COVID19 in the United States Report

Jacky Luo

8/3/2023

Introduction

This report will perform analysis to determine the rate of cases of and deaths due to COVID19 in the United States by county. The analysis will determine if there is any relationship between county population and county location on COVID19.

Import Packages and Data

The packages used for this document are listed below, please install any missing packages.

- `tidyverse`
- `lubridate`
- `ggplot2`
- `forecast`

The data used for this analysis comes from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University and can be found on their **github page**.

The specific files used will be the `UID_ISO_FIPS_LookUp_Table.csv` for lookup info, as well as the `time_series_covid19_confirmed_US.csv` and `time_series_covid19_deaths_US.csv` time series datasets.

```
#Import libraries, remove import outputs
verify_package <- function(package_name) {
    if (!eval(parse(text=paste("suppressPackageStartupMessages(require(",package_name,"))")))) {
        cat(package_name, " not detected, installing ", package_name, ".")
        install.packages(package_name, repos=mirror)
        library(package_name)
    }
}

packages_list = list("tidyverse", "lubridate", "ggplot2", "forecast")
for(package in packages_list){
    verify_package(package)
}

#Load data
base_url <- 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c
filenames <- c("time_series_covid19_confirmed_US.csv", "time_series_covid19_deaths_US.csv")
```

```
urls <- str_c(base_url,filenames)
us_cases <- read_csv(urls[1], show_col_types = FALSE)
us_deaths <- read_csv(urls[2], show_col_types = FALSE)

uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/U
uid <- read_csv(uid_lookup_url, show_col_types = FALSE) %>% select(-c(Lat, Long_, Combined_Key, code3, i
```

Clean and Preprocess Data

One of the first steps will be to tidy and reformat data to be of an appropriate format to be used for future analysis. In this step, steps will be performed similar to the processing steps from the class example.

The data we will use for this analysis will contain only COVID19 data from the United States.

First, the data will be pivoted to create a time series format where the dates are a column rather than an individual column for each date. The next step is to change the datatype of the `date` column to date rather than chr. Finally, we will select the relevant columns for the cases and deaths time series.

For analysis, the cases and deaths time series will be joined to create a multivariate time series by county.

The final step will be to rename the columns to keep some consistency. A preview of the data can be seen below.

```
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_, Country_Region, Combined_Key))

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_, Country_Region, Combined_Key))
```

```
## Warning in x:y: numerical expression has 1154 elements: only the first used
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `date = mdy(date)`.
## Caused by warning:
## !  3342 failed to parse.
```

```
us <- us_cases %>% full_join(us_deaths)
```

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

```
us <- us %>%
    rename(County=Admin2,State=Province_State,Date=date,Cases=cases,Deaths=deaths) %>%
    select(c(State, County, Population, Date, Cases, Deaths))

us %>% print()
```

```
## # A tibble: 3,823,248 x 6
##    State   County  Population Date       Cases Deaths
##    <chr>   <chr>        <dbl> <date>     <dbl>  <dbl>
##  1 Alabama Autauga      55869 2020-01-22     0      0
##  2 Alabama Autauga      55869 2020-01-23     0      0
##  3 Alabama Autauga      55869 2020-01-24     0      0
##  4 Alabama Autauga      55869 2020-01-25     0      0
##  5 Alabama Autauga      55869 2020-01-26     0      0
##  6 Alabama Autauga      55869 2020-01-27     0      0
##  7 Alabama Autauga      55869 2020-01-28     0      0
##  8 Alabama Autauga      55869 2020-01-29     0      0
##  9 Alabama Autauga      55869 2020-01-30     0      0
## 10 Alabama Autauga      55869 2020-01-31     0      0
## # i 3,823,238 more rows
```

