Novel Coronavirus (COVID-19) Data Analysis

VM

2024-06-18

Loading the Data

```
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(usmap)
library(prophet)

# Load the data
covid_df <- read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_da
```

Initial Data Exploration

Sample Data Display

Printing first few rows of the data to understand it's structure.

```
head(covid_df)
```

```
## # A tibble: 6 x 1,154
##        UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
## 1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 2 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## 6 84001011 US    USA     840  1011 Bullock Alabama        US              32.1
## # i 1,145 more variables: Long_ <dbl>, Combined_Key <chr>, `1/22/20` <dbl>,
## #   `1/23/20` <dbl>, `1/24/20` <dbl>, `1/25/20` <dbl>, `1/26/20` <dbl>,
## #   `1/27/20` <dbl>, `1/28/20` <dbl>, `1/29/20` <dbl>, `1/30/20` <dbl>,
## #   `1/31/20` <dbl>, `2/1/20` <dbl>, `2/2/20` <dbl>, `2/3/20` <dbl>,
## #   `2/4/20` <dbl>, `2/5/20` <dbl>, `2/6/20` <dbl>, `2/7/20` <dbl>,
## #   `2/8/20` <dbl>, `2/9/20` <dbl>, `2/10/20` <dbl>, `2/11/20` <dbl>,
## #   `2/12/20` <dbl>, `2/13/20` <dbl>, `2/14/20` <dbl>, `2/15/20` <dbl>, ...
```

Data Description

The dataset provided contains the time series data for confirmed COVID-19 cases in the United States. The data is maintained by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

Dataset Structure: - Province/State: The name of the state or province. - Country/Region: The name of the country or region (in this case, "US"). - Lat: Latitude coordinate of the location. - Long: Longitude coordinate of the location. - Date Columns: Each subsequent column represents the number of confirmed COVID-19 cases on a specific date, starting from the earliest date recorded.

This data also contain additional columns which are redundant and can be removed to simplify the analysis.

Summary Statistics

```
total_cases_per_state <- covid_df %>%
  select(-c(`UID`, `iso2`, `iso3`, `code3`, `FIPS`, `Admin2`, `Country_Region`, `Lat`, `Long_`, `Combine
  gather(key = "Date", value = "Cases", -`Province_State`) %>%
  group_by(`Province_State`) %>%
  summarise(Total_Cases = sum(Cases, na.rm = TRUE)) %>%
  arrange(desc(Total_Cases))

head(total_cases_per_state)
```

```
## # A tibble: 6 x 2
##   Province_State Total_Cases
##   <chr>                <dbl>
## 1 California      6166190335
## 2 Texas           4566537657
## 3 Florida         3978357707
## 4 New York        3392006819
## 5 Illinois        2122240785
## 6 Pennsylvania    1836846159
```

