COVID19_EX1

pzuloaga

2024-03-01

Objectives

The objective of this data analysis is to study the evolution of COVID 19 pandemic in Peru. We will study the trends in the contagion and also in the fatality rate to try to understand how do they correlated to some specific events.

Read and Import Data

We start reading and importing the date from John Hopkins data set.

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi

file_names <- c("time_series_covid19_confirmed_global.csv", "time_series_covid19_deaths_global.csv")

urls <- str_c(url_in, file_names)
global_cases <- read_csv(urls[1])
global_deaths <- read_csv(urls[2])
```

Tidy and transform Data

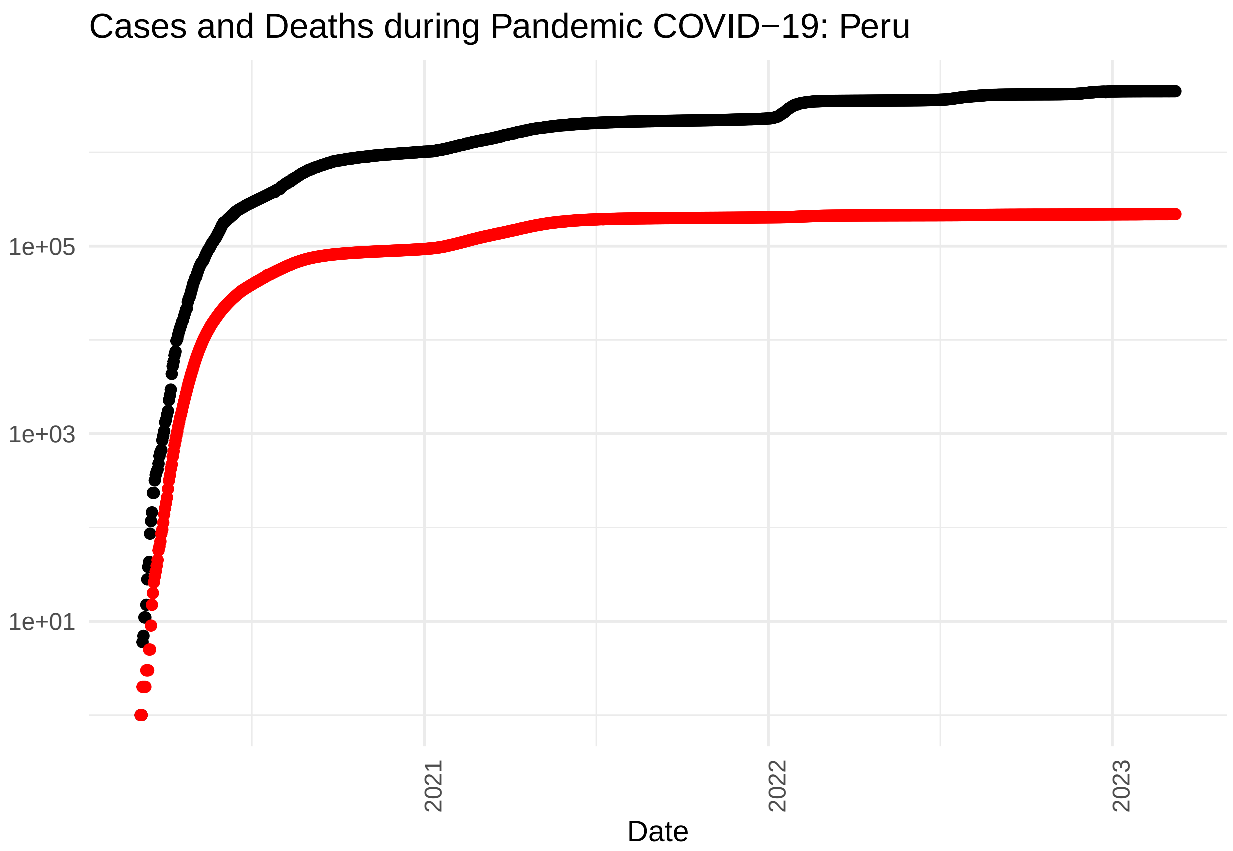
We also need to tidy and transform the data: data by date needed to be summarized and we also needed to joint the death and cases data sets, as well as group them by country.

