# I. Nature générale du Problème

Une variété d'informations peut être extraite d'une image d'un portrait, telle que l'ethnicité, l'âge et le sexe. En effet l'identification des images d'un portrait a été bien exploitée dans le monde réel, telles que dans le contrôle des passeports et des permis de conduire.

Dans notre projet, nous cherchons à réaliser un algorithme permettant l'estimation de l'âge d'une personne en se basant sur une image de son portrait.

Notre base de données, UTKFace, est constituée de 9778 images (dont deux images ont été supprimées à cause d'un manque de données). Sur chacune d'entre elles on voit un visage humain.

Les images sont constituées de 200*200 pixels si noirs ou blancs, et sont constituées de trois matrices 200*200 pour RGB.

Ainsi à chaque image est associée un nom contenant des données qui lui sont relatives, il s'agit de:

- L'âge de la personne sur l'image qui est compris entre 0 et 116.
- Son sexe, qui est soit 0 (les hommes), soit 1 (les femmes).
- Son ethnie qui est : 0 (les blancs), 1 (les noirs), 2 (les Asiatiques), 3 (les Indiens) ou 4(les autres ethnies).
- La date de la prise de la photo.

# II. Variable cible

La variable cible dans notre projet est l'âge, notre objectif est donc de déterminer l'âge ou l'Intervalle d'âge de la personne présentée dans l'image en question.

Nous avons remarqué que la répartition des âges en fonction des images est non uniforme. En effet, on observe un grand nombre d'images concernant des enfants de 1 an par rapport au reste de la population, comme indiqué dans l'histogramme ci-dessous.

Une telle distribution peut fausser les statistiques de l'algorithme de test, d'une façon que le modèle d'estimation sera moins performant pour les images des personnes ayant un âge important.

Pour cela nous envisageons de réduire la base de données afin d'avoir une distribution uniforme.

# III. Les variables explicatives

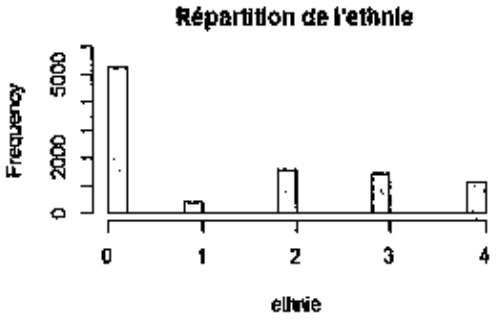
Les variables étudiées sont les pixels de chacune des images. Une image de 200*200 pixels a donc 120000 variables déterminées par un codage RGB (3 variables: rouge, vert et bleu). Les variables sont des réels variant entre 0 et 1.

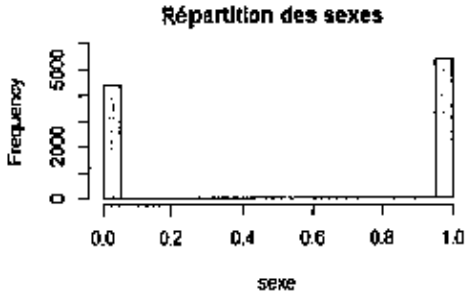
Par ailleurs, nous avons décidé de convertir les images en noir et blanc dans le but de diminuer les variables à analyser.

Ensuite, on envisage d'utiliser la méthode du Poor man ainsi qu'une méthode de pixellisation (regrouper un ensemble de pixels dans une variable) pour réduire encore le nombre de variables.

On souhaite alors déterminer l'âge d'une personne à partir de son image donc à partir de l'intensité des variables.

De plus, nous avons effectué des analyses de données associées à chaque image, il s'agit des données concernant le sexe et l'ethnie, afin d'en avoir une meilleure idée, mais on note qu'à ce stade du projet nous ignorons leurs influence sur l'âge.




Dans l'histogramme représentant le sexe ci-dessus, on remarque que le nombre d'image de femmes est plus important que le nombre des images contenant le portrait d'hommes.

Par ailleurs, dans l'histogramme de l'ethnie, on remarque que le nombre de personnes "Blancs" est supérieur par rapport aux autres ethnies.

# IV. Lien entre la variable cible et les variables explicatives

L'estimation de l'âge à partir des images fournis, va être conditionnée par le nombre de pixels conservés (pertinence de la base de données) et par la qualité de l'échantillon, la répartition n'étant pas équitable.

