# COVID-19 Data Analysis Using R: A Time Series Approach

**Introduction:** Since December 2019, an outbreak of coronavirus disease (COVID-19) has been recorded in Wuhan, China, caused by the severe acute respiratory coronavirus (SARS-CoV-2). Fever, dry cough, and pneumonia are common clinical signs, and they can lead to irreversible respiratory failure and death due to alveolar damage [2]. It is a pandemic; environmental factors such as temperature and relative humidity can affect coronavirus messages by disrupting the virus's survival in transmission routes; evidence for severe acute respiratory coronavirus (SARS-CoV) and Middle East syndrome coronavirus has been found (MERS-CoV) [1]. Coronavirus is spread primarily by respiratory droplets and people to people touch [2]. We looked into the associations between date and new COVID-19 cases using verified evidence.

**Dataset Description:** The Coronavirus disease (Covid-19) is quickly spreading across the world, including in Indonesia [3]. In DKI Jakarta, there has recently been an increase in the volume of information available about new events. We'll attempt to forecast how many points will rise in DKI Jakarta. For this project, data is given, including more than three linked time series variables with over 200 observations (per variable). The data is for the months of March 1, 2020, to July 31, 2020. This dataset includes data from open data sources such as covid19.go.id (pandemic data), projections. A timeline of COVID-19 pandemic incidents in Indonesia, from the national to the regional level, is included in this dataset. The data contains 4,578 observations and 37 columns.

**Project Description:** I solved this problem using RStudio. I add my R Script with the file. Here there are some steps that I follow for my analysis. Install necessary packages with the help of code install. Packages ().

```
library(dplyr)
library(lubridate)
library(forecast)
library(TTR)
library(ggplot2)
library(tseries)
library(gridExtra)
```

