Project Proposal

US Covid 19 cases tracking, management and analysis

Annesha De* Anwaya Wadnerkar† Namrata Galigari‡
Vidhyashree Murugesapandian§ Carlos Mahecha Parra¶

March. 20. 2024

```
library(tidyverse)
library(modelr)

# Load data
df <- read_csv("C:/Users/clmah/cmu_datasci_cool_team_repo/time_series_covid19_confirmed_US.csv")
#
#
#
#
# Prepare assigned dataframe for future plots
washington_data <- df %>%
  filter(Province_State == "Washington")
```

About this dataset

This data set was obtained from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The data sources used to build this data set include the World Health Organization (WHO), the European Centre for Disease Prevention and Control (ECDC), different US data sources per the state (eg. the Colorado Department of Public Health and Environment).

We'll be working with one of the daily time series summary data sets regarding confirmed cases in the US ("time_series_covid19_confirmed_US.csv"). This data is a compilation of daily case reports, also available in the given repository. This is the data structure of the chosen data set:

```
## Data structure
#UID            : num [1:3342] Registry ID
#iso2           : chr [1:3342] US (country code ISO2)
#iso3           : chr [1:3342] USA (country code ISO2)
#code3          : num [1:3342] USA country code
#FIPS           : num [1:3342] 1001 1003 1005 1007 1009 ...
#Admin2         : chr [1:3342] US county, like "Autauga" "Baldwin" ...
#Province_State: chr [1:3342] US State like "Alabama" "Alaska" "Arizona"
```

*anneshad@andrew.cmu.edu
†awadnerk@andrew.cmu.edu
‡ngaligar@andrew.cmu.edu
§vmuruges@andrew.cmu.edu
¶cmahecha@andrew.cmu.edu

```
#Country_Region: chr [1:3342] "US"
#Lat            : num [1:3342] 32.5 30.7 31.9 33 34 ... ()
#Long_          : num [1:3342] -86.6 -87.7 -85.4 -87.1 -86.6 ...
#Combined_Key   : chr [1:3342] US State and county: "Autauga, Alabama, US"...
#MM/DD/YY cols : num [1:3342] Number of cases per day between 2020-2023
#                 each column registers the number of confirmed cases per day
```

The data set is available as "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University" or "JHU CSSE COVID-19 Data" for short, and the url: https://github.com/CSSEGISandData/COVID-19.

Research questions

When the covid 19 outbreak occurred in the United States, the proliferation of the disease advanced to a large extent, and some states recorded high levels of new COVID-19 hospital admissions in the week. Given the selected data set which provides daily confirmed cases per county/state, it is possible to ask:

1. Which counties and states have the highest number of confirmed cases over time,
2. Is there a relationship between the transmission of the disease regarding proximity to these possible hotspots.

The exploration of the data set and the previous questions brought us to the following hypothesis:

- The geographic location, specifically latitude and longitude, is associated with the rate of COVID-19 spread in the United States and some states and counties can be identified as transmission hotspots

Given this hypothesis we want to: 1. Explore the distribution of confirmed cases over time, examining trends and patterns. 2. Visualize the spatial distribution of confirmed cases using latitude and longitude information. 3. Examine correlations between geographic location (latitude and longitude) and the number of confirmed cases.

Modeling:

Since we have daily counts of confirmed COVID-19 cases over time, a time series model would be more appropriate for this data set rather than a traditional linear regression model. The autoregressive integrated moving average (ARIMA) model is considered.

However, an additional linear regression model is also considered to make predictions regarding the number of confirmed cases based on counties latitude and longitude.

