

Predicting number of Covid19 deaths using Time Series Analysis (ARIMA Model)

By Navid Mashinchi

Executive Summary:

On March 11th, 2020 the World Health Organization (WHO) declared the novel coronavirus (Covid19) outbreak as a global pandemic. In this paper, a time series analysis to predict the number of deaths in the United States starting from August 1st – August 21st and August 1st – November 1st is modeled and studied. The time series model that was selected to make the prediction is called Auto Regressive Integrated Moving Average (ARIMA) model.

The paper is divided into the following sections:

1. Why & What is Time Series Analysis
2. When we can't use Time Series Analysis
3. Components of Time Series Analysis
4. Demonstration of Time Series Analysis
5. Conclusion

The data has been drawn from “Our World in Data” (<https://ourworldindata.org/covid-deaths>) and consists of the necessary information to conduct the time series analysis. The variables that are relevant to answer our research question are the dates (2019/12/31-2020-08/01), total deaths, new deaths and location (USA). The data has been cleaned and adjusted to satisfy all the necessary assumptions to use ARIMA to make the prediction.

The forecast of new deaths for the next 21 and 90 days reaches 18,589 (Total Deaths 171,903) and 82,653 (Total Deaths 235,967) respectively. The result of our projection has been very close when comparing it to CNN's projection. CNN projected on August 2nd that about 19,000 people could die in the next 2 weeks in the United States. In addition to that prediction, they also predicted on July 31st in their show “CNN Coronavirus Town Hall” the total numbers of death by November. CNN forecasted 231,000 death from Covid19 by November. The results of our ARIMA Model are very close when comparing it to CNN's projection.

Result:

Date:	Our Projection	CNN's Projection
2020/08/01 – 2020/08/21	New Deaths: 18,589	New Deaths: 19,000
2020/08/01 – 2020/10/31	Total Deaths: 235,967	Total Deaths: 231,000

1. Why & What is Time Series Analysis:

Time series analysis (TSA) is a statistical technique that consists of data points listed in time order. The x axis is made up of equally spaced points in time and the y axis contains the outcome values that are going to be projected from our model based on previous observed values. This technique is suitable for research questions such as forecasting future sales.

The reason why time series analysis exists, is due to the fact that the outcome variable in our model is dependent on one single explanatory variable only: *time*.

Suppose you run a shoe store and have the data available that tells you how many shoes you have sold in the past years. Given the data available, time series analysis would be applicable if you would like to predict how many shoes your store will sell in the future. In this case the outcome variable would be the number of shoes sold and the one and only explanatory variable would be time.

Other forecasting algorithms such as linear regression or logistic regression use one or more explanatory variables. Further there is a difference when it comes to the assumptions when comparing linear regression, logistic regression and the time series technique ARIMA.

In **Linear Regression** the following assumptions have to be met:

- Independence of observations.
- Homoscedasticity of errors (equal variance).
- A linear relationship.
- Errors are normally distributed.

In **Logistic Regression** the following assumptions have to be met:

- Dependent variable has to be binary.
- Independence of observations.
- Linearity in the logit for continuous variables.
- Lack of influential outliers
- Absence of multicollinearity

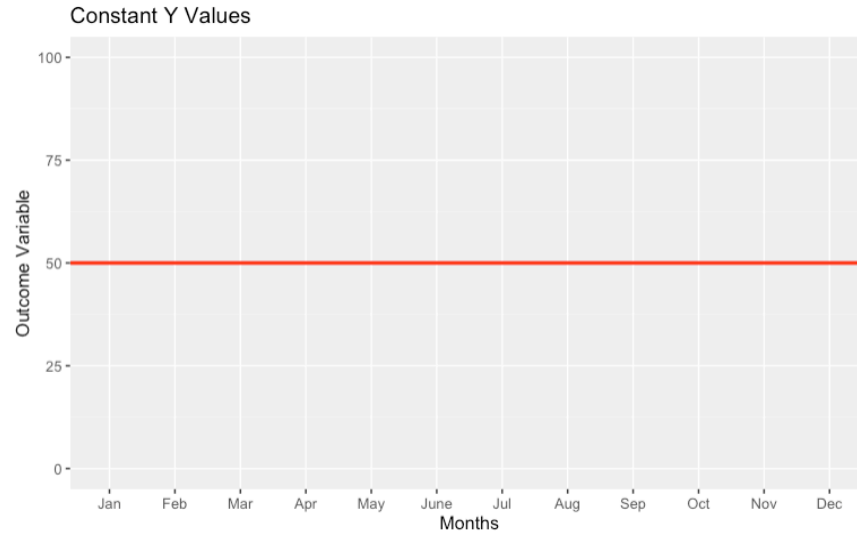
In **Time Series Analysis ARIMA** the following assumptions have to be met:

- Data has to be stationary.
- Data should be univariate. As mentioned above TSA ARIMA works on a single variable only.
- Data should be in time series data format.