# 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

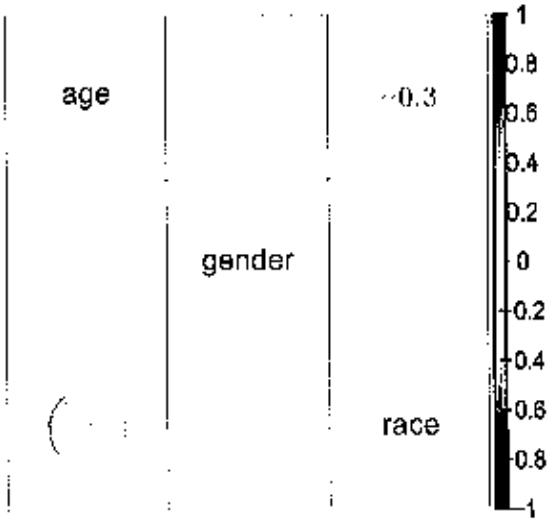
## The general nature of the problem

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.

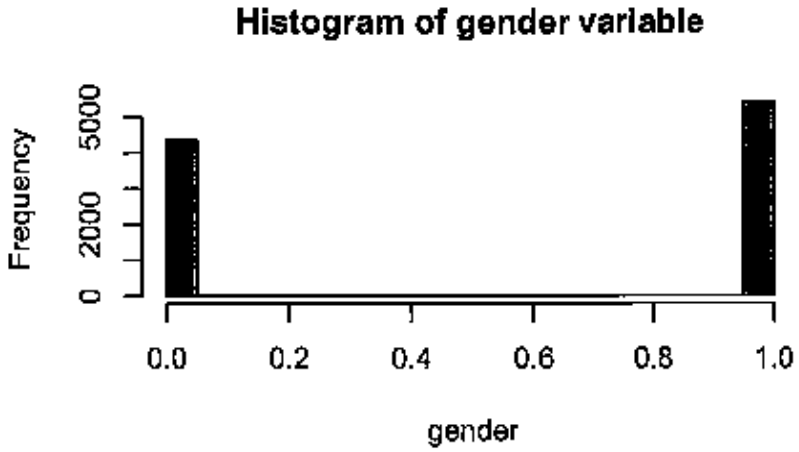
## Quick analysis of the target variable and its links to explanatory variables

The data is made up of 9778 face images.

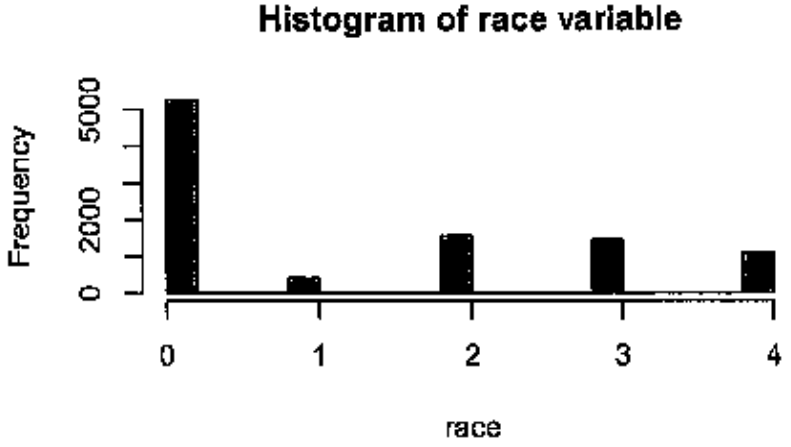
```
## corrplot 0.84 loaded
```



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.



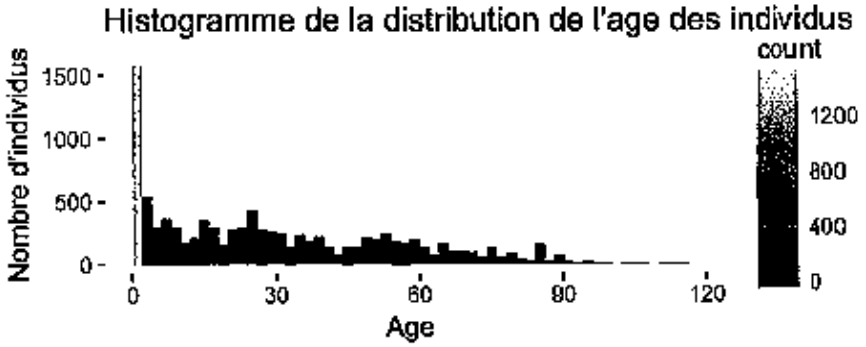

From this histogram, we can know that we have almost 50% of the sample are white people , but much less in black. That means we will have a more precise result in the recognition of white people.


