covid_analysis

Martyna Stajniak

2025-01-14

COVID-19 Data Visualization

```
library( "ggpmisc")
```

```
## Warning: pakiet 'ggpmisc' został zbudowany w wersji R 4.3.2
```

```
## Ładowanie wymaganego pakietu: ggpp
```

```
## Warning: pakiet 'ggpp' został zbudowany w wersji R 4.3.2
```

```
## Ładowanie wymaganego pakietu: ggplot2
```

```
## Warning: pakiet 'ggplot2' został zbudowany w wersji R 4.3.2
```

```
## Registered S3 methods overwritten by 'ggpp':
##   method                  from
##   heightDetails.titleGrob ggplot2
##   widthDetails.titleGrob  ggplot2
```

```
##
## Dołączanie pakietu: 'ggpp'
```

```
## Następujący obiekt został zakryty z 'package:ggplot2':
##
##     annotate
```

```
## Registered S3 method overwritten by 'ggpmisc':
##   method                  from
##   as.character.polynomial polynom
```

```
library("ggplot2")
library("plotly")
```

```
## Warning: pakiet 'plotly' został zbudowany w wersji R 4.3.2
```

```
##
## Dołączanie pakietu: 'plotly'
```

```
## Następujący obiekt został zakryty z 'package:ggplot2':
##
##     last_plot
```

```
## Następujący obiekt został zakryty z 'package:stats':
##
##     filter
```

```
## Następujący obiekt został zakryty z 'package:graphics':
##
##     layout
```

```r
library(readr)
```

```
## Warning: pakiet 'readr' został zbudowany w wersji R 4.3.2
```

```r
# I am using total_cases.csv dataset. This dataset contains only two columns: date and cases.

total_cases <- read_csv("C:/Users/Martynaa/Desktop/portfolio/analizy_R/total_cases.csv")
```

```
## Rows: 64 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (1): cases
## date (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(total_cases)
```

```
## # A tibble: 6 x 2
##   date        cases
##   <date>      <dbl>
## 1 2020-11-21 656305
## 2 2020-11-22 607893
## 3 2020-11-23 547681
## 4 2020-11-24 463730
## 5 2020-11-25 603506
## 6 2020-11-26 589316
```

```r
# Structure:
str(total_cases)
```

```
## spc_tbl_ [64 x 2] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ date : Date[1:64], format: "2020-11-21" "2020-11-22" ...
##  $ cases: num [1:64] 656305 607893 547681 463730 603506 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   date = col_date(format = ""),
##   ..   cases = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

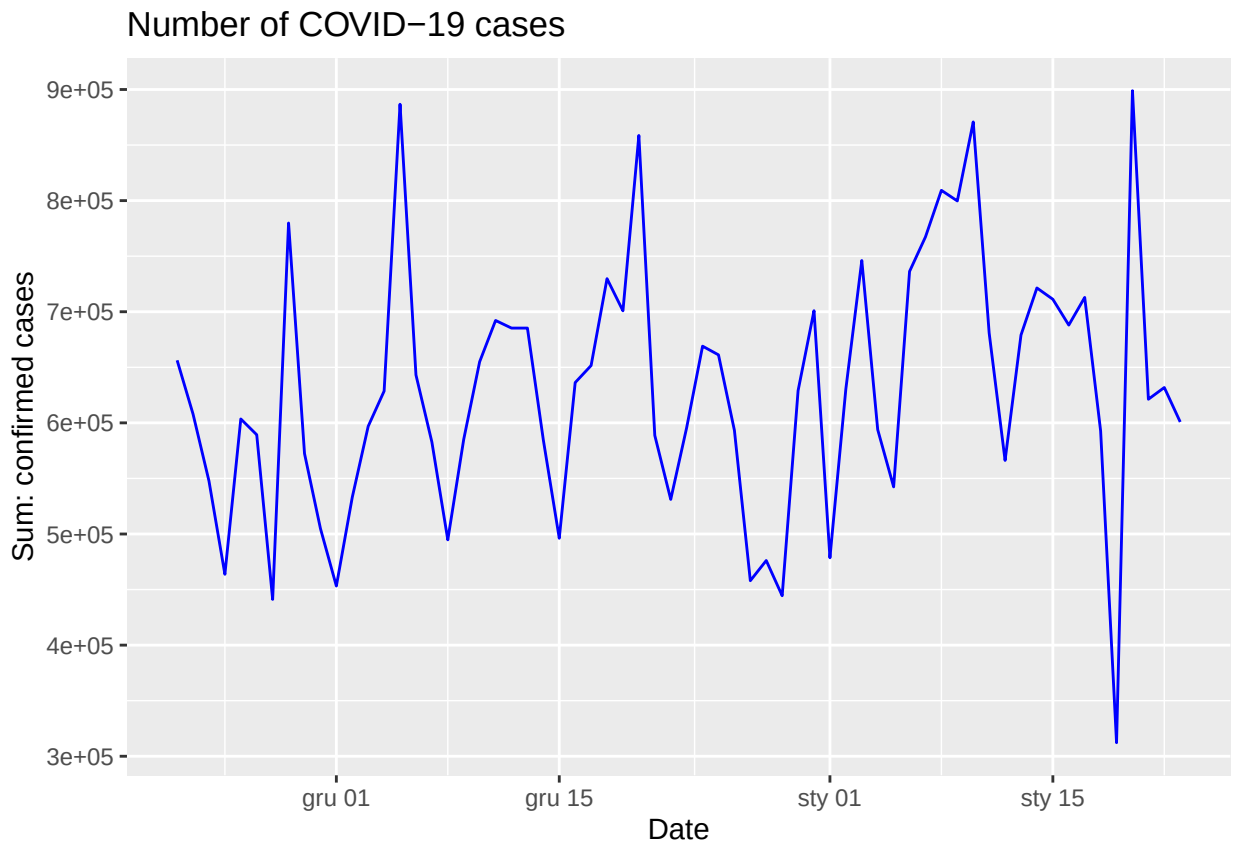
```r
# Summary:
summary(total_cases)
```

