

University of Connecticut School of Business

OPIM 5671 – Data Mining and Business Intelligence

**DAILY CLIMATE TIME SERIES AND FORECASTING**

**Instructor**

Prof. Jing Peng

jing.peng@uconn.edu

**Team - A7**

Rounak de

Rushikethan Reddy Dudipala

Vamsi Krishna Malempati

Vinay Reddy Poreddy

Satish kumar MNV

**Table of contents**

[**Introduction 3**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.gnoj9tcixlwx)

[**Literature 3**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.nd97462lnnyu)

[**Data Exploration 5**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.ot7u92l0crvd)

[**Data Description 6**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.x0owg5fie4t7)

[**Basic Statistics 7**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.h3gfy91nwxz)

[**Data Pre-Processing 7**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.7f8916hx0swh)

[**Data Analysis 8**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.uqsgt6o5dfyv)

[**Time Series tests 9**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.p6ghkfga5nh9)

[**Data Visualization 11**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.peaw3yx5ehix)

[4.   Relation between regressors 12](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.frnnjub42fmu)

[5.   Relation between the output variable and regressors: 13](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.deu55kfad3zu)

[**Predictive Modeling 15**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.fxt6e65i8bme)

[a. Teams Approach 15](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.hv7e6qa3r2th)

[b. Models Used 16](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.7dtf1660fo8b)

[c. Models Comparison 18](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.mj69krn63x8y)

[**Finding**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.5rf2quvxgn5g)**s** [**19**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.5rf2quvxgn5g)

[**Responsibilities of Each Team Member 19**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.e771he4mhy9f)

[**References 20**](https://docs.google.com/document/d/1EUhEnXRo_31OftJsCXUTm43WrZww90hy/edit#heading=h.fd7hbc2hz8qa)

**Introduction**

The "monsoon" type of climate that characterizes India is primarily found in South and South-East Asia. The Arabic term "Mausam," which denotes seasons, is the source of the English word "monsoon." The term "monsoon" was first used by Arab navigators several centuries ago to refer to a system of seasonal wind reversals that occurs along the Indian Ocean shorelines, particularly over the Arabian Sea, with the winds blowing from southwest to north-east during the summer and from north-east to south-west during the winter. In other terms, monsoons are periodic (seasonal) winds in which the wind direction completely reverses every six months.

The monsoon-influenced humid subtropical climate of Delhi has significant seasonal variation in both summer and winter temperatures and precipitation. Delhi's humid subtropical climate differs noticeably from that of many other humid subtropical cities, such as New Orleans and Brisbane, in that the city experiences dust storms, which are more typical of a desert climate, and wildfire haze during the dry season, which is typical of a tropical climate.

Average summertime temperatures are around 38 °C (100 °F), with the season's highs occurring in late May or early June. However, from time to time, heat waves can cause highs to occasionally reach as high as 45 °C (113 °F), giving the impression that the temperature is higher. With an average rainfall of 797.3 mm (31.39 inches), the monsoon begins in late June and lasts until mid-September. The temperature is often around 29 °C (84 °F), however on rainy days, it can drop to about 25 °C (77 °F) or rise to 32 °C (90 °F) during dry spells. With average temperatures falling from 29 to 21 °C (84 to 70 °F), the post-monsoon season lasts until late October after the monsoons have subsided.

With an average temperature of 14 °C (57 °F), winter begins in November and peaks in January. Even though Delhi experiences warm days, the Himalayas nearby cause cold waves that cause wind chill to drop the apparent temperature. Delhi is well known for its wintertime haze and severe fog. Road, air, and train traffic are all disrupted in December due to poor visibility. [3] The first week of March marks the end of winter, 2.2 to 49.2 °C (28.0 to 120.6 °F) of extreme temperatures have been recor

**Literature**

Weather forecasting is a very important aspect of modern human life where people plan a lot of things based on the upcoming weather situation. Especially in a city like Delhi which is notorious for its unpredictable climate, weather forecasting is something very important. We would be able to warn the residents of an upcoming heat or cold wave based on the time of the year for which we are predicting. It also helps people to prepare for other hazards like smog for which Delhi is notoriously famous for.