```
global_cases <- global_cases %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region', Lat, Long), names_to = "date", values_to =
  select(-c(Lat,Long))

global_deaths <- global_deaths %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region', Lat, Long), names_to = "date", values_to =
  select(-c(Lat,Long))

global <- global_cases %>% full_join(global_deaths) %>%
  rename(Country_Region = 'Country/Region',
         Province_State = 'Province/State') %>%
  mutate(date = mdy(date))

global <- global %>% filter(cases > 0)

global <- global %>%
  unite("Combined_Key", c(Province_State, Country_Region),
        sep = ", ",
        na.rm = TRUE,
        remove = FALSE)

uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/U

uid <- read_csv(uid_lookup_url) %>% select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))

global <- global %>%
  left_join(uid, by = c("Province_State", "Country_Region")) %>% select(-c(UID, FIPS)) %>% select(Provin


global_by_country <- global %>% group_by(Province_State, Country_Region, date) %>% summarize(cases = sun
```

Visualization

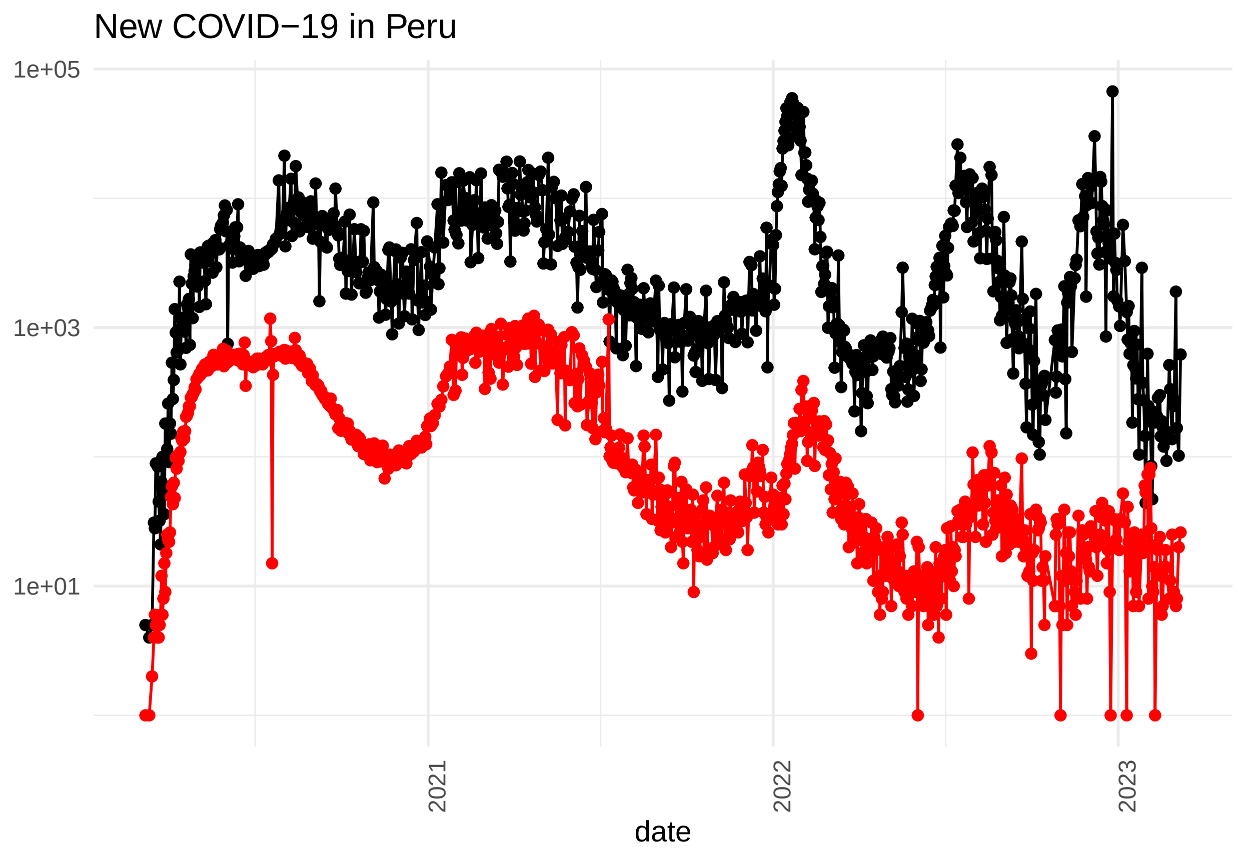
First, we will visualize the general evolution of cases and deaths in Peru:

```
country <- "Peru"
global_by_country %>% filter(Country_Region == country) %>% filter(cases > 0) %>%
  ggplot(aes(x=date, y=cases)) +
  geom_point(aes(color ="Total cases"),colour="black") +
  geom_point(aes(y=deaths, color="Total deaths"),colour="red") +
  xlab("Date") +
  ylab("Count") +
  scale_y_log10() +
  theme_minimal() +
  theme(legend.position="bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = str_c("Cases and Deaths during Pandemic COVID-19: ", country), y=NULL)
```



This plot is good to understand the evolution of the total cases, but it may be more revealing to analyze how the pandemic evolved in Peru in terms of the new cases, since this value may reveal the waves in the contagion process. This could be helpful to evaluate how the health system responded and what limitations did they have. So, will transform the data to get the new cases day by day.

