A study of Air Quality Index of Banaras Hindu University during different phases of Lockdown and fitting models for estimation and prediction.

A

PROJECT

Submitted in partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE

IN

STATISTICS

UNDER SUPERVISION OF

Dr. Manoj Kumar Chaudhary
Associate Professor
Department Of Statistics
Institute Of Science
BHU

SUBMITTED BY

Ankur Singh
M.Sc. Statistics
Department Of Statistics
Institute of Science
BHU
Enrolment No. 402407
Exam Roll No. 21420STA015

CERTIFICATE

This is to certify that this project entitled "**A study of Air Quality Index of Banaras Hindu University during different phases of Lockdown and fitting models for estimation and prediction.**" is an authentic record and has been successfully prepared and completed by **Ankur Singh, M.Sc. Statistics, Institute of Science, BHU during the session 2022-2023 under my supervision and guidance.**

Submitted on: 10-05-2023

Dr. Manoj Kumar Chaudhary
Associate Professor
Department Of Statistics
Institute Of Science, BHU

ACKNOWLEDGEMENT

This report is a result of study performed as part of our project works in the Department of Statistics, BHU.I would like to acknowledge and give my warmest thanks to my supervisor Dr. Manoj Kumar Chaudhary who made this work possible. His guidance and advice carried me through all the stages of writing my project.

I would also like to give special thanks to my family as a whole for their continuous support and understanding when undertaking my research and writing my project. Your prayer for me was what sustained me this far.

Finally, I would like to thank God, for letting me through all the difficulties. I have experienced your guidance day by day. You are the one who let me finish my degree. I will keep on trusting you for my future.

Ankur Singh

INDEX

INTRODUCTION

In addition to land and water, air is the prime resource for sustenance of life. With the technological advancements. Pollution is the introduction of harmful materials into the environment. Air pollution is the contamination of air due to the presence of substances in the atmosphere that are harmful to the health of humans and other living beings, or cause damage to the climate or to materials. It is also the contamination of indoor or outdoor surrounding either by chemical activities, physical or biological agents that alters the natural features of the atmosphere. There are many different types of air pollutants, such as gases (including ammonia, carbon monoxide, Sulphur dioxide, nitrous oxides, methane, carbon dioxide and chlorofluorocarbons), particulates (both organic and inorganic), and biological molecules. Air pollution can cause diseases, allergies, and even death to humans; it can also cause harm to other living organisms such as animals and food crops, and may damage the natural environment (for example, climate change, ozone depletion or habitat degradation) or built environment (for example, acid rain). Air pollution can be caused by both human activities and natural phenomena.

A vast amount of data on ambient air quality is generated and used to establish the quality of air in different areas. Earlier the air we breathe in use to be pure and fresh.
But, due to increasing industrialization and concentration of poisonous gases in the environment the air is getting more and more toxic day by day. Also, these gases are the cause of many respiratory and other disease.

Nowadays, air pollution in our country has become a bane of our existence. It is not only a problem in our country but the countries all over the world are affected by it. The effects of air pollution are felt especially in the metropolitan cities because of the growing levels of industrialization. The discharge of air pollutants such as concentrates, smog, solid materials etc. are settling over the cities. Excess amount of machinery, use of harmful gases, and use of methane are causing the air pollution. The population's heavy production of waste is the main reason that the air pollution is increasing rapidly. Carbon dioxide, these days, is solely responsible for 57% of the global warming that's taken place till now. Chlorofluorocarbons (CFC's) that are most commonly known for being used in refrigerators are responsible for lowering the concentration of ozone in the stratosphere. CFC's are also heavily in aerosol cans. They were banned in the United States, Canada and most of the Scandinavian countries because of its harmful properties and its capabilities of harming the environment. Aerosols are still used around the world on a regular basis and account for 25% of the global CFC use. The thing with chlorofluorocarbons is that it moves slowly and takes years to move to the stratosphere. This is the biggest reason for the breakdown of the ozone layer. Chlorofluorocarbons are a green house gas that contributes a lot to the depletion of the atmosphere and the environment.Smog is biggest outcome of extreme air pollution. It is the contraction of the words fog and smoke. It has been caused by smoke particles being condensed by water, which usually happens while burning coal. With the replacement of coal with petroleum in countries, pharmaceutical smog has become very common in many big cites. There is a growing need to combat this growing problem that is slowly running our lives as well as our environment. We need to take care of this before it gets worse.

Air Quality Index (AQI)

An **air quality index** (**AQI**) is used by government agencies to communicate to the public how polluted the air currently is or how polluted it is forecast to become Public health risks increase as the AQI rises. Different countries have their own air quality indices, corresponding to different national air quality standards.

The **Central Pollution Control Board** (CPCB) of India is a statutory organization under the Ministry of Environment, Forest and Climate Change (Mo.E.F.C.C.). It was established in 1974 under the Water (Prevention and Control of pollution) Act, 1974. The CPCB is also entrusted with the powers and functions under the Air (Prevention and Control of Pollution) Act, 1981. The National Air Quality Index (AQI) was launched in New Delhi on September 17, 2014, under the Swachh Bharat Abhiyan.
There are six AQI categories, namely Good, Satisfactory, Moderately polluted, Poor, Very Poor, and Severe. The proposed AQI will consider eight pollutants (PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, O_3, NH_3, and Pb) for which short-term (up to 24-hourly averaging period) National Ambient Air Quality Standards are prescribed. Based on the measured ambient concentrations, corresponding standards and likely health impact, a sub-index is calculated for each of these pollutants. The worst sub-index reflects overall AQI. Likely health impacts for different AQI categories and pollutants have also been suggested, with primary inputs from the medical experts in the group. The AQI values and corresponding ambient concentrations (health breakpoints) as well as associated likely health impacts for the identified eight pollutants are as follows:

Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution poses little or no risk
Moderate	51 to 100	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health warnings of emergency conditions. The entire population is more likely to be affected.
Hazardous	301 to 500	Health alert: everyone may experience more serious health effects.

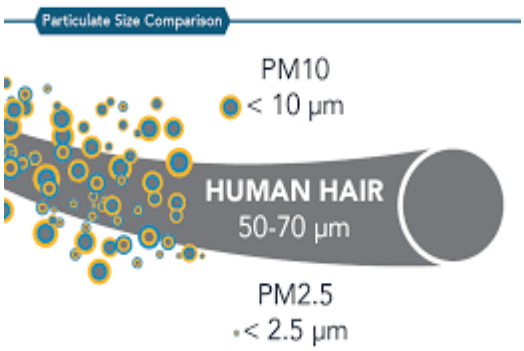
Major Components of air pollutions:

1. Particulate matter (PM10, PM2.5 and SPM):

Particulate matter, also known as particle pollution or PM, is a term that describes extremely small solid particles and liquid droplets suspended in air. Particulate matter can be made up of a variety of components including nitrates, sulphates, organic chemicals, metals, soil or dust particles, and allergens (such as fragments of pollen or mold spores). Particle pollution mainly comes from motor vehicles, wood burning heaters and industry. During bushfires or dust storms, particle pollution can reach extremely high concentrations

The size of particles affects their potential to cause health problems:

- **PM10** (particles with a diameter of 10 micrometers or less): these particles are small enough to pass through the throat and nose and enter the lungs. Once inhaled, these particles can affect the heart and lungs and cause serious health effects.



- **SPM (**Suspended particulate matter): are finely divided solids or liquids that may be dispersed through the air from combustion processes, industrial activities or natural sources.
- **PM2.5** (particles with a diameter of 2.5 micrometers or less): these particles are so small they can get deep into the lungs and into the bloodstream. There is sufficient evidence that exposure to PM2.5 over long periods (years) can cause adverse health effects. Note that PM10 includes PM2.5.

<u>Potential health effects from exposure to particulate matter:</u>

There are many health effects from exposure to particulate matter. Numerous studies have showed associations between exposure to particles and increased hospital admissions as well as death from heart or lung diseases. Despite extensive epidemiological research, there is currently no evidence of a threshold below which exposure to particulate matter does not cause any health effects. Health effects can occur after both short and long-term exposure to particulate matter.
Short-term and long-term exposure is thought to have different mechanisms of effect. Short-term exposure appears to exacerbate pre-existing diseases while long-term exposure most likely causes disease and increases the rate of progression.

<u>Short-term exposure (hours to days) can lead to:</u>

- Irritated eyes, nose and throat
- Worsening asthma and lung diseases such as chronic bronchitis (also called chronic obstructive pulmonary disease or COPD)
- Heart attacks and arrhythmias (irregular heart beat) in people with heart disease
- Increases in hospital admissions and premature death due to diseases of the respiratory and cardiovascular systems

Long-term exposure (many years) can lead to:

- Reduced lung function
- Development of cardiovascular and respiratory diseases
- Increased rate of disease progression
- Reduction in life expectancy

2. Carbon Monoxide:

Carbon monoxide (CO) is an odorless, colorless gas which forms when the carbon in fuels doesn't completely burn. It is usually generated by motor vehicles and industry but can also be formed during bushfires. Indoors, carbon monoxide is formed by unfluted gas heaters, wood-burning heaters, and contained in cigarette smoke.

Carbon monoxide levels are typically highest during cold weather, because cold temperatures make combustion less complete and traps pollutants close to the ground.

Carbon monoxide can cause harmful health effects by reducing the amount of oxygen reaching the body's organs (like the heart and brain) and tissues. At extremely high levels, carbon monoxide can cause death (carbon monoxide poisoning).

Potential health effects from exposure to carbon monoxide:

- Flu-like symptoms such as headaches, dizziness, disorientation, nausea and fatigue
- Chest pain in people with coronary heart disease
- At higher concentration: impaired vision and coordination, dizziness and confusion
- Potentially serious health effects on unborn babies when exposed to high levels

3. Sulphur Dioxide:

Sulphur dioxide is highly reactive gas with a pungent irritating smell. It is formed by fossil fuel combustion at power plants and other industrial facilities.

Natural processes that release Sulphur gases include decomposition and combustion of organic matter, spray from the sea, and volcanic eruptions. It contributes to the formation of particulate matter pollution. Sulphur dioxide irritates the lining of the nose, throat and lungs and may worsen existing respiratory illness especially asthma. It has also been found to exacerbate cardiovascular diseases.

Potential health effects from exposure to Sulphur dioxide:

- Narrowing of the airways leading to wheezing, chest tightness and shortness of breath

- More frequent asthma attacks in people with asthma
- Exacerbation of cardiovascular diseases

4. Nitrogen oxide:

Nitrogen dioxide is a highly reactive gas formed by emissions from motor vehicles, industry, unfluted gas-heaters and gas stove tops. High concentrations can be found especially near busy roads and indoors where unfluted gas-heaters are in use.

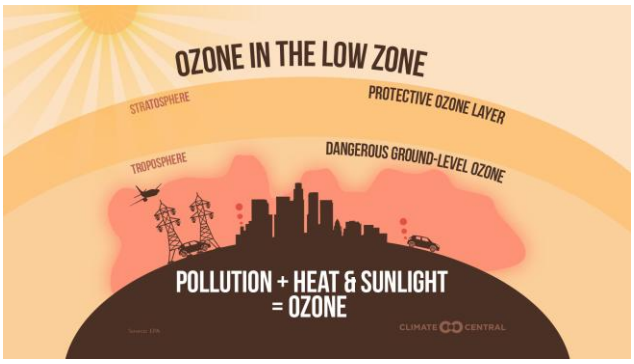
Other indoor sources can be from cigarette smoke or from cooking with gas. Outdoors, nitrogen dioxide contributes to the formation of ground-level ozone (O3) as well as particulate matter pollution. Nitrogen dioxide is a respiratory irritant and has a variety of adverse health effects on the respiratory system.

<u>Potential health effects from exposure to nitrogen dioxide:</u>

- Increased susceptibility to lung infections in people with asthma
- Increased susceptibility to asthma triggers like pollen and exercise
- Worsened symptoms of asthma – more frequent asthma attacks
- Airway inflammation in healthy people

5. Ozone:

Ozone, O3, is composed of three oxygen atoms joined together. Two oxygen atoms joined together form the basic oxygen molecule O2. The additional third atom makes ozone an unstable, highly reactive gas. Ozone is found in two areas of the Earth's atmosphere: in the upper atmosphere and at ground level. Ozone in the upper atmosphere protects us by filtering out damaging ultraviolet radiation from the sun.



On the other hand, ozone at ground level is damaging to our health. Ground level ozone is the main component of smog and is the product of the interaction between sunlight and emissions from sources such as motor vehicles and industry. Ground level ozone is more readily formed during the summer months and reaches its highest concentrations in the afternoon or early evening.

Ozone can travel long distances and accumulate to high concentrations far away from the sources of the original pollutants. Ground level ozone can be harmful to our health even at low levels. This includes ozone generated by ozone generators.

<u>Potential health effects from exposure to ozone:</u>

- Irritation and inflammation of eyes, nose, throat and lower airways: coughing, sore and scratchy throat or uncomfortable feeling in chest
- Reduced lung function: not able to breathe as deeply or vigorously as you normally would
- Exacerbation of asthma and chronic respiratory diseases such as chronic bronchitis (also called chronic obstructive pulmonary disease or COPD)
- Increased susceptibility to respiratory infections
- Can continue to damage lungs when symptoms have disappeared

6. Carbon dioxide:

Carbon dioxide (CO_2) is a colorless gas. In its solid form, it is used as dry ice. It can be found in spring water and is released when volcanoes erupt, trees are cut down, or fossil fuels and products made from them such as oil, gasoline, and natural gas are burned.

<u>Potential health effects from exposure to carbon dioxide:</u>

- Suffocation by displacement of air
- Incapacitation and unconsciousness
- Headache
- Vertigo and double vision
- Inability to concentrate
- Seizures
- Tinnitus

7. Lead:

Humans may be exposed to lead from air pollution directly, through inhalation, or through the incidental ingestion of lead that has settled out from the air onto soil or dust. Ingestion of lead settled onto surfaces is the main route of human exposure to lead originally released into the air.

Once taken into the body, lead distributes throughout the body in the blood and accumulates in the bones. Depending on the level of exposure, lead can adversely affect the nervous system, kidney function, immune system, reproductive and developmental systems, and the cardiovascular system. Lead exposure also affects the oxygen-carrying capacity of the blood.

Lead is persistent in the environment and accumulates in soils and sediments through deposition from air sources, direct discharge of waste streams to water bodies, mining, and erosion. Ecosystems near point sources of lead demonstrate a wide range of adverse effects, including losses in biodiversity, changes in community composition, decreased growth and reproductive rates in plants and animals, and neurological effects in vertebrates.

Potential health effects from exposure to lead:

- Neurological effects in children
- Cardiovascular effects in adults
- Behavioral problems, learning deficits and lowered IQ in infants and young children

8. Chlorofluoro carbon:

CFCs were used for many years as coolant in refrigerators and as cleaning agents. While generally chemically inert and non-toxic in these settings, CFCs diffuse into the upper atmosphere where they destroy the ultraviolet-absorbing ozone shield. Ozone depletion is a concern for the health of humans, as increased exposure to the sun's ultraviolet radiation can cause genetic damage that is associated with various cancers, especially skin cancer.

Potential health effects from exposure to chlorofluoro carbon (CFC):

Inhalation of CFCs results in:

- Frostbite on the skin or in the upper airway
- Symptoms of intoxication
- Reduced co-ordination
- Light-headedness and headache
- Tremors and convulsion
- Irregular heartbeat

Depletion of the ozone layer from CFCs creates dangerous environmental effects and increases exposure to dangerous ultraviolet rays, which can cause:

- Cataracts
- Weakened immune system
- Skin cancer

9. Ammonia:

Ammonia pollution is pollution by the chemical ammonia (NH_3) – a compound of nitrogen and hydrogen which is a byproduct of agriculture and industry. Common forms include air pollution by the ammonia gas emitted by agricultural slurry and fertilizer factories while natural sources include the burning coal mines. Gaseous ammonia reacts with other pollutants in the air to form fine particles of ammonium salts which affect human breathing.

Potential health effects from exposure to Ammonia:

- Skin or eye irritation.
- Bronchiolar and alveolar edema, and airway destruction resulting in respiratory distress or failure
- Ingestion of ammonia results in corrosive damage to the mouth, throat and stomach

Weather factor that also impact on Air Pollution

1.Temperature: Heat waves often lead to poor air quality. The extreme heat and stagnant air during a heat wave increase the amount of ozone pollution and particulate pollution. Drought conditions can also occur during a heat wave, meaning that soils are very dry. During a drought, forest fires are more common. Fires add carbon monoxide and particle pollution to the atmosphere.

2.Dew: Dew condensation processes reduce concentrations of gaseous and particulate pollutants in the near-surface layer.

3.Humidity: Humidity has an impact on the formation and dispersion of air pollutants. Humid air traps pollutants close to the ground, preventing them from dispersing into the atmosphere. This increases the concentrations of pollutants in the air, especially in urban areas. At the same time, high humidity can affect the chemical reactions that take place in the atmosphere, which can impact the formation of certain pollutants such as ozone. It can play an important role in the formation and chemical reactions of air pollutants, having a significant influence on air quality and human health.

4.Precipitate: Rain eases this problem by forcing down the most common air pollutants, like particulate matter and pollen down. Thereby, the quality of air becomes drastically better. This phenomenon is called wet deposition. It is the natural process that eliminates the material through atmospheric hydrometeors, like rain, hail, and snow. It delivers and deposits the contaminants to the ground. The process is also known as precipitation scavenging, rainout, wet removal, or simply washout.

5.Wind Speed: Higher wind speeds generally translate to a greater dispersion of air pollutants, resulting in lower air pollution concentrations in areas with stronger winds. As the ground heats up during the day, the air generally becomes more turbulent, causing air pollutants to disperse in the air.

6.Cloud Cover: clouds reduce the amount of sunlight hitting Earth's surface,which in turn can impact pollutants like ground-level ozone – a key component of smog. Ground-level ozone is formed from a chemical reaction between sunlight and the pollutants in the air.

7.UV-Index: UV lights purify the air by killing mold and bacteria, ultimately funneling clean air into the HVAC system for circulation. UV lights remove contaminants by emitting a light that destroys their DNA makeup.

METHEDOLOGY

In an attempt to get the more useful and descriptive results, it is necessary to begin with a pre-planned strategy. A systematic approach is required. So it's important to analyze questions that have been raised, to solve problems that have been posed or observed, to assess need and set goals, to determine whether or not specific objectives have been met, to establish baselines against which future comparisons can be made, to analyze trends across time, and generally, to describe what exists, in what amount, and in what context.

PLANNING:

Starting a project or research work with proper planning is very important. It include selection of topic and once the topic is decided the project work start automatically. Thus a proper planning is essential for the best results of the statistical data.

AIM:

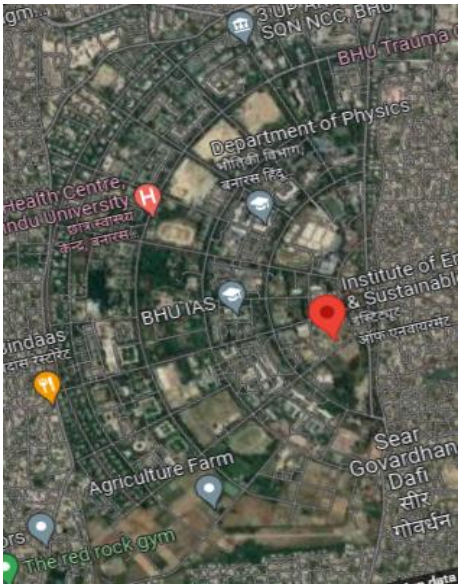
Project statement:- A study of air quality index of Banaras Hindu University during different phases of Lockdown and fitting models for estimation and prediction.

Objective:-

- Is their any difference in mean of AQI in different phases of lockdowns.
- Creating multiple linear regression model for the air quality of Banaras Hindu University.
- To check whether assumptions of multiple linear regression model are fulfilled.

    i. Linearity (A linear relationship between dependent and independent variables)
    ii. Multicollinearity (The independent variables are not highly correlated with each other)
    iii. Normality of residuals
    iv. Independence of residuals (no autocorrelation)
    v. Homoscedasticity (The variance of residuals is constant)

- Create a autoregressive integrated moving average time series model and predicting future air quality index of BHU.



Study area

Banaras Hindu University, Varanasi, Uttar Pradesh , India

25.2677°N 82.9891°E is one of the well known university of world.BHU is located on the southern edge of Varanasi, near the banks of the river Ganges, spread over 1,300 acres (5.3 km2) .

DATA COLLECTION:

To understand the temporal variation and episodic rise of the air pollution in the study region, real time quality monitoring was carried out by Institute of Environment & Sustainable Development (BHU) the data then transferred to Central Pollution Control Board (CPCB). In the present ambient air quality are measured by Continuous Ambient Air Monitoring Station for fine particulate matter PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3, CO and NH_3.
The Weather data is collected by Visual crossing weather site.
The AQI, weather and corresponding pollutants values are taken for the duration 23-June-2021 to 21-Feb-2023.

ANALYSIS AND REPORTING:

The major part of the project was compiled using Python (Jupiter notebook) and analyzed there about. Other programs used are MS-Excel and MS-Word.

DURATION OF THE PROJECT:

The project was started in December 2022 and it came to end in May 2023.

Theory

Multiple Linear Regression

Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change. A linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called *simple linear regression*; for more than one, the process is called multiple linear regression.

So a Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable

The formula for a multiple linear regression is-

$Y=\beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \varepsilon$

y = the predicted value of the dependent variable
β_0= the y-intercept (value of y when all other parameters are set to 0)
$\beta_1 x_1$= the regression coefficient (β_1)of the first independent variable (x_1)
.
.
.
$\beta_n x_n$= the regression coefficient of the last independent variable
ε= model error (how much variation there is in our estimate of y)

To find the best-fit line for each independent variable, multiple linear regression calculates three things:

- The regression coefficients that lead to the smallest overall model error.
- The *t* statistic of the overall model.
- The associated *p* value (how likely it is that the *t* statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).

It then calculates the *t* statistic and *p* value for each regression coefficient in the model.

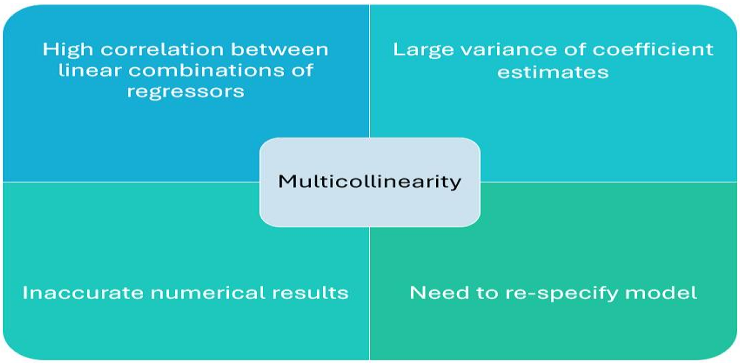
Assumptions of Multiple Linear Regression-

Five main assumptions underlying multiple regression models must be satisfied:

1. Linearity (A linear relationship between dependent and independent variables)
2. Multicollinearity (The independent variables are not highly correlated with each other)
3. Normality of residuals
4. Independence of residuals (no autocorrelation)
5. Homoscedasticity (The variance of residuals is constant)

Assumptions with there Causes, Consequences, Remedies and Detection Technique

1.Linearity: Multiple regressions can be linear and nonlinear. Multiple regressions are based on the assumption that there is a linear relationship between both the dependent and independent variables.
Scatterplots can show whether there is a linear or curvilinear relationship.



2.Multicollinearity: Collinearity is a linear association between *two* explanatory variables. Two variables are perfectly collinear if there is an exact linear relationship between them. For example, X_1 and X_2 are perfectly collinear if there exist parameters λ_0 and λ_1 such that, for all observations i,

$$X_{2i}=\lambda_0 + \lambda_1 X_{1i}$$

Multicollinearity refers to a situation in which *more than two* explanatory variables in a multiple regression model are highly linearly related.

Cause: High correlation between explanatory variables of regression model.

Consequences:

1. Least square estimators are indeterminant.
2. The variance and covariance of estimators become infinitely large.
3. Because of consequence 1 the confidence interval can be much wider leading to acceptance of zero null hypothesis more frequently.
4. Although the t-ratio of one or more coefficient is statistically insignificant, R^2 (the overall measure of goodness of fit) can be very high.
5. The ordinary least square estimators and their standard error can be sensitive toward small change in data.

Remedies:

- Increase the sample size by doing so, the variance of the OLS estimates tends to decrease, even if it is inflated by high correlation among regressors; furthermore, a larger sample size decreases over-fitting and tends to reduce.
- Drop one or more regressors that have a high VIF if they are not deemed to be essential.
- Replace highly correlated regressors with a linear combination of them.
- Use regularization methods such as ridge and lasso.
- Use Bayesian regression; if multicollinearity prevents a regression coefficient from being estimated precisely, then the prior on that coefficient will help to reduce its posterior variance.

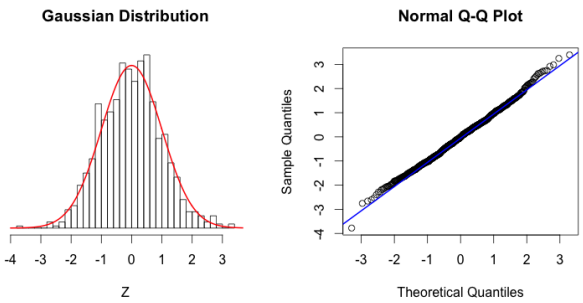
Detection Technique: A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

$VIF_i = 1/1 - R_i^2$

where: R_i^2=Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones.
When R_i^2 is equal to 0, and therefore, when VIF or tolerance is equal to 1, the ith independent variable is not correlated to the remaining ones, meaning that multicollinearity does not exist.1
In general terms,
VIF equal to 1 = variables are not correlated
VIF between 1 and 5 = variables are moderately correlated
VIF greater than 5 = variables are highly correlated

The higher the VIF, the higher the possibility that multicollinearity exists, and further research is required. When VIF is higher than 10, there is significant multicollinearity that needs to be corrected by using remedies listed above.

3. Normality of Residuals: Normality of the residuals is an assumption of running a linear model. So, if your residuals are normal, it means that your assumption is valid and model inference (confidence intervals, model predictions) should also be valid.
Prediction intervals are calculated based on the assumption that the residuals are normally distributed.



Consequences:
If the residuals are non normal, the prediction intervals may be inaccurate.
Violation of the normality assumption only becomes an issue with small sample sizes. For large sample sizes, the assumption is less important due to the central limit theorem, and the fact that the F and t-tests used for hypothesis tests and forming confidence intervals are quite robust to modest departures from normality.

Remedies:
If the data appear to have non-normally distributed random errors, but do have a constant standard deviation, you can always fit models to several sets of transformed data and then check to see which transformation appears to produce the most normally distributed residuals.

Transform the response variable to make the distribution of the random errors approximately normal.
Transform the predictor variables, if necessary, to attain or restore a simple functional form for the regression function.
Fit and validate the model in the transformed variables.
Transform the predicted values back into the original units using the inverse of the transformation applied to the response variable.

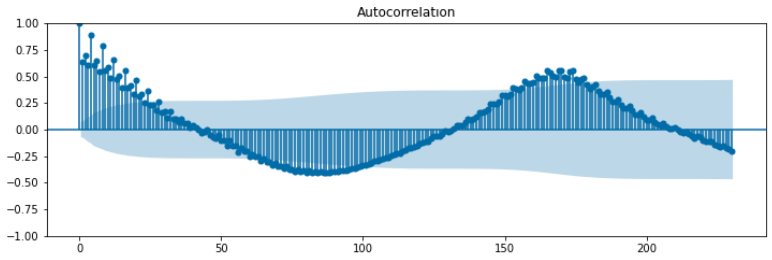
Detection Technique:

1. **"Residuals vs. Fitted" Plot-** A "Residuals vs. Fitted"-plot is a scatter plot of the residuals on the y-axis and the fitted (i.e., predicted) value on the x-axis. For the normality assumption to hold, the residuals should spread randomly around 0 and form a horizontal band.
2. **Q-Q Plot**- A Q-Q plot (or quantile-quantile plot) is a scatterplot that plots two sets of quantiles against one another. If the Q-Q plot forms a diagonal line, you can assume that the residuals follow a normal distribution.
3. **Histogram of the Residuals**-To not violate the normality assumption, the histogram should be centered around zero and should show a bell-shaped curve.
4. **Shapiro-Wilk test**- The Shapiro-Wilk test is a statistical test used to check if a continuous variable follows a normal distribution. The null hypothesis (H0) states that the variable is normally distributed, and the alternative hypothesis (H1) states that the variable is NOT normally distributed.