```
##       date                cases
##  Min.   :2020-11-21   Min.   :312202
##  1st Qu.:2020-12-06   1st Qu.:570826
##  Median :2020-12-22   Median :628794
##  Mean   :2020-12-22   Mean   :629472
##  3rd Qu.:2021-01-07   3rd Qu.:694328
##  Max.   :2021-01-23   Max.   :898893
```

```r
# Dimensions:
dim(total_cases)
```

```
## [1] 64  2
```

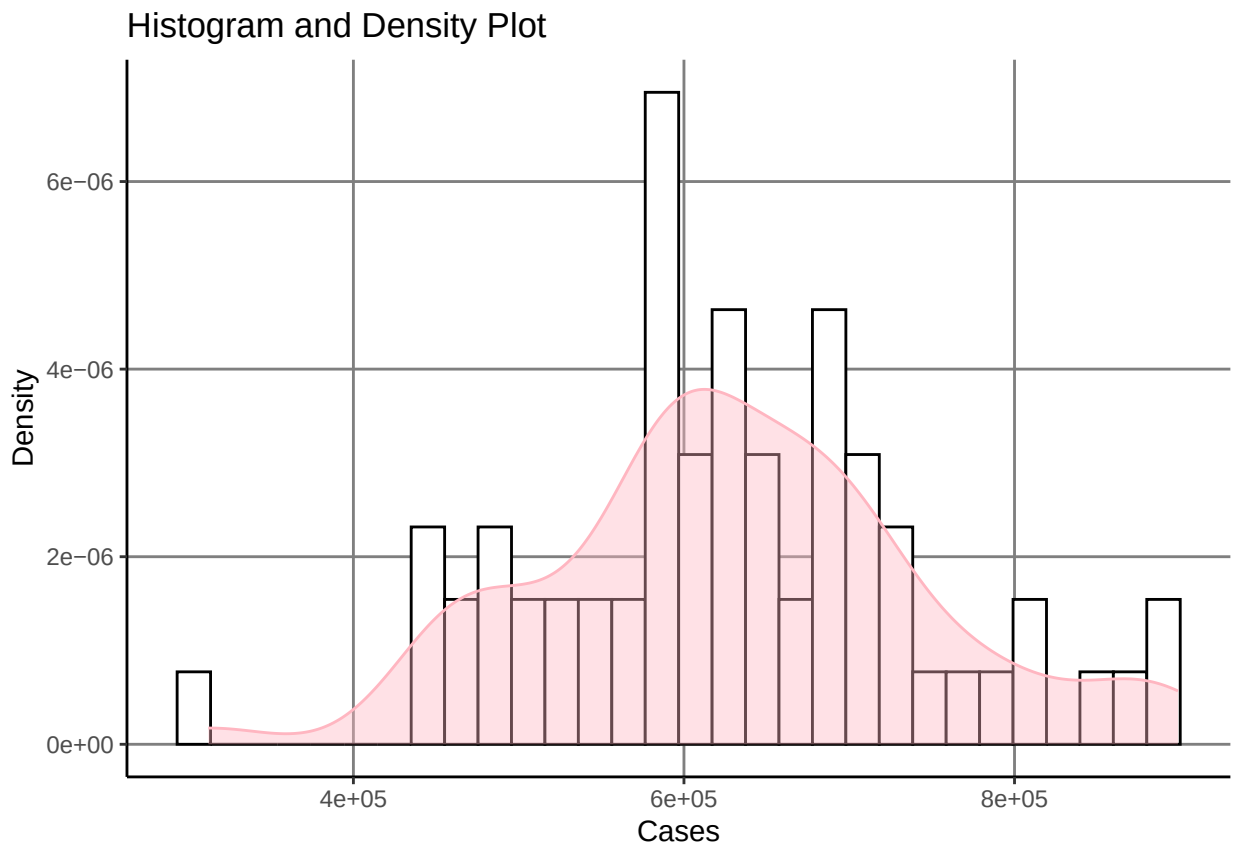
```
# chart
ggplot(total_cases, aes(x = date, y = cases)) +
  geom_line(color = "blue") +
  labs(title = "Number of COVID-19 cases",
       x = "Date",
       y = "Sum: confirmed cases")
```