2. When can we not use Time Series Analysis:

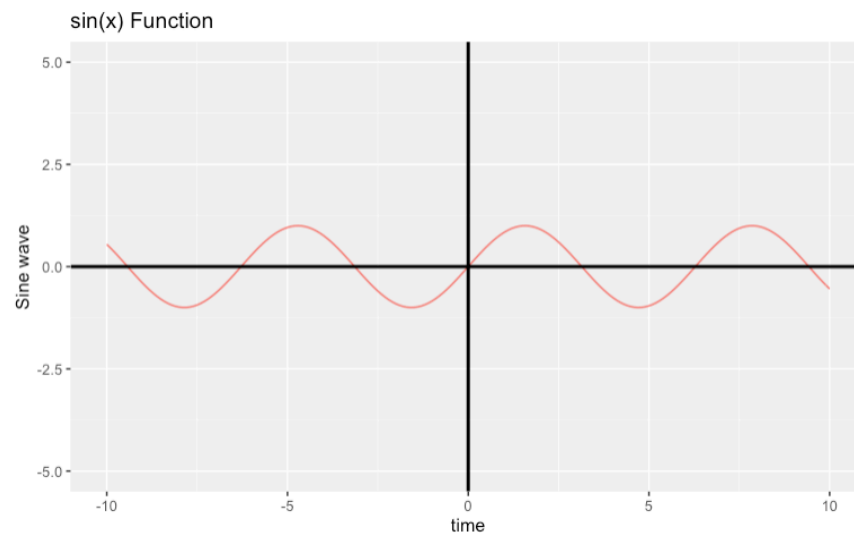
There are times when time series analysis is not the best statistical technique to answer a research question. There are 3 points that make this technique not suitable.

- a. When the data points are always constant.



The x axis has equally spaced points that represent the only variable time and the y values represent the outcome variable. Having constant data points throughout a period really makes this statistical technique not really useful. Consider the shoe store example again. If we sell in all months the same number of shoes, then there is no point of conducting a time series analysis to predict future sales, since the predicted outcome will always be the same.

b. When the data points represent a known function.

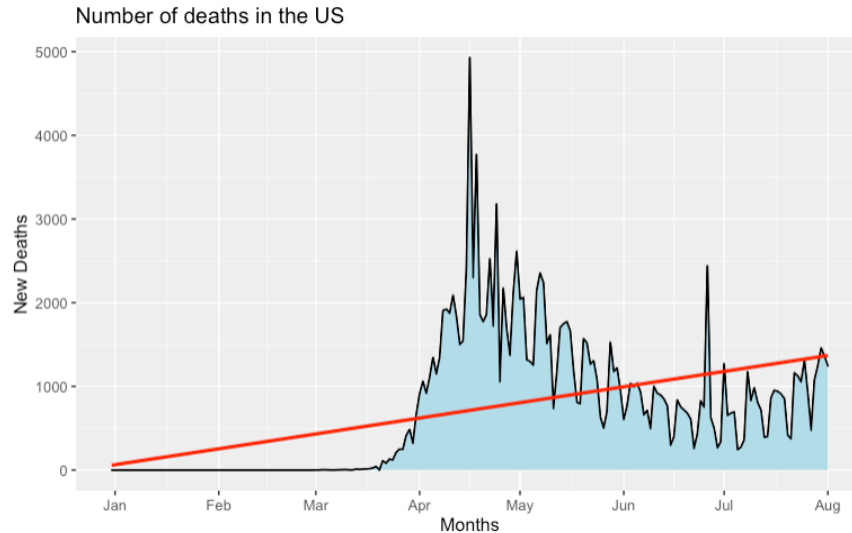


In the graph above we see the $\sin(x)$ function. If we add a value to this function, the predicted outcome can be easily computed. Using a time series analysis technique won't be necessary, since we will be able to calculate the predicted value by just plugging in the value into the $\sin(x)$ function. This applies for all other functions, such as $\cos(x)$ and etc. Hence, if the graph of our data points looks like a function then time series analysis doesn't become applicable.

c. When our data is not stationary.

