UNIVERSITÉ PARIS 8
VINCENNES-SAINT-DENIS

# Analyse des données Covid 19 d'Indonésie

**Omar HAMMOUCHE**

**Mohammed Amine Kaced**

**Université Paris 8, LIASD**

**Professeur : Akdag Herman**

**17 décembre 2020**

# SOMMAIRE

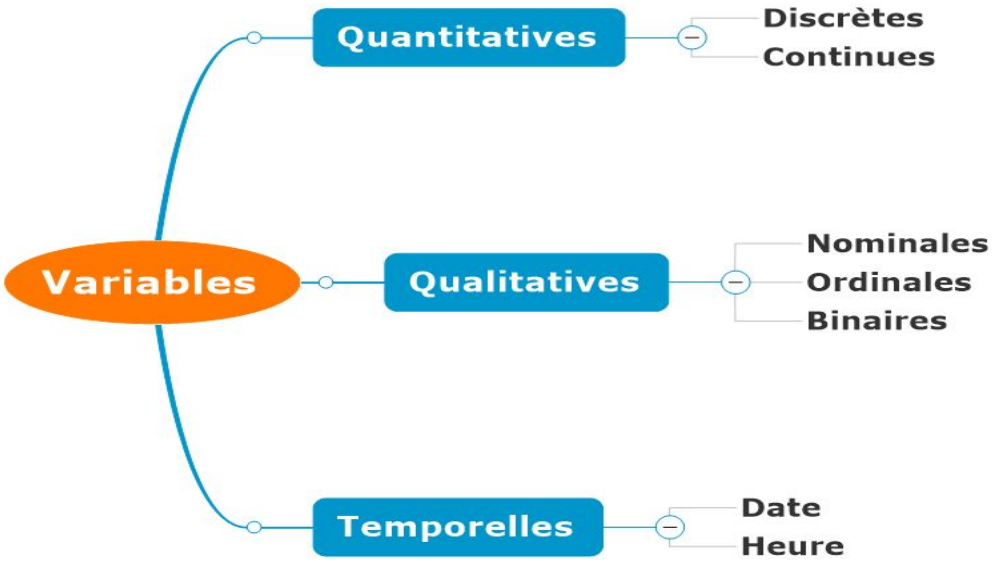
## Introduction :

Covid 19 est un nouveau virus qui a provoqué une pandémie et pour aider à gérer cette pandémie, nous devons analyser les données pour éviter la propagation du virus, mieux gérer la situation sanitaire et freiner la crise économique.Cela c'est le rôle des statisticiens qui utilisent les statistiques, les probabilité et des notions mathématiques afin de mieux expliquer les données et proposer des hypothèses notamment pour faire des prédictions.

Dans ce projet,nous allons essayer d'analyser cette base de données de l'Indonésie au cours du mois de mars.

## 1-Recherche des données :

On appelle variable une caractéristique commune à l'ensemble des individus d'une étude. La valeur de cette caractéristique varie entre les individus. C'est pour cela que nous parlons de variables. On distingue les **variables qualitatives, quantitatives et temporelles**.




Notre base de données se compose de 11 paramètre et 893 patients avec différentes types variables, on va détailler chaque paramètre comme suit :

**Variables qualitatives** :

***Binaires :**

-Gender : le sex du patient (homme ou femme).

***Nominales :**

- Province : 10 villes de l'Indonésie.
- Nationality : tous les patients viennent de l'Indonésie
- Hospital : les différents hôpitaux de l'Indonésie qui ont accueilli les malades du covid 19.
- Currente state : il existe 3 types d'état du patient :

    *Released : patient libérée
    *Isolated : patient isolée
    *Deceased : patient décédé

**Variables quantitatives discrètes :**

- Patient_id : le numéro du patient
- Age : entre 2 ans et 86 ans
- Contacted with : le nombre des gens qui ont contacté ce patient.