```
# Histogram and density plot
ggplot(total_cases, aes(x = cases)) +
  geom_histogram(aes(y = ..density..), color = "black", fill = "white") +
  geom_density(color = "lightpink", fill = "lightpink", alpha = 0.4) +
  theme(plot.background = element_rect("white"),
        panel.background = element_rect("white"),
        axis.line = element_line("black"),
        panel.grid.major = element_line(colour = "grey50")) +
  labs(title = "Histogram and Density Plot", x = "Cases", y = "Density")
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# This dataset presents 7 countries and total cases each day
top7<-read.csv("C:/Users/Martynaa/Desktop/portfolio/analizy_R/top7_02_2020.csv")
head(top7)
```

```
##        country       date cum_cases
## 1      Germany 2020-02-18        16
## 2         Iran 2020-02-18         0
## 3        Italy 2020-02-18         3
## 4 Korea, South 2020-02-18        31
## 5        Spain 2020-02-18         2
## 6           US 2020-02-18        13
```

```
# Structure:
str(top7)
```

```
## 'data.frame':    2030 obs. of  3 variables:
##  $ country  : chr  "Germany" "Iran" "Italy" "Korea, South" ...
##  $ date     : chr  "2020-02-18" "2020-02-18" "2020-02-18" "2020-02-18" ...
##  $ cum_cases: int  16 0 3 31 2 13 13 13 13 13 ...
```

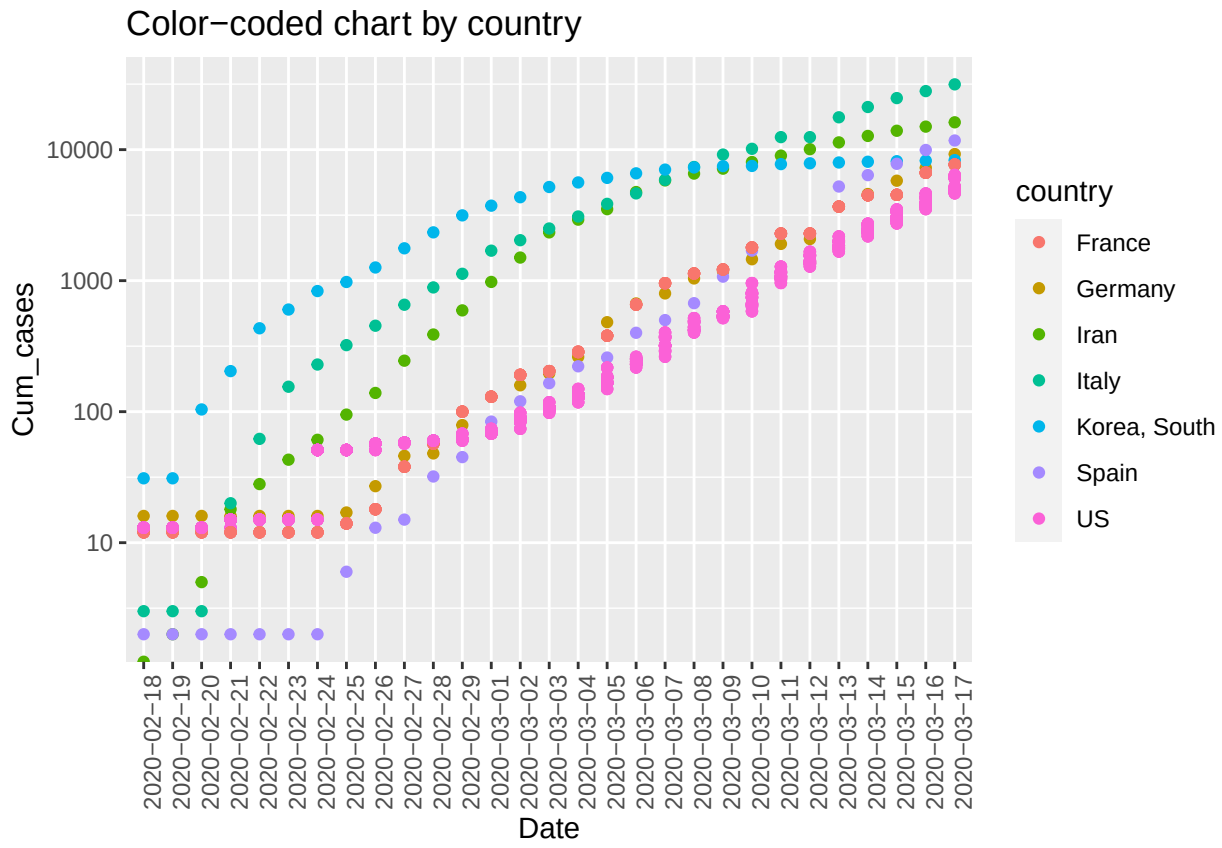
```
# Summarise:
summarise(top7)
```

```
## ramka danych z zerową liczbą kolumn oraz 1 wierszem
```

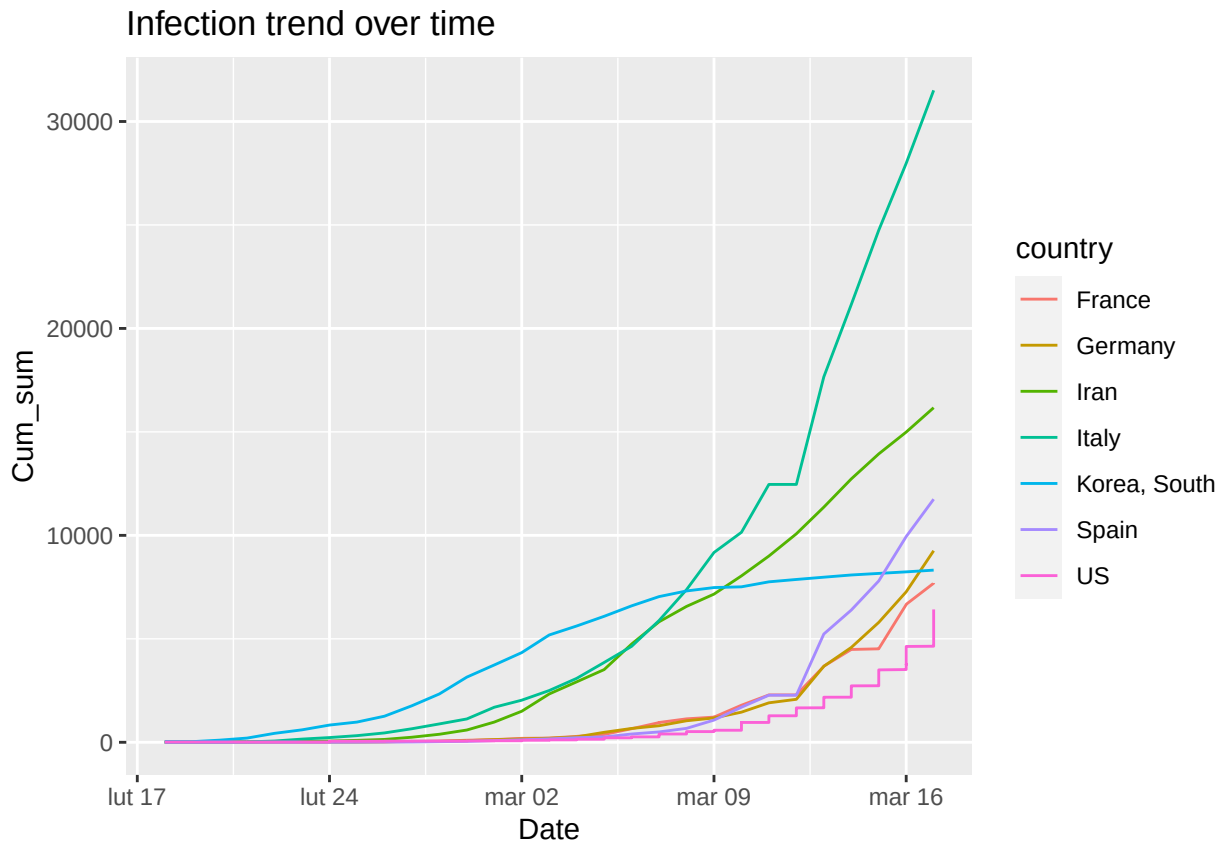
```
# Color-coded chart by country
ggplot(data = top7, aes(x = date, y = cum_cases, col = country)) +
  geom_point() +
  scale_y_log10() +
```

