
Sogang University

# Constructing a Forecasting model for Predicting the Level of Ozone in Air

What factors impact the level of ozone in air & analyzing the model in depth using R

Name: Esrat Maria

Student ID: 120190185

Name: Safdar Jamil

Student ID: 220180021

Time Series Data Analysis and Forecasting

Professor Myung Suk Kim

Due Date: June 18th, 2020

***Problem Motivation:***

Air is a natural resource and is available abundantly. It is an essential element of nature that support life on earth. Air is equally important for living organisms for their survival just like water. Air is very useful and has many applications. Uses of air are as like, sustaining life and growth, combustion, maintaining temperature, supplying energy, photosynthesis and many more.

Air is a mixture of gases like oxygen, nitrogen and carbon dioxide in higher proportions and much smaller levels of argon, water vapor, and other gases. Living things live and breathe in the clear gas. It has an indefinite shape and volume. It has no color or smell. Air is a matter so it has mass and weight. The weight of air creates atmospheric pressure. The Components of air are 78% nitrogen, 21% oxygen, 0.9% argon, 0.04% carbon dioxide, and very small amounts of other gases and water vapor. In aerobic respiration, animals need to breathe the oxygen in the air. In breathing, we inhale oxygen which reaches the lungs and from lungs blood capillaries absorb oxygen and carbon dioxide is breathed out into the air. Plants need carbon dioxide for photosynthesis.

Air pollution is a problem for all of us. However, some groups of people are especially sensitive to common air pollutants such as particulates and ground-level ozone. Sensitive populations include children, older adults, people who are active outdoors, and people with heart or lung diseases, such as asthma. If one is sensitive to air pollution, one need to be aware of steps they can take to protect their health[1].

Along with harming human health, air pollution can cause a variety of environmental effects as well:

- Acid rain[2]
- Eutrophication
- Haze[3]
- Effects on wildlife
- Ozone depletion
- Crop and forest damage
- Global climate change [4] etc.

When poor air quality affects our health it also affects our economy due to increasing medical costs and lost productivity when people are unable to work. So many factors are responsible for triggering Air pollution. Air pollution is caused by:

- human activities, e.g.:
    - burning of fuels for home heating
    - industrial processes
    - vehicle exhausts (particularly diesel)
    - road dust

[1] http://www.mass.gov/dep
[2] http://www.epa.gov/acidrain/
[3] http://www.epa.gov/oar/visibility/
[4] http://www.epa.gov/globalwarming/

    - quarrying
- natural sources
    - (e.g., wind-blown dust, pollen, sea salt and volcanic eruptions).

So many factors are polluting the air that we breathe in every day. Alongside many poisonous gases such as CO, NOx, titanium, sulfur dioxide, methane etc. the effect of ozone in the air is really severe for both humans and animals. Ground-level or "bad" ozone is not emitted directly into the air, but is created by chemical reactions between oxides of nitrogen (NOx) and volatile organic compounds (VOC) in the presence of sunlight. Emissions from industrial facilities and electric utilities, motor vehicle exhaust, gasoline vapors, and chemical solvents are some of the major sources of NOx and VOC.

If we take look at the image below we can see that how the level of Ozone in air has increased throughout the decades.

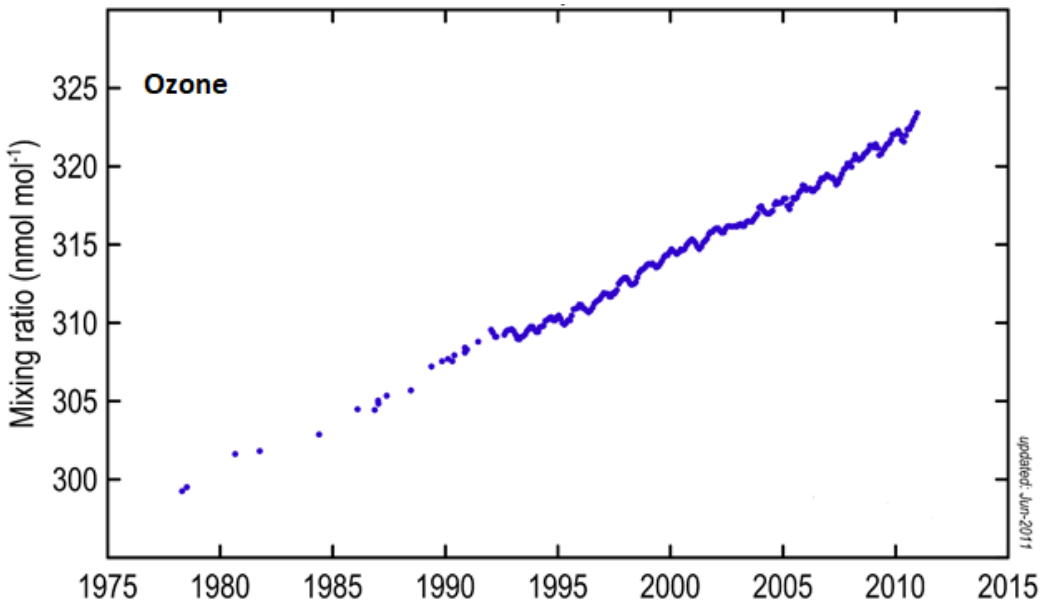



**Fig (1): Increase in the level of ozone in air**

**How the level of Ozone effecting the environment and human life?**

Breathing ozone can trigger a variety of health problems including chest pain, coughing, throat irritation, and congestion. It can worsen bronchitis, emphysema, and asthma. Ground-level ozone also can reduce lung function and inflame the linings of the lungs. Repeated exposure may permanently scar lung tissue.

Ground-level ozone also damages vegetation and ecosystems. In the United States alone, ozone is responsible for an estimated $500 million in reduced crop production each year. Breathing ozone can trigger a variety of health problems including chest pain, coughing, throat irritation, and congestion. It can

worsen bronchitis, emphysema, and asthma. Ground-level ozone also can reduce lung function and inflame the linings of the lungs. Repeated exposure may permanently scar lung tissue.

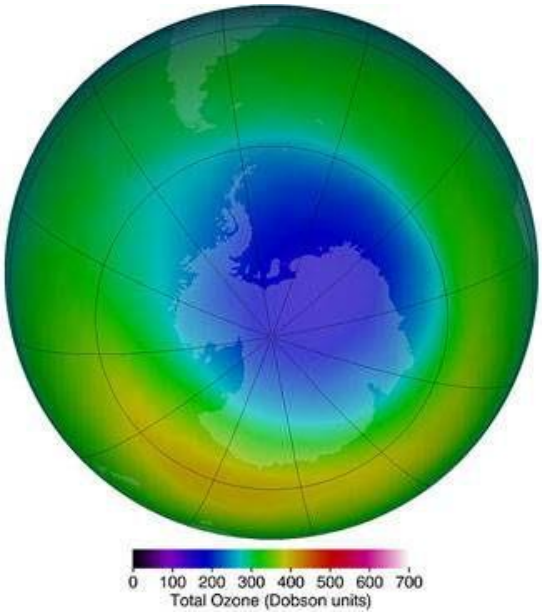



**Fig (2): Arctic stratospheric ozone reached its record low level of 205 Dobson units, shown in blue and turquoise, on March 12, 2020[5]**

Above image is an analysis of satellite observations which show that **ozone levels** reached their lowest point on March 12 at 205 Dobson units. While such low **levels** are rare, they are not unprecedented. Similar low **ozone levels** occurred in the upper atmosphere, or stratosphere, in 1997 and 2011.

