Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

Wizualizacja Danych
semestr 21Z

Praca domowa nr. 2

Hubert Kozubek

Warszawa, 2021

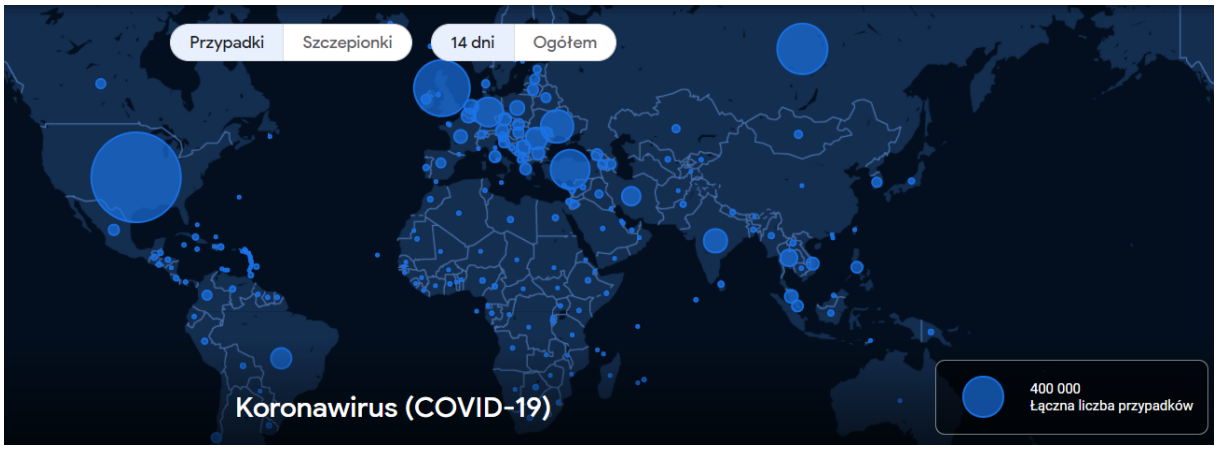
Spis treści

1. Wprowadzenie

Celem niniejszej pracy domowej było wykonanie wizualizacji danych na podstawie wykresu. Wykres ten musiał znajdować się w prasie, telewizji lub internecie i być opublikowany nie dawniej niż 2 tygodnie temu. Wizualizacja stworzona w ramach tej pracy domowej miała za zadanie poprawiać pewne błędy oraz niedociągnięcia wybranego wykresu.

2. Wykres oraz dane

Wizualizacja wybrana do poprawienia pochodzi z Google Wiadomości.



Rys. 1. Nowe przypadki zakażeń COVID-19, grafika z Google Wiadomości

Powodów wybrania tego wykresu jest kilka.

1. Nie jest możliwe wyświetlenie całej mapy świata jednocześnie.
2. Kontrast pomiedzy krajami, oceanami oraz danymi jest nieweilki, przez co elementy te w pewnym stopniu zelwają się ze sobą.
3. Kółka odwzorowujące nowe przypadki nachodzą na siebie, co zmniejsza ich czytelność.

Zaglądając do źródeł danych, na podstawie których powstał wykres, natrafiamy na stronę ourworldindata.org. Rownież z tej strony zostały pobrane dane wykorzystane dalej do narysowania poprawionego wykresu.