**Variables temporelles :**

- Confirmed date :la date à laquelle le patient s'est confirmé qu'il était malade.
- Released date : :la date à laquelle le patient s'est libéré.
- Deceased date : la date de décès du patient

| | patient_id | gender | age | nationality | province | current_state | contacted_with | confirmed_date | released_date | deceased_date | hospital |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | female | 31.0 | indonesia | DKI Jakarta | released | NaN | 2-Mar-20 | 13-Mar-20 | NaN | RSPI Sulianti Saroso |
| 1 | 2 | female | 64.0 | indonesia | DKI Jakarta | released | 1.0 | 2-Mar-20 | 16-Mar-20 | NaN | RSPI Sulianti Saroso |
| 2 | 3 | female | 33.0 | indonesia | DKI Jakarta | released | 1.0 | 6-Mar-20 | 13-Mar-20 | NaN | RSPI Sulianti Saroso |
| 3 | 4 | female | 34.0 | indonesia | DKI Jakarta | isolated | 1.0 | 6-Mar-20 | NaN | NaN | RSPI Sulianti Saroso |
| 4 | 5 | male | 55.0 | indonesia | DKI Jakarta | isolated | 1.0 | 8-Mar-20 | NaN | NaN | RS Persahabatan |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 888 | 889 | NaN | NaN | NaN | NaN | NaN | NaN | 26-Mar-20 | NaN | NaN | NaN |
| 889 | 890 | NaN | NaN | NaN | NaN | NaN | NaN | 26-Mar-20 | NaN | NaN | NaN |
| 890 | 891 | NaN | NaN | NaN | NaN | NaN | NaN | 26-Mar-20 | NaN | NaN | NaN |
| 891 | 892 | NaN | NaN | NaN | NaN | NaN | NaN | 26-Mar-20 | NaN | NaN | NaN |
| 892 | 893 | NaN | NaN | NaN | NaN | NaN | NaN | 26-Mar-20 | NaN | NaN | NaN |

893 rows × 11 columns

## 2-Nettoyage des données :

Dans cette étape, on a supprimé les données inutiles dans notre analyse et on n'a gardé que les données pertinentes. Comme la base de données contient les patients du même pays qui est l'Indonésie on n'a pas besoin de garder le paramètre « nationality ». En plus de ça on a éliminé la colonne « patient_id » parce qu'il existe déjà les indices des patients « lignes » dans python.

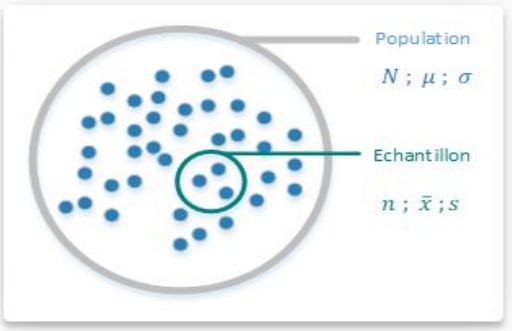
| | gender | age | province | current_state | contacted_with | confirmed_date | released_date | deceased_date | hospital |
|---|---|---|---|---|---|---|---|---|---|
| 0 | female | 31.0 | DKI Jakarta | released | NaN | 2-Mar-20 | 13-Mar-20 | NaN | RSPI Sulianti Saroso |
| 1 | female | 64.0 | DKI Jakarta | released | 1.0 | 2-Mar-20 | 16-Mar-20 | NaN | RSPI Sulianti Saroso |
| 2 | female | 33.0 | DKI Jakarta | released | 1.0 | 6-Mar-20 | 13-Mar-20 | NaN | RSPI Sulianti Saroso |
| 3 | female | 34.0 | DKI Jakarta | isolated | 1.0 | 6-Mar-20 | NaN | NaN | RSPI Sulianti Saroso |
| 4 | male | 55.0 | DKI Jakarta | isolated | 1.0 | 8-Mar-20 | NaN | NaN | RS Persahabatan |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 888 | NaN | NaN | NaN | NaN | NaN | 26-Mar-20 | NaN | NaN | NaN |
| 889 | NaN | NaN | NaN | NaN | NaN | 26-Mar-20 | NaN | NaN | NaN |
| 890 | NaN | NaN | NaN | NaN | NaN | 26-Mar-20 | NaN | NaN | NaN |
| 891 | NaN | NaN | NaN | NaN | NaN | 26-Mar-20 | NaN | NaN | NaN |
| 892 | NaN | NaN | NaN | NaN | NaN | 26-Mar-20 | NaN | NaN | NaN |

893 rows × 9 columns

## 3-Analyse des données :

Une population : L'ensemble des unités considérées par le statisticien. dans notre projet représente l'ensemble des patients affectés par le covid 19 au mois de mars dans l'Indonésie.