As mentioned above, one of the assumptions for time series analysis (ARIMA) is that the data has to be stationary. In order to have stationary data, the following conditions have to be met:

- Mean has to be constant according to the time.
- Variance has to be equal in different time intervals from the mean. In other words, the distance of the points should be the same from the mean.
- Covariance has also to be equal.



The graph above shows the number of deaths from January – August 1st due to Covid19. The red line displays the mean and as we can see above the mean is constantly increasing. First, we can see that the means aren't constant, therefore the data is violating the assumption for constant mean throughout the time. Secondly, we also see that the variances aren't equal. As we can see above the distance between the points and the mean line varies a lot. Consider the points spiking in mid-April and May. We can clearly see that the distances from the data point to the mean are not the same. Hence the data violates not only the equal mean assumptions, but also the equal variance assumption.

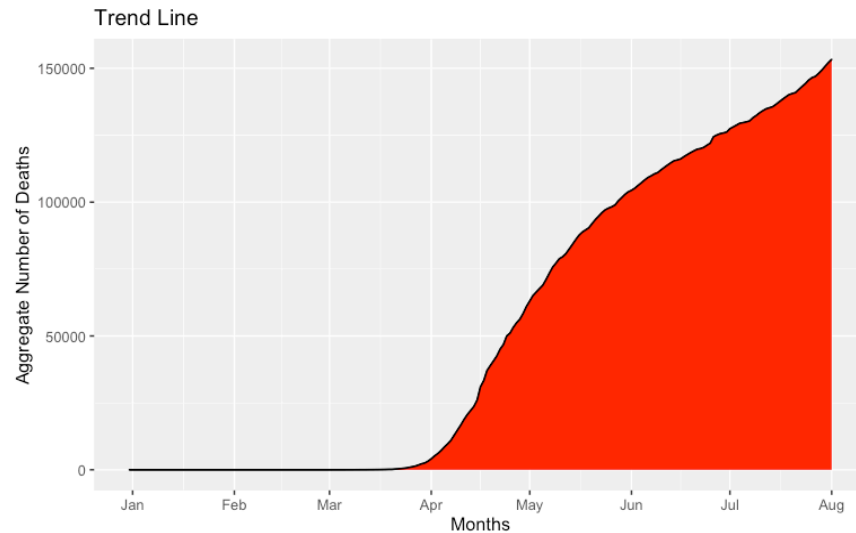
3. Components of Time Series Analysis:

There are three components that we need to understand when it comes to time series analysis.

a. General Trend - b. Seasonality - c. Irregular Fluctuations.

a. General Trend.

A general trend tells us how our data points are behaving. Behavioral trends can be described as increasing, decreasing or constant.



The above graph shows the trend of the total number of deaths, due to Covid19 in the United States between January – August 1st. Here we can clearly see an increasing trend. This also becomes evident when looking at the data points from the data set, which represent the number of deaths. (as seen below)

[1]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[27]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[53]	0	0	0	0	0	0	0	0	0	0	0	1	2	6	9
11	12	14	17	21	26	28	30	40	47	57	69	85			
[79]	108	150	150	260	340	471	590	801	1050	1296	1707	2191	2509		
3170	4079	5138	6053	7157	8501	9647	10989	12895	14817	16690	18777	20608			
[105]	22108	23649	26057	30985	33284	37054	38910	40682	42539	45063	46784	49963	51017		
53189	54876	56245	58355	60966	63006	65068	66385	67682	68934	71078	73431	75670			
[131]	77180	78794	79528	80684	82387	84133	85906	87568	88754	89562	90353	91921	93439		
94702	96007	97087	97720	98220	98916	100442	101617	102836	103781	104383	105147	106181			
[157]	107175	108211	109143	109802	110514	111007	112006	112924	113820	114669	115436	115732	116127		
116963	117717	118434	119112	119719	119975	120402	121228	121979	124416	125039	125539	125804			
[183]	126140	127410	128062	128740	129434	129676	129947	130306	131480	132309	133291	134097	134814		
135205	135605	136466	137419	138358	139266	140119	140534	140906	142066	143190	144242	145546			
[209]	146460	146935	148011	149256	150713	152070	153314								

b. Seasonality.