There is a unique situation that happens in Delhi during the winter time. This situation is known as smog. This happens when the temperature is low and the wind speed is also low. Normally during the winter time wind speed is on average higher. But during days of cold weather and low wind speed combined with the terrible pollution, this situation leads to the creation of smog.

In summer Delhi suffers from sudden heat waves. The government takes necessary precautions for this which includes shutting down of schools, warning the people about extreme temperatures etc. Naturally as a precautionary measure to plan about these events, it is very important to have an idea about the upcoming weather situation. Predicting the weather using past data gives us an idea about how the weather pattern may look in the upcoming year.

This helps the government to come up with adequate measures to combat extreme weather and helps the government to do policy making. Odd even number plate rules for taking out cars in winter to reduce the smog and pre planning of summer holidays for schools and other educational institutions during the summer heat wave are some of the government policies that depend on the prediction of the upcoming weather situation.

**Data Exploration**

Our Data Source is used to forecast the climate temperature in the location at Delhi, This Data contains 1462 Data entries and the Dataset can have 5 columns.

**A screenshot of a computer

Description automatically generated with low confidence**

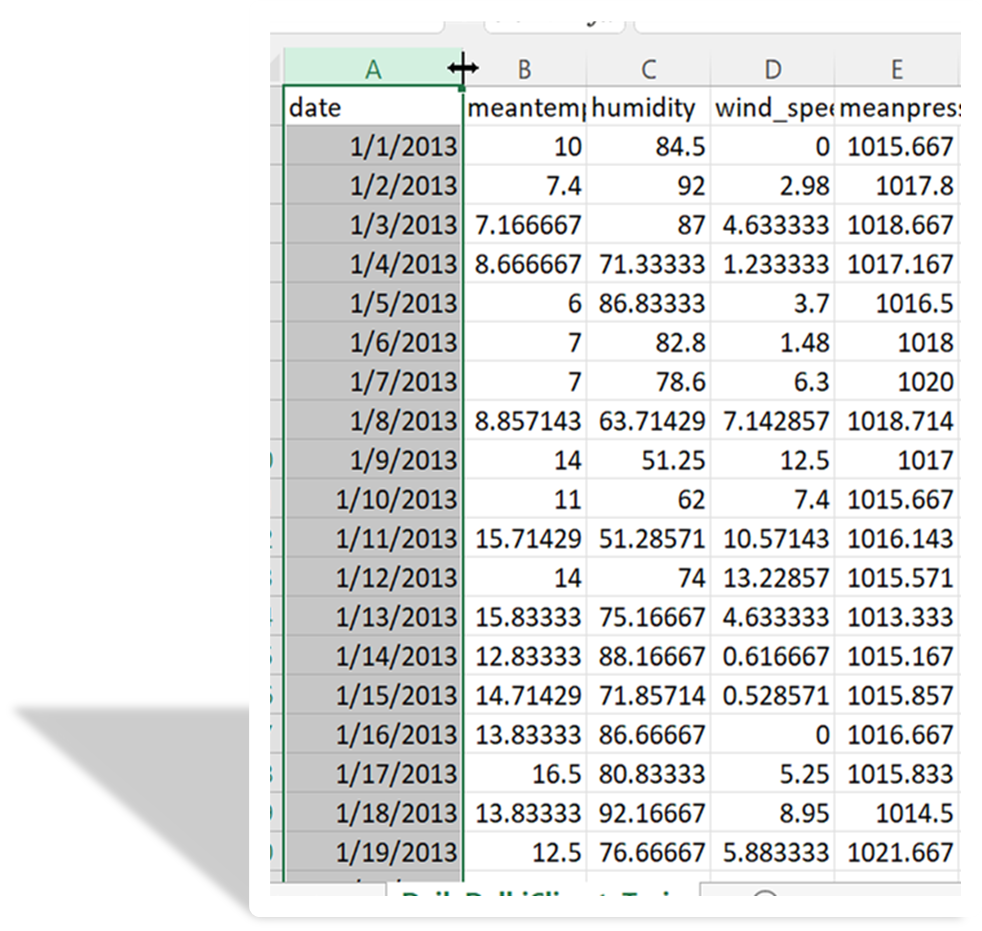
**Table

Description automatically generated**

Fig:Daily Climate Time Series Dataset from Jan 2013 to Jan 2017

**Data Description**

We have collected this data from Kaggle, and a respective link is provided below. This dataset contains the data from 2013 to 2016 through which we are about to forecast the temperature levels in Delhi. The parameters are mean temperature, humidity, wind speed, mean pressure.