Analysis and Plotting

Analysis of County Populations vs Infection and Deaths

First, as the US is a large country with over 3000 counties, 1919 of which are included in the data. The investigation into the data begins with determining whether the population of a county is related to the number of cases or deaths in the county. The top five most populous counties in the data will be plotted against the five least densely populated counties. These counties can be seen below.

```
county_populations <- us %>%
    group_by(County) %>%
    summarize(Population=mean(Population)) %>%
    filter(Population > 0) %>%
    arrange(desc(Population))

top_bot_counties <- county_populations %>%
    filter(County %in% c(county_populations$County[1:5], rev(county_populations$County)[1:5]))

top_bot_counties %>% print()
```

```
## # A tibble: 10 x 2
##    County      Population
##    <chr>            <dbl>
##  1 Los Angeles   10039107
##  2 Maricopa       4485414
##  3 San Diego      3338330
##  4 Miami-Dade     2716940
##  5 Riverside      2470546
##  6 Petroleum          487
##  7 Arthur             463
##  8 Kenedy             404
##  9 Loving             169
## 10 Kalawao             86
```

The metrics that will be used are deaths per million and cases per million.

A sample of the data from these counties below:

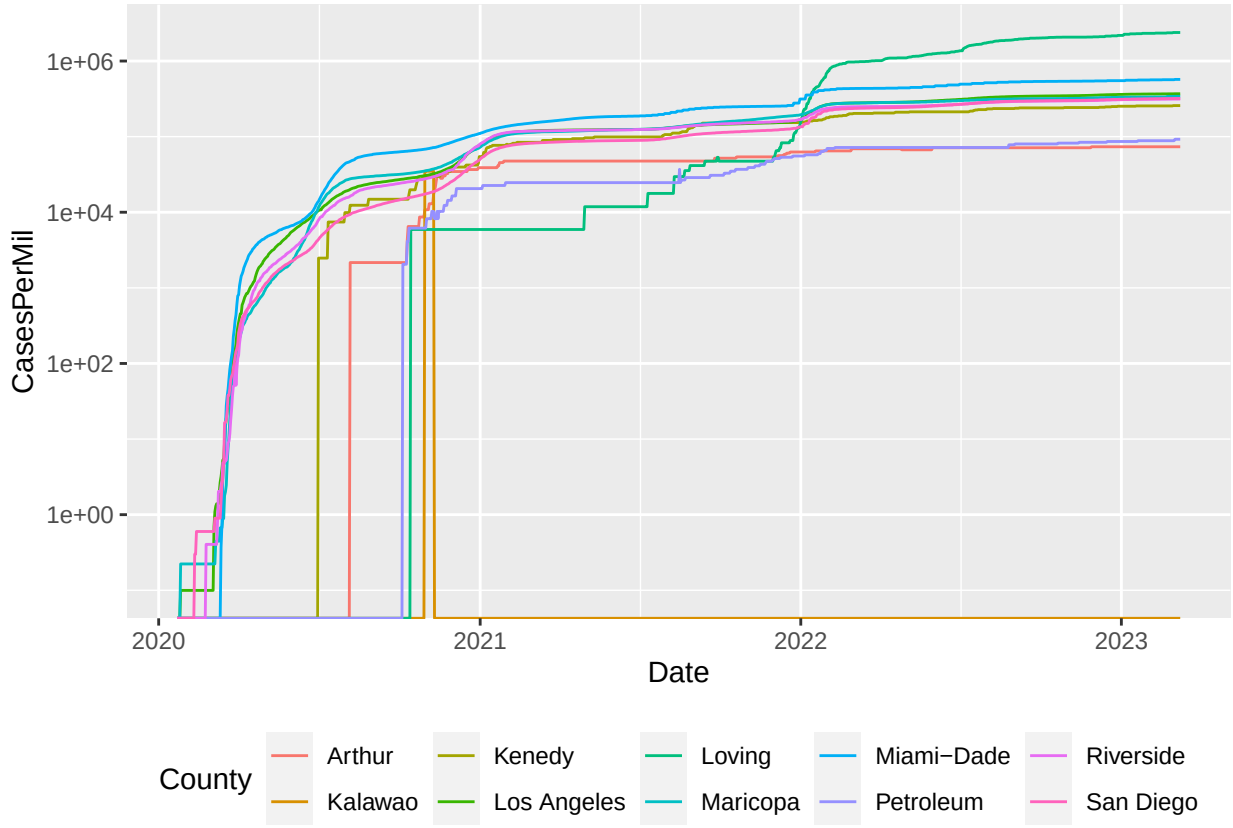
```
us_by_county <- us %>%
    filter(County %in% c(county_populations$County[1:5], rev(county_populations$County)[1:5])) %>%
    mutate(DeathsPerMil=Deaths/Population*1000000,CasesPerMil=Cases/Population*1000000)
us_by_county %>% print()
```

```
## # A tibble: 11,440 x 8
##    State   County   Population Date       Cases Deaths DeathsPerMil CasesPerMil
##    <chr>   <chr>         <dbl> <date>     <dbl>  <dbl>        <dbl>       <dbl>
##  1 Arizona Maricopa    4485414 2020-01-22     0      0            0       0
##  2 Arizona Maricopa    4485414 2020-01-23     0      0            0       0
##  3 Arizona Maricopa    4485414 2020-01-24     0      0            0       0
##  4 Arizona Maricopa    4485414 2020-01-25     0      0            0       0
##  5 Arizona Maricopa    4485414 2020-01-26     1      0            0       0.223
##  6 Arizona Maricopa    4485414 2020-01-27     1      0            0       0.223
##  7 Arizona Maricopa    4485414 2020-01-28     1      0            0       0.223
##  8 Arizona Maricopa    4485414 2020-01-29     1      0            0       0.223
##  9 Arizona Maricopa    4485414 2020-01-30     1      0            0       0.223
## 10 Arizona Maricopa    4485414 2020-01-31     1      0            0       0.223
## # i 11,430 more rows
```

