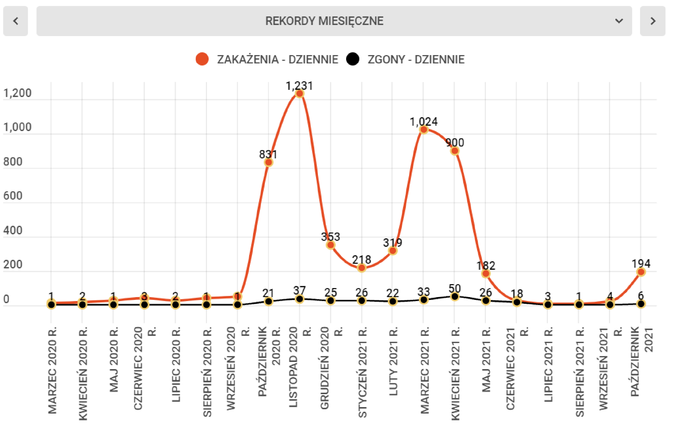
Rozwiązanie pracy domowej nr 2

Tomasz Modzelewski

Warszawa, 1 listopada 2021 r.

Wstęp

Celem niniejszej pracy domowej jest poprawienie wykresu opublikowanego na stronie internetowej Radia Kielce. Wykres ten z założenia ma przedstawiać największą dzienną liczbę zakażeń wirusem SARS-CoV-2 oraz największą dzienną liczbę zgonów z powodu COVID-19, oddzielnie dla każdego miesiąca od początku epidemii.



Źródło: http://m.radio.kielce.pl/pl/wiadomosci/juz-ponad-sto-zakazen-w-regionie-wykres,138175 (zakładka: REKORDY MIESIĘCZNE)

Wykres

Dane

Na początku przygotujmy dane niezbędne do utworzenia wykresu.

```
library(dplyr)
library(ggplot2)
library(tidyr)
```

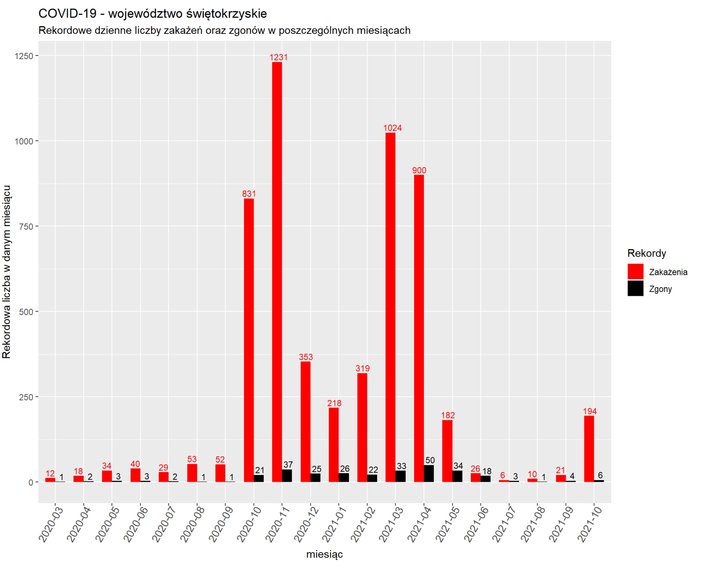
```
cases <- unlist(na.omit(read.table("zakazenia.txt")[-1]))
deaths <- unlist(na.omit(read.table("zgony.txt")[-1]))

names(cases) <- NULL
names(deaths) <- NULL

df <- data.frame(
  date = seq(
    as.Date("2020-03-04"),
    length.out=length(cases),
    by="day"
  ),
  cases = cases,
  deaths = deaths
)

df <- tibble(df)
result <- df %>% group_by(paste(format(date, format="%Y"), format(date, format="%m"), sep="-")) %>%
  summarize(Zakażenia = max(cases), Zgony = max(deaths))
colnames(result)[1] <- "miesiąc"
result <- result %>% pivot_longer(-miesiąc, names_to = "Rekordy", values_to = "rekord")

result
```

```
## # A tibble: 40 x 3
##    miesiąc Rekordy   rekord
##    <chr>   <chr>      <int>
##  1 2020-03 Zakażenia     12
##  2 2020-03 Zgony          1
##  3 2020-04 Zakażenia     18
##  4 2020-04 Zgony          2
##  5 2020-05 Zakażenia     34
##  6 2020-05 Zgony          3
##  7 2020-06 Zakażenia     40
##  8 2020-06 Zgony          3
##  9 2020-07 Zakażenia     29
## 10 2020-07 Zgony          2
## # ... with 30 more rows
```

Poprawiony wykres

Mając gotową ramkę danych, możemy wykonać poprawioną wizualizację.

```
ggplot(result, aes(x = miesiąc, y = rekord, fill = Rekordy)) +
  geom_col(position="dodge", stat="identity", width = 0.7) +
  geom_text(
    aes(label = rekord),
    position = position_dodge(1),
    vjust = -0.3,
    hjust=rep(c(0.3, 0.6), dim(result)[1]/2),
    size = 3,
    color = rep(c("red", "black"), dim(result)[1]/2)
  ) +
  ylab("Rekordowa liczba w danym miesiącu") +
  scale_fill_manual(values = c("red", "black")) +
  ggtitle(
    "COVID-19 - województwo świętokrzyskie",
    "Rekordowe dzienne liczby zakażeń oraz zgonów w poszczególnych miesiącach"
  ) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1, vjust = 1.09, size = 11))
```



Podsumowanie

Powyższy wykres zdecydowanie lepiej prezentuje dane niż ten opublikowany w internecie.

Przede wszystkim użyłem bardziej odpowiedniego typu wykresu, tj. kolumnowego. Wykorzystany w oryginale wykres liniowy jest mylący. Połączenie linią krzywą rekordów z dwóch sąsiednich miesięcy sugeruje, że między tymi dwoma wartościami były jakieś wielkości pośrednie - co oczywiście jest nieprawdą, celem wykresu jest jedynie pokazanie miesięcznych rekordów. Użycie słupków jednoznacznie wskazuje, że chodzi jedynie o dwie, przyporządkowane do każdego miesiąca liczby.

Mój wykres jest również czytelniejszy. W wersji dostępnej na stronie Radia Kielce liczby zakażeń i zgonów są umieszczone na jednym poziomie, co uniemożliwia ich jednoczesne odczytanie (zwłaszcza w miesiącach, w których zarówno zakażeń, jak i zgonów było bardzo mało), jest to możliwe dopiero po skorzystaniu z interaktywnych funkcjonalności wykresu. Na mojej grafice nie ma zaś żadnych problemów, by natychmiast odszukać interesujące nas dane z dowolnego miesiąca.

Na marginesie warto również wspomnieć, że wygenerowany przeze mnie wykres jest oparty na poprawnych danych - w przeciwieństwie do opublikowanego w sieci (np. raporty MZ jednoznacznie wskazują, że 1 maja 2021 mieliśmy 34 zgony w województwie świętokrzyskim, podczas, gdy Radio Kielce twierdzi, że rekord z tego miesiąca to 26).