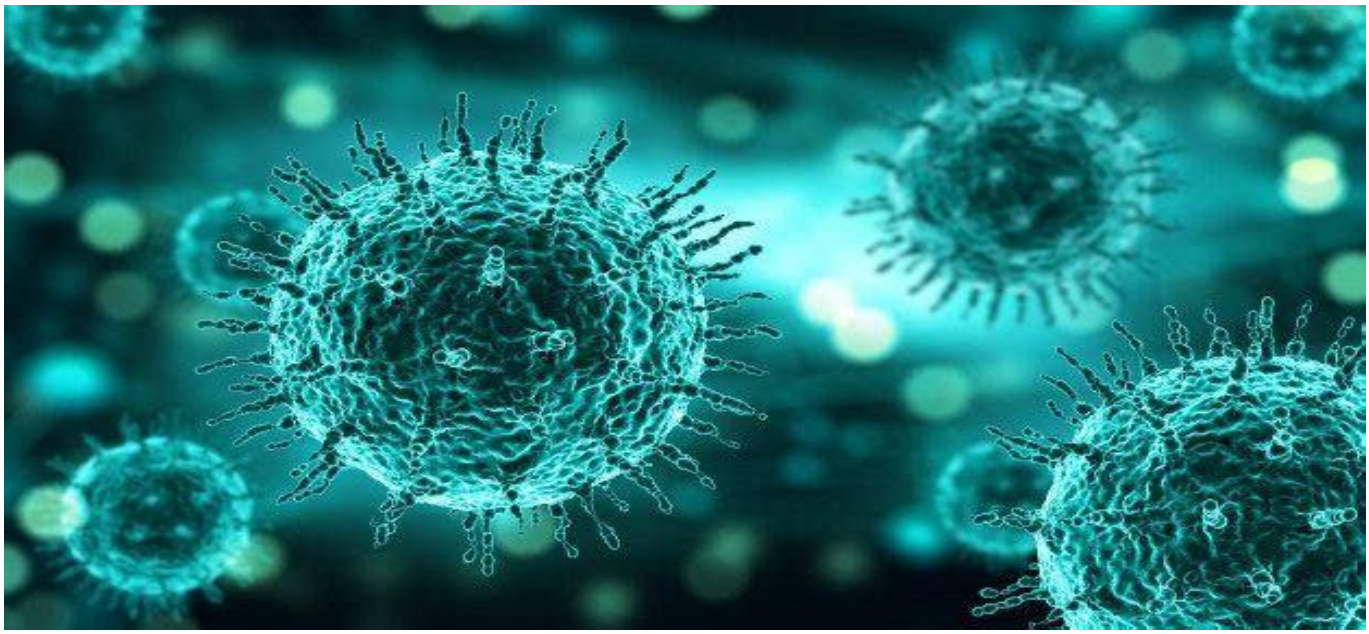
UNIVERSITE MOHAMMED V – RABAT

ECOLE MOHAMMADIA D'INGENIEURS

DEPARTEMENT GENIE INFORMATIQUE

Rapport : Projet R La Pandémie du Corona Virus



Réalisé Par :

Mehdi HARIM

Aymane KENBOUCH

Youssef ARGANE

Mohamed AHOU

Encadré Par :

Asmae EL KASSIRI

Table de Matière :

I. Introduction :

Le COVID-19, causé par le virus SARS-CoV-2, a été identifié pour la première fois en décembre 2019 et s'est rapidement propagé à travers le monde, déclenchant une pandémie mondiale. Le Maroc, comme beaucoup d'autres pays, a été touché par cette crise sanitaire, avec des conséquences significatives sur la santé publique et l'économie. Suivre l'évolution des cas et des décès liés au COVID-19 est essentiel pour informer les politiques de santé publique et les mesures de prévention.

Dans cette étude, nous allons procéder à un nettoyage des données (data cleaning) afin de préparer notre jeu de données. Ensuite, nous effectuerons des analyses descriptives pour calculer des statistiques et fournir un aperçu général de la situation épidémiologique au Maroc et l'état de son système santé. Nous visualiserons ces données à travers des graphiques illustrant l'évolution des différents indicateurs de la pandémie. Nous réaliserons également des analyses statistiques pour tester la validité d'une hypothèse nulle. Enfin, nous construirons des modèles de régression linéaire pour comprendre les relations entre certaines variables clés.

II. Data Cleaning :

Une étape très importante c'est le nettoyage de la data pour qu'on sera capable à la fin d'extraire des résultats significatifs, on import les libraires importantes pour la visualisation aussi .

```r
pacman::p_load(pacman, dplyr, rio, ggplot2, tidyr,scales)
data <- read.csv("C:/Users/MEHDI/Documents/JustTesting/owid-covid-data1.csv")
data_maroc <- subset(data, location == "Morocco")
data_maroc$gueris <- data_maroc$total_cases - data_maroc$total_deaths
data_maroc$new_gueris <- data_maroc$new_cases - data_maroc$new_deaths
replace <- c("total_cases", "total_deaths", "new_deaths", "new_cases"  , "total_tests" , "reproduction_rate", "gueris", "new_gueris" )
data_maroc[columns_to_replace] <- lapply(data_maroc[replace], function(x) {
  x[is.na(x)] <- 0
  return(x)
})
data_maroc$date <- as.Date(data_maroc$date)
Sys.setlocale("LC_TIME", "fr_FR.UTF-8")
data_maroc <- subset(data_maroc, weekdays(date) == "dimanche")
data_maroc <- data_maroc[, c("total_cases", "new_cases", "total_deaths", "new_deaths", "gueris", "new_gueris", "date" , "total_tests","reproduction_rate" )]

```

D'abord, nous extrairons les données concernant le Maroc. Ensuite, nous ajouterons une colonne des guéris en soustrayant le nombre de morts du nombre de cas. Nous créerons un vecteur contenant les colonnes avec lesquelles nous allons travailler et remplacerons les valeurs NA par 0 , Nous mettrons également la date de l'environnement RStudio en français et utiliserons la fonction (subset) pour ne conserver que les données des dimanches.