Time Series Analysis

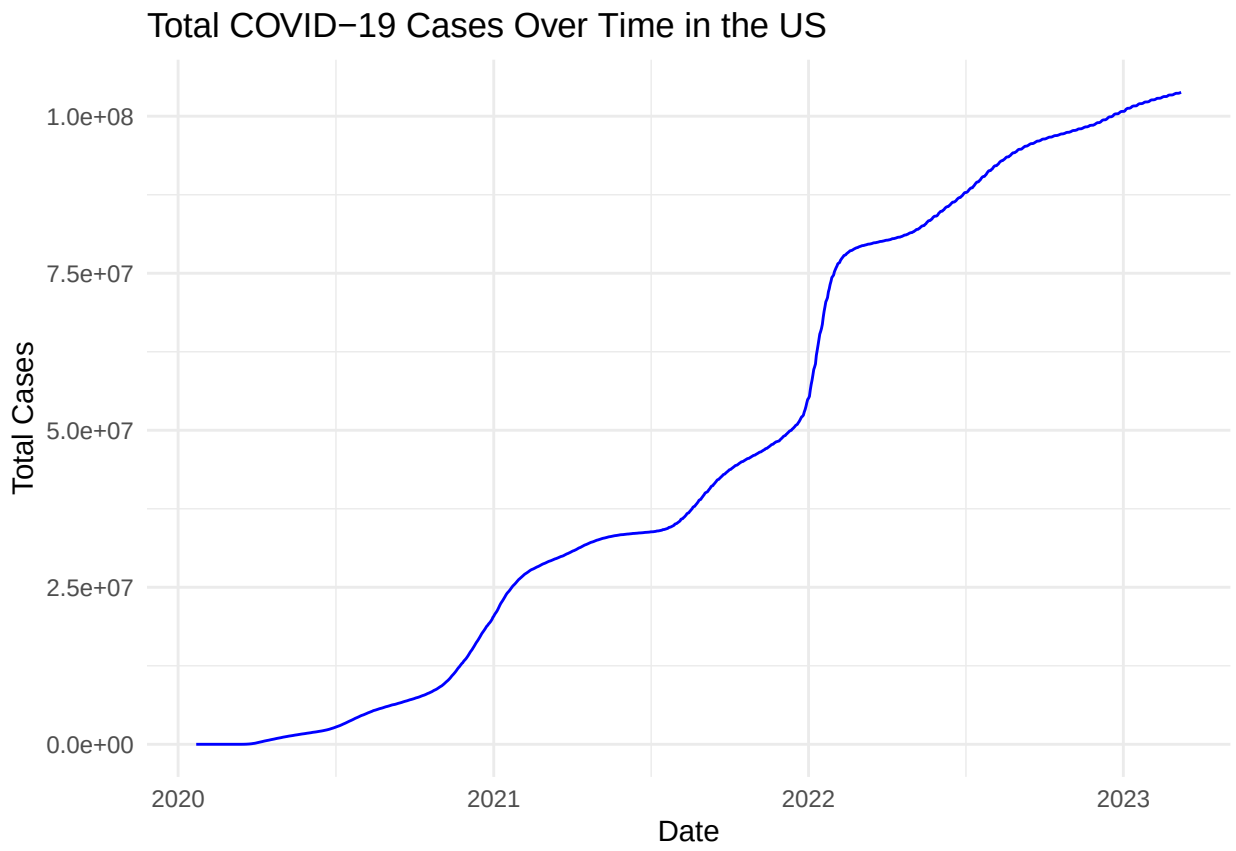
This is clear from the data that it is Time Series data, so we can perform some basis Time Series analysis.

Total Cases Over Time for the US:

```
us_total_cases <- covid_df %>%
  select(-c(UID, iso2, iso3, code3, FIPS, Admin2, Lat, Long_, Combined_Key, Country_Region)) %>%
  gather(key = "Date", value = "Cases", -Province_State) %>%
  group_by(Date) %>%
  summarise(Total_Cases = sum(Cases, na.rm = TRUE))

# Convert Date to proper date format
us_total_cases$Date <- as.Date(us_total_cases$Date, format = "%m/%d/%y")

# Plot total cases over time
ggplot(us_total_cases, aes(x = Date, y = Total_Cases)) +
  geom_line(color = "blue") +
  labs(title = "Total COVID-19 Cases Over Time in the US",
       x = "Date",
       y = "Total Cases") +
  theme_minimal()
```