**1. Answer:** Import dataset from covid_19_data.csv file. Prepared data for analysis; check null/missing value. I find no missing value in our dataset. Then clean and remove some variables from our dataset as we want to create a sub-sample from our dataset. After make subsample we did some descriptive statistics like mean, standard deviation and variance from our sub-sample data.

```
summary(Dataset)
```

```
##    ï..Date          Location.ISO.Code    Location           New.Cases
##  Length:10694       Length:10694       Length:10694       Min.   :    0.0
##  Class :character   Class :character   Class :character   1st Qu.:    3.0
##  Mode  :character   Mode  :character   Mode  :character   Median :   22.0
##                                                           Mean   :  177.9
##                                                           3rd Qu.:   83.0
##                                                           Max.   :14224.0
##    New.Deaths      New.Recovered    New.Active.Cases   Total.Cases
##  Min.   :  0.000   Min.   :   0.0   Min.   :-1762.0   Min.   :     1
##  1st Qu.:  0.000   1st Qu.:   0.0   1st Qu.:   -2.0   1st Qu.:   246
##  Median :  0.000   Median :  10.0   Median :    2.0   Median :  1766
##  Mean   :  5.018   Mean   : 142.6   Mean   :   30.3   Mean   : 14636
##  3rd Qu.:  2.000   3rd Qu.:  55.0   3rd Qu.:   26.0   3rd Qu.:  7057
##  Max.   :346.000   Max.   :9755.0   Max.   : 5279.0   Max.   :951651
##
##   Total.Deaths     Total.Recovered  Total.Active.Cases Location.Level
##  Min.   :    0.0   Min.   :     0   Min.   :  -128.0   Length:10694
##  1st Qu.:    7.0   1st Qu.:   138   1st Qu.:    81.0   Class :character
##  Median :   53.0   Median :   996   Median :   480.5   Mode  :character
##  Mean   :  498.1   Mean   : 11387   Mean   :  2751.1
##  3rd Qu.:  236.0   3rd Qu.:  4713   3rd Qu.:  1632.8
##  Max.   :27203.0   Max.   :772790   Max.   :151658.0
##
##  City.or.Regency   Province           Country           Continent
##  Mode:logical    Length:10694       Length:10694       Length:10694
```

```
##  NA's:10694      Class :character   Class :character   Class :character
##                  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Island           Time.Zone         Special.Status     Total.Regencies
##  Length:10694       Length:10694       Length:10694       Min.   :  1.00
##  Class :character   Class :character   Class :character   1st Qu.:  7.00
##  Mode  :character   Mode  :character   Mode  :character   Median : 11.00
##                                                           Mean   : 24.56
##                                                           3rd Qu.: 18.00
##                                                           Max.   :416.00
##   Total.Cities    Total.Districts Total.Urban.Villages Total.Rural.Village
##  Min.   : 1.000   Min.   :  44    Min.   :  35.0       Min.   :  275
##  1st Qu.: 1.000   1st Qu.: 103    1st Qu.:  99.0       1st Qu.:  928
##  Median : 2.000   Median : 169    Median : 175.0       Median : 1591
##  Mean   : 5.985   Mean   : 428    Mean   : 518.3       Mean   : 4574
##  3rd Qu.: 4.000   3rd Qu.: 289    3rd Qu.: 332.0       3rd Qu.: 2853
##  Max.   :98.000   Max.   :7230    Max.   :8488.0       Max.   :74953
##  NA's   :300                      NA's   :302          NA's   :327
##    Area..km2.        Population        Population.Density   Longitude
##  Min.   :    664   Min.   :   648407   Min.   :    8.59   Min.   : 96.91
##  1st Qu.:  16787   1st Qu.:  1999539   1st Qu.:   47.79   1st Qu.:106.11
##  Median :  42013   Median :  4216171   Median :  103.84   Median :113.42
##  Mean   : 113132   Mean   : 15801034   Mean   :  764.02   Mean   :113.66
##  3rd Qu.:  75468   3rd Qu.:  9095591   3rd Qu.:  262.70   3rd Qu.:121.20
##  Max.   :1916907   Max.   :265185520   Max.   :16334.31   Max.   :138.70
##
##     Latitude      New.Cases.per.Million Total.Cases.per.Million
##  Min.   :-8.682   Min.   :  0.00        Min.   :    0.01
##  1st Qu.:-6.205   1st Qu.:  0.72        1st Qu.:   66.20
##  Median :-2.462   Median :  5.07        Median :  403.36
##  Mean   :-2.737   Mean   : 12.78        Mean   : 1070.31
##  3rd Qu.: 0.212   3rd Qu.: 14.23        3rd Qu.: 1450.59
##  Max.   : 4.226   Max.   :404.35        Max.   :22056.32
##  New.Deaths.per.Million Total.Deaths.per.Million Case.Fatality.Rate
##  Min.   : 0.0000        Min.   :  0.00           Length:10694
##  1st Qu.: 0.0000        1st Qu.:  1.93           Class :character
##  Median : 0.0000        Median : 11.38           Mode  :character
##  Mean   : 0.2935        Mean   : 29.82
##  3rd Qu.: 0.3800        3rd Qu.: 44.55
##  Max.   :26.2200        Max.   :357.45
##  Case.Recovered.Rate Growth.Factor.of.New.Cases Growth.Factor.of.New.Death
##  Length:10694        Min.   :  0.000            Min.   : 0.000
##  Class :character    1st Qu.:  0.560            1st Qu.: 0.740
##  Mode  :character    Median :  1.000            Median : 1.000
##                      Mean   :  1.456            Mean   : 1.005
##                      3rd Qu.:  1.320            3rd Qu.: 1.000
##                      Max.   :120.500            Max.   :53.000
```

**2.Answer:** If the mean, variance, or timewise covariance of a time series shifts over time, it is said to be nonstationary [2]. Nonstationary time series can't be used in regression models because they can induce spurious regression, or a false association, due to a typical pattern in otherwise unrelated variables. If two or more nonstationary series are cointegrated, that is, if they are in any stationary relationship, they can also be used in a regression model. We take great caution when checking time series for non-stationarity and determining how to convert non-stationary time series so that they can be used in the analysis.

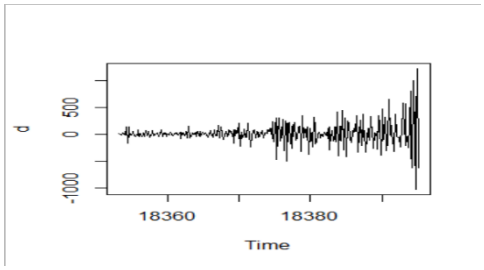



Fig.: Various time series to illustrate non-stationarity

**3.Answer:** And we all remember; two or more series are said to be cointegrated if they are independently integrated (in the time series sense). However, any linear combination of them has a lower order of integration. Individual series that are first-order interconnected are a typical example. In our dataset, there are no cointegrated variables.