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

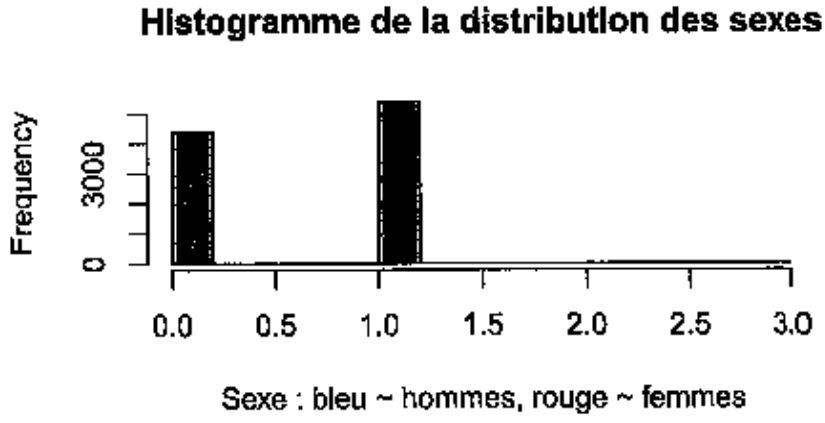
# Projet MRR

Le dataset à étudier est nommé UTKFace. Celui-ci est une base de 9780 photos d'individus en couleurs de taille 200px * 200px. La variable cible de notre problème est ici l'âge de nos individus, c'est à dire qu'à partir de ces photos nous devrons construire un algorithme capable de prédire l'âge d'un individu. Notre variable cible l'age nous est donné dans le nom de la photo de l'individu (ainsi que les données concernant la "race" et le sexe des individus) et comme nous avons la connaissance de cette variable l'algorithme d'apprentissage développé sera dit "supervisé".

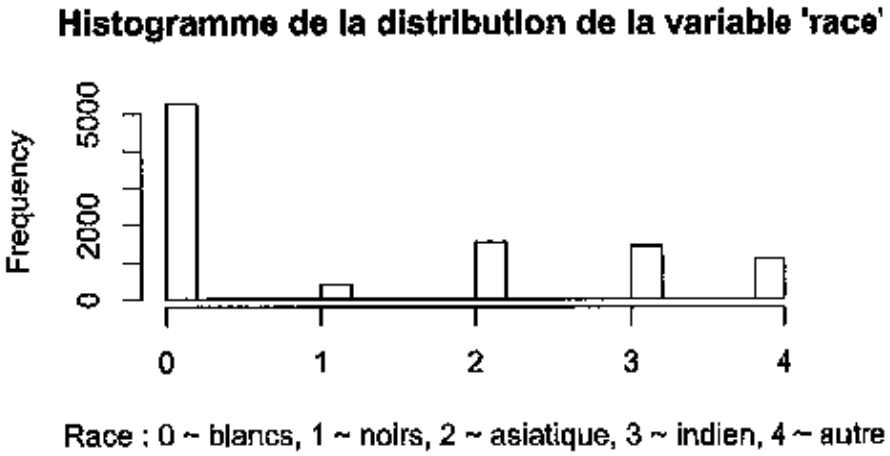
Description des variables du dataset :
- Age : entier entre 0 et 116 correspondant à l'âge d'un individu
- Race : entier entre 0 et 4 correspondant à la "race" de l'individu
- Gender : entier entre 0 et 1 correspondant au genre de l'individu
- Tous les pixels des photos