First proposed model:

We propose the following preliminary models fitted for a particular date. The first one uses as predictors (x) the variables State, County and Date, and as a dependent variable (y), the number of cases corresponding to that day

```r
# Preprocess the data
## The idea is to identify all columns named with a date and their values
date_columns <- grep("^\\d{2}/\\d{2}/\\d{2}$", names(df), value = TRUE)
df[date_columns] <- lapply(df[date_columns], as.Date, format = "%m/%d/%Y")

# Then we convert the data to long format
## we use pivot_longer to lengthen data, increasing the number of rows and
## decreasing the number of columns, starting with the first column/date
## this model is fitted for a particular date.
long_data <- df %>%
  pivot_longer(cols = starts_with("1/22/23"), names_to = "Date", values_to = "Cases") %>%
  mutate(Date = as.Date(Date, format = "%m/%d/%Y"))

# Fit the model
## Province_state: US states
## Admin2: Counties
model <- lm(Cases ~ Date + Province_State + Admin2, data = long_data)

# Store the summary result. Only part of the output is printed
summary_result <- summary(model)
summary_output <- capture.output(summary_result)
cat(head(summary_output, 20), sep = "\n")
```

```
##
## Call:
## lm(formula = Cases ~ Date + Province_State + Admin2, data = long_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -612960   -3436       0       0  979720
##
## Coefficients: (1 not defined because of singularities)
##                                                Estimate Std. Error t value
## (Intercept)                                     4278.21   72822.05   0.059
## Date                                                 NA         NA      NA
## Province_StateAlaska                           -10629.38   51248.28  -0.207
## Province_StateArizona                           10900.92   45007.79   0.242
## Province_StateArkansas                          -9091.83   15158.41  -0.600
## Province_StateCalifornia                        38463.48   29860.10   1.288
## Province_StateColorado                           1132.20   19099.19   0.059
## Province_StateConnecticut                       18519.09   43244.52   0.428
## Province_StateDelaware                          27242.64   46555.06   0.585
## Province_StateDistrict of Columbia              -1927.25   71713.08  -0.027
```

US_State coefficients: The coefficients for each US_State variable represent the estimated change in the response variable for a one-unit increase in that state, compared to the reference state (usually the first state listed alphabetically or numerically, depending on how the data is encoded). A positive coefficient indicates that the number of confirmed cases tends to be higher in that state compared to the reference state, while a negative coefficient indicates the opposite. The magnitude of the coefficient indicates the strength of the effect.

US_county coefficients: Similarly, the coefficients for each UD_county variable represent the estimated change in the response variable for a one-unit increase in that county, compared to the reference county.

Again, positive coefficients indicate higher confirmed cases in that county compared to the reference county, while negative coefficients indicate the opposite.

Second proposed model:

The second proposed model uses as predictors (x) the variables Latitude, Londitude and Date, and as a dependent variable (y), the number of cases corresponding to that particular day

```
# Preprocess the data
date_columns <- grep("^\\d{2}/\\d{2}/\\d{2}$", names(df), value = TRUE)
df[date_columns] <- lapply(df[date_columns], as.Date, format = "%m/%d/%Y")

long_data <- df %>%
  pivot_longer(cols = starts_with("3/9/23"), names_to = "Dates", values_to = "All_cases") %>%
  mutate(Date = as.Date(Dates, format = "%m/%d/%Y"))

# Fit the model
model_coords <- lm(All_cases ~ Date + df$Lat + df$Long_, data = long_data)

# Store the summary result. Only part of the output is printed
summary(model_coords)
```

```
##
## Call:
## lm(formula = All_cases ~ Date + df$Lat + df$Long_, data = long_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -73207  -28256  -22678   -9983 3669403
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23113.9     8410.7   2.748  0.00603 **
## Date             NA         NA      NA       NA
## df$Lat        -509.3      298.7  -1.705  0.08829 .
## df$Long_      -300.6      124.5  -2.414  0.01584 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 108400 on 3339 degrees of freedom
## Multiple R-squared:  0.001743,   Adjusted R-squared:  0.001145
## F-statistic: 2.915 on 2 and 3339 DF,  p-value: 0.05434
```

Latitude coefficient: The coefficient for the Latitude variable represents the estimated change in the response variable for a one-unit increase in latitude, holding all other variables constant. Since latitude represents north-south position, a positive coefficient suggests that moving north tends to be associated with higher confirmed cases, while a negative coefficient suggests the opposite.

