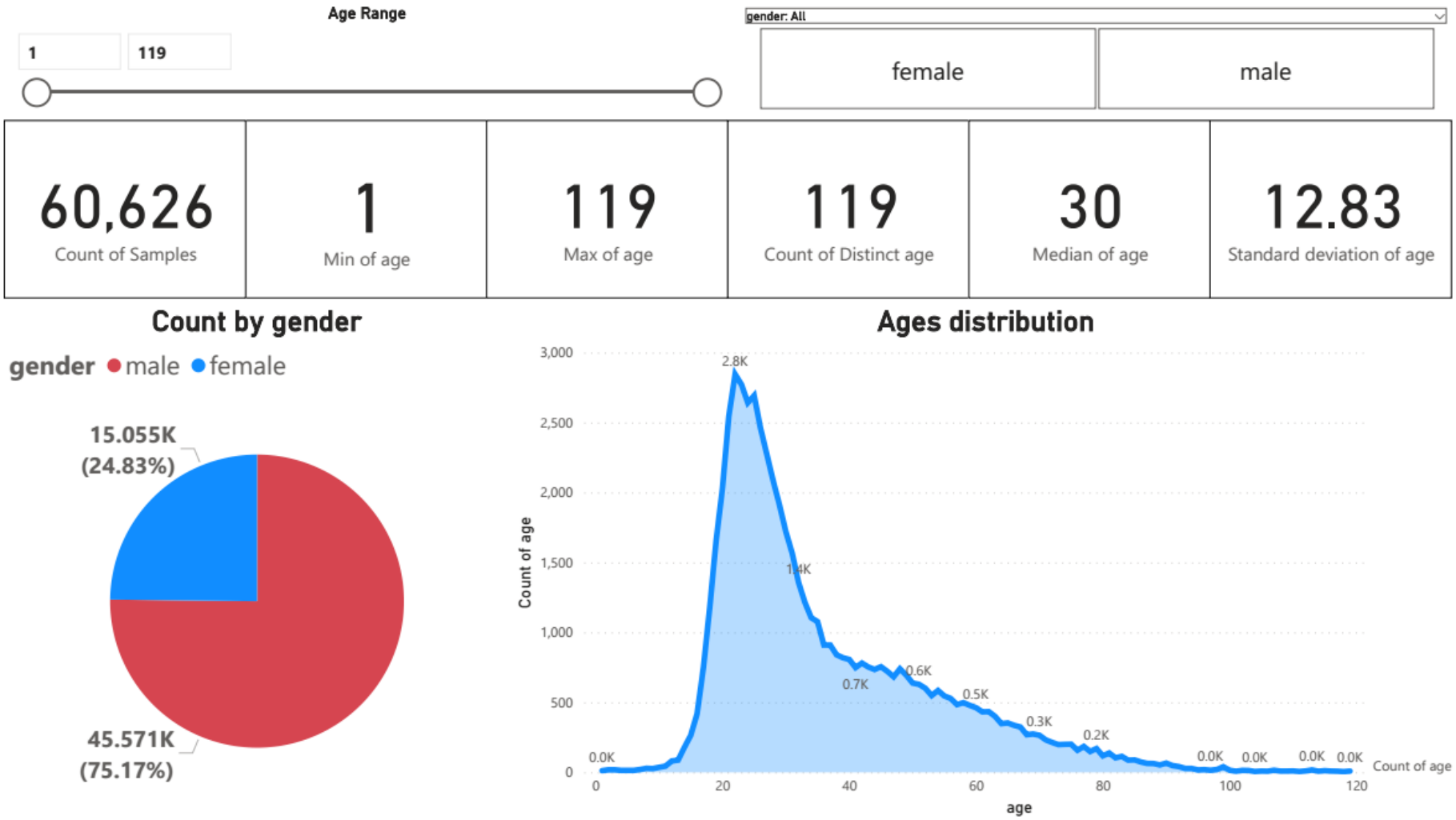
IMDB dataset is a face dataset of the most popular 100,000 actors as listed on the IMDb website and (automatically) crawled from their profiles date of birth, name, gender and all images related to that person.
Note: I have removed the outlier data, so the real total number of samples is 460,723