**4.Answer: Build an ARIMA model and dynamic regression model:**

```
covid_arima1 <- Arima(y = Covid19_data_time_series, order = c(1,1,1))
covid_arima1
```

```
## Series: Covid19_data_time_series
## ARIMA (1,1,1)
##
## Coefficients:
##          ar1      ma1
##       0.1047 -0.6572
## s.e.  0.0792   0.0511
```

```
## sigma^2 estimated as 32691:  log likelihood=-1951.04
## AIC=3908.08   AICc=3908.16   BIC=3919.14
```

p=1 (1st lag and 1st lag passed the threshold)

D=1 (we only do difference once to get stationary data)

Q=1 (1st lag on ACF plot passed the threshold)

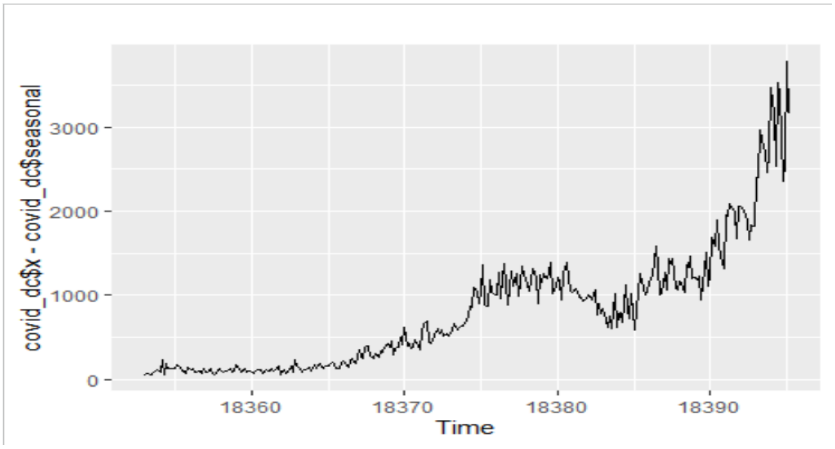
Then I try ARIMA (1,1,1) to see if it's accurate.

Now;Using acf() to compute the sample autocorrelations of the series data.

```
acf(na.omit(data1), lag.max = 1, plot = F)
Autocorrelations of series ‘na.omit(data1)’, by lag
Date
Date       New.Cases
 1.000 (0) 0.870 (0)
 0.990 (1) 0.864 (-1)
New.Cases
 Date       New.Cases
 0.870 (0) 1.000 (0)
 0.851 (1) 0.942 (1)
```

This is evidence that there is mild positive autocorrelation in the growth of data.

**Potential seasonality:**




**5.Answer:** Models that use both the time series to be forecasted and the past of another time series are known as dynamic regression models. It isn't suitable for our information.

```
fit_arima2 <- auto.arima(y=Covid19_data_time_series,d=0 )

fit_arima2

Coefficients:

         ar1     ar2    sma1    sma2       mean

      0.6328 0.3403 0.2827 0.1389 1068.9874

s.e.  0.0565 0.0570 0.0742 0.0601   526.0182

sigma^2 estimated as 36429:  log likelihood=-1973.99

AIC=3959.98   AICc=3960.27   BIC=3982.12
```

**Plot the Time Series:**

```
data1 <- data1 %>%
  subset(Date >= "2020-04-01")
Covid19_data_time_series <- ts(data = data1$New.Cases,
                               start = min(data1$Date),
                               frequency = 7)

Series: Covid19_data_time_series
ARIMA(2,0,0)(0,0,2)[7] with non-zero mean
```

