HW4 : Amount of Hospitalization for COVID-19

Ebenezer Nkum* University of Texas at El Paso (UTEP)

February 27, 2023

Contents

---

1 Load data

```
# Load data
dat <- read.csv("owid-covid-data.csv", header = T)
```

*enkum@miners.utep.edu

2 Step 1: Inspect data, assess it for completeness, good formatting, and any errors

2.1 Inspect data

2.2 Formatting

```
# obtain the amounts of hospitalizations for COVID-19
dat_hosp <- dat %>% select(c("iso_code","continent",
                             "location","date",
                             "hosp_patients","hosp_patients_per_million"))

# Subset for North America
dat_hosp_NaM <- dat_hosp %>%
                filter(continent =="North America") %>%
                filter(location=="United States" |
                       location=="Canada")

#convert date column to date
dat_hosp_NaM <- dat_hosp_NaM %>% mutate(date = as.Date(date))


# filter for 2021 to 2023
dat_hosp_NaM_2021 <- dat_hosp_NaM %>%
                     filter(between(date, as.Date('2021-01-01'),
                                          as.Date('2023-02-28')))
```

2.3 Check for missing values

```
miss.info <- function(dat, filename = NULL){
  vnames <- colnames(dat); vnames
  n  <- nrow(dat)
  out <- NULL
  for(j in 1: ncol(dat)){
    vname <- colnames(dat)[j]
    x <- as.vector(dat[,j])
    n1 <- sum(is.na(x), na.rm = T)
    n2 <- sum(x=="NA", na.rm = T)
    n3 <- sum(x==" ", na.rm = T)
    nmiss <- n1 + n2 + n3
    ncomplete <- n-nmiss

    #percentage for the highest count of the unique levels
    maxlevel_perc <-
```

```
      round(max(table(x, useNA="ifany"))/sum(table(x, useNA="ifany")),4)
    out <- rbind(out, c(col.number =j,vname =vname,
                  n.levels  = length(unique(x)),
                  miss.perc = nmiss/n, maxlevel_perc= maxlevel_perc))
  }

 out <- as.data.frame(out)
 row.names(out) <- NULL
 return(out)
}

# output the unique variables and check the missing rate
miss.info(dat_hosp_NaM_2021) %>%
  kbl(booktabs=T, linesep = "", longtable = T) %>%
  kable_styling(font_size = 12, latex_options = c("HOLD_position"))
```

col.number	vname	n.levels	miss.perc	maxlevel_perc
1	iso_code	2	0	0.5
2	continent	1	0	1
3	location	2	0	0.5
4	date	782	0	0.0013
5	hosp_patients	1504	0.00575447570332481	0.0058
6	hosp_patients_per_million	1499	0.00575447570332481	0.0058

We have some missing values in the 2023 daily hospitalization for Canada.

```
# There are complete data for both countries
# filter for 2021 to 2023
dat_hosp_NaM_2021 <- dat_hosp_NaM %>%
                     filter(between(date, as.Date('2021-01-01'),
                                          as.Date('2023-02-14')))
```

3 Final Data

The final data we look at is the daily COVID-19 hospitalization per million for Canada and the USA for the period between January 1, 2021 and February 14, 2023. We obtain this data to avoid the missing reported values.

4 Aim of the Visualization

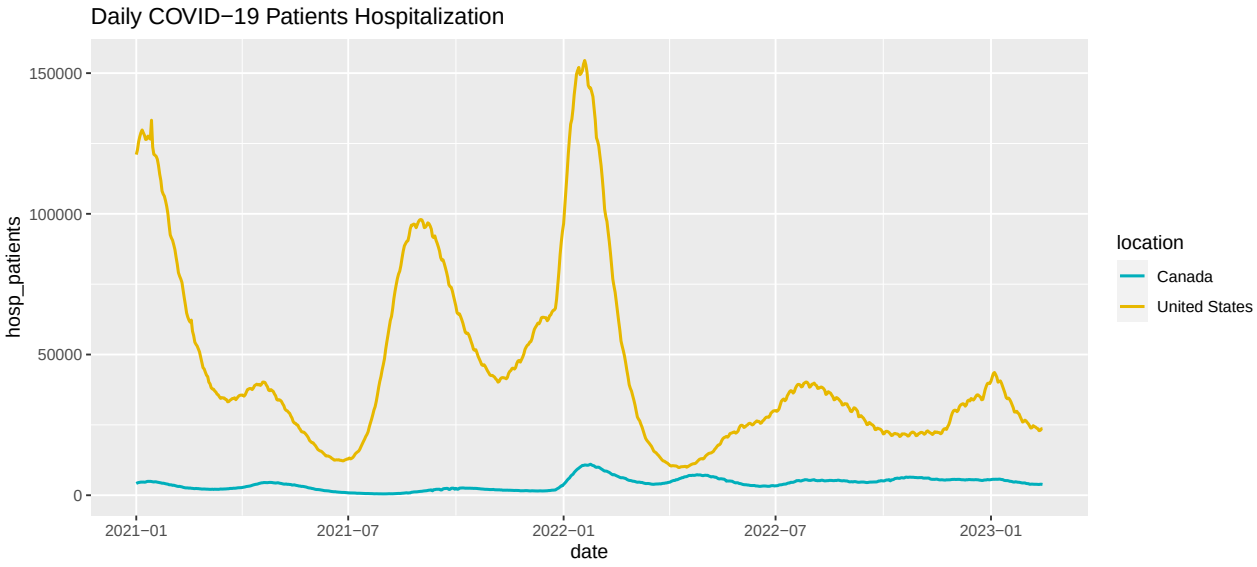
Canada and the United States, both North American countries, have been significantly impacted by the COVID-19 pandemic. One of the key measures of the impact of the pandemic has been the