2.An interactive Page from PowerBI

WIKI:

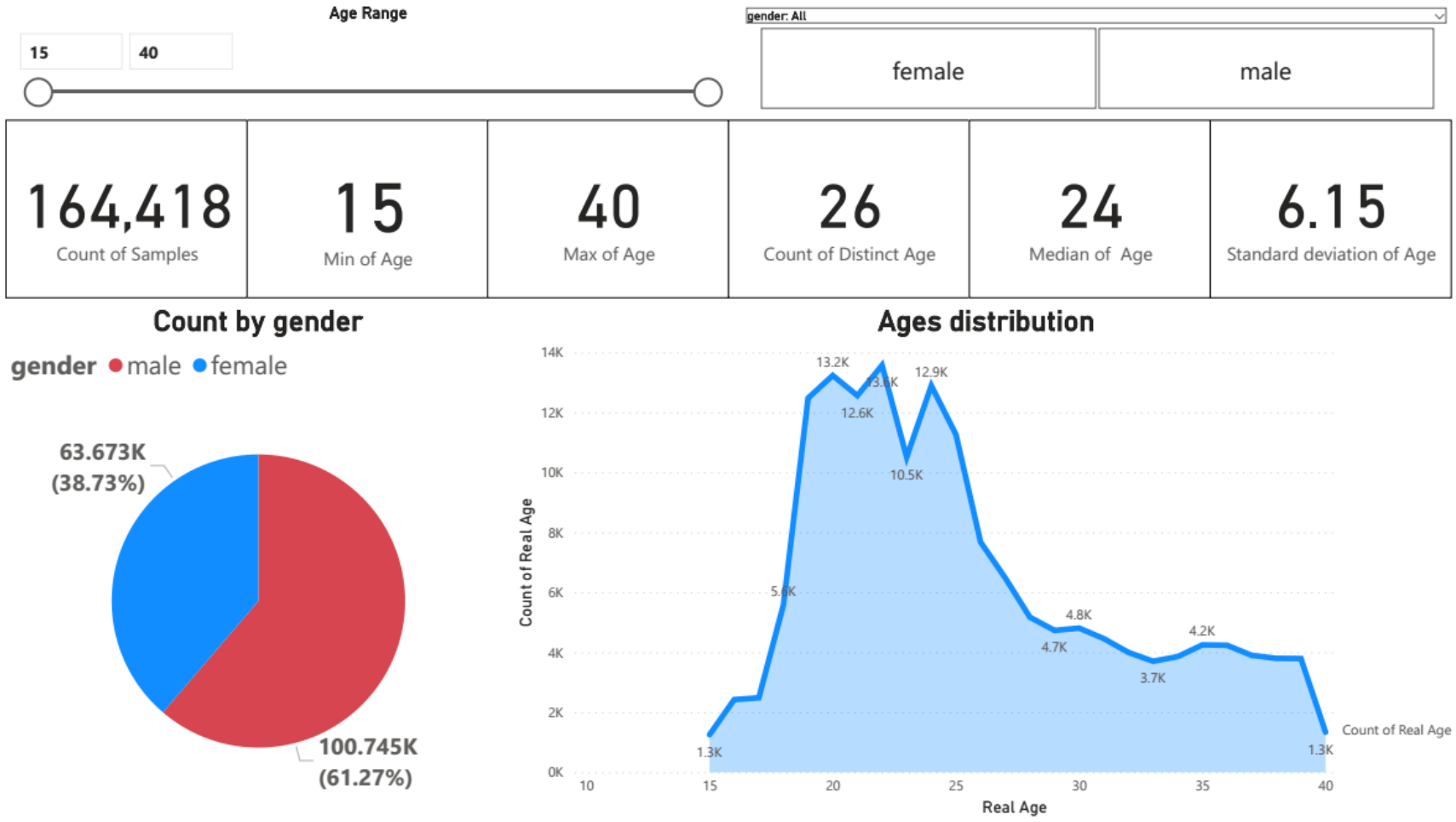
WIKI dataset is a face dataset contains 62,328 images with ages from Wikipedia.
Note: I have removed the outlier data, so the real total number of samples is 62,328



