John Hopkins COVID-19 Dataset Analysis

12/03/2021

About

The following is an analysis of COVID-19 data available from the John Hopkins University to generate figures and models. The data set is taken from 01/22/2020 to 12/03/21 for the first, second, and third visualizations and from 1/01/2021 to 12/03/21 (or if knitted, to the present) for the time series projections. Information is added on a daily basis and the CSV files are available on the John Hopkins University's GitHub. Attributes such as the incident date, state, confirmed cases, deaths, recovered cases, latitude, longitude, and others can be found within the files.

Code

Step 1: Import Libraries

Libraries are first called to help with pulling the data. Note that if knitting this from another computer, please make sure that the following libraries are installed for the code to compile correctly.

```
# Import Libraries
# Library for Knitting as a PDF
library(tinytex)

# Libraries for Tidying Data
library(tidyverse)
library(reshape2)
library(scales)

# Library for Generating Time Series
library(forecast)

# Library for Date Formation
library(lubridate)
```

Step 2: Import Data

Next, the information from the website is extracted. This is done using the GitHub links that provide the CSV files in a raw format. Below the links are read and the information is extracted.

```
# Read Raw GitHub Link
url_in<-"https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/
csse_covid_19_data/csse_covid_19_daily_reports_us/11-19-2021.csv"
url_in2<-"https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/
```

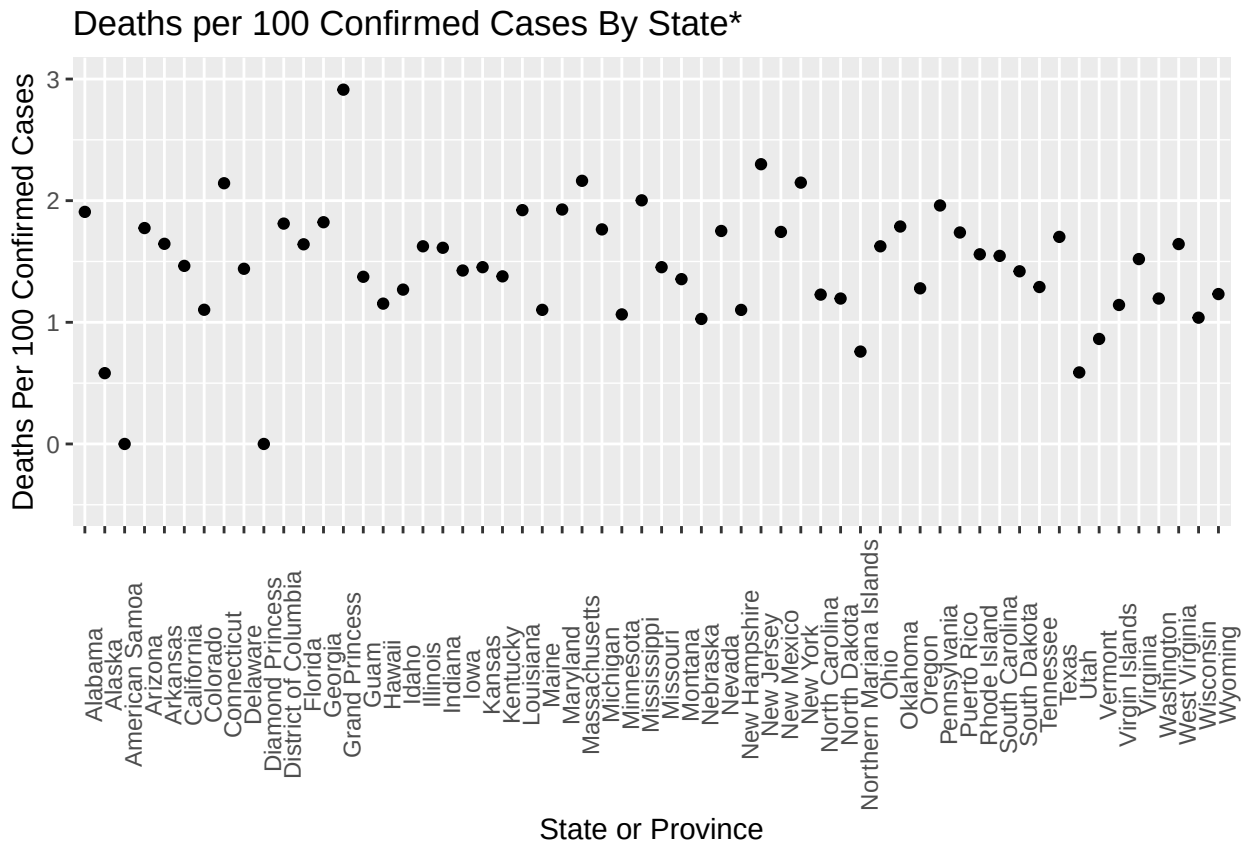
```
csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_
global.csv"

#Save Information Into Tibbles
COVID_cases<-read_csv(url_in)
Global_cases<-read_csv(url_in2)
```