The first plot will be of cases per million by date, separated by county.

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 10 rows containing missing values (`geom_line()`).
```



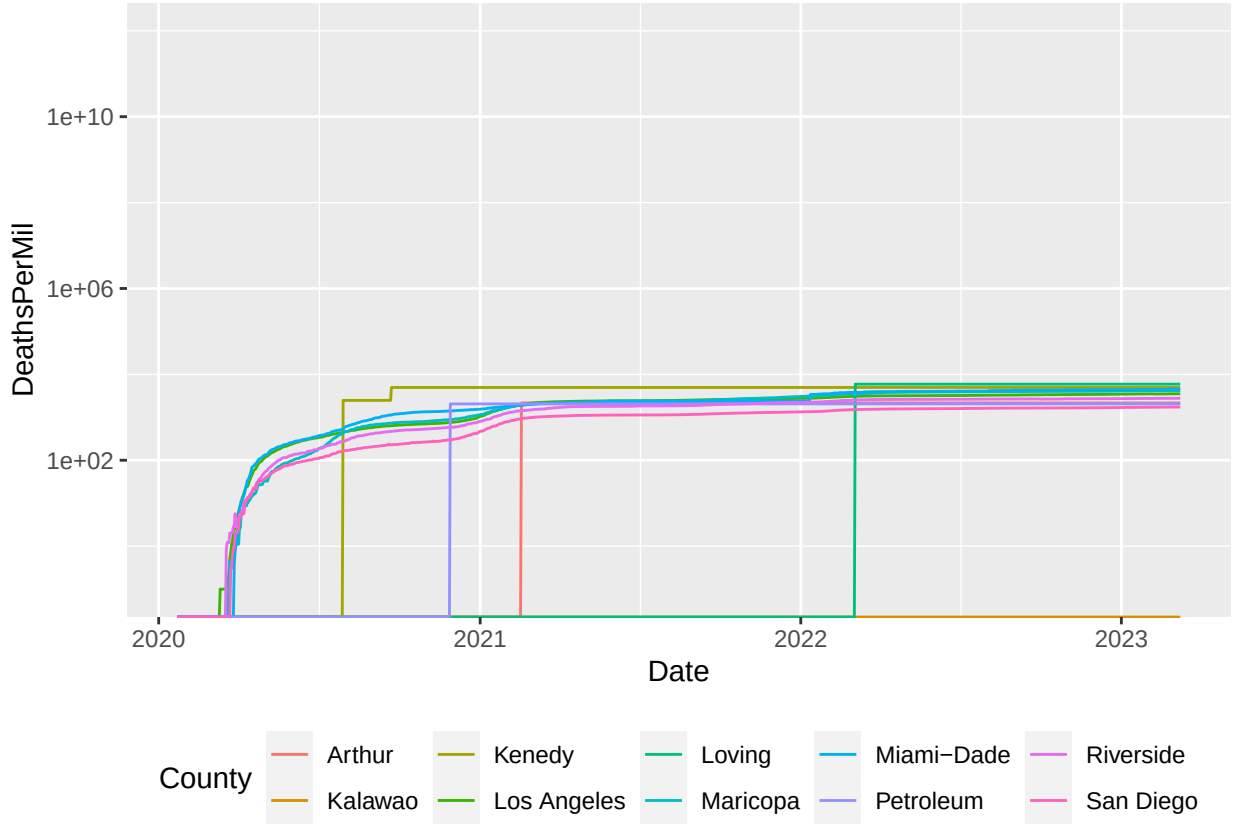
The first major observation is that Kalawao County had very few or no cases. As the population is only 86 and located fairly remotely in Hawaii, this could make sense. However, there is likely some erroneous data as there appears to be a spike in late 2020 which remains for a few days then returns to close to 0.

The second observation of note from this plot is that the less densely populated counties appear to be slower to get their first cases by months compared to the large, heavily populated counties. The smaller counties also appear to have sharper, rapid climbs in cases and longer periods of no activity as individual reported cases in these smaller counties will have significantly more impact.

Besides Loving County, Texas, all nine other counties appear to have similar number of cases per million as they are all within an order of magnitude when the data ends in early 2023. Loving County appears to be an outlier as a smaller county as it has approximately 9 months until the first case is reported, but ends with the highest cases per million.

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 10 rows containing missing values (`geom_line()`).
```



