**1. Exploratory Data Analysis**

**Objectif :**

* Comprendre du mieux possible nos données (un petit pas en avant vaut mieux qu'un grand pas en arriere)
* Développer une premiere stratégie de modélisation

**Checklist de base**

**Analyse de Forme :**

* **variable target** : SARS-Cov-2 exam result
* **lignes et colonnes** : 5644, 111
* **types de variables** : qualitatives : 70, quantitatives : 41
* **Analyse des valeurs manquantes** :
  + beaucoup de NaN (moitié des variables > 90% de NaN)
  + 2 groupes de données 76% -> Test viral, 89% -> taux sanguins

**Analyse de Fond :**

* **Visualisation de la target** :
  + 10% de positifs (558 / 5000)

**Analyse plus détaillée**

* **Relation Variables / Variables** :
  + blood\_data / blood\_data : certaines variables sont tres corrélées : +0.9 (a suveiller plus tard)
  + blood\_data / age : tres faible corrélation entre age et taux sanguins
  + viral / viral : influenza rapid test donne de mauvais résultats, il fauda peut-etre la laisser tomber
  + relation maladie / blood data : Les taux sanguins entre malades et covid-19 sont différents
  + relation hospitalisation / est malade :
  + relation hospitalisation / blood : intéressant dans le cas ou on voudrait prédire dans quelle service un patient devrait aller
* **NaN analyse** : viral : 1350(92/8), blood : 600(87/13), both : 90

**hypotheses nulle (H0):**

* Les individus atteints du covid-19 ont des taux de Leukocytes, Monocytes, Platelets significativement différents
  + H0 = Les taux moyens sont ÉGAUX chez les individus positifs et négatifs
* Les individus atteints d'une quelconque maladie ont des taux significativement différents