****

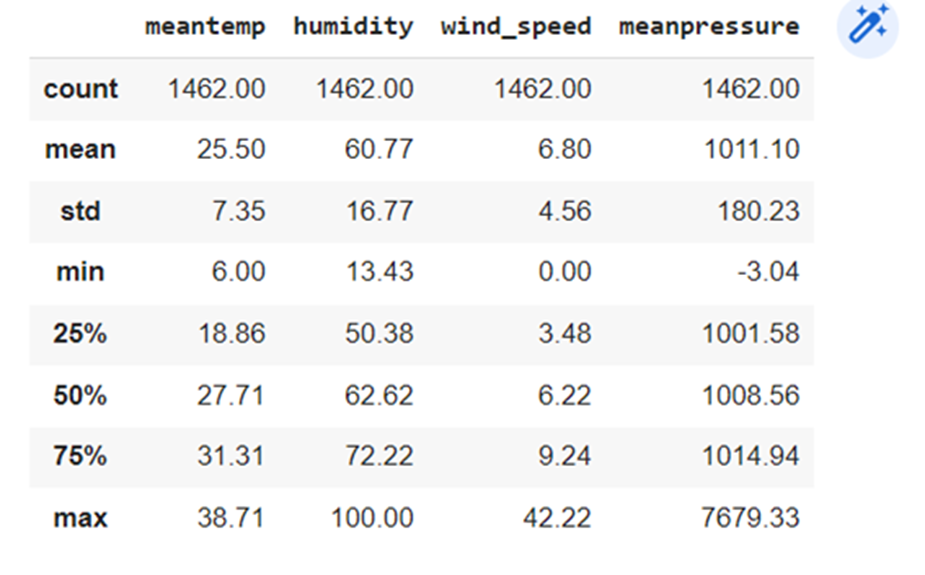
**Data Dictionary**

We have five parameters in our data set through which we are going to forecast the temperature in Delhi. The Date parameter plays an important factor in our project as we predict the temperature based on dates only. Temperature parameter is our target variable. Humidity, Wind speed and Mean pressure variables are dependent upon each other which varies with one another.

1. **Mean Temp:** Average of multiple temperature observations made at three-hour intervals throughout each day. Expressed in degree Celsius
2. **Humidity:** Amount of humidity, expressed in grams of water vapor/cubic centimeter
3. **Wind Speed:** Measured in kilometers per hour
4. **Mean Pressure:** Measured in atmospheres
5. **Year:** It represents which year is it
6. **Month:** It represents month number
7. **Month Name:** It represents the month name
8. **Day of the week:** It tells which day of the week is it
9. **Day Name:** It represents the day name