Un échantillon : Sous-ensemble de la population choisi pour l'étude. Un échantillon est par nature incomplet.



| Correspondance | Population | Echantillon |
|---|---|---|
| Moyenne | $\mu$ | $\bar{x}$ |
| Variance | $\alpha^2$ | $s^2$ |
| Ecart-type | $\alpha$ | s |
| Taille | *N* | n |
| Pourcentage | $\pi$ | *p* |

Dans cette étape, on va essayer de trouver des relations entres les colonnes et les lignes afin de comprendre ce qui se passe en Indonésie pendant la pandémie en utilisant des modèles ( graphes,diagrammes,piechart ...) pour l'extraction de la connaissance à partir de données qui sert à prédire l'évolution de la maladie et de prendre les bon mesures pour gérer la pandémie en indonésie.

**A - La relation entre chaque province et l'âge des patient affecté par le covid19:**

1 - l'âge min est l'âge minimum dans chaque province

2 - l'age max est l'âge maximum dans chaque province

3 - la médiane est l'âge médiane dans chaque province

4 - l'âge_mean est la moyenne d'âge des patients affectés par covid 19 dans chaque province

5- la variance : la dispersion des valeurs d'âge.

6- l'écart-type : mesure de la dispersion autour de la moyenne

| | province | Age_mean | Age_min | Age_max | Median | Variance | Ecart-type |
|---|---|---|---|---|---|---|---|
| 0 | Bali | 53.000000 | 53.0 | 53.0 | 53.0 | NaN | NaN |
| 1 | Banten | 51.666667 | 35.0 | 63.0 | 56.0 | 138.666667 | 11.775681 |
| 2 | DI Yogyakarta | 3.000000 | 3.0 | 3.0 | 3.0 | NaN | NaN |
| 3 | DKI Jakarta | 48.596774 | 2.0 | 86.0 | 48.5 | 285.250721 | 16.889367 |
| 4 | Jawa Barat | 45.333333 | 17.0 | 67.0 | 43.0 | 268.250000 | 16.378339 |
| 5 | Jawa Tengah | 51.750000 | 43.0 | 60.0 | 52.0 | 80.916667 | 8.995369 |
| 6 | Jawa Timur | 46.142857 | 22.0 | 56.0 | 49.0 | 132.476190 | 11.509830 |
| 7 | Kalimantan Barat | 26.500000 | 19.0 | 34.0 | 26.5 | 112.500000 | 10.606602 |
| 8 | Kepulauan Riau | 71.000000 | 71.0 | 71.0 | 71.0 | NaN | NaN |
| 9 | Sulawesi Utara | 51.000000 | 51.0 | 51.0 | 51.0 | NaN | NaN |