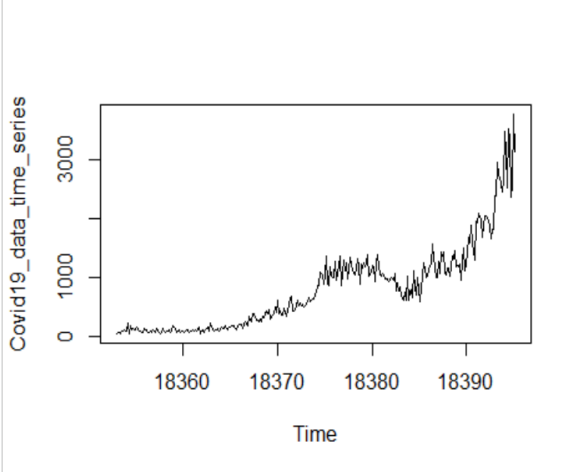



**Fig.: Time Series Plot of Covid 19 Dataset**

## 6. Answer:

**Build a VAR/VECM:**

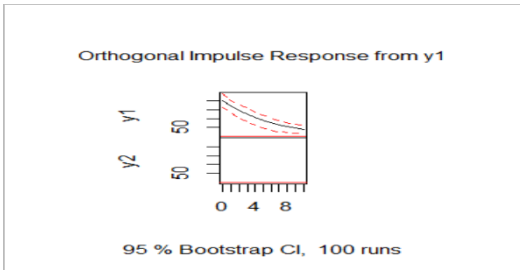
```
varmat <- as.matrix(cbind(data1$New.Cases,data1$Date))
varfit <- VAR(varmat)
varfit
VAR Estimation Results:
=======================

Estimated coefficients for equation y1:
=======================================
Call:
y1 = y1.l1 + y2.l1 + const

        y1.l1         y2.l1         const
 8.530118e-01  1.202560e+00 -2.211964e+04


Estimated coefficients for equation y2:
=======================================
Call:
y2 = y1.l1 + y2.l1 + const

       y1.l1        y2.l1        const
9.423246e-15 1.000000e+00 1.000000e+00
## sigma^2 estimated as 32691:  log likelihood=-1951.04
## AIC=3194.08   AICc=322908.16   BIC=39231
```



**7.Answer:** causality(varfit)$Granger

```
Granger causality H0: y1 do not Granger-cause y2

data:  VAR object varfit

F-Test = 0.97371, df1 = 1, df2 = 584, p-value = 0.3242
```

**Plot the detrended time series, generate variance decompositions and autocorrelation:**

```
adf.test(Covid19_data_time_series)

## Warning in adf.test(Covid19_data_time_series): p-value greater than printe
d p-
## value

##
## Augmented Dickey-Fuller Test
##
## data:  Covid19_data_time_series
## Dickey-Fuller = 1.0135, Lag order = 6, p-value = 0.99
## alternative hypothesis: stationary
```

**Impulse response functions:**

```
Impulse response coefficients
$y1
             y1           y2
 [1,] 200.02052 1.610648e-13
 [2,] 170.61987 2.045907e-12
 [3,] 145.54077 3.653700e-12
 [4,] 124.14800 5.025167e-12
 [5,] 105.89971 6.195044e-12
 [6,] 90.33370 7.192963e-12
 [7,] 77.05572 8.044200e-12
 [8,] 65.72944 8.770315e-12
 [9,] 56.06799 9.389699e-12
[10,] 47.82666 9.918042e-12
[11,] 40.79671 1.036872e-11
```

**Estimate the trend for our dataset and addressing seasonality:**

```
Covid19_data_time_series <- ts(data = data1$New.Cases,start =
min(data1$Date),frequency = 7)
```

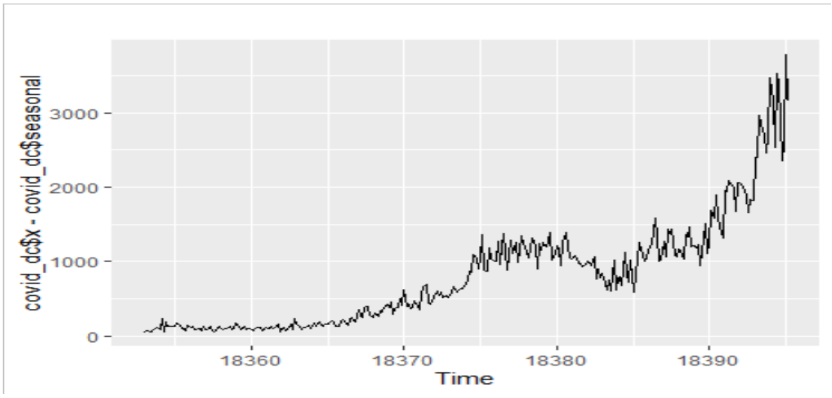



**Fig.: Trend Plot**

**Report:** From the plot and information above, we can see that the data's seasonal pattern is random or **no seasonal**. The trend patterns of data are increasing.

**Take logs and/or difference:**

```
diff(Covid19_data_time_series, lag = 1) %>% adf.test()
## Warning in adf.test(.): p-value smaller than printed p-value
##
## Augmented Dickey-Fuller Test
##
## data:.
## Dickey-Fuller = -9.0154, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

**Comment:** From adf.test (Covid19_data_time_series) we can see that; p-value = 0.2044, which means p-value is < 0.05 (not stationary). We'll use differencing to try to make the data more consistent. Then we plot detrended time series plot.

**8.Answer: Compare in-sample fit and Compare the out-of-sample forecasting performance:**

```
Model1:accuracy(covid_arima1)
##                     ME     RMSE      MAE         MPE     MAPE      MASE
ACF1
## Training set 27.139 179.8873 108.791 -0.8174189 17.23132 0.7189548 -0.0062
96387
```

**Model2: accuracy(fit_arima2)**

| | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|---|
| Training set | 5.279865 | 189.2442 | 116.0547 | -9.624561 | 22.34812 | 0.7669571 | -0.1401014 |

**Model3:**

