

Autour du SARS-CoV-2 (Covid-19)

Doévi BIAOU

2021-12-09

Contents

1	Présentation	1
2	Exercice proprement dit	1
2.1	Méthodes	1
2.2	Import des données	1
2.3	Manipulation des données	1
2.4	Représentation graphique proprement dite	2
2.5	Graphique des incidences journalières	4
2.6	Nombre total de cas confirmés	5

1 Présentation

Le but est ici de reproduire des graphes semblables à ceux du South China Morning Post (SCMP), sur la page The Coronavirus Pandemic et qui montrent pour différents pays le nombre cumulé (c'est-à-dire le nombre total de cas depuis le début de l'épidémie) de personnes atteintes de la maladie à coronavirus 2019.

2 Exercice proprement dit

2.1 Méthodes

Les données utilisées sont celles fournies par le Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) et accessibles sous ce lien.

Les données ainsi recueillies ont été épurées puis analysées avec R version 4.1.2 (2021-11-01) s'exécutant sous linux-gnu. Les graphiques ont été générés avec le package {ggplot2}.

2.2 Import des données

```
# Enregistrement du lien de la base dans l'objet "link"
link <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data"

# Import du contenu du fichier csv dans une base de données :
df <- read.csv(file = link, header = TRUE, sep = ",")
df <- as.data.frame(df) # Transformation en base de données
```

2.3 Manipulation des données

Nous utiliserons essentiellement le paquet {dplyr} de {Tidyverse} à cet effet.

En premier un aperçu de la base.

```
head(df[1:8]) # Affiche les dix premières lignes des colonnes 1 à 8.
```

```
## Province.State      Country.Region      Lat      Long X1.22.20 X1.23.20
## 1                Afghanistan 33.93911 67.70995      0      0
## 2                Albania    41.15330 20.16830      0      0
## 3                Algeria    28.03390 1.65960      0      0
## 4                Andorra    42.50630 1.52180      0      0
## 5                Angola     -11.20270 17.87390      0      0
## 6      Antigua and Barbuda 17.06080 -61.79640      0      0
## X1.24.20 X1.25.20
## 1      0      0
## 2      0      0
## 3      0      0
## 4      0      0
## 5      0      0
## 6      0      0
```

la manipulation va consister à :

- filtrer la base initiale pour ne retenir que les lignes concernant les pays qui nous intéressent
- pivoter de sorte à avoir trois colonnes : Pays ou région, date et incidence journalière.
- calculer les incidences nationales journalières (certains pays comme la Chine ont des données provenant de différentes provinces)
- mettre en forme les dates sous format iso : les dates étant sous format "Xmm.jj.aa"

2.4 Représentation graphique proprement dite

Le graphique sera réalisé avec le paquet {ggplot2} déjà chargé au sein du {Tidyverse}.

Un aperçu du graphique de la Belgique.

```
dff_p %>%
  filter(Country.Region == "Belgium") %>%
  ggplot(aes(x = Dates, y = Incidence)) +
  geom_line() +
  theme_bw() + scale_y_continuous(labels = scales::label_number()) +
  labs(title = "Nombre de cas confirmés, Belgique")
```

L'on se rend compte que les incidences présentées dans la base de données sont des incidences cumulées.

Nous pouvons recalculer les incidences quotidiennes.

```
# Calcul des incidences journalières
dff_p <- dff_p %>%
  group_by(Country.Region) %>% # Pour les regrouper par pays
  mutate(Incidence_journaliere = Incidence - lag(Incidence)) # calcul des incidences journalières

# Aperçu de la base
tail(dff_p)
```

```
## # A tibble: 6 x 4
## # Groups:   Country.Region [6]
##   Dates      Country.Region Incidence Incidence_journaliere
##   <date>      <chr>          <int>          <int>
## 1 2021-12-08 Korea, South      496584          7100
## 2 2021-12-08 Netherlands    2850604         18102
## 3 2021-12-08 Portugal      1177706          5286
## 4 2021-12-08 Spain         5246766           0
```