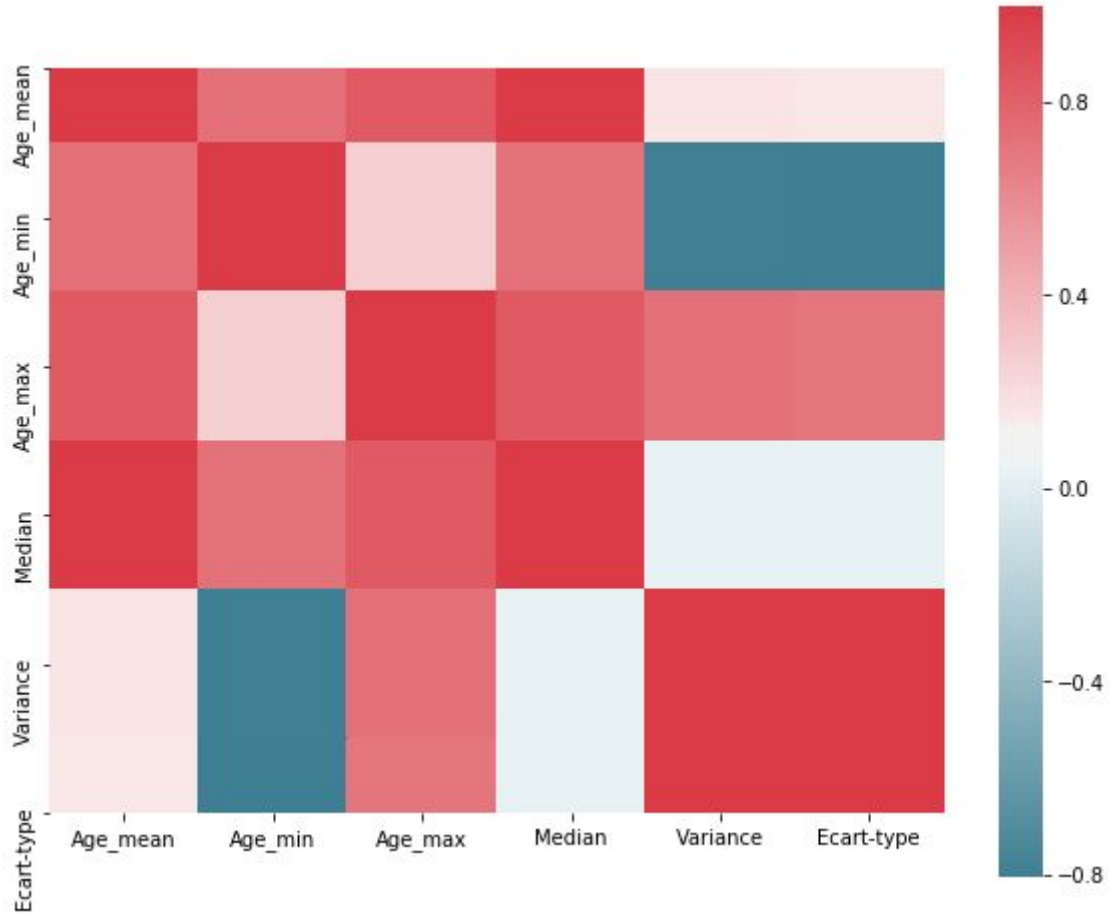
Nous trouvons de cette analyse que les âges diffèrent beaucoup mais la plupart des variantes ont entre 17 et 60 ans et cela est prouvé avec la médiane où les âges sont autour de la cinquantaine et la raison peut être due au déplacement de cette tranche dans les transports publics , supermarchés … .

On voit que la variance et l'écart type d'âge dans la province "Jawa Tangah" sont très bas avec des valeurs 80.91 et 8.99 respectivement ,ça veut dire que la différence entre les âges n'est pas grande par contre la variance d'âge et l'écart-type du province de "Dki Jakarta" est très grande avec des valeurs 285.25 et 16.88 respectivement et cela veut dire que les âges sont trop dispersé.

7-corrélation de pearson:

En probabilités et en statistique, la corrélation entre plusieurs variables aléatoires ou est une notion de liaison qui contredit leur indépendance.quotient de leur covariance par le produit de leurs écarts types.La valeur absolue du coefficient, toujours comprise entre 0 et 1.

| | Age_mean | Age_min | Age_max | Median | Variance | Ecart-type |
|---|---|---|---|---|---|---|
| Age_mean | 1.000000 | 0.728058 | 0.849039 | 0.995194 | 0.162809 | 0.148377 |
| Age_min | 0.728058 | 1.000000 | 0.267484 | 0.723593 | -0.795324 | -0.803360 |
| Age_max | 0.849039 | 0.267484 | 1.000000 | 0.838382 | 0.726806 | 0.707073 |
| Median | 0.995194 | 0.723593 | 0.838382 | 1.000000 | 0.034623 | 0.031249 |
| Variance | 0.162809 | -0.795324 | 0.726806 | 0.034623 | 1.000000 | 0.997436 |
| Ecart-type | 0.148377 | -0.803360 | 0.707073 | 0.031249 | 0.997436 | 1.000000 |




La corrélation entre les différents âges(écart type, moyenne, …) est élevée.

**Similarité :**

Âges avec une corrélation élevée:

age moyen et médian

Âges avec corrélation moyenne:

écart type et médiane

variance et médiane

Âges avec corrélation faible:

âge min et écart type

âge min et variance

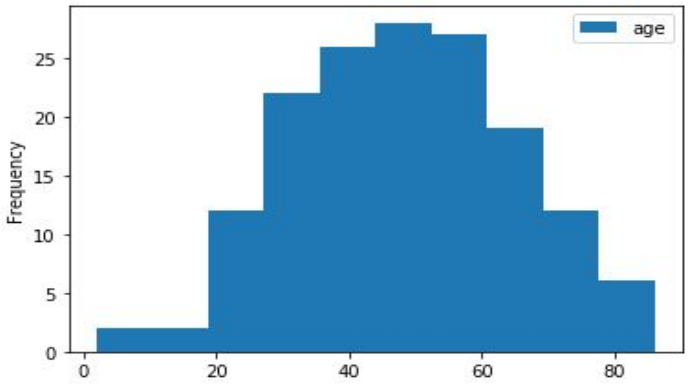
**C - La relation entre l'age et le sex des patients :**

| | gender | Age_mean | Age_min | Age_max | Median | Variance | Ecart-type |
|---|---|---|---|---|---|---|---|
| 0 | female | 47.292308 | 16.0 | 80.0 | 46.0 | 305.085096 | 17.466685 |
| 1 | male | 48.692308 | 2.0 | 86.0 | 50.0 | 258.393162 | 16.074612 |

On voit que l'âge moyen des femmes est autour de 48 ans et pareil pour les hommes.concernant la variance, on remarque qu'il y a une grande dispersion dans l'âge pour les 2 sexes.

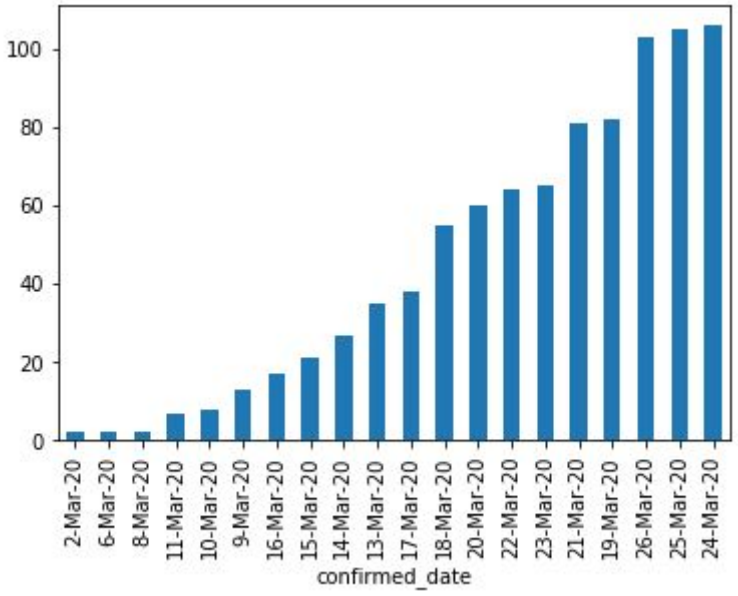
**D - Le nombre des patients affectés par le Covid-19 par tranche d'âge:**

```
Nombre des patients affectés par le Covid-19 qui ont age < 19 :  4
Nombre des patients affectés par le Covid-19 qui ont age entre 19 et 65 :  120
Nombre des patients affectés par le Covid-19 qui ont age > 65 :  26
```



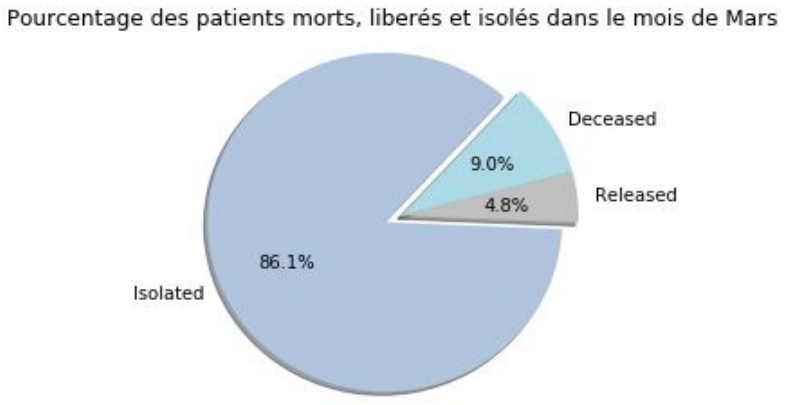

On constate que les personne moins de 19 ans sont peu affecté par le covid 19 et aussi les personne âgé, et on pense que c'est parce qu'ils bougent pas trop dans la vie quotidienne , on voit souvent la tranche d'âge entre 20 et 60 et ça c'est notre hypothèse pourquoi il sont plus affecté que les plus jeunes et ça peut être vrai puisque la première chose que les gouvernements font est le confinement afin de limiter les mouvements des gens et donc limiter la dispersion du virus.

