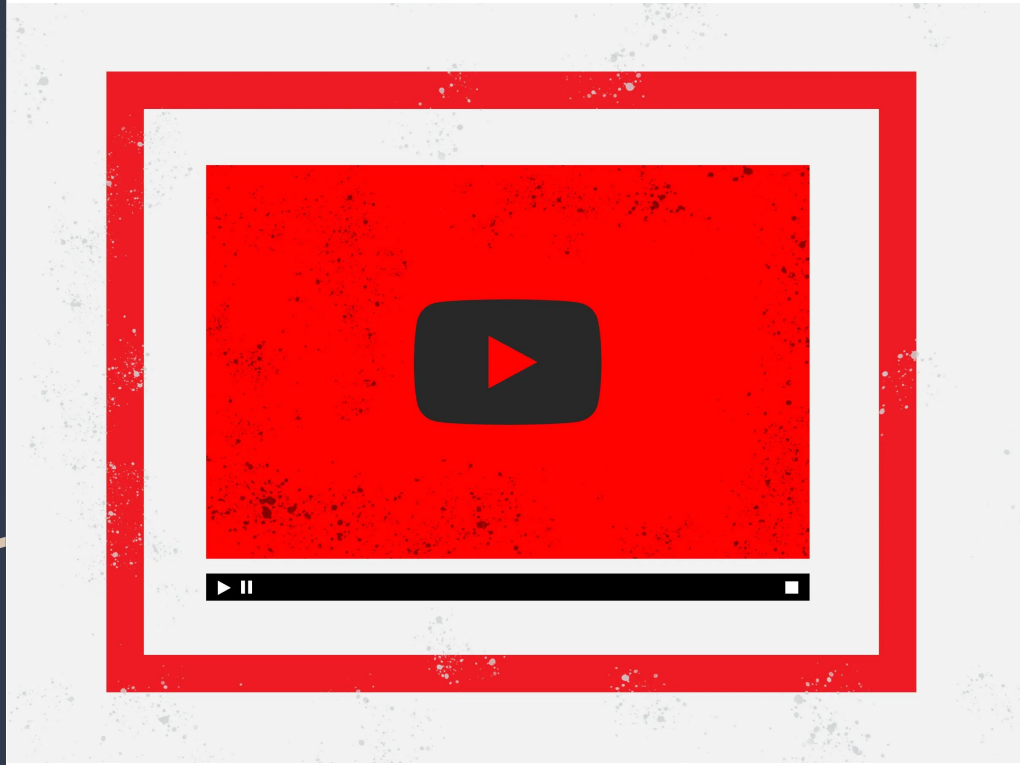
Online **Video** Characteristics and Transcoding Time Dataset Data Set

Florian Menielle et Matthieu Mac Nab

The dataset contains a million randomly sampled video instances listing 10 fundamental video characteristics along with the YouTube video ID



In a nutshell

There are two datasets :

- The first one is essential characteristics of 168 286 videos
- The second one measures the time and memory used to convert 68 784 videos from one codec to another

Here we will focus the analysis on the second dataset, and build models to predict necessary resources to convert a video given its characteristics

Dataset features 1/2

id = Youtube video id

duration = duration of video

bitrate bitrate(video) = video bitrate

height = height of video in pixels

width = width of video in pixels

frame rate = actual video frame rate

frame rate(est.) = estimated video frame rate

codec = coding standard used for the video

category = YouTube video category

url = direct link to video (has expiration date)