4.Independence of residuals (no autocorrelation): Autocorrelation measures the relationship between a variable's current value and its past values.

Correlation vs. Autocorrelation

Correlation measures the relationship between two variables, whereas autocorrelation measures the relationship of a variable with lagged values of itself.



In Linear Regression autocorrelation occurs when the residuals are not independent of each other. That is, when the value of e_{i+1} is not independent from e_i.

Consequences:

1. When the disturbance terms are serially correlated then the OLS estimators of the $\hat{\beta}$s are still unbiased and consistent but the optimist property (minimum variance property) is not satisfied.
2. The OLS estimators will be inefficient and therefore, no longer best linear unbiased estimator.
3. The estimated variance of the regression coefficients will be biased and inconsistent and will be greater than the variances of estimate calculated by other methods, therefore, hypothesis testing is no longer valid. In most of the cases, R^2 will be overestimated (indicating a better fit than the one that truly exists). The t- and F-statistics will tend to be higher.
4. The variance of random term e may be under-estimated if the e's are autocorrelated. That is, the random variance is likely to be under-estimate the true variance.

Detection Technique:

1. **Residual Plot**: If any pattern is spotted it means autocorrelation present.

2. **Run test of randomness:** A statistical test that is used to know the randomness in data. Run test of randomness is sometimes called the Geary test, and it is a nonparametric test. Run test

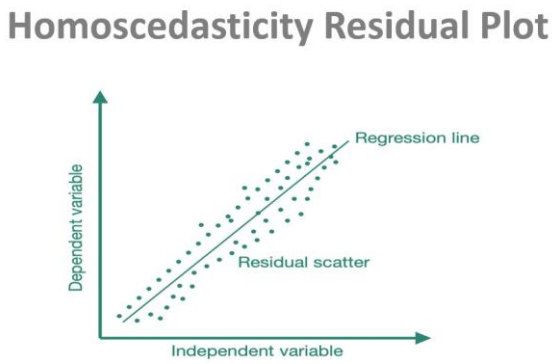
of randomness is an alternative test to test autocorrelation in the data. To confirm whether or not the data has correlation with the lagged value, run test of randomness is applied. Run test of randomness is basically based on the run. Run is basically a sequence of one symbol such as + or -. Run test of randomness assumes that the mean and variance are constant and the probability is independent.

3. **Durbin Watson Test:** The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical model or regression analysis. The Durbin-Watson statistic will always have a value ranging between 0 and 4. A value of 2.0 indicates there is no autocorrelation detected in the sample. Values from 0 to less than 2 point to positive autocorrelation and values from 2 to 4 means negative autocorrelation. It only test's first-order autocorrelation with assumptions that error term follow normal distribution and the explanatory variables or regressor are non stochastic.

4. **Breusch-Godfrey test**: is a statistical test that is used to detect autocorrelation in the residuals of a linear regression model. It helps to detect autocorrelation at different lags and it's applicable to both linear and non-linear models. It's benefit over DW test is that it does not restricted on any of the assumptions of DW test.

5. Homoscedasticity (The variance of residuals is constant): An important assumption of classical linear regression model is that the disturbance U_i appearing in the regression function are homoscedastic, i.e they all have the same variance. It may be the case however that all of the disturbance term do not have same variance this condition of non homogeneity of variance is known as heteroscedasticity.

$U_i^{'s}$ are heteroscedastic when variance of $U_i \neq \sigma^2_u$
(a constant value) $U_i = \sigma^2_{ui}$ (a value that varies)



Consequences:

Although the OLS estimator remains unbiased, the estimated standard error is wrong. Because of this, confidence intervals and hypotheses tests cannot be relied on. In addition, the OLS estimator is no longer BLUE.

Remedies:

1. Use OLS estimator to estimate the parameters of the model. Correct the estimates of the variances and covariances of the OLS estimates so that they are consistent.
2. Use an estimator other than the OLS estimator to estimate the parameters of the model.

Detection Technique:

1.Graphical method: In this method Residual square U_i^2 is plotted against the predicted value of the dependent variable. If the plot doesn't show any pattern then heteroscedasticity is said to be absent otherwise it is said to be absent.

2.Park test: Park had modeled the error variance as a function of explanatory variables defined as: or,

$\sigma^2_i = \sigma^2 X_i^{\beta} e^{ui}$

$\log_e \sigma^2_i = \log_e \sigma^2 + \beta \log_e X_i + U_i$

Where Ui is homoscedastic error term. However, since σ^2_i is unknown Park had been suggested the use of U_i^2 in its place. If β comes out to be significant then heteroscedasticity is said to be present in the data.

3.Spearman's Rank test:

1. 1. Fit the regression of Y on X and obtain the residuals.
2. 2. Compute the Spearman's rank correlation between absolute value of residuals and Xi (or $\hat{Y}i$)
3. 3. Test the null hypothesis that population correlation coefficient is zero using t-test. If the hypothesis is rejected then heteroscedasticity is said to be present. The t-statistics is given by:
   $t = r_s(n-2)^{1/2}/(1-r_s^2)^{1/2}$

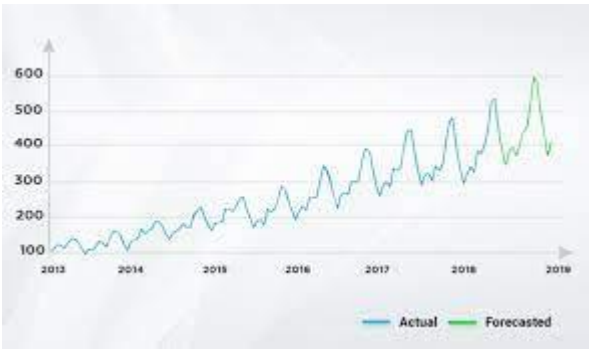
r_s = Spearman's rank correlation coefficient.

4.Goldfled Quandt test: This method is applicable only if the heteroscedastic variance (σ^2_i) is positively related with one of the explanatory variable. In this method it is assumed that σ^2_i is proportional to the square of the explanatory variable. Goldfled and Quandt had suggested a number of steps to detect the heteroscedasticity which are on next slide.

1. Order (or arrange) the observations in the ascending order of values of X.
2. Omit c central values (c is a specified a priori) and divide the remaining c central values into two equal halves having (n-c)/2 observations.
3. Fit separate OLS regression for both the halves and compute residual sum of squares for each (say RSS1 and RSS2) having (n-c-2k)/2 degrees of freedom. (k is no. of parameters estimated.)
4. Compute the ratio λ = RSS1 /RSS2 .
5. Heteroscedasticity is said to present if : $\lambda > F_{(n-c-2k)/2,\ (n-c-2k)/2,\ \alpha}$

Time Series

A time series is a sequence of data points that occur in successive order over some period of time. Time series *analysis* comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series *forecasting* is the use of a model to predict future values based on previously observed values.



Time series analysis

Time series analysis is a technique in statistics that deals with time series data and trend analysis

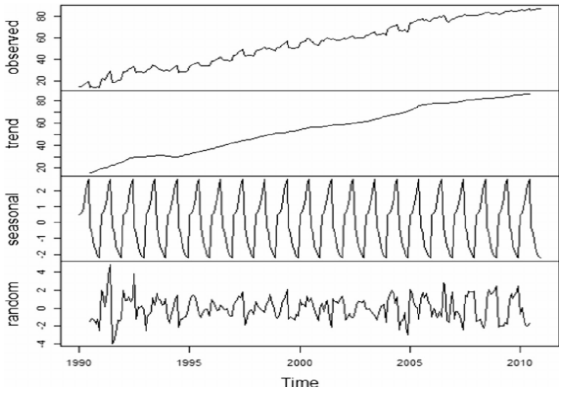
Certain aspects are an integral part of the time series analysis process. Analyst should be able to identify that the data is:

- **Stationarity** is a crucial aspect of a time series. A time series is determined to be stationary when its statistical properties such as the average (mean) and the variance do not alter over time. It has a constant variance and mean, and the covariance is separate from time.
- **Seasonality** refers to periodic fluctuations. For example, if you consider electricity consumption, it is typically high during the day and lowers during the night. In the case of shopping patterns, online sales spike during the holidays before slowing down and dropping.
- **Autocorrelation** is the similarity between observations as a function of the time lag between them. Plotting autocorrelated data yields a graph similar to a sinusoidal function.

Time Series Models

Models for time series data can have many forms and represent different stochastic processes. When modeling variations in the level of a process, three broad classes of practical importance are the *autoregressive* (AR) models, the *integrated* (I) models, and the *moving average* (MA) models. These three classes depend linearly on previous data points. Combinations of these ideas produce autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models. The autoregressive fractionally integrated moving average (ARFIMA) model generalizes the former three.

Components of Time Series Analysis



Trend: Trend shows a common tendency of data. It may move upward or increase or go downward or decrease over a certain, long period of time. The trend is a stable and long-term general tendency of movement of the data. To be a trend, it is not mandatory for the data to move in the same direction. The direction or movement may change over the long-term period but the overall tendency should remain the same in a trend.A Trend can be either linear or non-linear.

Seasonal Variations: Seasonal variations are changes in time series that occur in the short term, usually within less than 12 months. They usually show the same pattern of upward or downward

growth in the 12-month period of the time series. These variations are often recorded as hourly, daily, weekly, quarterly, and monthly schedules.Seasonal variations occur due to natural or manmade forces or variations. The numerous seasons and manmade variations play a vital role in seasonal variations. Seasonal variations can be clearly seen in some cases of man-made conventions. The festivals, customs, fashions, habits, and various occasions, such as weddings impact the seasonal variations. An increase in business during the seasonal variation period should not be considered a better business condition.

Cyclical Variations: Variations in time series that occur themselves for the span of more than a year are called Cyclical Variations. Such oscillatory movements of time serious often have a duration of more than a year. One complete period of operation is called either a cycle or a 'Business Cycle'.Cyclic variations contain four phases - prosperity, recession, depression, and recovery. It may be regular or non-periodic in nature. Usually, cyclical variations occur due to a combination of two or more economic forces and their interactions.

Random or Irregular Movements: There is another kind of movement that can be seen in the case of time series. It is pure Irregular and Random Movement. As the name suggests, no hypothesis or trend can be used to suggest irregular or random movements in a time series. These outcomes are unforeseen, erratic, unpredictable, and uncontrollable in nature.

**An additive model would be used when the variations around the trend do not vary with the level of the time series whereas a multiplicative model would be appropriate if the trend is proportional to the level of the time series

Y = T + S + C + I **Y = T * S * C * I**

Data: Types, Terms, and Concepts

Data, in general, is considered to be one of these three types:

1. **Time series data:** A set of observations on the values that a variable takes on at different points of time.
2. **Cross-sectional data:** Data of one or more variables, collected at the same point in time.
3. **Pooled data:** A combination of time series data and cross-sectional data.

These are some of the terms and concepts associated with time series data analysis:

- **Dependence:** Dependence refers to the association of two observations with the same variable at prior time points.
- **Stationarity:** This parameter measures the mean or average value of the series. If a value remains constant over the given time period, if there are spikes throughout the data, or if these values tend toward infinity, then it is not stationarity.
- **Differencing:** Differencing is a technique to make the time series stationary and to control the correlations that arise automatically. That said, not all time series analyses need differencing and doing so can produce inaccurate estimates.
- **Curve fitting:** Curve fitting as a regression method is useful for data not in a linear relationship. In such cases, the mathematical equation for curve fitting ensures that data that falls too much on the fringes to have any real impact is "regressed" onto a curve with a distinct formula that systems can use and interpret.

Types of Time Series Model

Autoregressive Model: AR model relies only on past period values to predict current ones. It's a linear model, where current period values are a sum of past outcomes multiplied by a numeric factor. We denote it as AR(p), where "p" is called the order of the model and represents the number of lagged values we want to include.
For instance, if we take X as time-series variable, then an AR(p), also known as a simple autoregressive model, would look something like this:

$$X_t = C + \Sigma_i \phi_i X_{t-i} + \epsilon_t \quad i=1,2,\ldots,p \quad (\epsilon t = Xt - \hat{X}t).$$

Moving Average Model: Rather than using past values of the forecast variable in a regression, a moving average model uses past forecast errors in a regression-like model.

$$X_t = c + \varepsilon_t + \Sigma_i \theta_i \varepsilon_{t-i}$$

where ε_t is white noise. We refer to this as an MA(q) model, a moving average model of order q.

ARMA (Auto Regressive Moving Average) Model: This is a model that is combined from the AR and MA models. In this model, the impact of previous lags along with the residuals is considered for forecasting the future values of the time series.

$$X_t = \Sigma_i \phi_i X_{t-i} + \Sigma_i \theta_i \varepsilon_{t-i} + \epsilon_t$$

ARIMA(autoregressive integrated moving average): ARIMA model is classified as an "ARIMA(p,d,q)" model, where:
p is the number of autoregressive terms
d is the number of non seasonal differences needed for stationarity
q is the number of lagged forecast errors in the prediction equation
The forecasting equation is constructed as follows. First, let y denote the d^{th} difference of Y, which means:
If d=0: $X_t = X_t$
If d=1: $X_t = X_t - X_{t-1}$
If d=2: $X_t = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) = X_t - 2X_{t-1} + X_{t-2}$
Note that the second difference of Y (the d=2 case) is not the difference from 2 periods ago. Rather, it is the *first-difference-of-the-first difference*, which is the discrete analog of a second derivative, i.e., the local acceleration of the series rather than its local trend.
In terms of X, the general forecasting equation is:
$\hat{X}_t = \mu + \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} - \theta_i e_{t-1} - \ldots - \theta_q e_{t-q}$
Here the moving average parameters (θ's) are defined so that their signs are negative in the equation, following the convention introduced by Box and Jenkins.
To identify the appropriate ARIMA model for X, we begin by determining the order of differencing (d) needing to stationaries the series and remove the gross features of seasonality, perhaps in conjunction with a variance-stabilizing transformation such as logging or deflating. If you stop at this point and predict that the differenced series is constant, you have merely fitted a random walk or

random trend model. However, the stationaries series may still have autocorrelated errors, suggesting that some number of AR terms ($p \geq 1$) and/or some number MA terms ($q \geq 1$) are also needed in the forecasting equation.

ARIMA(1,1,2) without constant

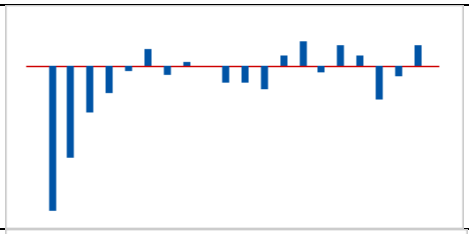
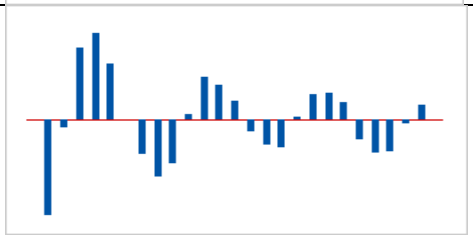
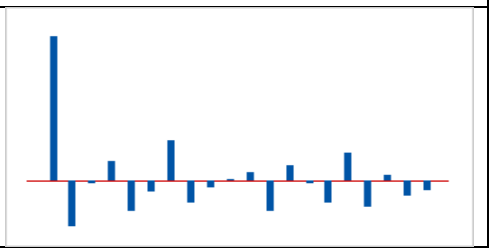
$\hat{X}_t = X_{t-1} + \phi_1 (X_{t-1} - X_{t-2}) - \theta e_{t-1} - \theta_1 e_{t-1}$

Autocorrelation function and Partial Autocorrelation function (ACF and PACF)

The coefficient of correlation between two values in a time series is called the autocorrelation function (ACF). When performing ACF it is advisable to remove any trend present in the data and to make sure the data is stationary.

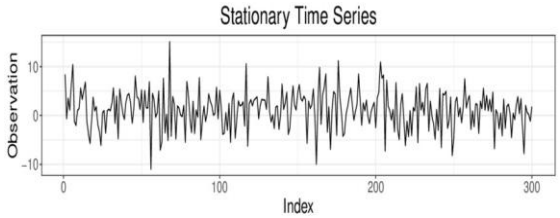
The partial autocorrelation function is a measure of the correlation between observations of a time series that are separated by k time units (y_t and y_{t-k}), after adjusting for the presence of all the other terms of shorter lag (y_{t-1}, y_{t-2}, ..., y_{t-k-1}).

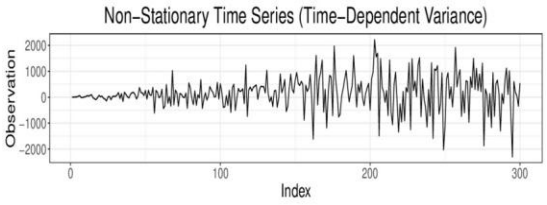
Analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) in conjunction is necessary for selecting the appropriate ARIMA model for any time series prediction.

Pattern	**Indicates**	**Example**
Large spike at lag 1 that decreases after a few lags.	A moving average term in the data. Use the autocorrelation function to determine the order of the moving average term.	
Large spike at lag 1 followed by a damped wave that alternates between positive and negative correlations.	A higher order moving average term in the data. Use the autocorrelation function to determine the order of the moving average term.	
Significant correlations at the first or second lag, followed by correlations that are not significant.	An autoregressive term in the data. The number of significant correlations indicate the order of the autoregressive term.	

Stationarity of a Time Series

A stationary time series is one whose properties do not depend on the time at which the series is observed . The time series data that mean and variance do not vary across time. The data is considered non-stationary if there is a strong trend or seasonality observed from the data.





1. **<u>Statistical methods to check stationarity</u>**

   Augmented Dickey-Fuller Test: The Augmented Dickey-Fuller Test (ADF) is a stationarity unit root test. The ADF test is a modified version of the Dickey Fuller exam. In the time series analysis, unit-roots might produce unexpected findings.With serial correlation, the Augmented Dickey-Fuller test may be utilized. The ADF test is more powerful and can handle more complicated models than the Dickey-Fuller test. However, like with other unit root tests, it should be used with caution because it has a somewhat high Type I error rate.

   The following are the test hypotheses:

   Null hypothesis (H0): The time series data is non-stationary.

   Alternate hypothesis (H1): The time series is stationary (or trend-stationary).

   The ADF test extends the Dickey-Fuller test equation to include in the model a high order regressive process. It adds extra differencing terms, but the rest of the equation stays unchanged. This increases the thoroughness of the test. The null hypothesis, on the other hand, remains the same as in the Dickey-Fuller test. To reject the null hypothesis, the p-value produced should be less than the significance level (say, 0.05). As a result, we may conclude that the series is stationary.

2. **<u>Kwiatkowski Phillips Schmidt Shin (KPSS) test:</u>**

   The Kwiatkowski Phillips Schmidt Shin (KPSS) test determines if a time series is stationary around a mean or linear trend, or non-stationary as a result of a unit root. A stationary time series has statistical features such as mean and variance that remain constant across time.

   The following are the test hypotheses:

   Null hypothesis (H0): The data is stationary.

   Alternate hypothesis (H1): The data is not stationary.

   The linear regression underpins the KPSS test. With the regression equation, it divides a series into three parts: a deterministic trend, a random walk, and a stationary error. If the data is

stationary, the intercept will have a fixed element or the series will be stationary around a fixed level.

The test uses OLS to compute the equation, which varies significantly depending on whether you want to test for level or trend stationarity. To assess level stationarity, a reduced version lacking the temporal trend component is used.

(Visualization method) Rolling Statistics : A rolling analysis of a time series model is often used to assess the model's stability over time. Plot the moving average or moving standard deviation to see if it varies with time. It's a visual technique.
If rolling mean and variance of a time series is not constant over time then it its a non stationary time series.

Techniques to make Non-Stationary time series Stationary

- **Transformation** techniques used and they are as follows: Log transforming of the data, Taking the square root of the data, Taking the cube root, Proportional change.

- **Differencing** is performed by subtracting the previous observation from the current observation or we can say, by subtracting previous day demand from current day demand.By differencing, stationarity can be achieved easily. This means time-series does not depends on time. It's like white noise, no matter when we observe it looks same at any point of time. Whereas, trends and seasonality affect the time-series at different times. Stationary time-series does not have any predictable pattern.

TABULATION AND ANALYSIS

Data are recorded in Microsoft Excel sheet files that were compiled in tables where the relevant information is extracted, analyzed and mathematically treated using Python programming software. The Exploratory Data Analysis (EDA) is used for analyzing the datasets.

Variable Summary-

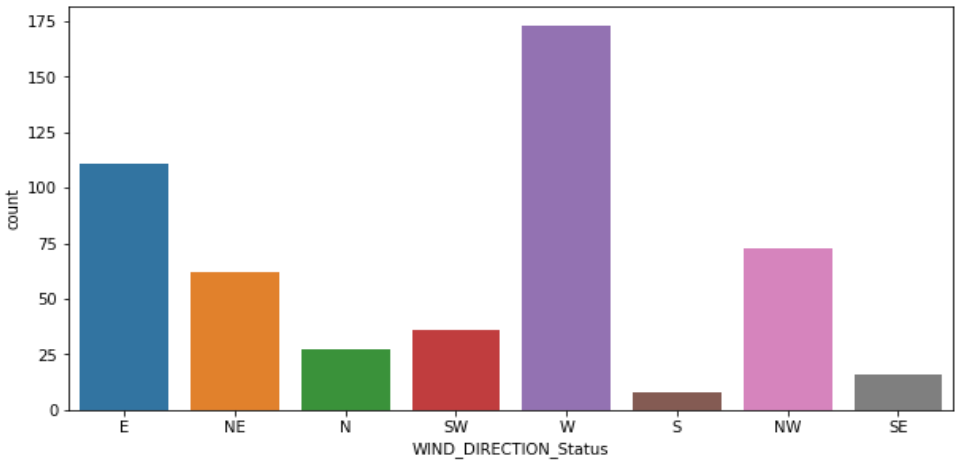
Variable	Unit	Count	Mean	Standard Deviation	1st Quartile	Median	3rd Quartile	Minimum	Maximum
PM2.5	Micrograms per cubic meter of air	548	67.17	54.29	29	51	86	6	333
PM10	micrograms per cubic meter of air	574	70.98	32.39	44	69	94	7	204
NO_2	ppb	587	20.94	11.11	12	19	29	3	79
NH_3	ppm	573	4.27	2.89	2	3	6	1	32
SO_2	ppb	595	21.42	9.06	15.5	20	25	1	48
CO	ppm	606	17.44	13.12	9	14	22	1	74
OZONE	Dobson unit	589	25.79	32.37	6	10	33	1	193
Temperature maximum	Fahrenheit	609	88.71	11.76	80.7	91.5	96.9	57.2	115.6
Temperature minimum	Fahrenheit	609	67.85	12.70	55.5	73.5	78.9	40.2	90.8
Temperature	Celsius	609	25.31	6.79	19.05	27.83	30.5	9.27	38.22
Temperature Feels like	Fahrenheit	609	82.5	16.75	66.3	86.8	97	48.6	113.3
Dew	Percentage	609	64.96	12.32	54	64.7	78	37.1	83.1
Humidity	Percentage	609	70.24	16.59	62.5	74	81.9	19.5	97.4
Precipitate	Millimeter(24hrs)	609	0.09	0.36	0	0	0.004	0	4.13
Precipitate probability	Percentage	609	26.76	44.30	0	0	100	0	100
Precipitate cover	Millimeter	609	1.95	3.96	0	0	4.17	0	29.17
Snow	Millimeter(24hrs)	609	0	0	0	0	0	0	0
Snow depth	Millimeter	609	0	0	0	0	0	0	0
Wind gust	Kmph	408	15.28	7.63	8.9	14.5	20.1	3.4	40
Wind speed	Kmph	609	9.55	4.37	6.5	9.2	11.4	2.2	46.2
Wind Direction	Degree	609	193.58	105.84	81.6	247.9	282.5	0.7	359
Sea level pressure	hPa	609	1008.08	6.99	1002.2	1007	1014.3	995	1023.3
Cloud cover	Percentage	609	35.14	30.87	4.6	29.8	63.3	0	98.4
Visibility	Km	609	2.68	1.37	1.8	2.4	3.5	0.2	15
Solar radiation	kWh/m^2	609	218.99	57.19	182.8	212.3	250.3	15.9	344.8
Solar energy	kWh/m^2	609	18.9	4.93	15.8	18.3	21.5	1.2	29.9
UV index	Unitless	609	7.83	1.5	7	8	9	1	10
Sever risk	Unitless	408	14.89	11.4	10	10	10	10	60
Daylight duration	Hours	609	12.02	1.10	10.94	11.91	13.1	10.94	13.71
Moon phase	Percentage	609	0.5	0.28	0.25	0.50	0.75	0	1

There are few more variables like (”FEELS LIKE MAX”,”FEELS LIKE MIN”,”SUNRISE”,“SUNSET”,“ICON”,“stations”)
Which are of no use and variables like “CONDITION“ and ”WEATHER DESCRIPTION” which are converted to categorical variable.

Transformation

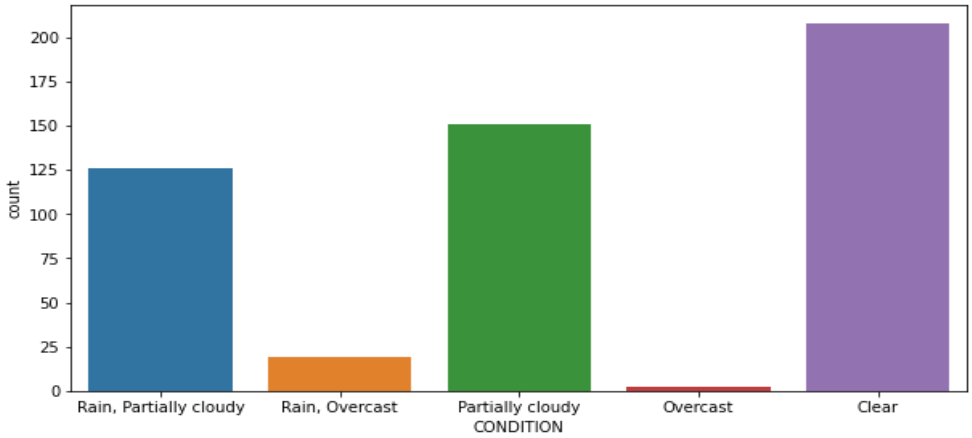
1.The direction of flow of wind in Banaras Hindu University in the total duration of study

NE= Wind Direction > 22.5 & Wind Direction <= 67.5 For NE Wind Direction_Status = 1
E = Wind Direction > 67.5 & Wind Direction <= 112.5 For E Wind Direction_Status= 2
SE= Wind Direction > 112.5 & Wind Direction <= 157.5 For SE Wind Direction_Status = 3
S = Wind Direction > 157.5 & Wind Direction <= 202.5 For S Wind Direction_Status = 4
SW= Wind Direction > 202.5 & Wind Direction <= 247.5 For SW Wind Direction_Status = 5
W = Wind Direction > 247.5 & Wind Direction <= 292.5 For W Wind Direction_Status = 6
NW= Wind Direction > 292.5 & Wind Direction <= 337.5 For NW Wind Direction_Status = 7
N = Wind Direction > 337.5 & Wind Direction <= 360 also Wind Direction >= 0 & Wind Direction <= 22.5 For N Wind Direction_Status = 8



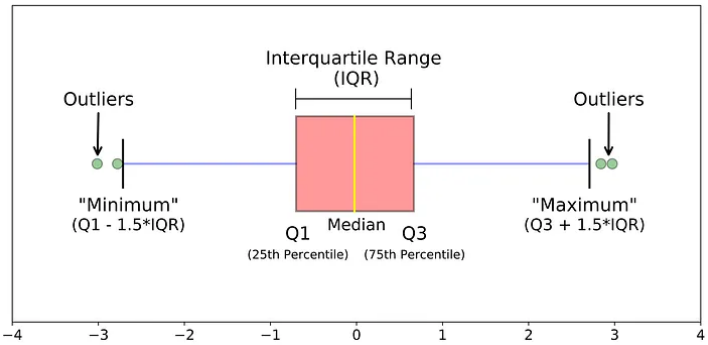
2. The weather condition in Banaras Hindu University in the total duration of study

For “Rain,Partially cloudy” CONDITION_status= 1
For “Rain, Overcast” CONDITION_status= 2
For Partially cloudy CONDITION_status= 3
For Overcast CONDITION_status= 4
For Clear CONDITION_status= 5



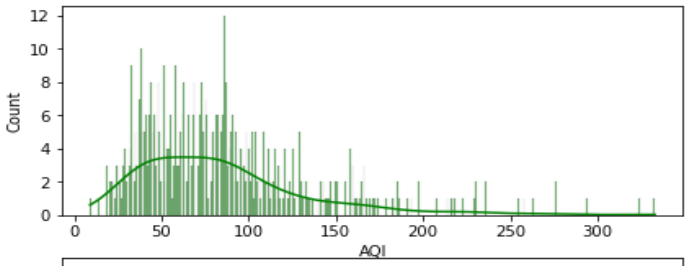
The distribution and presence of outliers in the factors deciding Air Quality Index

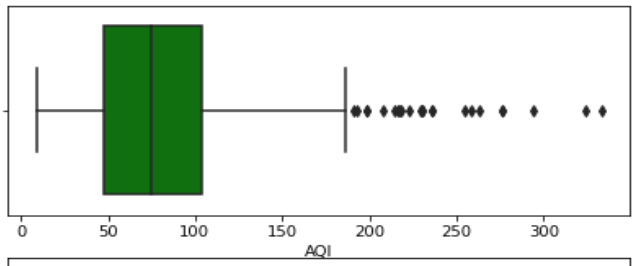
Boxplots are a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).