3.An interactive Page from PowerBI

AFAD:

The Asian Face Age Dataset (AFAD) is a new dataset proposed for evaluating the performance of age estimation, which contains more than 160K facial images and the corresponding age and gender labels. This dataset is oriented to age estimation on Asian faces

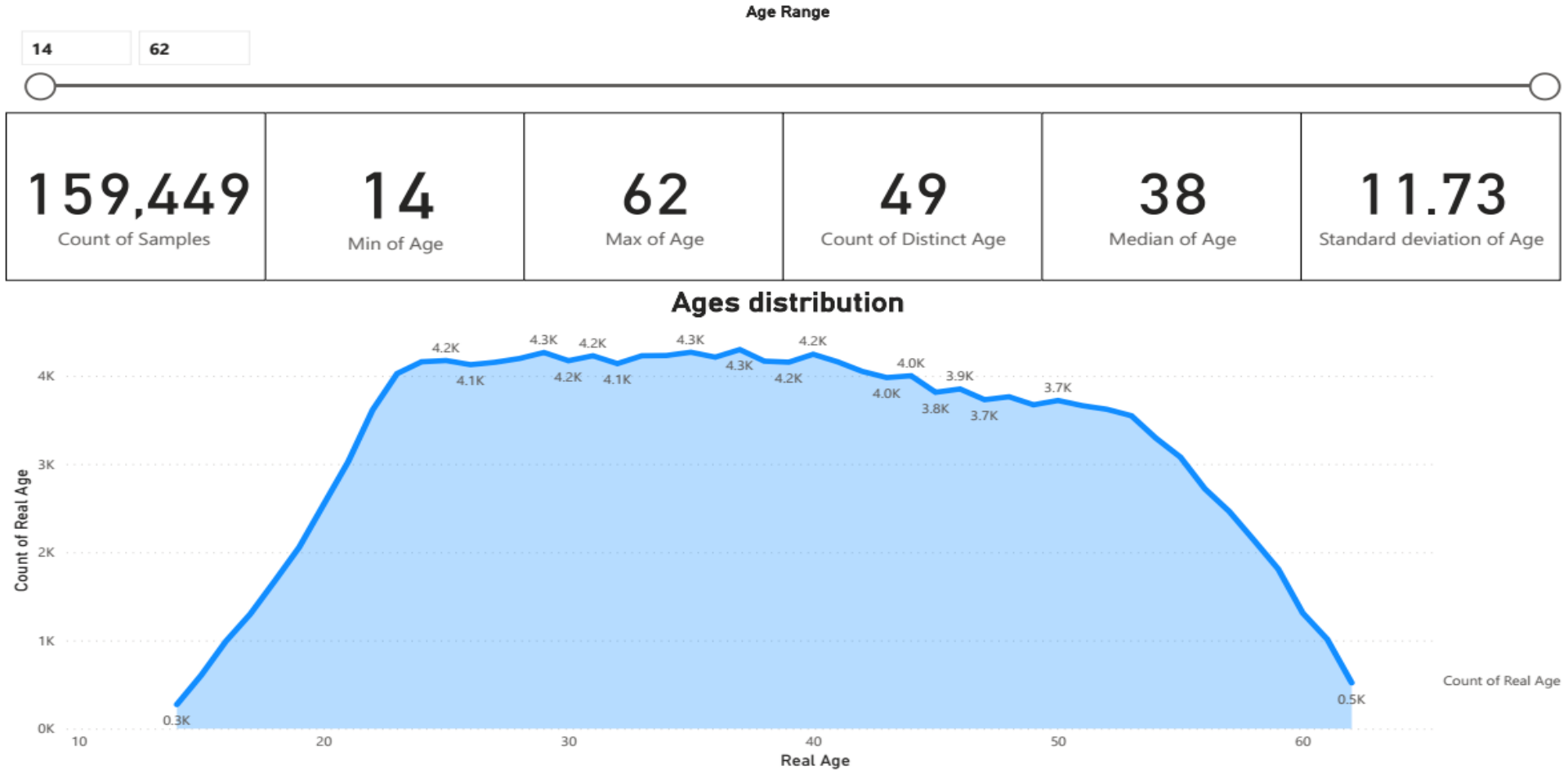


4.An interactive Page from PowerBI

CACD:

Cross-Age Celebrity Dataset (CACD) contains 163,446 images from 2,000 celebrities collected from the Internet.