	total_cases	new_cases	total_deaths	new_deaths	gueris	new_gueris	date	total_tests	reproduction_rate
235102	0	0	0	0	0	0	2020-01-05	0	0.00
235109	0	0	0	0	0	0	2020-01-12	0	0.00
235116	0	0	0	0	0	0	2020-01-19	0	0.00
235123	0	0	0	0	0	0	2020-01-26	0	0.00
235130	0	0	0	0	0	0	2020-02-02	0	0.00
235137	0	0	0	0	0	0	2020-02-09	0	0.00
235144	0	0	0	0	0	0	2020-02-16	0	0.00
235151	0	0	0	0	0	0	2020-02-23	13	0.00
235158	0	0	0	0	0	0	2020-03-01	0	0.00
235165	2	2	0	0	0	2	2020-03-08	0	0.00
235172	17	15	1	1	16	14	2020-03-15	0	0.00
235179	96	79	3	2	93	77	2020-03-22	627	0.00
235186	390	294	25	22	365	272	2020-03-29	2273	1.78
235193	919	529	59	34	860	495	2020-04-05	4848	1.50
235200	1545	626	111	52	1434	574	2020-04-12	8604	1.38
235207	2685	1140	137	26	2548	1114	2020-04-19	15123	1.33
235214	3897	1212	159	22	3738	1190	2020-04-26	27399	0.96
235221	4729	832	173	14	4556	818	2020-05-03	42112	1.07
235228	5910	1181	186	13	5724	1168	2020-05-10	65924	1.00
235235	6741	831	192	6	6549	825	2020-05-17	89957	0.77
235242	7406	665	198	6	7208	659	2020-05-24	142882	0.70
235249	7780	374	204	6	7576	368	2020-05-31	209139	0.75
235256	8151	371	208	4	7943	367	2020-06-07	305953	1.17
235263	8692	541	212	4	8480	537	2020-06-14	422520	1.35
235270	9839	1147	213	1	9626	1146	2020-06-21	538191	1.84
235277	11877	2038	220	7	11657	2031	2020-06-28	646195	1.26
235284	13822	1945	232	12	13590	1933	2020-07-05	765580	1.15
235291	15542	1720	245	13	15297	1707	2020-07-12	878626	0.92

III. Analyses Statistiques Descriptives :

On commence par créer la fonction calcule_descriptives, qui calcule pour une colonne le min, le max, le mean (moyenne) et l'écart-type.

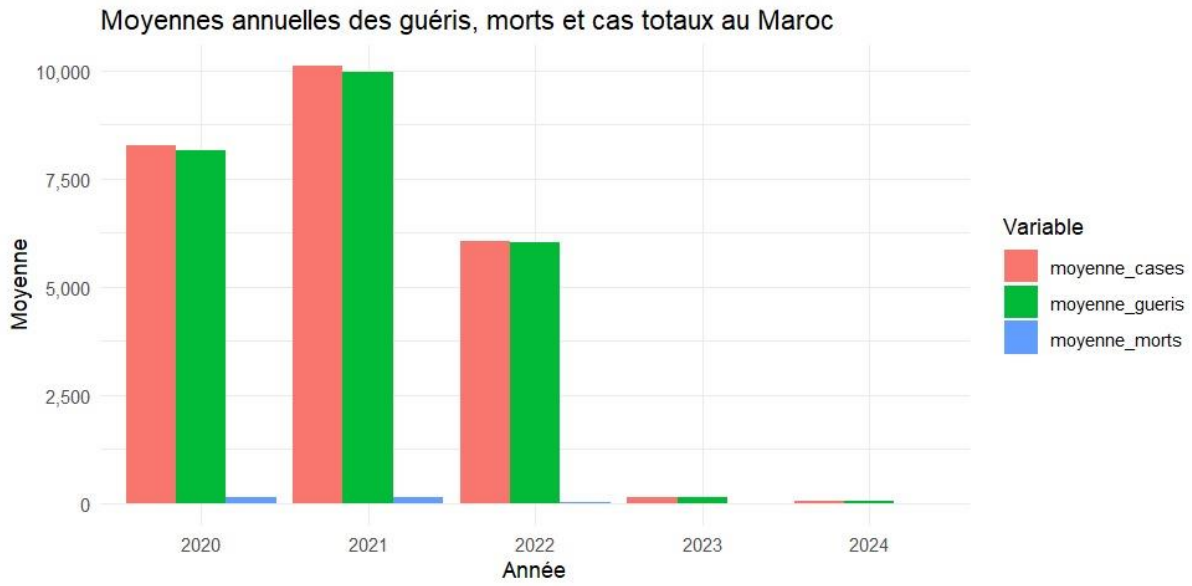
```
calculate_descriptives <- function(column) {
  min_val <- min(column)
  max_val <- max(column)
  mean_val <- mean(column)
  sd_val <- sd(column)
  return(c(min = min_val, max = max_val, mean = mean_val, sd = sd_val))
}
descriptive_stats <- sapply(data_maroc, calculate_descriptives)
descriptive_stats
```

```
     total_cases new_cases total_deaths new_deaths    gueris new_gueris      date total_tests reproduction_rate
min          0.0     0.000        0.000    0.00000       0.0       0.00 18266.000          0          0.0000000
max    1279115.0 64784.000    16305.000  744.00000 1262810.0   64106.00 19848.000   11738659          2.1900000
mean    847020.7  5634.868    11588.978   71.82819  835431.7    5563.04 19057.000    2529984          0.6566520
sd      492336.9 11408.380     6164.523  142.40675  486247.4   11291.57   459.714    3683713          0.5706767
>
```

On remarque que l'écart-type des variables est grand, ce qui signifie que la dispersion des valeurs est importante. De plus, au niveau des valeurs maximales et minimales, les variables "total des cas" et "total des guéris" coïncident, ce qui peut montrer qu'elles ont le même comportement.

IV. Visualisation de données :

1. Moyennes annuelles des guéris, morts et cas totaux au Maroc :



On remarque que les moyennes des nouveaux cas et des guéris sont très proches alors que le nombre de nouveaux morts par rapport aux ces deux est négligeable, De plus, on observe un pic en 2021.