```r
country <- "Peru"
global_by_country %>% filter(Country_Region == country) %>% filter(cases > 0) %>%
    mutate(new_cases = cases - lag(cases),
      new_deaths = deaths - lag(deaths)) %>%
      filter(new_cases > 0, new_deaths > 0) %>%
    ggplot(aes(x = date, y = new_cases)) +
      geom_line(aes(color = "new_cases"),colour="black")+
      geom_point(aes(color= "new_cases"),colour="black") +
      geom_line(aes(y = new_deaths, color = "new_deaths"),colour="red") +
      geom_point(aes(y = new_deaths, color = "new_deaths"),colour="red")+
      scale_y_log10()+
      theme_minimal() +
      theme(legend.position = "bottom",
      axis.text.x = element_text(angle = 90))+
      labs(title = "New COVID-19 in Peru", y = NULL)
```



From this new plot we can discovered there was five big waves in contagion. We can associate the first two to the variants alpha and delta, and we can clearly see how after 2022 the waves started to mimic as a seasonal flu, quite different from the two first cycles. We can also see that the third wave -at the begging of 2022- had the historical peak in cases, however, the deaths in this period were quite lower compared to the first two waves. This is related to the fact that by the beginning of 2022 Peru reached a vaccination of 80% of the total population.

To further analyze this effect, we will try to model the fatality rate.

Modeling

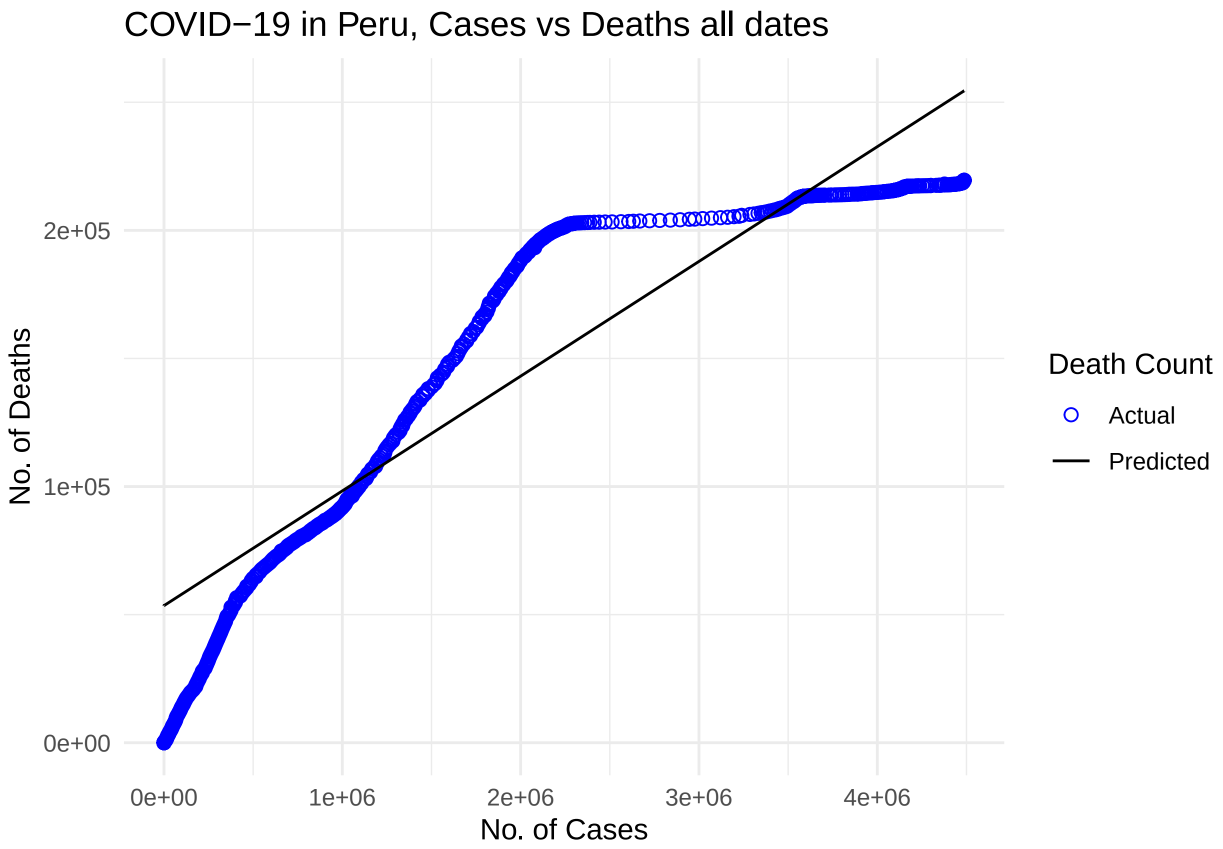
We will define the fatality rate as deaths/cases. If we try to adjust the whole data to a linear trend we will see that it is not possible:

```
mod = lm(deaths ~ cases, data = global_by_country %>% filter(Country_Region == "Peru"))
summary(mod)
```

```
##
## Call:
## lm(formula = deaths ~ cases, data = global_by_country %>% filter(Country_Region ==
##     "Peru"))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53511 -22576  -4944  28321  48501
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 5.351e+04  1.736e+03   30.83   <2e-16 ***
## cases       4.479e-02  6.427e-04   69.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 31030 on 1097 degrees of freedom
## Multiple R-squared:  0.8157, Adjusted R-squared:  0.8155 
## F-statistic:  4856 on 1 and 1097 DF,  p-value: < 2.2e-16
```

```r
Per_pred = global_by_country %>% filter(Country_Region == "Peru")%>%
  mutate(pred = predict(mod), year=year(date))
  Per_pred %>% ggplot() +
  geom_point(aes(x = cases, y = deaths, color = "Actual"), shape = 1, size = 2) +
  geom_line(aes(x = cases, y = pred, color = "Predicted"))+
  scale_color_manual(name = "Death Count", values = c("Actual" = "blue", "Predicted" = "black"))+
  xlab("No. of Cases")+
  ylab("No. of Deaths")+
  theme_minimal()+
  ggtitle("COVID-19 in Peru, Cases vs Deaths all dates")
```

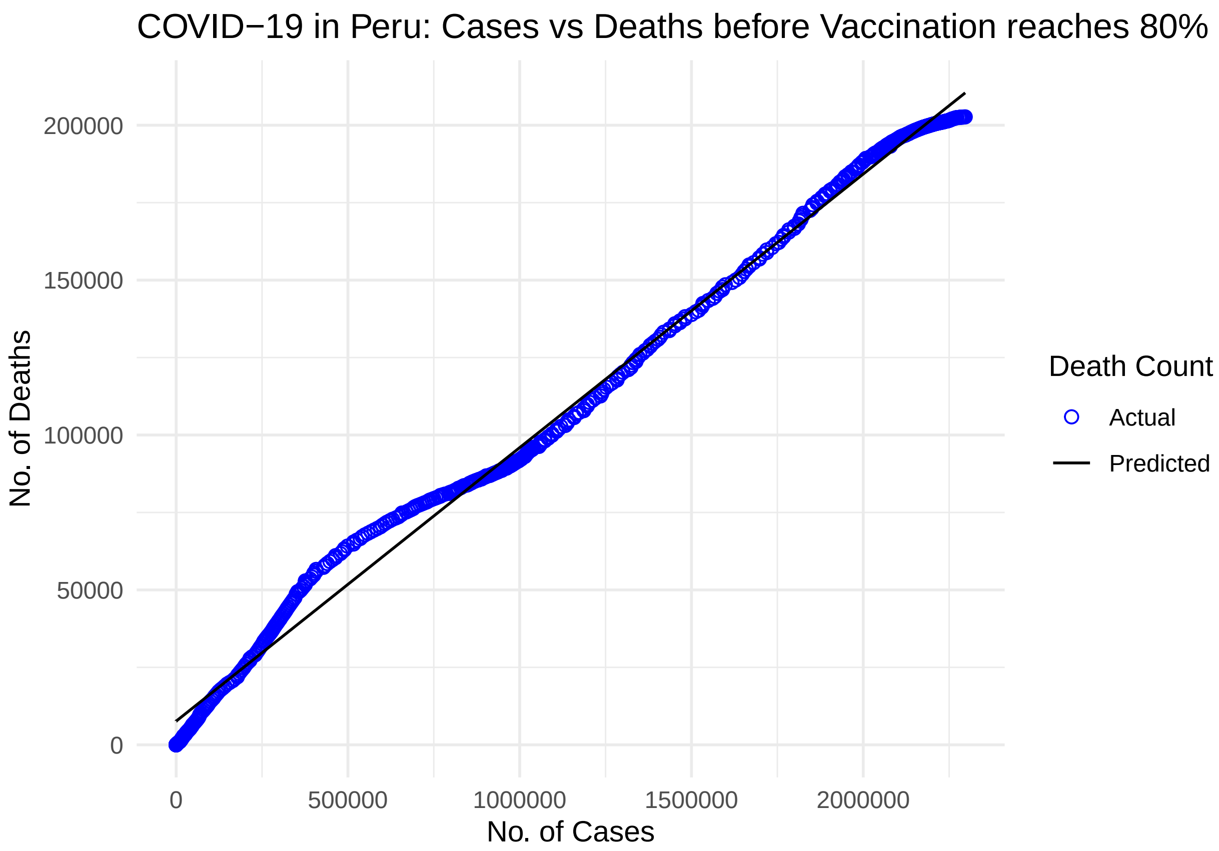


There are two main obvious trends, that clearly shows a change in the fatality rate, i.e. much lower deaths for the same amount of cases. Then, we can filter the data to the first period:

```r
mod = lm(deaths ~ cases, data = global_by_country %>% filter(Country_Region == "Peru", date < '2022-01-0
summary(mod)
```

```
##
## Call:
## lm(formula = deaths ~ cases, data = global_by_country %>% filter(Country_Region ==
##     "Peru", date < "2022-01-01"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7762.4 -3513.3  -554.5  2742.4 12997.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.634e+03  3.453e+02   22.11   <2e-16 ***
## cases       8.830e-02  2.349e-04  375.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4786 on 664 degrees of freedom
## Multiple R-squared:  0.9953, Adjusted R-squared:  0.9953
## F-statistic: 1.413e+05 on 1 and 664 DF,  p-value: < 2.2e-16
```

```r
Per_pred = global_by_country %>% filter(Country_Region == "Peru", date < '2022-01-01')%>%
  mutate(pred = predict(mod))
  Per_pred %>% ggplot() +
  geom_point(aes(x = cases, y = deaths, color = "Actual"), shape = 1, size = 2) +
  geom_line(aes(x = cases, y = pred, color = "Predicted"))+
  scale_color_manual(name = "Death Count", values = c("Actual" = "blue", "Predicted" = "black"))+
  xlab("No. of Cases")+
  ylab("No. of Deaths")+
  theme_minimal()+
  ggtitle("COVID-19 in Peru: Cases vs Deaths before Vaccination reaches 80%")
```



If we filter the data for the date before 01/01/2022 we can see that linear trend has a better matching. So, in this period we can actually model the number of death based on the number of cases, since this represent a quite similar contagion situation.

Conclusion and Possible Biases

From this analysis we can see the data set can be helpful to understand the development of the COVID-19. In particular, for Peru we can identify five waves of contagion, and two different behaviors of the disease, each one with a different fatality rate, that is correlated with the change in the spread due to a high rate of vaccination. There may be some bias the recording of the data, since many methodologies were used during its evolution. There was also a high involvement form the government to try to minimize the reported cases and fatalities specially during the peaks. Since the John Hopkins data set takes the official reports by country, they may have an inherent bias because of the country's transparency to record and report this data.

```
sessionInfo()
```

```
## R version 4.3.3 (2024-02-29 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
##
## locale:
```

```
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: America/Buenos_Aires
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.2     readr_2.1.5     tidyr_1.3.1     tibble_3.2.1
##  [9] ggplot2_3.5.0   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bit_4.0.5         gtable_0.3.4      highr_0.10        crayon_1.5.2
##  [5] compiler_4.3.3    tidyselect_1.2.0  parallel_4.3.3    scales_1.3.0
##  [9] yaml_2.3.8        fastmap_1.1.1     R6_2.5.1          labeling_0.4.3
## [13] generics_0.1.3    curl_5.2.1        knitr_1.45        munsell_0.5.0
## [17] pillar_1.9.0      tzdb_0.4.0        rlang_1.1.3       utf8_1.2.4
## [21] stringi_1.8.3     xfun_0.42         bit64_4.0.5       timechange_0.3.0
## [25] cli_3.6.2         withr_3.0.0       magrittr_2.0.3    digest_0.6.34
## [29] grid_4.3.3        vroom_1.6.5       rstudioapi_0.15.0 hms_1.1.3
## [33] lifecycle_1.0.4   vctrs_0.6.5       evaluate_0.23     glue_1.7.0
## [37] farver_2.1.1      fansi_1.0.6       colorspace_2.1-0  rmarkdown_2.25
## [41] tools_4.3.3       pkgconfig_2.0.3   htmltools_0.5.7
```