Midterm Project

Guok Wei Jie

2024-04-07

COVID19 EDA

In this EDA, I will be exploring how **vaccinations have affected the case fatality rate of COVID-19 patients across the lower 48 states of USA.**

Data Visualizing and Cleaning

This creates a data frame `df` for the state-level COVID-19 data.

```
df <- covid19(level = 2, verbose = FALSE)
```

Out of all the countries in the data set, **USA** has the most number of observations. Let's focus on this country for our EDA.

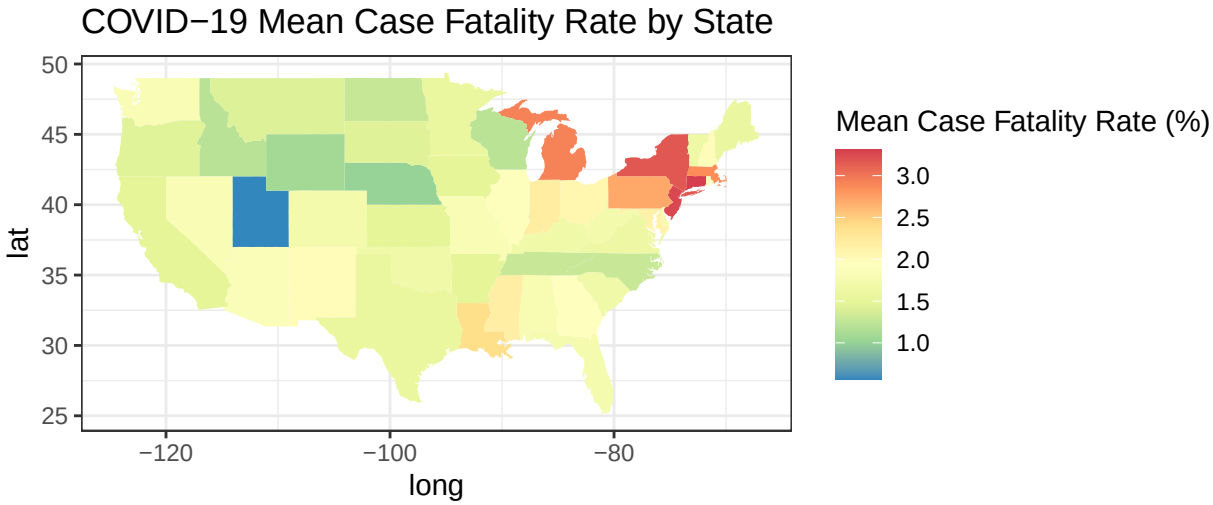
```
df_grouped_country <- df %>%
  group_by(administrative_area_level_1) %>%
  summarise(NumberOfRows = n()) %>%
arrange(desc(NumberOfRows))

df_USA <- filter(df, administrative_area_level_1 == "United States")

# Removing rows with NA values for vaccination policy but non-NA values for vaccines
df_USA <- df_USA[!(is.na(df_USA$vaccination_policy) & !is.na(df_USA$vaccines)),]
```

To visualize how vaccinations affect the case fatality rate, I plotted a geospatial state map of the lower 48 states showing the mean case fatality rate, as well as the mean percentage of fully vaccinated individuals, on separate graphs.