# **Data Summary Statistics**

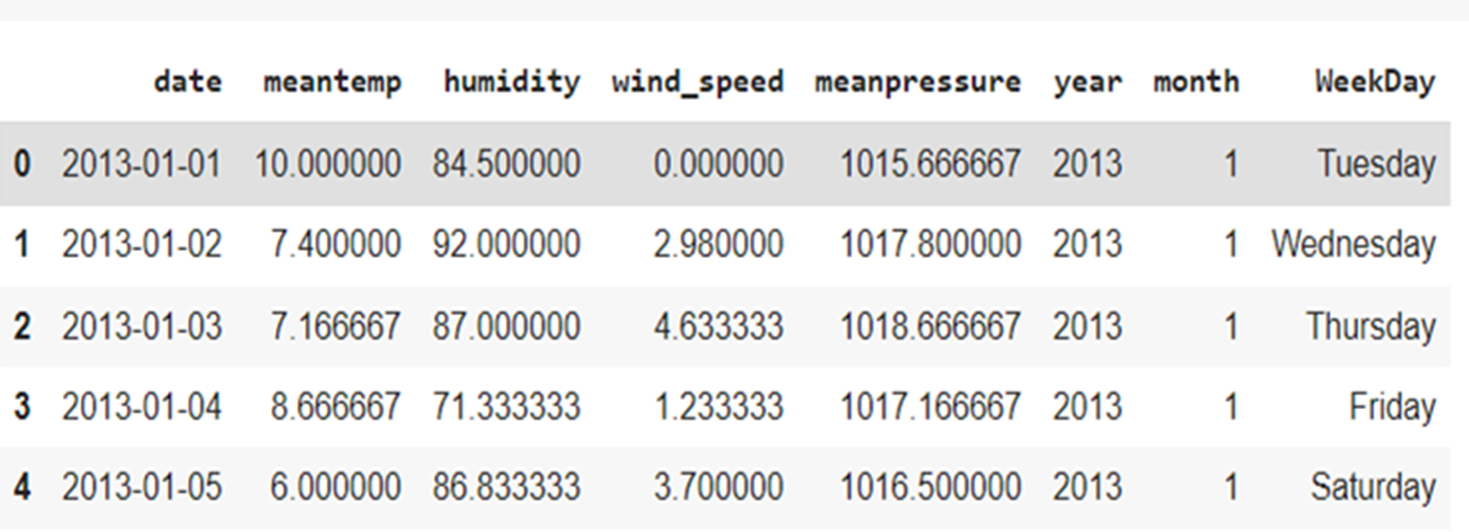
From our below figure we can see the statistics of our dataset. The mean temperature is 25.50 which is an average indicator of temperature. The standard deviation is very low which indicates that there are other variables which play a substantial role in both the growth and decline of the temperature levels. Delhi wind speed ranges from 0 kilometers per hour to 42.2 kilometers per hour, demonstrating the importance of climatic changes in Delhi, which have an impact on temperature levels.



# **Data Pre-Processing**

The most crucial stage of any project using data is data preprocessing. The dataset we used in this study also required pre-processing activities to make it model ready. Therefore, we carried out several data preprocessing steps.

The data we received was on an hourly basis. We transformed the data daily and reduced the granularity due to the size of the data. From this we didn't get any null values, so we did not replace them with zero or mean values, but we have added a few parameters like month, weekday to forecast the temperature levels on a monthly bias.



**Feature Engineering**

Data transformation is called feature engineering, and it is used to improve the prediction capabilities of machine learning models. Python and Pandas should be familiar to you by this point. On Dataquest, you may interactively learn about both. How to conduct various feature engineering tasks with NumPy and Pandas in Python

Frequently more accurate prediction: By improving the weighting of variables, feature engineering approaches like standardization and normalization frequently increase accuracy and occasionally hasten convergence.

Improved relationships in the data can be interpreted: Our comprehension of the data expands as we construct additional features and comprehend how they connect to our outcome of interest. We may still earn a good assessment score if we omit the feature engineering step and utilize sophisticated models, but we would better understand our data and how it relates to the goal variable.

Most models do not support all possible data representations, hence feature engineering is required. Missing values must be imputed in order for models like linear regression to manage them (filled in). Examples of this can be found in the section that follows.

A picture containing graphical user interface

Description automatically generated

Table

Description automatically generated

**Data Analysis**

After completing the data preprocessing stages, the next item we were interested in and something that needed to be checked was to see if the time series plot of our target variable, that is the mean temp plot is stationary or not. As a result, we must determine whether the time series plot has trend and seasonality. By observing the below figure, we can say that there are no continuous ups and downs in the graph which makes it difficult to determine whether the data has seasonality. So, to find that we must test for the white noise, unit root, and seasonality by using SAS.