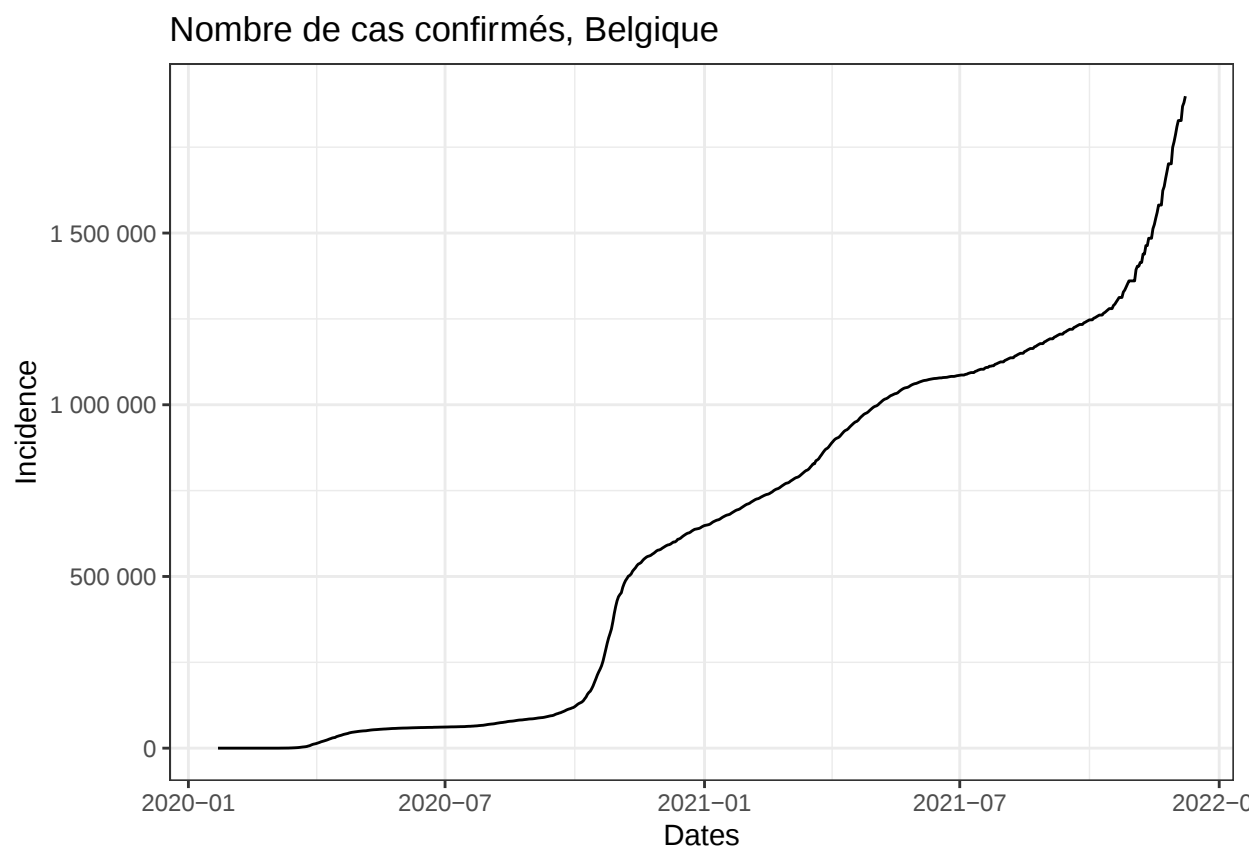


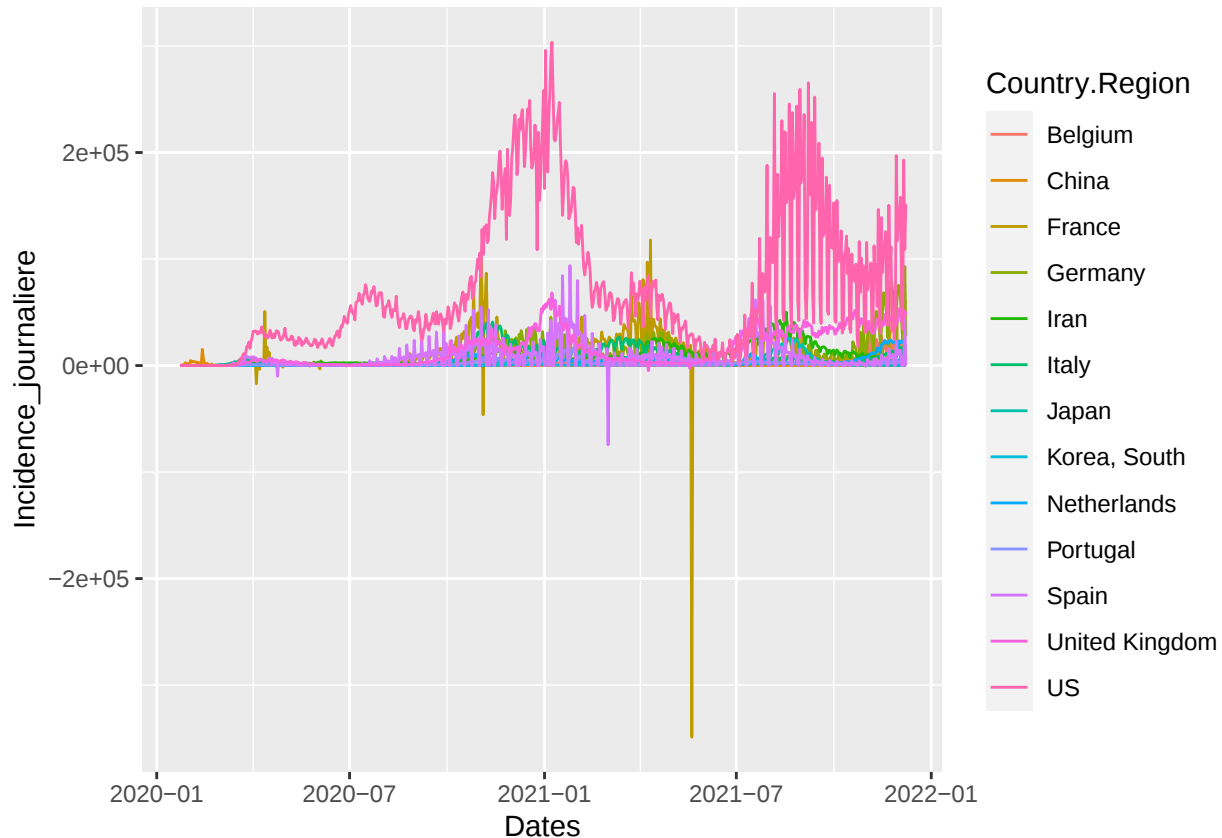
Figure 1: Nombre de cas confirmés, Belgique

```
## 5 2021-12-08 United Kingdom 10671538 51003
## 6 2021-12-08 US 49539734 151739
```

2.5 Graphique des incidences journalières

```
dff_p %>%
  ggplot(aes(x = Dates, y = Incidence_journaliere, color = Country.Region)) +
  geom_line()
```

```
## Warning: Removed 13 row(s) containing missing values (geom_path).
```



Quelques cas d'incidences négatives pouvant correspondre en réalité à des rectifications de statistiques nationales. Nous les définirons comme données manquantes.

```
# Aperçu, dates problématiques :
subset(dff_p, Incidence_journaliere < 0, c(Dates, Country.Region, Incidence_journaliere))
```

```
## # A tibble: 20 x 3
## # Groups:   Country.Region [7]
##   Dates      Country.Region Incidence_journaliere
##   <date>     <chr>                <int>
## 1 2020-04-04 France                -17076
## 2 2020-04-07 France                 -3489
## 3 2020-04-23 France                 -1709
## 4 2020-04-24 Spain                -10034
## 5 2020-04-29 France                 -1455
## 6 2020-05-02 Portugal                -161
## 7 2020-05-24 France                 -439
```

```
## 8 2020-05-25 Spain -372
## 9 2020-06-02 France -647
## 10 2020-06-03 China -1
## 11 2020-06-03 France -3226
## 12 2020-06-19 Italy -148
## 13 2020-06-28 France -407
## 14 2020-11-04 France -46121
## 15 2021-03-02 Spain -74347
## 16 2021-04-09 United Kingdom -4787
## 17 2021-05-18 United Kingdom -2362
## 18 2021-05-20 France -348840
## 19 2021-07-24 China -1
## 20 2021-10-31 Germany -554

# Les définir comme données manquantes
dff_p <- dff_p %>%
  mutate(Incidence_journaliere_2 = ifelse(Incidence_journaliere < 0, NA, Incidence_journaliere))
```