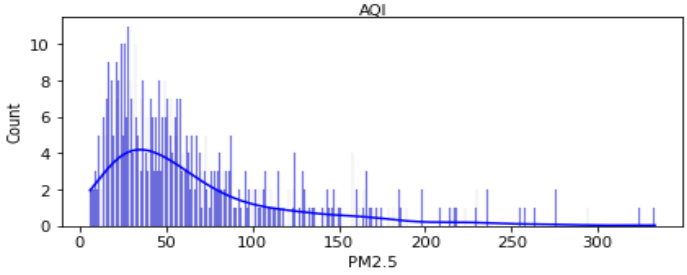


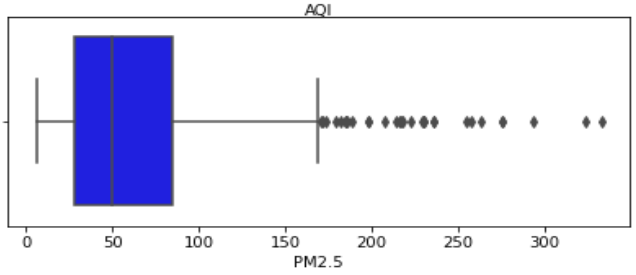
This type of plot is used to easily detect outliers. It can also tell us if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

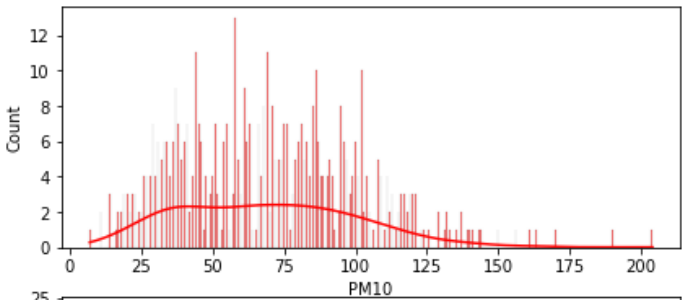
A histogram, on the other hand, is a graph that shows the distribution of numerical data.

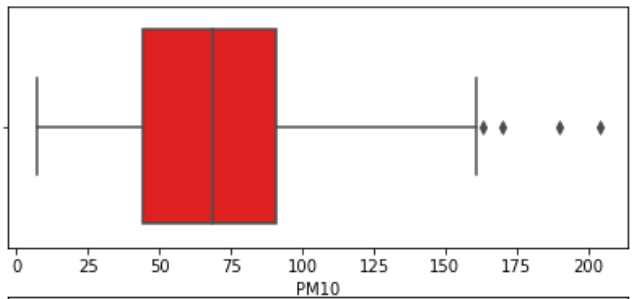


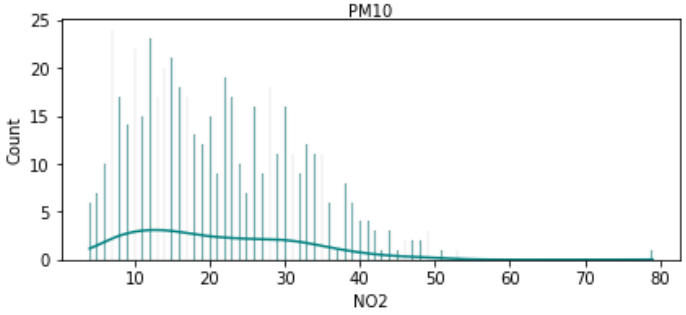


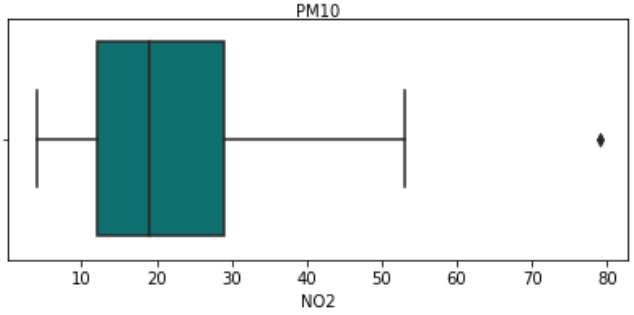


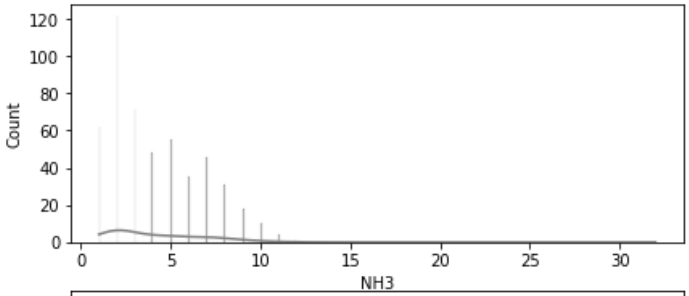


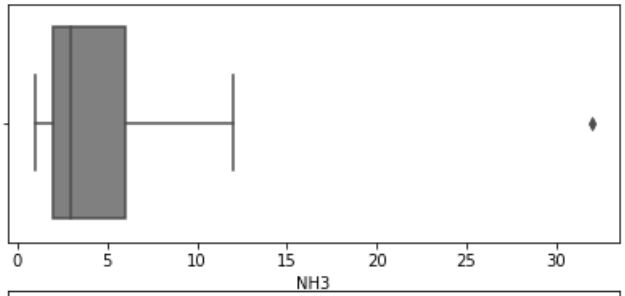


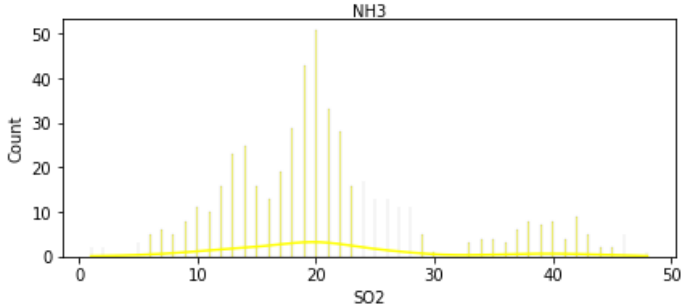


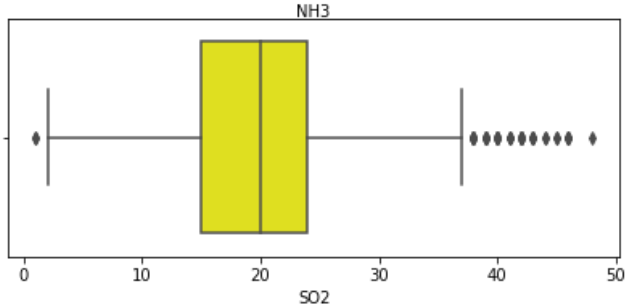


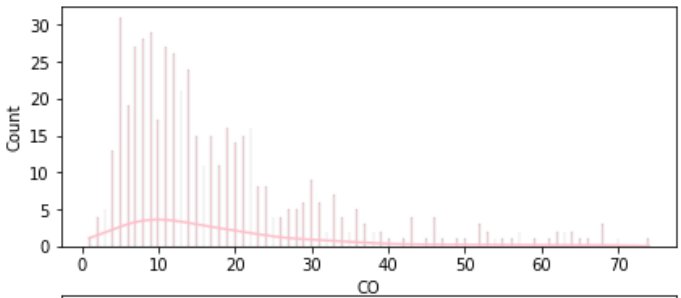


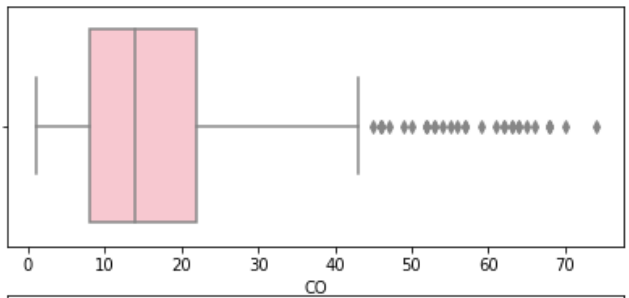

Count
NH3

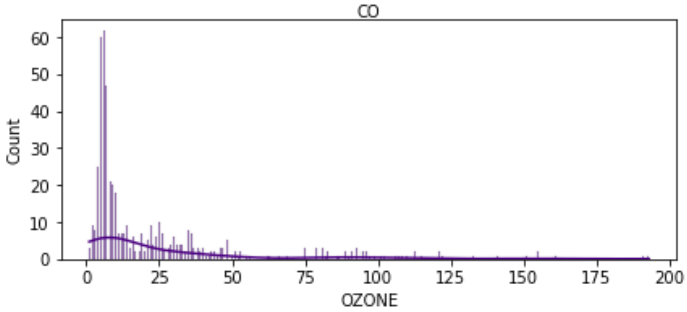

NH3

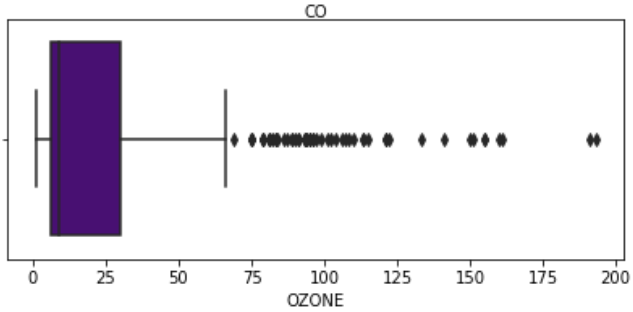

Count
SO2

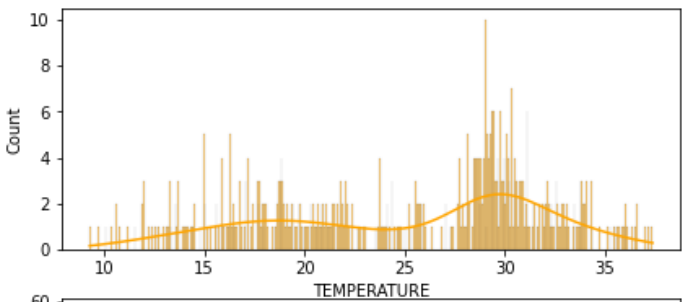

SO2

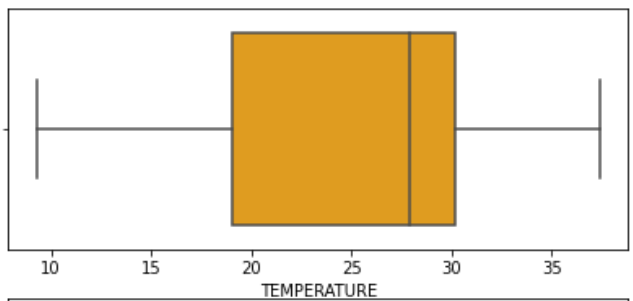

Count
CO

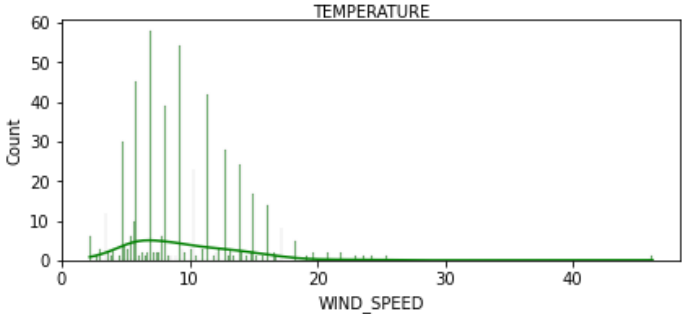

CO

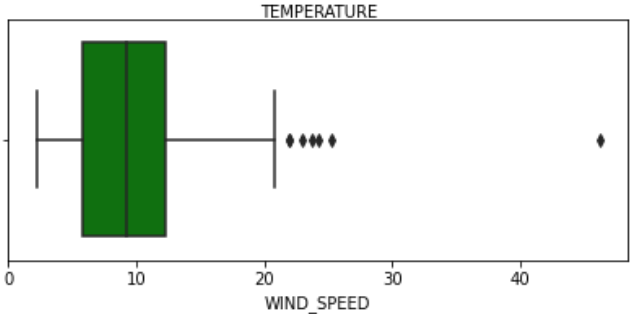

Count
OZONE

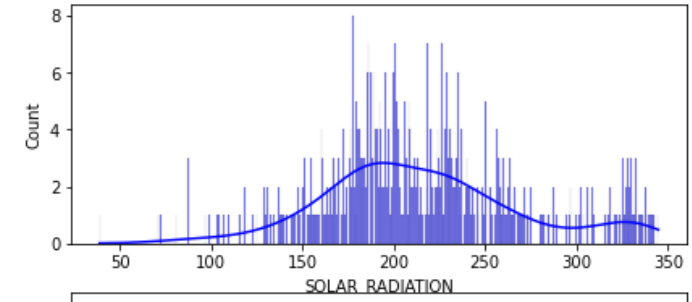

OZONE

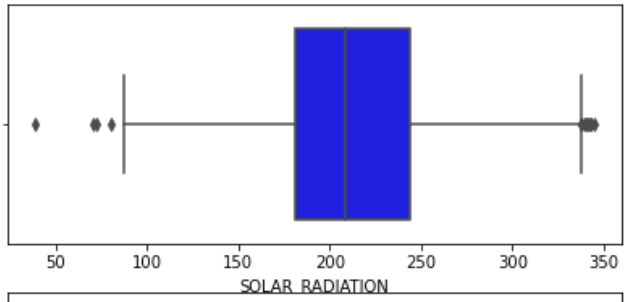

Count
TEMPERATURE

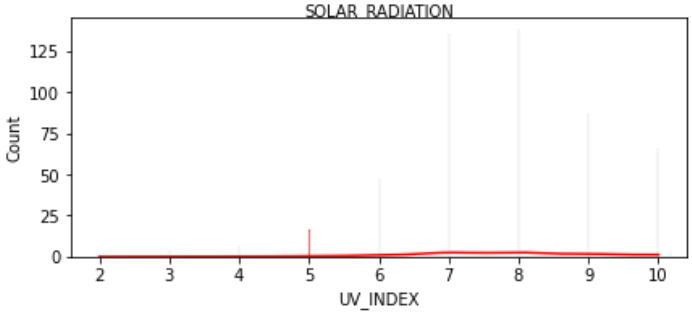

TEMPERATURE

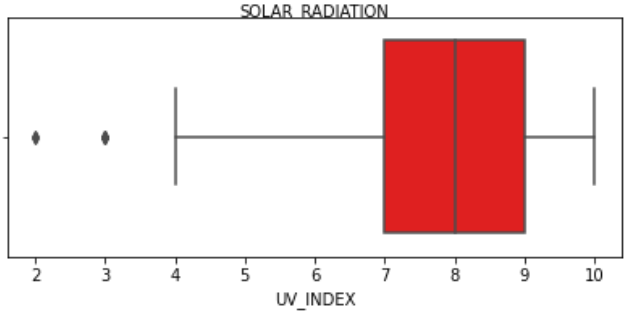

Count
WIND_SPEED

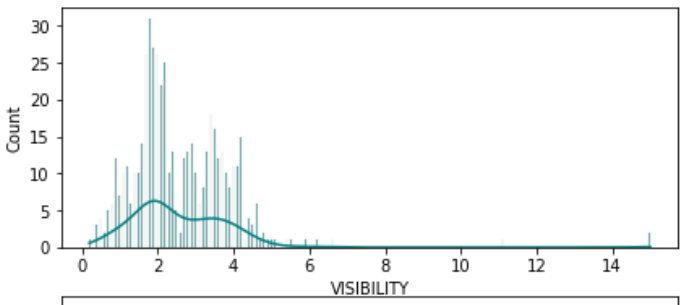

WIND_SPEED

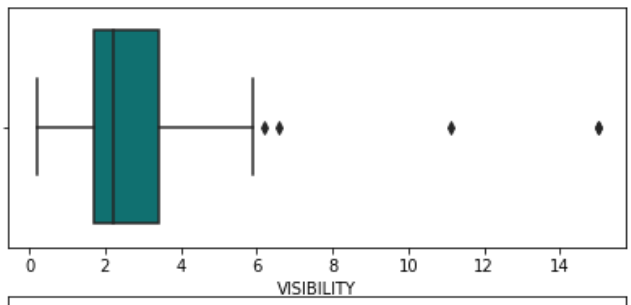

Count
SOLAR_RADIATION

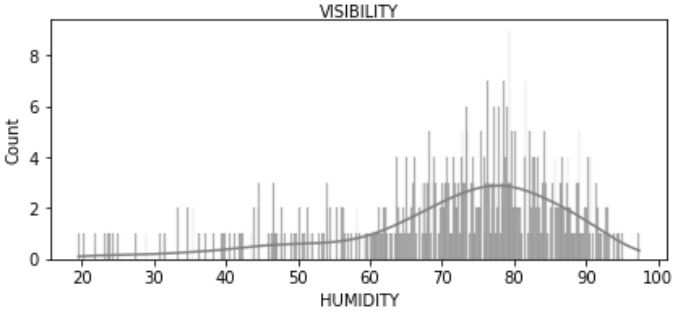

SOLAR_RADIATION

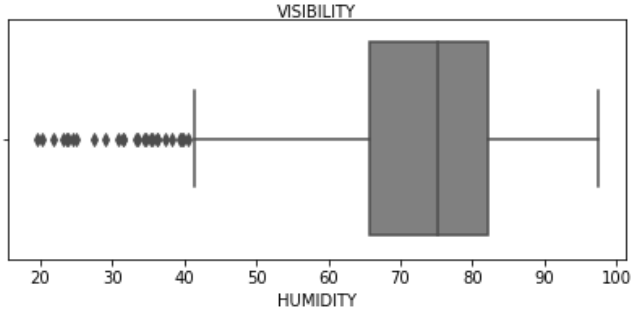

Count
UV_INDEX

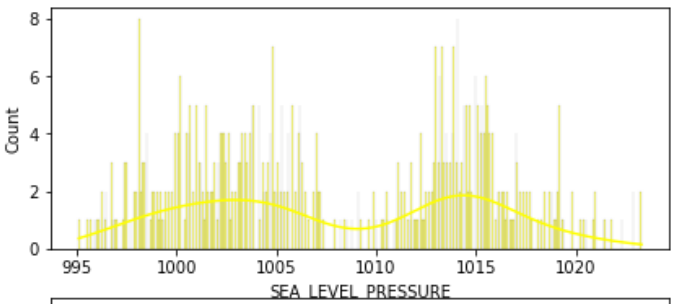

UV_INDEX

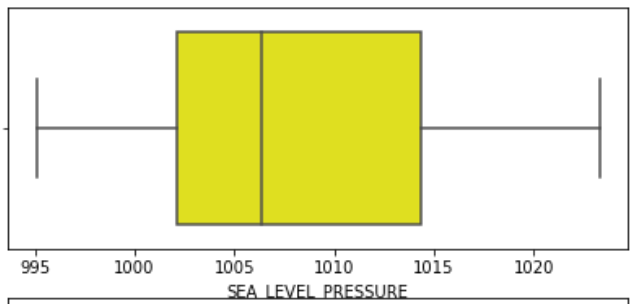

Count
VISIBILITY

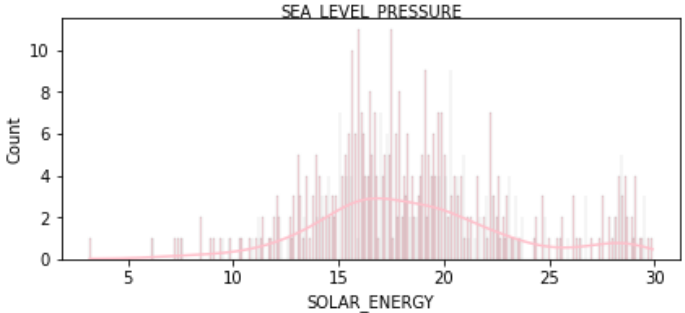

VISIBILITY

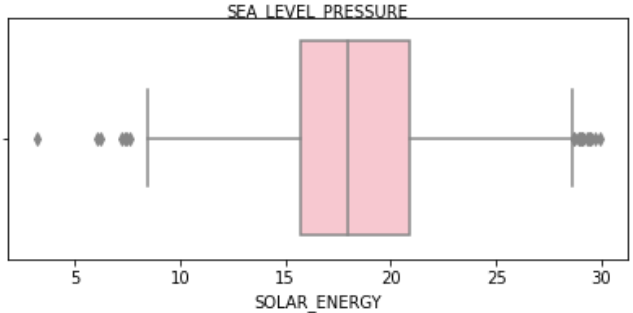

Count
HUMIDITY


HUMIDITY


Count
SEA_LEVEL_PRESSURE


SEA_LEVEL_PRESSURE


Count
SOLAR_ENERGY


SOLAR_ENERGY

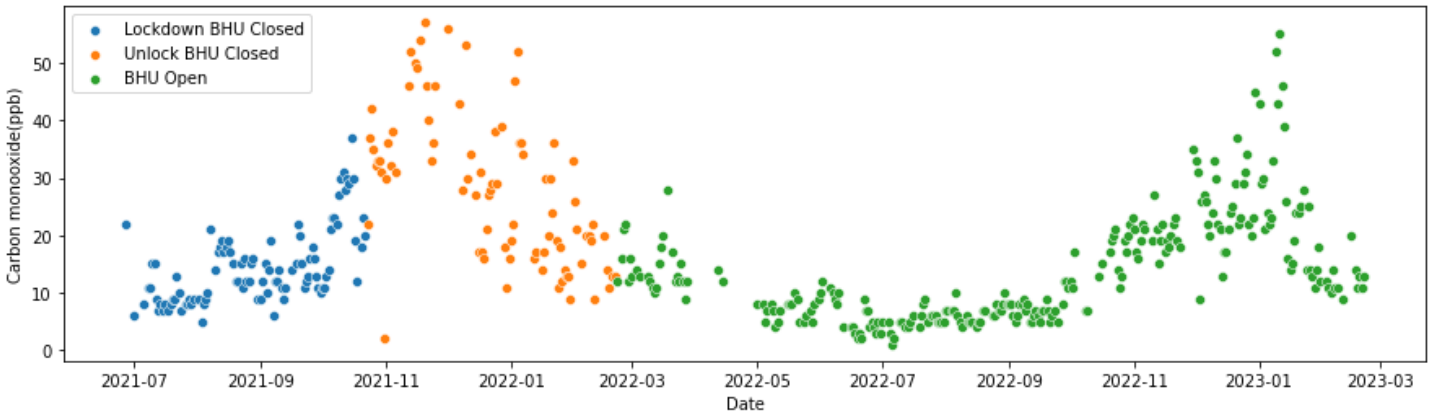
Variables summary after Cleaning-

Variable	Unit	Count	Mean	Standard Deviation	1st Quartile	Median	3rd Quartile	Minimum	Maximum
PM2.5	Micrograms per cubic meter of air	442	65.25	45.89	28	50	82.75	6	250
PM10	micrograms per cubic meter of air	442	67.22	29.18	43	67.5	88	11	143
NO_2	ppb	442	20.15	10.68	12	18	28	4	53
NH_3	ppm	442	3.98	2.56	2	3	6	1	12
SO_2	ppb	442	20.89	8.40	16	20	24	1	46
CO	ppm	442	16.65	11.20	8	14	22	1	57
OZONE	Dobson unit	442	22.45	26.96	6	9	28.75	1	122
Temperature	Celsius	442	24.80	6.78	18.77	27.77	30.05	9.27	37.38
Dew	Percentage	442	65.82	12.24	54.15	66.15	78.3	41	83.1
Humidity	Percentage	442	73.03	13.66	66.25	75.80	82.67	27.50	97.40
Precipitate	Millimeter(24hrs)	442	0.10	0.38	0	0	0.008	0	4.21
Precipitate cover	Millimeter	442	2.05	3.95	0	0	4.17	0	20.83
Wind speed	Kmph	442	9.22	3.89	5.85	9.2	11.4	2.2	21.9
Wind Direction status	Unitless	442	4.55	2.26	2	6	6	1	8
Sea level pressure	hPa	442	1008.37	7.02	1002.3	1007.05	1014.4	995.5	1023.3
Cloud cover	Percentage	442	36.33	30.80	5.07	33.9	63.57	0	98.4
Visibility	Km	442	2.54	1.08	1.8	2.2	3.4	0.2	6.6
Solar radiation	kWh/m^2	442	215.28	49.06	183.8	208.95	240.75	80.7	341.7
Solar energy	kWh/m^2	442	18.59	4.22	15.82	18.05	20.77	7.2	29.5
UV index	Unitless	442	7.81	1.30	7	8	9	4	10
Daylight duration	Hours	442	11.96	1.10	10.89	11.84	13.07	10.55	13.71
Moon phase	Percentage	442	0.51	0.28	0.27	0.53	0.75	0	1
Condition status	Unitless	442	3.27	1.61	1	3	5	1	5

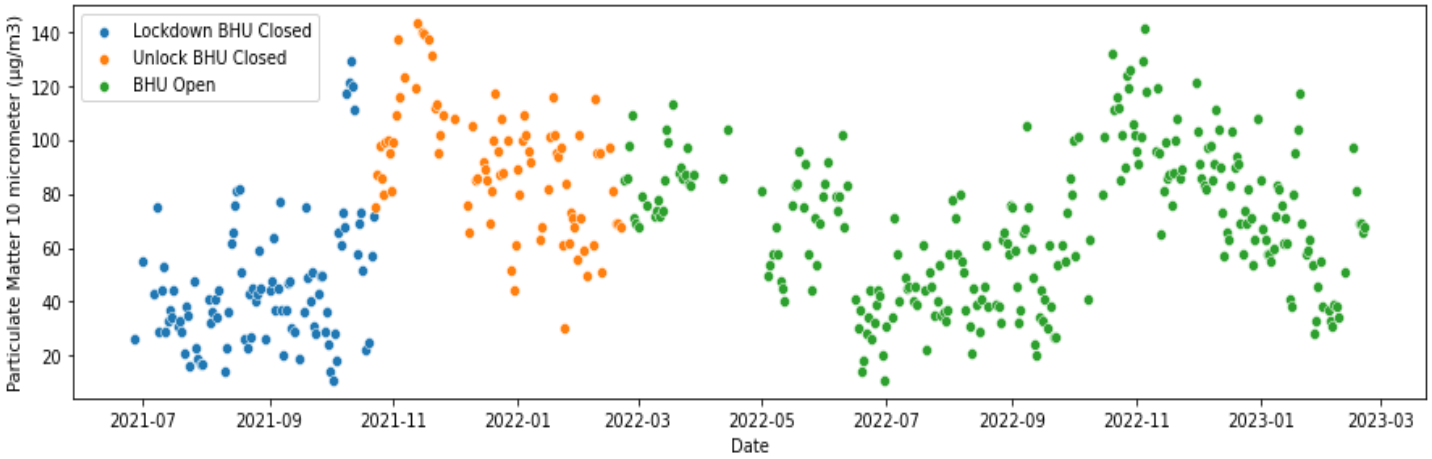
Visualization

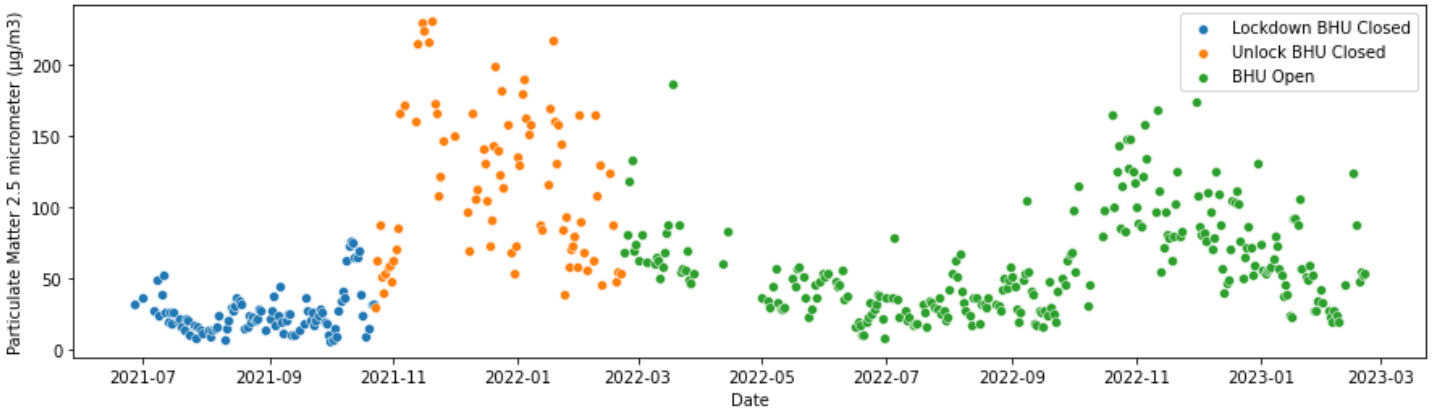
Variation in the factors deciding Air Quality Index in different phases of Lockdowns

We can see AQI is quite high during the unlock phase corresponding to low temperature because dew percentage is high which resist pollutant to settle down fastly.