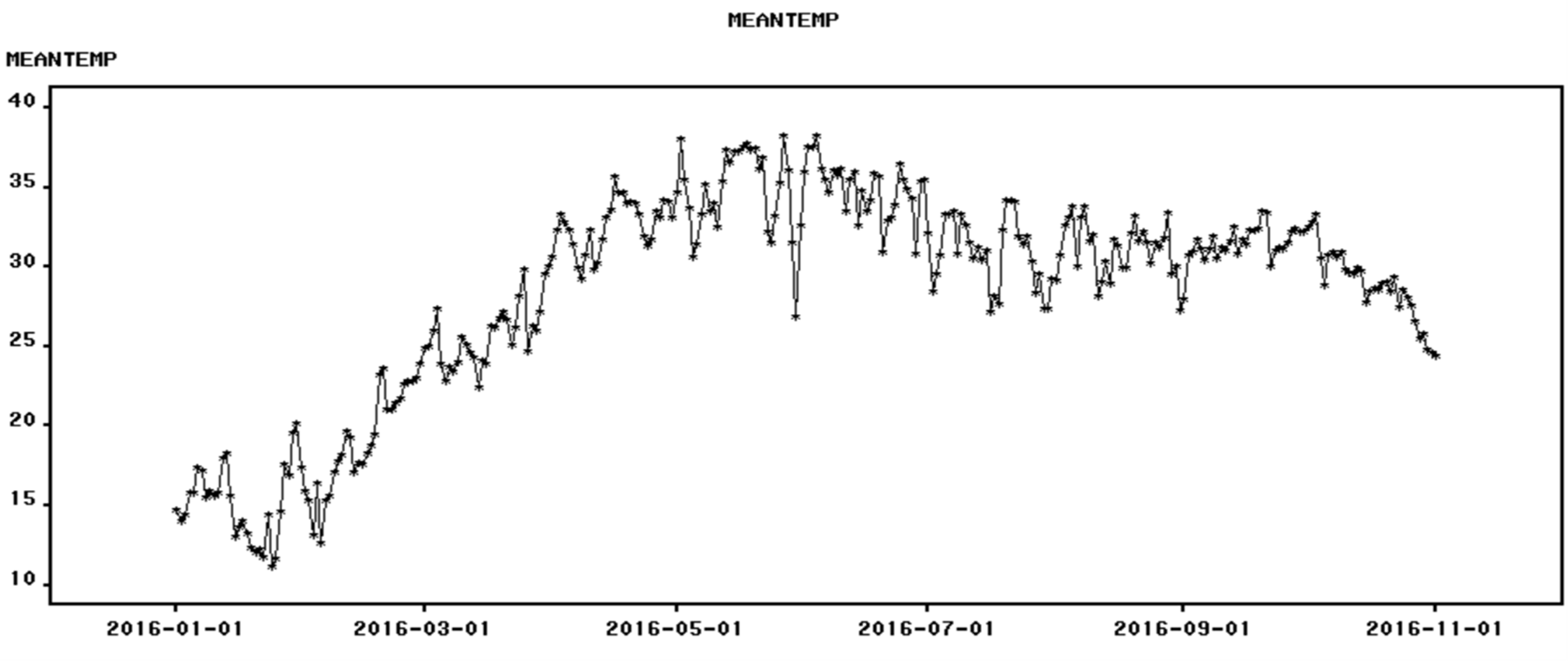


Fig: Figure shows no trend in the time series plot.