i = number of i frames in the video

p = number of p frames in the video

Dataset features 2/2

b = number of b frames in the video

frames = number of frames in video

i_size = total size in byte of i videos

p_size = total size in byte of p videos

b_size = total size in byte of b videos

size = total size of video

o_codec = output codec used for transcoding

o_bitrate = output bitrate used for transcoding

o_framerate = output frame rate used for transcoding

o_width = output width in pixel used for transcoding

o_height = output height used in pixel for transcoding

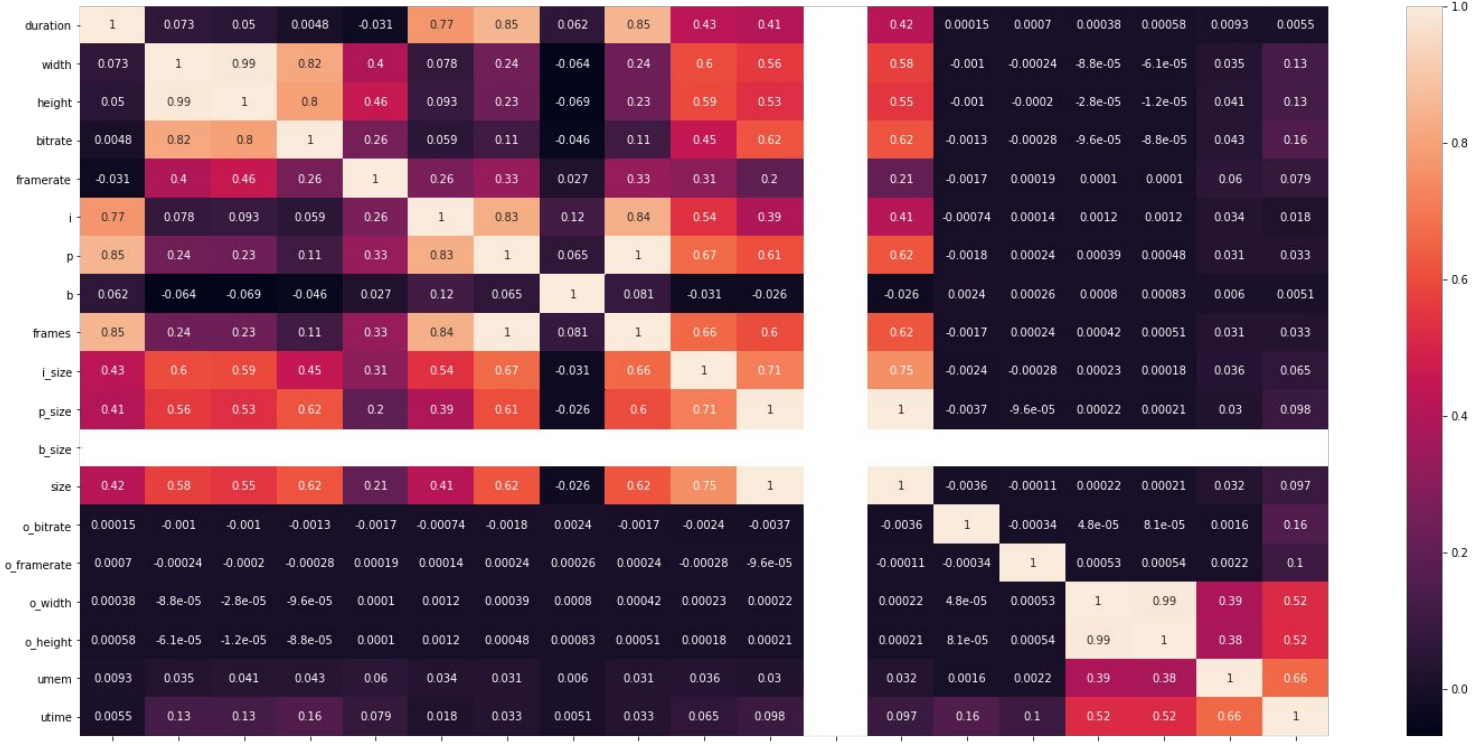
umem = total codec allocated memory for transcoding