**E - Le nombre des patients affectés par Covid 19 par jour :**




Le nombre de cas s'augmente intensivement dans les deux semaines qui viennent et la raison pour ça c'est que le taux de contamination est très élevé par jours car les symptômes n' apparaître que dans les deux semaine qui viennent après la contamination et les gens continue de disperser le virus sans qu'ils sachent.

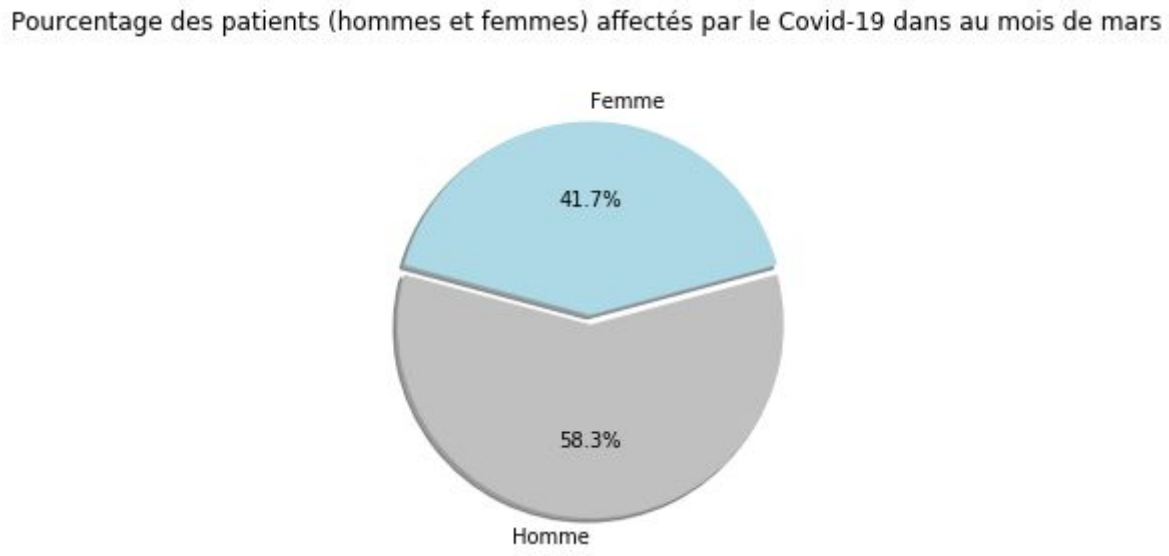
**F - Le pourcentage des patients libérés,isolés et décès au mois de mars :**



On voit que le nombre de personnes isolées est bien supérieur au nombre de décès et libérés, est tout à fait normal car le patient reste entre une semaine et 3 semaines à l'hôpital. Le pourcentage de personnes libérées est faible par rapport au décès est parce que la plupart des patients sont encore isolés, et ce pourcentage des patients libéré augmentera par rapport au décès dans les semaines à venir parce qu'on ne voit les résultats de la semaine que dans les deux semaines suivantes.

Mais le fait que le nombre des patients isolés est très grand ça peut causer une saturation dans les lits des hôpitaux ce qui va causer une augmentation intensif des morts dans les semaines suivantes.

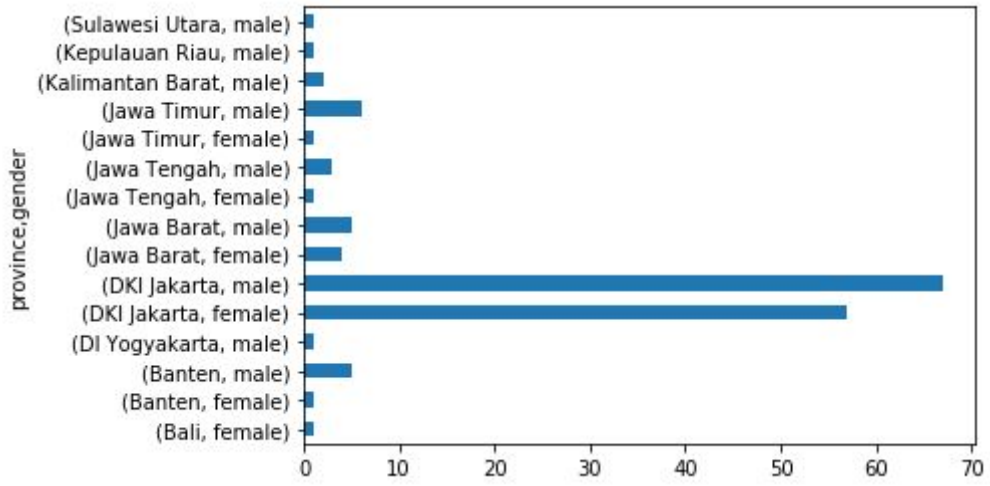
**G - Le pourcentage des patients hommes/femmes affectés par le covid 19:**