Note: This Dataset doesn't have the Gender Feature

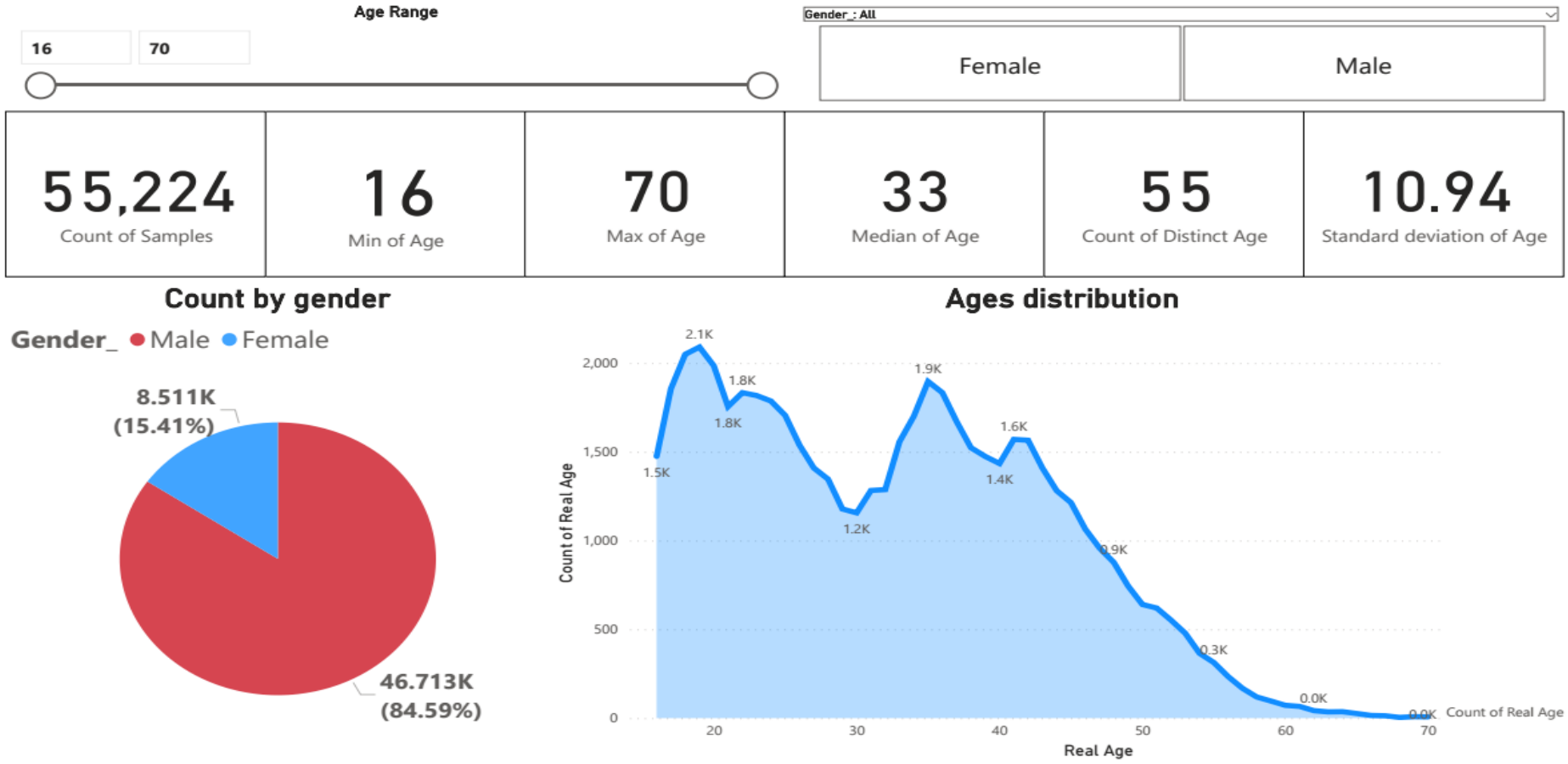


5.An interactive Page from PowerBI

MORPH:

The MORPH Longitudinal Database comprises two core datasets, each containing approximately 200,000 images, resulting in a total size of 400,000+ images! These datasets are further segmented into racial and gender subsets

Note: This analysis is on a sample of the data only MORPH(2).

