**Objectif :**

* Comprendre du mieux possible nos données (un petit pas en avant vaut mieux qu'un grand pas en arrière)
* Développer une première 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)
* **Signification des variables** :
  + variables continues standardisées, skewed (asymétriques), test sanguin
  + age quantile : difficile d'interpreter ce graphique, clairement ces données ont été traitées, on pourrait penser 0-5, mais cela pourrait aussi etre une transformation mathématique. On peut pas savoir car la personne qui a mit ce dataset ne le précise nul part. Mais ca n'est pas tres important
  + variable qualitative : binaire (0, 1), viral, Rhinovirus qui semble tres élevée
* **Relation Variables / Target** :
  + target / blood : les taux de Monocytes, Platelets, Leukocytes semblent liés au covid-19 -> hypothese a tester
  + target/age : les individus de faible age sont tres peu contaminés ? -> attention on ne connait pas l'age, et on ne sait pas de quand date le dataset (s'il s'agit des enfants on sait que les enfants sont touchés autant que les adultes). En revanche cette variable pourra etre intéressante pour la comparer avec les résultats de tests sanguins
  + target / viral : les doubles maladies sont tres rares. Rhinovirus/Enterovirus positif - covid-19 négatif ? -> hypothese a tester ? mais il est possible que la région est subie une épidémie de ce virus. De plus on peut tres bien avoir 2 virus en meme temps. Tout ca n'a aucun lien avec le covid-19

**Analyse plus avancée :**

**Relation variables / variables :**

* blood\_data / blood\_data : Certaines variables très corrélées -> +0.9
* blood\_data / age : très faible correlation
* viral / viral : influenza rapide test donne de mauvais résultats
* relation maladie / blood\_data : les taux sanguins entre malades et covid19 sont différents
* relation hospitalisation / est malade :
* relation hospitalisation / blood : intéressant si prédiction du service dans lequel 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
  + REJETEE H0 = Les taux moyens sont ÉGAUX chez les individus positifs et négatifs
* Les individus atteints d'une quelconque maladie ont des taux significativement différents