```
  labs(title = "Color-coded chart by country", x = "Date", y = "Cum_cases") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

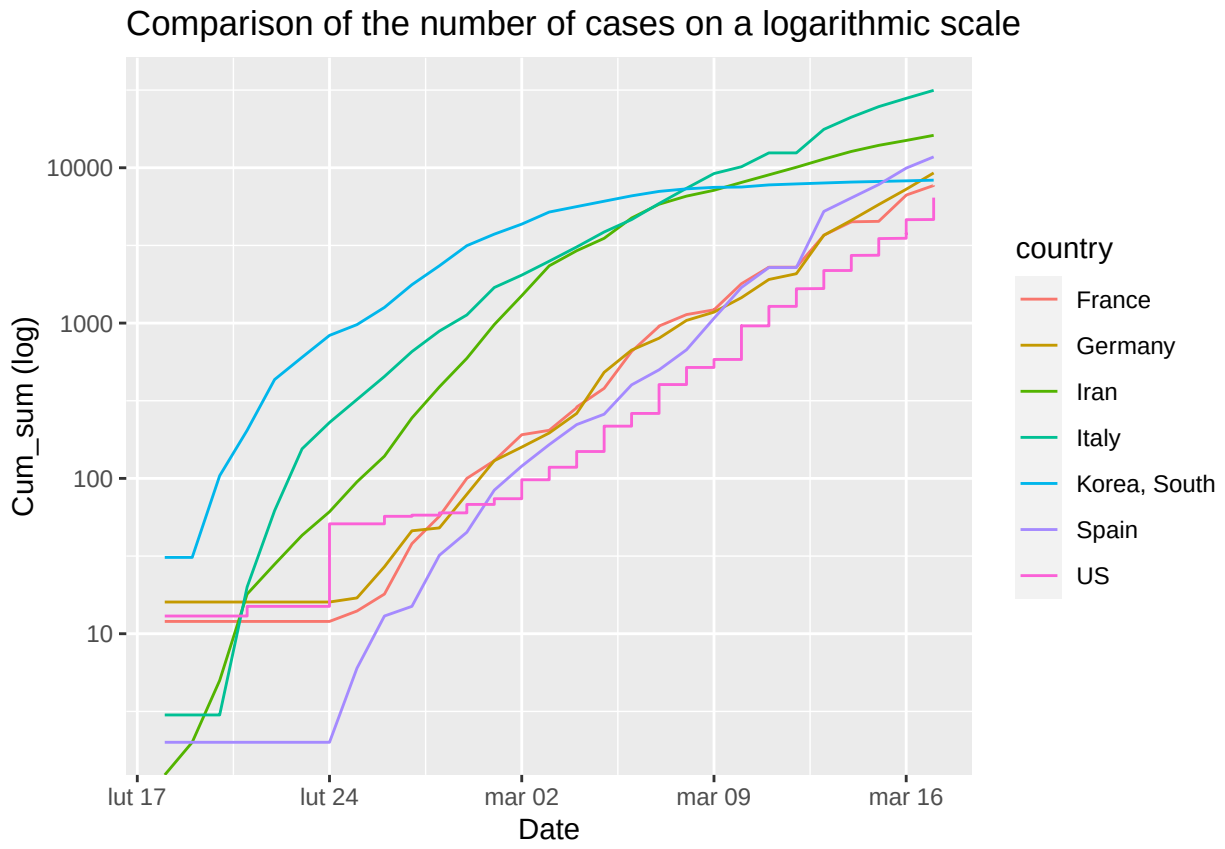



```
# Infection trend over time
ggplot(data = top7, aes(x = as.Date(date), y = cum_cases, color = country)) +
  geom_line() +
  labs(title = "Infection trend over time", x = "Date", y = "Cum_sum")
```