**Environmental Effects**

Ground-level ozone can have detrimental effects on plants and ecosystems. These effects include:

- interfering with the ability of sensitive plants to produce and store food, making them more susceptible to certain diseases, insects, other pollutants, competition and harsh weather;
- damaging the leaves of trees and other plants, negatively impacting the appearance of urban vegetation, as well as vegetation in national parks and recreation areas; and
- Reducing forest growth and crop yields, potentially impacting species diversity in ecosystems.

***Literature Review:***

The objective of many present study [6]is to investigate, through measurements and model simulation; average level of ozone and of its precursor gases like $NO_x$ and CO and to delineate the significant mechanisms that ensures the level of ozone over the region.

[5] https://www.nasa.gov/feature/goddard/2020/nasa-reports-arctic-stratospheric-ozone-depletion-hit-record-low-in-march

[6] Sink mechanism for significantly low level of ozone over the Arabian sea https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2008JD011256

The study on ozone effects on vegetation [7]says that the critical levels set for ozone may have considerable policy significance, but the arguments on which they are based, and the uncertainties inherent in setting the values have not, to date, been systematically set out in the open scientific literature. However there are some limitations and uncertainties in the methods used to define the critical levels.

In recent days it has been showed by NASA that the overall level of ozone in the atmosphere has been reduced in some extent. Stratospheric ozone is nicknamed "good" ozone, because the ozone layer plays a vital role in absorbing ultraviolet (UV-B) rays that are harmful to living beings on the earth. Since direct contact with ozone at the ground level can cause damages to living cells, organs, and species including humans, animals, and plants, tropospheric or ground-level ozone is nicknamed "bad" ozone[8].

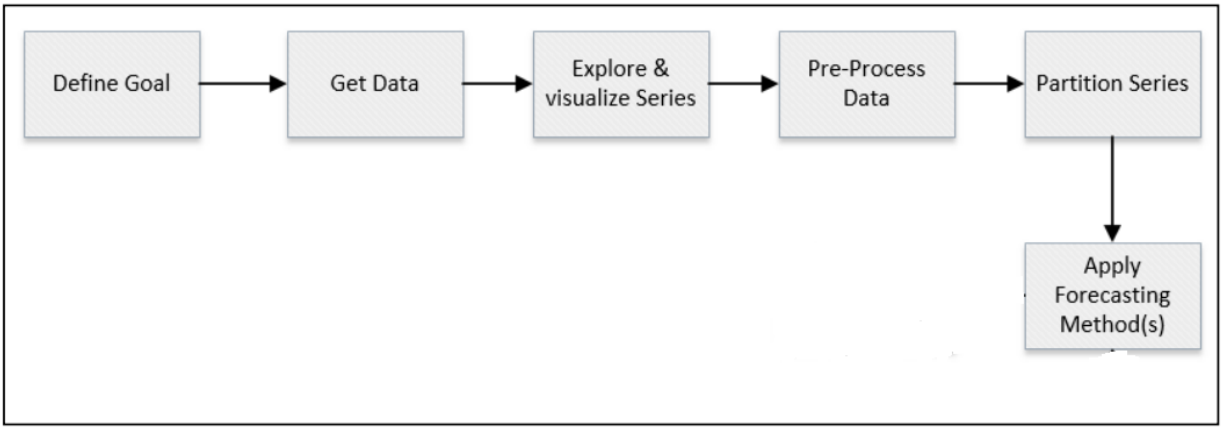
The reports on this topic were found on the internet when searched extensively. Most of them reported very elaborately the effects of ozone on environment and health. Our goal is to make a prediction model considering the level of ozone in the air and report how the future look like when measuring ozone level in our atmosphere.

***Statement of Research Objectives:***

Which factors are affecting the rise of ozone level in the air? How has this behavior been over the past few years? The objective of this research is to make an ARIMA model to predict seasonality of our dataset and to make a forecast model to determine the behavior of level of ozone in the air.

We used R Language and R Studio to integrate all the datasets then we performed exploratory data analysis, data preparation and performed Autoregressive Integrated Moving Average. In this study, we focus only on Ozone pollutants.

The steps involved to make our forecasting model is like below:




Driven by a major improvement and the unique challenges of air quality estimates in the past two decades, this study aims to show exploratory data analysis and evaluate the performance of the Autoregressive Integrated Moving Average model.

[7] Critical levels for ozone effects on vegetation in Europe- J. Fuhrer, L. Skarby and M. R. Ashmore

[8] Ozone Pollution: A Major Health Hazard Worldwide - Junfeng (Jim) Zhang, Yongjie Wei and Zhangfu Fang

***Description of Data:***

The data were obtained from the New York State Department of Conservation (ozone data) and the National Weather Service (meteorological data). It has been modified slightly and obtained from here[9]. This data set has daily air quality measurements in New York from January 1995 to September 2007 over a period of 156 months. In this research we will analyze one of the attributes, $O_3$. The dataset contains 153 instances of daily averaged responses spreading from Jan 1995 to Sep 2007.

***Data Set Exploration:***

This should include summary statistics, means, medians, quartiles, or any other relevant information about the dataset.

***Data Set Information:***

A data frame with 153 observations on 6 variables.

Below is the peek of our dataset that we are working with:

| Ozone | Month | Year | Solar.R | Wind | Temp | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 41 | Jan | 1995 | 190 | 7.4 | 67 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 36 | Feb | 1995 | 118 | 8 | 72 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | Mar | 1995 | 149 | 12.6 | 74 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | Apr | 1995 | 313 | 11.5 | 62 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | May | 1995 | 0 | 14.3 | 56 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | Jun | 1995 | 0 | 14.9 | 66 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 23 | Jul | 1995 | 299 | 8.6 | 65 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 19 | Aug | 1995 | 99 | 13.8 | 59 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 8 | Sep | 1995 | 19 | 20.1 | 61 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | Oct | 1995 | 194 | 8.6 | 69 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 7 | Nov | 1995 | 0 | 6.9 | 74 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 16 | Dec | 1995 | 256 | 9.7 | 69 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | Jan | 1996 | 290 | 9.2 | 66 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | Feb | 1996 | 274 | 10.9 | 68 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | Mar | 1996 | 65 | 13.2 | 58 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | Apr | 1996 | 334 | 11.5 | 64 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 34 | May | 1996 | 307 | 12 | 66 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | Jun | 1996 | 78 | 18.4 | 57 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 30 | Jul | 1996 | 322 | 11.5 | 68 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 11 | Aug | 1996 | 44 | 9.7 | 62 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 1 | Sep | 1996 | 8 | 9.7 | 59 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 11 | Oct | 1996 | 320 | 16.6 | 73 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 4 | Nov | 1996 | 25 | 9.7 | 61 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 32 | Dec | 1996 | 92 | 12 | 61 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | Jan | 1997 | 66 | 16.6 | 57 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | Feb | 1997 | 266 | 14.9 | 58 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | Mar | 1997 | 0 | 8 | 57 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 23 | Apr | 1997 | 13 | 12 | 67 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 45 | May | 1997 | 252 | 14.9 | 81 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 115 | Jun | 1997 | 223 | 5.7 | 79 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 37 | Jul | 1997 | 279 | 7.4 | 76 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | Aug | 1997 | 286 | 8.6 | 78 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | Sep | 1997 | 287 | 9.7 | 74 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | Oct | 1997 | 242 | 16.1 | 67 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | Nov | 1997 | 186 | 9.2 | 84 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | Dec | 1997 | 220 | 8.6 | 85 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | Jan | 1998 | 264 | 14.3 | 79 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 29 | Feb | 1998 | 127 | 9.7 | 82 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Once we import the dataset in R the structure of our dataset looks like below:

[9] http://vincentarelbundock.github.io/Rdatasets/

```
> str(air)
'data.frame':   153 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 0 28 23 19 8 0 ...
 $ Month  : chr  "Jan" "Feb" "Mar" "Apr" ...
 $ Year   : int  1995 1995 1995 1995 1995 1995 1995 1995 1995 1995 ...
 $ Solar.R: int  190 118 149 313 0 0 299 99 19 194 ...
 $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
>
```

The dimension and structure of our dataset is like below:

```
> class(air)
[1] "data.frame"
>
> dim(air)
[1] 153   6
```

We choose to work with ***all*** rows of records to complete this project. We figured that this amount of record is enough to make our prediction and also it makes the whole process a lot smoother and faster.

***Dataset characteristics:***

| Variable | Type | Unit | Description | Detail |
|---|---|---|---|---|
| **Ozone** | numeric | parts per billion | mean Ozone concentration | Roosevelt island |
| **Solar.R** | numeric | | Solar radiation | central park |
| **Wind** | numeric | miles per hour | average wind speed | LaGuardia airport |
| **Temp** | numeric | Fahrenheit | maximum daily temperature | LaGuardia airport |
| **Month** | numeric | | Month of observation | 5 values |
| **Year** | numeric | | All years 1995 - 2007 | 153 days total |

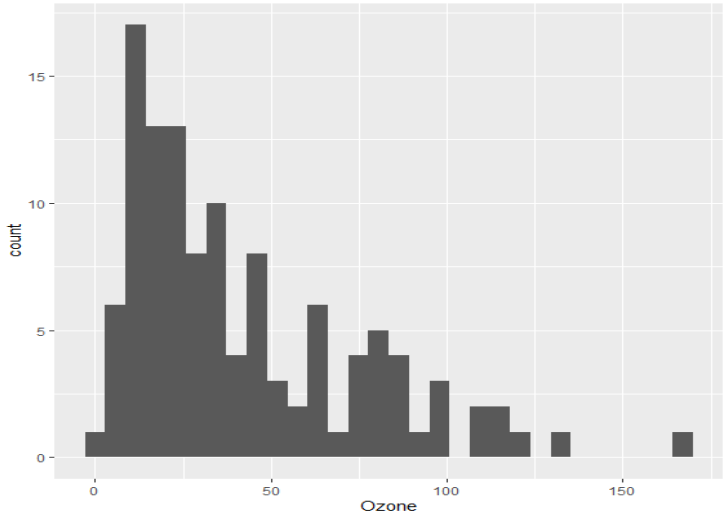
***Attribute Information:***

- **Ozone**: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
- **Solar.R**: Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park
- **Wind**: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
- **Temp**: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

From this dataset we can see the factors involved to analyze the air that surround us. In this work we will analyze how the level of Ozone ($O_3$) has increased in the air and how is the future case scenario looking in this regard.

Now we will analyze our dataset by plotting different diagrams to know clearly what kind of data we are working with.

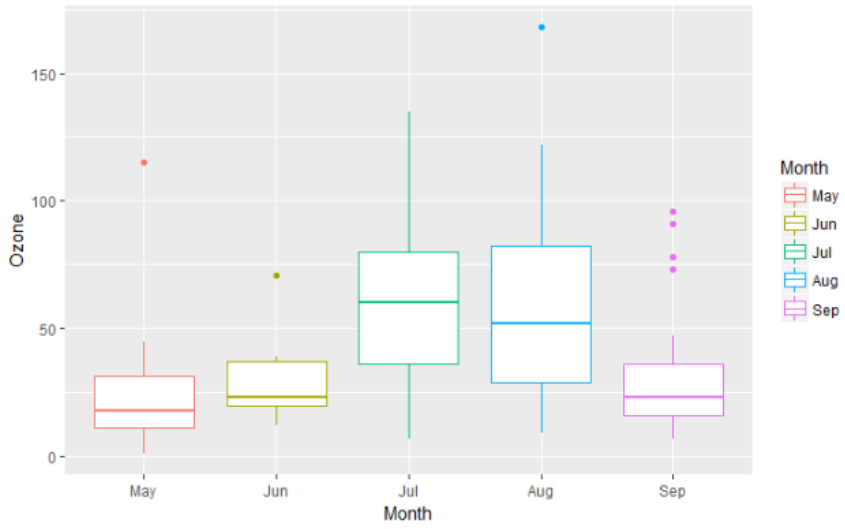
Now we will plot a histogram to count the overall measure of Ozone in the air in correspond to our dataset. The plot is like below:




From above diagram we can see that our histogram is asymmetric.