D'après la comparaison entre le nombre de femmes et d'hommes infectés par le covid 19 on voit que les femmes sont plus touchées que les hommes, et la raison principale que les savant disent c'est que les femmes ont un système immunitaire plus fort que les hommes mais il y a une autre l'hypothèse selon laquelle les hommes ne prennent tout simplement pas soin de leur corps, avec des niveaux plus élevés de tabagisme, de consommation d'alcool et d'obésité et c'est la même constaté en chine et en italie.

**H- Patient affecté par le covid " homme et femme" par province :**

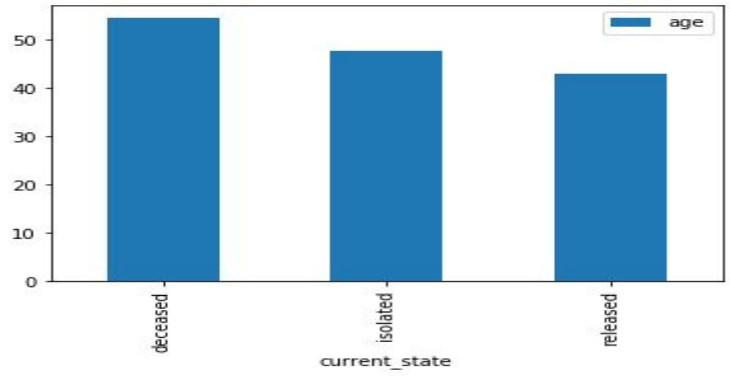
| province | gender | |
|---|---|---|
| Bali | female | 1 |
| Banten | female | 1 |
| | male | 5 |
| DI Yogyakarta | male | 1 |
| DKI Jakarta | female | 57 |
| | male | 67 |
| Jawa Barat | female | 4 |
| | male | 5 |
| Jawa Tengah | female | 1 |
| | male | 3 |
| Jawa Timur | female | 1 |
| | male | 6 |
| Kalimantan Barat | male | 2 |
| Kepulauan Riau | male | 1 |
| Sulawesi Utara | male | 1 |




On voit que même dans chaque région, le nombre d'hommes touchés est toujours supérieur au nombre de femmes, et cela est cohérent avec les résultats de contamination de l'ensemble du pays ( hommes > femmes ).

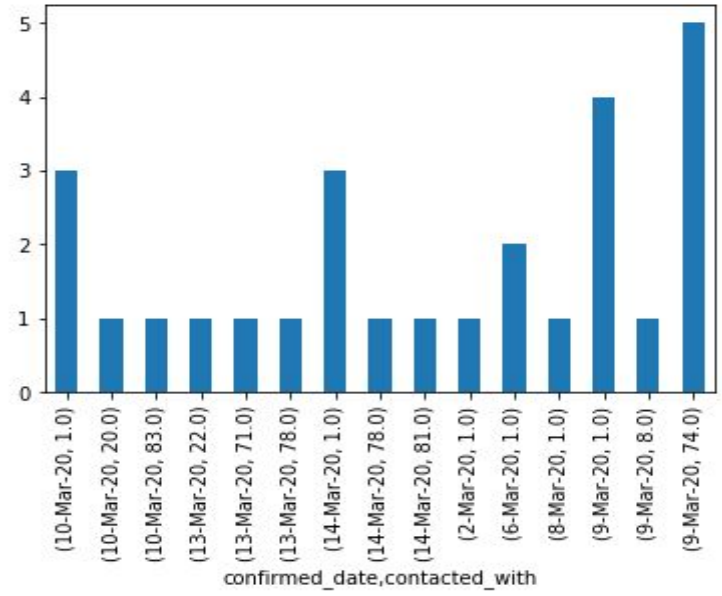
**I- La moyenne d'âge des patients morts /isolés /libérés :**

| current_state | age |
|---|---|
| deceased | 54.466667 |
| isolated | 47.699248 |
| released | 43.000000 |




On voit que l'âge moyen des patients décédés est d'environ la cinquantaine, et on peut en déduire que les personnes âgées sont plus susceptibles de mourir, et c'est peut-être à cause de leur système immunitaire qui est plus faible que les jeunes.

