# Covid 19 data

Akikat

6/2/2021

## Covid 19 data analysis

This is a template for analysis of Covid 19 data gathered in US. Source of the data: COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University https://github.com/CSSEGISandData/COVID-19.

### Project setup

In order to reproduce the analysis following libraries should be used:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)
```

### Loading data

Next code chuck takes care of data downloading from website.

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

file_names <- c("time_series_covid19_confirmed_US.csv","time_series_covid19_confirmed_global.csv", "time

urls <- str_c(url_in, file_names)


global_cases <- read.csv(urls[2], head = TRUE, sep=",", check.names=FALSE)
global_deaths <- read.csv(urls[4], sep=",", check.names=FALSE)
us_cases <- read.csv(urls[1], check.names=FALSE)
us_deaths <- read.csv(urls[3], check.names=FALSE)
global_recovered <- read.csv(urls[5], check.names=FALSE)
```

### Prepare data

Prior to analysis sata needs to be manipulated. Columns unnecessary for analysis are removed. Pivot longer function allows to see statistical change over time. Population data is added to the dataset to make analysis more relevant.

```
global_cases <-  global_cases[c(-3, -4)]
global_cases <- global_cases %>%
  pivot_longer(cols = -c(1:2),
               names_to = "date",
               values_to = "cases")

global_deaths <-  global_deaths[c(-3, -4)]
global_deaths <- global_deaths %>%
  pivot_longer(cols = -c(1:2),
               names_to = "date",
               values_to = "deaths")

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

```
## Joining, by = c("Province/State", "Country/Region", "date")
```

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


us_cases <- us_cases[c(-(1:5))]
us_cases <- us_cases[c(-(4:5))]

us_cases <- us_cases %>%
  pivot_longer(cols = -c(1:4),
                          names_to = "date",
                        values_to = "cases")
us_cases <- us_cases %>%
  mutate(date = mdy(date))
```

```
us_deaths <- us_deaths[c(-(1:5))]
us_deaths <- us_deaths[c(-(4:5))]

us_deaths <- us_deaths %>%
  pivot_longer(cols = -c(1:5),
               names_to = "date",
               values_to = "deaths")%>%
  mutate(date = mdy(date))

US <- us_cases %>%
  full_join(us_deaths)  %>%
  rename(Country_Region = "Country_Region",
         Province_State = "Province_State")
```

```
## Joining, by = c("Admin2", "Province_State", "Country_Region", "Combined_Key", "date")
```

```
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, iso2, iso3, Admin2, code3))

global <- global %>%
  left_join(uid, by = c("Province_State", "Country_Region")) %>%
  select(-c(UID, FIPS)) %>%
  select(Province_State, Country_Region, date, cases, deaths, Population, 'Combined key')
```

### Visualizations

Initial view on number of cases and mortality rates are visualized with the help of ggplot2 library.

The following plots show mortality, that is the death rate compared to number of registered cases. Basically, this shows how dangerous it is to get Covid 19.

```
US_by_state <- US %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths), Population = sum(Population))  %>%
  mutate(deaths_per_mill = deaths*1000000/Population, mortality = deaths/cases) %>%
  select(Province_State, Country_Region, date, cases, deaths, deaths_per_mill, Population, mortality) %
  ungroup()
```