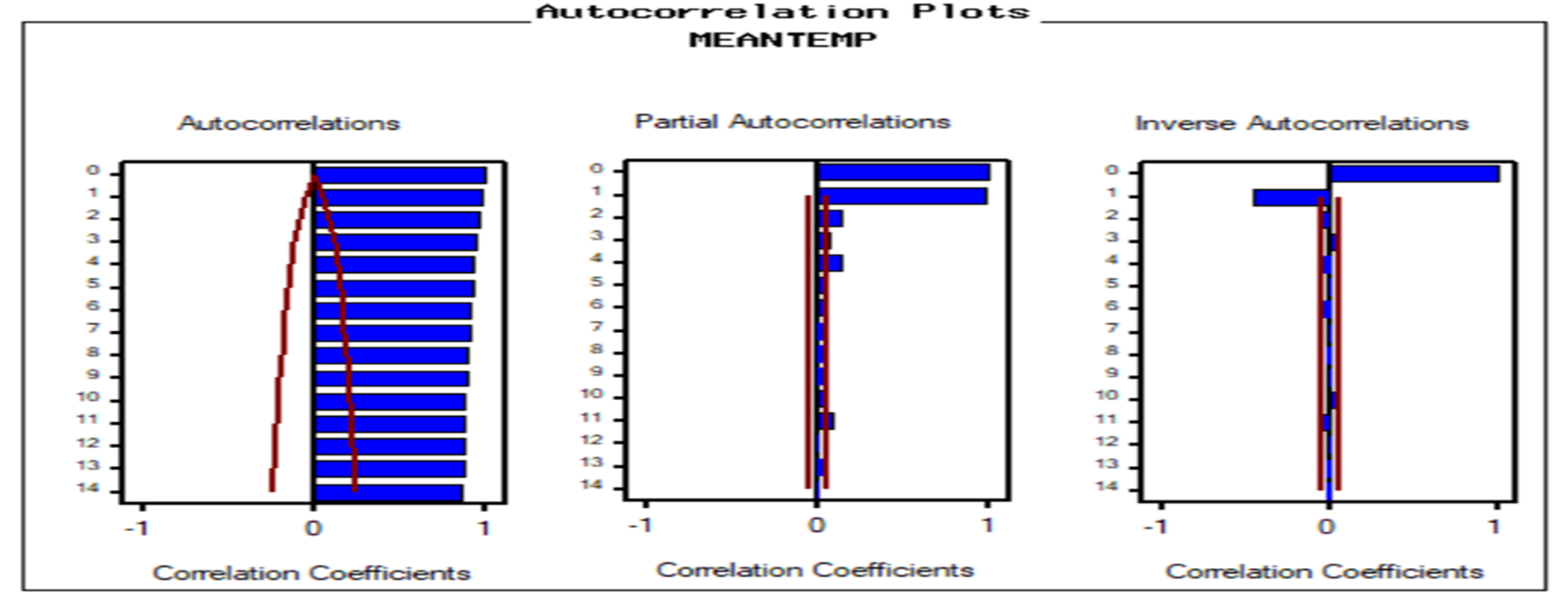
# **Time Series tests**

We must first look for trends, seasonality, and stationary conditions before forecasting the data for a time series. If a time series' mean, variance, and covariance remain constant across time, it is said to be stationary. If a time series is stationary, it lacks trend and seasonality.

Chart, diagram, bar chart

Description automatically generated

We can see white noise testing, unit root tests, and second root tests in the figures above. From the unit root test plot, we can infer that the time series is stationary throughout time and that there is no white noise present in it. This is evident from the white noise test plot. There is no seasonality, according to the seasonal root test plot for our time series.



From the above autocorrelation points we can conclude that there is no trend and seasonality. Similarly from the Partial Autocorrelations, we can observe there are the two stripes, based on this we proceeded to ARIMA models, and not included the Winters Addi

**Time Series Exploration steps**

In the Time Series We design a Workflow How we approach the Model building, we collected data by the dataset ,we removed the null values in the Data that process is called as Data Cleaning and after we will visualization of data after that we will start with Model building so that first we should do model identification after that we are doing estimation and validation finally we did Time series Forecasting.

Diagram

Description automatically generated

Diagram

Description automatically generated

Fig: Finding Mean Temperature based on different years in the location

From the above graph we are saying that there is no seasonality for comparing with the first and the third waves so there is no seasonality in the above graph and then I zoomed the graph and we took another screenshot of it.

A picture containing diagram

Description automatically generated

# **Data Visualization**

The study of data visualization involves the visual display of data. It effectively communicates findings from data by plotting data graphically.

A bar graph is a useful tool for categorical data representation. With the use of bars, which stand in for value on the y-axis and category on the x-axis, a bar graph presents data. To display data that falls into a particular category, bar graphs utilize bars of varied heights.

In the below graph there is a difference between the Different months with the Mean temp and we are comparing the years with the Temperature with the highest temperature held in the month and least temperature held in which month then we are divided The different month into season’s.

Chart, bar chart

Description automatically generated

A bar graph is a useful tool for categorical data representation. With the use of bars, which stand in for value on the y-axis and category on the x-axis, a bar graph presents data. To display data that falls into a particular category, bar graphs utilize bars of varied heights.

Mean Temperature over the Span of 5 years Line chart is a graph that represents information as a series of data points connected by a straight line. In line charts, each data point or marker is plotted and connected with a line or curve.

**Humidity Levels Over the Span Of 5 Years**

Chart, line chart

Description automatically generated

We want to know about the difference between the years with the Humidity level in the years of 2013 to 2017 if the temperature is high at the same time the humidity levels can also get high.