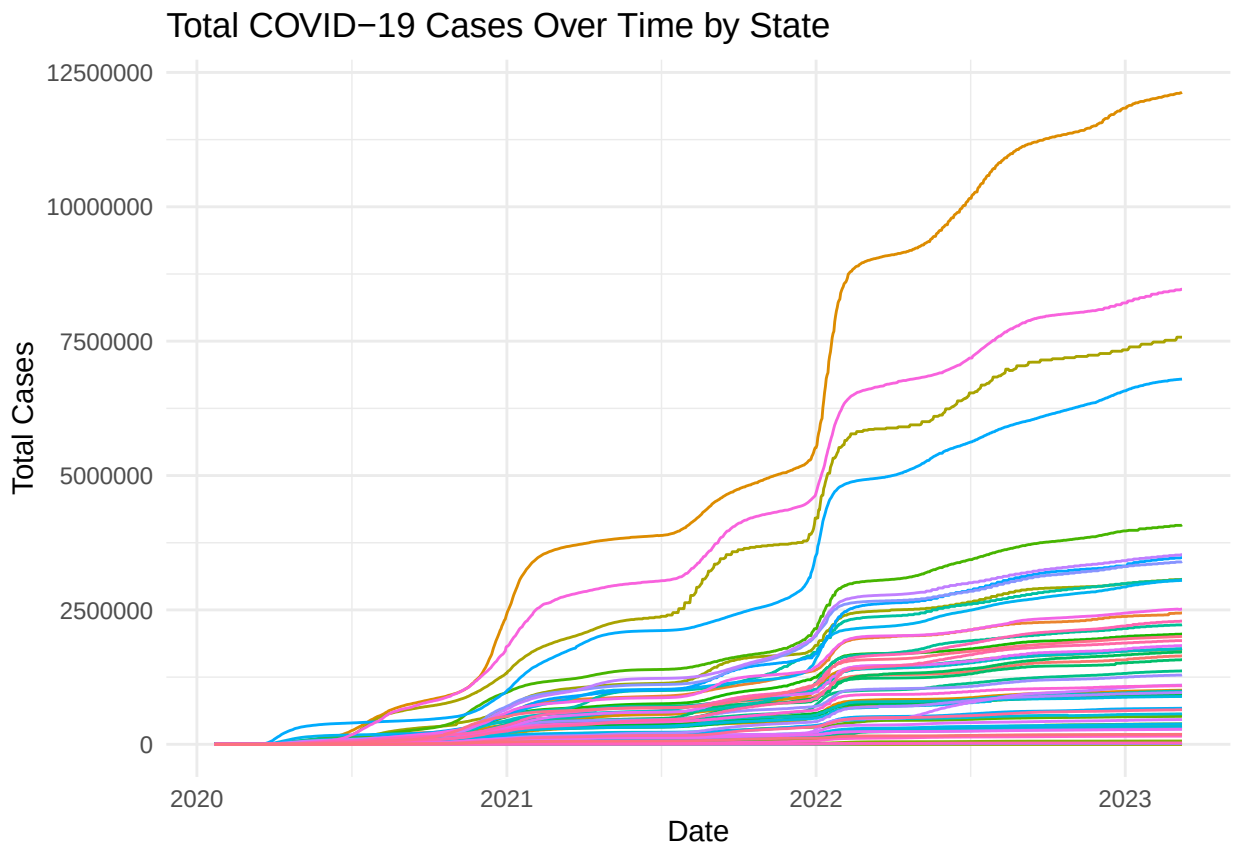


Total Cases Over Time by State

```
state_cases_over_time <- covid_df %>%
  select(-c(UID, iso2, iso3, code3, FIPS, Admin2, Lat, Long_, Combined_Key, Country_Region)) %>%
  gather(key = "Date", value = "Cases", -Province_State) %>%
  group_by(Province_State, Date) %>%
  summarise(Total_Cases = sum(Cases, na.rm = TRUE))

# Convert Date to proper date format
state_cases_over_time$Date <- as.Date(state_cases_over_time$Date, format = "%m/%d/%y")

# Plot total cases over time by state
ggplot(state_cases_over_time, aes(x = Date, y = Total_Cases, color = Province_State)) +
  geom_line() +
  labs(title = "Total COVID-19 Cases Over Time by State",
       x = "Date",
       y = "Total Cases") +
  theme_minimal() +
  theme(legend.position = "none")
```