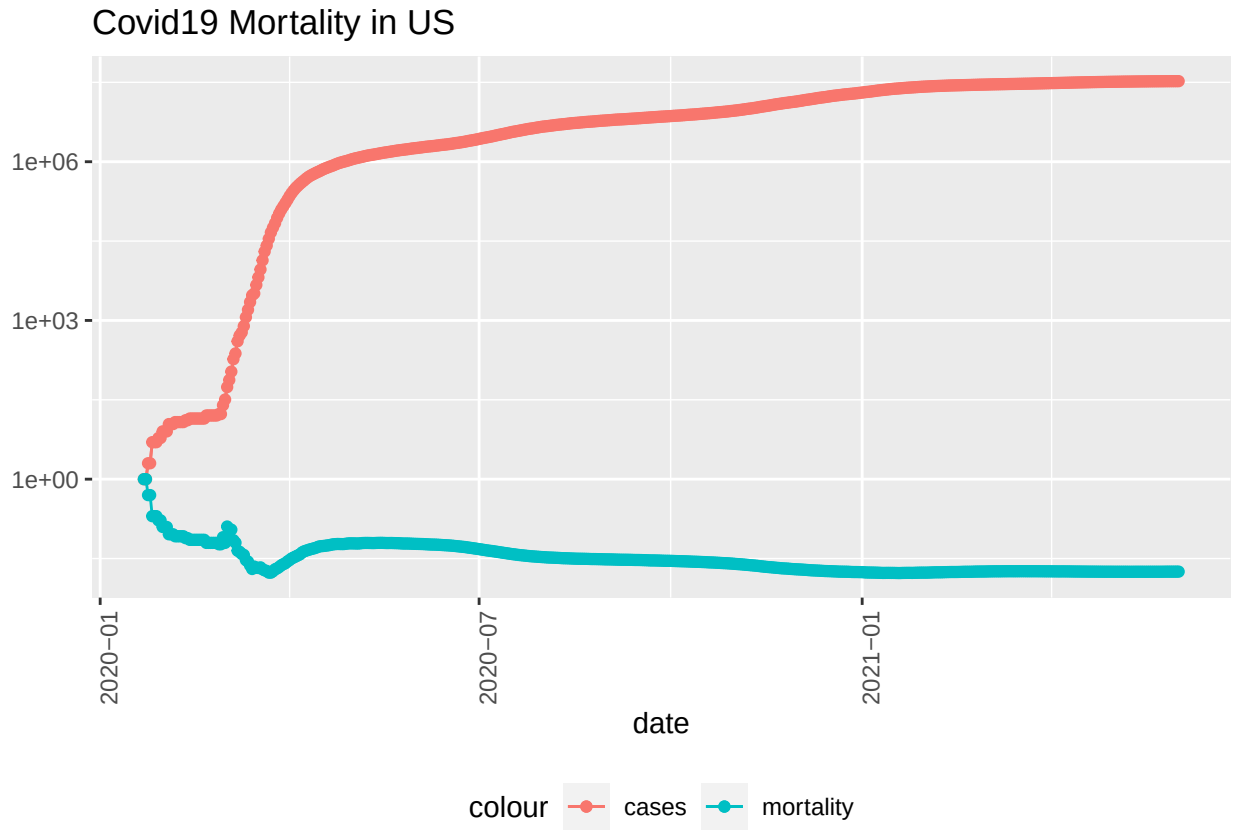
```
## `summarise()` has grouped output by 'Province_State', 'Country_Region'. You can override using the `
```

```
US_Totals <- US_by_state %>%
  group_by(Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths), Population = sum(Population))  %>%
```

```
  mutate(deaths_per_mill = deaths*1000000/Population,  mortality = deaths/cases) %>%
  select(Country_Region, date, cases, deaths, deaths_per_mill, Population, mortality) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Country_Region'. You can override using the `.groups` argument.
```

```
US_Totals %>%
  filter(cases>0) %>%
  ggplot(aes(x=date, y=cases)) +
  geom_line(aes(color="cases")) +
  geom_point(aes(color="cases")) +
  geom_line(aes(color="cases")) +
  geom_line(aes(y = mortality, color="mortality")) +
  geom_point(aes(y = mortality, color="mortality")) +
  scale_y_log10() +
  theme(legend.position = "bottom", axis.text.x = element_text(angle=90)) +
  labs(title = "Covid19 Mortality in US", y = NULL)
```