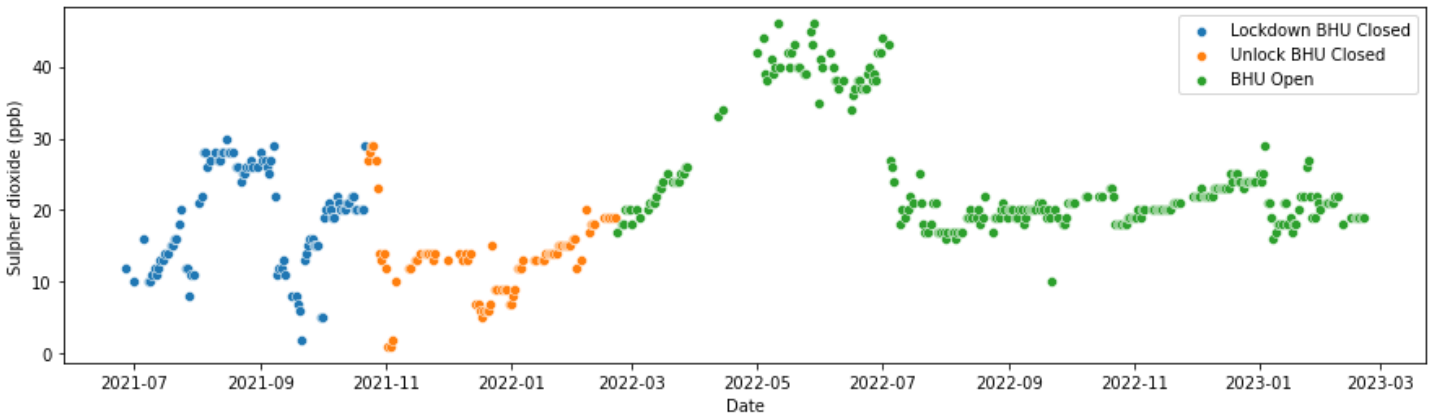


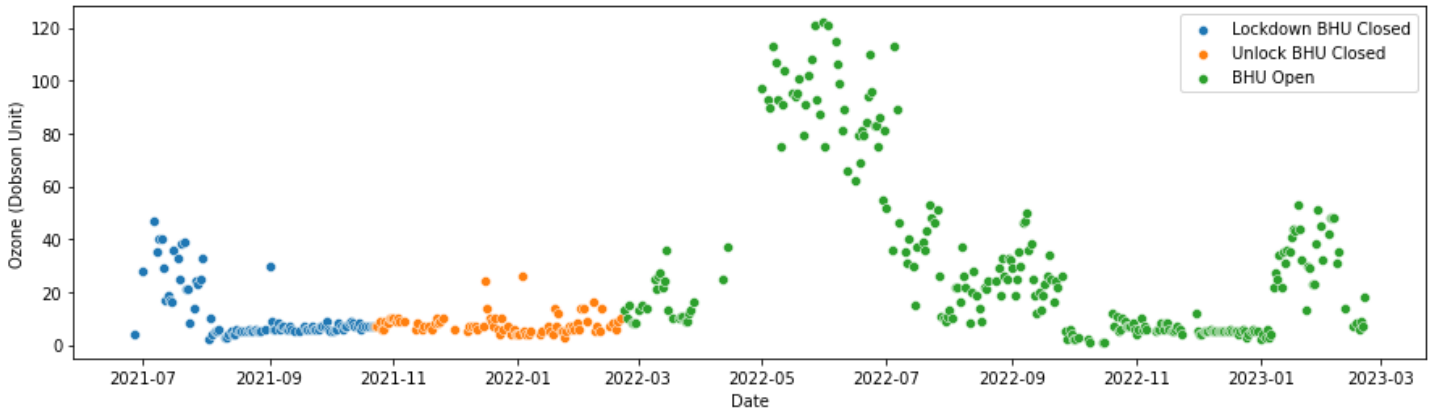
Carbon monoxide emitted when fuel's are not completely burned and not showing any significant relation between phases of lockdown rather showing trend with temperature.

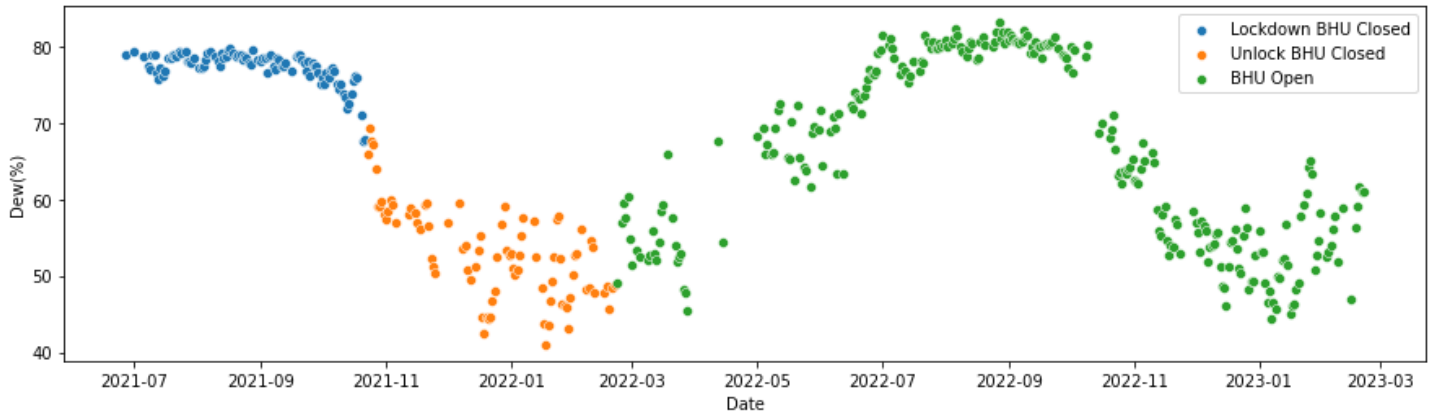


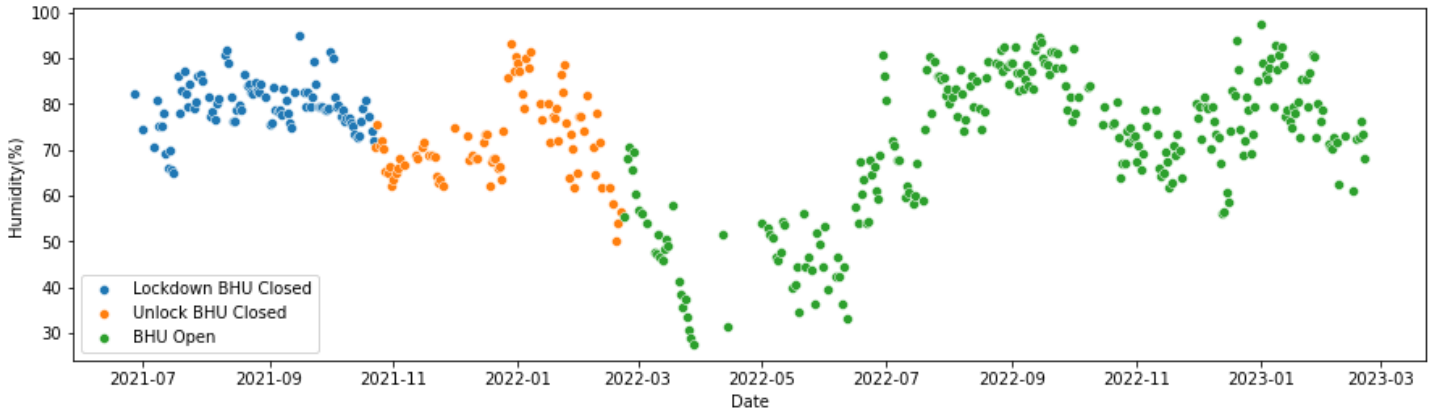


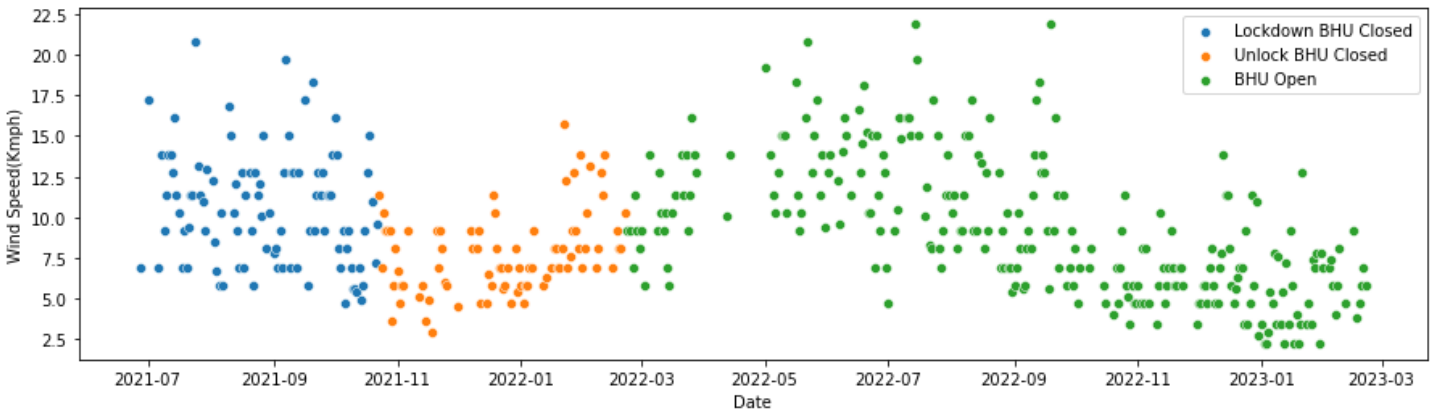
The PM2.5 and PM10 showing similar behavior as AQI.

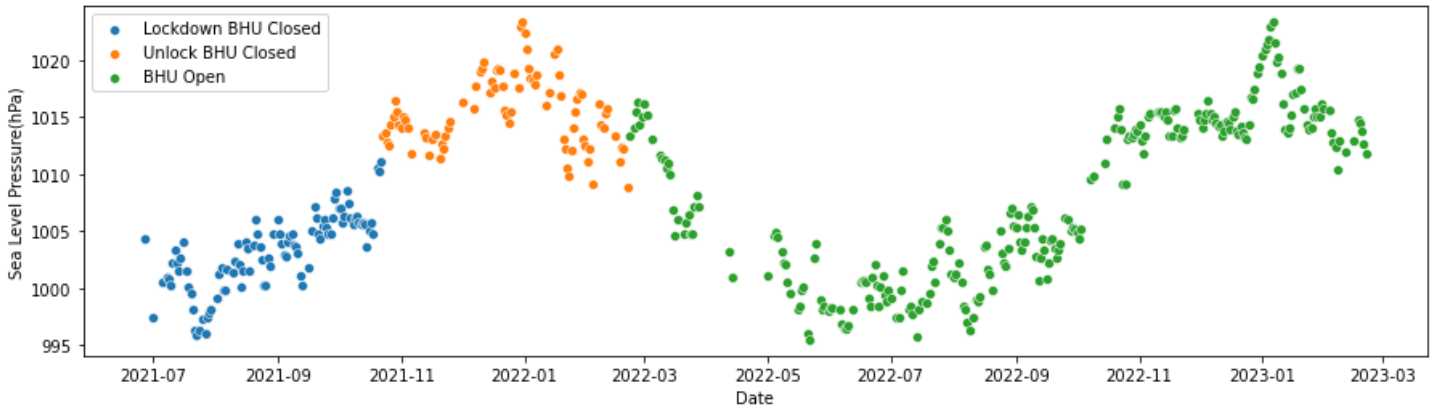




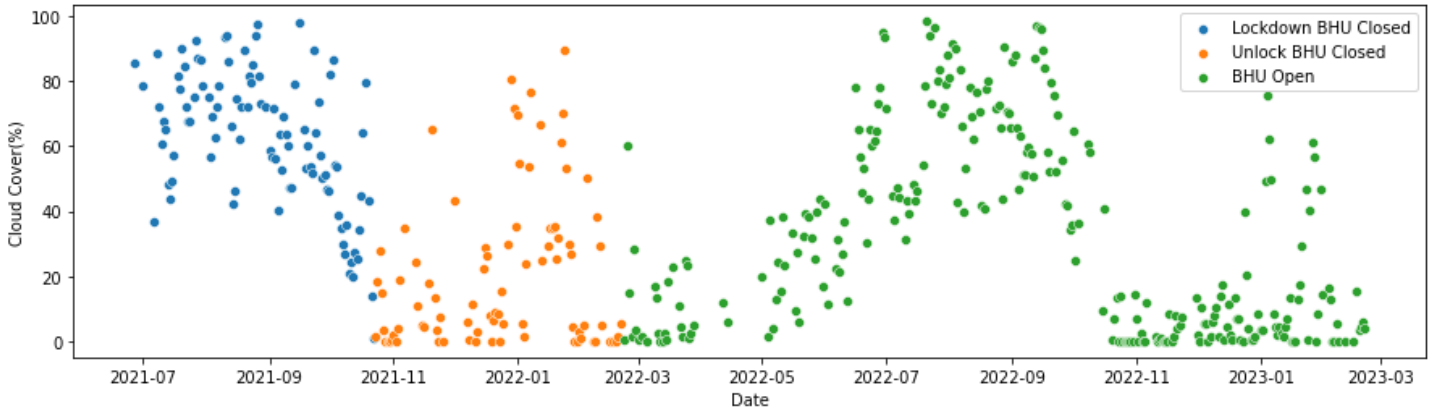


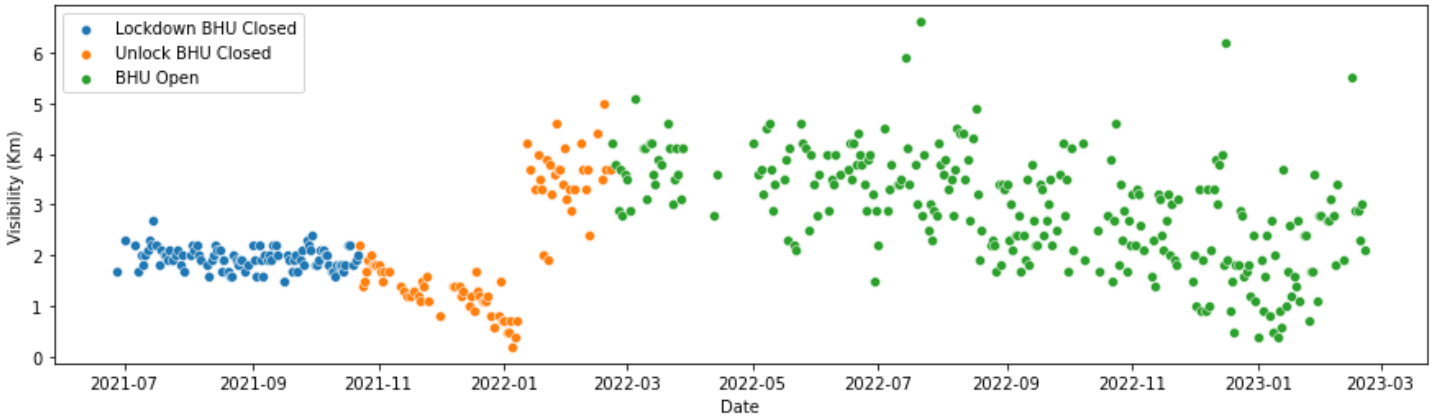


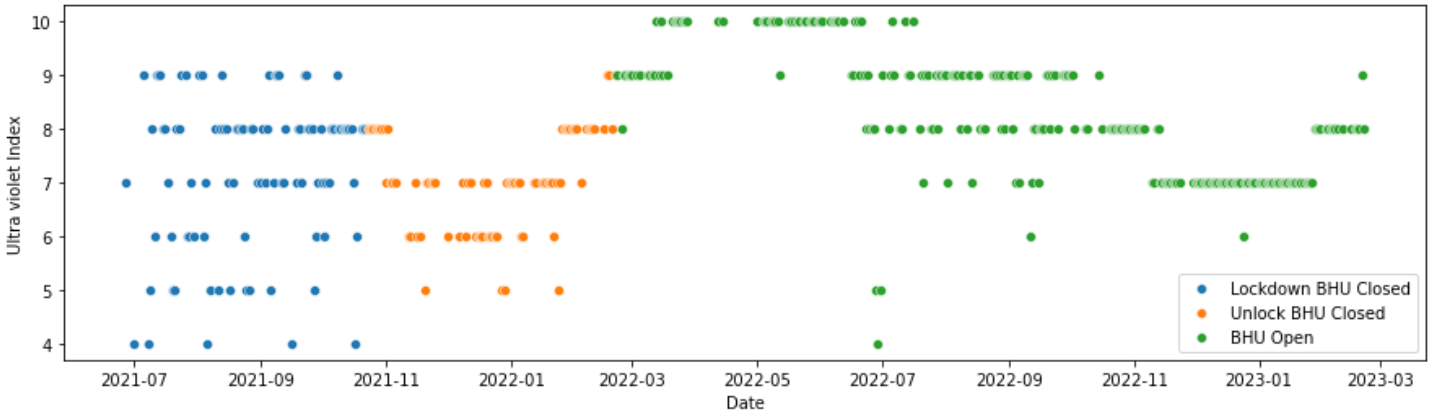


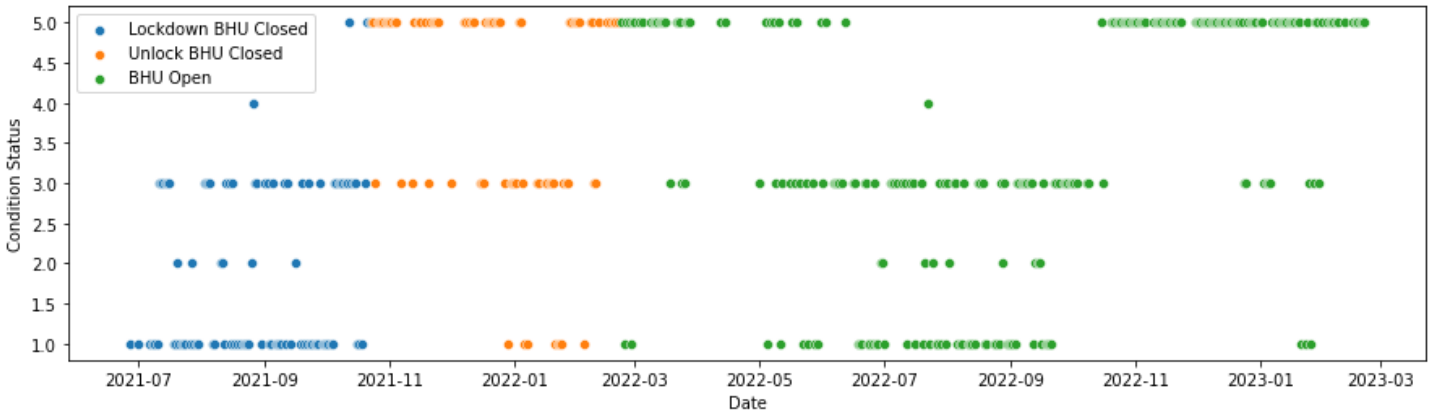


Sea level pressure also depends on atmospheric temperature, so don't seems to be significant for estimating AQI.

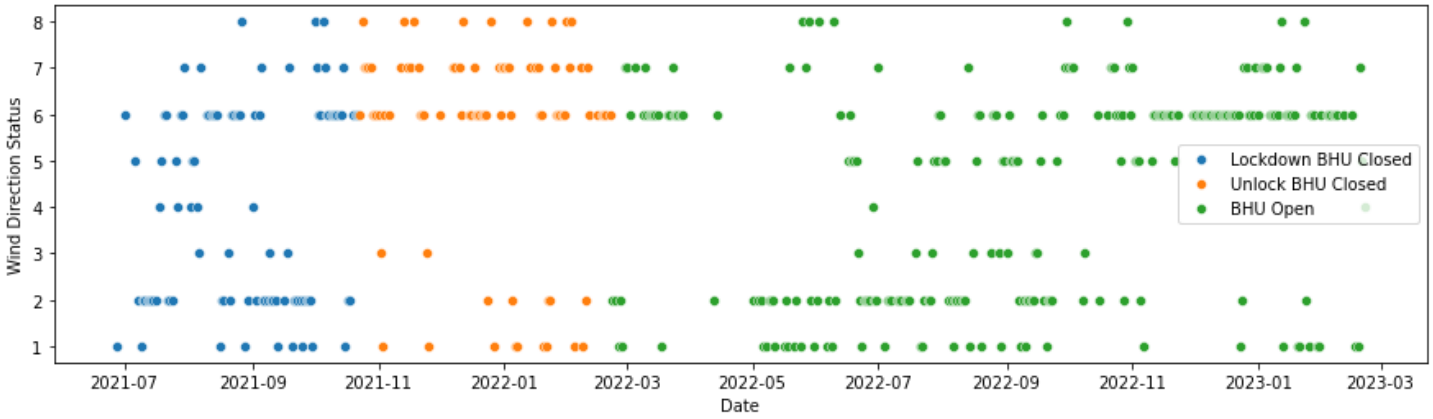




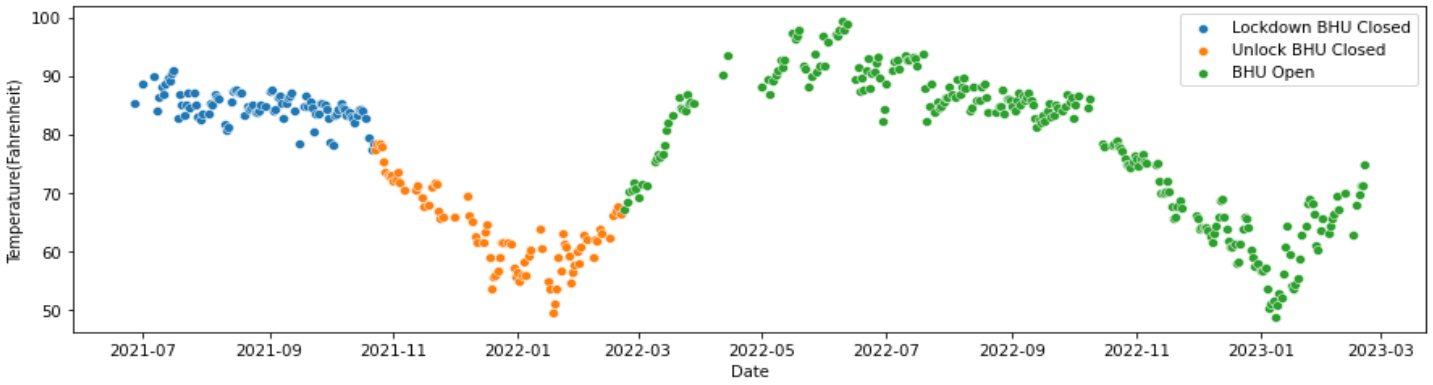


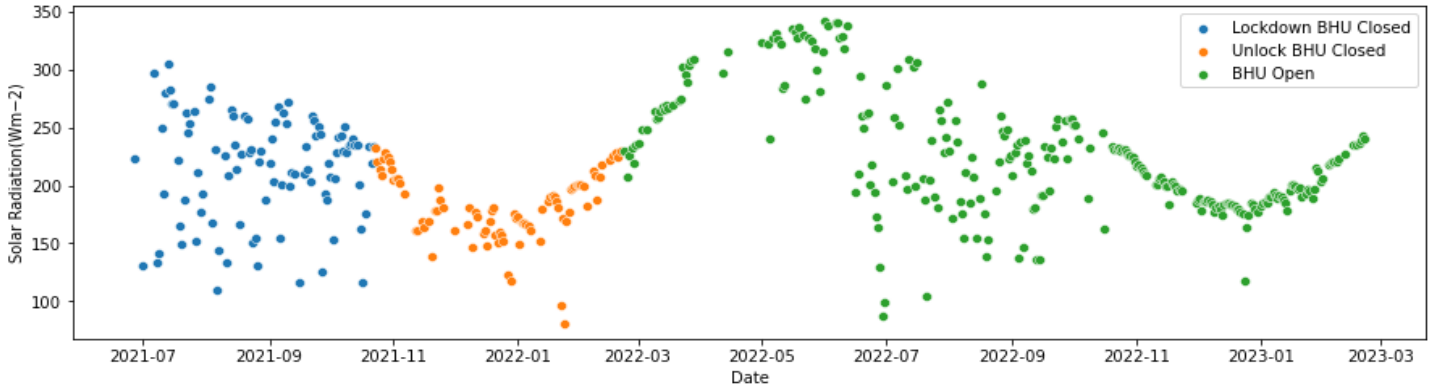


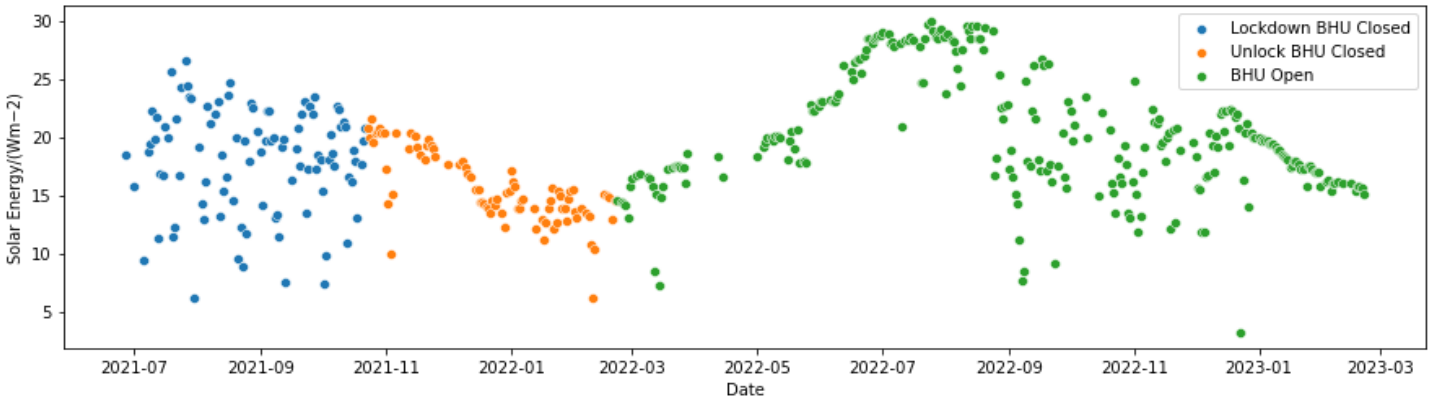
Here 1 represent Rain with partial cloudy condition, 2 represent Rain with overcast condition, 3 represent Partially cloudy, 4 represent Overcast condition and 5 represent Clear weather.