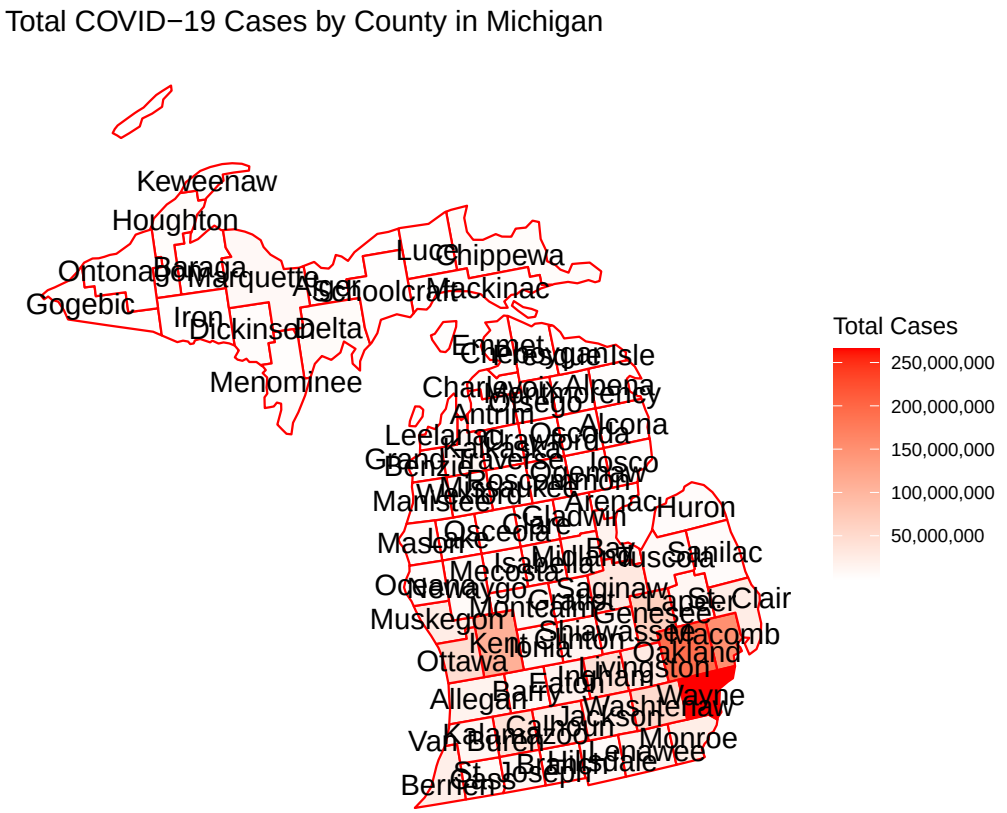


Geographical Analysis - Heatmap of Total Cases by State

I wanted to analyse the Cases in Michigan State. I wanted to see the case distribution in different county as heatmap.

```
michigan_data <- covid_df %>%
  filter(Province_State == "Michigan")
michigan_total_cases <- michigan_data %>%
  select(-c(UID, iso2, iso3, code3, Lat, Long_, Combined_Key, Country_Region)) %>%
  gather(key = "Date", value = "Cases", -c(Admin2, Province_State, FIPS)) %>%
  group_by(Admin2, FIPS) %>%
  summarise(Total_Cases = sum(Cases, na.rm = TRUE)) %>%
  rename(county = Admin2, fips = FIPS)

plot_usmap(regions = "counties", include = c("MI"), data = michigan_total_cases, values = "Total_Cases"
  scale_fill_continuous(low = "white", high = "red", name = "Total Cases", label = scales::comma) +
  labs(title = "Total COVID-19 Cases by County in Michigan") +
  theme(legend.position = "right")
```



Time Series Analysis of Michigan Data

To understand more about Michigan State Cases, I wanted to create a simple Time Series Forecasting. I am not doing all the Time Series analysis as it is not in the scope of the work.

Here I am using Prophet Model to create Uni-variant Time Series forecasting.

Data Preparation

```
# Aggregate the data by date
michigan_time_series <- michigan_data %>%
  select(-c(UID, iso2, iso3, code3, FIPS, Admin2, Lat, Long_, Combined_Key, Country_Region)) %>%
  gather(key = "Date", value = "Cases", -Province_State) %>%
  group_by(Date) %>%
  summarise(Total_Cases = sum(Cases, na.rm = TRUE))

# Convert Date to proper date format
michigan_time_series$Date <- as.Date(michigan_time_series$Date, format = "%m/%d/%y")

# Prepare data for prophet
michigan_prophet <- michigan_time_series %>%
  rename(ds = Date, y = Total_Cases)
```