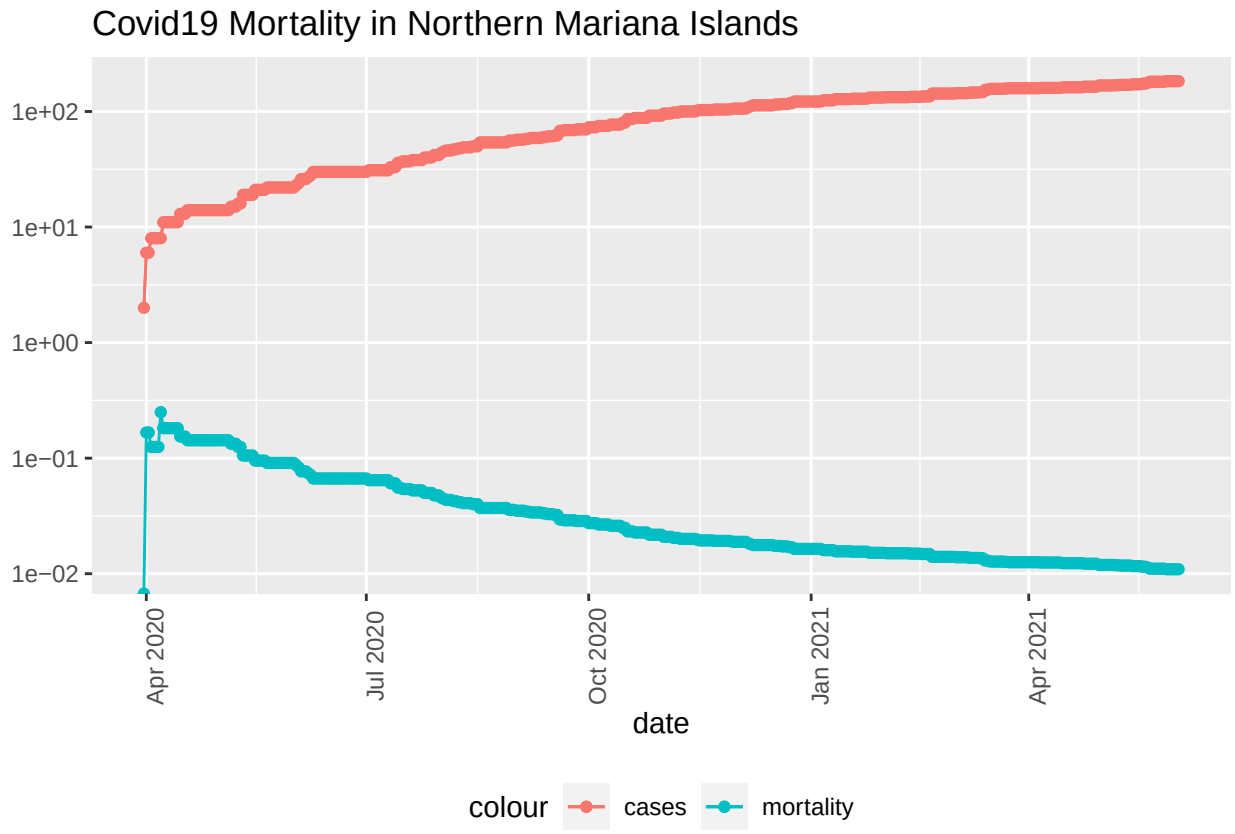



```
state <- "Northern Mariana Islands"
US_by_state %>%
  filter(Province_State == state) %>%
  filter(cases>0) %>%
  ggplot(aes(x=date, y=cases)) +
  geom_line(aes(color="cases")) +
```

```
  geom_point(aes(color="cases")) +
  geom_line(aes(color="cases")) +
  geom_line(aes(y = mortality, color="mortality")) +
  geom_point(aes(y = mortality, color="mortality")) +
  scale_y_log10() +
  theme(legend.position = "bottom", axis.text.x = element_text(angle=90)) +
  labs(title=str_c("Covid19 Mortality in ", state), y = NULL)
```

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

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




```
US_by_state <- US_by_state %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths))

US_Totals <- US_Totals %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths))
tail(US_Totals %>% select(new_cases, new_deaths, everything()))
```

```
## # A tibble: 6 x 9
##   new_cases new_deaths Country_Region date          cases deaths deaths_per_mill
##       <int>      <int> <chr>          <date>        <int>  <int>           <dbl>
```

```
## 1     21858        567 US             2021-05-28 33242999 593976          1784.
## 2     11999        343 US             2021-05-29 33254998 594319          1785.
## 3      6733        124 US             2021-05-30 33261731 594443          1786.
## 4      5776        142 US             2021-05-31 33267507 594585          1786.
## 5     22943        638 US             2021-06-01 33290450 595223          1788.
## 6     16913        610 US             2021-06-02 33307363 595833          1790.
## # ... with 2 more variables: Population <int>, mortality <dbl>
```

```
US_state_totals <- US_by_state %>%
  group_by(Province_State) %>%
  summarize(deaths = max(deaths), cases = max(cases), Population = max(Population),
            cases_per_thou = cases * 1000 / Population, deaths_per_thou = deaths * 1000/Population, mor
  filter(cases>0, Population >0)
```

```
## `summarise()` has grouped output by 'Province_State'. You can override using the `.groups` argument.
```