2.5.1 Le nouveau graphe, tous les pays

```
dff_p %>%
  ggplot(aes(x = Dates, y = Incidence_journaliere_2, color = Country.Region)) +
  geom_line() + scale_color_viridis_d() +
  scale_y_continuous(labels = scales::label_number()) +
  theme_bw() +
  labs(x = "", y = "Incidence journalière", color = "Pays",
       title = "Incidence journalière de covid 19")
```

2.5.2 Incidence journalière : Pays par pays, sur le même graphique

```
dff_p %>%
  ggplot(aes(x = Dates, y = Incidence_journaliere_2, color = Country.Region)) +
  geom_line() + scale_color_viridis_d() +
  scale_y_continuous(labels = scales::label_number()) +
  facet_wrap(~Country.Region, ncol = 3, scales = "free_y") +
  theme_bw() +
  labs(x = "", y = "Incidences journalières", color = "Pays",
       title = "Incidences journalières de covid 19") +
  theme(legend.position = "none") # Pour supprimer la légende
```

2.6 Nombre total de cas confirmés

Les incidences présentées par défaut étant des incidences cumulées, le nombre total de cas correspond donc à l'incidence cumulée à la date la plus récente, ou à l'incidence maximale renseignée (au cas où un pays serait en retard dans la notification de ses données).

```
max(dff_p$Dates)

## [1] "2021-12-08"

dff_p %>% filter(Dates == max(dff_p$Dates)) %>% # Date la plus récente uniquement
  ggplot(aes(x = fct_reorder(Country.Region, Incidence), y = Incidence)) +
  geom_col() +
  coord_flip() +
  theme_bw() +
```