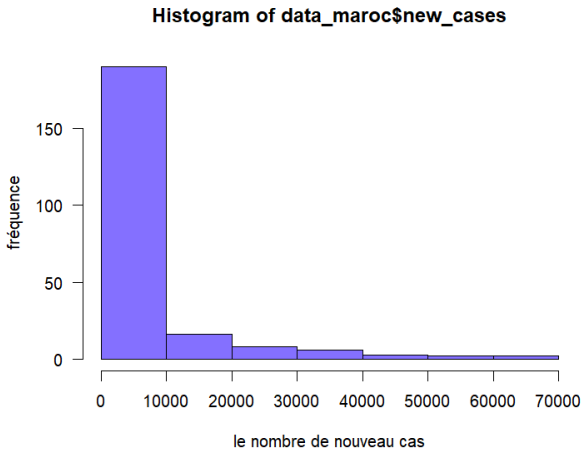
Le code :

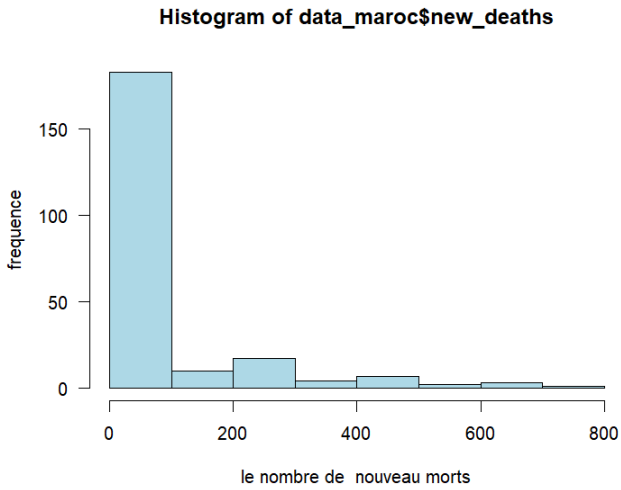
D'abord, nous avons transformé le dataset data_maroc en ajoutant une colonne d'année extraite des dates. Ensuite, regrouper les données par année et calculer les moyennes annuelles pour chaque variable. Ces moyennes sont réorganisées en un format long pour faciliter la visualisation.

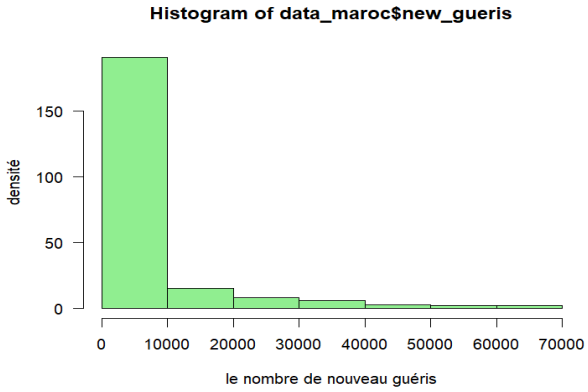
```r
moyennes_long <- data_maroc %>%
  mutate(year = format(date, "%Y")) %>%
  group_by(year) %>%
  summarise(
    moyenne_gueris = mean(new_gueris),
    moyenne_morts = mean(new_deaths),
    moyenne_cases = mean(new_cases)
  ) %>%
  gather(
    key = "variable",
    value = "value",
    -year
  )
ggplot(moyennes_long, aes(x = year, y = value, fill = variable)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Moyennes annuelles des guéris, morts et cas totaux au Maroc",
    x = "Année", y = "Moyenne", fill = "Variable"
  ) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
```

2. Histogrammes de Totales des Cas, des morts et des guéries :

```r
hist(data_maroc$new_cases,
     xlab = "le nombre de nouveau cas",
     ylab = "fréquence",
     las = 1,
     col="lightslateblue")

hist(data_maroc$new_deaths,
     xlab = "le nombre de  nouveau morts",
     ylab = "frequence",
     col="lightblue",
     las = 1)

hist(data_maroc$new_gueris,
     xlab ="le nombre de nouveau guéris",
     ylab = "densité",
     col ="lightgreen",
     las = 1)
```



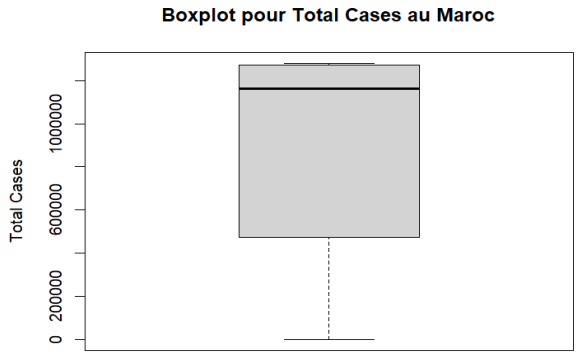


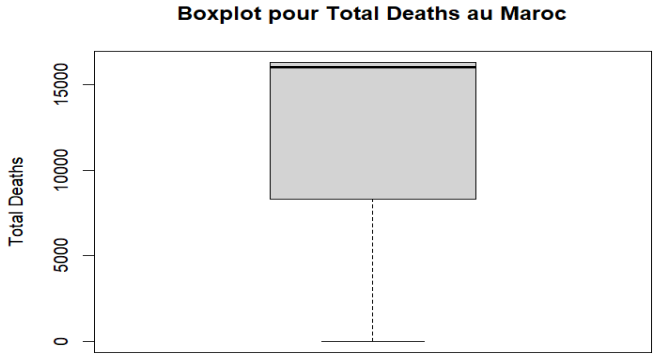


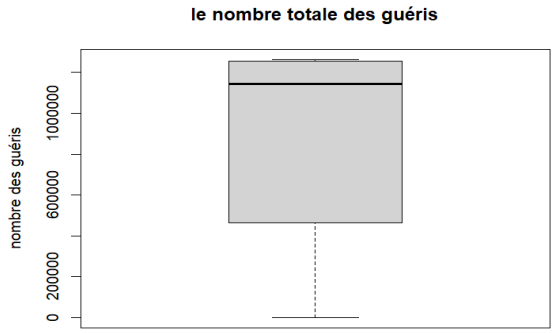
On remarque que les valeurs entre 0 et 1000 sont les plus fréquentes par rapport aux observations générales, contrairement aux valeurs élevées, ce qui indique une augmentation exponentielle des valeurs.