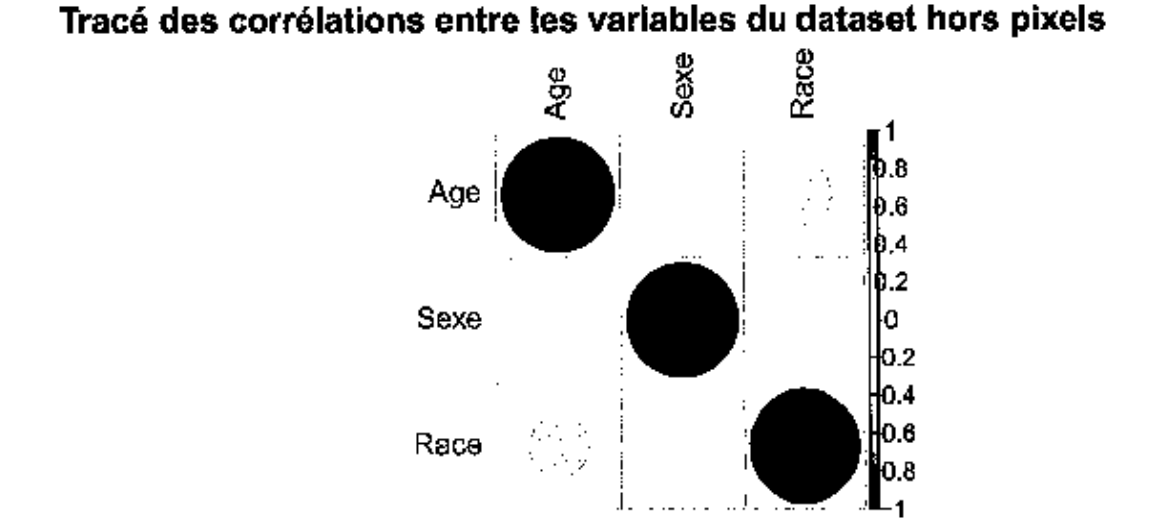

On observe ici une sur-représentation des individus agés d'1 an et les individus plus âgés sont eux très peu représentés. Cette sous-représentation des personnes très âgées peut nuire au potentiel de généralisation de notre algorithme si l'on souhaite qu'il prédise des âges très élevés.




La distribution des sexes est quasiment équivalente ce qui est une bonne chose pour l'apprentissage si cette variable est importante.




Ici, les blancs sont sur-représentés alors que les noirs sont sous-représentés. Ainsi, si cette variable est très importante pour prédire l'âge, il peut y avoir des risques de sous apprentissage (underfit) pour les noirs.




Comme attendu, il n'y a pas de corrélation entre les variables et il est évident que la seule connaissance du sexe et de la "race" d'un individu ne soit pas suffisante pour prédire son âge.

Nous envisageons la démarche suivante pour résoudre le problème :
- nous allons travailler sur des échantillons plus petits (1000 par exemple). Nous serons vigilants quant au fait de prendre des échantillons représentatifs du dataset (distribution des différents groupes similaire)
- Pour évaluer notre modèle, nous allons créer des classes d'âge (ex :la classe 0 pourrait correspondre aux bébés de 1 à 3 ans). Ainsi plutôt que d'évaluer notre modèle avec par exemple la somme des moindres carrés entre l'âge prédit et l'âge réel , nous regarderons comme pour une classification, l'appartenance de la prédiction à une classe d'âge.
- les photos étant alignés et recadrés, nous pourrions découper l'image en plusieurs régions (front, yeux ...) pour éventuellement trouver une corrélation entre un groupe de pixels d'une même région et l'âge de la personne.

# Projet de MRR - UTKFace

[redacted]

## Nature du problème

Le but de notre projet est de réaliser un programme permettant de déterminer l'âge d'un individu en fonction d'une photo de son visage. Nous avons à notre disposition une base de données, UTKFace, contenant 9778 photos d'individus. Ces photos sont prises de face et ne contiennent que le visage des individus, certaines sont en noir et blanc mais la majorité est en couleur.

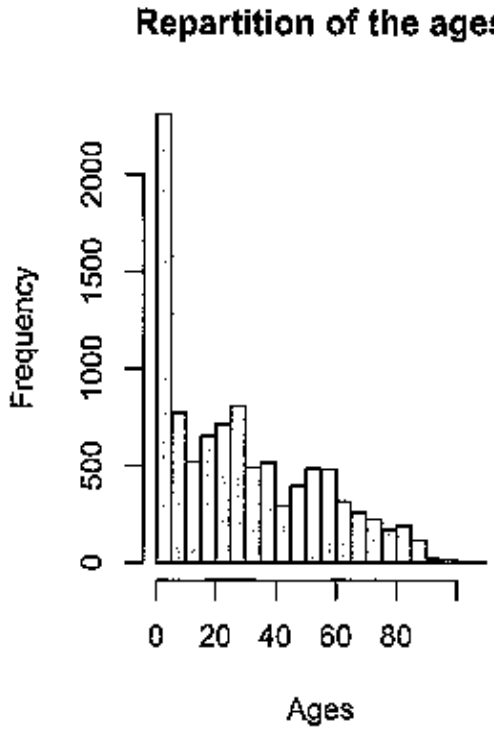
Toutes les images de la base sont d'une résolution 200*200 pixels. À ces images sont associées 4 variables numériques elles sont toutes stockées dans le nom des fichiers, séparées par des '_' et comprennent :