Seasonality plays a big part when it comes to time series analysis. One example of seasonality is when you see a spike in our data points. Suppose you have the data set of air travelers throughout a year. The data points of that graph would show a big spike in the month of December, due to the Christmas break. On the other hand, we could potentially see a big dip in the month of March and April 2020 where Covid19 just began to spread. This would lead to a major decrease in the number of air travelers.

c. Irregular Fluctuations.

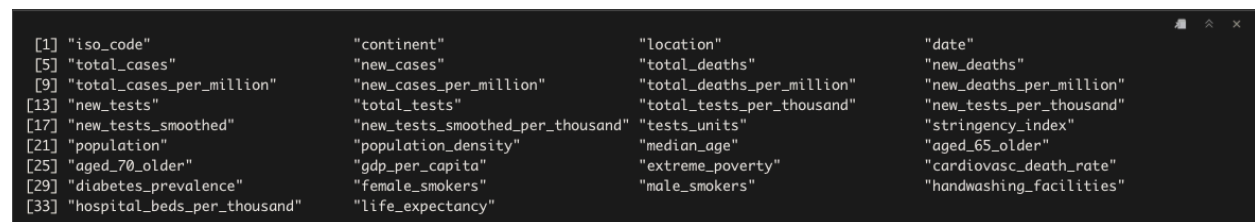
This component is referred to the random component of time series analysis. Irregular fluctuations are uncontrolled situations in which the y values would change. In the air travelers example, flight cancelations due to weather conditions such as a storm would result in the number of air travelers to decrease. The cancelation of flights due to the storm would be an example of an uncontrolled situation in which the y value would be affected. This wouldn't mean necessarily that there will be a storm at the same time next year. This is where the random effect comes into place.

4. Demonstration of Time Series Analysis

Research Question: How many people are going to die due to Covid19 in the United States from August 1st – August 21st and August 1st – November 1st. We are going to use time series analysis to answer this question. In particular we are going to use the ARIMA model. Time series analysis is the perfect statistical technique to answer this question, since we are going to project how many people are going to lose their lives unfortunately due to the disease Covid19.

Data Source: “Our World in Data” - <https://ourworldindata.org/covid-deaths>

The data consists of 34,033 rows and 34 columns. Below is a list of variables that are in the data.



[1] "iso_code"	"continent"	"location"	"date"
[5] "total_cases"	"new_cases"	"total_deaths"	"new_deaths"
[9] "total_cases_per_million"	"new_cases_per_million"	"total_deaths_per_million"	"new_deaths_per_million"
[13] "new_tests"	"total_tests"	"total_tests_per_thousand"	"new_tests_per_thousand"
[17] "new_tests_smoothed"	"new_tests_smoothed_per_thousand"	"tests_units"	"stringency_index"
[21] "population"	"population_density"	"median_age"	"aged_65_older"
[25] "aged_70_older"	"gdp_per_capita"	"extreme_poverty"	"cardiovasc_death_rate"
[29] "diabetes_prevalence"	"female_smokers"	"male_smokers"	"handwashing_facilities"
[33] "hospital_beds_per_thousand"	"life_expectancy"		

Collection method: Observations

Data definitions:

In order to work on the research question, we had to clean the data and extract the relevant variables that help us conduct the time series analysis. We created a data frame that consisted of the dates (2019/12/31-2020-08/01), total deaths, new deaths and location (USA).

“date” variable: Represents the dates that we are going to use for our analysis. Luckily the class of the date column was already date. If the date column wasn't a date, we would have to convert the class to date.

“total_deaths” variable: Represents the total number of Covid19 deaths in the United States. The class of this variable is numeric.

“new_deaths” variable: Represents the number of Covid19 deaths in the United States per day. The class of this variable is numeric.

“location” variable: Represents the name of the country. The class of this variable is factor.