Model Development

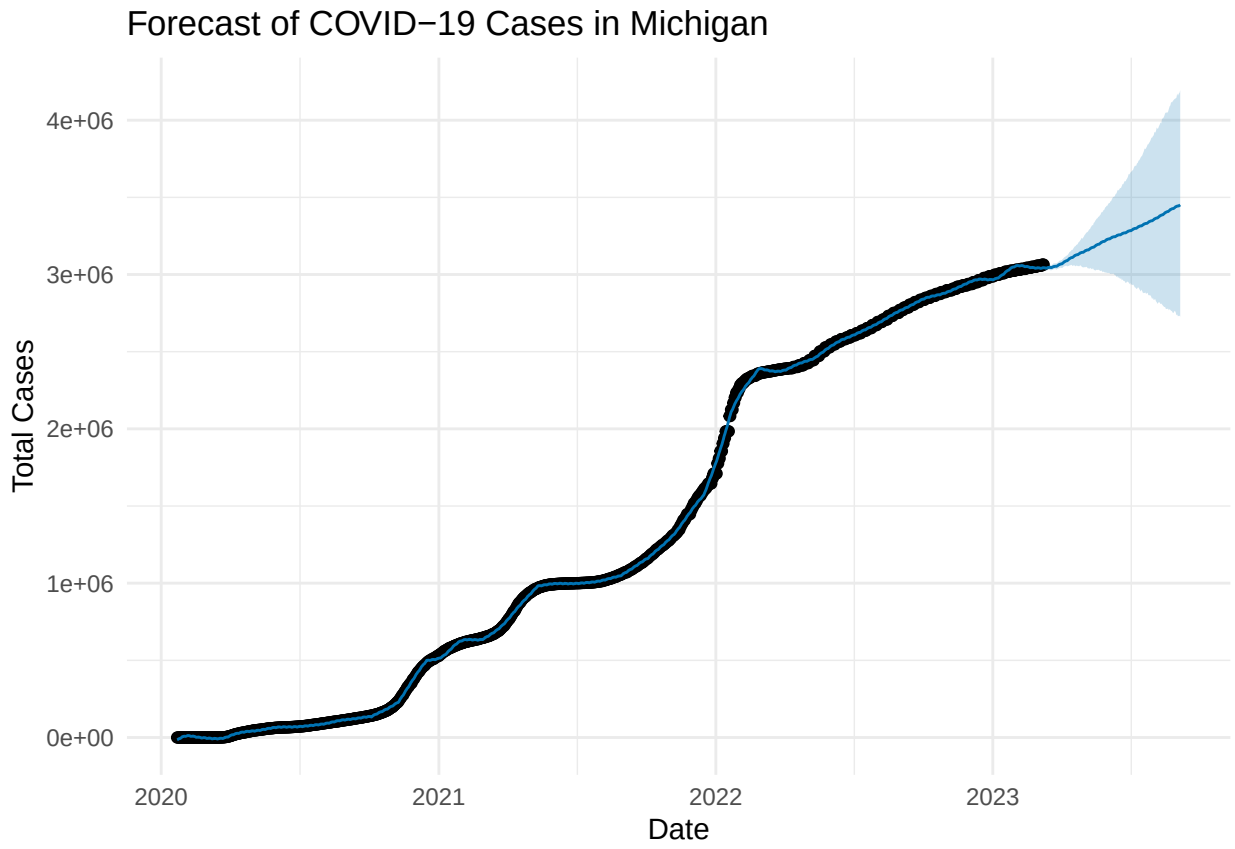
```
# Fit the prophet model
m <- prophet(michigan_prophet)
```

```r
# Create a dataframe for future dates
future <- make_future_dataframe(m, periods = 180)  # Forecasting for the next 180 days

# Forecast future cases
forecast <- predict(m, future)
```

Forecasted Results

```r
plot(m, forecast) +
  ggtitle("Forecast of COVID-19 Cases in Michigan") +
  xlab("Date") +
  ylab("Total Cases") +
  theme_minimal()
```



Investigating Possible Bias in the data

Identifying bias in data, especially in a dataset as large and complex as COVID-19 case counts, requires a systematic approach. Here are a few ways to investigate potential bias in the data:

1. **Data Collection Methods**: Verify how the data was collected. Bias can occur if there are differences in testing rates, reporting practices, or data collection methods across different regions or over time.
2. **Data Completeness**: Check for missing data. Inconsistent reporting or missing data can introduce bias.
3. **Temporal Consistency**: Look for changes in data reporting practices over time. For instance, if there were changes in testing availability or public health policies, these could introduce bias.

4. **Geographical Consistency**: Compare data across different regions (states, counties) to identify inconsistencies that could indicate bias.