3. Tworzenie danych na podstawie wykresu

Do stworzenia poprawionego wykres został napisany następujący kod.

```
library(ggplot2)
library(dplyr)
library(readxl)

# Path to xlsx file
xl <- "D:\\Hubert\\Dokumenty\\Politechnika Warszawska\\IIAD\\Semestr III\\WD\\
    ↪ Laby\\HW2\\owid-covid-data.xlsx"


# Choosing data only younger than 2 weeks ago, and adjusting location to mach region data
    ↪  from world map
read_excel(xl) %>%
  select(location, date, new_cases) %>%
  filter(date>="2021-10-21" & date <="2021-11-03") %>%
  mutate(location = case_when(location == "United States" ~ "USA",
                              location == "United Kingdom" ~ "UK",
                              location == "Samoa" ~ "American Samoa",
                              location == "Czechia" ~"Czech Republic",
                              location == "Micronesia (country)" ~ "Micronesia",
                              location == "Georgia" ~ "South Georgia",
                              location == "Sint Maarten (Dutch part)" ~ "Sint Maarten",
                              location == "British Virgin Islands" ~ "Virgin Islands",
                              TRUE ~ location)) %>%
  group_by(location) %>%
  summarise(sum.from.2weeks = sum(new_cases)) -> d1

# Adjusting region data to match d1 location
map_data("world") %>%
  mutate(region = case_when(region == "Antigua" ~ "Antigua and Barbuda",
                            region == "Barbuda" ~ "Antigua and Barbuda",
                            region == "Democratic Republic of the Congo" ~ "Congo",
                            region == "Republic of the Congo" ~ "Congo",
                            region == "Nevis" ~ "Saint Kitts and Nevis",
                            region == "Saint Kitts" ~ "Saint Kitts and Nevis",
                            region == "Bonaire" ~ "Bonaire Sint Eustatius and Saba",
                            region == "Sint Eustatius" ~ "Bonaire Sint Eustatius and Saba
                                ↪ ",
                            region == "Saba" ~ "Bonaire Sint Eustatius and Saba",
                            region == "Trinidad" ~ "Trinidad and Tobago",
                            region == "Tobago" ~ "Trinidad and Tobago",
                            region == "Saint Vincent" ~ "Saint Vincent and the Grenadines
                                ↪ ",
                            region == "Grenadines" ~ "Saint Vincent and the Grenadines",
                            TRUE ~ region)) -> world_map

# Joining d1 and World by country (region == location)
merge(world_map, d1, by.x="region", by.y="location",all.x = TRUE) %>%
  arrange(order) %>%
```

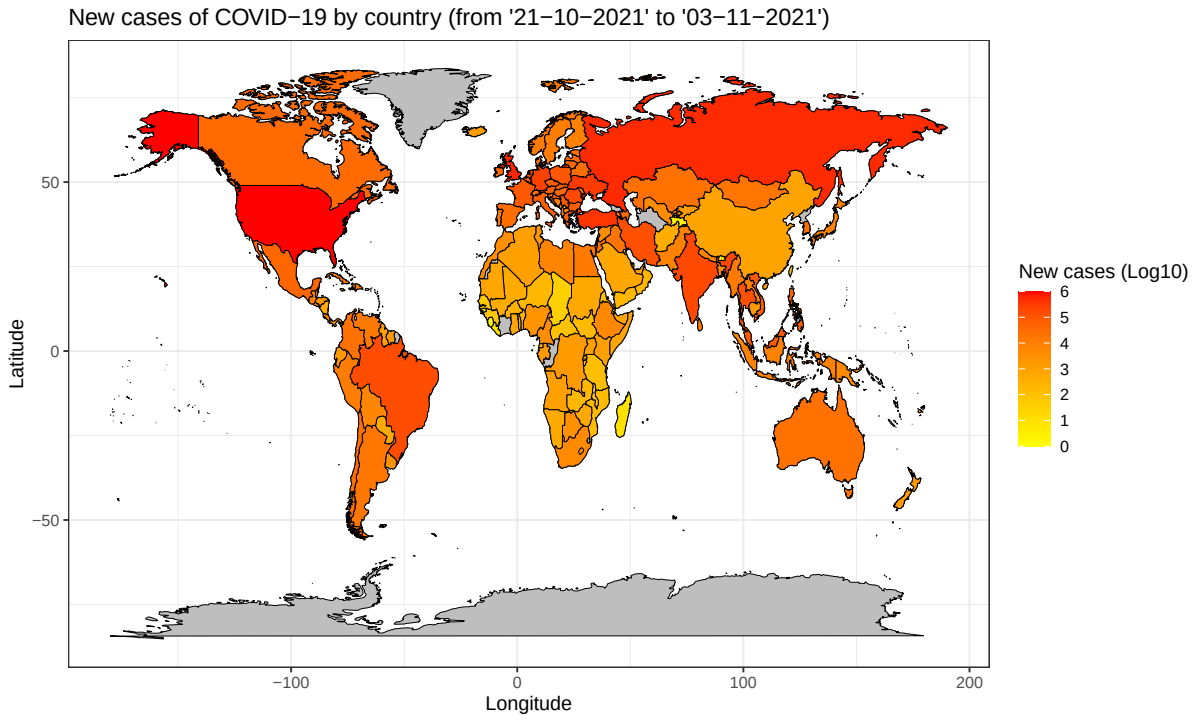
```
  mutate(corrected.sum = ifelse(is.na(sum.from.2weeks)| sum.from.2weeks<=0,-1,sum.from
      ↪ .2weeks)) %>%
  filter(corrected.sum!=-1) -> data

# Drawing world map and mapping new cases
ggplot() +
  geom_map(
    data = world_map, map = world_map,
    aes(long, lat, map_id = region),
    color = "black", fill = 'grey',size = 0.1
  ) +
  theme_bw() +
  geom_map(
    data = data, map = data,
    aes(long, lat, map_id = region, fill = log10(corrected.sum)),
    color = "black",size = 0.1
  ) +
  scale_fill_gradient(
    low = "yellow", high = "red"
  ) +
  labs(title = "New cases of COVID-19 by country (from '21-10-2021' to '03-11-2021')",
      ↪ x = "Longitude",y = "Latitude" ,fill = "New cases (Log10)")
```