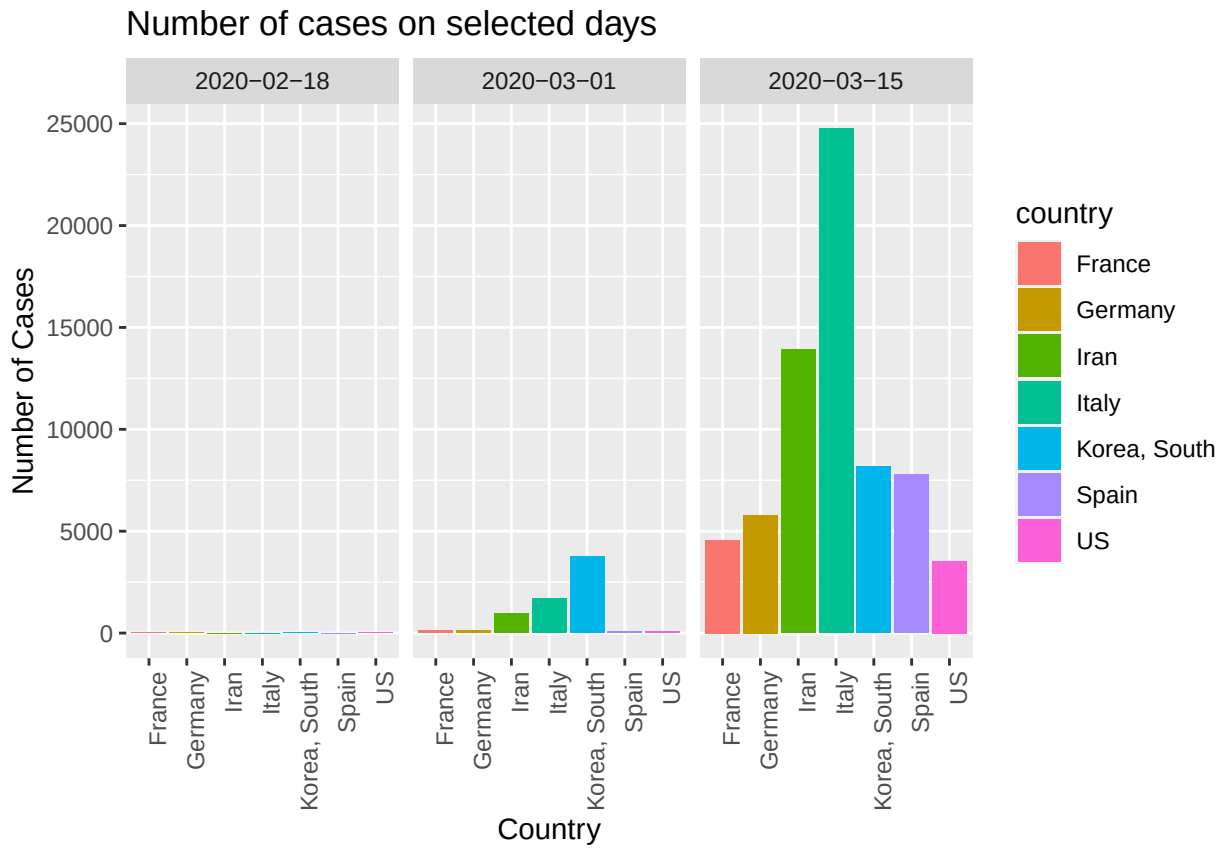



```
# Comparison of the number of cases on a logarithmic scale, now we can better capture growth proportion
ggplot(data = top7, aes(x = as.Date(date), y = cum_cases, color = country)) +
  geom_line() +
  scale_y_log10() +
  labs(title = "Comparison of the number of cases on a logarithmic scale", x = "Date", y = "Cum_sum (log
```

