Covid 19 Data Analysis

2025-03-04

Import Libraries

```
library(tidyverse)
library(lubridate)
library(ggplot2)
library(dplyr)
```

Introduction

Project Purpose

This is the Final Project for the course DTSA 5301: Data Science as a Field. We are demonstrating our ability to complete all steps in the data science process by creating a reproducible report on the COVID19 data set from the John Hopkins GitHub site.

Question of Interest

- How many covid Cases do we have in Illinois and what is the mortality rate for covid 19 in Illinois
- Can we predict future COVID19 cases and deaths in Illinois with a Linear Regression Model?

Project Step 1: Describe and Import the Dataset

Data Description

CSSE COVID19 Time Series Data

Two of the datasets are time series tables for the US confirmed cases and deaths, reported at the county level.

The other two datasets are for the global confirmed cases and deaths. Australia, Canada, and China are reported at the province/state level. Dependencies of the Netherlands, the UK, France and Denmark are listed under the province/state level. The US and other countries are at the country level.

Source https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

Import Datasets

```r
# All files begin with this string.
url_in <- ("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_co

# Vector containing four file names.
file_names <-
  c("time_series_covid19_confirmed_global.csv",
  "time_series_covid19_deaths_global.csv",
  "time_series_covid19_confirmed_US.csv",
  "time_series_covid19_deaths_US.csv")

# String concatenate url_in and each of the file names.
urls <- str_c(url_in, file_names)
```

```r
# Store each dataset in a variable.
global_cases <- read_csv(urls[1])
global_deaths <- read_csv(urls[2])
US_cases <- read_csv(urls[3])
US_deaths <- read_csv(urls[4])
```

Step 2: Tidy and Transform Data

We want to begin our analysis by cleaning and tidying our data. We will drop any unnecessary columns and convert our data types to relevant formats.

Tidy Global Data

```r
global_cases <- global_cases %>%
  pivot_longer(cols =
                 -c('Province/State',
                    'Country/Region', Lat, Long),
               names_to = "date",
               values_to = "cases")
```

```r
global_deaths <- global_deaths %>%
  pivot_longer(cols =
                 -c('Province/State',
                    'Country/Region', Lat, Long),
               names_to = "date",
               values_to = "deaths")
```

```r
global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(Country_Region = 'Country/Region',
         Province_State = 'Province/State') %>%
  mutate(date = mdy(date))
```

```
## Joining with ‘by = join_by(‘Province/State‘, ‘Country/Region‘, Lat, Long,
## date)‘
```

Summary of Global Data (Descriptive Statistics)

```r
summary(global)
```

```
##  Province_State     Country_Region          Lat              Long
##  Length:330327      Length:330327      Min.   :-71.950   Min.   :-178.12
##  Class :character   Class :character   1st Qu.:  3.934   1st Qu.: -42.60
##  Mode  :character   Mode  :character   Median : 21.513   Median :  20.94
##                                        Mean   : 19.719   Mean   :  22.18
##                                        3rd Qu.: 40.464   3rd Qu.:  90.36
##                                        Max.   : 71.707   Max.   : 178.06
##                                        NA's   :2286      NA's   :2286
##       date                cases              deaths
##  Min.   :2020-01-22   Min.   :        0   Min.   :      0
##  1st Qu.:2020-11-02   1st Qu.:      680   1st Qu.:      3
##  Median :2021-08-15   Median :    14429   Median :    150
##  Mean   :2021-08-15   Mean   :   959384   Mean   :  13380
##  3rd Qu.:2022-05-28   3rd Qu.:   228517   3rd Qu.:   3032
##  Max.   :2023-03-09   Max.   :103802702   Max.   :1123836
##
```

Tidy US Data

```r
US_cases <- US_cases %>%
  pivot_longer(cols = -(UID:Combined_Key),
               names_to = "date",
               values_to = "cases")  %>%
  select(Admin2:cases) %>%
  mutate(date = mdy(date)) %>%
  select (-c(Lat, Long_))
```

```r
US_deaths <- US_deaths %>%
  pivot_longer(cols = -(UID:Population),
               names_to = "date",
               values_to = "deaths")  %>%
  select(Admin2:deaths) %>%
  mutate(date = mdy(date)) %>%
  select (-c(Lat, Long_))
```

```r
US <- US_cases %>%
  full_join(US_deaths)
```

```
## Joining with `by = join_by(Admin2, Province_State, Country_Region,
## Combined_Key, date)`
```