3. Boxplots des totales des cas, des morts et des guéries :

```r
boxplot(data_maroc$total_cases,
        main = "Boxplot pour Total Cases au Maroc",
        ylab = "Total Cases")

boxplot(data_maroc$total_deaths,
        main = "Boxplot pour Total Deaths au Maroc",
        ylab = "Total Deaths")

boxplot(data_maroc$gueris,
        main ="le nombre totale des guéris",
        ylab = "nombre des guéris")
```





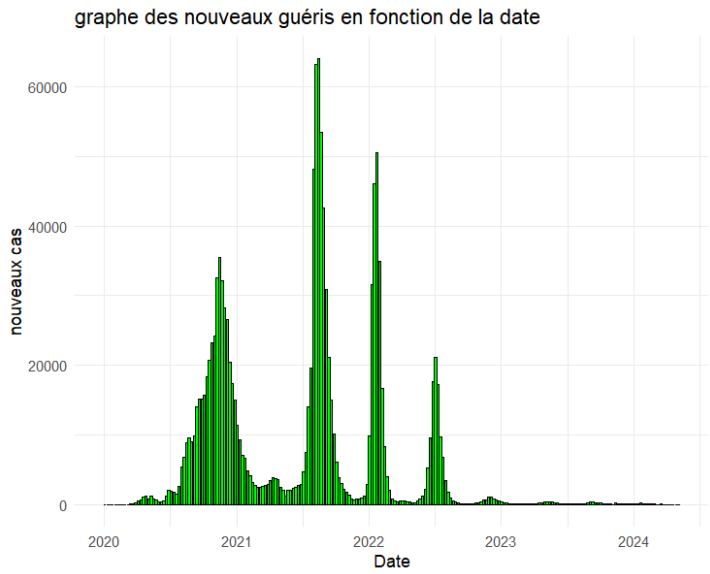


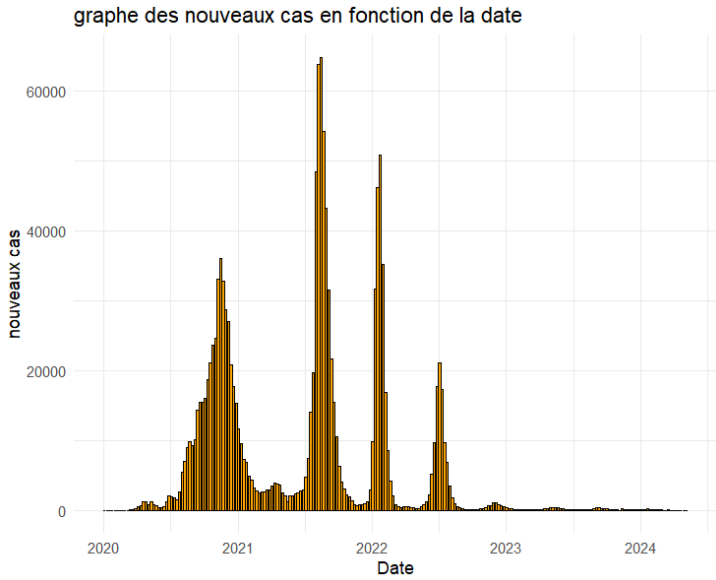
On voit que les valeurs sont dispersées autour de la médiane, avec une prédominance des petites valeurs, car la majorité des valeurs qui est sous le médians sont les valeurs petites, ce qui concorde avec les résultats des histogrammes.

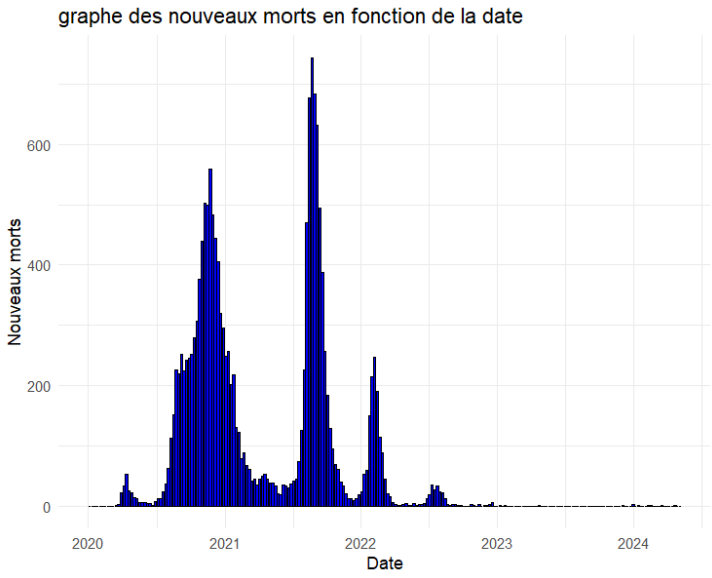
4. Diagramme des barres :

   Objectif Exploratoire :

```r
ggplot(data_maroc, aes(x = date, y = new_deaths)) +
  geom_col(fill = "blue", color = "black") +
  labs(x = "Date", y = "Nouveaux morts", title = "graphe des nouveaux morts en fonction de la date") +
  theme_minimal()

ggplot(data_maroc, aes(x = date, y = new_cases)) +
  geom_col(fill = "orange", color = "black") +
  labs(x = "Date", y = "nouveaux cas", title = "graphe des nouveaux cas en fonction de la date") +
  theme_minimal()

ggplot(data_maroc, aes(x = date, y = new_gueris)) +
  geom_col(fill = "green", color = "black") +
  labs(x = "Date", y = "nouveaux cas", title = "graphe des nouveaux guéris en fonction de la date") +
  theme_minimal()