Kod zawiera komentarze w celu wyjaśnieniające poszczególne bloki kodu, niemniej jednak aby lepiej zrozumieć sposób działania programu, można wyjaśnić jego funkcjonowanie w kilku prostych krokach.

1. Określenie ścieżki do pliku xlsx zawierającego dane.
2. Odfiltrowanie danych starszych niż 2 tygodnie, zmiana lokalizacji na kompatybilną z map_data("world"), suma nowych przypadków ze względu na kraj.
3. Stworzenie ramki danych niezbędnej do narysowania mapy, modyfikacja regionu na kompatybilny z d1.
4. Połączenie ramek danych d1 i world_map w jedną, odfitrowanie krajów o których nie ma aktualnych danych o zakażeniach.
5. Rysowanie wykresu.

Odpalając powyższy kod otrzymujemy następujący wykres:



Rys. 2. Nowe przypadki zakażeń COVID-19, grafika wygenerowana na podstawie danych z ourworldindata.org

4. Wnioski

Porównując obydwie wizualizaje można zauważyć, że błędy zaadresowane w tabeli 2 zostały poprawione. Na rys. 2 widzimy mapę całego świata naraz, do narysowania wybrano bardzij kontrastujące kolory, a problem nachodzących kropek został rozwiązany poprzez pomalowanie krajów z użyciem skali różnokolorowej.

Warto jednak nadmienić, że poprawiając niektóre cechy wykresu, musieliśmy zrezygnowa z innych. Skala na rys. 2 jest skalą logarytmiczną. Wykorzystanie skali logarytmicznej powoduje, że Rosja i USA wyglądają na wykresie podobnie, pomimo że dzieli je 500000 nowych przypadków. Natomiast gdyby zastosować skalę liniwą natrafiamy na inny problem. Skala liniowa jest zdominowana przez nowe przypadki w USA, gdzie ich liczba przekracza milion, przez co ciężko jest odróżnić kraje z 20000 nowo zakażonych a 2000. Problem tan był mniejszy na rys. 1, pomimo skali liniowej.

5. Pliki

Program Homework2.R służący do narysowania wykresu został zawarty w tym samym folderze co sprawozdanie z pracy domowej.