Summary of US Data (Descriptive Statistics)

```r
summary(US)
```

```
##     Admin2          Province_State     Country_Region     Combined_Key
##  Length:3819906     Length:3819906     Length:3819906     Length:3819906
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       date                cases           Population           deaths
##  Min.   :2020-01-22   Min.   :  -3073   Min.   :       0   Min.   :  -82.0
##  1st Qu.:2020-11-02   1st Qu.:    330   1st Qu.:    9917   1st Qu.:    4.0
##  Median :2021-08-15   Median :   2272   Median :   24892   Median :   37.0
##  Mean   :2021-08-15   Mean   :  14088   Mean   :   99604   Mean   :  186.9
##  3rd Qu.:2022-05-28   3rd Qu.:   8159   3rd Qu.:   64979   3rd Qu.:  122.0
##  Max.   :2023-03-09   Max.   :3710586   Max.   :10039107   Max.   :35545.0
```

Visualize United States Cases vs Deaths

The graph below displays the cases and deaths in the United States on a logarithmic scale. It appears that the growth rates for both cases and deaths tend to stabilize and approach a limit over time.

```r
US <- US_cases %>% full_join(US_deaths)
```

```
## Joining with `by = join_by(Admin2, Province_State, Country_Region,
## Combined_Key, date)`
```

```r
US_by_state <- US %>% group_by(Province_State,Country_Region,date) %>% summarize(cases=sum(cases),death
```

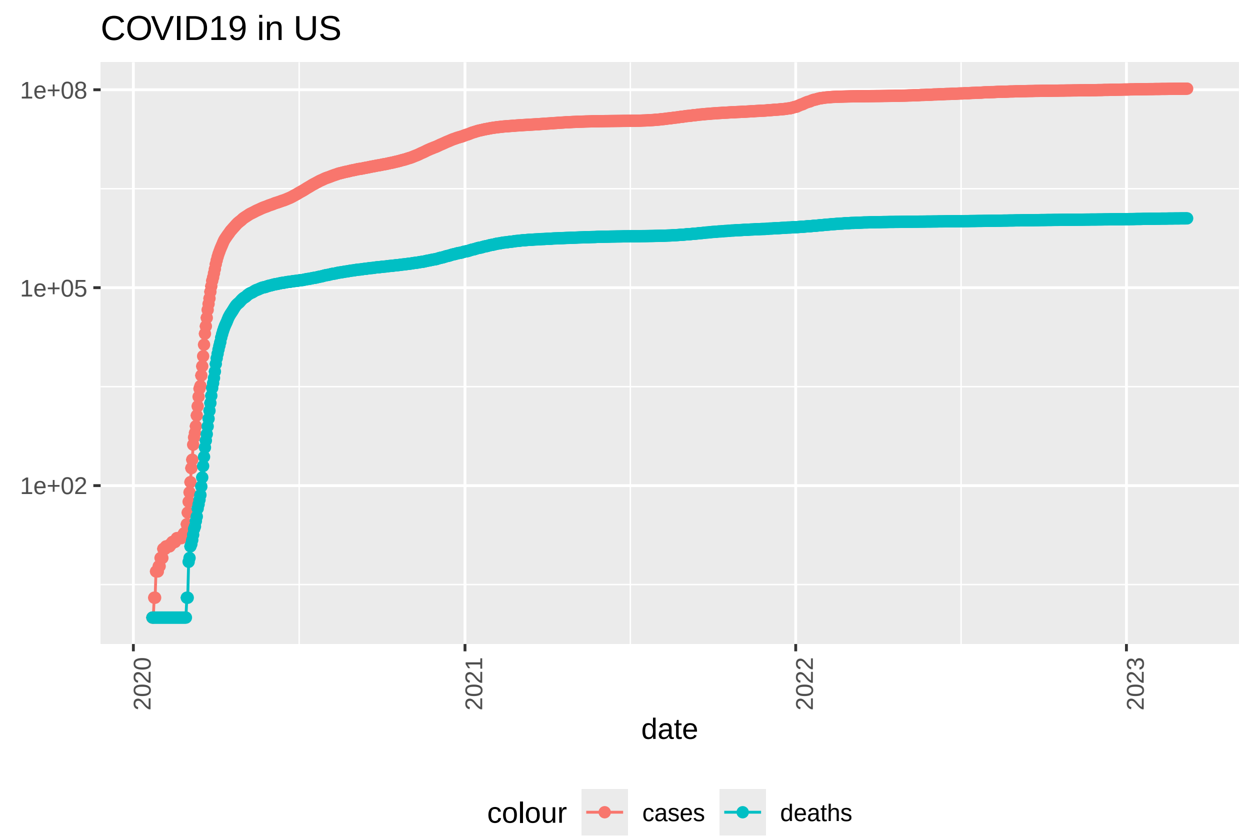
```
## `summarise()` has grouped output by 'Province_State', 'Country_Region'. You can
## override using the `.groups` argument.
```

```r
US_by_state <- US_by_state %>% mutate(new_cases = cases-lag(cases), new_deaths = deaths-lag(deaths))

US_totals <- US_by_state %>% group_by(Country_Region,date) %>% summarize(cases=sum(cases),deaths=sum(de
```