```
Model1: accuracy(covid_arima3)
                    ME     RMSE      MAE         MPE     MAPE      MASE         C
## Training set 277.1300 169.8558 208.761 -0.844174189 19.23132 0.7189548 -0.
00296387
```

**9.Answer:** Testing for serial correlation: A common goal of time series analysis is extrapolating past behavior into the future. The forecasting procedures include random walks, moving averages, trend models, simple, linear, quadratic, and seasonal exponential smoothing, and ARIMA parametric time series models.

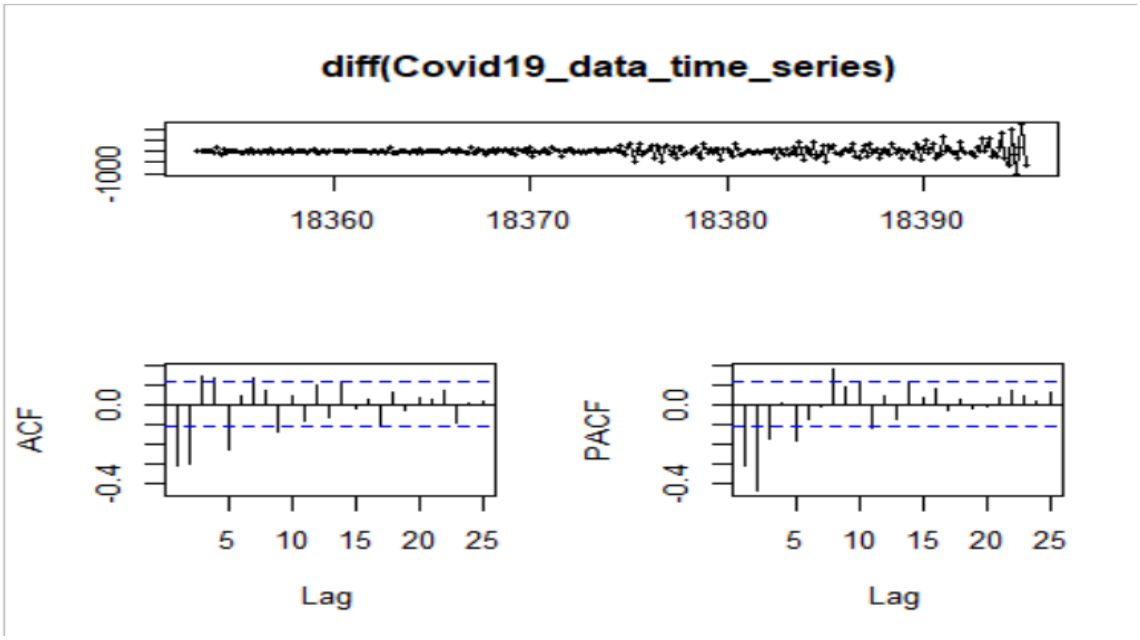
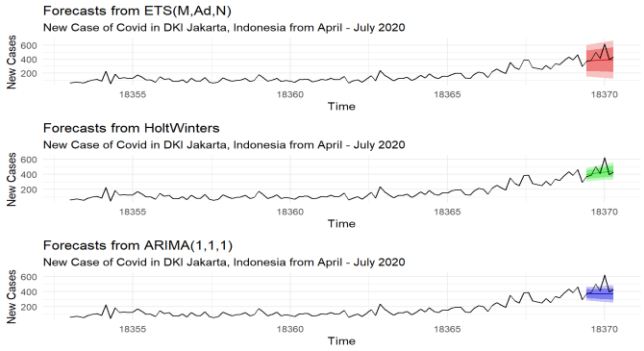



**Fig.: Detrended, ACF, PACF plot**

**10.Answer:** SARIMA, ARIMA, models, including exponential smoothing, are some of the most commonly used methods for time series forecasting. The term "Auto-Regressive Integrated Moving Average" is an acronym for "Auto-Regressive Integrated Moving Average." Forecasts in an Auto-Regressive model equate to a linear combination of the variable's past values. The Arima model is going well comparing out-of-sample RMSPE , ME ,RMSE ,MAE , MPE ,MAPE, MASE etc.

**11.Answer:** When this happens, the better is and try to figure out which forecast is better (or best). A linear fusion of the two sets of data will yield the cumulative prediction... If a time series model can be precisely defined, then it. The Arima model is the perfect fit for our dataset.




**The End**