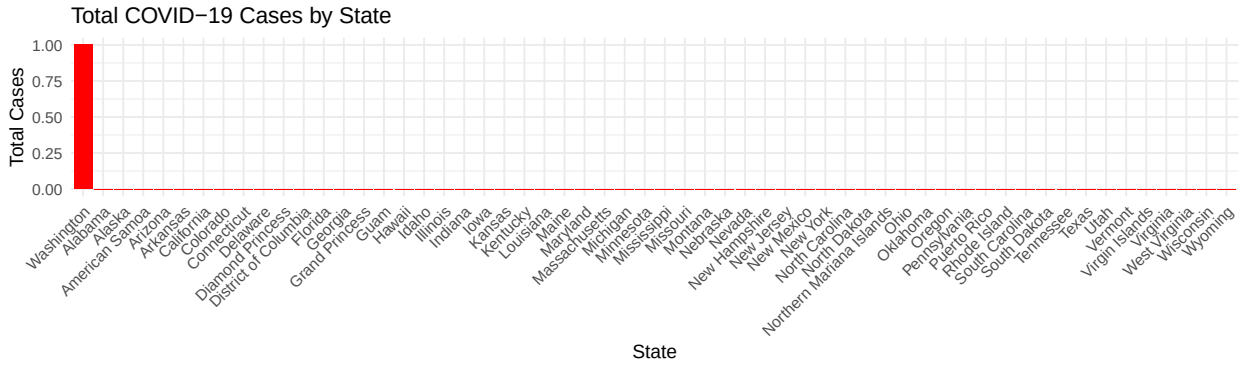
Longitude coefficient: Similarly, the coefficient for the Longitude variable represents the estimated change in the response variable for a one-unit increase in longitude, holding all other variables constant. Since longitude represents east-west position, a positive coefficient suggests that moving east tends to be associated with higher confirmed cases, while a negative coefficient suggests the opposite.

Plots and data exploration

Plot 1: This ggplot visualizes the total number of COVID-19 cases for each state, providing insights into the distribution of cases across different states.

```
# Summarize total confirmed cases by state
total_cases_by_state <- df %>%
  group_by(Province_State) %>%
  summarise(Total_Cases = sum(`1/23/20`, na.rm = TRUE)) %>%
  arrange(desc(Total_Cases))

# Plot the bar plot of total confirmed cases for each state
ggplot(total_cases_by_state, aes(x = reorder(Province_State, desc(Total_Cases)),
                                 y = Total_Cases)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(title = "Total COVID-19 Cases by State",
       x = "State",
       y = "Total Cases") +
  theme_minimal() +
  # Rotate state names for better readability
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The bar for Washington is noticeably high while the bars for other states are negligible, it suggests that Washington had a significantly higher number of COVID-19 cases compared to the other top 10 states on the selected date (in this case, January 23, 2020).

This could be due to various factors such as:

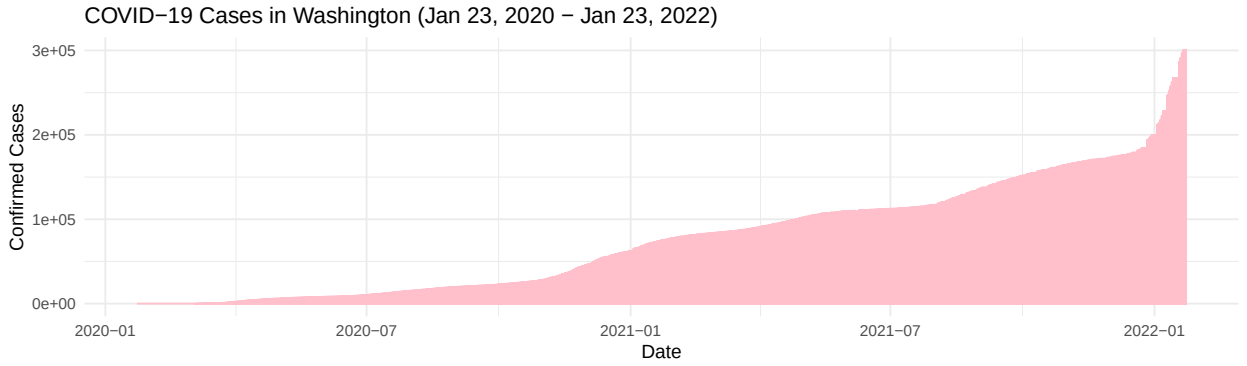
1. Early outbreak: Washington might have experienced an earlier and more significant outbreak compared to other states, resulting in a higher number of cases by January 23, 2020.
2. Population density: Washington might have a higher population density or other demographic factors that contributed to the rapid spread of the virus in the early stages of the pandemic.
3. Testing availability: Differences in testing availability and testing strategies between states could have influenced the number of reported cases.
4. Reporting discrepancies: Variations in reporting standards and practices between states might have affected the accuracy and consistency of case reporting.
5. Public health measures: Differences in public health interventions, policies, and compliance with preventive measures could have impacted the spread of the virus in each state.

Plot 2: This plot visualizes the trend of COVID-19 cases in Washington from January 23, 2020, to January 23, 2022.