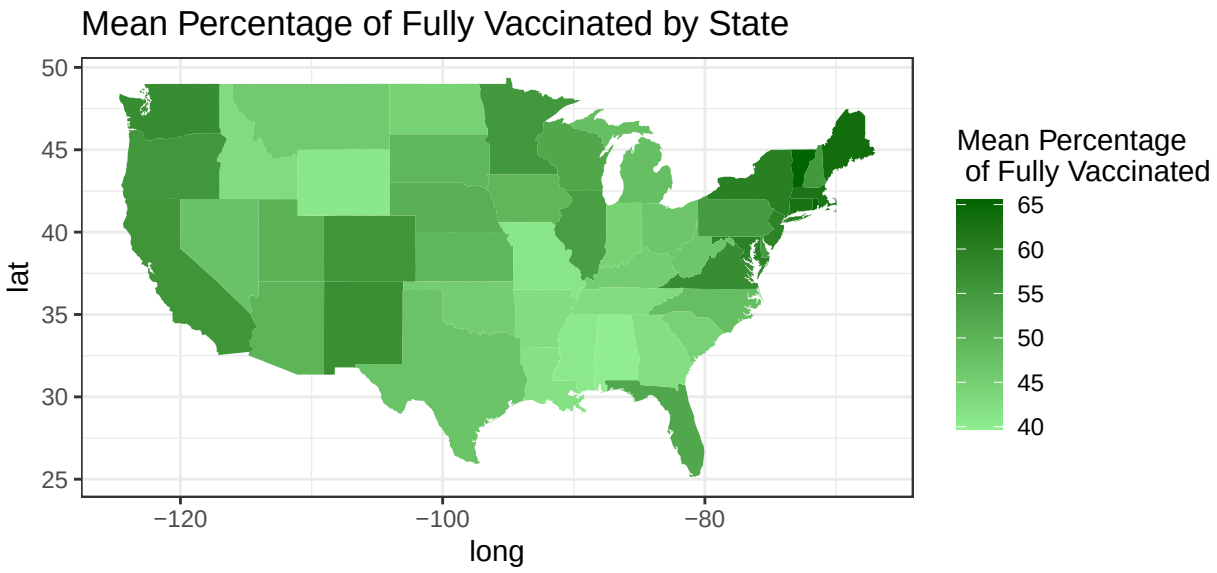
```
df_USA$administrative_area_level_2 <- tolower(df_USA$administrative_area_level_2)

us_states <- map_data("state") %>%
  select(long, lat, group, region)

# Only focus on lower 48 states as the others do not have valid data for vaccination
df_extra_states <- filter(df_USA, tolower(administrative_area_level_2)
                          %in% c("alaska", "hawaii", "northern mariana islands",
                                 "virgin islands", "guam", "american samoa",
                                 "puerto rico"))
```

```r
df_lower_48 <- anti_join(df_USA, df_extra_states, by = "administrative_area_level_2")

df_cases <- df_lower_48 %>%
  mutate(case_fatality_rate = deaths/confirmed * 100,
         percentage_fully_vaccinated =
           people_fully_vaccinated/population * 100) %>%
  group_by(administrative_area_level_2) %>%
  summarise(
    case_fatality_rate = mean(case_fatality_rate, na.rm = TRUE),
    percentage_fully_vaccinated = mean(percentage_fully_vaccinated, na.rm =TRUE)
  )

top_5_states <- df_cases %>%
  top_n(5, case_fatality_rate)

merged_data <- left_join(us_states, df_cases, by =
                           c("region"= "administrative_area_level_2"))

ggplot(merged_data, aes(x = long, y = lat, group = group)) +
  geom_polygon(aes(fill = case_fatality_rate), size = 0.25) +
  coord_quickmap() +
  scale_fill_distiller(palette = "Spectral", name = "Mean Case Fatality Rate (%)") +
  labs(title = "COVID-19 Mean Case Fatality Rate by State") +
  theme_bw()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
ggplot(merged_data, aes(x = long, y = lat, group = group)) +
  geom_polygon(aes(fill = percentage_fully_vaccinated), size = 0.25) +
  coord_quickmap() +
  scale_fill_gradient(low = "lightgreen", high = "darkgreen",
                      name = "Mean Percentage \n of Fully Vaccinated") +
  labs(title = "Mean Percentage of Fully Vaccinated by State") +
  theme_bw()
```