### Modeling

Simple predictive model was used to compare actual mortality data with expected mortality rate.

```
mod <- lm(mortality ~ cases, data = US_Totals)
summary(mod)
```

```
##
## Call:
## lm(formula = mortality ~ cases, data = US_Totals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04562 -0.02134 -0.00534  0.00613  0.93721
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.279e-02  4.443e-03   14.13  < 2e-16 ***
## cases       -1.755e-09  2.566e-10   -6.84 2.34e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06914 on 496 degrees of freedom
## Multiple R-squared:  0.0862, Adjusted R-squared:  0.08436
## F-statistic: 46.79 on 1 and 496 DF,  p-value: 2.335e-11
```

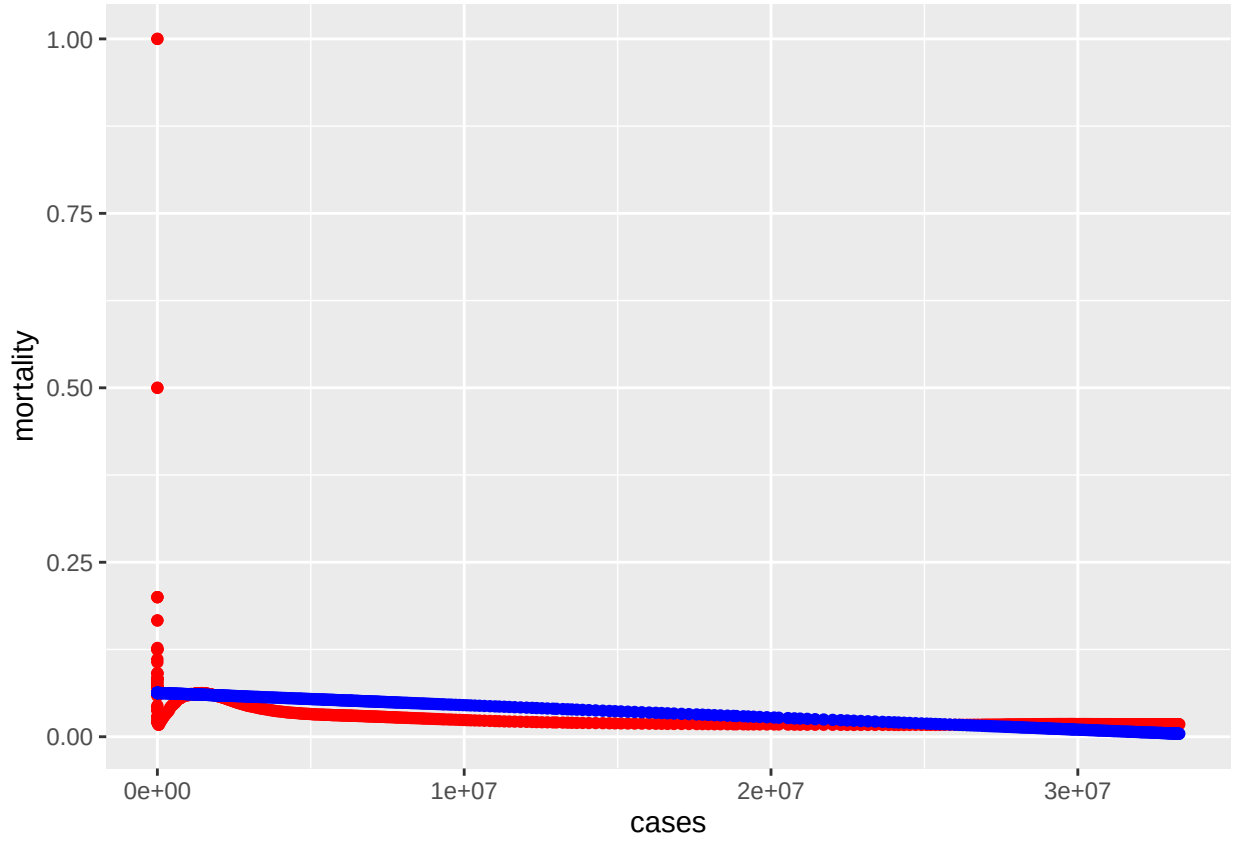
```
x_grid <- seq(1, 151)
new_df <- tibble(cases_per_thou = x_grid)
US_Totals %>% mutate(pred = predict(mod))
```

```
## # A tibble: 498 x 10
##    Country_Region date       cases deaths deaths_per_mill Population mortality
##    <chr>          <date>     <int>  <int>           <dbl>      <int>     <dbl>
##  1 US             2020-01-22     1      1         0.00300  332875137     1
##  2 US             2020-01-23     1      1         0.00300  332875137     1
```

```
##  3 US             2020-01-24     2      1         0.00300  332875137     0.5
##  4 US             2020-01-25     2      1         0.00300  332875137     0.5
##  5 US             2020-01-26     5      1         0.00300  332875137     0.2
##  6 US             2020-01-27     5      1         0.00300  332875137     0.2
##  7 US             2020-01-28     5      1         0.00300  332875137     0.2
##  8 US             2020-01-29     6      1         0.00300  332875137     0.167
##  9 US             2020-01-30     6      1         0.00300  332875137     0.167
## 10 US             2020-01-31     8      1         0.00300  332875137     0.125
## # ... with 488 more rows, and 3 more variables: new_cases <int>,
## #   new_deaths <int>, pred <dbl>
```

```
US_total_with_pred <- US_Totals %>% mutate(pred = predict(mod))


US_total_with_pred %>% ggplot() +
  geom_point(aes(x = cases, y = mortality), color = "red") +
  geom_point(aes(x = cases, y = pred), color = "blue")
```