Similar to the cases plot, Kalawao County, Hawaii appears to have no interaction with COVID19.

Also similar to the cases plot, the smaller counties have a delay before reporting any deaths but then reach a steady state where they closely match the death rate of larger counties.

Analysis of County Geography vs Infection and Deaths

The next plots will attempt to determine whether regionality of the United States has an impact on COVID19 transmission and deaths. For this, two cities on the east coast are selected along with two cities on the west

coast, and one from the center of the continental United States. It was also decided for each group of two cities to select one from northern city and one southern city.

Large, metropolitan cities were selected as they reduce the variance and outliers seen in the smaller counties and give more data points to work with. In some cases, smaller counties than the largest county of a metropolitan area were selected due to an unknown bug causing an error in the plots.

The cities selected are as follows:

1. Philadelphia, PA (Philadelphia County)
2. Miami, FL (Miami-Dade County)
3. Minneapolis, MN (Hennepin County)
4. Fort Worth, TX (Tarrant County)
5. Seattle, WA (Snohomish County)
6. Los Angeles, CA (Los Angeles County)

A preview of the data can be seen below:

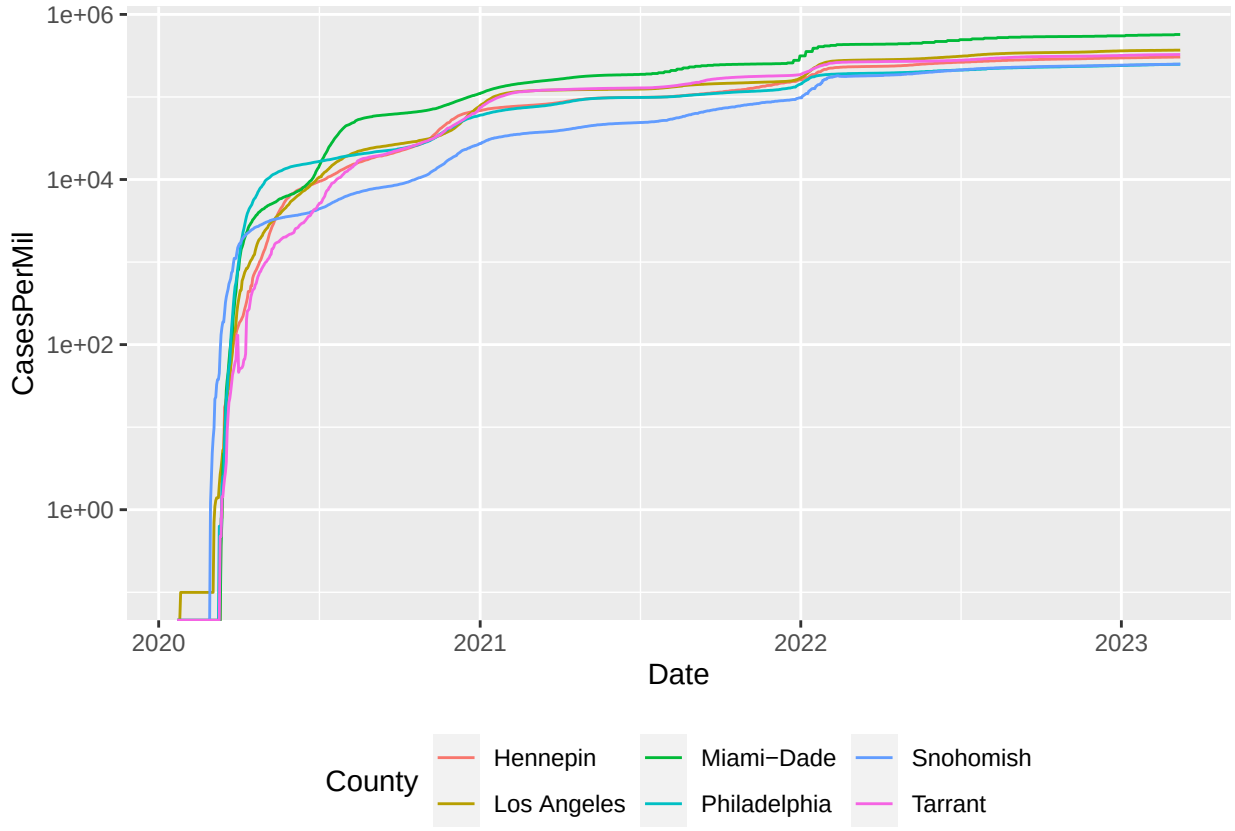
```
geo_counties <- c("Philadelphia", "Miami-Dade", "Hennepin", "Tarrant", "Snohomish", "Los Angeles")

us_by_loc <- us %>%
    filter(County %in% geo_counties) %>%
    mutate(DeathsPerMil=Deaths/Population*1000000,CasesPerMil=Cases/Population*1000000)
us_by_loc %>% print()
```

```
## # A tibble: 6,864 x 8
##    State      County Population Date       Cases Deaths DeathsPerMil CasesPerMil
##    <chr>      <chr>       <dbl> <date>     <dbl>  <dbl>        <dbl>       <dbl>
##  1 California Los A~   10039107 2020-01-22     0      0            0      0
##  2 California Los A~   10039107 2020-01-23     0      0            0      0
##  3 California Los A~   10039107 2020-01-24     0      0            0      0
##  4 California Los A~   10039107 2020-01-25     0      0            0      0
##  5 California Los A~   10039107 2020-01-26     1      0            0      0.0996
##  6 California Los A~   10039107 2020-01-27     1      0            0      0.0996
##  7 California Los A~   10039107 2020-01-28     1      0            0      0.0996
##  8 California Los A~   10039107 2020-01-29     1      0            0      0.0996
##  9 California Los A~   10039107 2020-01-30     1      0            0      0.0996
## 10 California Los A~   10039107 2020-01-31     1      0            0      0.0996
## # i 6,854 more rows
```

The first plot will be of cases per million by date, separated by county.