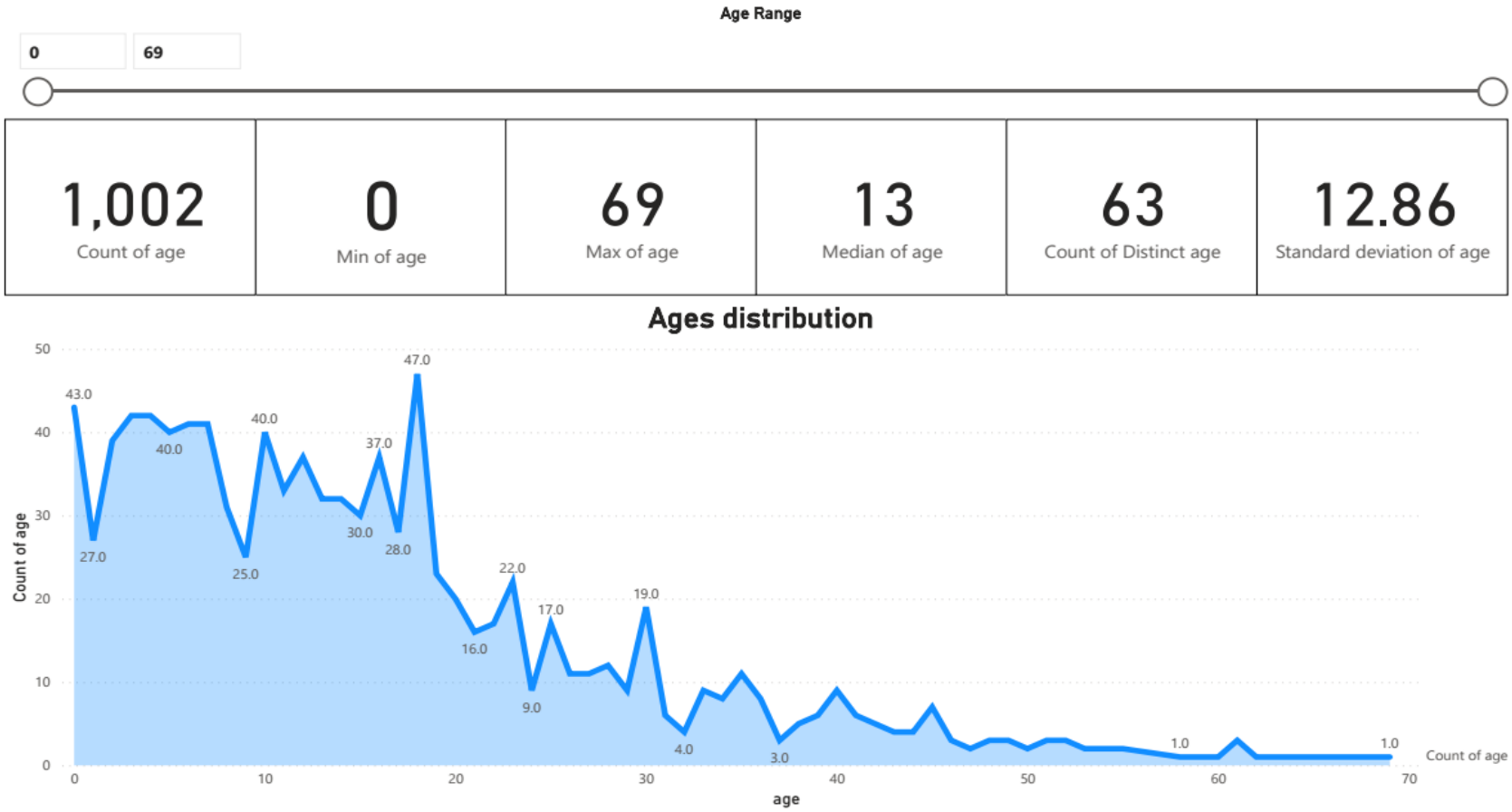


6.An interactive Page from PowerBI

FGNET:

FGNet is a dataset for age estimation and face recognition across ages. It is composed of a total of 1,002 images of 82 people with age range from 0 to 69 and an age gap up to 45 years.

Note : the original FG-NET website does not provide this data any more, but someone provide them in his homepage, this dataset don't have the Gender Feature