Here **Northeast** direction represented by **1**, **East** direction represented by **2** ,**Southeast** direction represented by **3**, **South** direction represented by **4**, **Southwest** direction represented by **5**, **West** direction represented by **6**, **Northwest** direction represented by **7** and **North** direction represented by **8.**







The Temperature, Solar radiation and Solar energy show similar trend as we know they are correlated with each other.

Hypothesis Testing

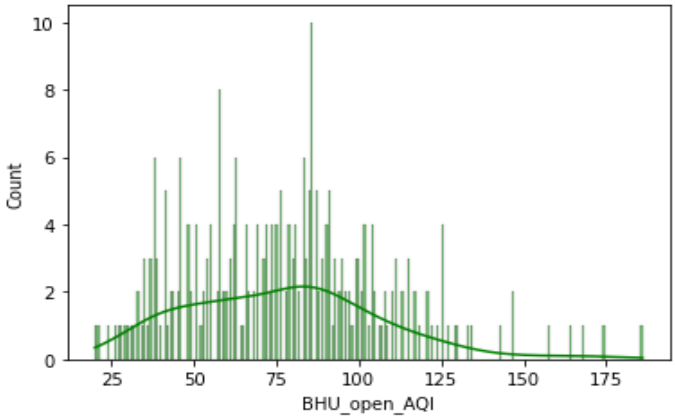
Testing whether mean AQI for different phases are same?

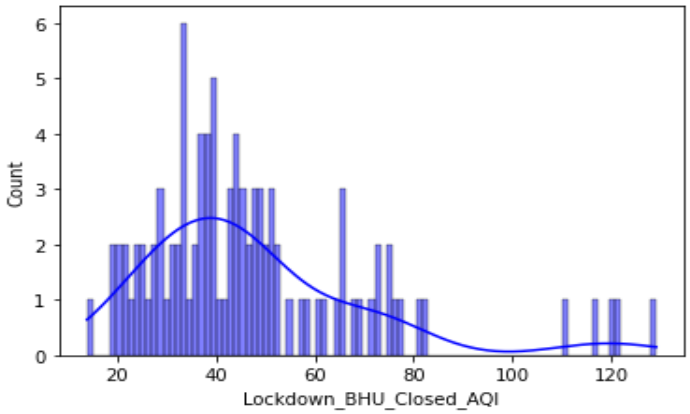
To compare means of different samples we can use t-test which is a parametric test with 3 basic assumptions

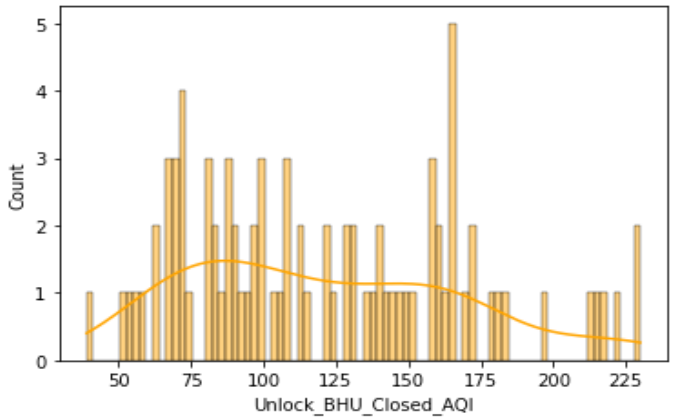
1. Samples are independent
2. Samples are (approximately) normally distributed
3. Groups have similar amount of variance

Also,

Phases	Number of days	AQI Total	Mean AQI	Variance
BHU open	261	20147	77.1916	894.7323
BHU closed during lockdown	97	4628	47.7113	533.8533
BHU closed during unlock	84	10197	121.3928	2255.2052







We can see that all the groups are not approximately normally distributed and variance between group differ a lot, hence t-test can't be used.

We can use a non parametric test named as **Kruskal-Wallis Test(U test)** which can be used when the normality and equality of variance assumptions are voilated, but having few assumptions

- There are at least three independently drawn random samples group
- Each group has at least 5 observations

Test Statistic:

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j} \right] - 3(n+1)$$

Where,

n_j=Number of observation in j^{th} group
T_j=Sum of rank of j^{th} sample
C=Number of groups
$n=\sum n_j$

Steps:

- Define Null hypothesis and alternative hypothesis
- Rank the group observations in the combined series and sum the ranks of each group observations
- Compute Test Statistics
- Take the tabulated value of $\chi^2_{(c-1,\alpha)}$
- If the calculated value of $H>\chi^2_{(c-1,\alpha)}$, we reject null hypothesis (Kruskal_Wallis Test is a Right-Tailed test)

Testing

Null hypothesis : Means are equal i.e $\mu_{\text{BHU open}} = \mu_{\text{BHU closed in lockdown}} = \mu_{\text{BHU closed in unlock}}$

Alternative hypothesis : At least two μ_i^s are different

Phases	(n_j)	T_j	T_i^2/n_j
BHU open	261	58883	13284320.65
BHU closed during lockdown	97	10166	1065438.722
BHU closed during unlock	84	28047	9364692.964

H=124.3437586 and $\chi^2_{(2,0.05)} = 0.196$

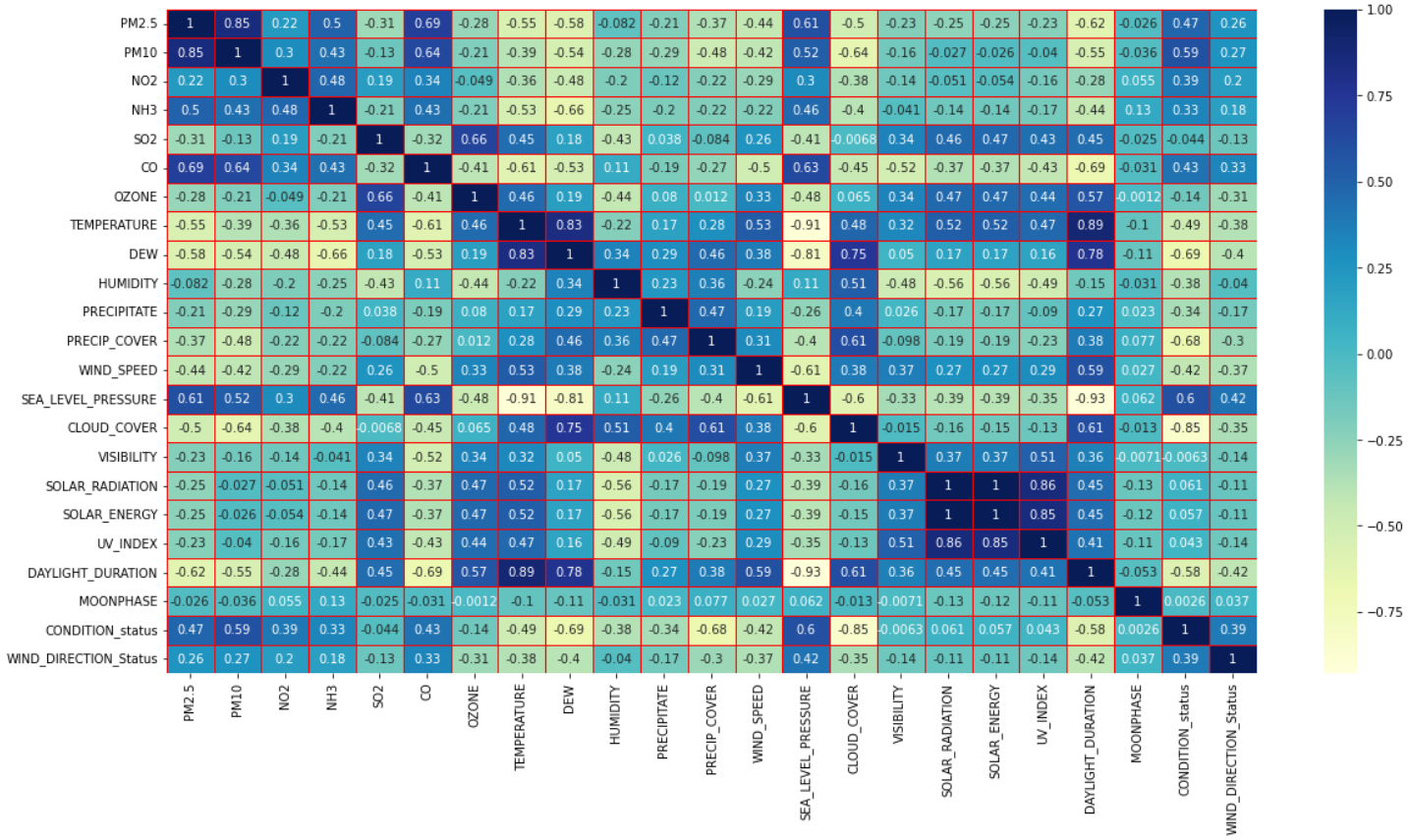
So we reject the null hypothesis, Hence different phases have different means of AQI.

Fitting Multiple Linear Regression Model

1.Deducing Predictor and Target Variables

Our target variable or dependent variable is AQI where as predictor or independent variables are to be deduced after checking multicollinearity between all available variables.

Checking for possible non independent variables through Correlation plot-



As we know both PM2.5 and PM10 are important part for estimating Air Quality Index so we can't drop instead of having high correlation, so transforming them in new Variable named PM-comp as we know transformation can be done between highly dependent variables to get rid from the case or multicollinearity.
Let,

PM-comp=0.5*PM2.5 +0.5*PM10

Also its hard to figure out the dropping variables because there are many highly correlated variable, So moving to Variance Inflation Factor(VIF) method to check amount of correlation between variables.

Feature	VIF
PM_comp	11.75807
NO2	9.444295
NH3	8.642902
SO2	19.051058
CO	11.100581
OZONE	4.677473
TEMPERATURE	1401.390747

DEW	3220.318103
HUMIDITY	1080.975193
PRECIPITATE	1.264188
PRECIP_COVER	3.055903
WIND_SPEED	12.159964
SEA_LEVEL_PRESSURE	829.396449
CLOUD_COVER	17.179848
VISIBILITY	9.075108
SOLAR_RADIATION	2075.991702
SOLAR_ENERGY	1884.363797
UV_INDEX	188.783953
DAYLIGHT_DURATION	814.802506
MOONPHASE	1.061326
CONDITION_status	26.497076
WIND_DIRECTION_Status	7.369136

After dropping variables one by one having maximum VIF we came at point where the left variables have VIF less than 20.

Feature	VIF
PM_comp	10.731261
NO2	8.601273
NH3	6.381217
SO2	15.680389
CO	9.61766
OZONE	3.664622
PRECIPITATE	1.182719
PRECIP_COVER	2.467156
WIND_SPEED	8.870698
CLOUD_COVER	4.923067
VISIBILITY	7.046403
MOONPHASE	1.04341
CONDITION_status	13.580909
WIND_DIRECTION_Status	6.409448

These are the variables having VIF<20 hence these variables having very small correlation, So using these variables for model creation.

2. Fitting of Multiple Linear Regression model

Multiple linear regression model also uses Ordinary least square method for estimation of coefficients.

Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression). Least squares stand for the minimum squares error (SSE). Maximum likelihood and Generalized method of moments estimator are alternative approaches to OLS.

We used 85% of observations to train our model-

OLS Regression Results

Dep. Variable:	AQI	R-squared:	0.946
Model:	OLS	Adj. R-squared:	0.944
Method:	Least Squares	F-statistic:	525.0
Date:	Thu, 30 Mar 2023	Prob (F-statistic):	1.49e-220
Time:	01:11:01	Log-Likelihood:	-1377.2
No. Observations:	375	AIC:	2780
Df Residuals:	362	BIC:	2831
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.9375	4.962	-4.421	0.000	-31.696	-12.179
PM_comp	1.0973	0.023	48.071	0.000	1.052	1.142
NO2	-0.1666	0.064	-2.620	0.009	-0.292	-0.042
NH3	0.6284	0.285	2.203	0.028	0.067	1.189
SO2	0.3522	0.093	3.788	0.000	0.169	0.535
CO	0.4497	0.081	5.582	0.000	0.291	0.608
OZONE	0.3791	0.027	14.279	0.000	0.327	0.431
PRECIP_COVER	0.0307	0.174	0.176	0.860	-0.312	0.373
WIND_SPEED	0.4536	0.172	2.632	0.009	0.115	0.793
CLOUD_COVER	0.0836	0.033	2.533	0.012	0.019	0.148
VISIBILITY	0.2980	0.644	0.463	0.644	-0.968	1.564
CONDITION_status	-0.7938	0.666	-1.192	0.234	-2.103	0.515
WIND_DIRECTION_Status	0.4983	0.258	1.929	0.054	-0.010	1.006

Probability(F-Statistic) and F-Statistics: This tells the overall significance of the regression.
The "F value'' and "Prob(F)'' statistics test the overall significance of the regression model. Specifically, they test the null hypothesis that *all* of the regression coefficients are equal to zero. This tests the full model against a model with no variables and with the estimate of the dependent variable being the mean of the values of the dependent variable. The F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares. Its value will range from zero to an arbitrarily large number.
The value of Prob(F) is the probability that the null hypothesis for the full model is true (i.e., that all of the regression coefficients are zero). For example, if Prob(F) has a value of 0.01000 then there is 1 chance in 100 that all of the regression parameters are zero. This low a value would imply that at least some of the regression parameters are nonzero and that the regression equation does have some validity in fitting the data (i.e., the independent variables are not purely random with respect to the dependent variable).
*So our regression model is significant

R-squared: is a statistical measure that indicates how much of the variation of a dependent variable is explained by an independent variable.

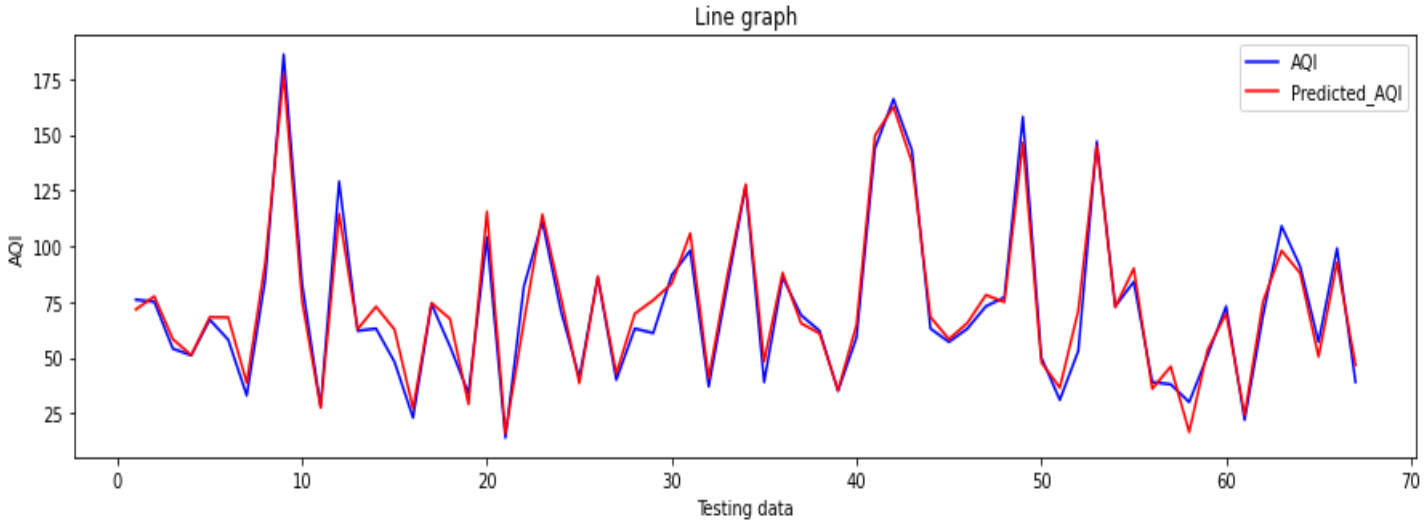
Adjusted R-squared: This measures the variation for a multiple regression model, and helps you determine goodness of fit.

Model is given by-

AQI = -21.937488445489095 + 1.09734313*PM_comp* -0.16656863NO2 + 0.62840198*NH3* + *0.35216351*SO2 + 0.44972472*CO* + *0.37906451*OZONE + 0.03069609*PRECIP_COVER* + *0.45359676*WIND_SPEED + 0.08356036*CLOUD_COVER* + *0.29800233*VISIBILITY - 0.7937893*CONDITION_status* + *0.49830492*WIND_DIRECTION_status

We can depict from the table that p-value (i.e., Pr (>|t|)) for all the exploratory variables PM-comp, NO_2, NH_3, SO_2, O3, CO, WIND_SPEED and CLOUD_COVER are statistically significant. The R^2 value for the model is 0.946 and adjusted R^2 value is 0.944. Since adjusted R^2 is approximately equal to R^2 and both are closer to 1 this means that most of the variability is explained by the model.

Testing of Model



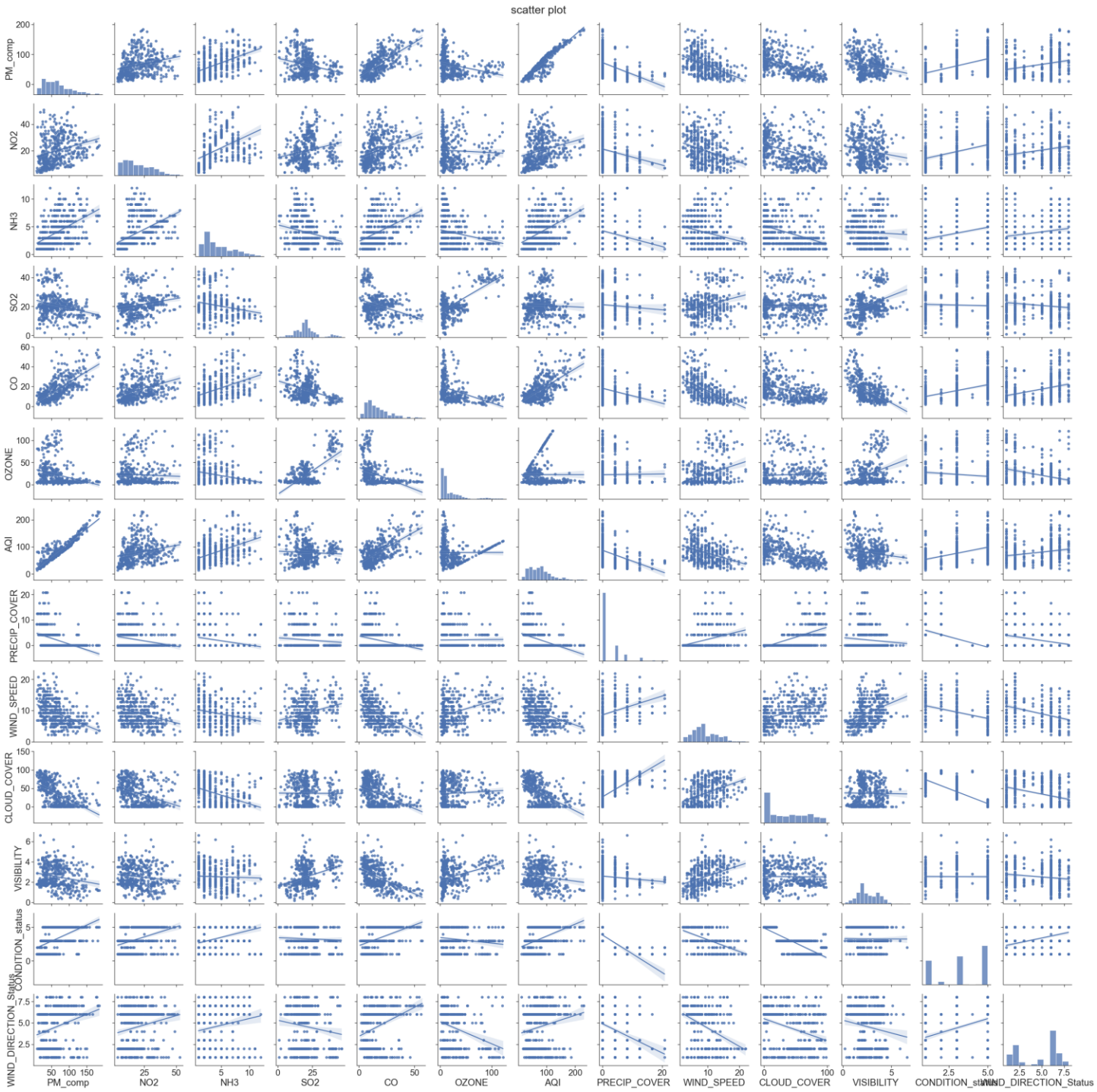
The above line graph shows the Actual AQI and Predicted AQI based on testing set.
Also Correlation between AQI and Predicted AQI = 0.9823246
Which tell us how better is the fitted model works.

Verifying Assumption of Multiple Linear Regression

Assumptions are-

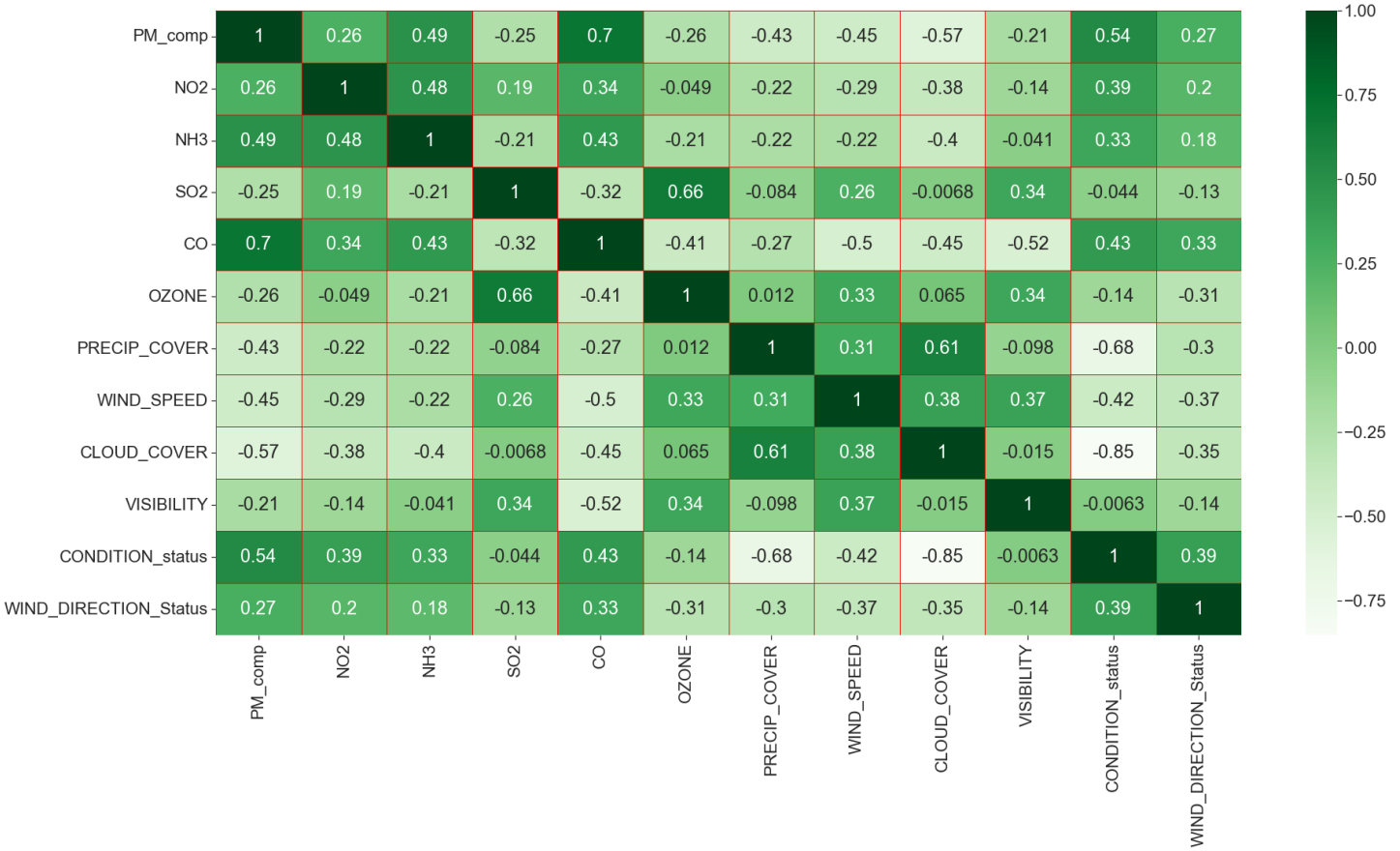
Linearity (A linear relationship between dependent and independent variables)
Multicollinearity (The independent variables are not highly correlated with each other)
Normality of residuals
Independence of residuals (no autocorrelation)
Homoscedasticity (The variance of residuals is constant)

Linearity (A linear relationship between dependent and independent variables)



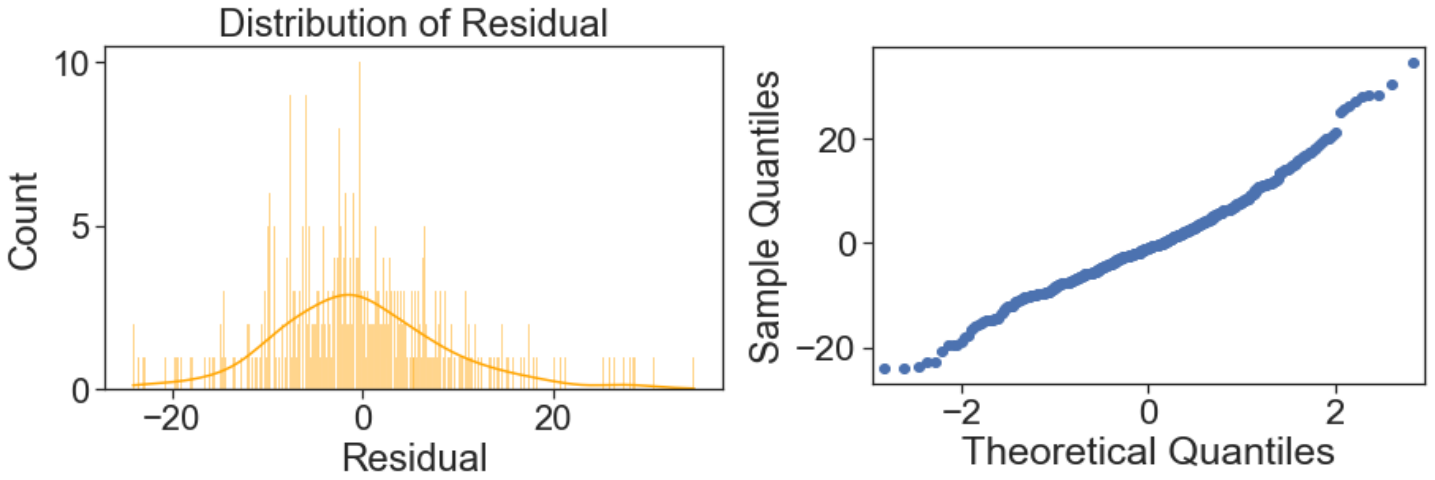
We can see from above pairplot that each of the explanatory variable are not linearly realated to the dependent variable i.e AQI.