After this step, the data is transformed into useful states for graphing and modeling.

Step 3: Clean, Tidy, Select Important Data for a Figure

The columns for the state, confirmed cases of COVID-19, and deaths caused by the virus are isolated from the first tibble. A ratio is then taken of the deaths per confirmed cases and multiplied by 100 to get the amount of deaths per 100 confirmed cases. Then the death to 100 confirmed cases ratio is plotted for each state. The data set also includes the Grand Princess cruise which was isolated off the border of the US once authorities were alerted that passengers contracted COVID-19. Thus an asterisk is placed in the figure's title.

```
# Select the State, Confirmed Cases, and Death Count
CRDA<-COVID_cases%>%
    select("Province_State","Confirmed","Deaths")

# Get the Ratio of COVID-19 Deaths to 100 Confirmed Cases and Save to a Column
CRDA$Ratio<-100*CRDA$Deaths/CRDA$Confirmed

# Plot the Information of the Ratio for Each State
p1<-ggplot(CRDA,aes(x=Province_State,y=Ratio))+
  geom_point()+
  theme(axis.text.x = element_text(angle =90))+
    labs(title = "Deaths per 100 Confirmed Cases By State*",
    x="State or Province",
    y="Deaths Per 100 Confirmed Cases")+
  coord_cartesian(ylim = c(-0.5,3))
p1
```



Step 4: Analysis of the Figure

According to the graph, most of the states tend to have a ratio of about 1 to 2 deaths per 100 cases. This however doesn't take into consideration that the data provided might not be completely reported. There is still the possibility that the amount of deaths or cases could be higher, other COVID-19 deaths could have been falsely attributed to different causes, or that carriers of COVID-19 didn't report being ill to the state. The graph also doesn't speak about the health of the remaining 98 - 99 people who survive COVID-19. Although the death rate is low, there is the chance that those who were infected with the virus didn't return to perfect health. Thus the 1 - 2 deaths per COVID-19 cases is auspicious, but should still be taken with a bit of skepticism.

Another thing to note is that Alaska, Delaware, and Georgia have abnormally low or high rates. This could be caused by peculiar behavior or traits of these states. These outliers should be investigated further in more detail to determine what makes their ratio outside of the normal values and whether these findings could benefit other states to decrease death rates.

Step 5: Generate Figures for COVID-19 Cases in Canada, Mexico, and the US

Next, the three most populated countries within North America (Canada, Mexico, and the US) are investigated. In this example, the amount of cases over the past year is graphed with population taken and not taken into consideration. Populations used in the calculations were taken at the beginning of the year

