MSDS Hackathon 2020

Andrew J. Graves

7/10/2020

Load packages

```
library(tidyverse)
```

```
## Warning: replacing previous import 'vctrs::data_frame' by 'tibble::data_frame'
## when loading 'dplyr'
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.3
```

```
library(lubridate)
library(ggthemes)
```

Load and tidy data

```
covid_dat <- read_csv("data/owid-covid-data.csv") %>%
  select(location, date, total_cases_per_million) %>%
  mutate(date = mdy(date)) %>%
  filter(date <= as_date("2020-05-08") &
         date >= as_date("2020-01-20"))

# Find diffs between country labels
world_dat <- map_data("world")

world_grp <- world_dat %>%
  group_by(region) %>%
  tally()

diff_world <- setdiff(world_grp$region, covid_dat$location)
diff_covid <- setdiff(covid_dat$location, world_grp$region)

# Re-coded these manually by inspecting the diffs!
join_names <- covid_dat %>%
  mutate(location = recode(str_trim(location),
       "United States" = "USA",
```

```
        "United States Virgin Islands" = "Virgin Islands",
        "British Virgin Islands" = "Virgin Islands",
        "United Kingdom" = "UK",
        "Gibraltar" = "UK",
        "Democratic Republic of Congo" =
          "Democratic Republic of the Congo",
        "Congo" = "Republic of Congo",
        "Trinidad and Tobago" = "Trinidad",
        "Timor" = "Timor-Leste",
        "Saint Maarten (Dutch part)" = "Saint Maarten",
        "Saint Vincent and the Grenadines" = "Saint Vincent",
        "Saint Kitts and Nevis" = "Saint Kitts",
        "Faeroe Islands" = "Faroe Islands",
        "Bonaire Sint Eustatius and Saba" = "Bonaire",
        "Antigua and Barbuda" = "Antigua",
        "Cote d'Ivoire" = "Ivory Coast",
        "Hong Kong" = "China"
  ))

map_dat <- map_data("world") %>%
  rename(location = region) %>%
  inner_join(join_names, by = "location") %>%
  filter(date == as_date("2020-05-08")) %>%
  select(-order, -location, -subregion) %>%
  rename(y = total_cases_per_million) %>%
  mutate(binned_cases = case_when(
    y < 5 ~ 1, y < 10 ~ 2, y < 50 ~ 3, y < 100 ~ 4,
    y < 500 ~ 5, y < 1000 ~ 6, y < 2000 ~ 7, y < 5000 ~ 8,
    y >= 5000 ~ 9
  )
)
```

Set global plot parameters

```
titles <- c("Total confirmed COVID-19 cases per million people",
            paste("The number of confirmed cases is lower than the",
                  "number of total cases. The main reason for this",
                  "is limited testing.")
  )

covid_caption <- paste("Source: European CDC- Situation Update",
                       "Worldwide - Last updated 7th May, 11:15",
                       "(London time)",
                       "\nOurWorldInData.org/coronavirus * CC BY"
)

hack_theme <- theme(
        text = element_text(family = "serif"),
        plot.title = element_text(size = 16),
        plot.subtitle = element_text(size = 8),
        plot.caption = element_text(hjust = 0, size = 8),
        plot.title.position = "plot",
```

```
        plot.caption.position = "plot",
)
```

Set parameters for line-plot figure

```
loc_fig1 <- c("United States", "United Kingdom",
              "World", "South Korea", "China")

hex <- RColorBrewer::brewer.pal(n = length(loc_fig1), name = "Set1")
hex_colors <- hex[c(4, 5, 3, 1, 2)]

date_breaks <- paste0("2020-",
                      c("01-22", "03-01", "04-10", "05-08")) %>%
  as_date()
date_labels <- paste0(c("Jan 22", "Mar 1", "Apr 10", "May 8"), ", 2020")
```

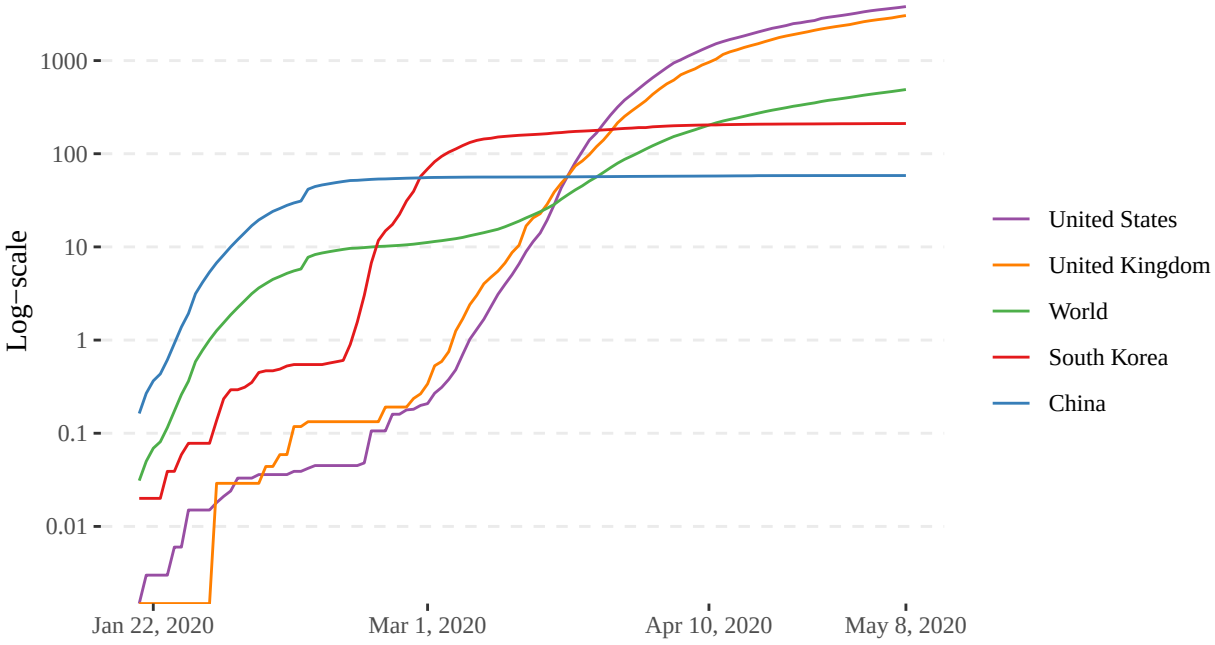
Create the line-plot figure

```
line_plot <- covid_dat %>%
  filter(location %in% loc_fig1) %>%
  mutate(location = fct_relevel(
                      as_factor(location), loc_fig1)
         ) %>%
  ggplot(aes(x = date, y = total_cases_per_million,
             color = location)) +
  geom_line() +
  scale_x_continuous(breaks = date_breaks,
                     labels = date_labels) +
  scale_y_log10(n.breaks = 6,
    labels = function(x) sprintf("%g", x)) +
  scale_color_manual(values = hex_colors) +
  labs(x = "", y = "Log-scale", color = "",
       title = titles[1], subtitle = titles[2],
       caption = covid_caption) +
  theme_classic() +
  theme(panel.grid.major.y = element_line(linetype = "dashed"),
        axis.line = element_blank()) +
  hack_theme

ggsave("output/line_plot.png")
line_plot
```