utime = total transcoding time for transcoding

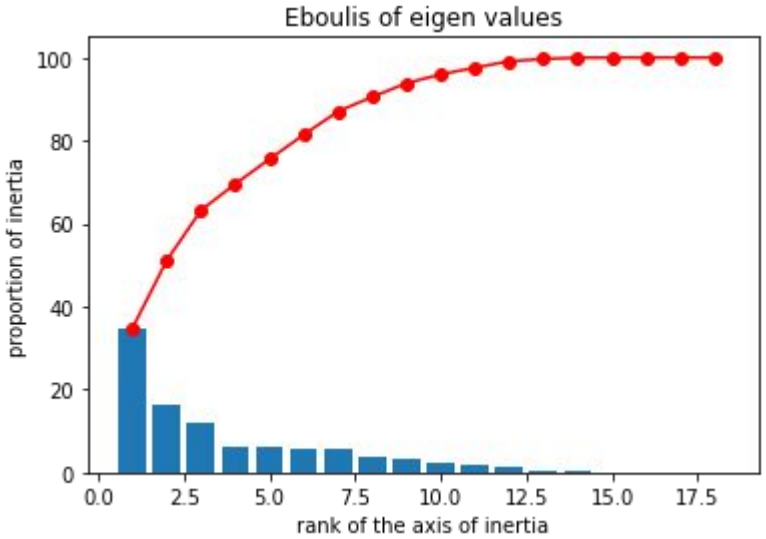
Understanding the data

- 68784 rows
- most features are numerical, some categorical (id, link, codec)
- Associated task is regression
- Possible value to predict are **utime** (conversion time) and **umem** (needed memory)

Correlation matrix

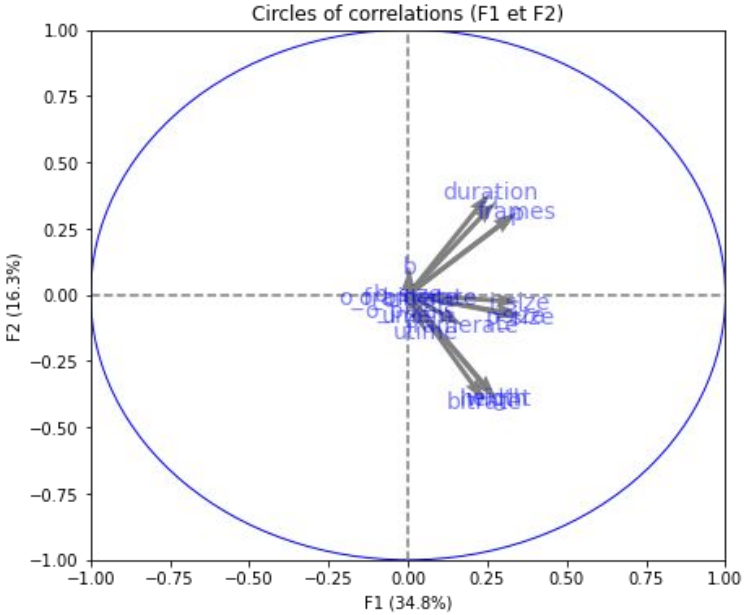


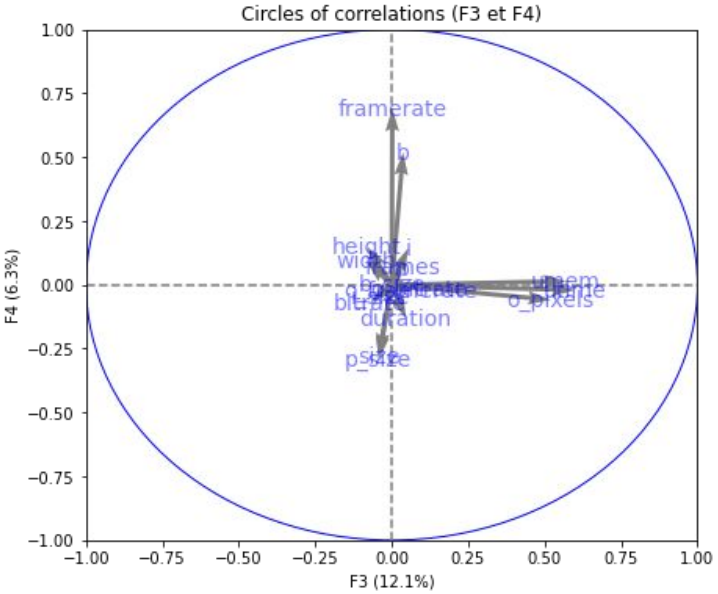
Principal components analysis



- The elbow shows at the third component, but four ones are needed to draw two correlation circles, so we will choose the first four ones.

Principal components analysis





The models

We will use two models for this regression task :

- multi-linear regression
- Random forest

Random forest are known to perform better in classification than regression, so we expect best results from the linear model.

Model comparison

Model	MSE	Params
Multi-linear	3.3225e-27	None
Random forest	0.0269	max depth : 9 number of estimators : 5

Random forest params was found using a cross validation grid, but the linear model still outperform it.