Summary of data post data cleaning:

```

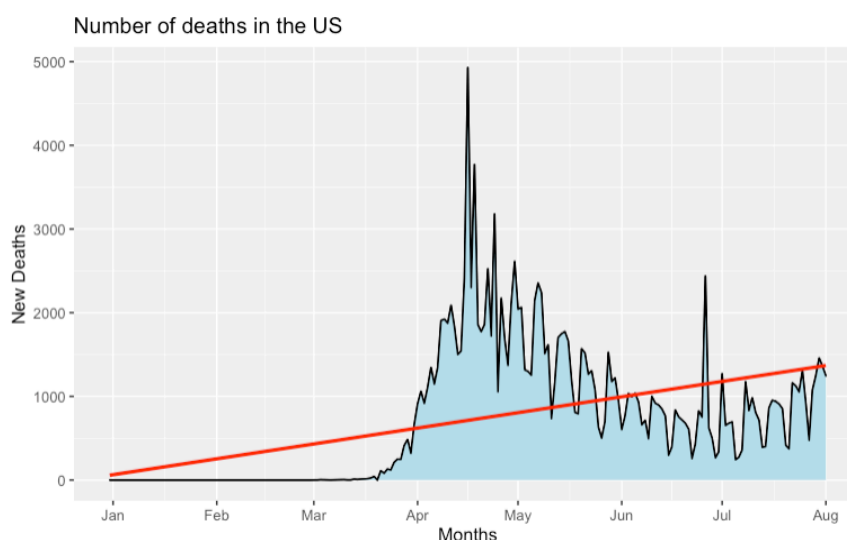
Index          dat_demo
Min.   :2019-12-31   Min.    :  0.0
1st Qu.:2020-02-22   1st Qu.:  0.0
Median :2020-04-16   Median : 500.0
Mean   :2020-04-16   Mean    : 713.1
3rd Qu.:2020-06-08   3rd Qu.:1167.0
Max.   :2020-08-01   Max.    :4928.0

```

When looking at the summary of our data, we can see that our time series index starts from 2019/12/08 and ends in 2020/08/01. Further we can see that the average number of deaths per day is 713 in the United States. The maximum number of deaths reached to 4928 on April 16th.

Time Series Analysis:

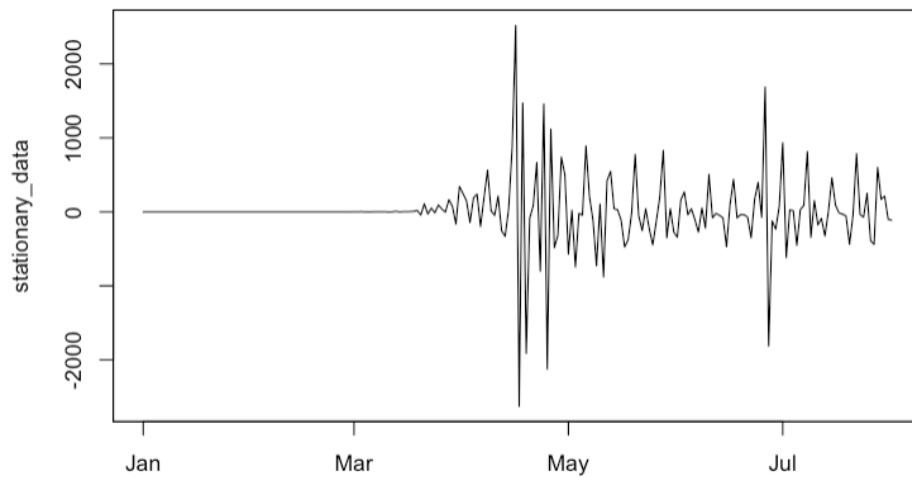
When conducting the time series analysis on our data, we first created the date sequence using the zoo library which is a well-known R package when it comes to this particular statistical technique. After that, we check if our data doesn't violate any assumption for time series analysis. We have to make sure that the data is stationary. In other words, it needs to have equal mean, variance and covariance across the different time intervals.



As mentioned on page 4, the graph above shows that we don't have stationary data. The red line, which is the mean line is increasing. The variance isn't constant as well, since the distance between the data points and the mean line varies across the x axis. As a result, we have to transform our data to be stationary.

Depending on the data there are different techniques to make the data stationary. Data transformations such as logarithms can help stabilize the variance for time series analysis. On the other hand, differencing can help stabilize the mean. Every dataset has its own need and on our data, we just used differencing. Differencing means computing the difference between

consecutive observations. Once we applied differencing to our data, the graph of our data changed to the following:



The data transformation took place and we can see here how the mean is constant now. In order to check if our data is now stationary, we conducted the Augmented Dickey-Fuller Test, which tells us if our data is stationary or not.

H_0 : The null hypothesis is that there is a unit root.

H_A : The alternative hypothesis is that the time series is stationary.

The p-value of our test was 0.01 on a significance level of 0.05. Hence, we rejected the null hypothesis. In other words, our data is now stationary and we can proceed with the analysis.