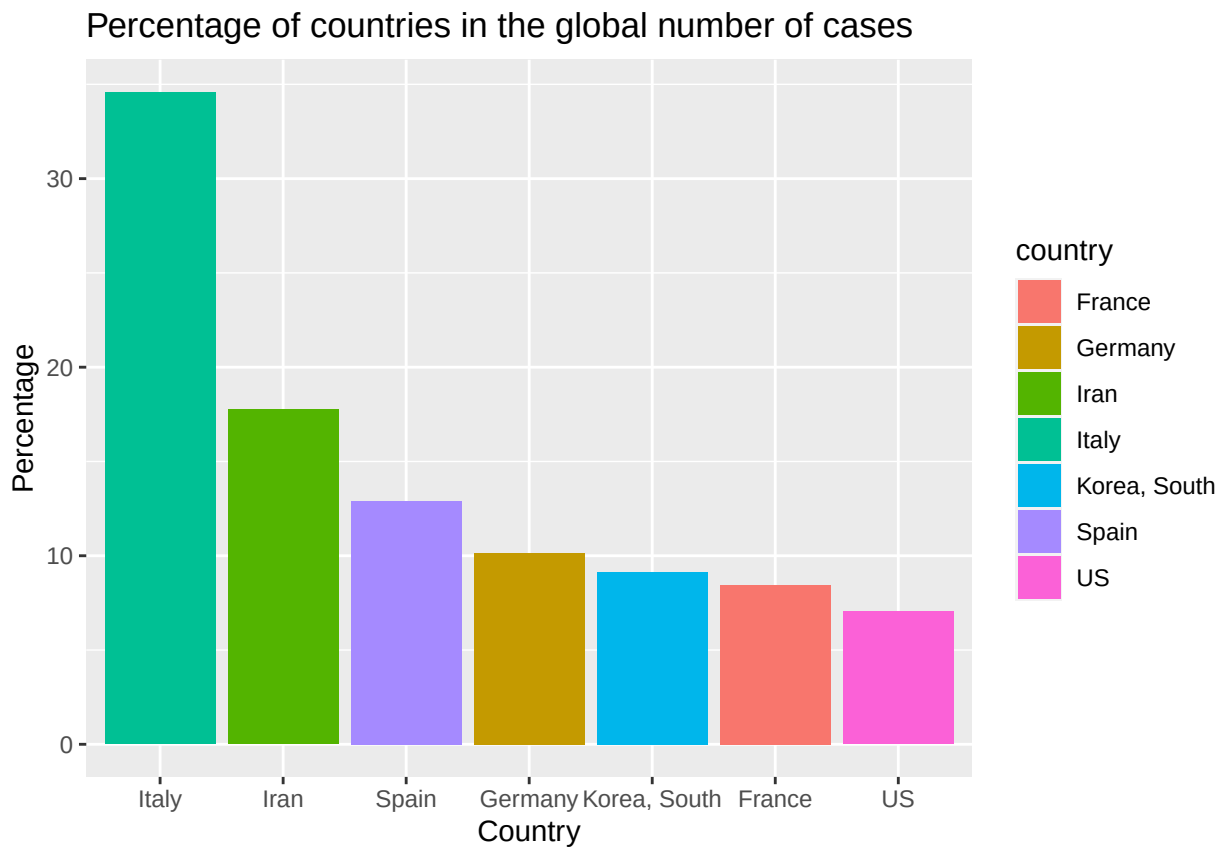
```
## Warning: Transformation introduced infinite values in continuous y-axis
```