Multicollinearity (The independent variables are not highly correlated with each other



Hence none of the predictor variables are highly correlated, so no multicollinearity.

3.Normality of residuals

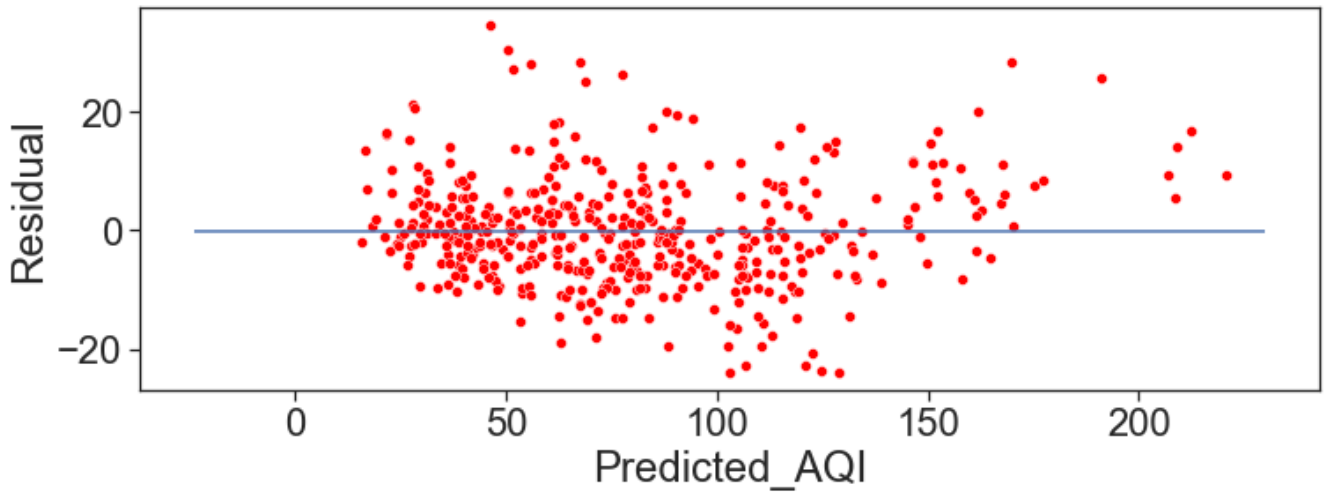


From above two plot (Histcount plot and QQ plot) we can say that residuals follow approximately normal distribution but its hard to tell whether residuals follow normal distribution or not, so now by **Shapiro wilk test-**

ShapiroResult(statistic=0.9775571823120117, p-value=2.4721909994696034e-06)

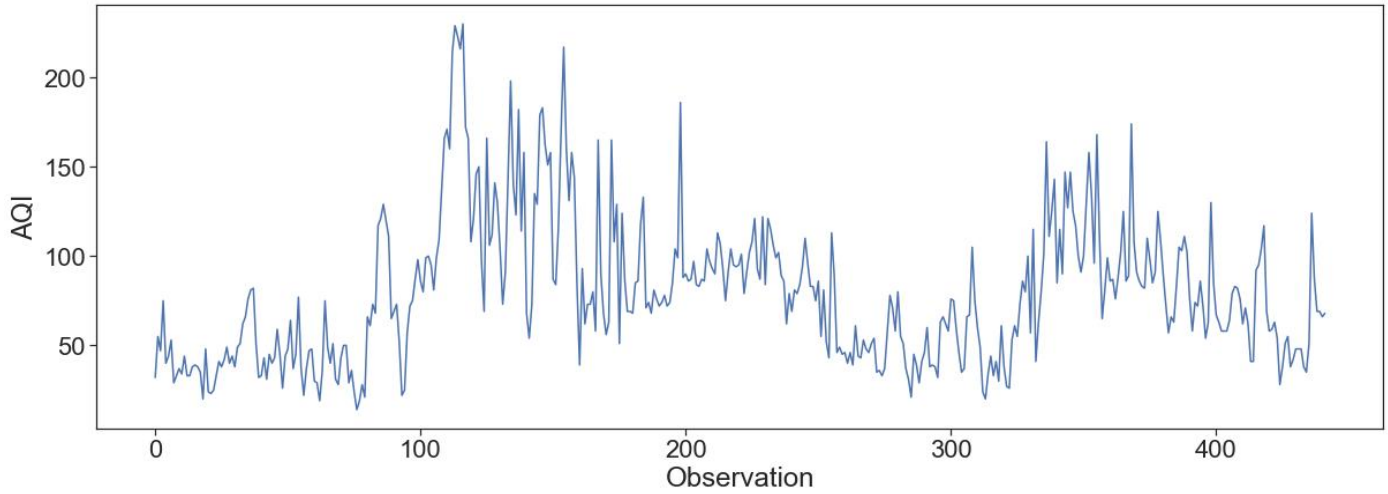
Since p-value is less than 0.05, we reject null hypothesis of Shapiro wilk test. This mean we have sufficient evidence to say that residual term does not follow normal distribution.

Independence of residuals (no autocorrelation)



From Residual vs estimated AQI plot we can see a slightly 'U' shape hence might be presence of autocorrelation.

Now using Durbin watson test-Durbin Watson Test is a test statistic used to detect the presence of autocorrelation at lag 1 in the residuals from a regressive analysis, but the series should be stationary.



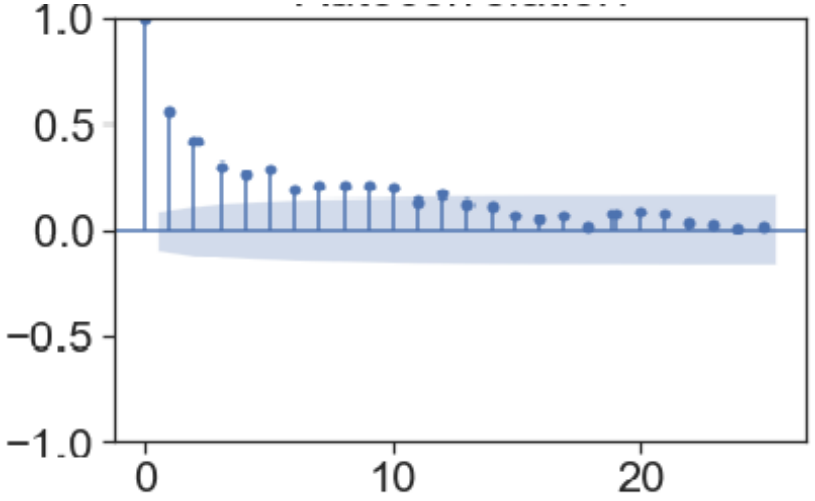
From the above plot its hard to tell whether AQI possesses Stationary Time series or not, so applying **Dickey-Fuller test stationarity of time series**

```
1. ADF : -2.941198195520525
2. P-Value : 0.04076268881097018
3. critical values :
      1% : -3.445613745346461
      5% : -2.868269325317112
      10% : -2.5703544951308404
```

As p-value is greater than critical values hence we cannot reject null hypothesis, Hence the AQI over Time is Non Stationary Time Series.

A Autoregressive model can be both stationary as well as non stationary where as finite Moving Average model is always Stationary. So it is quite obvious that we can only fit AR model for this AQI Data. The AR part involves regressing the variable on its own lagged (i.e., past) values. The MA part involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past.

From above Autocorrelation plot we can say that this will not follow an AR(1) model, hence can't use Durbin_Watson test.



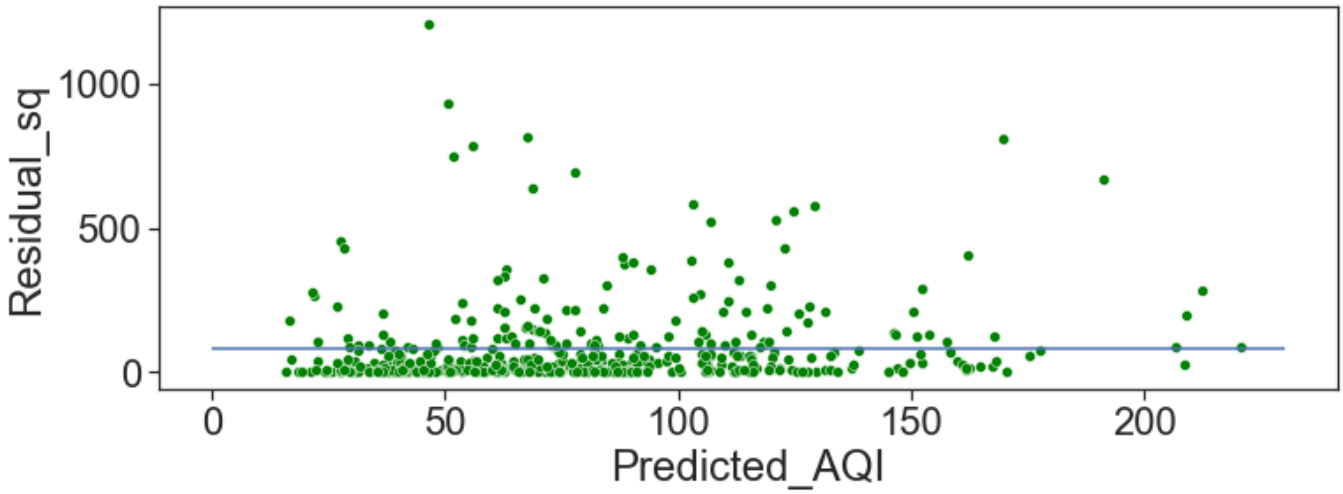
Breusch_Godfrey LM Test

Null hypothesis that no Autocorrelation

Lag	Breusch_Godfrey LM Test Statistic	Breusch_Godfrey LM Test P_Value
1	148.442618	0.0
2	158.347942	0.0
3	158.91049	0.0
4	161.955075	0.0
5	168.99387	0.0

For different lags the Breusch Godfrey Statistic is greater than P-value hence we can say that Autocorrelation is present.

Homoscedasticity (The variance of residuals is constant)



From residual square vs Predicted plot we are not sure about presence of homoscedasticity.

Goldfeld-Quandt test to check Homoscedasticity of Residuals

Step 1:Arrange the observations in ascending order of Xi. If there are more than one explanatory variables(X) then you choose the one regarding which you have a concern that with this variable the error variance is positively related and arrange in ascending order according to this variable. In other words, you can choose any one of them to arrange.

Step 2: Omit 'c' central observations and divide the remaining (n-c) observations into two groups containing (n-c)/2 observations each. The first (n-c)/2 observations belong to the first group(the smaller variance group) and the remaining (n-c)/2 observations belong to the second group(the larger variance group).

Step 3: Fit a separate regression model for the first group and obtain RSS1. Also, fit a separate regression model on the second group and obtain RSS2.

RSS = Residual Sum of Squares = u_i^2
$u_i = Y_{predicted} - Y_{calculated}$
This RSS each have (n-c)/2 – k or (n-c-2k)/2 degrees of freedom, where k is the number of parameters to be estimated.

Step 4: Compute the Test Statistic

$F_{cal}=\{RSS_2 / df\}/\{RSS_1 / df\}$

Step 5: Find out the critical value

Use the F Table to find out the critical value for the given level of significance(alpha). In this test, the values of df_1 and df_2 are the same($df_1=df_2$).

Step 6: Compare $F_{critical}$ and $F_{calculated}$ and state the result.

If $F_{calculated} < F_{critical}$; Accept the Null Hypothesis.
If $F_{calculated} > F_{critical}$; Reject the Null Hypothesis.

Observation-

RSS_1 = 11340.537568549102 RSS_2 = 24968.55540030007
Degree of freedom for both residual sum of square is (N-C-2K)/2=195

F_{cal}=2.201708274354275
$F_{critical}$ (195,195) at Probability level of 0.05=1.266

F_{cal} greater than $F_{critical}$ at chosen level of significance so we can reject the hypothesis of homoscedasticity i.e, heteroscedasticity is very likely.

Select ARIMA Model for Time Series Using Box-Jenkins Methodology

Box-Jenkins Methodology

The Box-Jenkins methodology is a five-step process for identifying, selecting, and assessing conditional mean models (for discrete, univariate time series data).

Determine whether the time series is stationarity. If the series is not stationary, successively difference it to attain stationarity. The sample autocorrelation function (ACF) and partial autocorrelation function (PACF) of a stationary series decay exponentially (or cut off completely after a few lags).

Identify a stationary conditional mean model for the series. The sample ACF and PACF functions can help with this selection. For an autoregressive (AR) process, the sample ACF decays gradually, but the sample PACF cuts off after a few lags. Conversely, for a moving average (MA) process, the sample ACF cuts off after a few lags, but the sample PACF decays gradually. If both the ACF and PACF decay gradually, consider an ARMA model.

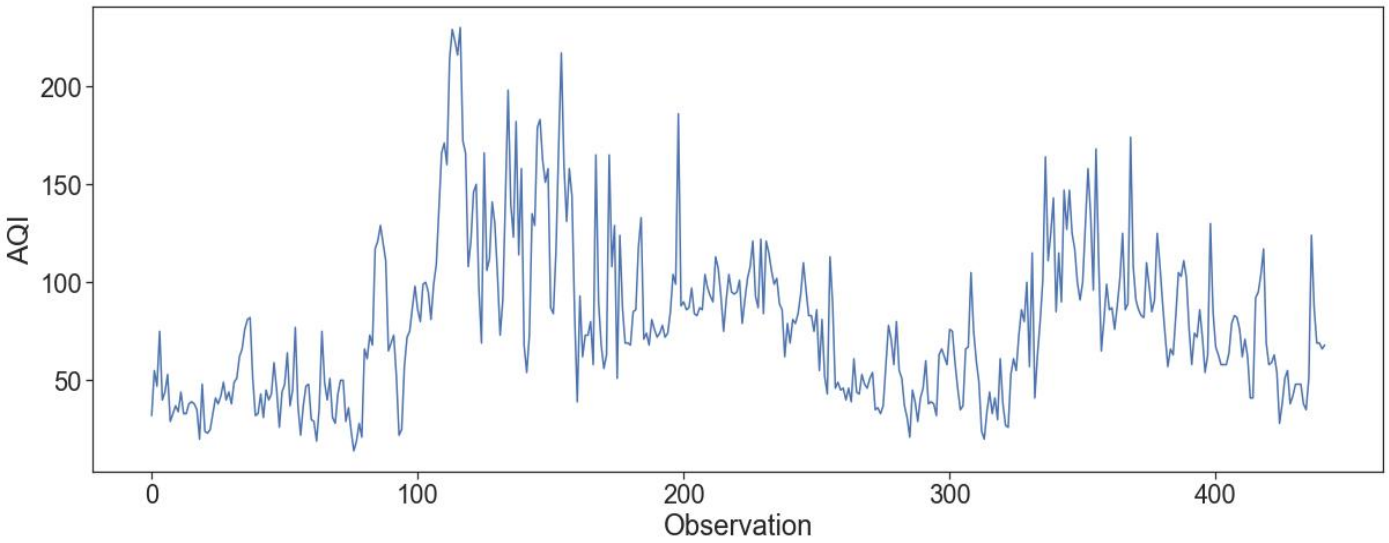
Create a model template for estimation, and then fit the model to the series. When fitting nonstationary models in Econometrics Toolbox™, you do not need to manually difference the series and fit a stationary model. Instead, you can use the series on the original scale, and create an arima model object with the desired degree of nonseasonal and seasonal differencing. Fitting an ARIMA model directly is advantageous for forecasting: forecasts are returned on the original scale (not differenced).

Conduct goodness-of-fit checks to ensure the model describes the series adequately. Residuals should be uncorrelated, homoscedastic, and normally distributed with constant mean and variance. If the residuals are not normally distributed, you can change the innovation distribution to a Student's t.

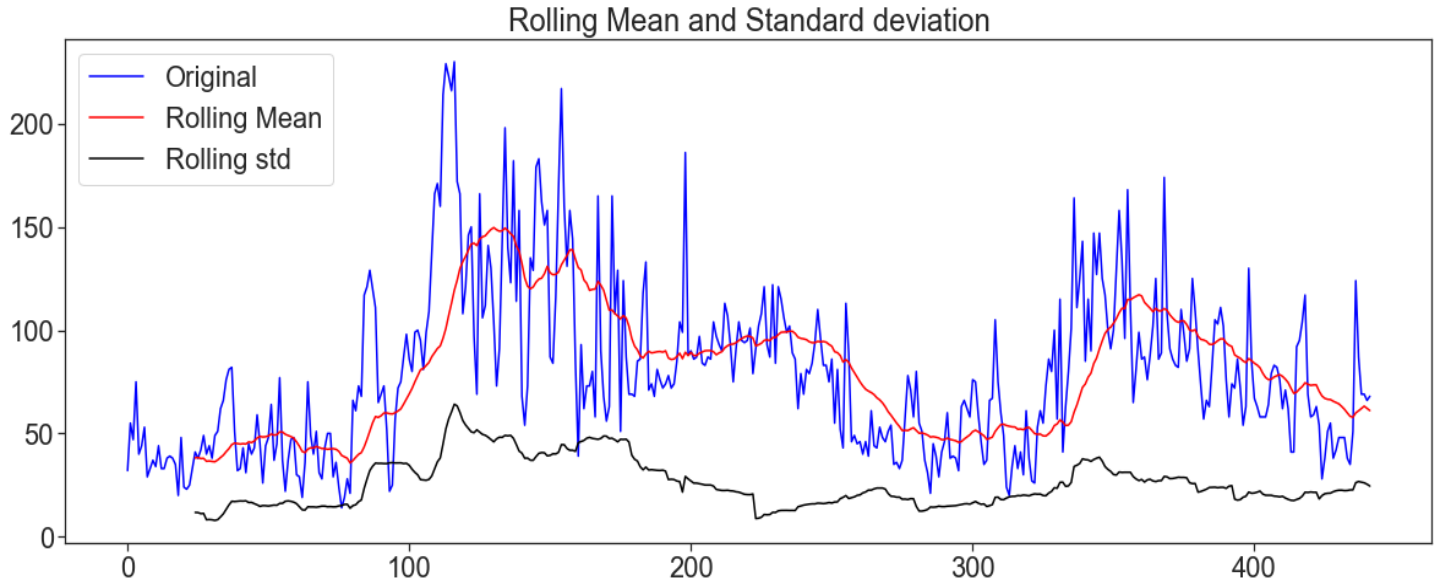
After choosing a model—and checking its fit and forecasting ability—you can use the model to forecast or generate Monte Carlo simulations over a future time horizon.

Fitting a Time Series Model

1. **Checking Stationarity of AQI series**

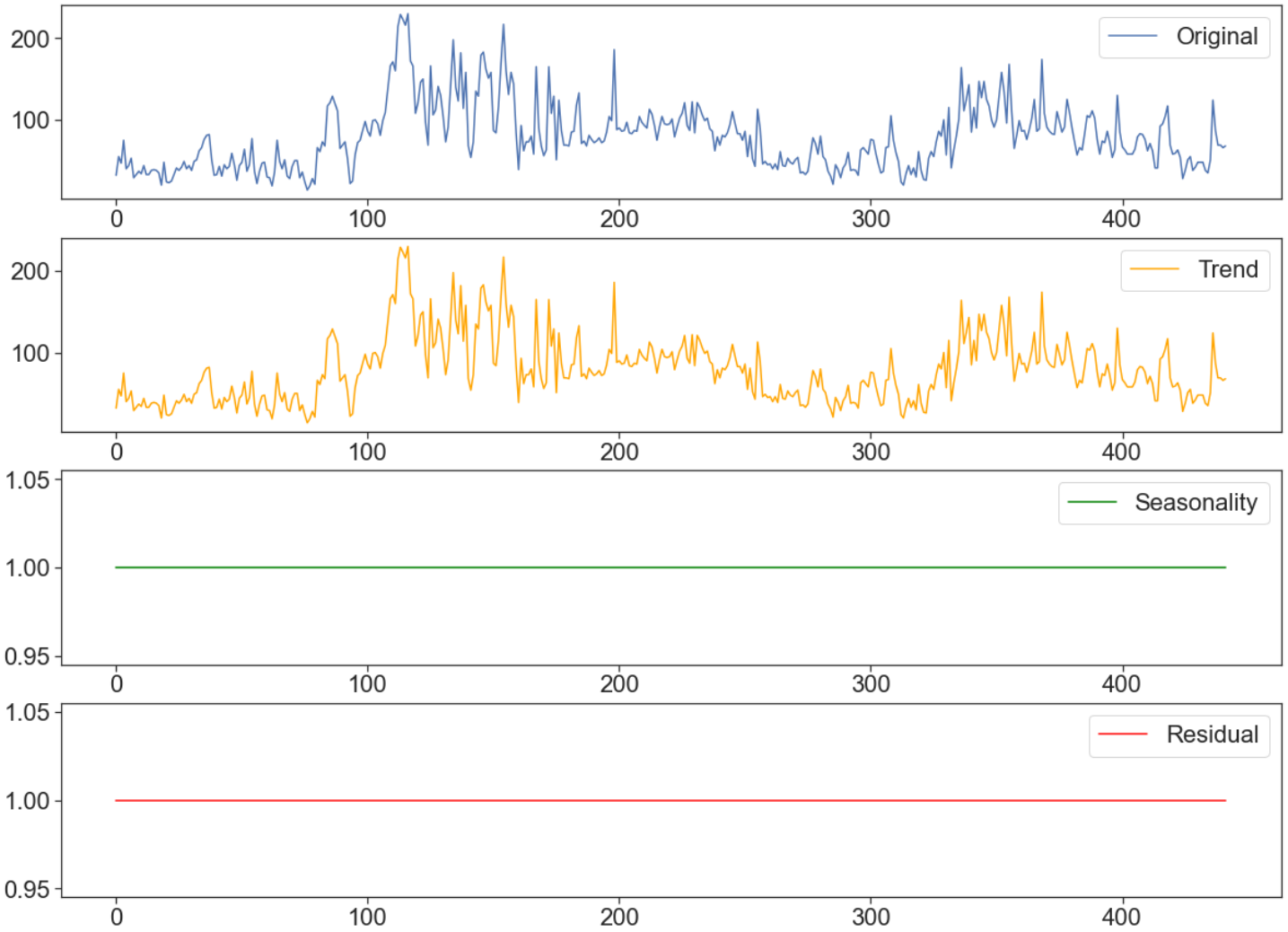


Can't say whether Stationary or not, So using Rolling Method to check Stationarity of Time Series Model.



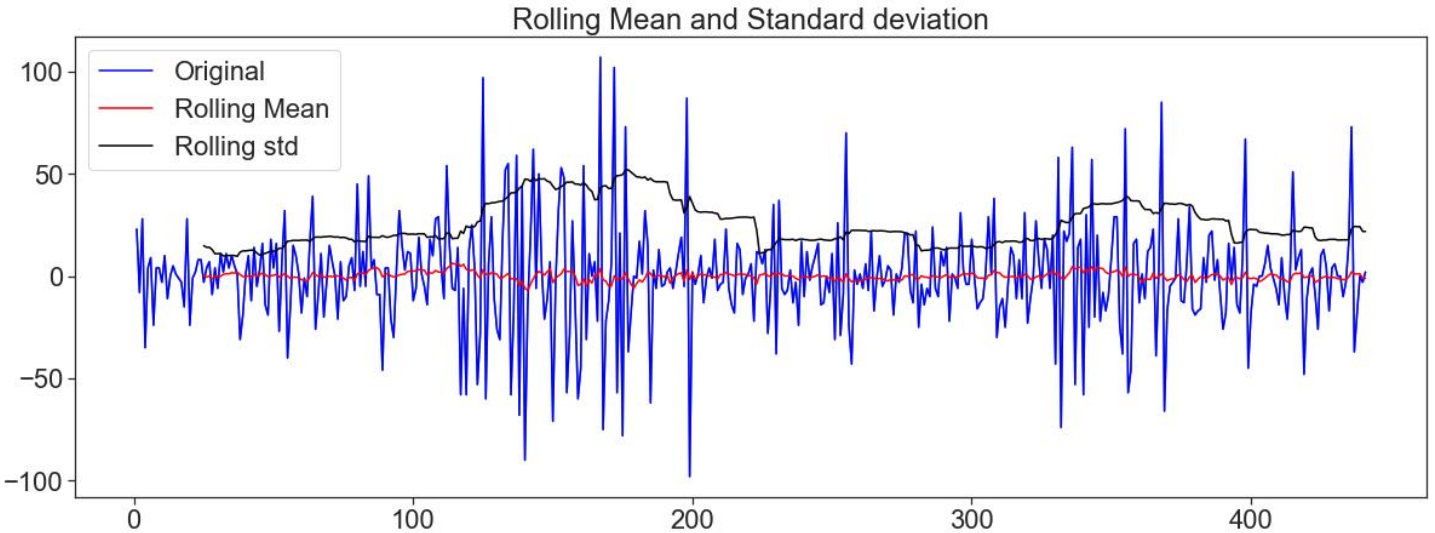
From above plot we can say that this Time Series does not have constant mean and variance hence Non-Stationary time series.

Checking for possible components of time series



Hence AQI is a Non Stationary Time series with trend component.
ARIMA models are applied in some cases where data show evidence of non-stationarity.

Trying Differencing to make Time Series Stationary (diff=1)



Rolling mean and rolling variance look like constant if we see small part at a time, So can be said as stationary time series.

Checking Stationarity of Time Series model by Argumented Dickey Fuller Test

Null hypothesis: The time series data is non stationary

Alternative hypothesis: That is stationary

Level of significance = 0.05, meaning 95% confidence. The test are interrupted with p-value if p>0.05 reject null hypothesis

Observation-

1. **ADF : -12.275296412155138**
2. **P-Value : 8.459469502689792e-23**
3. **critical values :**
   - **1% : -3.445542818501549**
   - **5% : -2.868238133603207**
   - **10% : -2.5703378690483176**

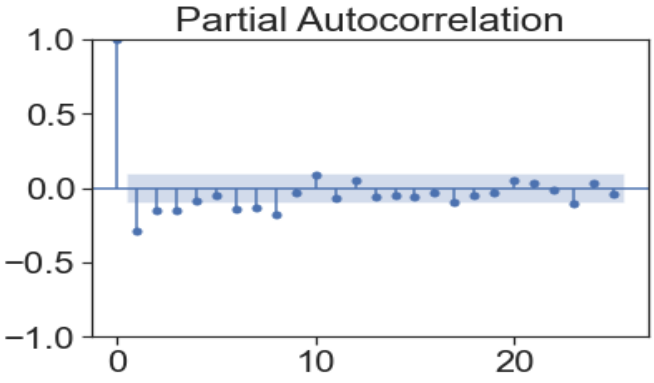
P_value greater than critical value so null hypothesis is rejected hence Time series is now Stationary.

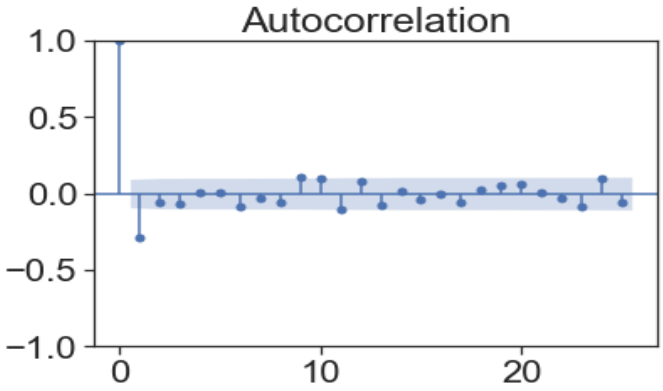
2.Figuring out order for ARIMA model

From the Autocorrelation plot we can say the MA(1) might fit, hence p might equal to 1.

From Partial Autocorrelation plot we can say AR(1) might fit, hence q might equal to 1.

So we can say that our Time series model might follow ARIMA(1,1,1).