Now if we want to see the measure of ozone level in the air from our working dataset the plot is like below and we write the following in R:

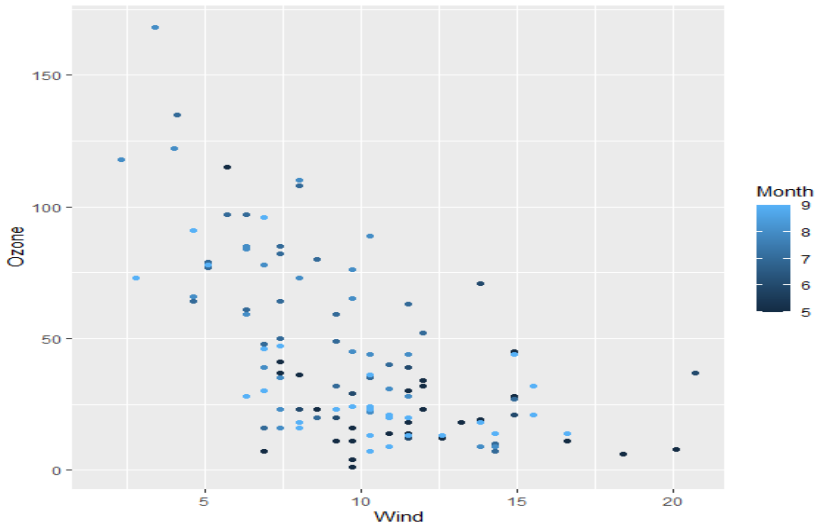
$qplot(Month, Ozone, data = air, geom = "boxplot", color = Month)$




Now if we wat to check the correlation among the attributes then we can see that Ozone and the attribute Wind has a very strong correlation. We write the following in R:

$qplot(Wind, Ozone, data = airquality, color = Month, geom = 'point')$

The output is like below:




## ***Applied Methodology:***

To forecast the level of ozone in the air we first make a model based on trend and seasonality. The months are considered as binary observations and the month of December acts as a dummy variable. In our model the trend ranges from 1 to 154.

$$\boldsymbol{level\ of\ ozone = \beta_0 + \beta_1 Solar.R + \beta_2 Wind + \beta_3 Temp + \beta_4 Jan + \cdots + \beta_{14} Nov + \varepsilon}$$

Using Excel the regression output is like below:

| SUMMARY OUTPUT | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| *Regression Statistics* | | | | | | | | |
| Multiple R | 0.644971467 | | | | | | | |
| R Square | 0.415988194 | | | | | | | |
| Adjusted R Square | 0.352045295 | | | | | | | |
| Standard Error | 27.30773909 | | | | | | | |
| Observations | 153 | | | | | | | |
| | | | | | | | | |
| ANOVA | | | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| Regression | 15 | 72769.84246 | 4851.32283 | 6.505619911 | 2.47099E-10 | | | |
| Residual | 137 | 102162.6281 | 745.7126141 | | | | | |
| Total | 152 | 174932.4706 | | | | | | |
| | | | | | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -1546.084651 | 1316.018754 | -1.174819619 | 0.242104796 | -4148.421211 | 1056.25191 | -4148.421211 | 1056.251909 |
| Year | 0.757318953 | 0.662003355 | 1.143980536 | 0.254626589 | -0.551747144 | 2.06638505 | -0.551747144 | 2.066385051 |
| Solar.R | 0.047298777 | 0.025625292 | 1.845784888 | 0.06708203 | -0.003373475 | 0.09797103 | -0.003373475 | 0.097971029 |
| Wind | -3.215570719 | 0.731304959 | -4.397031195 | 2.18524E-05 | -4.661675971 | -1.76946547 | -4.661675971 | -1.769465467 |
| Temp | 1.088492642 | 0.304417641 | 3.575655598 | 0.000483246 | 0.486527697 | 1.69045759 | 0.486527697 | 1.690457587 |
| Jan | 3.875007227 | 11.0816969 | 0.349676341 | 0.72711856 | -18.0382861 | 25.7883006 | -18.0382861 | 25.78830056 |
| Feb | 14.85848445 | 10.99175743 | 1.351784239 | 0.178671953 | -6.876959765 | 36.5939287 | -6.876959765 | 36.59392866 |
| Mar | -1.120989747 | 10.96935114 | -0.102192895 | 0.918752992 | -22.81212706 | 20.5701476 | -22.81212706 | 20.57014757 |
| Apr | 10.46360971 | 10.94535606 | 0.955986233 | 0.340763065 | -11.18007899 | 32.1072984 | -11.18007899 | 32.1072984 |
| May | 5.97460491 | 10.99831524 | 0.543229102 | 0.58785485 | -15.77380692 | 27.7230167 | -15.77380692 | 27.72301674 |
| Jun | -1.149937414 | 10.9838645 | -0.104693336 | 0.916772269 | -22.86977391 | 20.5698991 | -22.86977391 | 20.56989908 |
| Jul | -7.457739803 | 11.22976935 | -0.66410445 | 0.507739605 | -29.66383622 | 14.7483566 | -29.66383622 | 14.74835662 |
| Aug | 1.075624284 | 11.06480847 | 0.097211288 | 0.922700675 | -20.80427334 | 22.9555219 | -20.80427334 | 22.95552191 |
| Sep | 9.146550054 | 10.95266327 | 0.835098261 | 0.405116336 | -12.51158815 | 30.8046883 | -12.51158815 | 30.80468826 |
| Oct | -0.934009994 | 11.17814775 | -0.083556777 | 0.933530802 | -23.03802825 | 21.1700083 | -23.03802825 | 21.17000826 |
| Nov | -17.65456735 | 11.26363699 | -1.567394915 | 0.119329299 | -39.9276347 | 4.6185 | -39.9276347 | 4.618500001 |

From the above diagram we can see that the adjusted $R^2$ value is quite low, 0.352. It shows that around 65% of our model is unexplained. Looking at the p-values which should be less than 0.05 we can say that only significant ones are $Wind$ and $Temp$. Therefore we remove the non-significant variables. The model now looks like below:

$$\boldsymbol{level\ of\ ozone = \beta_0 + \beta_1 Wind + \beta_2 Temp + \varepsilon}$$

Now our regression output is like below:

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.580053729 |
| R Square | 0.336462329 |
| Adjusted R Square | 0.32761516 |
| Standard Error | 27.81777419 |
| Observations | 153 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 2 | 58858.18646 | 29429.09323 | 38.03050795 | 4.36434E-14 |
| Residual | 150 | 116074.2841 | 773.8285608 | | |
| Total | 152 | 174932.4706 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -43.65730778 | 25.09826924 | -1.739454915 | 0.084006103 | -93.24911177 | 5.9344962 | -93.24911177 | 5.934496198 |
| Wind | -2.91803721 | 0.7204541 | -4.05027497 | 8.17757E-05 | -4.341586308 | -1.49448811 | -4.341586308 | -1.494488112 |
| Temp | 1.343756108 | 0.268155144 | 5.011114417 | 1.50493E-06 | 0.813906929 | 1.87360529 | 0.813906929 | 1.873605286 |

Now the adjusted $R^2$ value is still quite low, 0.327. This kind of behavior may be problematic in terms of making precise prediction.

$$\boldsymbol{level\ of\ ozone = -43.65730778 - 2.91803721(Wind) + 1.343756108(Temp)}$$