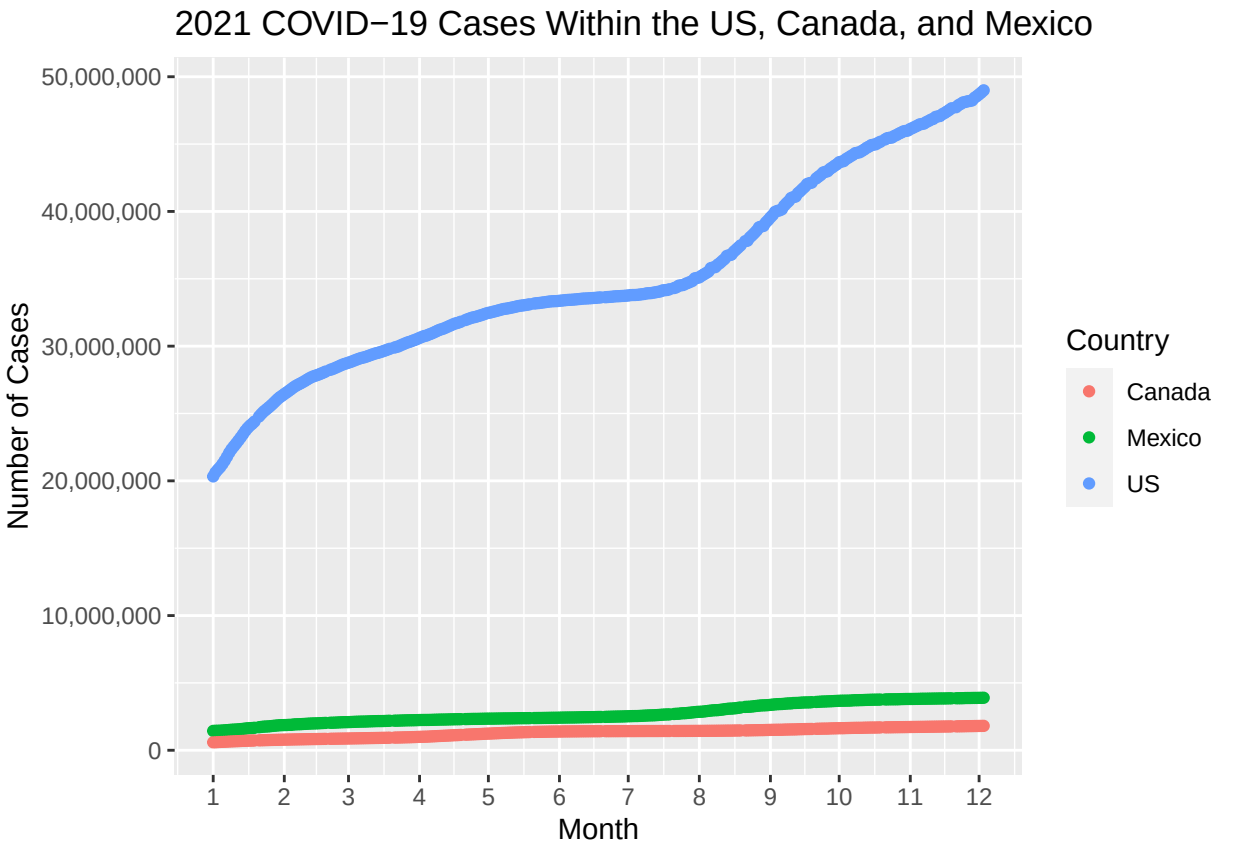
and held constant to make computing easier. With a more throrough evaluation, the growth rate of the populations should be taken into account.