Data Completeness

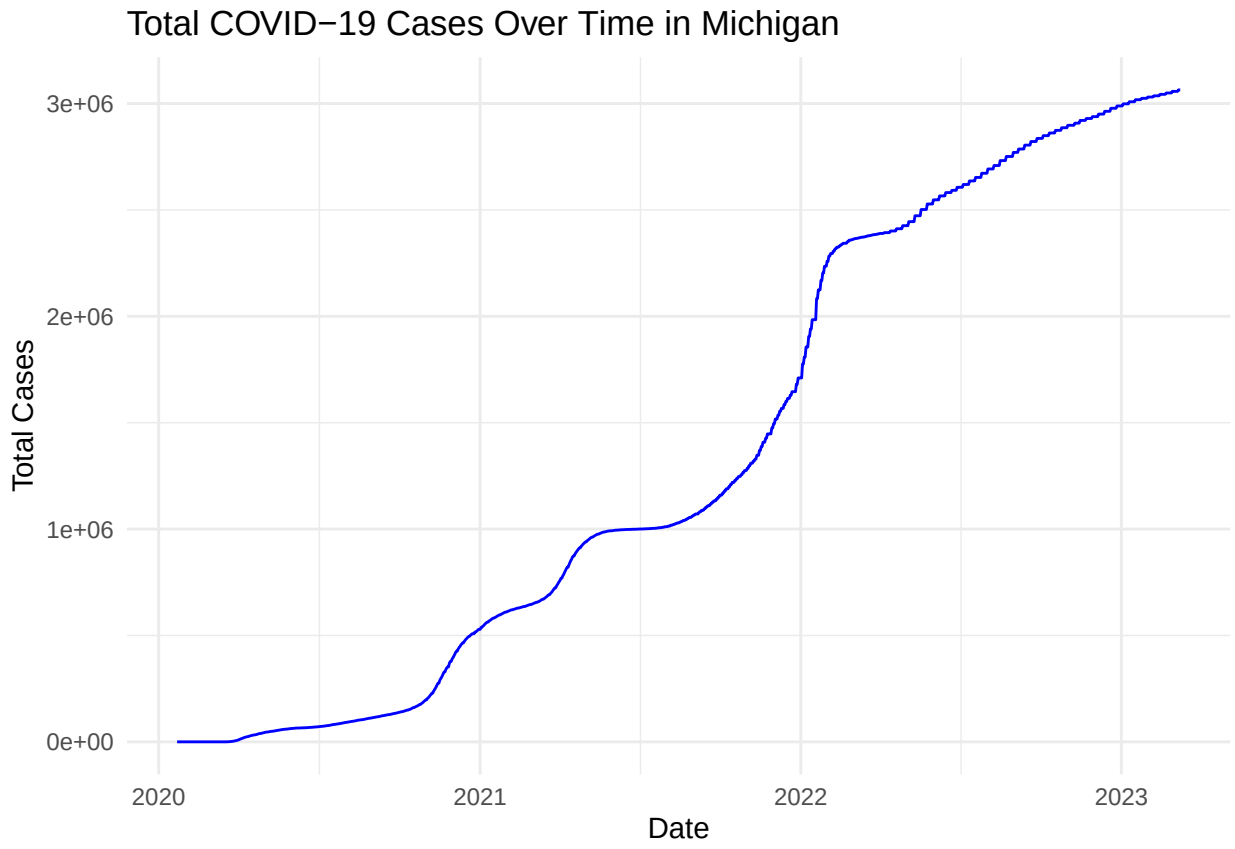
```
sum(is.na(michigan_data))
```

```
## [1] 2
```

This indicates that, there is few missing data, so looks like the data collection is nearly complete.

Temporal Analysis

```
ggplot(michigan_time_series, aes(x = Date, y = Total_Cases)) +
  geom_line(color = "blue") +
  labs(title = "Total COVID-19 Cases Over Time in Michigan",
       x = "Date",
       y = "Total Cases") +
  theme_minimal()
```