The next step was to conduct the modeling. The ARIMA model has been selected to do the Time Series Analysis. This particular model forecasts based on its previous values and it has 3 parameters. The parameters are p, d and q which are required to do the forecasting. Before we jump into how to choose the right parameter numbers, we first need to understand what the parameters mean.

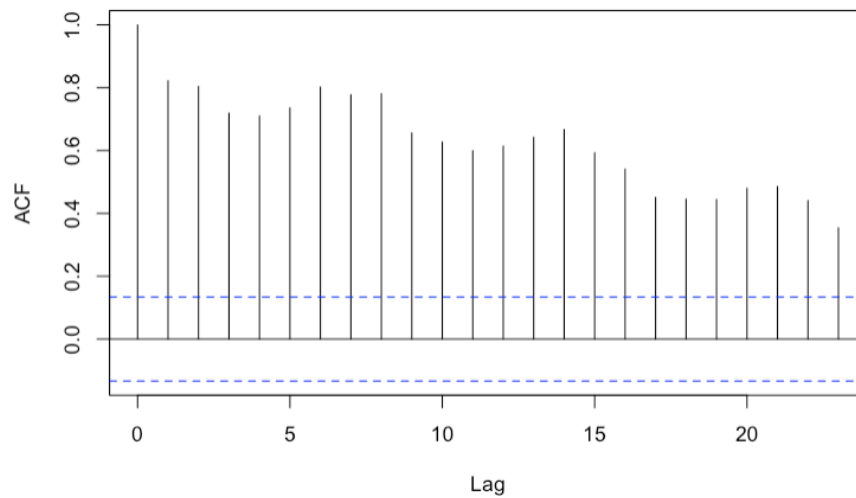
“**p**”: Stands for auto regressive. Auto regression is referring to when past values are being used to predict the future values.

“**d**”: Stands for integration. Integration takes in the amount of differencing that will be used for the time series.

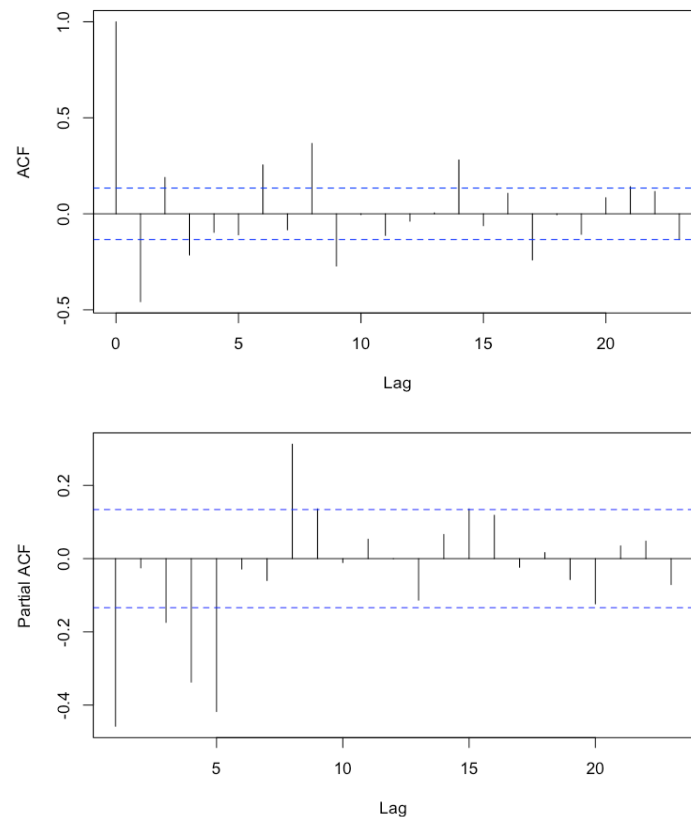
“**q**”: Stands for moving average. The moving average is the average you calculate when you take the different intervals.

The way we select our p,d and q values depend on the ACF and PACF functions, which stand for autocovariance or autocorrelation function and partial autocorrelation function respectively.

When looking at the ACF graph when our data isn't stationary, we get the following:



Here we see that our values are all exceeding the blue line. The goal is to have the values under the blue line, and they should be inverted as well. Now when we look at the ACF and PACF of our stationary data we get the following.