```



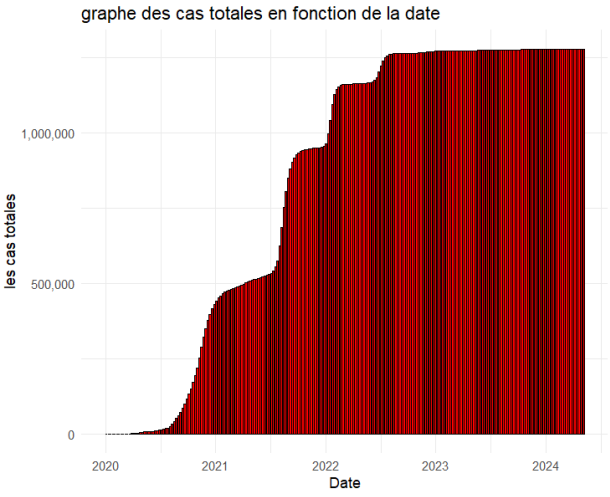


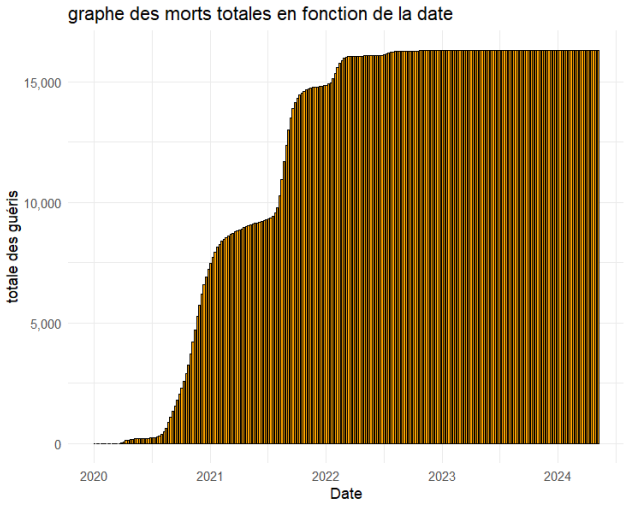


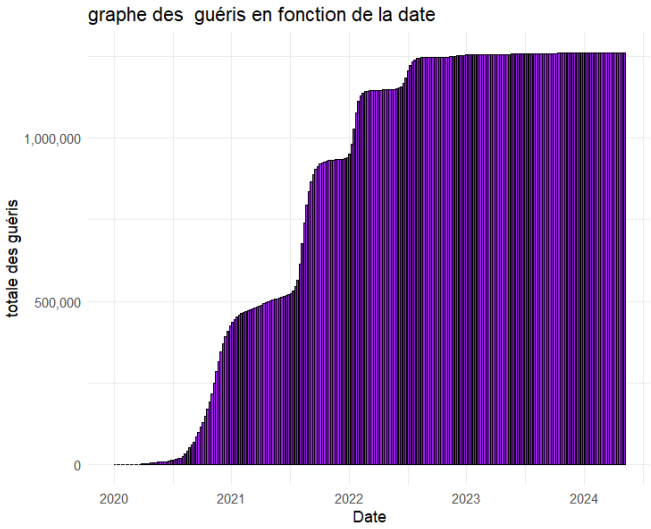
On remarque la présence de multiples vagues (courbes plurimodales), ce qui peut être dû aux mutations répétées du virus ainsi qu'aux alternances de relâchements et d'application des lois par l'État. On remarque aussi que le pic est atteint à la fin de l'année 2021, les Totales des cas, morts et guéris en fonction de la date :

5. Comparaison entre les totaux :

```r
ggplot(data_maroc, aes(x = date, y = total_cases)) +
  geom_col(fill = "red", color = "black") +
  labs(x = "Date", y = "les cas totales", title = "graphe des cas totales en fonction de la date") +
  theme_minimal()+
  scale_y_continuous(labels = scales::comma)

ggplot(data_maroc, aes(x = date, y = total_deaths)) +
  geom_col(fill = "orange", color = "black") +
  labs(x = "Date", y = "totale des guéris", title = "graphe des morts totales en fonction de la date") +
  theme_minimal()+
  scale_y_continuous(labels = scales::comma)

ggplot(data_maroc, aes(x = date, y = gueris)) +
  geom_col(fill = "purple", color = "black") +
  labs(x = "Date", y = "totale des guéris", title = "graphe des  guéris en fonction de la date") +
  theme_minimal()+
  scale_y_continuous(labels = scales::comma)
```



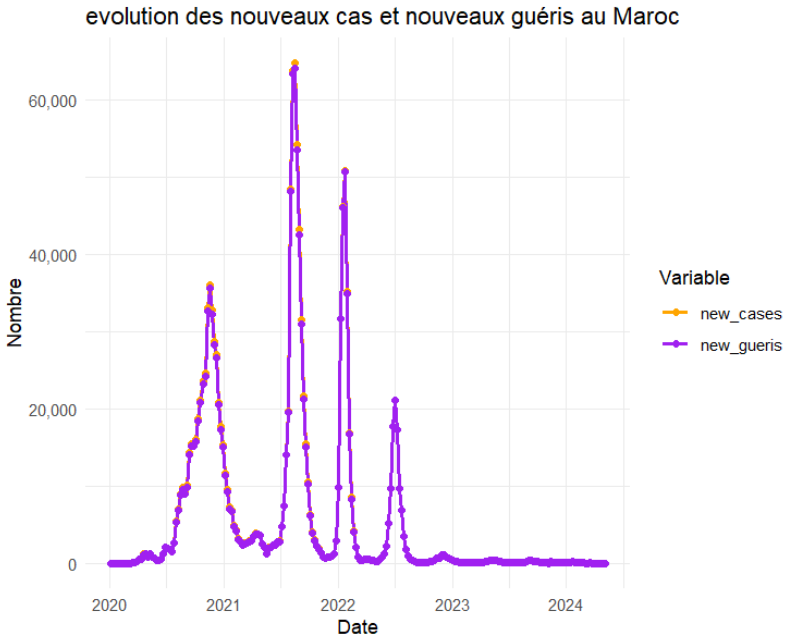




On remarque : que les variables des cas totale et total des guéris ont le même comportement

6. Evolution de nouveaux cas et nouveaux guéris au Maroc :

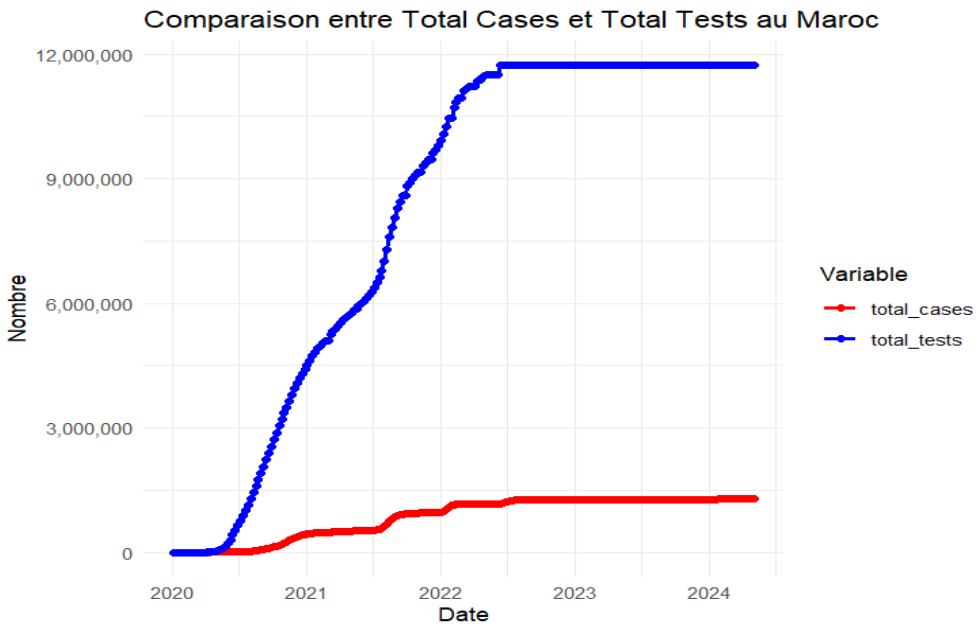
```r
data_maroc2 <- data_maroc %>% select(date, new_cases, new_gueris)
data_long <- gather(data_maroc2, key = "variable", value = "value", -date)
ggplot(data_long, aes(x = date, y = value, color = variable)) +
  geom_line(size = 1) +
  geom_point(size = 1.5) +
  labs(title = "evolution des nouveaux cas et nouveaux guéris au Maroc",
       x = "Date", y = "Nombre",
       color = "Variable") +
  theme_minimal() +
  scale_y_continuous(labels = scales::comma) +
  scale_color_manual(values = c("new_cases" = "orange", "new_gueris" = "purple"))

nonZero <- function(x) {
```