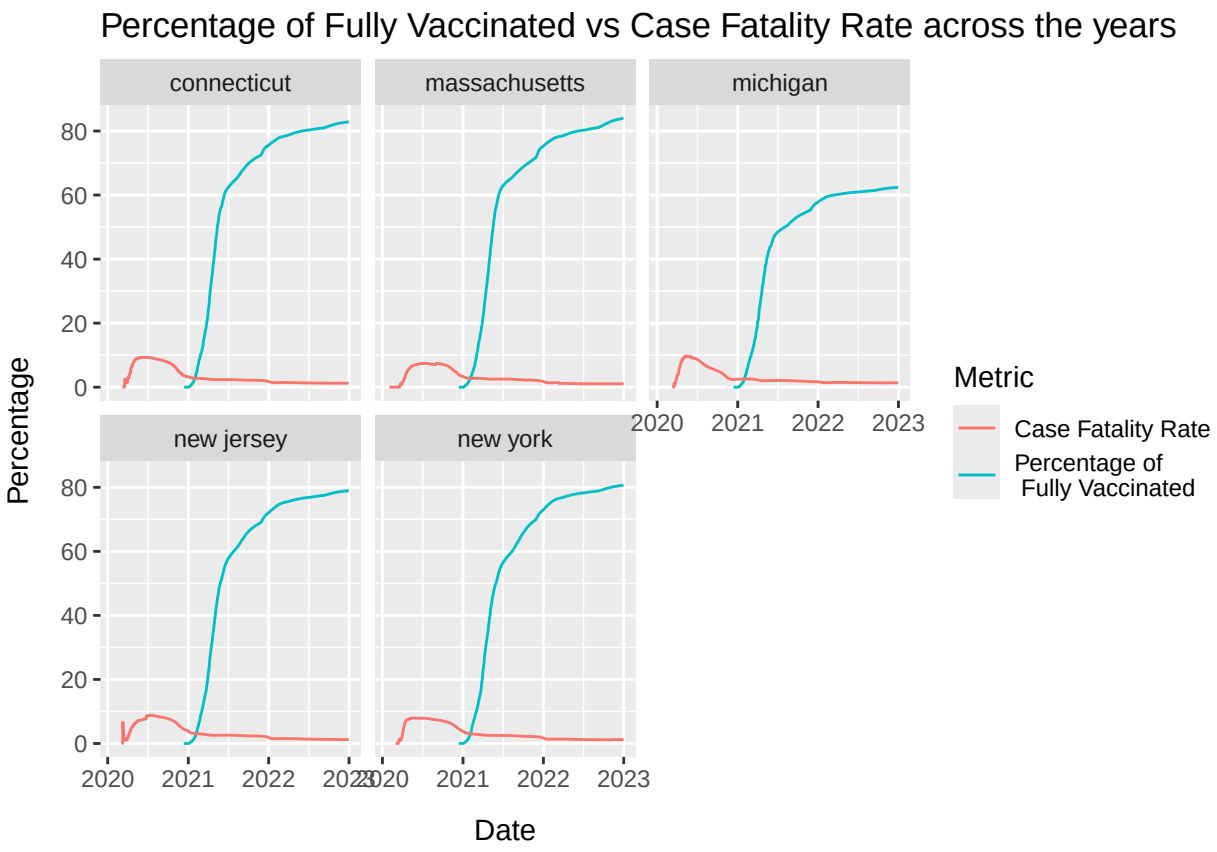


To better analyse the outliers, I plotted a line plot faceted by the top 5 states which are mostly outliers, in order to see how case fatality rate changes with percentage of fully vaccinated.

```
df_highest_fatality <- filter(df_USA, administrative_area_level_2
                              %in% c('new york', 'connecticut','new jersey',
                                     'massachusetts', 'michigan') &
                                !is.na(vaccination_policy))

df_highest_fatality <- df_highest_fatality %>% arrange(date) %>%
  mutate(fully_vaccinated = people_fully_vaccinated/population * 100,
         fatality_rate = deaths/confirmed * 100)

df_highest_fatality <- df_highest_fatality %>%
  select(date, fully_vaccinated, fatality_rate,
         administrative_area_level_2)

ggplot(df_highest_fatality, aes(x = date)) +
  geom_line(aes(y = fully_vaccinated, color =
                  "Percentage of \n Fully Vaccinated")) +
  geom_line(aes(y = fatality_rate, color = "Case Fatality Rate")) +
  labs(x = "Date", y = "Percentage", color = "Metric",
       title = "Percentage of Fully Vaccinated vs Case Fatality Rate across the years") +
  facet_wrap(~administrative_area_level_2) +
  theme( axis.title.x = element_text(margin = margin(t = 10)),
         axis.title.y = element_text(margin = margin(r = 10)))
```

```
## Warning: Removed 288 rows containing missing values or values outside the scale range
## (`geom_line()`).
```

```
## Warning: Removed 8 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



To analyse how different levels of vaccination policy levels affect the percentage of fully vaccinated and therefore the case fatality rate, I made a faceted plot by vaccination policy for all of the states.

