## Group 9 Report 2: Kruthi, Pawan, Haleigh, Sanjeev, Thrishna

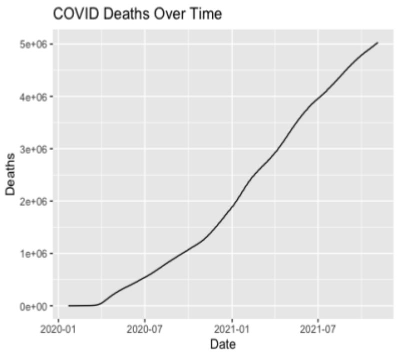
### Introduction to the data set and research problems

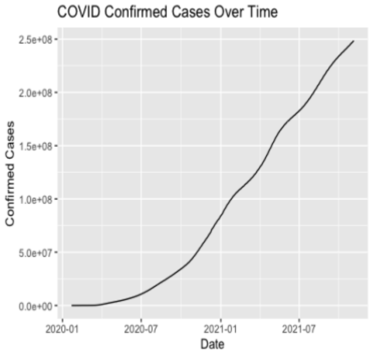
The COVID-19 pandemic has dramatically affected the daily lives of everyone in the world. As such, our group wanted to investigate the toll the pandemic has taken on different places across the globe. The dataset we have selected consists of worldwide COVID-19 data, broken down for each country, and broken down for each state within the US. The attributes within this data set are confirmed cases, deaths, recoveries, the rate of increase, and their corresponding dates. Within this dataset, we plan to use both the countries-aggregated.csv and worldwide-aggregated.csv to combine all the attributes and discover the relationships and trends within this combined data set.

Specifically, we want to research a few areas like, are confirmed COVID-19 cases correlated with deaths? And if so, is there a lag between high confirmed COVID-19 cases and high death cases? Additionally, we want to look at how the trend of COVID-19 cases change based on geographic location around the world. This delves into how strict countries were in quarantining and lockdown, so it would be interesting to see if there was an optimal quarantining method. Furthermore, we also want to look at if we can predict future COVID-19 cases using forecasting. Finally, we want to look at the times of the year where COVID-19 cases and deaths occur the most and if it is seasonal.

### Exploratory Data Analysis

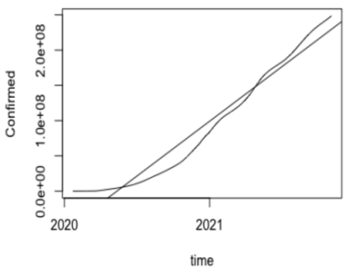
```
library(readr)
library(ggplot2)
library(astsa)
library(lubridate)
library(tidyverse)
covid = read_csv("worldwide-aggregate.csv")
countries<- read_csv("countries-aggregated.csv")
world <- read_csv("worldwide-aggregate.csv")
ggplot(covid, aes(x=Date, y=Deaths)) +geom_line() + xlab("Date") + ylab("Deaths") + ggtitle("COVID Deaths Over Time")
ggplot(covid, aes(x=Date, y=Confirmed)) +geom_line() + xlab("Date") + ylab("Confirmed Cases") +  ggtitle("COVID Confirmed Cases Over Time")
```

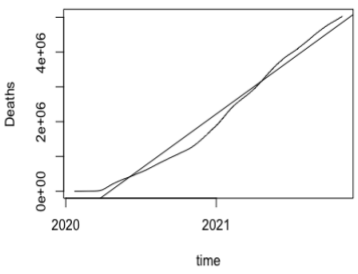




We can see that over time the deaths are increasing as well as the confirmed COVID cases.