Observation : Les deux variables new cases et new guéries sont totalement superposées, Donc le même comportement.

7. Comparaison entre Totale des Cases et Totales des Testes Faites au Maroc :

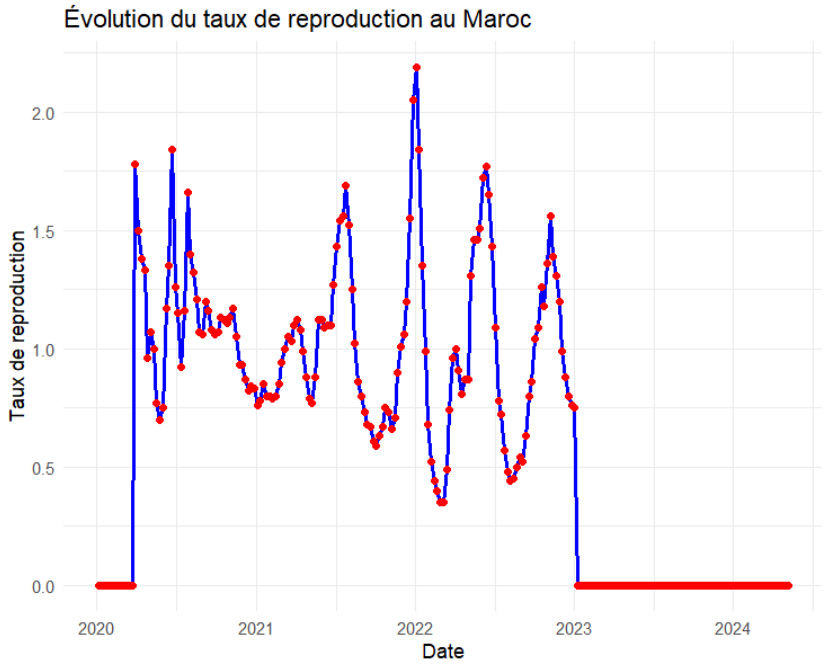
```r
nonZero <- function(x) {
  for (i in 2:length(x)) {
    if (x[i] == 0) {
      x[i] <- x[i - 1]
    }
  }
  return(x)
}
data_maroc$total_tests <- nonZero(data_maroc$total_tests)

data_maroc_long <- data_maroc %>%
  select(date, total_cases, total_tests) %>%
  gather(key = "variable", value = "value", -date)
ggplot(data_maroc_long, aes(x = date, y = value, color = variable)) +
  geom_line(size = 1) +
  geom_point(size = 1.5) +
  labs(title = "Comparaison entre Total Cases et Total Tests au Maroc",
       x = "Date", y = "Nombre",
       color = "Variable") +
  theme_minimal() +
  scale_y_continuous(labels = scales::comma) +
  scale_color_manual(values = c("total_cases" = "red", "total_tests" = "blue"))
```



Remarque : Lorsque la pandémie a été découverte, de nombreuses personnes ont commencé à se faire tester.

8. Evolution du taux de reproduction en fonction de la date au Maroc :

```
ggplot(data_maroc, aes(x = date, y = reproduction_rate)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 1.5) +
  labs(title = "Évolution du taux de reproduction au Maroc",
       x = "Date", y = "Taux de reproduction") +
  theme_minimal()
```



Taux de reproduction : est une mesure épidémiologique clé qui indique combien de personnes une personne infectée peut en infecter d'autres en moyenne.

Il existe deux principaux types de taux de reproduction :

a. Le Taux de reproduction de base (r0) : Il représente le taux de reproduction dans une population où personne n'est immunisé et aucune intervention n'a été mise en place. Un r0 supérieur à 1 signifie que l'infection peut se propager dans la population, tandis qu'un r0 inférieur à 1 indique que l'infection finira par s'éteindre.
b. Le Taux de reproduction effectif (re) : Il représente le taux de reproduction à un moment donné, prenant en compte les interventions (comme la distanciation sociale, le port de masques) et l'immunité acquise (par infection ou vaccination). Un re supérieur à 1 signifie que l'infection se propage, tandis qu'un re inférieur à 1 signifie que l'épidémie est sous contrôle.

Ici : le pic de valeur supérieur à 2 est atteint en 2022 qui signifie qu'une personne peut infecter plus que deux personnes

9. Le Taux de mortalité :

Objectif : Evaluer l'état de système santé au Maroc.

```
data_maroc$taux_mortalite <- ifelse(data_maroc$total_cases == 0, 0, data_maroc$total_deaths / data_maroc$total_cases)
ggplot(data_maroc, aes(x = date, y = taux_mortalite)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 1.5) +
  labs(title = "Évolution du taux de mortalité au Maroc",
       x = "Date", y = "Taux de mortalité") +
  theme_minimal()
```

Le Taux de mortalité d'une maladie est une mesure qui indique la proportion de décès parmi les cas confirmés de cette maladie.