Total confirmed COVID−19 cases per million people

The number of confirmed cases is lower than the number of total cases. The main reason for this is limited testing.



Source: European CDC− Situation Update Worldwide − Last updated 7th May, 11:15 (London time)
OurWorldInData.org/coronavirus * CC BY

Set parameters for map figure

```
map_breaks <- 1:9

map_labels <- c(0, 5, 10, 50, 100, 500, 1000, 2000, 5000) %>%
  as.character()
map_labels[length(map_labels)] <- ">5000"
```

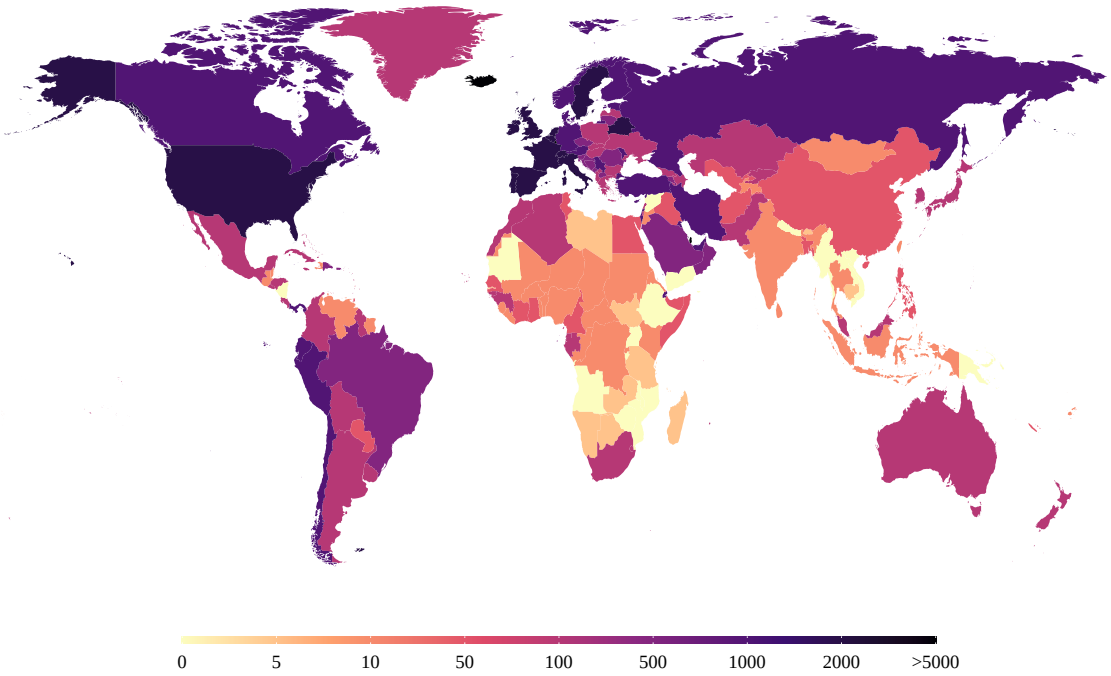
Create the map figure

```
map_fig <- map_dat %>%
  ggplot(aes(x = long, y = lat, group = group)) +
  geom_polygon(aes(fill = binned_cases)) +
  scale_fill_viridis_c(option = "magma", direction = -1,
                       breaks = map_breaks, labels = map_labels) +
  labs(fill = "",
       title = paste0(titles[1], ", May 8, 2020"),
                      subtitle = titles[2],
       caption = covid_caption) +
  theme_map() +
  theme(legend.position = "bottom",
        legend.justification = "center",
        legend.key.height = unit(0.1, "cm"),
        legend.key.width = unit(2, "cm")) +
  hack_theme

ggsave("output/world_map.png")
map_fig
```

Total confirmed COVID−19 cases per million people, May 8, 2020

The number of confirmed cases is lower than the number of total cases. The main reason for this is limited testing.



Source: European CDC− Situation Update Worldwide − Last updated 7th May, 11:15 (London time)
OurWorldInData.org/coronavirus * CC BY