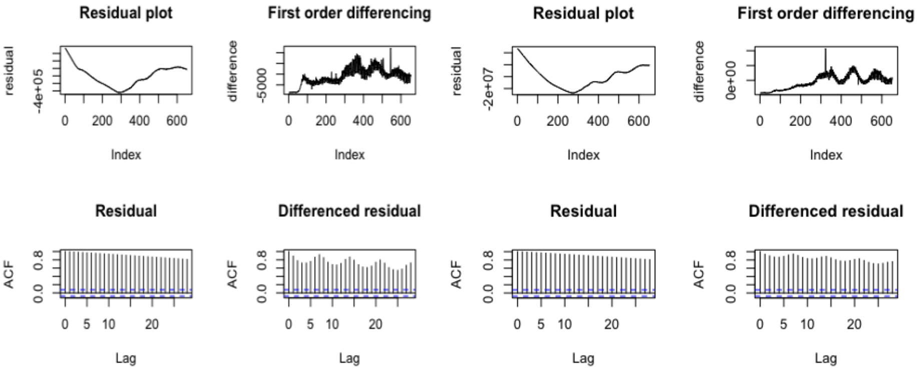
```
y=world$Confirmed
x=world$Deaths
time=world$Date
y.lm=lm(y~time)
plot(world$Date,y, type='l', xlab='time', ylab='Confirmed')
abline(reg=y.lm)x.lm=lm(x~time)
plot(world$Date,x, type='l', xlab='time', ylab='Deaths')
abline(reg=x.lm)
```



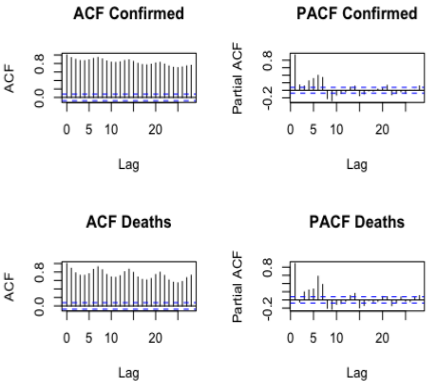


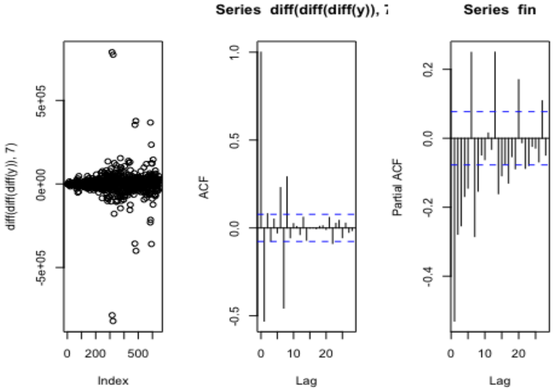
```
par(mfrow=c(2,2))
res=resid(x.lm)
```

```
plot(res, type='l', ylab = 'residual', main='Residual plot')
plot(diff(res), type='l', ylab="difference", main = 'First order differencing')
acf(res, main='Residual')
acf(diff(res), main="Differenced residual")
par(mfrow=c(2,2))
res=resid(y.lm)
plot(res, type='l', ylab = 'residual', main='Residual plot')
plot(diff(res), type='l', ylab="difference", main = 'First order differencing')
acf(res, main='Residual')
acf(diff(res), main="Differenced residual")
```

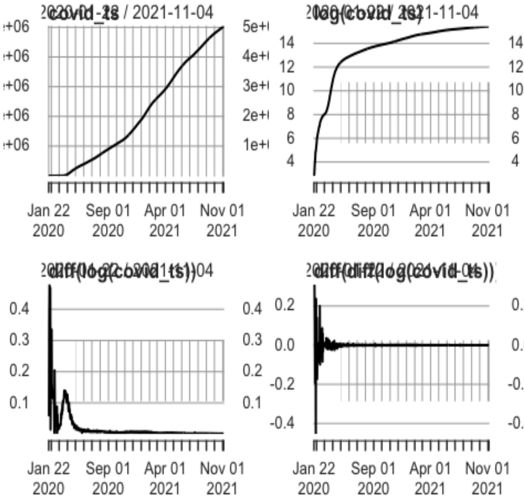


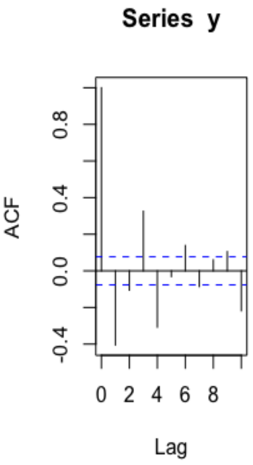
```
#1st difference, acf, pacf
par(mfrow=c(2,2))
acf(diff(y), main='ACF Confirmed')
pacf(diff(y), main='PACF Confirmed')
acf(diff(x), main='ACF Deaths')
pacf(diff(x), main='PACF Deaths')
par(mfrow=c(1,3))
#plot 1st seasonal difference, acf, pacf
plot(diff(diff(diff(y)), 7))
acf(diff(diff(diff(y)), 7))
fin = diff(diff(diff(y)), 7)
pacf(fin)
```