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 6 rows containing missing values (‘geom_line()‘).
```



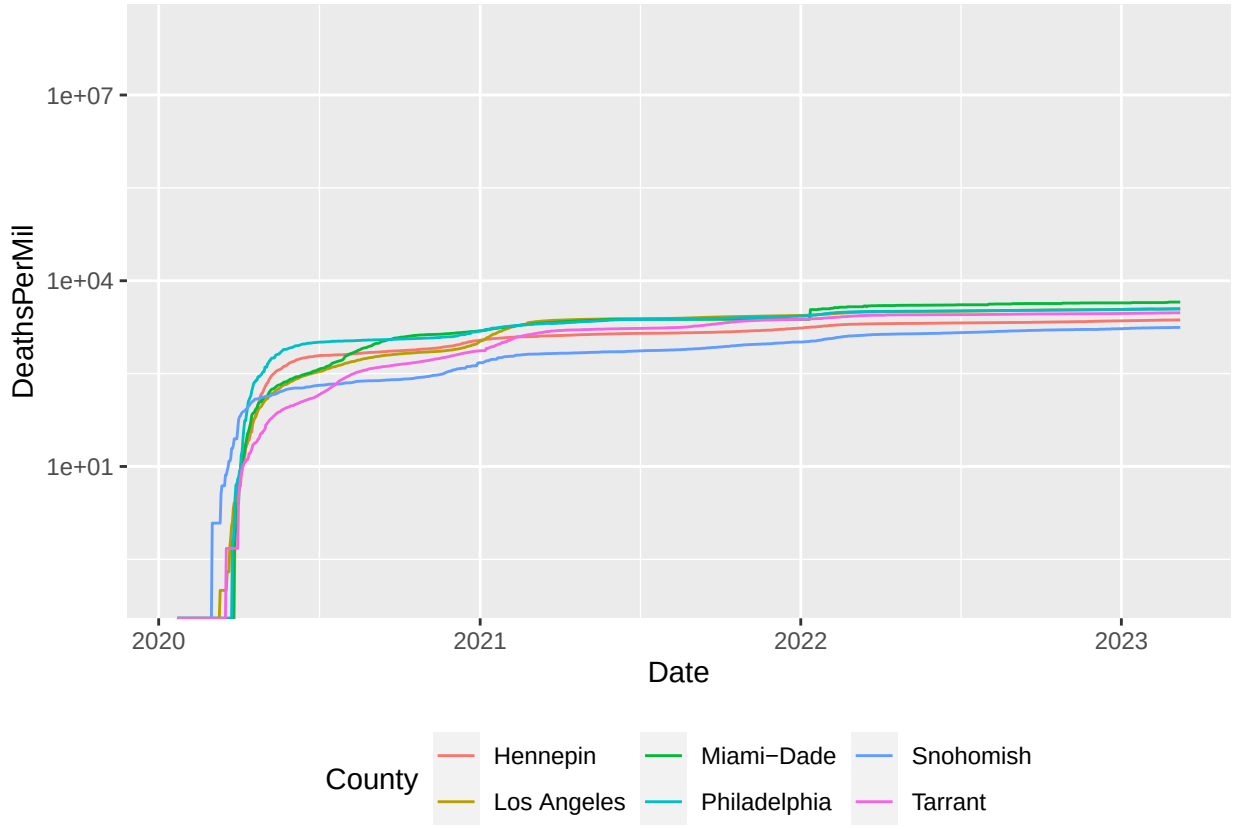
The plot appears to show another anomaly as the data for Tarrant County, Texas shows the number cases decreasing in the early months of 2020 before returning to normal.

Miami-Dade appears to be the county with the highest rate of transmission while Snohomish County appears to be have the lowest. This could be due to selection of county as Snohomish County is not the most populous county for Seattle as King County, WA caused an error in the plot.

There does not appear to be any relation between regionality and COVID19 transmission.

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 6 rows containing missing values (‘geom_line()‘).
```



The spike and then dip in deaths for Tarrant County does not appear for deaths as it did with cases, which further supports erroneous data.

The same counties with the highest and lowest rates of cases per million are the same counties for deaths per million. However, the large early spike for Los Angeles county does not appear on the deaths plot as it did on the cases plot.

Other than this, there is no distinct difference in deaths per million by region.

Time Series Modelling and Forecasting

As the data is time series data, univariate time series forecasting will be performed on the total number of cases across the United States. This begins by grouping all counties by date to form a daily time series.

```
tsdata <- us %>%
    group_by(Date) %>%
    summarize(Cases=sum(Cases)) %>%
    filter(!is.na(Cases))

tsdata %>% print()
```

```
## # A tibble: 1,143 x 2
##    Date       Cases
##    <date>     <dbl>
##  1 2020-01-22     1
```