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

```r
US_totals <- US_totals %>% mutate(new_cases = cases-lag(cases), new_deaths = deaths-lag(deaths))
US_totals %>% filter(cases > 0) %>% ggplot(aes(x=date,y=cases)) + geom_line(aes(color="cases")) + geom_
```

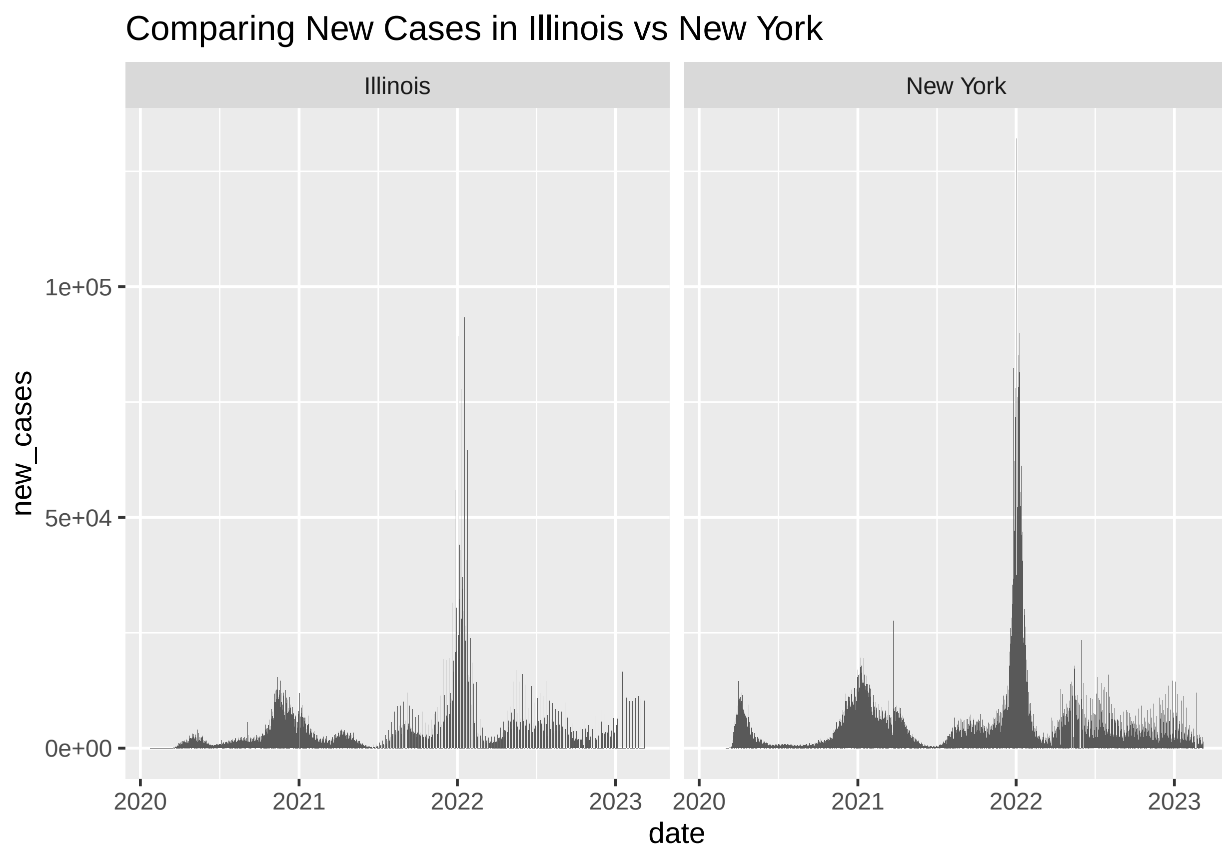


Visualize and Compare State Data: Illinois vs New York

This comparison examines the new COVID-19 cases in Illinois and New York throughout the course of the pandemic. It is clear that New York experienced a significantly higher number of new cases at the start of 2022. Additionally, it appears that Illinois stopped consistently recording data around mid-2022. Both datasets show gaps in reporting, which may be due to changes in data recording practices or a reduction in reported infections.

```
AZvsNY <- US_by_state %>% filter(Province_State=="Illinois" | Province_State=="New York") %>% filter(ca

ggplot(data=AZvsNY,aes(x=date,y=new_cases)) +
  geom_bar(stat="identity") +
  facet_wrap("Province_State") +
  labs(title="Comparing New Cases in Illinois vs New York")
```