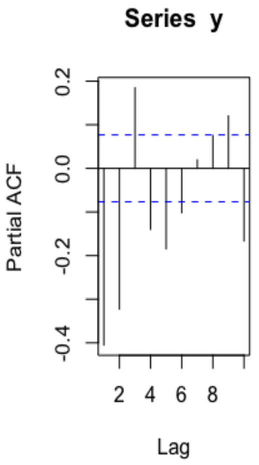




```
covid_ts = xts(covid$Deaths, as.Date(covid$Date, format='%Y-%m-%d'))
par(mfrow=c(2,2))
plot(covid_ts)
plot(log(covid_ts))
plot(diff(log(covid_ts)))
plot(diff(diff(log(covid_ts))))
y = diff(diff(log(covid_ts)))
par(mfrow=c(1, 2))
acf(y, lag.max=10, na.action=na.pass)
pacf(y, lag.max=10, na.action=na.pass)
```





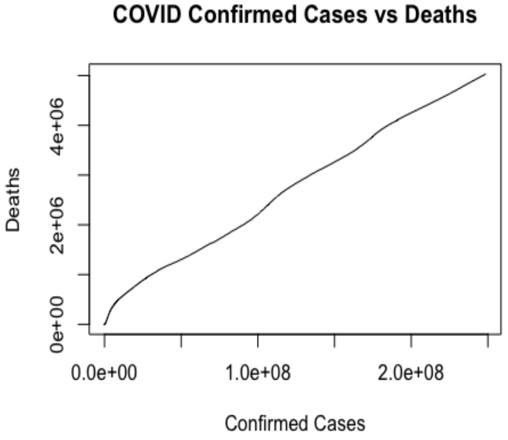


From this we can see that the data is not normally distributed and it becomes stationary after being differenced twice. We can also see that there is no clear cut off for the ACF nor PACF.

## Statistical results and conclusions from models and methods

### Research Question #1: Are confirmed cases linearly correlated with deaths?

```
cor(covid$Confirmed, covid$Deaths)
## [1] 0.9975087
plot(covid$Confirmed, covid$Deaths, xlab="Confirmed Cases", ylab="Deaths", main="COVID Confirmed Cases vs Deaths", type='l')
```

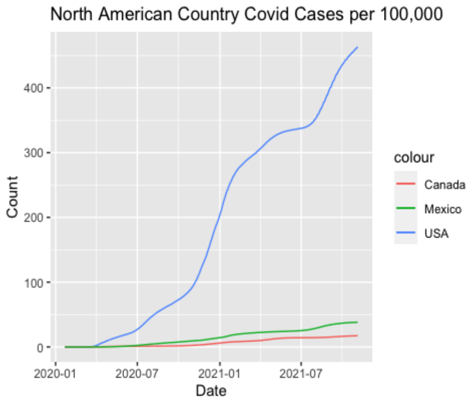


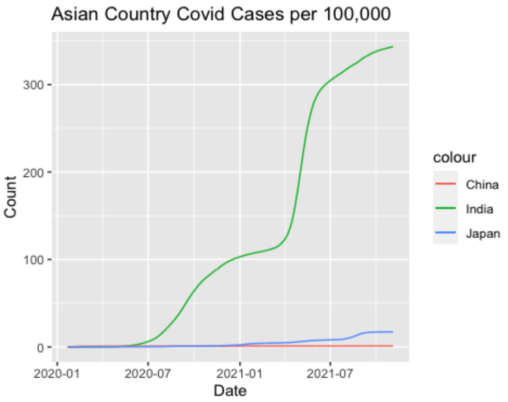
Confirmed cases and deaths appear to have a linear relationship and have a high correlation of 0.997.