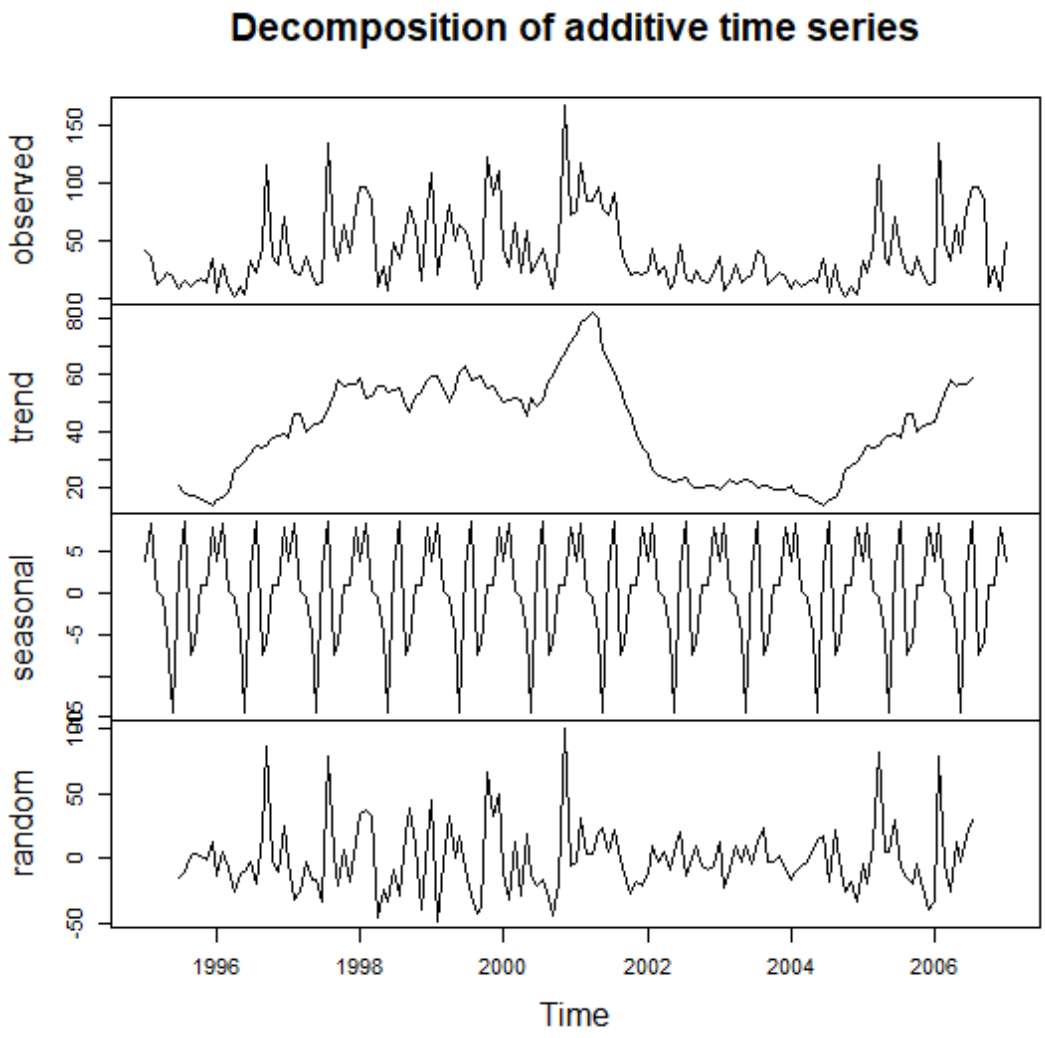
***Decomposition:***

Now our aim is to decompose the model using R to find a trend and seasonality. We first have to omit all the NA values in our dataset. We replace all the NA values with zeros in our dataset. Since we have data ranging from 1995 to 2007 we write the following lines of code in R:

$$trend\_(seasonality\_check) <- ts(airquality\$Ozone, start = c(1995, 3), end = c(2007, 6), frequency = 52/4)$$

$$plot(decompose(trend\_seasonality\_check))$$

The output is like below:




From the above diagram we can see that the ozone level has both an increasing and decreasing pattern. The pattern repeats with higher and lower values over time.

***Dickey-Fuller test:***

To determine if the data is stationary or not we did a Dickey-Fuller test for our time series data. For Dickey-Fuller test we know:

$$H_0: The\ model\ is\ non-stationary$$

$$H_1: The\ model\ is\ stationar$$

```
> adf.test(airquality$Ozone)

        Augmented Dickey-Fuller Test

data:  airquality$Ozone
Dickey-Fuller = -2.9119, Lag order = 4, p-value = 0.1982
alternative hypothesis: stationary
```

As we can see, the p-value of our data is larger than 0.05 so we couldn't reject $H0$. It means the data is not stationary, so we have to make our data stationary before doing any further analysis. So, we did a transformation by taking log to our data and performed the Dickey-Fuller test again.

```
> adf.test(log(airquality$Ozone))

        Augmented Dickey-Fuller Test

data:  log(airquality$Ozone)
Dickey-Fuller = -3.0103, Lag order = 4, p-value = 0.1574
alternative hypothesis: stationary
```