The function is as follows:

```
# Declare the function
val1<-function(ex1){

  # Declare a List of the Names of the Countries, the Populations, and
  # Empty DataFrames
  lis1=list('Mexico', 'Canada', 'US')
  pop1=list(128900000,38010000,329500000)
  df3=data.frame()
  df3=df3[FALSE,]

  # Take the GitHub Raw Data and Remove Dates Before 2021
  df<-Global_cases%>%
    select(-'Province/State',-'Lat',-'Long',-contains('/20'))
  df2<-aggregate(.~`Country/Region`,df,sum)

  # Get the State Rows from the Data
  GR1<-function(nam1){
    dat2<-filter(df2,`Country/Region`==nam1)
    return(dat2)
  }

  # Create a Loop for Each Country
  for (c1 in 1:length(lis1)){

    # Get Rows for Graphing Depending on the State
    sav2<-GR1(lis1[c1])

    # If Population Taken Into Account, Divide Value By Population and
    # Multiply By 100 to Get the Percentage of Infected Individuals
    if(ex1=="Pop"){
      sav2[,2:ncol(sav2)]<-100*sav2[,2:ncol(sav2)]/pop1[c1]
      df3<-rbind(df3,sav2)
    }

    # If Population Isn't Taken Into Account, Return Values as is
    else{
      df3<-rbind(df3,sav2)
    }
  }

  # Melt the List in Preparation for Graphing
  df_melted<-melt(df3)

  # Rename the Columns and Format DataFrame for Plotting
  require(scales)
  colnames(df_melted)<-c("Country","Month","Deaths")
  df_melted$Month<-as.Date(df_melted$Month,'%m/%d/%y')

  # Plot the Graph With Population Taken Into Account with the Correct Labeling
  if(ex1=="Pop"){
```

```
  p2<-ggplot(df_melted, aes(x=Month,y=Deaths))+geom_point(aes(color=Country))+
  labs(title = "Percent of COVID-19 Cases Within the US, Canada, and Mexico",
  x="Month",
  y="Percent of COVID-19 Cases in Given Population (%)")+
  scale_y_continuous(labels=comma)+
  scale_x_date(date_breaks = "months", date_labels="%b")
  }

  # Plot the Graph Without Population Taken Into Account with the Correct
  # Labeling
  else{
  p2<-ggplot(df_melted, aes(x=Month,y=Deaths))+geom_point(aes(color=Country))+
  labs(title = "2021 COVID-19 Cases Within the US, Canada, and Mexico",
  x="Month",
  y="Number of Cases")+
  scale_y_continuous(labels=comma)+
  scale_x_date(date_breaks = "months", date_labels="%b")
  }

  # Show Graph and Return Tibble for Later
  print(p2)
  return(df3)
}
```

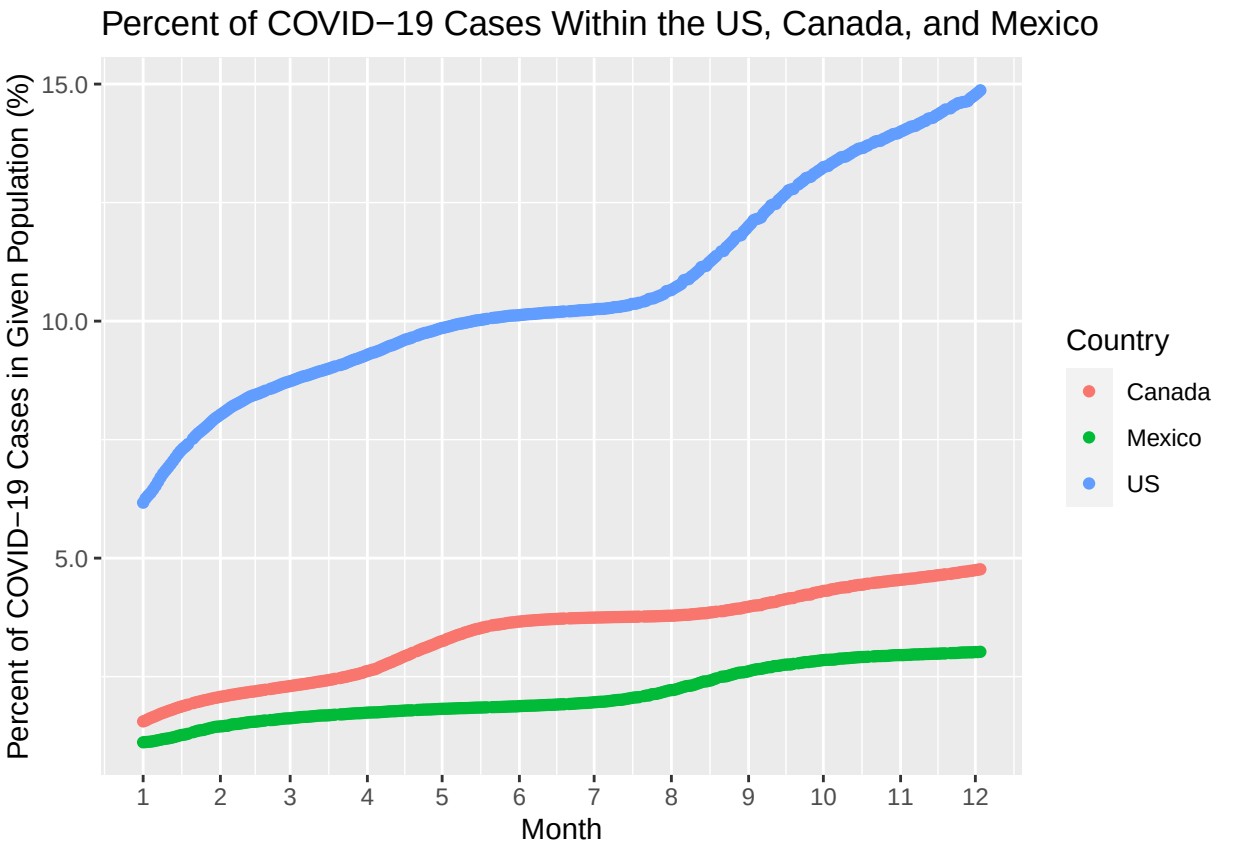
```
# Run Without Population Taken Into Account
sav1<-val1("None")
```



