M.R.R. Project 2018 - Statistical Analysis and Description - Binomial 8

Presentation of the data : UTKFace - Large Scale Face Dataset

The UTKFace dataset is a large scale face dataset. It contains over 20 000 face images. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc.

The information on the JPG images provided are :
- age : an interger between 0 and 116
- gender : 0 (male) or 1 (female)
- race : an integer from 0 to 4, denoting White, Black, Asian, Indian, and Others (like Hispanic, Latino, Middle Eastern)
- date & time: is in the format of yyyymmddHHMMSSFFF, showing the date and time an image was collected to UTKFace

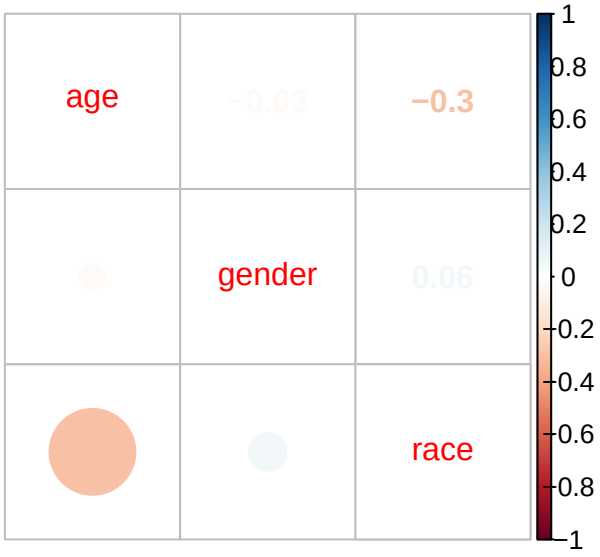
The general nature of the problem

The aim of the problem is to estimate the age of a face from an image. We will construct a model from the variables provided. The target variable is `age`. The co-variable are `gender`, `race` and we also will include information about the images such as grayscale or RGB color intensity per pixel.
In the problem, we will use about 10 000 face images.

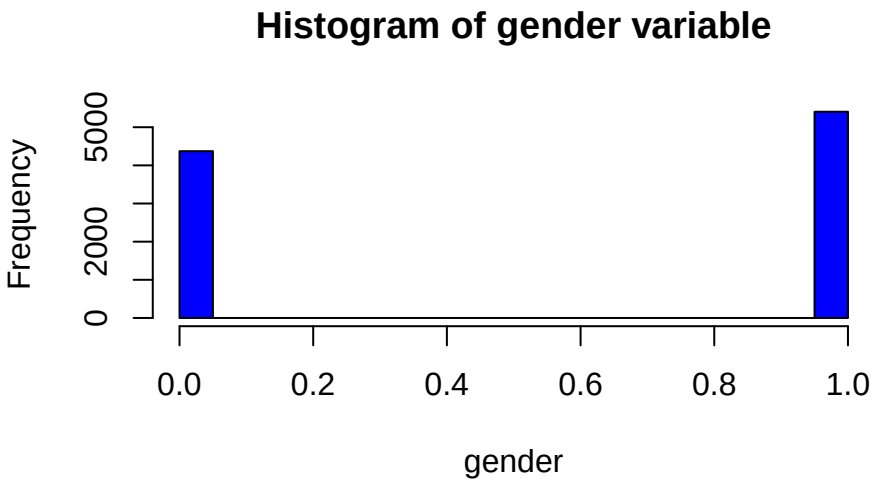
Quick analysis of the target variable and its links to explanatory variables

The data is made up of 9778 face images.

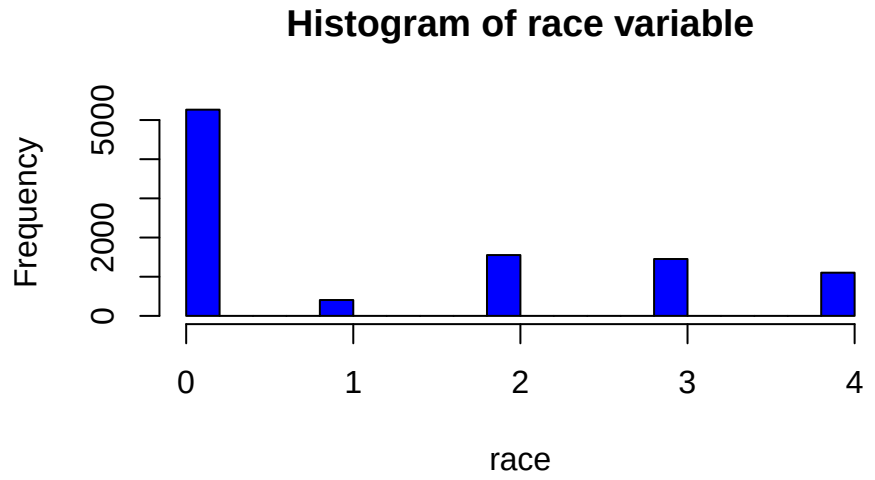
```
## corrplot 0.84 loaded
```



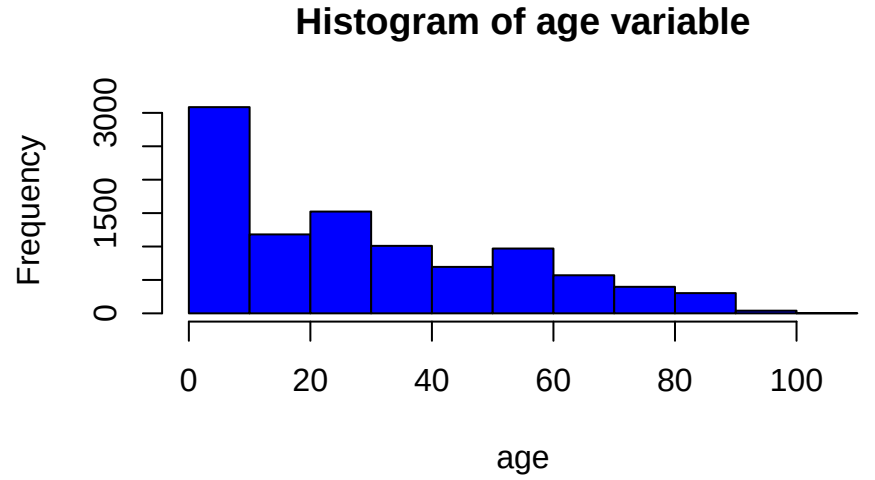
The correlation matrix shows that the `age` variable is more correlated to the `race` variable than the `gender` variable.



We can notice that there are slighty more women than men in the images.



From this histogram, we can know that we have almost 50% of the sample are white people , but much less in black. That means we will have a more precise result in the recognition of white people.



There are more young people than aged people. It will affect the model that we will compute.