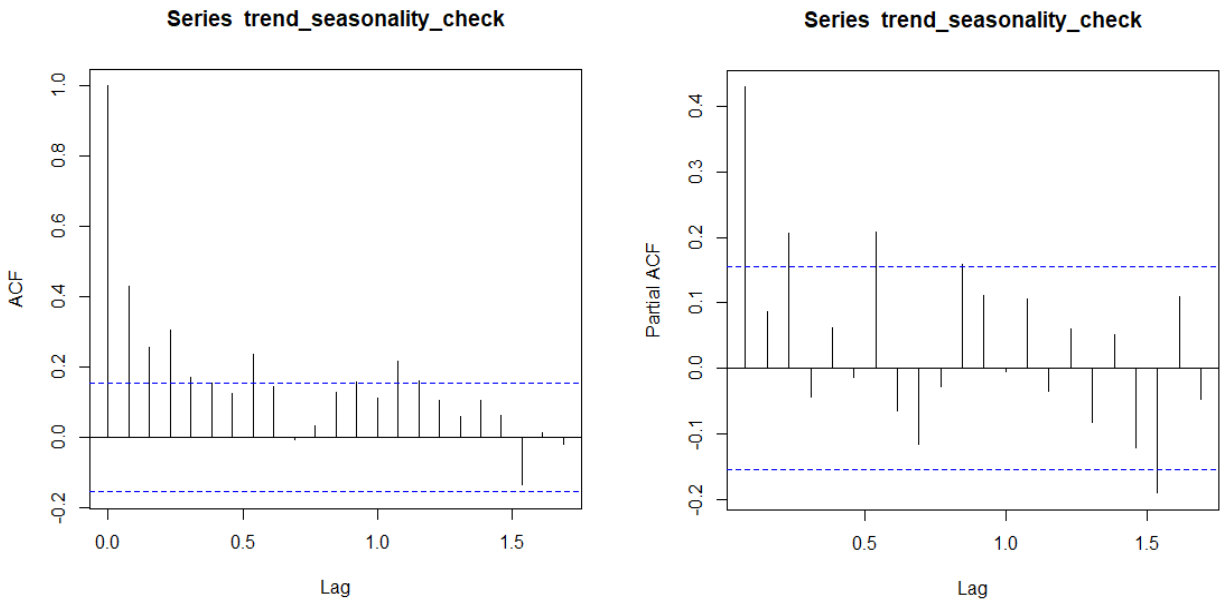
From above we can see that the p-value is still bigger than 0.05. We decided to take not only the log but also adding the difference effect.

```
> adf.test(diff(log(airquality$Ozone)))

        Augmented Dickey-Fuller Test

data:  diff(log(airquality$Ozone))
Dickey-Fuller = -6.2235, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Now we can see that the p-value is smaller than 0.05 and we can reject $H_0$. We can now say that our data is stationary.

***ACF and PACF test:***

The plot is like below. The plot on the left size is (ACF) and the plot on the right side is (PACF).



***ARIMA Test:***

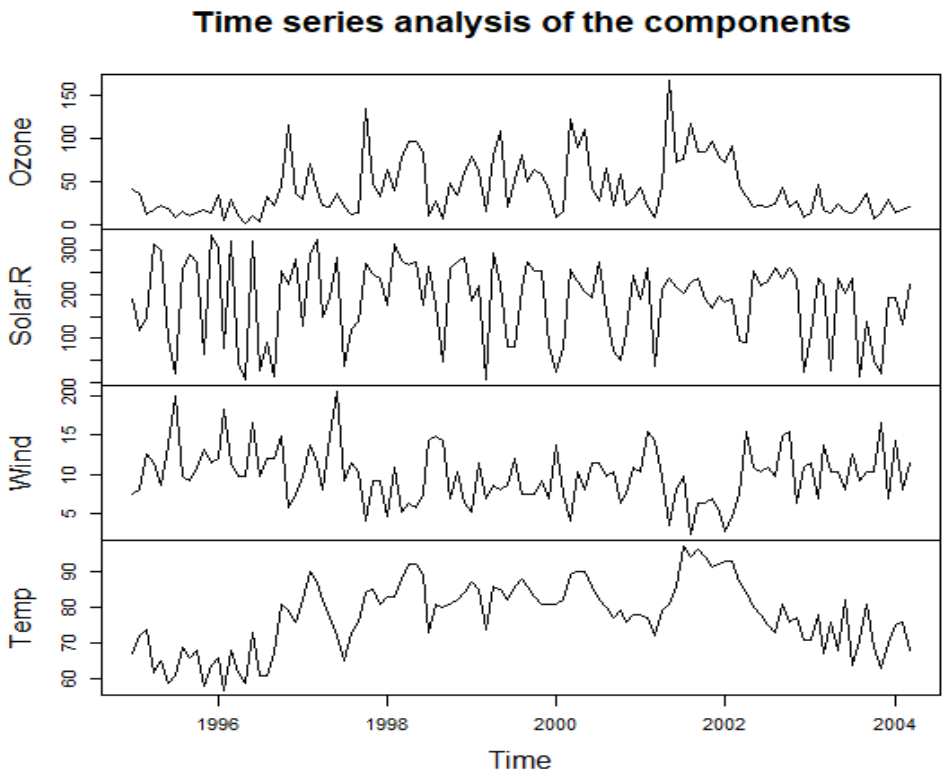
To transform our model into a time series model we write the following lines of code in R:

ptime <- ts(airquality, frequency=12, start=c(1995,1))

The variable looks like below:

```
> head(ptime)
         Ozone Solar.R Wind Temp Month Day
Jan 1995    41     190  7.4   67     5   1
Feb 1995    36     118  8.0   72     5   2
Mar 1995    12     149 12.6   74     5   3
Apr 1995    18     313 11.5   62     5   4
May 1995    23     299  8.6   65     5   7
Jun 1995    19      99 13.8   59     5   8
```

If we plot the above variable we get an output like below:




We will make 4 models for doing the ARIMA analysis.

Model 1: ARIMA(2, 0, 2)

Model 2: ARIMA(1, 0, 0)

Model 3: ARIMA(0, 1, 1)

Model 4: ARIMA(0, 1, 0)

We first attach a library called “forecast” to proceed with our tests. To compute we write the following lines of code in R:

```
> arima_fit = auto.arima(ptime[,1])
> tsdiag(arima_fit)
> arima_fit = auto.arima(ptime[,3])
> tsdiag(arima_fit)
> arima_fit = auto.arima(ptime[,4])
> tsdiag(arima_fit)
> arima_fit = auto.arima(ptime[,5])
> tsdiag(arima_fit)
```

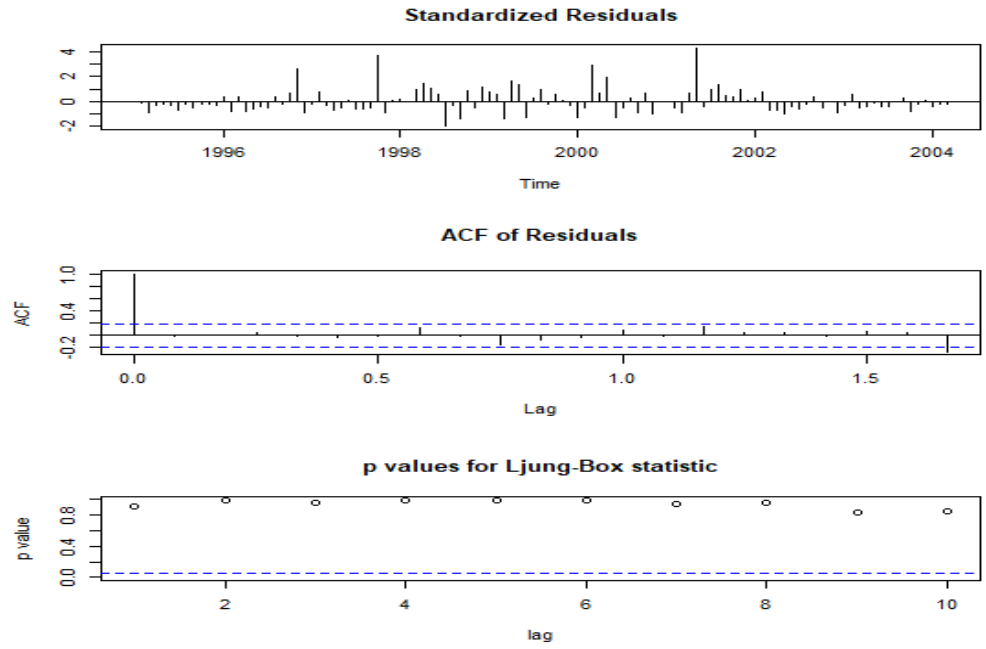
For example,

For model 2, the summary of the variable arima_fit looks like below:

```
Series: ptime[, 3]
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1    mean
      0.2254  9.9371
s.e.  0.0923  0.4216

sigma^2 estimated as 12.12:  log likelihood=-294.97
AIC=595.94   AICc=596.17   BIC=604.07