```
df_vac_policy <- df_lower_48 %>%
  filter(!is.na(vaccination_policy)) %>%
  mutate(case_fatality_rate = deaths/confirmed * 100,
         vaccination_rate = people_fully_vaccinated/population * 100) %>%
  group_by(administrative_area_level_2, vaccination_policy) %>%
  summarise(
    case_fatality_rate = mean(case_fatality_rate, na.rm = TRUE),
    vaccination_rate = mean(vaccination_rate, na.rm =TRUE)
  )
```

```
## `summarise()` has grouped output by 'administrative_area_level_2'. You can
## override using the `.groups` argument.
```

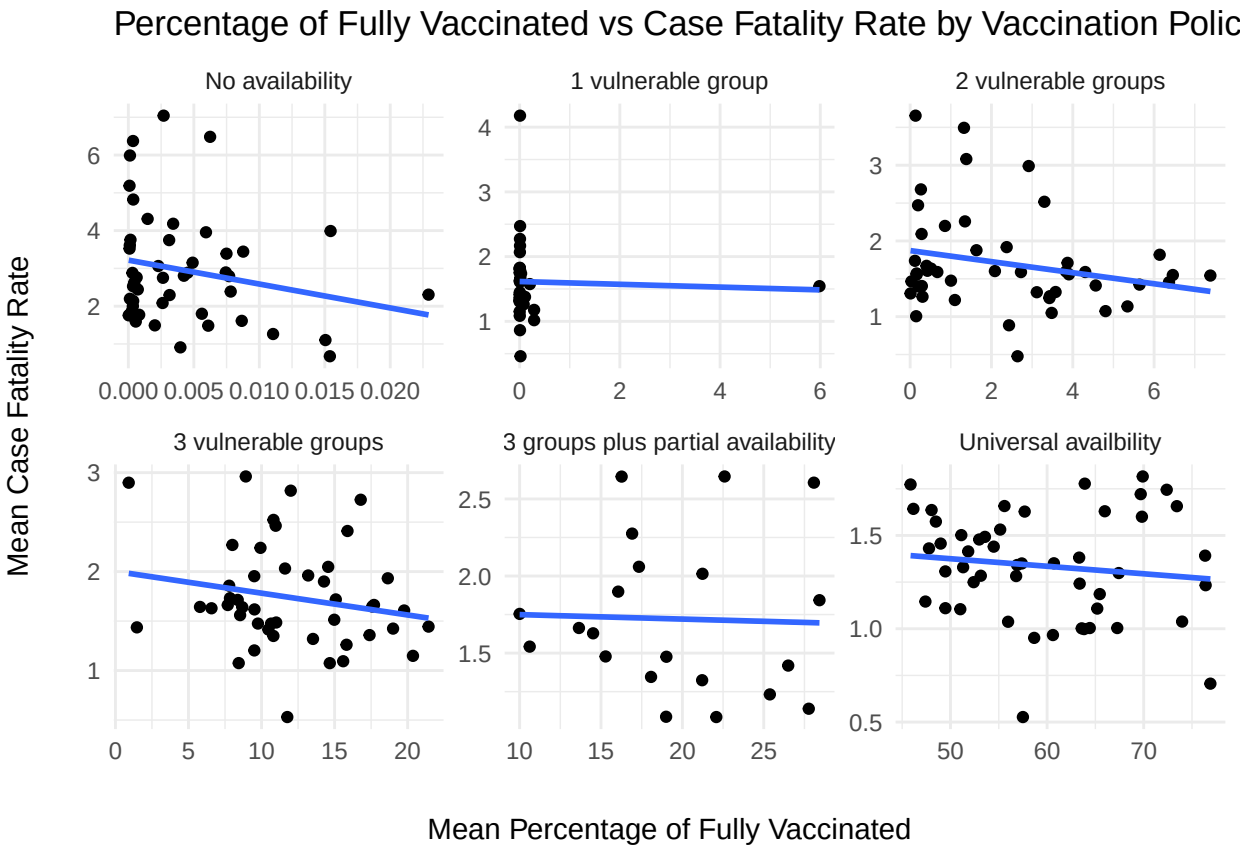
```
df_vac_policy$vaccination_policy <- factor(df_vac_policy$vaccination_policy,
                levels = 0:5,
                labels = c("No availability",
```