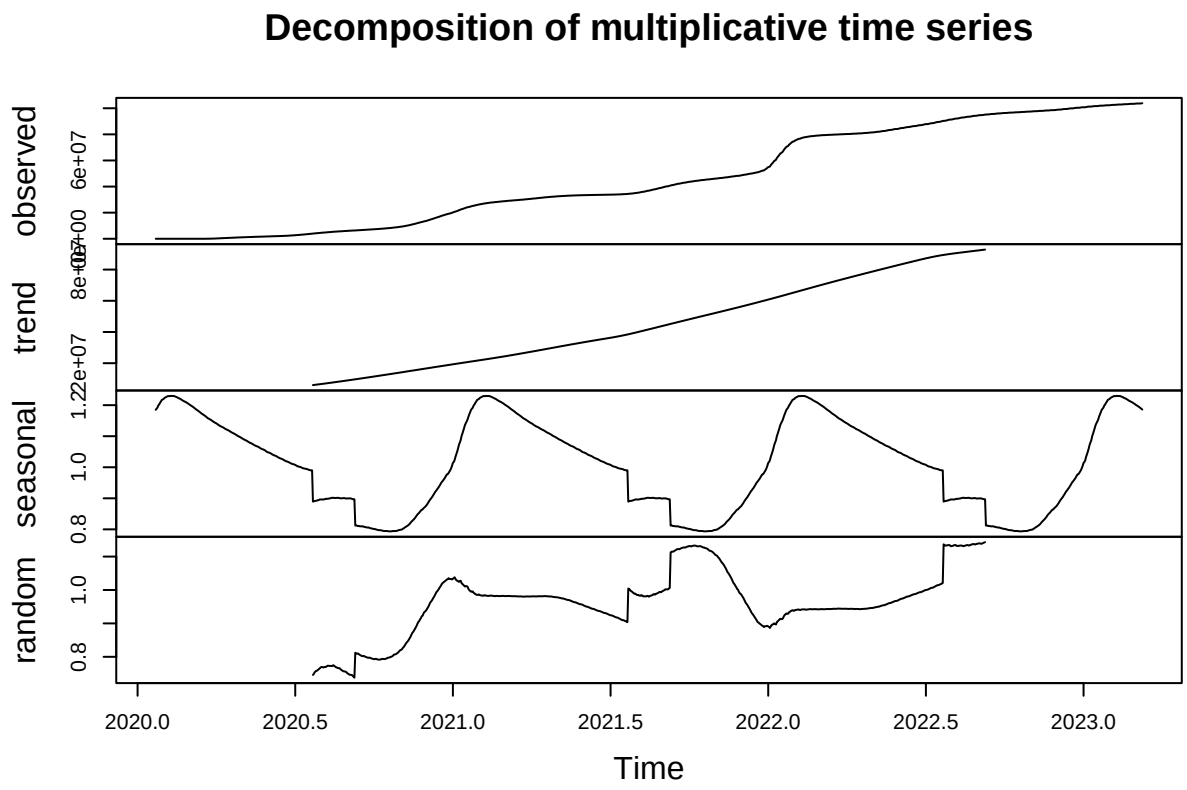
```
##  2 2020-01-23     1
##  3 2020-01-24     2
##  4 2020-01-25     2
##  5 2020-01-26     5
##  6 2020-01-27     5
##  7 2020-01-28     5
##  8 2020-01-29     6
##  9 2020-01-30     6
## 10 2020-01-31     8
## # i 1,133 more rows
```

Next, our data will be converted into a time series object to be used by the model.

```
covid_ts <- ts(tsdata$Cases, frequency=365, start=c(2020,22))
head(covid_ts, 10) %>% print()
```

```
## Time Series:
## Start = c(2020, 22)
## End = c(2020, 31)
## Frequency = 365
##  [1] 1 1 2 2 5 5 5 6 6 8
```

Next, the time series will be decomposed as analysis to view the trend, seasonalities and residuals of the multiplicative time series.