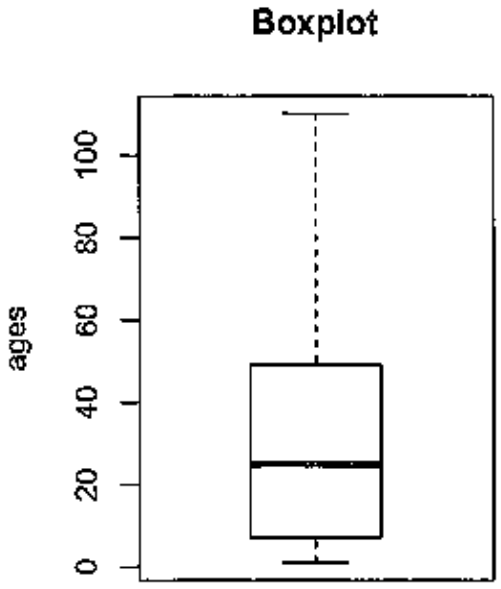
- L'âge de l'individu, un entier compris entre 0 et 116 ;
- Son genre, un entier qui vaut 0 pour les hommes, 1 pour les femmes ;
- Son ethnie est un entier compris entre 0 et 4 respectivement pour les blancs, les noirs, les asiatiques, les indiens et les autres ethnies ;
- La date de mise en ligne de la photo dans la base de données.

Voici un exemple d'une des images de la base de données:

On remarque sur le graphique ci-dessous que la répartition de photos n'est pas uniforme. En effet on observe une surproportion d'images concernant des enfants de moins de 5 ans compte tenu du reste de la population. Étant donné disproportion, on peut s'attendre à ce que notre modèle soit moins entraîné et donc moins performant pour déterminer l'âge des humains de plus de 5 ans. En particulier, le nombre de données disponibles diminue avec l'âge, le diagramme en boite met en avant un échantillon particulièrement jeune, avec 75% de la population ayant moins de 49 ans et une médiane à 25 ans.

Nous avons par ailleurs fait le choix de faire abstraction de la date de mise en ligne de la photo (qui n'influe pas sur l'âge perçu), et dans un premier temps, nous ferons également abstraction de l'ethnie et du sexe. Bien que ceux-ci peuvent influer sur les résultats, nous faisons le choix de nous concentrer d'abord sur un modèle plus simple pour ensuite l'affiner si nécéssaire.




## Les variables explicatives disponibles

Les variables explicatives sont les pixels de l'image, ce qui pour une image de 200*200 pixels représente un total de 40000 variables. Par ailleurs pour des images en couleurs, chaque pixel est déterminé par 3 variables : rouge, vert et bleu (codage RGB) qui représentent la couleur du pixel en synthèse additive. Ces trois variables sont des niveaux d'intensités variant entre 0 et 255.

Cependant certaines images de la base de données sont en noir et blanc. Par conséquent nous avons pris la décision de d'abord convertir les images en noir et blanc avant de les analyser. Cette conversion à de plus l'avantage de diviser par 3 le nombre de variables explicatives. On passe alors de 120000 variables à 40000.

Le but de notre programme est donc à partir d'une image noir et blanc de 200*200 pixels de renvoyer un entier correspondant à l'âge de l'individu.

## La variable cible et son lien avec les variables explicatives

La variable cible est donc l'âge de l'individu pris en photo. Les variables explicatives sont les valeurs des différents pixels de l'image (convertie en noir et blanc).

La valeur de chaque pixel, prise individuellement, a un impact sur l'image finale et par conséquent sur l'âge de l'individu. Cependant un modèle linéaire classique (somme des valeurs des pixels pondérées par des poids) n'est ici pas envisagable car il suffit d'imaginer une photo d'un individu ayant subit une translation de quelques pixels : chaque pixels sera alors complétement différents bien que l'individu soit le même et que la différence est imperceptible entre les deux photos.

Nous envisageons ainsi de baisser la résolution des photos afin d'une part de subdiviser l'image en groupe de pixel ce qui permet de parer à ce problème, et d'autre part de réduire considérablement le nombre de variables explicatives.