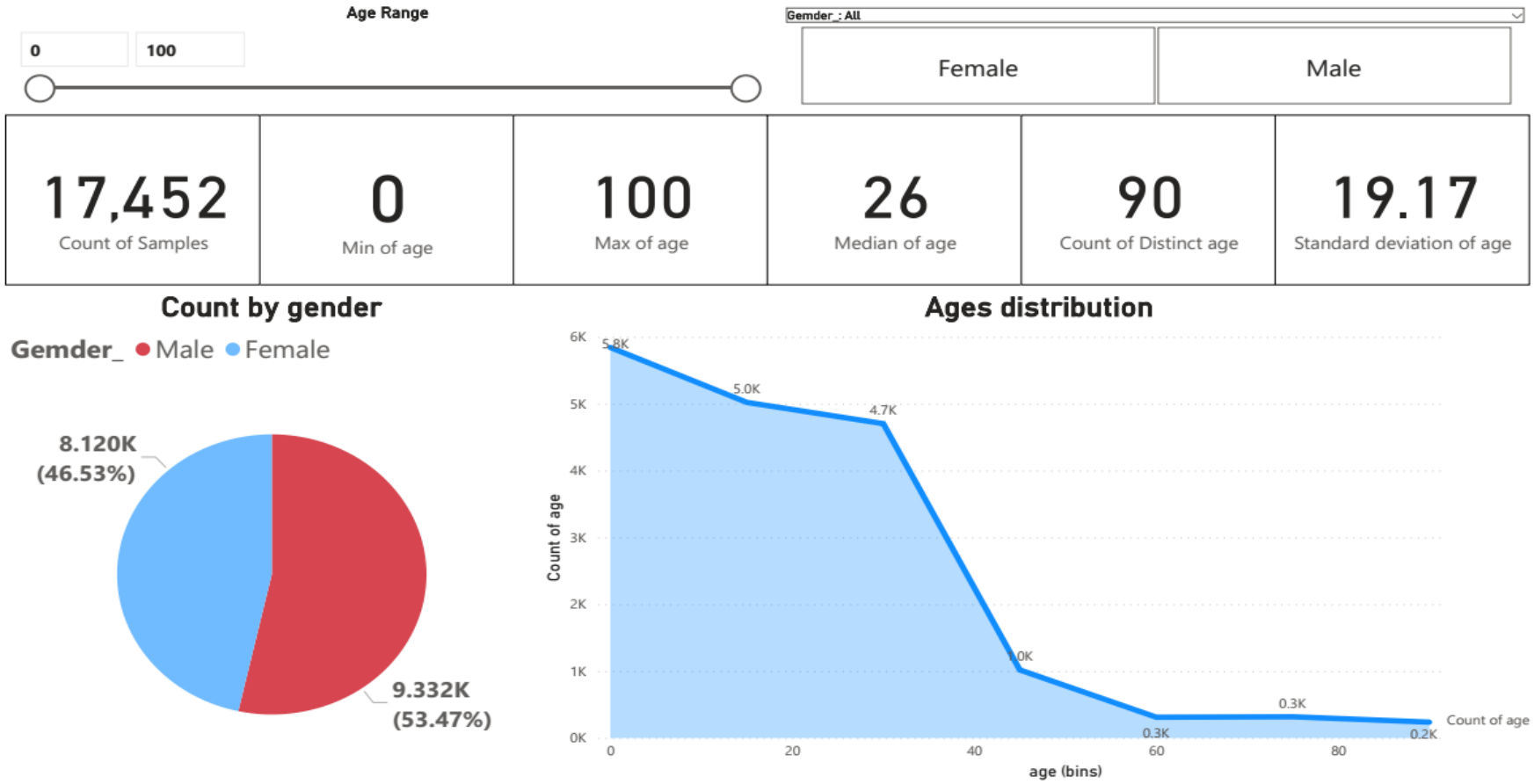


7.An interactive Page from PowerBI

Adience:

The sources of the images included in our set are Flickr albums, assembled by automatic upload from iPhone5 (or later) smart-phone devices, and released by their authors to the general public under the Creative Commons (CC) license.

Note: Total number of photos: 26,580 , but I have found the a metadata of this dataset have 17,452 photo only!