**J- Le nombre de contamination Par jour :**




2 mars : 1

6 mars : 2

8 mars : 1

9 mars : 382

10 mars : 106

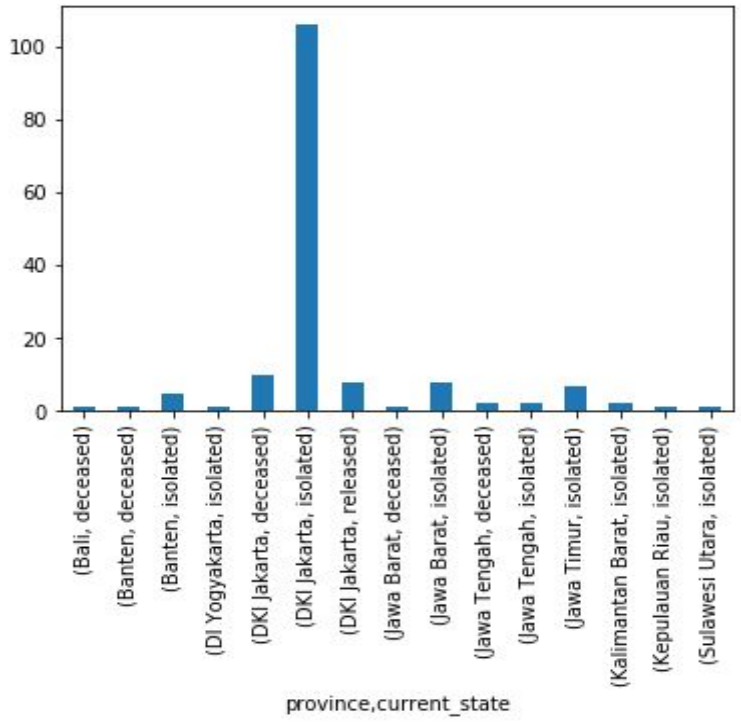
13 mars : 171

14 mars : 162

On constate que les premier jours de mars la contamination était faible mais quelques jours après les contaminations sont devenu intensifs, donc ça montre qu'il y aura de forte contaminations les semaines prochaine.

**K-Le nombre des patients morts,isolés et libérés par province :**

| province | current_state | |
|---|---|---|
| Bali | deceased | 1 |
| Banten | deceased | 1 |
| | isolated | 5 |
| DI Yogyakarta | isolated | 1 |
| DKI Jakarta | deceased | 10 |
| | isolated | 106 |
| | released | 8 |
| Jawa Barat | deceased | 1 |
| | isolated | 8 |
| Jawa Tengah | deceased | 2 |
| | isolated | 2 |
| Jawa Timur | isolated | 7 |
| Kalimantan Barat | isolated | 2 |
| Kepulauan Riau | isolated | 1 |
| Sulawesi Utara | isolated | 1 |




On voit que la ville la plus touchée est la ville de Jakarta(nombr de morts est 10, nombre des patients isolés est 106), et c'est la capitale, ce qui est normal car les capitales sont les villes les plus peuplées du monde, donc le nombre de contaminations sera forcément plus élevé.

## Conclusion :

Le mois à traiter est le début de la pandémie, il est donc normal que le nombre de décès et de contaminations soit assez faible par rapport au mois suivant, mais cela donne une idée de ce qui va se passer dans les prochaines semaines, et puisque les contaminations se sont intensifiées chaque jour, il est très probable que les mois à venir seront très difficiles pour l'Indonésie et le pays doit prendre ses propres précautions en augmentant le nombre de lits de réanimation, le nombre d'agents de santé qui prennent en charge les malades et prennent les bonnes décisions pour gérer la situation sanitaire.