Training set error measures:
                     ME     RMSE      MAE       MPE     MAPE      MASE
Training set 0.005737414 3.449467 2.738685 -15.93358 35.10801 0.6955613
                    ACF1
Training set -0.01072183
```

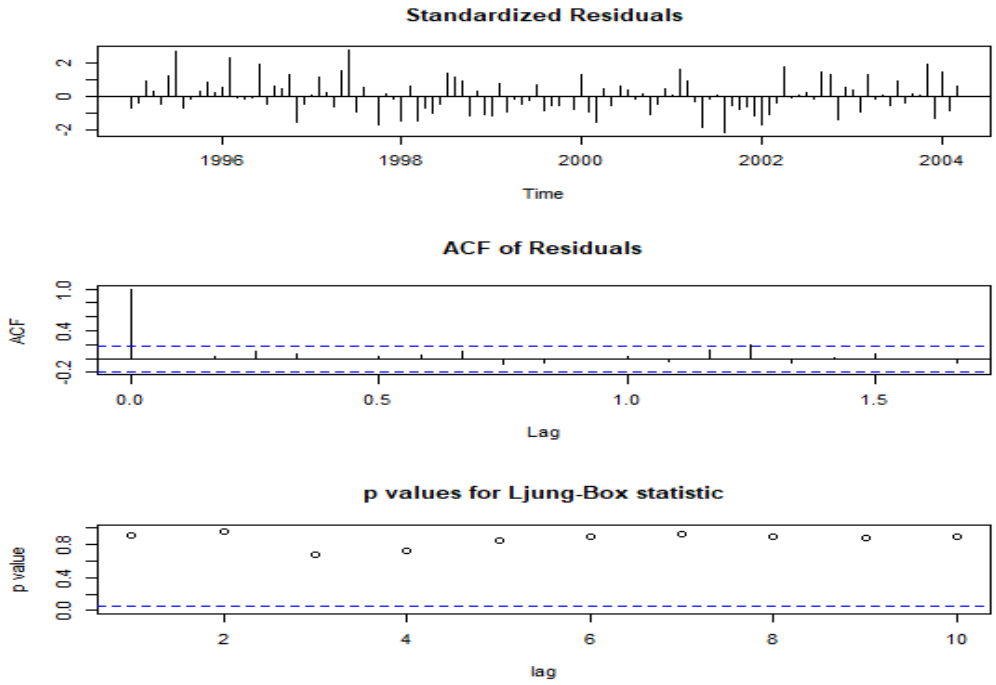
- **Model 1**

The output:



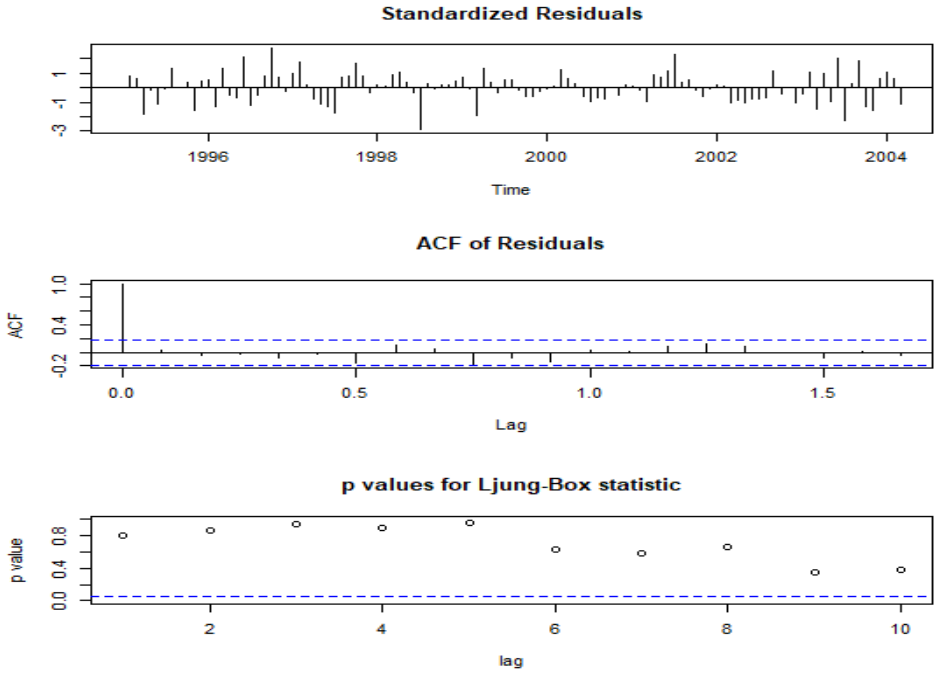
- **Model 2**

The output:



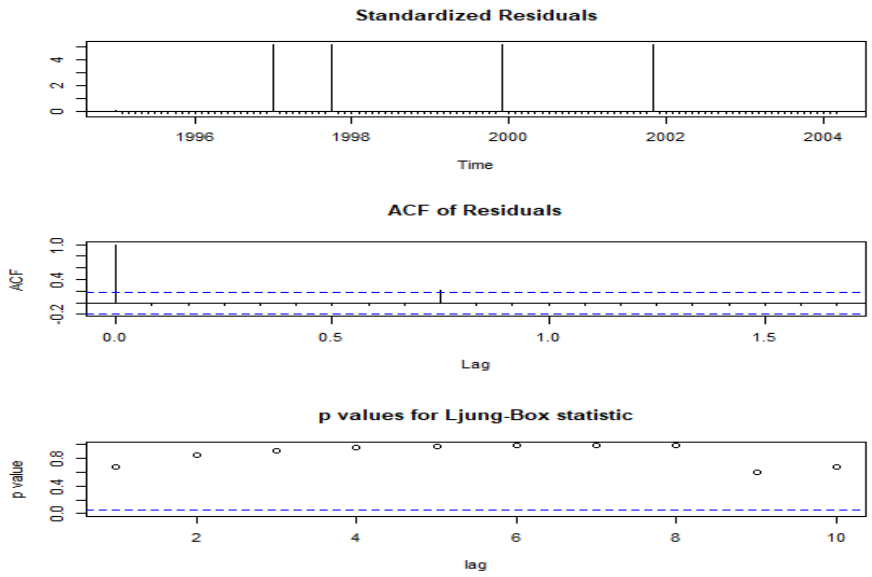

- **Model 3**

The output:



- **Model 4**

The output:




***White noise check:***

To test white noise process more accurately, we run the Box-test. The output is like below:

```
> model_1 = auto.arima(ptime[,1])
> Box.test(model_1$residuals,lag=1)

        Box-Pierce test

data:  model_1$residuals
X-squared = 0.010131, df = 1, p-value = 0.9198
```

```
> model_2 = auto.arima(ptime[,3])
> Box.test(model_2$residuals,lag=1)

        Box-Pierce test

data:  model_2$residuals
X-squared = 0.01276, df = 1, p-value = 0.9101
```

```
> model_3 = auto.arima(ptime[,4])
> Box.test(model_3$residuals,lag=1)