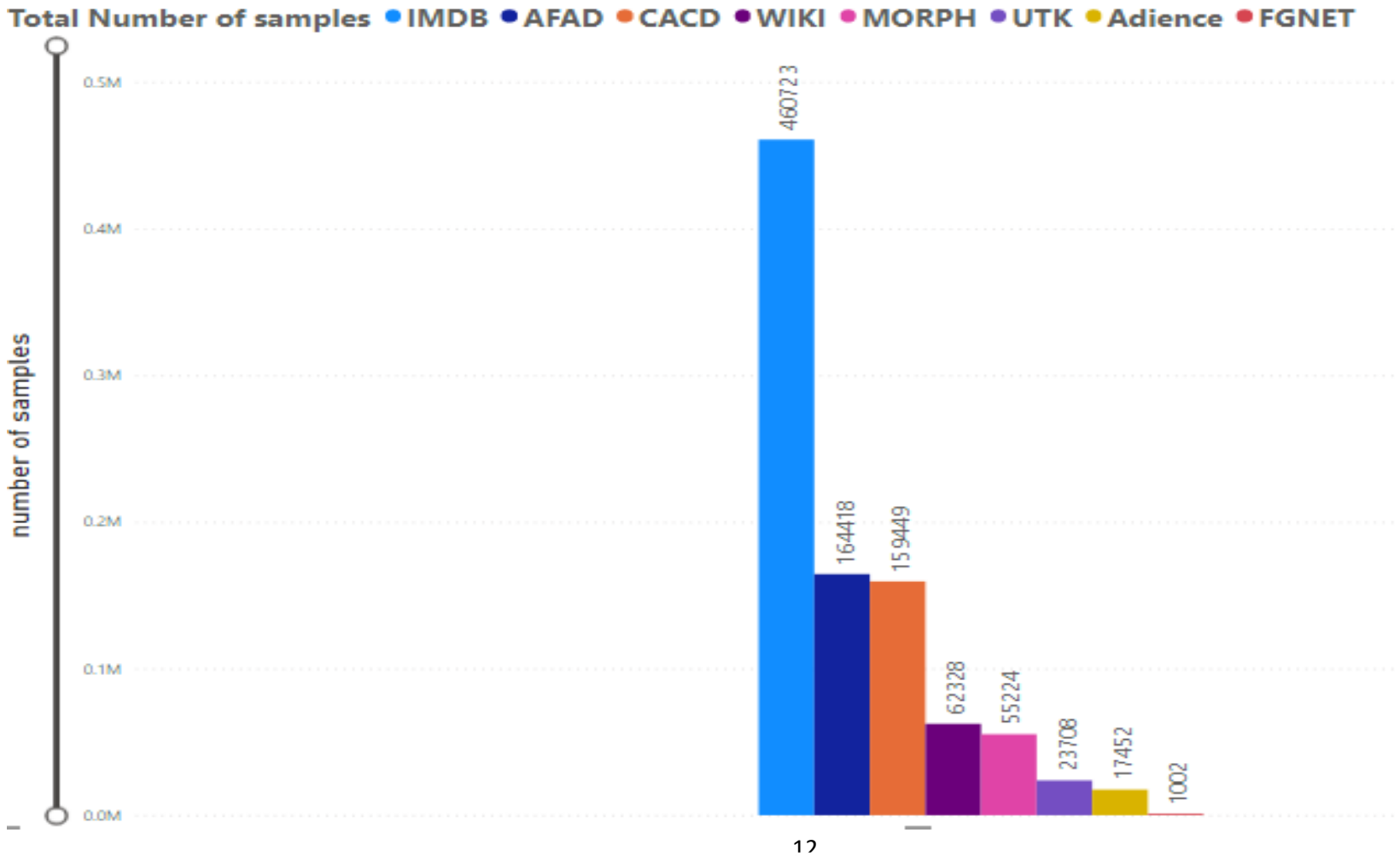
8.An interactive Page from PowerBI-

KANFace:

KANFace consists of 40K still images and 44K sequences (14.5M video frames in total) captured in unconstrained, real-world conditions from 1,045 subjects. The dataset is manually annotated in terms of identity, exact age, gender and kinship.

- Need need an owner contact.

Size Comparison:



Capturing state:

No.	Dataset	Capturing state
1	UTK	in the wild, provides the correspondingly aligned and cropped faces
2	IMDB	in the wild
3	WIKI	in the wild
4	AFAD	in the wild
5	CACD	in the wild
6	MORPH	Controlled
7	FGNET	in the wild
8	Adience	in the wild
9	KANFace	in the wild

Availability:

No.	Dataset	Availability
1	UTK	Available
2	IMDB	Available
3	WIKI	Available
4	AFAD	Available
5	CACD	Available
6	MORPH	need an owner contact
7	FGNET	Available
8	Adience	need an owner contact
9	KANFace	need an owner contact

Usage constraints:

No.	Dataset	Usage constrains
1	UTK	research purposes only
2	IMDB	research purpose only
3	WIKI	research purpose only
4	AFAD	research purpose only
5	CACD	research purpose only
6	MORPH	research use and commercial
7	FGNET	research purposes only
8	Adience	research purposes only
9	KANFace	research purposes only

Conclusion:

- the whole datasets I have found are almost 9 if we consider imdb-wiki dataset as two separate datasets
- the imdb-wiki dataset is the most complete dataset I have found, it has large number of images, the wider range of ages, the best distribution of ages, and good equality of gender distribution.
- MORPH dataset has the largest number of images and a very good distribution based on the race, but the Gender distribution in the samples I have worked on bias toward Males.