Checking for best (p,d,q) for our model

Using a module named “pmdarima” we get performance of several ARIMA model in terms of their AIC and BIC.

Performing stepwise search to minimize AIC

ARIMA(1,1,2)(0,0,0)[0]	intercept	: AIC=4078.368, Time=0.27 sec
ARIMA(0,1,0)(0,0,0)[0]	intercept	: AIC=4152.351, Time=0.02 sec
ARIMA(1,1,0)(0,0,0)[0]	intercept	: AIC=4117.160, Time=0.04 sec
ARIMA(0,1,1)(0,0,0)[0]	intercept	: AIC=4101.023, Time=0.15 sec
ARIMA(0,1,0)(0,0,0)[0]		: AIC=4150.355, Time=0.02 sec
ARIMA(0,1,2)(0,0,0)[0]	intercept	: AIC=4090.372, Time=0.11 sec
ARIMA(1,1,1)(0,0,0)[0]	intercept	: AIC=4078.168, Time=0.12 sec
ARIMA(2,1,1)(0,0,0)[0]	intercept	: AIC=4078.263, Time=0.25 sec
ARIMA(2,1,0)(0,0,0)[0]	intercept	: AIC=4109.025, Time=0.16 sec
ARIMA(2,1,2)(0,0,0)[0]	intercept	: AIC=4080.193, Time=0.36 sec
ARIMA(1,1,1)(0,0,0)[0]		: AIC=4076.202, Time=0.10 sec
ARIMA(0,1,1)(0,0,0)[0]		: AIC=4099.031, Time=0.04 sec
ARIMA(1,1,0)(0,0,0)[0]		: AIC=4115.165, Time=0.03 sec
ARIMA(2,1,1)(0,0,0)[0]		: AIC=4076.302, Time=0.27 sec
ARIMA(1,1,2)(0,0,0)[0]		: AIC=4076.408, Time=0.30 sec
ARIMA(0,1,2)(0,0,0)[0]		: AIC=4088.388, Time=0.16 sec
ARIMA(2,1,0)(0,0,0)[0]		: AIC=4107.031, Time=0.16 sec
ARIMA(2,1,2)(0,0,0)[0]		: AIC=4078.231, Time=0.45 sec

Best model: ARIMA(1,1,1)(0,0,0)[0]
Total fit time: 3.050 seconds
ARIMA(order=(1, 1, 1), scoring_args={}, suppress_warnings=True,
with_intercept=False)
So our Time Series model will follow ARIMA(1,1,1).

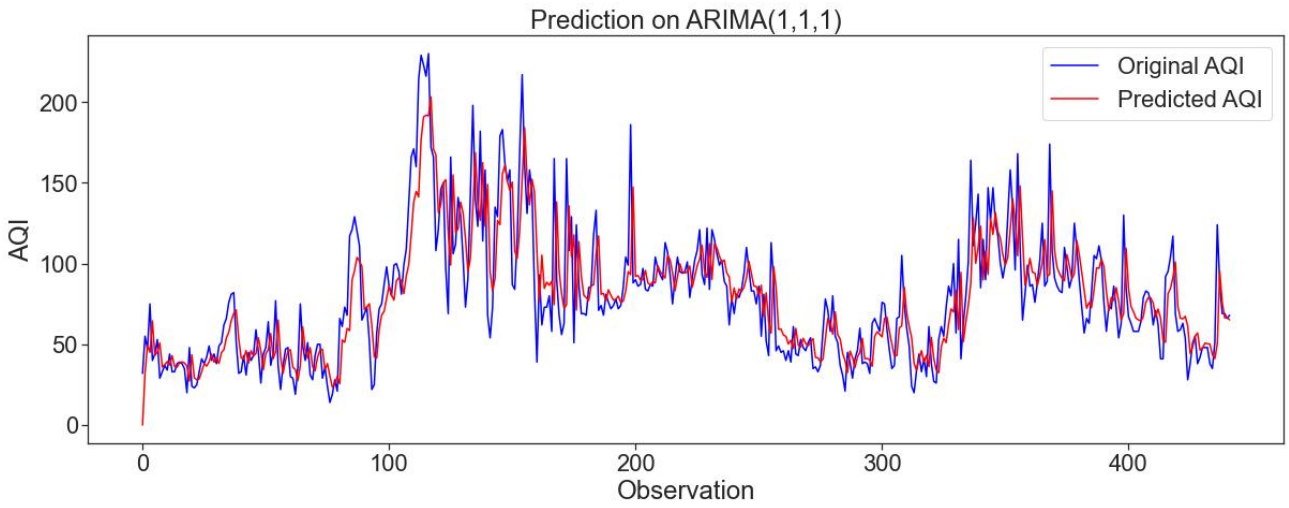
SARIMAX Results

Dep. Variable:	**AQI**	**No. Observations:**	**442**
Model:	**ARIMA(1, 1, 1)**	**Log Likelihood**	**-2035.101**
Date:	**Thu, 13 Apr 2023**	**AIC**	**4076.202**
Time:	**00:22:54**	**BIC**	**4088.469**
Sample:	**0**	**HQIC**	**4081.041**
	- 442		
Covariance Type:	**opg**		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	**0.4891**	**0.048**	**10.245**	**0.000**	**0.396**	**0.583**
ma.L1	**-0.8738**	**0.027**	**-32.636**	**0.000**	**-0.926**	**-0.821**
sigma2	**596.0025**	**28.706**	**20.762**	**0.000**	**539.739**	**652.266**

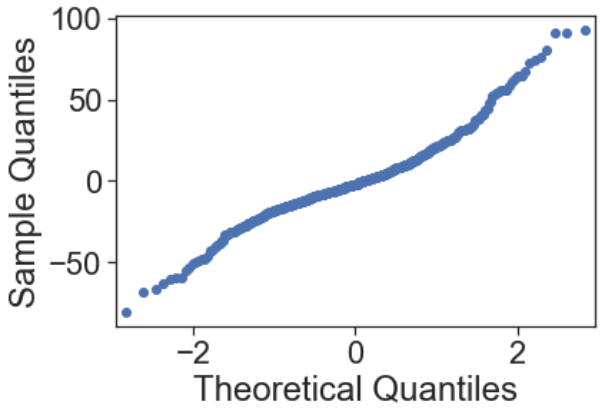
Ljung-Box (L1) (Q):	**0.37**	**Jarque-Bera (JB):**	**102.28**
Prob(Q):	**0.54**	**Prob(JB):**	**0.00**
Heteroskedasticity (H):	**0.85**	**Skew:**	**0.58**
Prob(H) (two-sided):	**0.33**	**Kurtosis:**	**5.06**

3.Visualization of Original Vs Predicted AQI



Correlation between AQI and Predicted AQI = 0.79751633, hence we can say model is not much of use.

4. Checking for normality of residuals

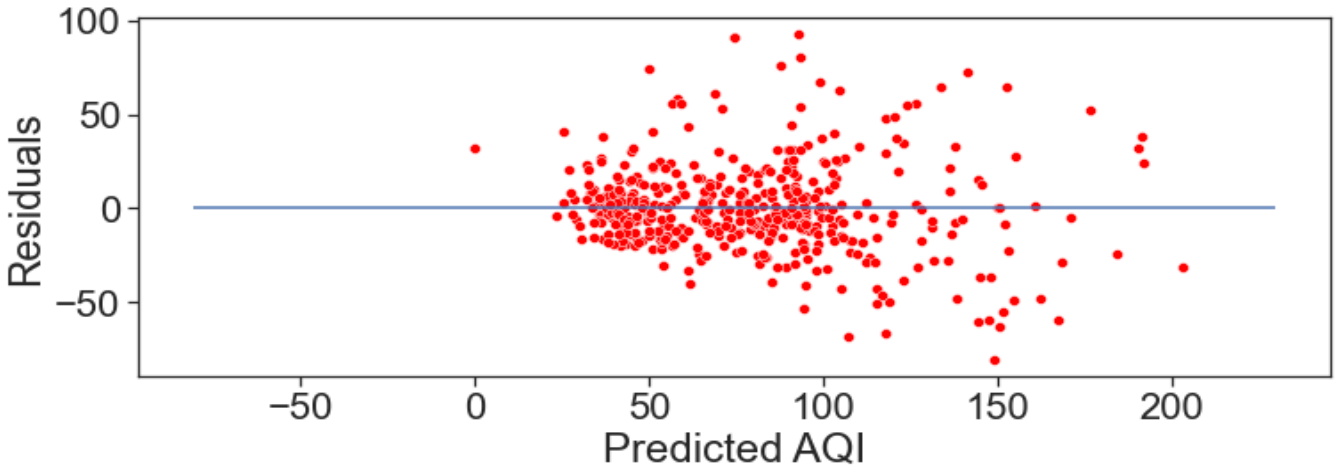


From QQ plot we can say that residuals follow approximately normal distribution but its hard to tell whether residuals follow normal distribution or not, so now by **Shapiro wilk test-**

ShapiroResult(statistic=0.9775571823120117, p-value=2.4721909994696034e-06)

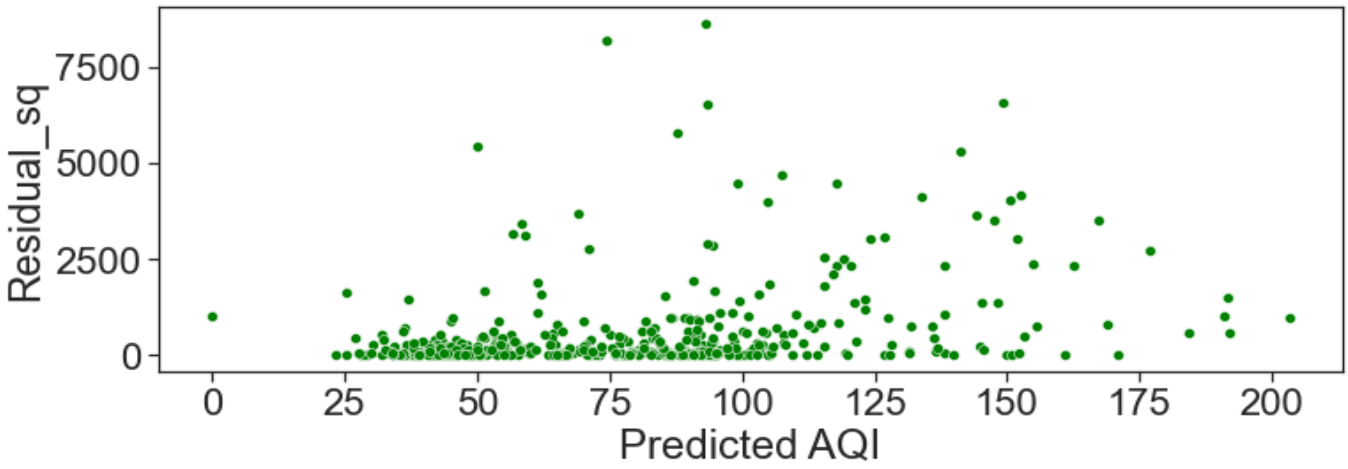
Since p-value is less than 0.05, we reject null hypothesis of Shapiro wilk test. This mean we have sufficient evidence to say that residual term does not follow normal distribution.

4.1 Checking for Autocorrelation



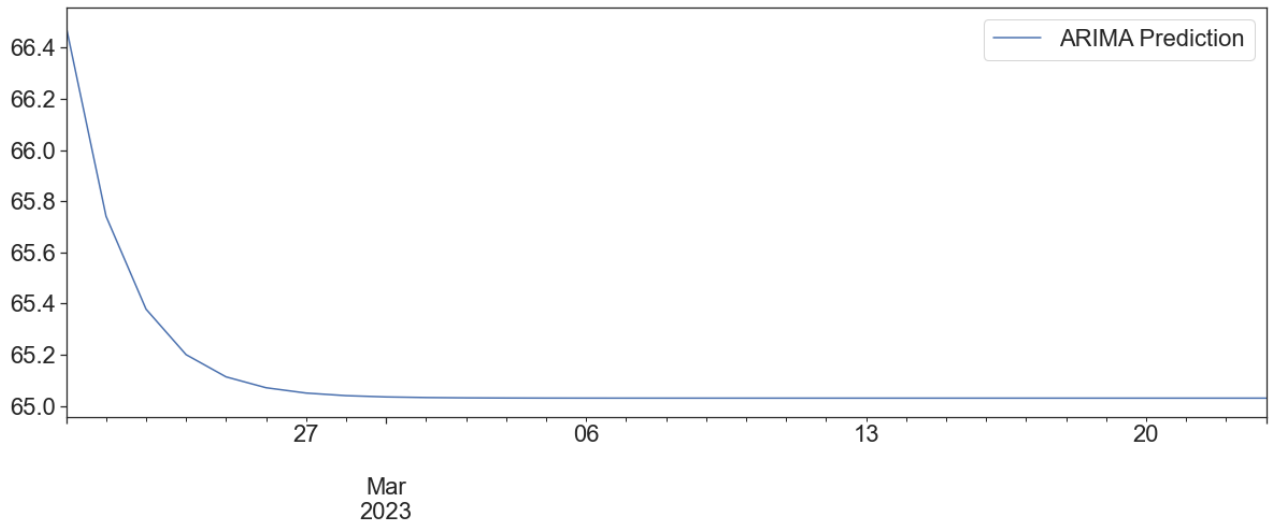
From Residual vs estimated AQI plot we can see a slightly 'C' shape hence might be presence of autocorrelation.

4.2 Checking for homoscedasticity



Hence presence of heteroscedasticity.

5. Visualization of Prediction for next 30 days-



Conclusion

1. There is difference in mean AQI of Banaras Hindu University in different phases of lockdown.

2.The formed Multiple Linear Regression Model is given by-

AQI= -21.937488445489095 + 1.09734313*PM_comp -0.16656863*NO2 + 0.62840198*NH3 + 0.35216351*SO2 + 0.44972472*CO+ 0.37906451*OZONE + 0.03069609*PRECIP_COVER + 0.45359676*WIND_SPEED + 0.08356036*CLOUD_COVER + 0.29800233*VISIBILITY-0.7937893*CONDITION_status + 0.49830492*WIND_DIRECTION_status

3.The Multiple linear regression model have R square value of 0.946 and Adjusted R square value as 0.944.

4.The Residuals of Multiple Linear Regression Model does not follow the assumption of normality.

5.The Residuals of Multiple Linear Regression Model are correlated, hence Autocorrelation Present.

6.The Residuals of Multiple Linear Regression Model does not have constant Variance throughout, hence Heteroscedasticity most likely present.

7.The Time Series(AQI of BHU) Model is not Stationary.

8.The Time series Model is Stationary after applying differencing of order 1

9.The Time Series is following ARIMA(1,1,1) model

10.The ARIMA(1,1,1) model is not much of a use as its Akaike Information Criterion(AIC) is High.

11.The ARIMA model have problems of Autocorrelation, Heteroscedasticity and Non-Normality of Residuals.

References

- https://en.wikipedia.org/wiki/Air_pollution
- https://education.nationalgeographic.org/resource/pollution/
- https://scied.ucar.edu/learning-zone/air-quality/how-weather-affects-air-quality#:~:text=Heat%20waves%20often%20lead%20to,forest%20fires%20are%20more%20common
- https://airly.org/en/how-does-humidity-affect-air-quality-all-you-need-to-know/#:~:text=Increase%20in%20airborne%20pollutants%20%E2%80%93%20high,translates%20into%20mold%20and%20mildew
- https://www.aqi.in/blog/effect-of-rain-on-air-pollution/#:~:text=Rain%20eases%20this%20problem%20by,phenomenon%20is%20called%20wet%20deposition
- https://www.statlect.com/fundamentals-of-statistics/multicollinearity
- https://en.wikipedia.org/wiki/Multicollinearity#Consequences
- https://www.google.com/maps/place/Institute+of+Environment+%26+Sustainable+Development/@25.2684424,82.9773559,5670m/data=!3m1!1e3!4m6!3m5!1s0x398e3180507fb08b:0x5a524cf653c5083f!8m2!3d25.2622975!4d82.995184!16s%2Fg%2F11dxbdl7zx?hl=en
- https://www.investopedia.com/terms/v/variance-inflation-factor.asp#:~:text=A%20variance%20inflation%20factor%20(VIF)%20is%20a%20measure%20of%20the,adversely%20affect%20the%20regression%20results
- https://analyse-it.com/docs/user-guide/fit-model/linear/residual-normality#:~:text=Normality%20is%20the%20assumption%20that,Shapiro%2DWilk%20or%20similar%20test
- https://www.itl.nist.gov/div898/handbook/pmd/section4/pmd453.htm#:~:text=If%20the%20data%20appear%20to,the%20most%20normally%20distributed%20residuals
- https://www.codingprof.com/5-ways-to-check-the-normality-of-residuals-in-r-examples/
- https://itfeature.com/time-series-analysis-and-forecasting/autocorrelation/consequences-of-autocorrelation
- https://prepnuggets.com/glossary/breusch-godfrey-test/
- https://itfeature.com/heteroscedasticity/heteroscedasticity-tests-and-remedies
- https://www.lkouniv.ac.in/site/writereaddata/siteContent/202003271457478511akash_Heteroscedasticity.pdf
- https://www.tutorialspoint.com/time-series-analysis-definition-and-components
- https://en.wikipedia.org/wiki/Decomposition_of_time_series
- https://www.tibco.com/reference-center/hat-is-time-series-analysis
- https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/time-series/how-to/partial-autocorrelation/interpret-the-results/partial-autocorrelation-function-pacf/#:~:text=The%20partial%20autocorrelation%20function%20is,t%E2%80%93k%E2%80%931)
- https://analyticsindiamag.com/how-to-make-a-time-series-stationary/

Code

1. Importing libraries

```
import pandas as pd
import numpy as np
import seaborn as sns
from scipy import stats
import matplotlib.pyplot as plt
import scipy as sp
import os
import warnings
from sklearn.model_selection import train_test_split
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.linear_model import LinearRegression
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from statsmodels.graphics import tsaplots
from statsmodels.stats.stattools import durbin_watson
import statsmodels.api as sm
from statsmodels.regression.linear_model import OLS
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_pacf
from statsmodels.graphics.tsaplots import plot_acf
from scipy.stats import shapiro
from statsmodels.tsa.seasonal import seasonal_decompose
import statsmodels.tools.tools as smt
import statsmodels.stats.diagnostic as smd
from sklearn.metrics import mean_squared_error
from math import sqrt
```

2. Performing Exploratory Data Analysis¶

3.1 Importing Dataset

```
os.chdir("E:\\Project")
#Renaming the AQI dataset as AQI
AQI=pd.read_excel("AQI.xlsx")
```

3.2 Exploring Dataset

```
AQI.describe()
#Searching for null values in columns
AQI.isnull().sum()
AQI.columns
```

3.3 Data Cleaning

//Dropping few variables which seems to be of no use or by which any other variable is generated. Also dropping rows associated by null values//

```
AQI=AQI.drop(['PRCIP_TYPE',"WIND_GUST","SEVER_RISK",'SUNRISE',"SUNSET","TEMPMAX", "TEMPMIN",
       'FEELS_LIKE_MAX', 'FEELS_LIKE_MIN',"FEELS_LIKE",'PRECI_PROB','SNOW', 'SNOW_DEPTH',
       'WEATHER_DESCRIPTION', 'ICON', 'stations'],axis=1)
AQI=AQI.dropna()
AQI.info()
#Saving the datadet wint no null values in new excel workbook as BHU_AQI
```

```
AQI.to_excel(r'E:\\Project\BHU_AQI.xlsx',index=True)
BHU_AQI=pd.read_excel("BHU_AQI.xlsx")
BHU_AQI.columns,BHU_AQI.shape
BHU_AQI.WIND_DIRECTION.head()
#As we can see the variable WIND_DIRECTION take continues values [0-360] so converting them in classes, and
assigning them numeric values in new column named as WIND_DIRECTION_status.

NE= BHU_AQI[(BHU_AQI.WIND_DIRECTION>22.5)&(BHU_AQI.WIND_DIRECTION<=67.5)].index
E = BHU_AQI[(BHU_AQI.WIND_DIRECTION>67.5) & (BHU_AQI.WIND_DIRECTION<=112.5)].index
SE= BHU_AQI[(BHU_AQI.WIND_DIRECTION>112.5) & (BHU_AQI.WIND_DIRECTION<=157.5)].index
S = BHU_AQI[(BHU_AQI.WIND_DIRECTION>157.5) & (BHU_AQI.WIND_DIRECTION<=202.5)].index
SW= BHU_AQI[(BHU_AQI.WIND_DIRECTION>202.5)&(BHU_AQI.WIND_DIRECTION<=247.5)].index
W = BHU_AQI[(BHU_AQI.WIND_DIRECTION>247.5) & (BHU_AQI.WIND_DIRECTION<=292.5)].index
NW= BHU_AQI[(BHU_AQI.WIND_DIRECTION>292.5) & (BHU_AQI.WIND_DIRECTION<=337.5)].index
N = BHU_AQI[((BHU_AQI.WIND_DIRECTION>337.5) & (BHU_AQI.WIND_DIRECTION<=360))|
      ((BHU_AQI.WIND_DIRECTION>=0) & (BHU_AQI.WIND_DIRECTION<=22.5))].index
BHU_AQI.loc[NE,'WIND_DIRECTION_Status'] = 'NE'
BHU_AQI.loc[E,'WIND_DIRECTION_Status'] = 'E'
BHU_AQI.loc[SE,'WIND_DIRECTION_Status'] = 'SE'
BHU_AQI.loc[S,'WIND_DIRECTION_Status'] = 'S'
BHU_AQI.loc[SW,'WIND_DIRECTION_Status'] = 'SW'
BHU_AQI.loc[W,'WIND_DIRECTION_Status'] = 'W'
BHU_AQI.loc[NW,'WIND_DIRECTION_Status'] = 'NW'
BHU_AQI.loc[N,'WIND_DIRECTION_Status'] = 'N'
BHU_AQI.WIND_DIRECTION_Status.head()

fig, ax=plt.subplots(figsize=(10,5))
sns.countplot(x='WIND_DIRECTION_Status', data=BHU_AQI)
plt.show()

BHU_AQI.loc[NE,'WIND_DIRECTION_Status'] = '1'
BHU_AQI.loc[E,'WIND_DIRECTION_Status'] = '2'
BHU_AQI.loc[SE,'WIND_DIRECTION_Status'] = '3'
BHU_AQI.loc[S,'WIND_DIRECTION_Status'] = '4'
BHU_AQI.loc[SW,'WIND_DIRECTION_Status'] = '5'
BHU_AQI.loc[W,'WIND_DIRECTION_Status'] = '6'
BHU_AQI.loc[NW,'WIND_DIRECTION_Status'] = '7'
BHU_AQI.loc[N,'WIND_DIRECTION_Status'] = '8'
BHU_AQI.WIND_DIRECTION_Status.head()
#The datatype of column WIND_DIRECTION_Status is object so converting it to integer type.

BHU_AQI['WIND_DIRECTION_Status']=BHU_AQI['WIND_DIRECTION_Status'].astype(int)
#The Condition variable also take values as object datatype, so converting them into numeric values.

BHU_AQI.CONDITION.unique()

fig, ax=plt.subplots(figsize=(10,5))
sns.countplot(x='CONDITION', data=BHU_AQI)
```

```
plt.show()
fig, ax=plt.subplots(figsize=(10,5))
sns.countplot(x='CONDITION', data=BHU_AQI)
plt.show()

BHU_AQI.insert(25,"CONDITION_status",BHU_AQI.CONDITION.replace(['Rain, Partially cloudy',
                                        'Rain, Overcast', 'Partially cloudy',
                                        'Overcast', 'Clear'],[1,2,3,4,5]))
BHU_AQI.CONDITION_status.head()

                //Now dropping the columns named CONDITION and WIND_DIRECTION//
BHU_AQI=BHU_AQI.drop(['CONDITION','WIND_DIRECTION'],axis=1)
BHU_AQI.shape
BHU_AQI=BHU_AQI.drop(['CONDITION','WIND_DIRECTION'],axis=1)
BHU_AQI.shape

#Checking for outlier and dropping them

fig, axes = plt.subplots(16,2 , figsize=(15, 50))
sns.histplot(data=BHU_AQI,x="AQI",bins=442,kde=True,color="green",ax=axes[0,0])
sns.boxplot(data=BHU_AQI,x="AQI",color="green",ax=axes[0,1])
sns.histplot(data=BHU_AQI,x="PM2.5",bins=442,kde=True,color="blue",ax=axes[1,0])
sns.boxplot(data=BHU_AQI,x="PM2.5",color="blue",ax=axes[1,1])
sns.histplot(data=BHU_AQI,x="PM10",bins=442,kde=True,color="red",ax=axes[2,0])
sns.boxplot(data=BHU_AQI,x="PM10",color="red",ax=axes[2,1])
sns.histplot(data=BHU_AQI,x="NO2",bins=442,kde=True,color="teal",ax=axes[3,0])
sns.boxplot(data=BHU_AQI,x="NO2",color="teal",ax=axes[3,1])
sns.histplot(data=BHU_AQI,x="NH3",bins=442,kde=True,color="gray",ax=axes[4,0])
sns.boxplot(data=BHU_AQI,x="NH3",color="gray",ax=axes[4,1])
sns.histplot(data=BHU_AQI,x="SO2",bins=442,kde=True,color="yellow",ax=axes[5,0])
sns.boxplot(data=BHU_AQI,x="SO2",color="yellow",ax=axes[5,1])
sns.histplot(data=BHU_AQI,x="CO",bins=442,kde=True,color="pink",ax=axes[6,0])
sns.boxplot(data=BHU_AQI,x="CO",color="pink",ax=axes[6,1])
sns.histplot(data=BHU_AQI,x="OZONE",bins=442,kde=True,color="indigo",ax=axes[7,0])
sns.boxplot(data=BHU_AQI,x="OZONE",color="indigo",ax=axes[7,1])
sns.histplot(data=BHU_AQI,x="TEMPERATURE",bins=442,kde=True,color="orange",ax=axes[8,0])
sns.boxplot(data=BHU_AQI,x="TEMPERATURE",color="orange",ax=axes[8,1])
sns.histplot(data=BHU_AQI,x="WIND_SPEED",bins=442,kde=True,color="green",ax=axes[9,0])
sns.boxplot(data=BHU_AQI,x="WIND_SPEED",color="green",ax=axes[9,1])
sns.histplot(data=BHU_AQI,x="SOLAR_RADIATION",bins=442,kde=True,color="blue",ax=axes[10,0])
sns.boxplot(data=BHU_AQI,x="SOLAR_RADIATION",color="blue",ax=axes[10,1])
sns.histplot(data=BHU_AQI,x="UV_INDEX",bins=442,kde=True,color="red",ax=axes[11,0])
sns.boxplot(data=BHU_AQI,x="UV_INDEX",color="red",ax=axes[11,1])
sns.histplot(data=BHU_AQI,x="VISIBILITY",bins=442,kde=True,color="teal",ax=axes[12,0])
sns.boxplot(data=BHU_AQI,x="VISIBILITY",color="teal",ax=axes[12,1])
sns.histplot(data=BHU_AQI,x="HUMIDITY",bins=442,kde=True,color="gray",ax=axes[13,0])
sns.boxplot(data=BHU_AQI,x="HUMIDITY",color="gray",ax=axes[13,1])
sns.histplot(data=BHU_AQI,x="SEA_LEVEL_PRESSURE",bins=442,kde=True,color="yellow",ax=axes[14,0])
```

```
sns.boxplot(data=BHU_AQI,x="SEA_LEVEL_PRESSURE",color="yellow",ax=axes[14,1])
sns.histplot(data=BHU_AQI,x="SOLAR_ENERGY",bins=442,kde=True,color="pink",ax=axes[15,0])
sns.boxplot(data=BHU_AQI,x="SOLAR_ENERGY",color="pink",ax=axes[15,1])
plt.show()

#Checking for outliers in possible variables using Z-score
z=np.abs(stats.zscore(BHU_AQI[['PM2.5', 'PM10', 'NO2', 'NH3', 'SO2', 'CO',
     'OZONE', 'AQI', 'TEMPERATURE', 'HUMIDITY', 'WIND_SPEED', 'SOLAR_RADIATION', 'UV_INDEX',"VISIBILITY"]]))
outliers=np.where(z>3)
print(np.unique(outliers))
#Dropping the outliers
BHU_AQI_new=BHU_AQI.drop([ 0  , 1  , 2  , 3  , 5 ,  6 ,  7  , 9 , 10 , 11 , 12 , 23 ,
                   42  ,61  ,76  ,77 ,126 ,128,129 ,130 ,131 ,132, 135, 139,
                   146 ,147, 148, 149, 151, 152 ,153 ,154, 155 ,158 ,174 ,217
                   ,245 ,246, 247, 248, 249, 250, 251, 252, 253, 254, 256, 257,
                   261, 267, 268, 269, 274, 277,286, 292,310,341, 346, 391, 459,
                   481,497,498],axis=0)
BHU_AQI_new.shape

                              //Checking for any duplicate rows//
#Checking for duplicate rows
duplicates = BHU_AQI_new[BHU_AQI_new.duplicated()]
print("Duplicate Rows : ",len(duplicates))
duplicates.head()

#Saving the datadet wint no null values in new excel workbook as BHU_AQI
BHU_AQI_new.to_excel(r'E:\\Project\BHU_AQI_clean.xlsx',index=False)
BHU_AQI_clean=pd.read_excel("BHU_AQI_clean.xlsx")
BHU_AQI_clean.columns,BHU_AQI_clean.shape
```