Step 3: Analysis

I am focusing my research on Illinois so I will create four new dataframes with only Illinois data.

```
# Filter US dataset for only the rows where Province_State is Illinois.
wisc <- US %>%
  filter(Province_State == "Illinois", cases > 0) %>%
  group_by(date, Admin2)

# Group Wisconsin data by county and add mortality rate column.
wisc_counties <- wisc %>%
  group_by(Admin2, date) %>%
  mutate(mortality_rate = deaths / cases) %>%
  select(Admin2, date, cases, deaths, Population, mortality_rate)

# Sum all Illinois county cases, deaths, and populations.
wisc_totals <- wisc %>%
  group_by(date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths), Population = sum(Population)) %>%
  select(date, cases, deaths, Population) %>%
  ungroup()

# Create a dataframe that contains the most recent statistics for each Illinois county.
# Updated April 22, 2022.
current_counties <- wisc_counties %>%
```

```
  filter(date == "2022-04-22") %>%
  group_by(Admin2) %>%
  mutate(county_mortality_rate = deaths/cases) %>%
  select(date, Admin2, cases, deaths, Population, county_mortality_rate) %>%
  ungroup()
```

I will now analyze these datasets to find information relevant to my question of interest.

```
# Total Illinois cases to date.
max(wisc_totals$cases)
```

```
## [1] 4083292
```

```
# Total Illinois deaths to date.
max(wisc_totals$deaths)
```

```
## [1] 41496
```

```
# Illinois mortality rate:
max(wisc_totals$deaths) / max(wisc_totals$cases)
```

```
## [1] 0.01016239
```

We can see that Illinois had a total of 4,083,292 cases with 41,496 deaths. Illinois had a mortality rate of about 1%.

Modeling with Linear Regression

My objective is to determine whether I can predict future Illinois COVID19 cases and deaths with a Linear Regression Model.

Linear regression is a statistical method used to predict the value of a dependent variable (Y) based on an independent variable (X). The aim is to establish a linear relationship between the predictor variable (X) and the outcome variable (Y). In the case of simple linear regression, this relationship is represented as a straight line on a graph. If the exponent of the predictor variable differs from 1, the relationship becomes nonlinear, resulting in a curve rather than a straight line.

```
# Prepare the data set for modeling.
wisc_county_totals <- wisc_counties %>%
  group_by(Admin2) %>%
  summarize(deaths = sum(deaths, na.rm = TRUE),
            cases = sum(cases, na.rm = TRUE),
            Population = max(Population, na.rm = TRUE)) %>%  # Assuming population is constant across r
  mutate(cases_per_hundred = 100 * cases / Population,
         deaths_per_hundred = 100 * deaths / Population) %>%
  filter(!is.na(cases) & !is.na(deaths) & !is.na(Population) &
         cases > 0 & deaths > 0 & Population > 0) %>%  # Exclude rows with 0 or NA values
  select(Admin2, cases, deaths, Population, cases_per_hundred, deaths_per_hundred)

# Build the linear regression model.
lr_model <- lm(deaths_per_hundred ~ cases_per_hundred, data = wisc_county_totals)
```

```r
# Display summmary for model analysis.
summary(lr_model)
```

```
##
## Call:
## lm(formula = deaths_per_hundred ~ cases_per_hundred, data = wisc_county_totals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -177.948  -43.805   -1.393   35.684  248.222
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.003660  52.857087   0.151     0.88
## cases_per_hundred 0.011636   0.002806   4.147 7.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.61 on 100 degrees of freedom
## Multiple R-squared:  0.1468, Adjusted R-squared:  0.1382
## F-statistic:  17.2 on 1 and 100 DF,  p-value: 7.064e-05
```

The model and its coefficients are statistically significant, with p-values well below the 0.05 threshold, suggesting that the model is highly effective in predicting COVID-19-related deaths. This indicates that the model can be a valuable tool for governments to forecast potential fatalities based on the number of confirmed cases during an outbreak. By leveraging this model, policymakers can better plan for and respond to COVID-19 outbreaks, ensuring more informed decision-making and resource allocation.

Step 4: Report Conclusion and Sources of Bias

Conclusion

I found that the total number of cases is a strong predictor of the total deaths per 100 individuals, making it a valuable tool for policymakers. Additionally, the total number of COVID-19 cases in Illinois surpassed 4 million, which, while significant, remains considerably lower than that of New York. Illinois also had a COVID-19 mortality rate of approximately 1%.

Sources of Bias

COVID-19 has become a highly politicized issue, and strong opinions on this topic can introduce bias into analysis. To mitigate this risk, I remained objective and avoided making assumptions, focusing solely on the data rather than the political context surrounding the pandemic. Additionally, bias can arise from how data is collected. However, this particular dataset comes with comprehensive documentation detailing its collection process and the organizations involved, which increases its credibility. While there may be some inconsistencies in how COVID-19 cases were reported, such challenges are common in any data related to infectious diseases. It is important to recognize this as an inherent issue and work with the available data to the best of our ability.