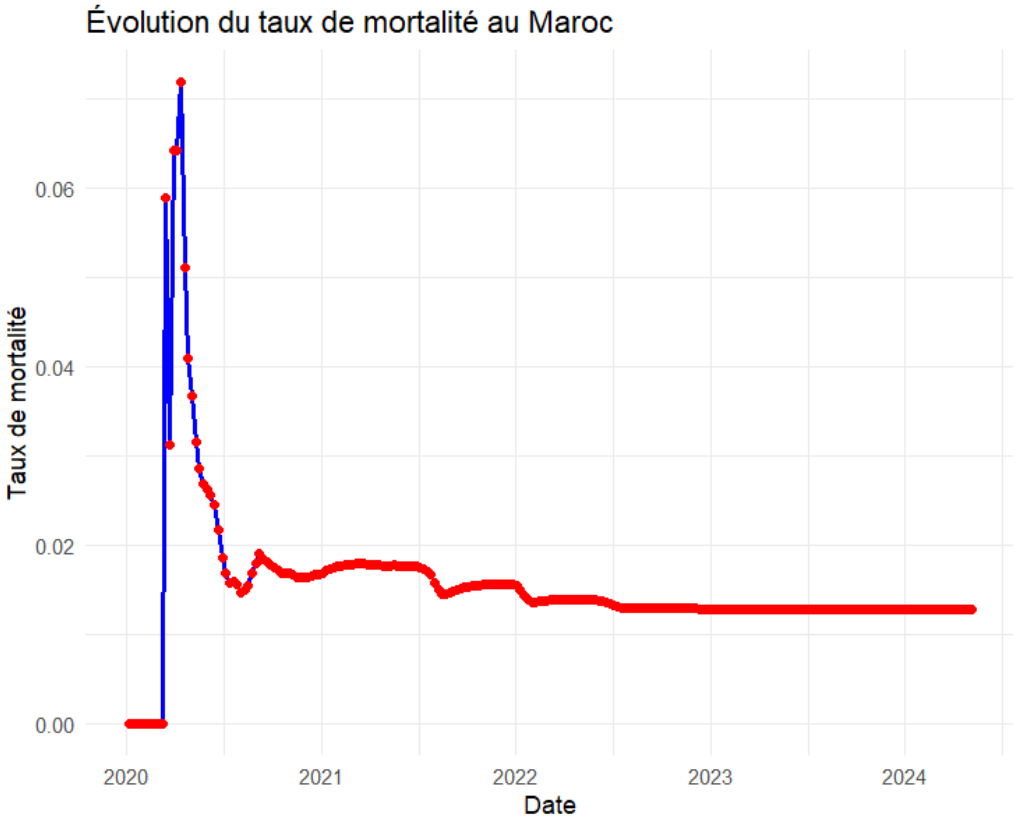
Dans le contexte de l'analyse des données au Maroc, ce taux permet de comprendre la gravité de la maladie et l'efficacité des systèmes de santé et des interventions médicales.

a. Un taux de mortalité élevé peut indiquer :

- Une forte virulence du virus.
- Des capacités limitées de traitement et d'hospitalisation.
- Des retards dans la détection et le traitement des cas.

b. Un taux de mortalité bas indique :

- Un bon système de soins de santé avec des capacités suffisantes.
- Des interventions médicales efficaces.
- Une population relativement moins vulnérable ou une meilleure gestion des cas.



Dans notre cas :

Le taux de mortalité a atteint son pic au début de 2020, indiquant que l'État a été pris par surprise par la pandémie et n'était pas préparé à l'affronter. Il y avait des capacités limitées en termes de traitement et d'hospitalisation pour faire face à la situation.

V. Les Coefficients de Corrélation Linéaire :

```
correlation1 <- cor(data_maroc$total_cases , data_maroc$gueris)
correlation2 <- cor (data_maroc$total_deaths , data_maroc$gueris)
correlation3 <- cor(data_maroc$total_cases , data_maroc$gueris)
correlation4 <- cor(data_maroc$new_cases , data_maroc$new_deaths)

print(correlation1)
print(correlation3)
print(correlation2)
print(correlation4)
```

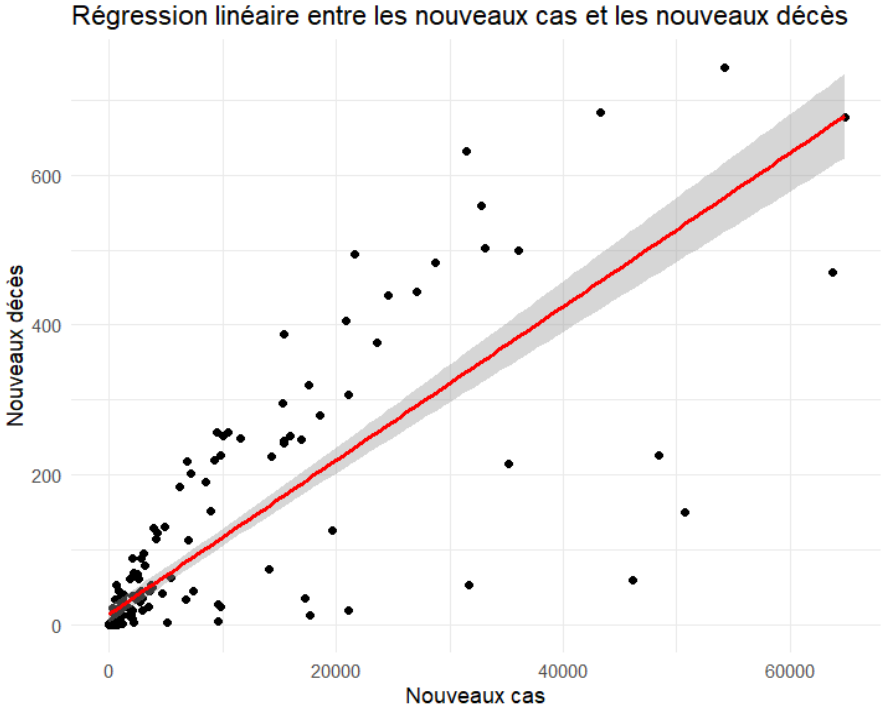
```
> print(correlation1)
[1] 0.9999981
> print(correlation3)
[1] 0.9999981
> print(correlation2)
[1] 0.9876717
> print(correlation4)
[1] 0.8223277
>
```

La corrélation du total des cas avec le total des décès ou le total des guérisons est toujours proche de 1, avec une valeur de 0.998. Pour la corrélation des nouveaux cas avec les nouveaux décès, elle est de 0.822. Pour la corrélation des nouveaux cas avec les nouvelles guérisons, elle est de 0.999. Cela signifie qu'il y a une forte corrélation entre les nouveaux cas et les nouvelles guérisons, plus qu'entre les nouveaux cas et les nouveaux décès.