The first graph shows the rates of COVID-19 cases over the past year without population taken into account. The US has the greatest amount of COVID-19 cases throughout the year with Mexico and then Canada following. Relatively speaking, Canada and Mexico's rates are comparatively stable to the US' which climbs over the year.

To get a further clarification on this figure, population is taken into account for the next one.

```
# When Taken Into Account for Population
sav2<-val1("Pop")
```



Similar to the last graph, the COVID-19 rates for the US continues to be greater and climb faster than Canada and Mexico's going from 5% to 15% of the country's population within a year. However, compared to the last figure, Canada's rates climb faster than Mexico's and are higher overall. The place of Mexico and Canada switch and in this graph, Mexico has lower COVID-19 cases than Canada. Throughout the year, Canada and Mexico both remain below the US' minimum of approximately 5%.

Step 6: Generate Time Series Projections for the US

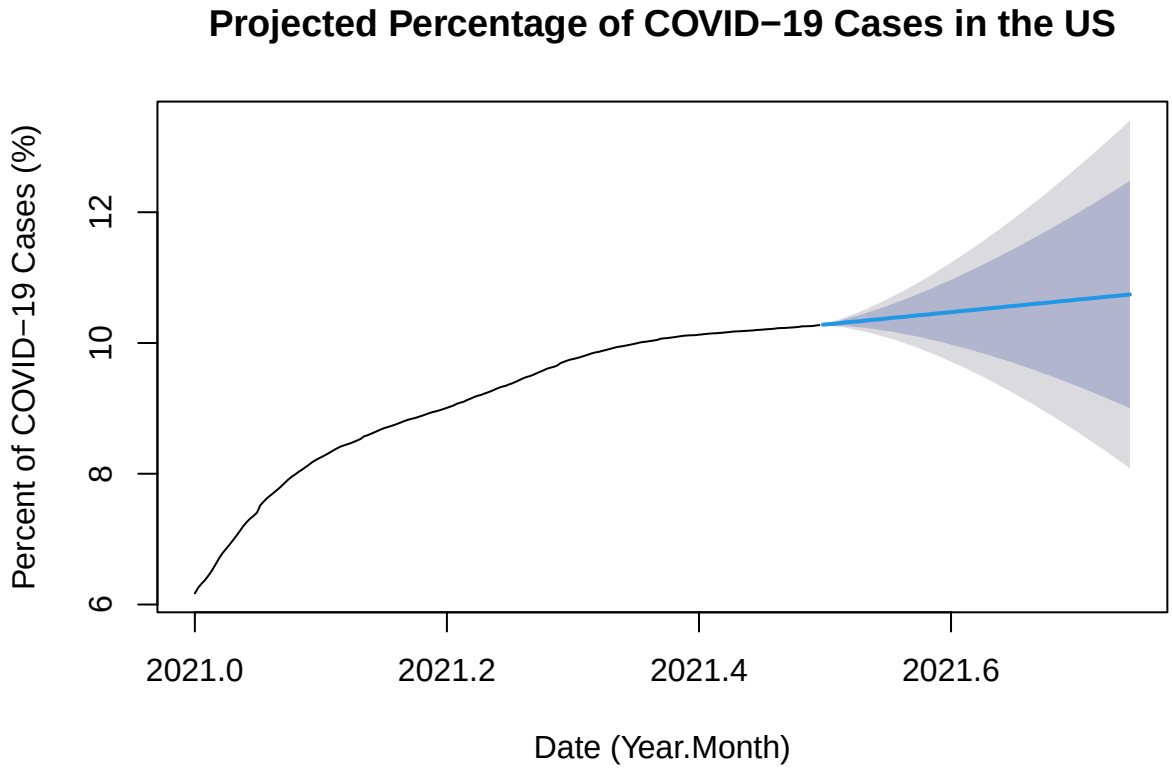
Next, the percentage of COVID-19 cases within the US is predicted. A range is projected three months out at different times within this past year. First an overall function is created that has the time series function for predictions.

```
# Declare Function for Time Series, Input Day in Year Where Prediction Starts
# and Data for Curve
val2<-function(span1,sav1){
  # Convert Information Into Numeric and Assign Dates for Each Value
  df3_ts1<-as.numeric(unlist(unname(sav1[3,2:length(sav1)])))
  df3_ts2<-df3_ts1[1:span1]

  # Place Into a Time Series Starting on the 1rst of January of 2021 with the
  # Data Separated at Daily Intervals
  df3_ts<-ts(df3_ts2,start=decimal_date(ymd("2021-01-01")),frequency=365.25)

  # Apply ARIMA to Values
  df3_fit<-auto.arima(df3_ts)
```

```
  # Generate Graph With Forecasts Into the Future
  proj1=90# Days in the Future
  plot(forecast(df3_fit,proj1),xlab="Date (Year.Month)",ylab="Percent of COVID-19 Cases (%)",
       main="Projected Percentage of COVID-19 Cases in the US")
}
```