**Results**

We see that after a rapid rise of infection there is a tendency to plateau. Also, despite the rise of number of cases, mortality rate is falling. Which can be explained by general rise in awareness, better medical procedures. We see that predictive model and actual mortality rates match quite ok as time goes by.

### Source of bias

There are still fluctations that can be explained by a lot of factors, including local policies, lockdown practices, traditional models of interactions within communitities of various states. All these parameters are hidden when we consider big picture. And they can affect the results a lot. Therefore, better approach would be to analyse smaller regions separately. This would allow to find better remedies and strategies to infection spread.

```
sessionInfo()
```

```
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19041)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
## system code page: 1251
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.7.10 forcats_0.5.1    stringr_1.4.0    dplyr_1.0.5
##  [5] purrr_0.3.4      readr_1.4.0      tidyr_1.1.3      tibble_3.1.1
##  [9] ggplot2_3.3.3    tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.0  xfun_0.22         haven_2.4.1       colorspace_2.0-0
##  [5] vctrs_0.3.7       generics_0.1.0    htmltools_0.5.1.1 yaml_2.2.1
##  [9] utf8_1.2.1        rlang_0.4.10      pillar_1.6.0      glue_1.4.2
## [13] withr_2.4.2       DBI_1.1.1         dbplyr_2.1.1      modelr_0.1.8
## [17] readxl_1.3.1      lifecycle_1.0.0   munsell_0.5.0     gtable_0.3.0
## [21] cellranger_1.1.0  rvest_1.0.0       evaluate_0.14     labeling_0.4.2
## [25] knitr_1.33        fansi_0.4.2       highr_0.9         broom_0.7.6
## [29] Rcpp_1.0.6        scales_1.1.1      backports_1.2.1   jsonlite_1.7.2
## [33] farver_2.1.0      fs_1.5.0          hms_1.0.0         digest_0.6.27
## [37] stringi_1.5.3     grid_4.0.5        cli_2.4.0         tools_4.0.5
## [41] magrittr_2.0.1    crayon_1.4.1      pkgconfig_2.0.3   ellipsis_0.3.1
## [45] xml2_1.3.2        reprex_2.0.0      assertthat_0.2.1  rmarkdown_2.7
## [49] httr_1.4.2        rstudioapi_0.13   R6_2.5.0          compiler_4.0.5
```