3. Visualization of Data

#Using excel created 3 new columns ("Lockdown_BHU_Closed","Unlock_BHU_Closed","BHU_Open") having date as data, reflecting values of all variables for different sessions of BHU.

```
BHU_AQI_Session=pd.read_excel("sess.xlsx")

sns.histplot(data=BHU_AQI_Session,x="BHU_open_AQI",bins=261,kde=True,color="green")
<AxesSubplot:xlabel='BHU_open_AQI', ylabel='Count'>

sns.histplot(data=BHU_AQI_Session,x="Lockdown_BHU_Closed_AQI",bins=97,kde=True,color="blue")
<AxesSubplot:xlabel='Lockdown_BHU_Closed_AQI', ylabel='Count'>

sns.histplot(data=BHU_AQI_Session,x="Unlock_BHU_Closed_AQI",bins=84,kde=True,color="orange")
<AxesSubplot:xlabel='Unlock_BHU_Closed_AQI', ylabel='Count'>

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.AQI)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.AQI)
```

```
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.AQI)
plt.ylabel("Air Quality Index")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)
<matplotlib.legend.Legend at 0x1a7457eae80>
```

From above plot we can see that the first unlock phase have significant impact on Air Pollution.

```
fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.PM25)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.PM25 )
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.PM25)
plt.ylabel("Particulate Matter 2.5 micrometer (µg/m3)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.PM10)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.PM10)
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.PM10)
plt.ylabel("Particulate Matter 10 micrometer (µg/m3)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.NO2 )
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.NO2 )
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.NO2 )
plt.ylabel("Nitrogen dioxide (ppb)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.NH3 )
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.NH3 )
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.NH3 )
plt.ylabel("Ammonia (ppm)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.SO2 )
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.SO2 )
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.SO2 )
plt.ylabel("Sulpher dioxide (ppb)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)
```

```
fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.CO )
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.CO )
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.CO )
plt.ylabel("Carbon monooxide(ppb)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.OZONE)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.OZONE )
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.OZONE )
plt.ylabel("Ozone (Dobson Unit)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.TEMPERATURE)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.TEMPERATURE)
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.TEMPERATURE)
plt.ylabel("Temperature(Fahrenheit)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.DEW)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.DEW)
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.DEW)
plt.ylabel("Dew(%)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.HUMIDITY)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.HUMIDITY)
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.HUMIDITY)
plt.ylabel("Humidity(%)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.PRECIP_COVER )
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.PRECIP_COVER )
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.PRECIP_COVER )
plt.ylabel("Precipitate Cover(mm)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)
```

```
fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.WIND_SPEED )
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.WIND_SPEED )
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.WIND_SPEED )
plt.ylabel("Wind Speed(Kmph)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.SEA_LEVEL_PRESSURE)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.SEA_LEVEL_PRESSURE)
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.SEA_LEVEL_PRESSURE)
plt.ylabel("Sea Level Pressure(hPa)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.CLOUD_COVER )
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.CLOUD_COVER)
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.CLOUD_COVER)
plt.ylabel("Cloud Cover(%)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.VISIBILITY)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.VISIBILITY )
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.VISIBILITY)
plt.ylabel("Visibility (Km)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.SOLAR_RADIATION)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.SOLAR_RADIATION)
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.SOLAR_RADIATION)
plt.ylabel("Solar Radiation(Wm−2)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.SOLAR_ENERGY)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.SOLAR_ENERGY)
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.SOLAR_ENERGY)
plt.ylabel("Solar Energy/(Wm−2)")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)
```

```
fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.UV_INDEX)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.UV_INDEX)
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.UV_INDEX)
plt.ylabel("Ultra violet Index")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.CONDITION_status)
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.CONDITION_status )
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.CONDITION_status)
plt.ylabel("Condition Status")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)

fig, ax = plt.subplots(figsize=(15, 4))
sns.scatterplot(x=BHU_AQI_Session.Lockdown_BHU_Closed,y=BHU_AQI_Session.WIND_DIRECTION_Status )
sns.scatterplot(x=BHU_AQI_Session.Unlock_BHU_Closed,y=BHU_AQI_Session.WIND_DIRECTION_Status)
sns.scatterplot(x=BHU_AQI_Session.BHU_Open,y=BHU_AQI_Session.WIND_DIRECTION_Status)
plt.ylabel("Wind Direction Status")
plt.xlabel("Date")
plt.legend(["Lockdown BHU Closed","Unlock BHU Closed","BHU Open"],loc=0)
```

4. Deducing Predictor and Target Variables

5.1 Checking for Correlated variables and droping them.

```
#Defining the Predictor variable
X=BHU_AQI_clean.drop(["DATE","AQI","Unnamed: 0"],axis=1)
#Defining the Target variable
Y=BHU_AQI_clean.AQI

plt.figure(figsize=(20,10))
sns.heatmap(X.corr(), cmap="YlGnBu",annot=True,linewidths=1,linecolor="red")

#Checking for significance importance of both PM values with AQI
np.corrcoef(BHU_AQI_clean['AQI'],BHU_AQI_clean['PM2.5']),np.corrcoef(BHU_AQI_clean['AQI'],BHU_AQI_clean['PM10'
])

BHU_AQI_clean.insert(0,"PM_comp",BHU_AQI_clean["PM2.5"]*0.5+BHU_AQI_clean["PM10"]*0.5)
BHU_AQI_clean.head(2)

#Defining the Predictor variable
X=BHU_AQI_clean.drop(["DATE","AQI","Unnamed: 0","PM2.5","PM10"],axis=1)
#Defining the Target variable
Y=BHU_AQI_clean.AQI
X.head()
plt.figure(figsize=(20,10))
sns.heatmap(X.corr(), cmap="YlGnBu",annot=True,linewidths=1,linecolor="red")
```

```
plt.show()

5.2 Checking for multicollinearity and dropping the correlated variables using Variance inflation factor
Int=X.astype(int)
VIF= pd.DataFrame()
VIF['feature'] = X.columns
VIF['VIF']= [variance_inflation_factor(Int.values, i) for i in range(X.shape[1])]
VIF
#Dropping the variable with maximum VIF that is variable ("DEW")
X1=X.drop(["DEW"],axis=1)
Int=X1.astype(int)
VIF= pd.DataFrame()
VIF['feature'] = X1.columns
VIF['VIF']= [variance_inflation_factor(Int.values, i) for i in range(X1.shape[1])]
VIF
X2=X.drop(["DEW","SOLAR_RADIATION"],axis=1)
Int=X2.astype(int)
VIF= pd.DataFrame()
VIF['feature'] = X2.columns
VIF['VIF']= [variance_inflation_factor(Int.values, i) for i in range(X2.shape[1])]
VIF
X3=X.drop(["DEW","SOLAR_RADIATION","DAYLIGHT_DURATION"],axis=1)
Int=X3.astype(int)
VIF= pd.DataFrame()
VIF['feature'] = X3.columns
VIF['VIF']= [variance_inflation_factor(Int.values, i) for i in range(X3.shape[1])]
VIF
X4=X.drop(["DEW","SOLAR_RADIATION","DAYLIGHT_DURATION","SEA_LEVEL_PRESSURE"],axis=1)
Int=X4.astype(int)
VIF= pd.DataFrame()
VIF['feature'] = X4.columns
VIF['VIF']= [variance_inflation_factor(Int.values, i) for i in range(X4.shape[1])]
VIF
X5=X.drop(["DEW","SOLAR_RADIATION","DAYLIGHT_DURATION","SEA_LEVEL_PRESSURE",
           "UV_INDEX"],axis=1)
Int=X5.astype(int)
VIF= pd.DataFrame()
VIF['feature'] = X5.columns
VIF['VIF']= [variance_inflation_factor(Int.values, i) for i in range(X5.shape[1])]
VIF
X6=X.drop(["DEW","SOLAR_RADIATION","DAYLIGHT_DURATION","SEA_LEVEL_PRESSURE",
           "UV_INDEX","TEMPERATURE"],axis=1)
Int=X6.astype(int)
VIF= pd.DataFrame()
VIF['feature'] = X6.columns
VIF['VIF']= [variance_inflation_factor(Int.values, i) for i in range(X6.shape[1])]
VIF
X7=X.drop(["DEW","SOLAR_RADIATION","DAYLIGHT_DURATION","SEA_LEVEL_PRESSURE",
```

```
                "UV_INDEX","TEMPERATURE","HUMIDITY"],axis=1)
Int=X7.astype(int)
VIF= pd.DataFrame()
VIF['feature'] = X7.columns
VIF['VIF']= [variance_inflation_factor(Int.values, i) for i in range(X7.shape[1])]
VIF
X_final=X.drop(["DEW","SOLAR_RADIATION","DAYLIGHT_DURATION","SEA_LEVEL_PRESSURE",
                "UV_INDEX","TEMPERATURE","HUMIDITY","SOLAR_ENERGY"],axis=1)
Int=X_final.astype(int)
VIF= pd.DataFrame()
VIF['feature'] = X_final.columns
VIF['VIF']= [variance_inflation_factor(Int.values, i) for i in range(X_final.shape[1])]
VIF

X_FINAL=X_final.drop(["MOONPHASE","PRECIPITATE"],axis=1)
X_FINAL.columns
```

5. Model Building

```
x_train, x_test, y_train, y_test = train_test_split(X_FINAL,Y, test_size=0.15, random_state=0)
x_train_sm=sm.add_constant(x_train)
x_train_sm

#Method 1:Fitting a Linear regression model using ordinary least square method
lr=sm.OLS(y_train,x_train_sm).fit()
lr

#Checking wether our model is significant or not
1.49e-220<0.05

#Method 2: Using Multiple Linear Regression
x_train, x_test, y_train, y_test = train_test_split(X_FINAL,Y, test_size=0.15, random_state=0)
Lin_reg = LinearRegression()
Lin_reg.fit(x_train, y_train)
print(Lin_reg.intercept_)
print(Lin_reg.coef_)
train_accuracy=Lin_reg.score(x_train,y_train)
test_accuracy=Lin_reg.score(x_test,y_test)
print(Lin_reg.score(x_train, y_train))
print(Lin_reg.score(x_test, y_test)
AQI= -21.937488445489095 + 1.09734313PM_comp -0.16656863NO2 + 0.62840198NH3 + 0.35216351SO2 +
0.44972472CO + 0.37906451OZONE + 0.03069609PRECIP_COVER + 0.45359676WIND_SPEED +
0.08356036CLOUD_COVER + 0.29800233VISIBILITY -0.7937893CONDITION_status +
0.49830492WIND_DIRECTION_status

BHU_AQI_clean.columns
BHU_AQI_clean=BHU_AQI_clean.drop(['Unnamed: 0', 'PM10', 'PM2.5','TEMPERATURE', 'DEW', 'HUMIDITY',
'SEA_LEVEL_PRESSURE', 'SOLAR_RADIATION', 'SOLAR_ENERGY', 'UV_INDEX','DAYLIGHT_DURATION','MOONPHASE',
 'PRECIPITATE'],axis=1)
```

```
BHU_AQI_clean.columns

#Inserting a new column named as Predicted_AQI
BHU_AQI_clean.insert(7,"Predicted_AQI",-21.937488445489095 + 1.09734313*(BHU_AQI_clean['PM_comp']) -
0.16656863*(BHU_AQI_clean['NO2']) + 0.62840198*(BHU_AQI_clean['NH3']) + 0.35216351*(BHU_AQI_clean['SO2']) +
0.44972472*(BHU_AQI_clean['CO']) + 0.37906451*(BHU_AQI_clean['OZONE'])
+0.03069609*(BHU_AQI_clean['PRECIP_COVER']) + 0.45359676*(BHU_AQI_clean['WIND_SPEED']) +
0.08356036*(BHU_AQI_clean['CLOUD_COVER']) + 0.29800233*(BHU_AQI_clean['VISIBILITY']) -
0.7937893*(BHU_AQI_clean['CONDITION_status']) + 0.49830492*(BHU_AQI_clean['WIND_DIRECTION_Status']))
BHU_AQI_clean.columns

lin=BHU_AQI_clean.drop(["Predicted_AQI","DATE"],axis=1)
lin.head()

#Using scatter plot for determining Linearity
sns.set(style='ticks',color_codes=True,font_scale=2)
pl0t=sns.pairplot(lin,height=3,palette='Accent',diag_kind='hist',kind='reg')
pl0t.fig.suptitle('scatter plot',y=1)

AQI_Predicted=Lin_reg.predict(x_test)

                        //Visualization of Predicted_AQI vs AQI//
x= np.arange(1, 68)
y1 = y_test
y2 = AQI_Predicted
plt.figure(figsize=(15,4))
plt.title("Line graph")
plt.ylabel("AQI")
plt.xlabel("Testing data")
plt.plot(x, y1, color ="blue")
plt.plot(x, y2, color ="red")
plt.legend(["AQI","Predicted_AQI"],loc=0)
plt.show()

#Correlation between AQI and Predicted AQI
np.corrcoef(AQI_Predicted,y_test)

BHU_AQI_clean=BHU_AQI_clean.drop([ 'PM_comp', 'NO2', 'NH3', 'SO2', 'CO', 'OZONE',
       'PRECIP_COVER', 'WIND_SPEED', 'CLOUD_COVER',
      'VISIBILITY', 'CONDITION_status', 'WIND_DIRECTION_Status'],axis=1)
BHU_AQI_clean.columns
#Inserting a column named as Residual
BHU_AQI_clean.insert(3,"Residual",(BHU_AQI_clean["AQI"])-(BHU_AQI_clean["Predicted_AQI"]))
BHU_AQI_clean.head(2)

#Inserting a column of squared residual named as Residual_sq
BHU_AQI_clean.insert(4,"Residual_sq",BHU_AQI_clean["Residual"]*BHU_AQI_clean["Residual"])
BHU_AQI_clean.head(2)
```

```
Predicted=BHU_AQI_clean.to_excel(r'E:\\Project\final.xlsx',index=False)
Predicted=pd.read_excel("final.xlsx")
Predicted.head(2)
df=pd.DataFrame([["Mean",Predicted.loc[:,"AQI"].mean(),Predicted.loc[:,"Predicted_AQI"].mean(),Predicted.loc[:,"Residual"].mean()],["StandardDeviation",Predicted.loc[:,"AQI"].std(),Predicted.loc[:,"Predicted_AQI"].std(),Predicted.loc[:,"Residual"].std()],],columns=["","AQI","Predicted_AQI","Residual"])
df

rmse=sqrt(mean_squared_error(BHU_AQI_clean['Predicted_AQI'],BHU_AQI_clean['AQI']))
print(rmse)
```

6. Verifying Assumption of Multiple Linear Regression

```
#Using scatter plot for determining Linearity
sns.set(style='ticks',color_codes=True,font_scale=2)
pl0t=sns.pairplot(X_final,height=3,palette='Accent',diag_kind='hist',kind='reg')
pl0t.fig.suptitle('scatter plot',y=1)

#Multicollinearity (The independent variables are not highly correlated with each other)
plt.figure(figsize=(30,15))
sns.heatmap(X_FINAL.corr(), cmap="Greens",annot=True,linewidths=1,linecolor="red")

#Normality of residuals
plt.figure(figsize=(7,4))
sns.histplot(data=Predicted,x="Residual",bins=442,kde=True,color="orange")
plt.xlabel("Residual")
plt.ylabel("Count")
plt.title("Distribution of Residual")
sns.displot(Predicted.Residual,kind="kde")
#Using QQ-plot
fig, ax = plt.subplots(figsize=(6,4))
sp.stats.probplot(Predicted.Residual,plot=ax,fit=True)
plt.show()
#Using QQ-Plot
fig=sm.qqplot(Predicted.Residual)
plt.show()
#Null hypothesis that Residual follow normal distribution
shapiro(Predicted.Residual)

#Independence of residuals (no autocorrelation)
                        //VISUALIZING RESIDUALS FOR CHECKING AUTOCORRELATION
fig, ax = plt.subplots(figsize=(12, 4))
sns.scatterplot(x=Predicted.Predicted_AQI,y=Predicted.Residual,color="red" )
plt.hlines(Predicted.loc[:,"Residual"].mean(),Predicted.loc[:,"Residual"].min(),230)

                         //Applying Dubin_Watson Test
# Checking Whether our AQI Time series data is Stationary or not
plt.figure(figsize=(20,7))
Predicted["AQI"].plot()
```

```
plt.ylabel("AQI")
plt.xlabel("Observation")
dftest=adfuller(Predicted['AQI'],autolag='AIC')
print("1. ADF :",dftest[0])
print("2. P-Value :",dftest[1])
print("3. critical values :")
for key, val in dftest[4].items():
   print("\t",key,":",val)

acf=plot_acf(Predicted["Residual"],lags=25)

                    //Applying Breusch_Godfrey Test for testing Autocorrelation
BHU_AQI_Clean=X
BHU_AQI_Clean_smt=smt.add_constant(BHU_AQI_Clean)
import statsmodels.regression.linear_model as rg
ivar=['const','PM_comp', 'NO2', 'NH3', 'SO2', 'CO', 'OZONE',
     'PRECIP_COVER', 'WIND_SPEED', 'CLOUD_COVER', 'VISIBILITY',
     'CONDITION_status', 'WIND_DIRECTION_Status']
reg=rg.OLS(BHU_AQI_clean['AQI'],BHU_AQI_Clean_smt[ivar],hasconst=bool).fit()
#Null hypothesis that no Autocorrelation
for i in [1,2,3,4,5]:
   print(i)
   print("Breusch_Godfrey LM Test Statistic:",np.round(smd.acorr_breusch_godfrey(reg, nlags=i)[0], 6))
   print("Breusch_Godfrey LM Test P_Value:",np.round(smd.acorr_breusch_godfrey(reg, nlags=i)[1], 6))

#Homoskedasticity (The variance of residuals is constant)
                    //Visualizing for presence of Homoscedasticty/Heteroscedasticity
fig, ax = plt.subplots(figsize=(12, 4))
sns.scatterplot(x=Predicted.Predicted_AQI,y=Predicted.Residual_sq,color="green" )
plt.hlines(Predicted.loc[:,"Residual_sq"].mean(),Predicted.loc[:,"Residual_sq"].min(),230)

#Sorting residuals in ascending order
sorted_AQI=Predicted.sort_values('AQI',ascending=True)
sorted_AQI
#N=442,C=22,K=15
#SR1=sorted residual 1
SR1=sorted_AQI["Residual_sq"].iloc[:210]
#RSS1 is residual sum of square for first sorted residuals
RSS1=SR1.sum()
#SR2=sorted residual 2
SR2=sorted_AQI["Residual_sq"].iloc[232:442]
#RSS2 is residual sum of square for second sorted residuals
RSS2=SR2.sum()
RSS1,RSS2

#F calculated
F_cal=RSS2/RSS1
F_cal
```

7. Prediction

```
PM_comp=float(input("Enter the measured value of (PM2.5 + PM10)*0.5 (micrograms per cubic meter of air)"))
NO2=float(input("Enter the measured value of NO2(parts per billion in air)"))
NH3=float(input("Enter the measured value of NH3(parts per million in air)"))
SO2=float(input("Enter the measured value of SO2(parts per billion in air)"))
CO=float(input("Enter the measured value of CO(parts per million in air)"))
OZONE=float(input("Enter the measured value of OZONE(Dobson unit)"))
PRECIP_COVER=float(input("Enter the measured value of PRECIPITATE COVER(mm in 24 hours)"))
WIND_SPEED=float(input("Enter the measured value of WIND SPEED(KM/H)"))
CLOUD_COVER=float(input("Enter the observed value of CLOUD COVER(Percentage)"))
VISIBILITY=float(input("Enter the measured value of VISIBILITY(KM)"))
CONDITION_status=float(input("Enter the observed value of CONDITION status(1 for Rain and partially cloudy, 2 for Rain and overcast,3 for Partially coludy,4 for Overcast,5 for Clear))"))
WIND_DIRECTION_status=float(input("Enter the observed value of WIND_DIRECTION_status(Direction(1 for NE, 2 for E,3 for SE,4 for S,5 for SW,6 for W,7 for NW,8 for N))"))

Predicted_AQI= -21.937488445489095 + 1.09734313*PM_comp -0.16656863*NO2 + 0.62840198*NH3 + 0.35216351*SO2 + 0.44972472*CO + 0.37906451*OZONE +0.03069609*PRECIP_COVER + 0.45359676*WIND_SPEED + 0.08356036*CLOUD_COVER + 0.29800233*VISIBILITY -0.7937893*CONDITION_status + 0.49830492*WIND_DIRECTION_status
Predicted_AQI
Enter the measured value of (PM2.5 + PM10)*0.5 (micrograms per cubic meter of air)
Enter the measured value of NO2(parts per billion in air)
Enter the measured value of NH3(parts per million in air)
Enter the measured value of SO2(parts per billion in air)
Enter the measured value of CO(parts per million in air)
Enter the measured value of OZONE(Dobson unit)
Enter the measured value of PRECIPITATE COVER(mm in 24 hours)
Enter the measured value of WIND SPEED(KM/H)
Enter the observed value of CLOUD COVER(Percentage)
Enter the measured value of VISIBILITY(KM)
Enter the observed value of CONDITION status(1 for Rain and partially cloudy, 2 for Rain and overcast,3 for Partially coludy,4 for Overcast,5 for Clear))
Enter the observed value of WIND_DIRECTION_status(Direction(1 for NE, 2 for E,3 for SE,4 for S,5 for SW,6 for W,7 for NW,8 for N))
```

8. ARIMA Model

```
# Checking Whether our AQI Time series data is Stationary or not
plt.figure(figsize=(40,10))
BHU_AQI_clean["AQI"].plot()
plt.ylabel("AQI")
plt.xlabel("Observation")

#Using Rolling Method to check Stationarity of Time Series Model
rollmean=BHU_AQI_clean['AQI'].rolling(25).mean()
rollstd=BHU_AQI_clean['AQI'].rolling(25).std()
plt.figure(figsize=(20,7))
```

```
fig=plt.figure(1)
original=plt.plot(BHU_AQI_clean['AQI'],color='blue',label='Original')
mean=plt.plot(rollmean,color='red',label='Rolling Mean')
std=plt.plot(rollstd,color='black',label='Rolling std')
plt.legend(loc='best')
plt.title('Rolling Mean and Standard deviation')
plt.show()

#Checking for possible components of time series
dec=seasonal_decompose(BHU_AQI_clean.AQI,period=1,model='multiplicative')
trend=dec.trend
seasonal=dec.seasonal
residual=dec.resid
plt.figure(figsize=(20,15))
fig=plt.figure(1)
plt.subplot(411)
plt.plot(BHU_AQI_clean['AQI'],label='Original')
plt.legend(loc='best')
plt.subplot(412)
plt.plot(trend,label='Trend',color='orange')
plt.legend(loc='best')
plt.subplot(413)
plt.plot(seasonal,label='Seasonality',color='green')
plt.legend(loc='best')
plt.subplot(414)
plt.plot(residual,label='Residual',color='red')
plt.legend(loc='best')
plt.show()

#Trying Differencing to make Time Series Stationary
plt.figure(figsize=(20,7))
AQI_diff=BHU_AQI_clean['AQI']-BHU_AQI_clean['AQI'].shift(1)
plt.plot(AQI_diff)
rollmean=AQI_diff.rolling(25).mean()
rollstd=AQI_diff.rolling(25).std()
fig=plt.figure(1)
original=plt.plot(AQI_diff,color='blue',label='Original')
mean=plt.plot(rollmean,color='red',label='Rolling Mean')
std=plt.plot(rollstd,color='black',label='Rolling std')
plt.legend(loc='best')
plt.title('Rolling Mean and Standard deviation')
plt.show()

#Checking whether lag 1 can make series Stationary
AQI_diff=BHU_AQI_clean['AQI']-BHU_AQI_clean['AQI'].shift(1)
AQI_diff1=AQI_diff.dropna()

#Null hypo that the time series data is non stationary
```

```
#Alter hypo that is stationary
#Assume alpha=0.05, meaning 95% confidence. The test are interepted with p-value if p>0.05 reject null
#Null hypo that the time series data is non stationary
#Alter hypo that is stationary
#Assume alpha=0.05, meaning 95% confidence. The test are interepted with p-value if p>0.05 reject null
dftest=adfuller(AQI_diff1,autolag='AIC')
print("1. ADF :",dftest[0])
print("2. P-Value :",dftest[1])
print("3. critical values :")
for key, val in dftest[4].items():
   print("\t",key,":",val)

#Checking what value of p and q
pacf=plot_pacf(AQI_diff1,lags=25)
acf=plot_acf(AQI_diff1,lags=25)

#Checking for best (p,d,q) for our model
!pip install pmdarima
import pmdarima as pm
def arimamodel(timeseries):
   automodel=pm.auto_arima(timeseries,start_p=1,sttart_q=1,max_p=5,max_q=5,test='adf',seasonal=True,trace=True)
   return automodel
arimamodel(BHU_AQI_clean['AQI'])

#Visualization of Original Vs Predicted AQI
plt.figure(figsize=(20,7))

model=sm.tsa.arima.ARIMA(BHU_AQI_clean['AQI'],order=(1,1,1))
result=model.fit()
print(result.summary())

#Visualization of Actual and Predicted AQI through ARIMA(1,1,1) model
fig=plt.figure(figsize=(20,7))
original=plt.plot(BHU_AQI_clean['AQI'],color='blue',label='Original AQI')
predicted=plt.plot(result.fittedvalues,color='red',label='Predicted AQI')
plt.xlabel('Observation')
plt.ylabel('AQI')
plt.legend(loc='best')
plt.title('Prediction on ARIMA(1,1,1)')
plt.show()

#Correlation between AQI and Predicted AQI
np.corrcoef(BHU_AQI_clean['AQI'],result.fittedvalues)

Predicted=result.fittedvalues
Residual=BHU_AQI_clean['AQI']-result.fittedvalues
fig, ax = plt.subplots(figsize=(12, 4))
sns.scatterplot(x=BHU_AQI_clean['AQI'],y=Residual,color="red" )
```

```
plt.ylabel("Residuals")
plt.hlines(Residual.mean(),Residual.min(),230)

BHU_AQI_clean['AQI'].mean()
rmse=sqrt(mean_squared_error(result.fittedvalues,BHU_AQI_clean['AQI']))
print(rmse)

ARIMA_model=sm.tsa.arima.ARIMA(BHU_AQI_clean['AQI'],order=(1,1,1))
Prediction_ARIMA=ARIMA_model.fit()
BHU_AQI_clean.tail()
```

9. Prediction for next 30 days using ARIMA(1,1,1) model

```
index_future_dates=pd.date_range(start='2023-2-21',end='2023-3-23')
#print(index_future_dates)
Predicted_ARIMA=Prediction_ARIMA.predict(start=len(BHU_AQI_clean),end=len(BHU_AQI_clean)+30,type='levels').ren
ame('ARIMA Prediction')
#print(comp_pred)
Predicted_ARIMA.index=index_future_dates
print(Predicted_ARIMA)
Predicted_ARIMA.plot(figsize=(20,7),legend=True)
```