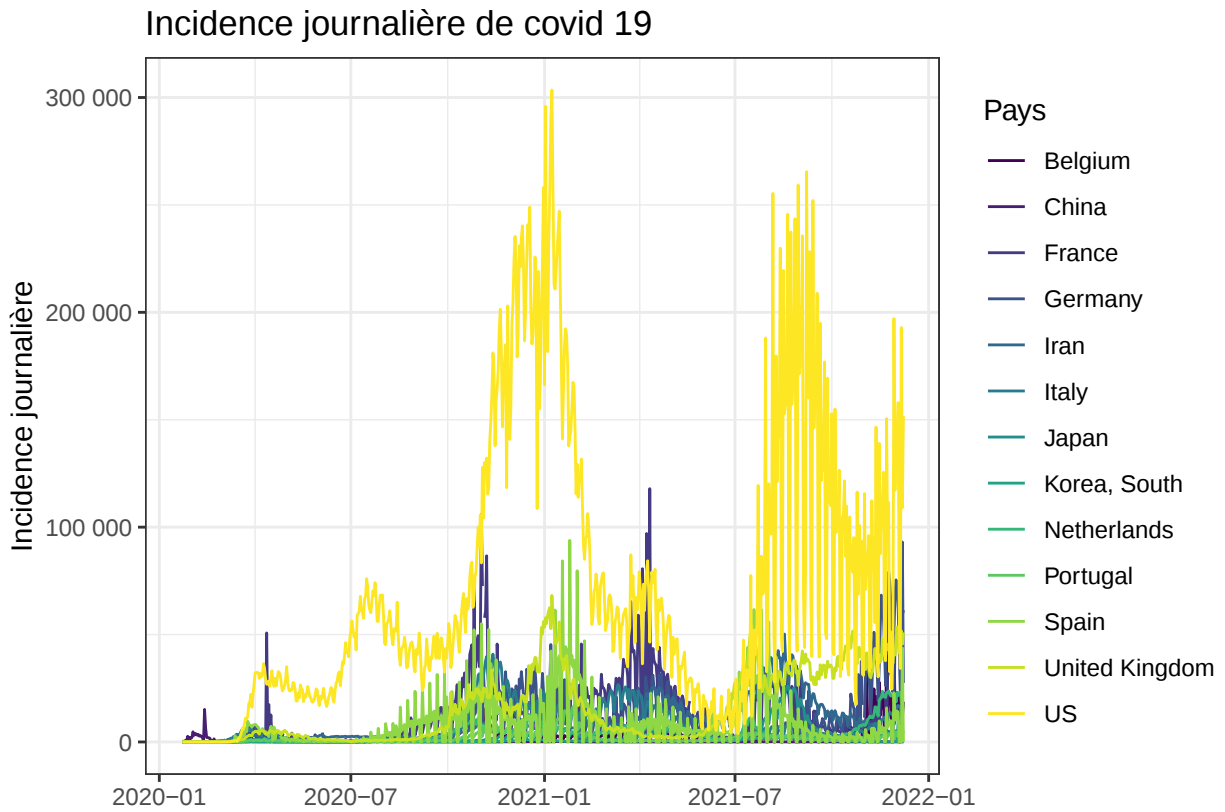


Figure 2: Incidence journalière de covid 19, tous les pays

```
labs(x = "", title = paste("Nombre de total de cas confirmés de covid 19 au", Sys.Date())) +
scale_y_continuous(labels = scales::number_format())
```