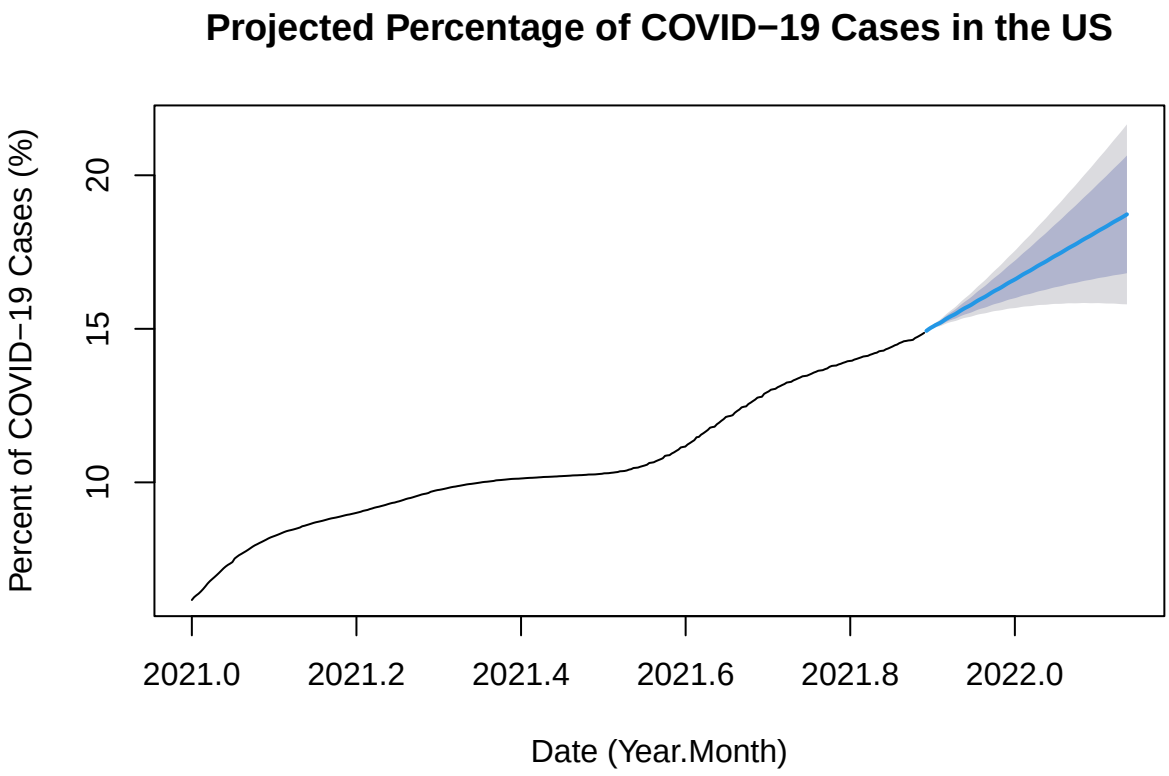
The first time series model shows a model that would have been generated about halfway through this year. The range of cases for the US over the next three months is expected to be between 8% and 14%. Since this is total cases and the amount can't decrease this means that the range extends from 10% to 14%.

```
val2(182,sav2)
```



