**content**

The dataset was provided by the Mexican government [(link)](https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico). This dataset contains an enormous number of anonymized patient-related information including pre-conditions. The raw dataset consists of 21 unique features and 1,048,576 unique patients. **In the Boolean features, 1 means "yes" and 2 means "no". values as 97 and 99 are missing data**.

* sex: 1 for female and 2 for male.
* age: of the patient.
* classification: covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.
* patient type: type of care the patient received in the unit. 1 for returned home and 2 for hospitalization.
* pneumonia: whether the patient already have air sacs inflammation or not.
* pregnancy: whether the patient is pregnant or not. (All men 97 – not pregnant)
* diabetes: whether the patient has diabetes or not.
* copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.
* asthma: whether the patient has asthma or not.
* inmsupr: whether the patient is immunosuppressed or not.
* hypertension: whether the patient has hypertension or not.
* cardiovascular: whether the patient has heart or blood vessels related disease.
* renal chronic: whether the patient has chronic renal disease or not.
* other disease: whether the patient has other disease or not.
* obesity: whether the patient is obese or not.
* tobacco: whether the patient is a tobacco user.
* usmr: Indicates whether the patient treated medical units of the first, second or third level.
* medical unit: type of institution of the National Health System that provided the care.
* intubed: whether the patient was connected to the ventilator.
* icu: Indicates whether the patient had been admitted to an Intensive Care Unit.
* date died: If the patient died indicate the date of death, and 9999-99-99 otherwise.

**Analysis:**

1. The dataset says all values like 97 & 99 are missing values but **this is incorrect**.
2. Only 97 value represents missing data for most of the features except PREGNANT.
3. 3 columns have this value 97. (INTUBED, PREGNANT, ICU)
4. All men in feature SEX have value = 97 for PREGNANT feature. Value 98 is missing data.
5. PATIENT\_TYPE = 1 (returned home) and 2 (hospitalized). That means for all patients who are returned to home (value 1) they are not in ICU nor they INTUBED.
6. So, for all ICU and INTUBED value 97 means (Returned home) should be replaced with 2. Value 99 is missing data.