number of hospitalizations due to COVID-19. The two countries share the world's longest international border, which spans over 8,891 kilometers, and are bound by the largest trading relationship in the world, with goods and services crossing the border in both directions. Given their geographical proximity and shared border, comparing COVID-19 hospitalization rates between these two countries is important to understand the similarities and differences in the pandemic's impact on their healthcare systems. Furthermore, the two countries have a long history of collaboration in various fields, including defense, energy, and environmental management, and share many cultural and social ties. Through a comparison of the daily COVID-19 hospitalization rates per million people in Canada and the US over the course of the pandemic, we can gain insight into the effectiveness of each country's response efforts and inform public health policies to mitigate the impact of the pandemic.

5 Cook book approach

```
# plot the graph : cookbook
# Multiple line plot
p <- ggplot(dat_hosp_NaM_2021, aes(x = date, y = hosp_patients)) +
  geom_line(aes(color = location), size = 0.8) +
  scale_color_manual(values = c("#00AFBB", "#E7B800")) +
  ggtitle("Daily COVID-19 Patients Hospitalization")


# Turn it interactive with ggplotly
# p <- ggplotly(p)
p
```



In my previous approach, I created a graph to compare the daily COVID-19 hospitalizations between Canada and the USA. However, I realized that the population density in these regions renders a direct comparison inaccurate. This is because the total number of hospitalizations can be influ-

enced by population size and may not provide a reliable reflection of the pandemic's impact on a population.

To address this issue, I have decided to use hospitalizations per million people instead as this provides a more meaningful measure of the pandemic's impact. It accounts for differences in population size and allows for more accurate comparisons between populations of different sizes.

Furthermore, I considered removing the background theme from the graph as this could be considered redundant and adds unnecessary visual clutter. Also, I opted to add direct labels to the graph which can reduce the cognitive load require to interpret the data. This approach not only improves accessibility for colorblind readers but also maximizes the data-ink ratio by minimizing unnecessary visual elements.

6 Final plot and improvement

We observe that after March in 2022, the impact of COVID for Canada has been higher than in USA when we measure by the impact by Hospitalization.

```
# plot the graph : cookbook


# Multiple line plot
p <- ggplot(dat_hosp_NaM_2021, aes(x = date, y = hosp_patients_per_million)) +
  geom_line(aes(color = location), size = 0.8) +
  ylab("Hospital Patients Per Million")+
  xlab("Date")+
  scale_color_manual(values = c("#00AFBB", "#E7B800")) +
  ggtitle("Daily COVID-19 Patients Hospitalization Per Million")+
  geom_vline(xintercept = as.numeric(as.Date("2022-03-15")), color = "red")+
  theme_classic() +
  theme(legend.position = "none")

# Add text to the left and right sides of the plot
p <- p + annotate("text", x = as.Date("2021-05-05"), y = 500,
      label = "The impact of COVID-19 is higher\n in the USA compared to Canada.",
      hjust = 0, size = 3, color = "#00AFBB") +
  annotate("text", x = as.Date("2022-10-01"), y = 500,
    label = "The impact of COVID-19 is higher\n in the Canada compared to USA",
        hjust = 1, size = 3, color = "#E7B800") +
    annotate("text", x = as.Date("2023-03-30"), y = 100,
    label = "Canada", hjust = 1, size = 3, color = "#00AFBB") +
      annotate("text", x = as.Date("2023-03-15"), y = 68,
    label = "USA", hjust = 1, size = 3, color = "#E7B800")



# Turn it interactive with ggplotly
```

```
# p <- ggplotly(p)
 p
```