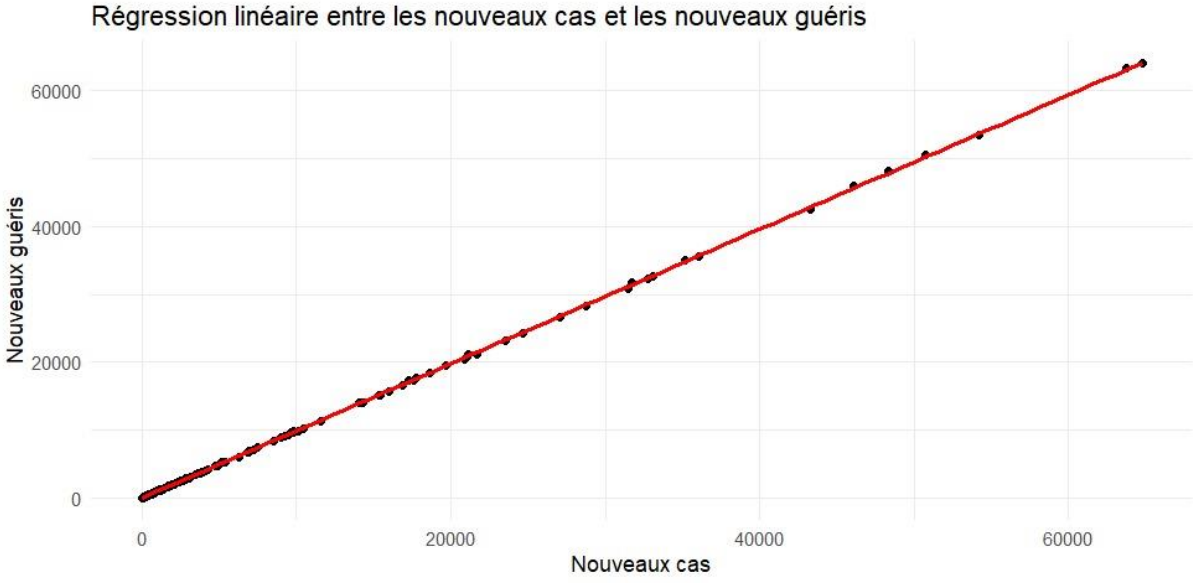
VI. La régression linéaire :

```
ggplot(data_maroc, aes(x = new_cases, y = new_deaths)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  theme_minimal() +
  labs(title = "Régression linéaire entre les nouveaux cas et les nouveaux décès", x = "Nouveaux cas", y = "Nouveaux décès")


ggplot(data_maroc, aes(x = new_cases, y = new_gueris)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  theme_minimal() +
  labs(title = "Régression linéaire entre les nouveaux cas et les nouveaux guéris", x = "Nouveaux cas", y = "Nouveaux guéris")
```

1. Régression linéaire entre nouveaux morts et nouveaux cas au Maroc



2. Régression linéaire entre nouveaux guéries et nouveaux cas au Maroc



Pour la régression linéaire entre les morts et les nouvelles cases : Il y'a un comportement dans le même sens mais pas de force corrélation, Malgré que la corrélation soit égale à 0.822.

Pour la régression linéaire entre les guéris et les nouvelle cases sont bien compatible.

VII. L'hypothèse null :

On suppose que : Les mesures de sécurité faites pendant le confinement était efficace.

D'abord on crée deux data frames qui contient une colonne qui concerne la valeur de reproduction pendant le confinement et une autre hors le confinement, le confinement était dans la période entre 2020/03/29 et 2020/07/05

```
data_maroc_confinement <- subset(data_maroc,date >= as.Date("2020-03-29") & date <= as.Date("2020-07-05"))
data_maroc_apres_confinement <- subset(data_maroc,date >= as.Date("2020-07-05") & date <= as.Date("2020-10-05"))

test <- t.test(data_maroc_confinement$reproduction_rate,data_maroc_apres_confinement$reproduction_rate)
print(test)
```

```
        Welch Two Sample t-test

data:  data_maroc_confinement$reproduction_rate and data_maroc_apres_confinement$reproduction_rate
t = 0.20353, df = 21.55, p-value = 0.8406
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1901683  0.2315016
sample estimates:
mean of x mean of y
 1.200667  1.180000
```

Les résultats obtenus ne montrent pas une grande différence entre la période de confinement et après le confinement (1.2 et 1.18)

Donc : On dit que les mesures de sécurité n'ont pas été si efficaces.

Les mesures de sécurité qui sont :

- Port du Masque Obligatoire
- Distanciation Sociale
- Confinement National
- Restrictions de Voyage

Et d'autres …

VIII. Conclusion :

Pour conclure ce projet, nous avons exploré et analysé la dynamique de la pandémie de COVID-19 au Maroc à travers divers aspects des données disponibles. Nos analyses ont révélé plusieurs points clés :

Asymétrie de la Distribution : Les données de nouveaux cas, de décès et de guérisons montrent une distribution asymétrique, avec des valeurs principalement concentrées entre 0 et 1000. Cela suggère une variabilité significative dans l'évolution de la pandémie.

Taux de Reproduction : Le taux de reproduction a atteint son pic à la fin de l'année 2021, où une personne infectée pouvait transmettre le virus à deux autres, confirmant ainsi une transmission efficace du virus à cette période.

Tests et Détection : Une augmentation significative du nombre de tests réalisés a permis de révéler un nombre plus important de cas, ce qui a contribué à une meilleure compréhension de l'extension de la pandémie et à une gestion plus efficace des ressources sanitaires.

Relations entre Variables :

Il existe une relation forte et significative entre le nombre total de décès et le nombre total de cas, indiquant que le nombre de décès est étroitement lié à l'ampleur de l'épidémie.

Une relation parfaite a été observée entre le nombre total de guéris et le nombre total de cas, ce qui montre que presque tous les cas diagnostiqués ont conduit à des guérisons.

Ces résultats mettent en évidence l'importance de la surveillance continue et de l'analyse des données épidémiologiques pour guider les politiques de santé publique et les interventions ciblées. Pour l'avenir, il est crucial de maintenir une collecte de données précise et de poursuivre l'analyse afin d'adapter rapidement les stratégies de lutte contre la pandémie et de minimiser ses impacts sur la société et l'économie.