### Research Question #2: How does the trend of covid cases change based on geographic location?

```
### North American Countries
us_country <- countries[countries$Country == "US",]
canada <- countries[countries$Country == "Canada",]
mexico <- countries[countries$Country == "Mexico",]
us_country$Date=with_tz(us_country$Date, "America/New_York")
canada$Date=with_tz(canada$Date, "America/New_York")
mexico$Date=with_tz(mexico$Date, "America/New_York")
us_country$per = us_country$Confirmed/100000
canada$per = canada$Confirmed/100000
mexico$per = mexico$Confirmed/100000
(ggplot(us_country, aes(x = Date, y = `per`)) +geom_line(aes(y = `per`, color = "USA"))  +  geom_line(data = canada, aes(color = "Canada")) + geom_line(data = mexico, aes(color = "Mexico")) +xlab('Date') + ylab('Count') +  ggtitle('North American Country Covid Cases per 100,000'))
### Asian Countries
india <- countries[countries$Country == "India",]
china <- countries[countries$Country == "China",]
japan <- countries[countries$Country == "Japan",]
india$Date=with_tz(india$Date, "America/New_York")
china$Date=with_tz(china$Date, "America/New_York")
```

```
japan$Date=with_tz(japan$Date, "America/New_York")
india$per = india$Confirmed/100000
china$per = china$Confirmed/100000
japan$per = japan$Confirmed/100000
(ggplot(india, aes(x = Date, y = `per`)) + geom_line(aes(y = `per`, color = "India"))  + geom_line(data = china, aes(color = "China"))
+  geom_line(data = japan, aes(color = "Japan")) + xlab('Date') +  ylab('Count') + ggtitle('Asian Country Covid Cases per 100,000'))
```

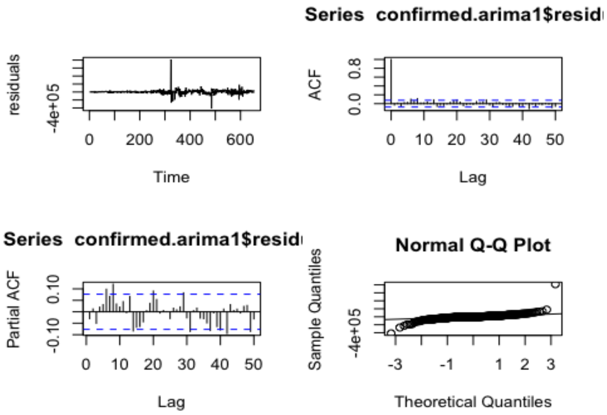


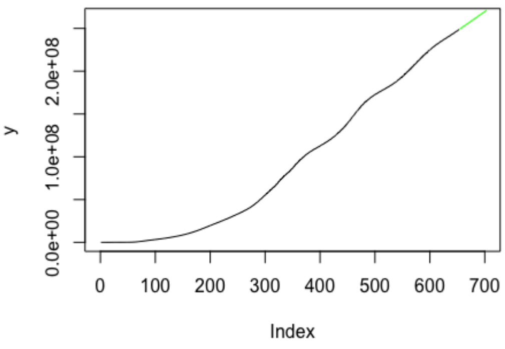


As seen by the graphs, there seems to be a high count of COVID-19 cases in the United States and India. While these two countries are in different geographic locations, it seems as though the quarantining policy wasn’t strict enough, wasn’t enforced properly, or wasn’t followed in these two countries. Further research will need to be done to see if there is a change in trend based on geographic location. However, it is more likely based on this analysis that a refinement to the research question could be made to measure the difference in COVID-19 case trends in countries based on the strictness of their quarantining policies.

### Research Question #3: How can we predict the number of covid cases with forecasting?

```
par(mfrow=c(2,2))
#explaratory analysis (de-trend confirmed cases)
y=world$Confirmed
time=world$Date
#model 1
confirmed.arima1 = arima(y, order=c(0, 2,1), seasonal=list(order=c(0, 1,1), period=7))
plot(confirmed.arima1$residuals, ylab = 'residuals', type = 'l')
acf(confirmed.arima1$residuals, lag.max  = 50)
pacf(confirmed.arima1$residuals, lag.max  = 50)
qqnorm(confirmed.arima1$residual)
qqline(confirmed.arima1$residual)
Box.test(confirmed.arima1$residuals, fitdf=1, lag = 20, type="Ljung")
confirmed.arima1$aic
#Forecasting with model 1 confirmed cases ARIMA(0,2,1)x(0,1,1)7
forecast=predict(confirmed.arima1, n.ahead =50)
par(mfrow=c(1,1))
plot(y, xlim = c(0,700), ylim = c(0, 262659973), type='l')
lines(forecast$pred, col="green")
```