In both ACF and PCF graphs we can now see that most of the lines don't exceed the blue line and also have inverted lines. The q value can be selected when using the acf graph and the p value can be found from the pacf graph. In both cases we pick the number before the first inverted line. Depending on the data you can use that rule and see how the model fits or you can

also use the `auto.arima` function, which returns the best model according to AIC or BIC. Our group decided to proceed with the `auto.arima` function which returned us the following model.

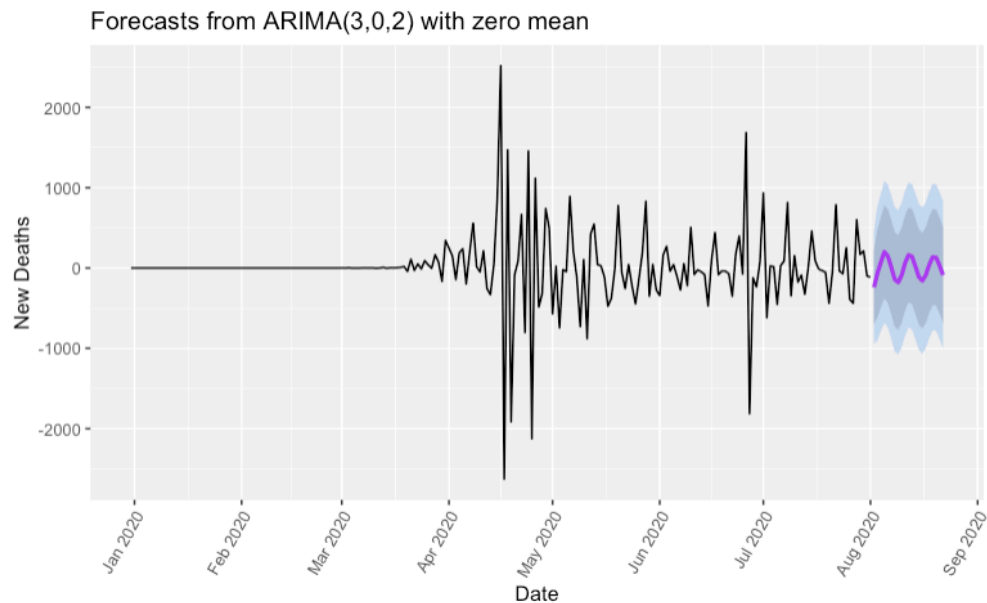
```
Series: stationary_data
ARIMA(3,0,2) with zero mean

Coefficients:
      ar1      ar2      ar3      ma1      ma2
      0.6082 -0.1778 -0.6058 -1.2614  0.8282
s.e.    0.0599  0.0686  0.0563  0.0460  0.0476

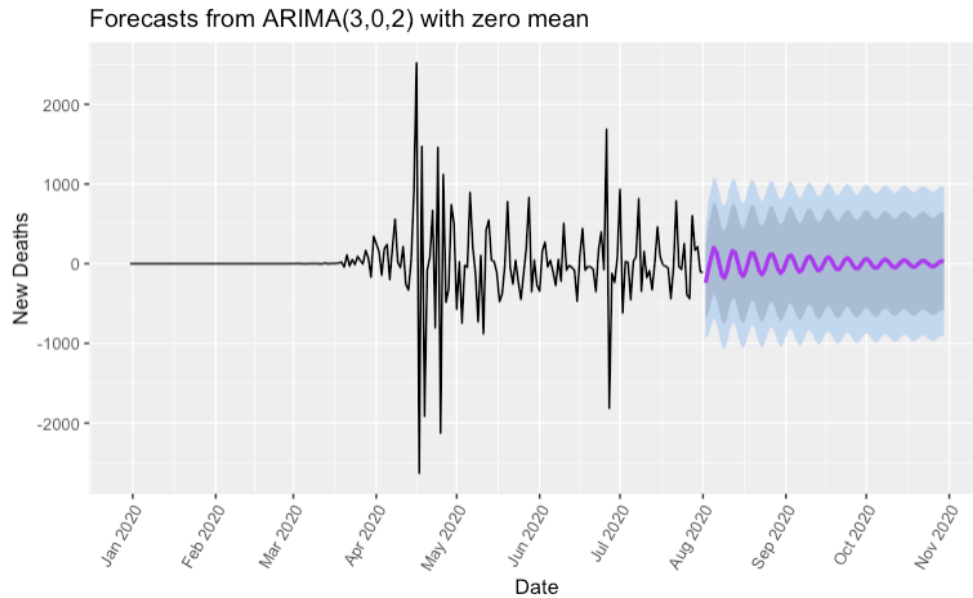
sigma^2 estimated as 126057:  log likelihood=-1558.96
AIC=3129.92  AICc=3130.32  BIC=3150.11
```