```
                          "1 vulnerable group",
                          "2 vulnerable groups",
                          "3 vulnerable groups",
                          "3 groups plus partial availability",
                          "Universal availbility"))

# Facet the plot by vaccination policy
ggplot(df_vac_policy, aes(x = vaccination_rate, y = case_fatality_rate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~ vaccination_policy, scales = "free") +
  labs(
    title = "Percentage of Fully Vaccinated vs Case Fatality Rate by Vaccination Policy",
    x = "Mean Percentage of Fully Vaccinated",
    y = "Mean Case Fatality Rate"
  ) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(margin = margin(t = 20)),
    axis.title.y = element_text(margin = margin(r = 20))
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_smooth()`).
```

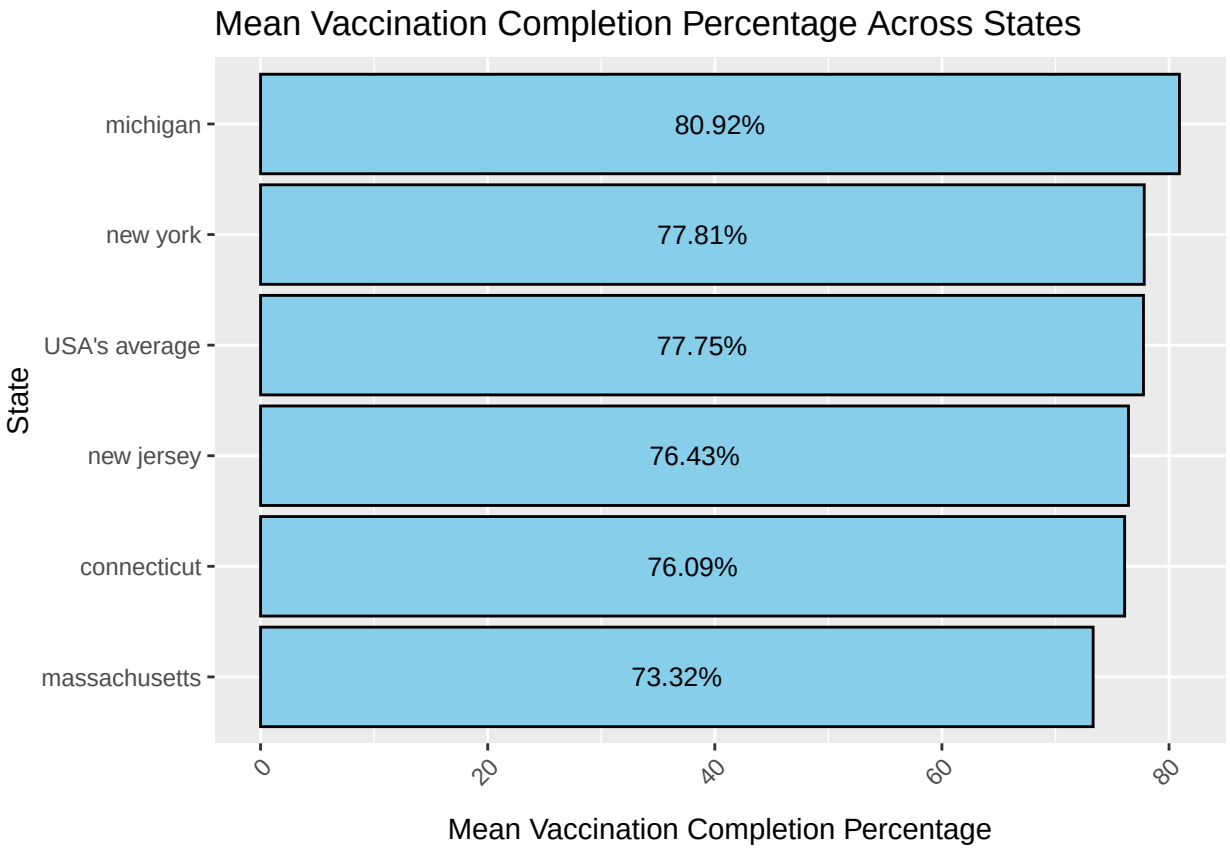
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



Vaccination completion can affect the efficacy of the vaccine, and hence it is important to compare to see how many vaccinated individuals are fully vaccinated. I plotted a bar plot to compare the top 5 states with highest case fatality rates with the USA's average.

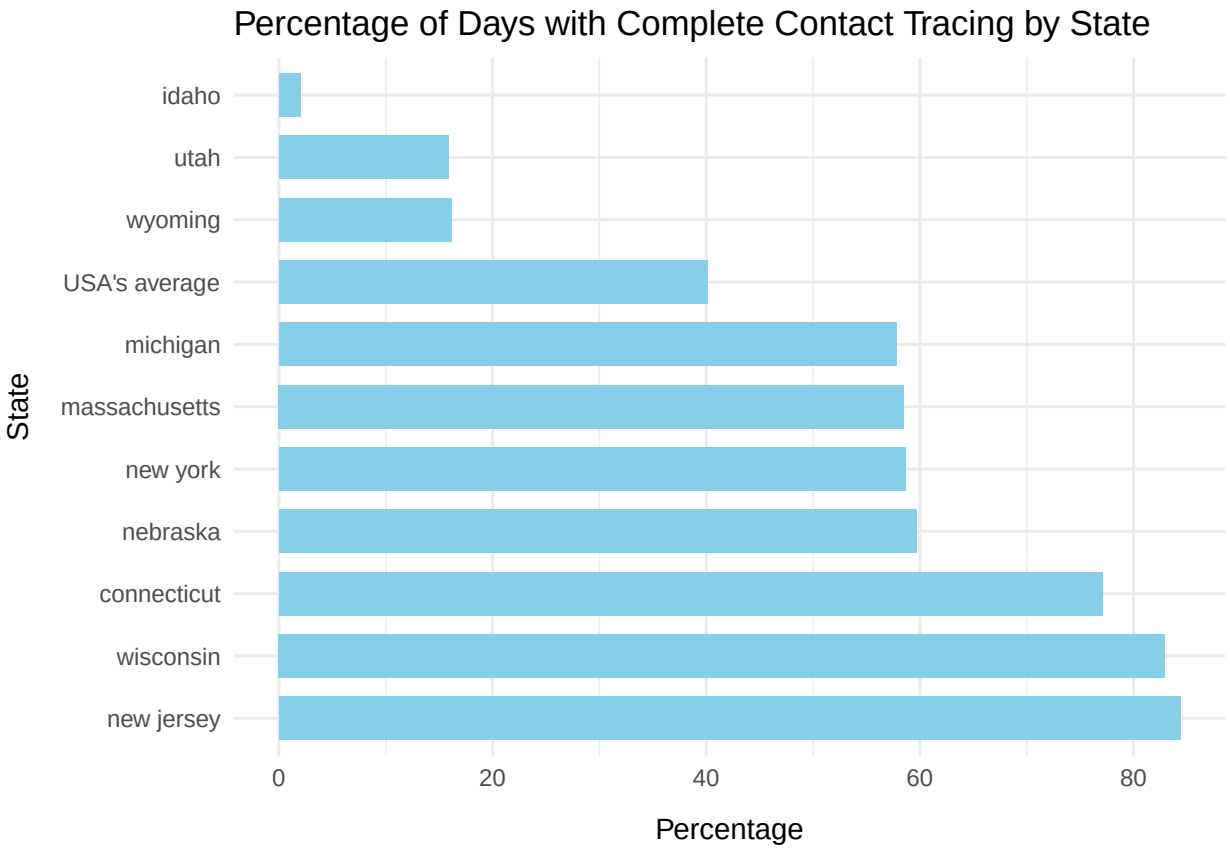
```
df_fully_vaxxed <- df_lower_48 %>%
  filter(!is.na(people_fully_vaccinated) & !is.na(people_vaccinated)) %>%
  mutate(vaccination_completion = people_fully_vaccinated / people_vaccinated * 100) %>%
  group_by(administrative_area_level_2) %>%
  summarise(
    vaccination_completion = mean(vaccination_completion, na.rm = TRUE)
  )