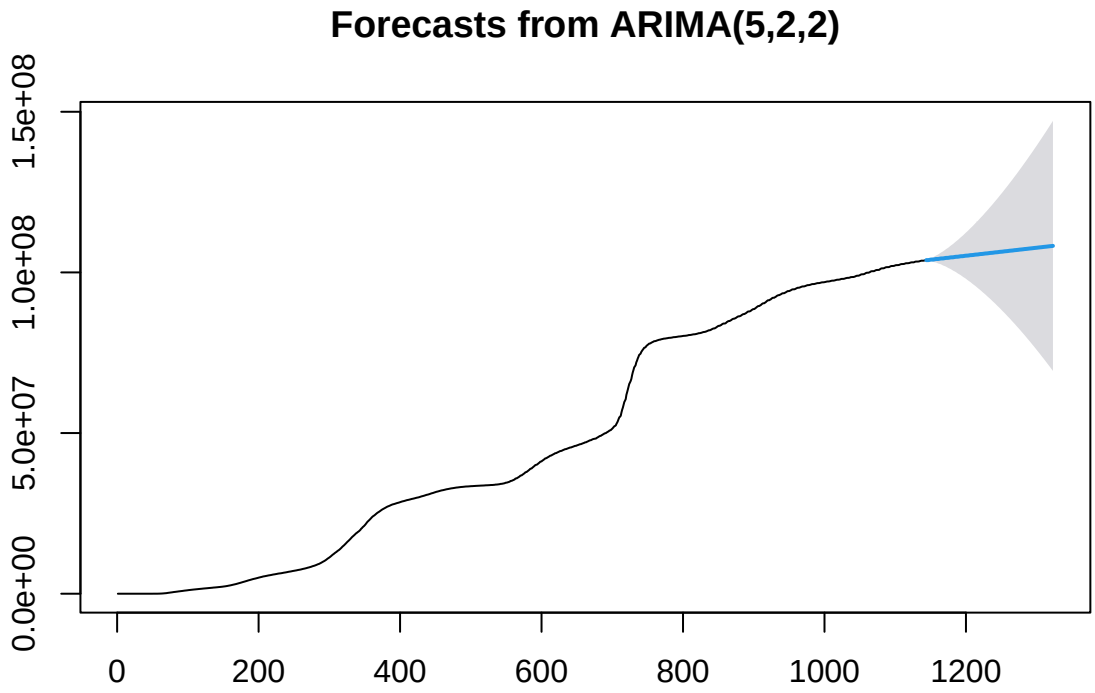


A model can then be created. For this analysis, an auto ARIMA model will be used from the `forecast` package.

```r
model <- auto.arima(tsdata$Cases)
print(model)
```

```
## Series: tsdata$Cases
## ARIMA(5,2,2)
##
## Coefficients:
##          ar1      ar2      ar3      ar4      ar5      ma1     ma2
##       0.1537  -0.6663  -0.3264  -0.3559  -0.5004  -0.9612  0.7471
## s.e.  0.0306   0.0247   0.0301   0.0238   0.0283   0.0267  0.0213
##
## sigma^2 = 2.305e+09:  log likelihood = -13916.8
## AIC=27849.6   AICc=27849.72   BIC=27889.91
```

An ARIMA(5,2,2) model is created. This model can then be used to predict future cases from the data supplied. 6 months (30-day periods) of forecasting will be performed.



The plot shows a mostly steady increase in cases.

Performing a Ljung-Box test on the residuals of the model will give a p-value to determine model performance.

```r
Box.test(model$resid, lag=15, type="Ljung-Box")
```

```
##
##  Box-Ljung test
```

```
##
## data:  model$resid
## X-squared = 553.38, df = 15, p-value < 2.2e-16
```

As the p-value is very close to zero, the model can be considered accurate for this application.

Bias

Sources of bias:

It is possible that there are positive cases of COVID19 that are not reported. I have personal knowledge of friends and family who tested positive but did not inform others which could mean unreported cases.

COVID19 deaths could also be undertracked as COVID19 complications could result in other conditions that cause death, and then is not reported as COVID19 causing the death.

There are potential other sources of bias in the data, as well as personal bias in the analysis.

Conclusion

From the analysis performed in this report, population size of a county has an effect on the delay of COVID19 infections and deaths where smaller counties will have a larger time before any cases or deaths are reported.

The report also finds there is no regional bias in COVID19 infections and deaths as the rate of cases and deaths per million are very similar between cities all over the continental United States.

Finally, an ARIMA model was built to forecast COVID19 cases across the United States with a high degree of accuracy according to the Ljung-Box test.