```
# Select columns for the specified date range
washington_data <- washington_data %>%
  select(starts_with("1/23/20"):starts_with("1/23/22"))

# Convert data to long format
washington_data_long <- washington_data %>%
  pivot_longer(cols = everything(), names_to = "Date_", values_to = "Cases_") %>%
  mutate(Date_ = mdy(gsub("X", "", Date_)))

# Plot the time series of confirmed cases in Washington
ggplot(washington_data_long, aes(x = Date_, y = Cases_)) +
  geom_line(color = "pink") +
  labs(title = "COVID-19 Cases in Washington (Jan 23, 2020 - Jan 23, 2022)",
       x = "Date",
       y = "Confirmed Cases") +
  theme_minimal()
```



While similar trend is followed for other states, this graph shows the exponential surge of active cases in Washington (with the highest number of cases as seen in the previous plot) over a span of 2 years. It is observed that there was a significant surge in the active cases in the span of these years with a consistently upward graph.

Additional questions

Using the COVID-19 dataset, there are numerous questions that can be explored using plots and modeling techniques.

Trend Analysis:

1. What is the overall trend of COVID-19 cases over time globally, nationally, or within specific regions/states?
2. How do the trends vary across different countries, states, or regions?

Spatial Analysis:

1. Where are the COVID-19 hotspots geographically located?
2. How does the distribution of COVID-19 cases vary spatially across different regions, states, or countries?

Demographic Analysis:

1. Are there any demographic factors (age, gender, ethnicity) associated with higher rates of COVID-19 transmission or mortality?
2. How does the impact of COVID-19 vary across different demographic groups?

Public Health Interventions:

1. How effective are public health interventions (e.g., lockdowns, mask mandates, vaccination campaigns) in reducing the spread of COVID-19?
2. What strategies can be implemented to mitigate the impact of COVID-19 and prevent future outbreaks?

Comparative Analysis:

1. What can we learn from comparing the responses to COVID-19 across different countries or regions?

To provide significant insights and conclusions, additional datasets can be integrated with the COVID-19 dataset. Some potential datasets include:

Google COVID-19 Community Mobility Reports:

Dataset: Google COVID-19 Community Mobility Reports

Questions:

1. How does human mobility (e.g., visits to workplaces, retail, parks) correlate with COVID-19 transmission rates?
2. What is the impact of mobility restrictions and lockdown measures on reducing COVID-19 spread?

Centers for Disease Control and Prevention (CDC) COVID-19 Vaccination Data:

Dataset: Centers for Disease Control and Prevention (CDC) COVID-19 Vaccination Data

Questions:

1. What is the impact of COVID-19 vaccination campaigns on reducing infection rates, hospitalizations, and mortality?
2. Are there disparities in COVID-19 vaccination coverage based on demographic factors, socioeconomic status, or geographic location?

United Nations World Population Prospects:

Dataset: United Nations World Population Prospects

Questions:

1. How does the age distribution of the population affect COVID-19 transmission and mortality rates?
2. Are there disparities in COVID-19 outcomes based on population density or urban-rural divide?

Our team repository

You can find the .rmd file, the pdf document and the data set considered for this project in the following repository: https://github.com/cmahechacmu/cmu_datasci_cool_team_repo.git

References

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. The Lancet infectious diseases, 20(5), 533-534.