The `auto.arima` function selected $p=3$, $d=0$ and $q=2$. When applying this model to predict the number of deaths for the next 21 and 90 days the results have been very close to the prediction that CNN has conducted. As mentioned in the executive summary, our model projected for the next 21 and 90 days 18,589 (Total Deaths 171,903) and 82,653 (Total Deaths 235,967) deaths respectively. CNN projected on August 2nd that about 19,000 people could die in the next 2 weeks in the United States and that 231,000 Americans will lose their lives from Covid19 by November. Below we can see the graphs for the forecast and the results in a table. Note the purple line in the graphs represent the projected values.

Projection graph next 21 days:



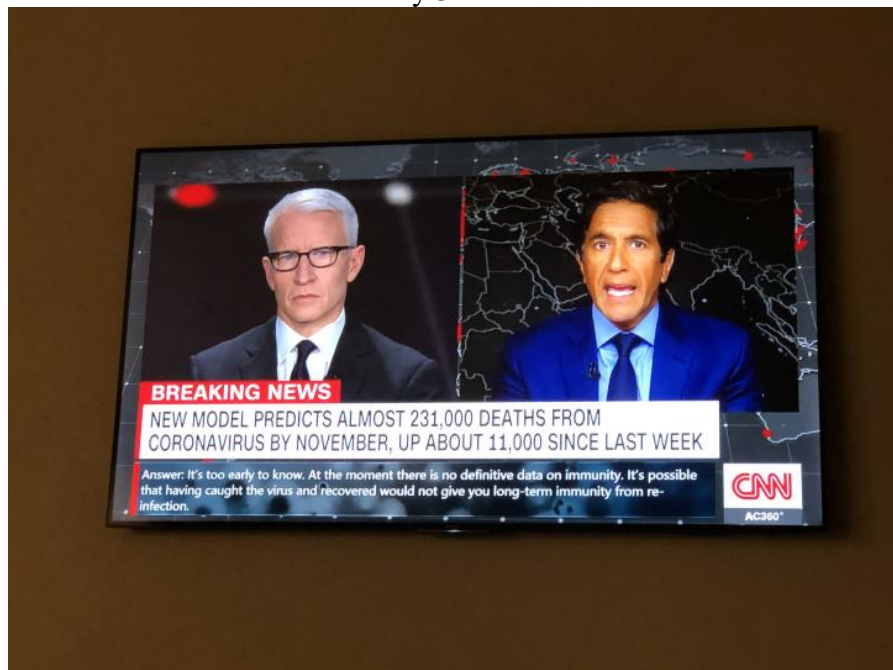
Projection graph next 90 days:



Date:	Our Projection	CNN's Projection
2020/08/01 – 2020/08/21	New Deaths: 18,589	New Deaths: 19,000
2020/08/01 – 2020/10/31	Total Deaths: 235,967	Total Deaths: 231,000

The motivation behind this project was to see how our model would do in comparison to CNN's model. Below we see the announcements that have been made from CNN.

TV Screenshot from July 31st Global Town Hall Show



19,000 more Americans could die from Covid-19 in the next 20 days, CDC composite forecast shows

By Christina Maxouris, [Holly Yan](#) and [Dakin Andone](#), CNN

🕒 Updated 8:18 PM ET, Sun August 2, 2020

5. Conclusion:

Based on our ARIMA model and our forecast, we can see a slight difference but close enough result when comparing it to CNN's forecast. The difference for the 21-day forecast can be explained by CNN rounding their numbers since it's a news headline and make it easier to read for the reader. There is a difference of only $19,000 - 18,589 = 411$ people, which is very close. When it comes to the forecast for the next 90 days, we can see a bigger difference $235,967 - 231,000 = 4,967$. The difference of 4,967, is reasonable. In the end there is no perfect model. As a famous British statistician George E.P Box once said, "*All models are wrong, but some are useful.*" These models are useful, and it was a great experience learning about time series analysis. I am sure there is much more to learn about this statistical technique to get a deeper understanding how we can improve our model, but there is no perfect model. It's very sad seeing the results in regard to the number of lives that will be lost, because of the pandemic in the near future.