# Calculate the overall average
overall_average <- mean(df_fully_vaxxed$vaccination_completion, na.rm = TRUE)

# Arrange the dataframe by percentage_fully_vaxxed
df_fully_vaxxed <- df_fully_vaxxed %>%
  arrange(vaccination_completion)

# Select the top 5 and worst 5 states
top_bottom_states <- bind_rows(head(df_fully_vaxxed, 5), tail(df_fully_vaxxed, 5))

# Create a dataframe for the overall average
overall_average_df <- data.frame(administrative_area_level_2 =
                                   "USA's average",
                                 vaccination_completion = overall_average)
```

```r
df_states_of_interest <- filter(df_fully_vaxxed,
                                administrative_area_level_2
                                %in% top_5_states$administrative_area_level_2)

# Combine the overall average dataframe with top_bottom_states
final_plot_data <- rbind(df_states_of_interest, overall_average_df)

ggplot(final_plot_data, aes(x = reorder(administrative_area_level_2,
                                        vaccination_completion), y =
                            vaccination_completion)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  geom_text(aes(label = paste0(round(vaccination_completion, 2), "%")),
            position = position_stack(vjust = 0.5),
            size = 3.5, color = "black") +
  labs(x = "State", y = "Mean Vaccination Completion Percentage",
       title = "Mean Vaccination Completion Percentage Across States") +
  theme(axis.title.x = element_text(margin = margin(t = 10))) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```



The level of contact tracing affects the number of cases confirmed and also the number of deaths from COVID-19. Hence, I wanted to analyse how it affected the top 5 and bottom 5 states for case fatality rate, as compared to the average for the country.

```r
bottom_5_states <- df_cases %>%
  arrange(case_fatality_rate) %>%
  slice(1:5)

states_of_interest <- bind_rows(bottom_5_states, top_5_states)

df_check <- df_lower_48 %>%
  filter(contact_tracing < 0)

df_contact_tracing <- df_lower_48 %>%
  filter(!is.na(contact_tracing)) %>%
  group_by(administrative_area_level_2) %>%
  summarise(
    percentage_2 = sum(contact_tracing == 2) / n() * 100
  ) %>%
  arrange(desc(percentage_2))

overall_average <- mean(df_contact_tracing$percentage_2, na.rm = TRUE)

percentage_2_average_df <- data.frame(administrative_area_level_2 =
                                        "USA's average",
                                      percentage_2 = overall_average)

df_states_of_interest <- filter(df_contact_tracing,
                                administrative_area_level_2
                                %in% states_of_interest$administrative_area_level_2)

final_plot <- rbind(df_states_of_interest, percentage_2_average_df)

ggplot(final_plot, aes(x = reorder(administrative_area_level_2,
                                   -percentage_2), y = percentage_2)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  labs(
    title = "Percentage of Days with Complete Contact Tracing by State",
    x = "State",
    y = "Percentage"
  ) +
  theme_minimal() +
  coord_flip() +
  theme(
    axis.title.x = element_text(margin = margin(t = 10)),
    axis.title.y = element_text(margin = margin(r = 10))
  )
```


Percentage of Days with Complete Contact Tracing by State
idaho
utah
wyoming
USA's average
michigan
massachusetts
new york
nebraska
connecticut
wisconsin
new jersey
State
0
20
40
60
80
Percentage