        Box-Pierce test

data:  model_3$residuals
X-squared = 0.064567, df = 1, p-value = 0.7994
```

```
> model_4 = auto.arima(ptime[,5])
> Box.test(model_4$residuals,lag=1)

        Box-Pierce test

data:  model_4$residuals
X-squared = 0.16135, df = 1, p-value = 0.6879
```

According to this test, all p- value of model is bigger than 0.05.

The Ljung–Box test's hypotheses defined as:

$H_0$: *Independently distributed* (*i. e.the correlation in the population from which the sample is taken are* 0, *so that any observed correlation in teh data result from randomness of the sampling process*. )

$H_1$: $Not\ independently\ distributed$ ($They\ exhibit\ serial\ correlation$)

So, all models are independently distributed. Among them, model 1 and model 2 has the higher value for the p-value.

***Picking the best model from RMSE:***

As we calculate the RMSE for all 4 models as shown above for model 2, we get the values as shown below:

| MODEL | RMSE VALUE |
|---|---|
| **Model 1** | **28.50332** |
| **Model 2** | **3.449467** |
| **Model 3** | **5.807943** |
| **Model 4** | **0.1863488** |

According to above figure, Model 4 has the smallest number of RMSE which show the accuracy of modeling and it means model 4 is the best model.

For model 4, the output:

```
Series: ptime[, 5]
ARIMA(0,1,0) with drift

Coefficients:
       drift
      0.0364
s.e.  0.0178

sigma^2 estimated as 0.03536:  log likelihood=28.23
AIC=-52.47   AICc=-52.36   BIC=-47.07

Training set error measures:
                      ME      RMSE       MAE          MPE      MAPE      MASE
Training set 4.471742e-05 0.1863488 0.06949599 -0.02880418 0.9779688 0.1433355
                     ACF1
Training set -0.03812591
```

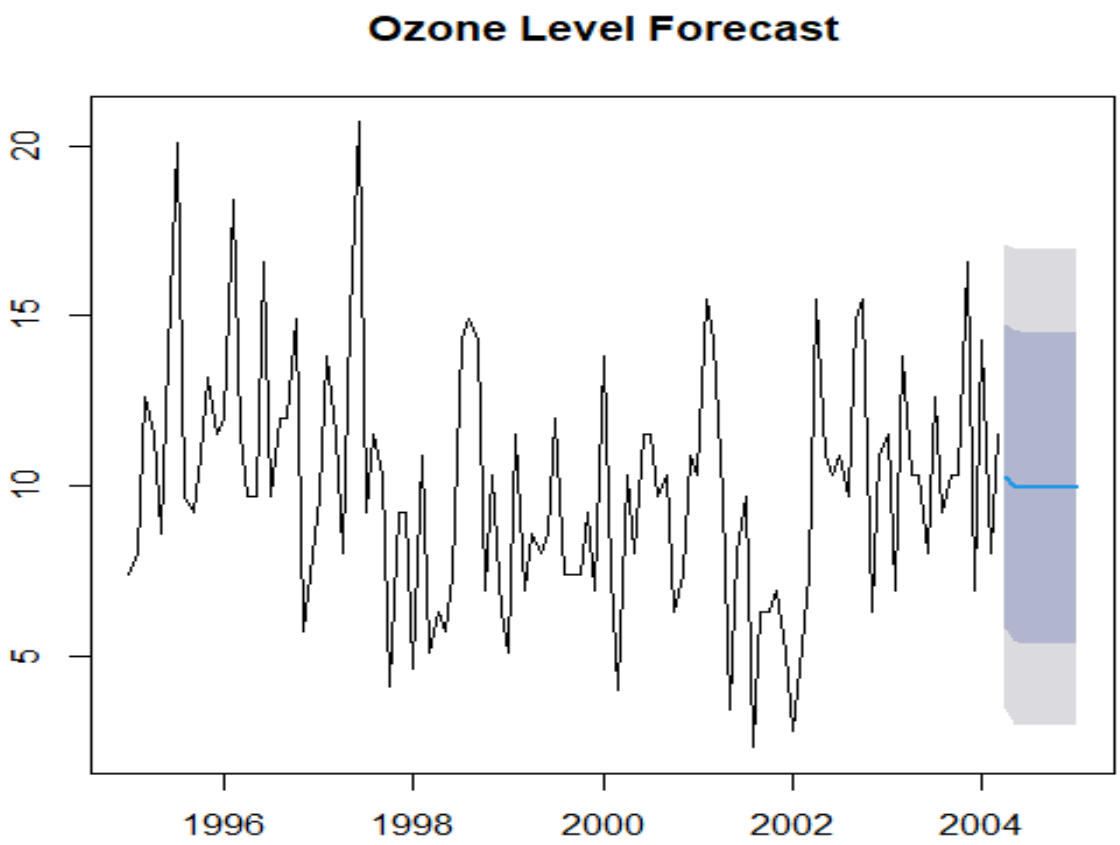
***Forecasting:***

Finally to make our forecasting model, we write the following lines of code in R:

$forecast\ =\ forecast(model\_4, h\ =\ 10)$

$plot(forecast, main\ =\ "Ozone\ Level\ Forecast")$

So, in the end our forecasting model using ARIMA to predict the level of ozone in the air over the next few years look like below:




***Expected original contribution:***

In this research we analyzed the reasons that trigger the level of Ozone in the air. Our aim was also to make a proper forecasting model to get a good view of this behavior pattern in future. Previous research works didn't focus much on the methods that are mostly used to make such forecasting model. In our work we tried making an ARIMA based prediction model to determine the level of $O_3$ in the air.

Some limitations are witnessed during the whole research process. In the dataset there are data available from the year **1995-2007**. We only worked with 153 rows of data. To us it seemed like to do an in-depth analysis this amount of data is not enough and may not give a proper analyzation of the whole data.

For further research we think it would be better if we can work with some recent data. On the other hand some detailed focus can be paid on the factors triggering level of ozone in the air. In this dataset there are only few factors mentioned. If we could do a further analysis on what kind of causes or factors are initiating increased amount of Ozone in the air then it would have been easier to take necessary steps to stop spreading such phenomena. Even though from recent news it has been showed that there has been a significant reduce in the level of Ozone in the air.

***References:***

1. http://www.mass.gov/dep
2. http://www.epa.gov/acidrain/
3. http://www.epa.gov/oar/visibility/

4. http://www.epa.gov/globalwarming/
5. https://www.nasa.gov/feature/goddard/2020/nasa-reports-arctic-stratospheric-ozone-depletion-hit-record-low-in-march
6. Sink mechanism for significantly low level of ozone over the Arabian sea https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2008JD011256
7. Critical levels for ozone effects on vegetation in Europe- J. Fuhrer, L. Skarby and M. R. Ashmore
8. Ozone Pollution: A Major Health Hazard Worldwide - Junfeng (Jim) Zhang, Yongjie Wei and Zhangfu Fang
9. http://vincentarelbundock.github.io/Rdatasets/