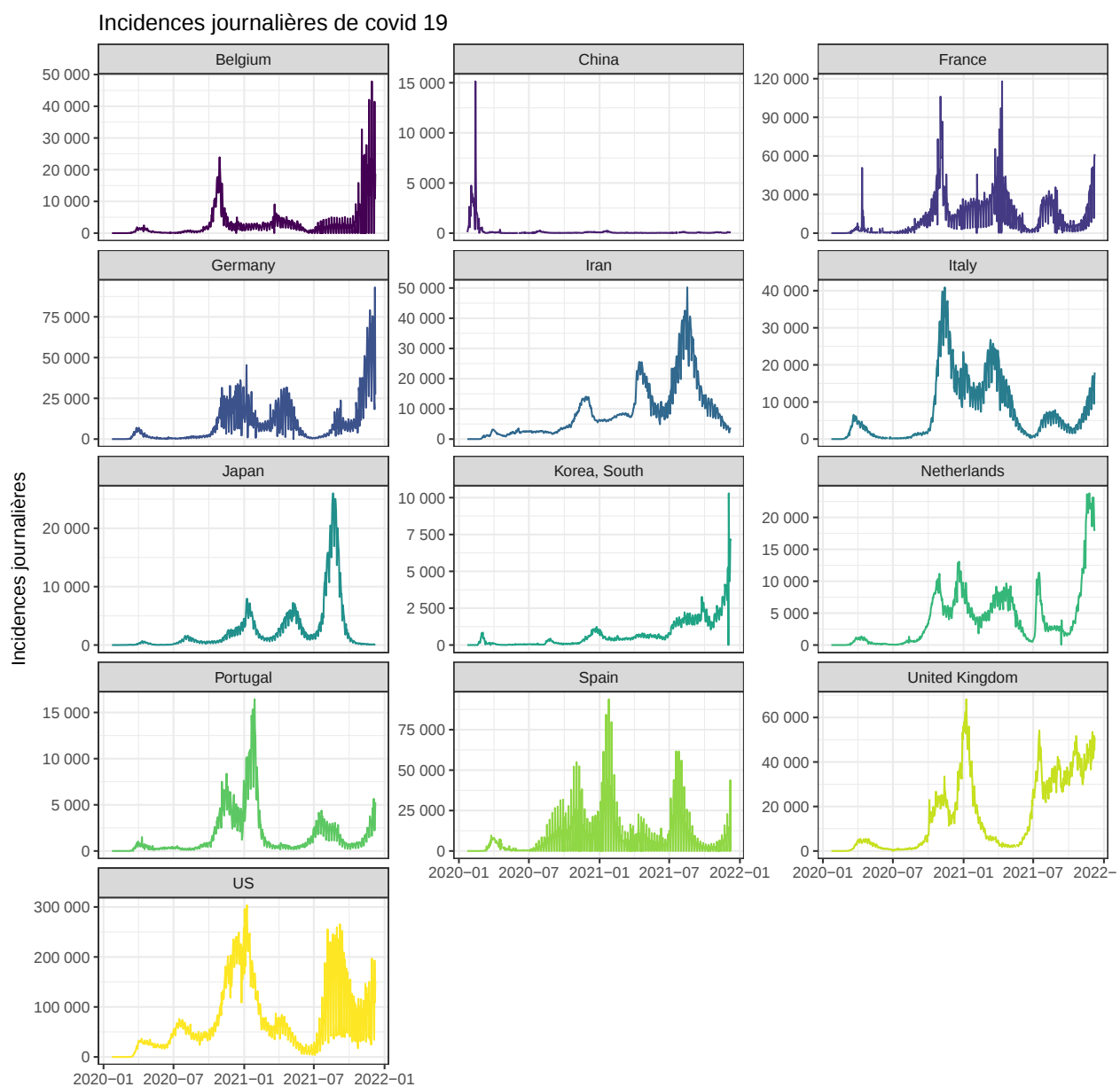


Figure 3: Incidence journalière, pays par pays

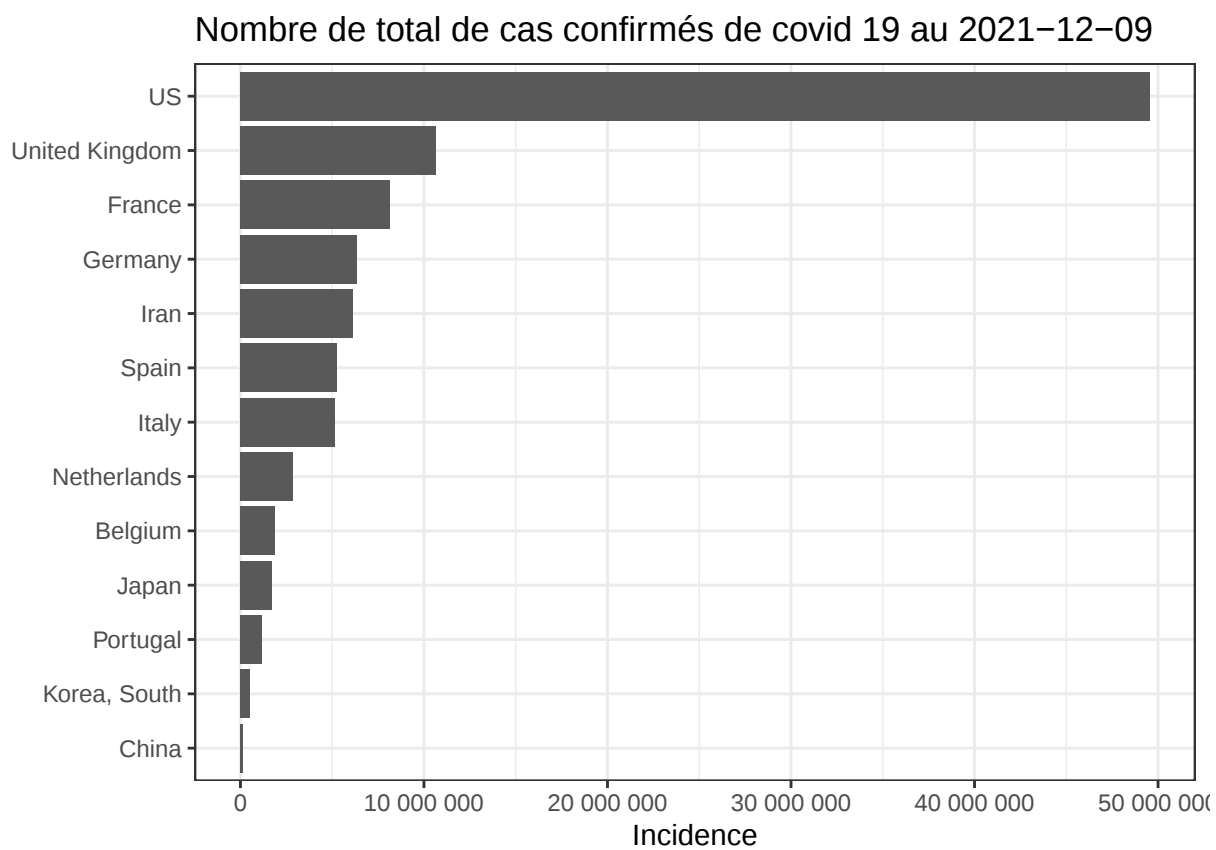


Figure 4: Nombre de cas confirmés de covid 19