The next time series model is taken at the present. The expected amount of cases varies between 16% and 20%.

```
val2(length(sav2),sav2)
```



These graphs show that at varying stages in the year, there was a potential for the cases and spread of the virus to change by different amounts. Halfway through the year, COVID-19 cases in the US were projected to grow to 14% or remain constant at 10%. However, since the virus continued to spread the range moved upwards. Now the range extends from 16% to 20%.

Conclusions

Figure 1: Most of the states have a ratio of 1 to 2 deaths per 100 COVID-19 cases. Although this number is low, the health of the remaining 98 to 99 people that remain per 100 cases isn't mentioned. States that serve as outliers should be investigated more to see if there are peculiar laws enforced or traits associated with the residents. Findings could benefit the remaining states and decrease the cases of death.

Figure 2 and 3: For the US, the amount of COVID-19 cases climbed faster than for Mexico and Canada in both figures. Thus, the US should look into the measures that Canada and Mexico have taken to reduce the spread of the virus and apply this nationally. When the figure didn't take percentage into account, Canada and Mexico's rates were less than the 10% maximum for the US infected cases. When percentage is taken into account, the lowest amount of COVID-19 cases went from Canada to Mexico and both states were about 1/3rd to 1/6th that of the US'. Thus, it's important to keep in mind ratios when comparing different countries of variable sizes.

Model 1 and 2: At different times during the year, there was a potential for the rates of COVID-19 cases to either increase or decrease. The model is generated on previous data and predicts future behaviors based on past ones. Thus, the range for the COVID-19 projected cases is higher at the end of the year compared to the middle.

Forms of Bias

All: All figures and models are the mercy of reported information by the clinics and hospitals of the amount of cases, deaths, and where the events took place. If there is stigma for being tested or associated with COVID-19, this would bias the data and make the results less accurate. There is also an inherent bias when the cases are segregated by state or country. Behaviors towards COVID-19 vary greatly within a country or state which can simplify the results and overlook outliers such as states within Mexico, counties in US states, or providences in Canada.

Figure 1: In the case of the first figure 1, the death rate might be higher or lower depending on whether the amount of deaths or cases are underreported. As stated earlier, the results also can be misleading and deceptively comforting. The 1% to 2% death rate per 100 cases tells nothing about the condition of the suriving invdivuals or whether there are long lasting effects.

Figure 2 and 3: Treating the US, Canada, and Mexico as single entities biases the figures since there are providences, states, and counties that have unique behaviours towards following COVID-19 regulations which might skew the overall results. For a future more in-depth report, analysis of each country should be looked into with greater detail. The bias with these figures is the simplification that naturally comes with countries that are collections of unique groups of demographics and people.

Model 1 and 2: The last two models are biased for similar reasons to the last two figures: the graphs simplify the US as an agglomerate whole. Certain states would perhaps have larger or smaller projections depending on characteristics and qualities that would predispose the states towards COVID-19 elimination or spread.