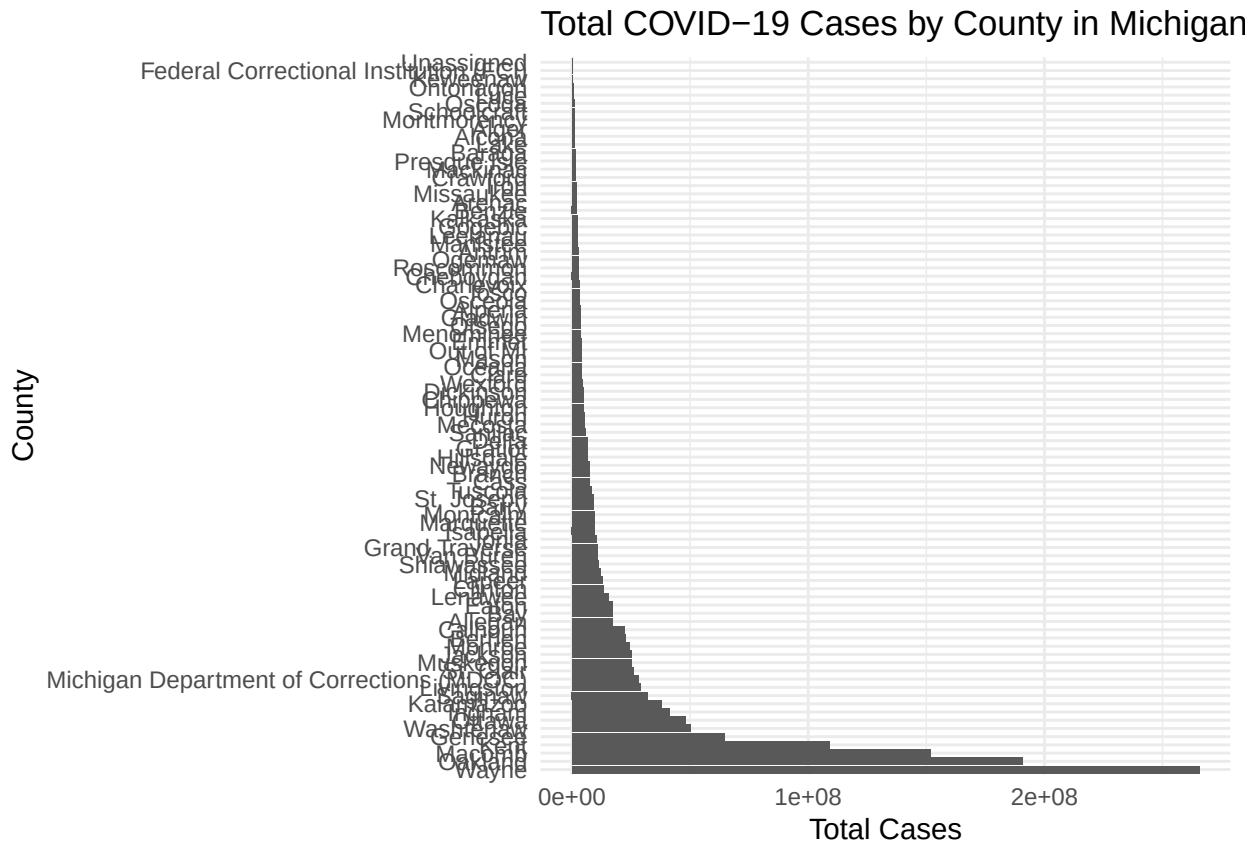
This also indicates that there is very less likely for abrupt changes happened for the long term between 2020-2023. We can also do Monthly and Daily analysis to find if there is jumps in there as further analysis.

Geographical Comparison

```
michigan_data <- covid_df %>%
  filter(Province_State == "Michigan")
```

```r
michigan_total_cases <- michigan_data %>%
  select(-c(UID, iso2, iso3, code3, Lat, Long_, Combined_Key, Country_Region)) %>%
  gather(key = "Date", value = "Cases", -c(Admin2, Province_State, FIPS)) %>%
  group_by(Admin2, FIPS) %>%
  summarise(Total_Cases = sum(Cases, na.rm = TRUE)) %>%
  rename(county = Admin2, fips = FIPS)

# Plot cases by county
ggplot(michigan_total_cases, aes(x = reorder(county, -Total_Cases), y = Total_Cases)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Total COVID-19 Cases by County in Michigan",
       x = "County",
       y = "Total Cases") +
  theme_minimal()
```



Here also it looks like we will not able to conclude there is significant discrepancies.

Conslusion

Based on the investigation, it looks like we need to do further analysis to find if there is bias in the data. If we can consider testing, sex, race and other details it may be possible to find any bias in the data.