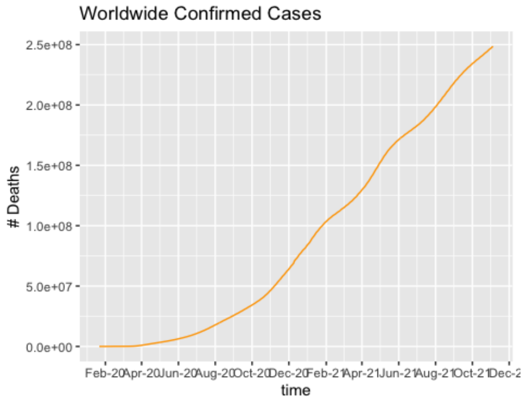




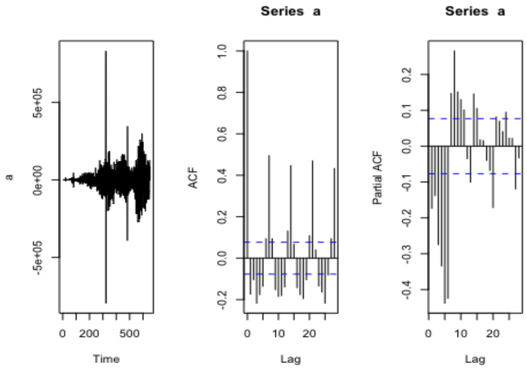
The overall conclusion regarding this question is that confirmed cases can be forecasted using an ARIMA(0,2,1)x(0,1,1)7 model. The period of 7 for seasonality of confirmed cases can be explained by the trends we see weekly and the way the data is collected. Oftentimes the patterns are dependent on the week.

### Research Question #4: What times of the year do the most covid cases and/or deaths occur?

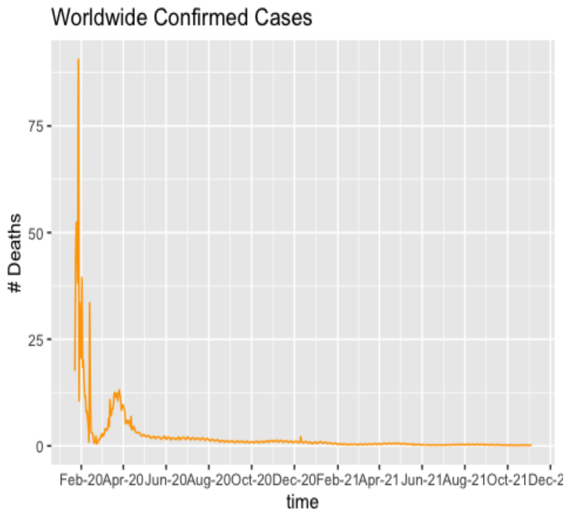
```
#Worldwide Confirmed Cases
par(mfrow=c(1, 3))
worldwide_data <- read_csv("worldwide-aggregate.csv")
worldwide_data$Date=as.Date(with_tz(worldwide_data$Date, "America/New_York"))
ggplot(worldwide_data, aes(x=Date, y=Confirmed))+geom_line(color = 'orange')+labs(x='time', y='# Deaths', title='Worldwide Confirmed Cases')+scale_x_date(date_breaks = "2 months" , date_labels = "%b-%y")
```



```
plot.ts(diff(diff(worldwide_data$Confirmed)))
acf(diff(diff(worldwide_data$Confirmed)))
pacf(diff(diff(worldwide_data$Confirmed)))
```



```
ww_arima = arima(worldwide_data$Confirmed, order= c(0,2,0))
#Worldwide Increase Rate
worldwide_data$Date=as.Date(with_tz(worldwide_data$Date, "America/New_York"))
ggplot(worldwide_data, aes(x=Date, y=`Increase rate`))+geom_line(color = 'orange')+labs(x='time', y='# Deaths', title='Worldwide Confirmed Cases')+ scale_x_date(date_breaks = "2 months" , date_labels = "%b-%y")
```


Worldwide Confirmed Cases
# Deaths
Feb-20
Apr-20
Jun-20
Aug-20
Oct-20
Dec-20
Feb-21
Apr-21
Jun-21
Aug-21
Oct-21
Dec-2
time