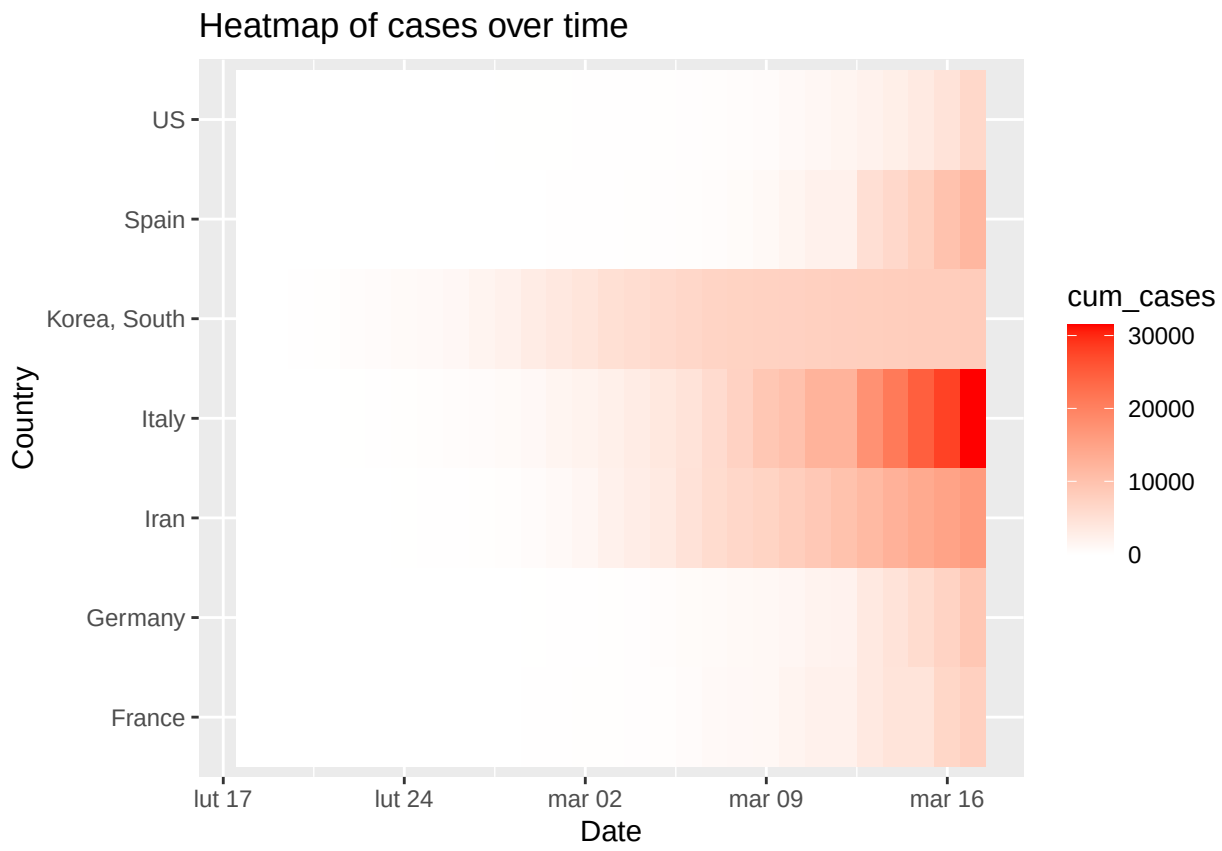

```
# Number of cases on selected days:  start, middle, end
selected_dates <- top7 %>% filter(date %in% c("2020-02-18", "2020-03-01", "2020-03-15"))
ggplot(data = selected_dates, aes(x = country, y = cum_cases, fill = country)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~date) +
  labs(title = "Number of cases on selected days", x = "Country", y = "Number of Cases") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```r
# Percentage of countries in the global number of cases
total_cases_by_country <- top7 %>%
  group_by(country) %>%
  summarise(total_cases = max(cum_cases, na.rm = TRUE))

total_cases_by_country <- total_cases_by_country %>%
  mutate(percentage = (total_cases / sum(total_cases)) * 100)

ggplot(data = total_cases_by_country, aes(x = reorder(country, -percentage), y = percentage, fill = cou
  geom_bar(stat = "identity") +
  labs(title = "Percentage of countries in the global number of cases", x = "Country", y = "Percentage"
```




```
#Heatmap

ggplot(data = top7, aes(x = as.Date(date), y = country, fill = cum_cases)) +
  geom_tile() +
  labs(title = "Heatmap of cases over time", x = "Date", y = "Country") +
  scale_fill_gradient(low = "white", high = "red")
```


Heatmap of cases over time
US
Spain
Korea, South
Italy
Iran
Germany
France
Country
lut 17
lut 24
mar 02
mar 09
mar 16
Date
cum_cases
30000
20000
10000
0