Chart, line chart

Description automatically generated

In the above graph we are comparing the data with the temperature with the span of 5 years and there are the 3 seasons in the location of Delhi. They are different temperatures like summer, which are very high and so in the graph it has ups and downs.

**Chart, line chart

Description automatically generated**

Now we are discussing the Year with the wind speed and there are some high strikes in the graph so that means in the winter season in India, at that time the temperature is constant but the wind speed is high that particular time.

**2) Mean Temperature over the year**

Chart, bar chart

Description automatically generated

In the above graph we can see that there is a month wise data and there is 12 months with 3 different seasons so that it compares the data with the temperature with the different months and seasons and in the month of 11,12,1,2 are the winter months in India so the temperature is low.

**Chart, bar chart

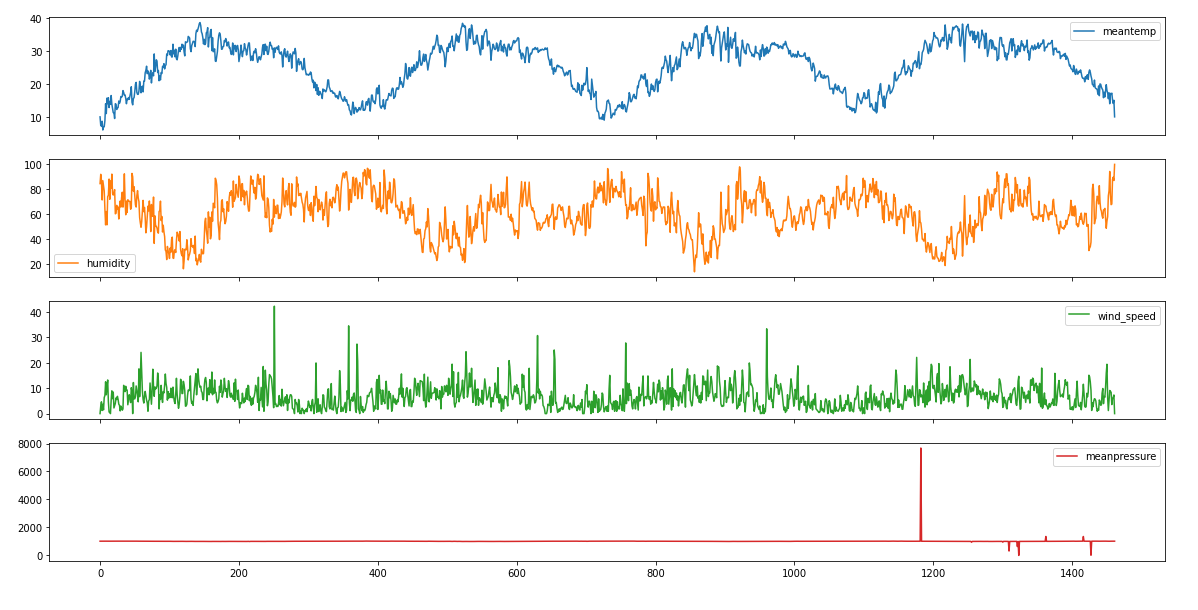
Description automatically generated**

Here we are discussing the Humidity with the different months and if the temperature is high the humidity is also high and sometimes the temperature goes down but the humidity remains the same or sometimes it may be high.

**Chart, bar chart

Description automatically generated**

In the above graph we can see the difference between the months and the Wind\_speed according to the season, so in the rainy season the wind speed is high compared to the other months in a year.

****

Here we can see that the difference between the mean temperature with humidity and wind speed so that the temperature is very high in summer by that the humidity is very low and also there is no wind speed and in the remaining two seasons like rainy and the winter seasons at that time or the a that months the temperature will get down so the wind speed goes up and humidity remains constant.

**3) Seasonal Difference with Mean**

We added a new column in the table as a season and we split our months data into different seasons such as summer, winter, and rainy seasons. So finally the below graph says that the mean value of the total months can be shown below.

**Chart, box and whisker chart

Description automatically generated**

**Predictive Modeling**

The predictive model is a data model that uses inferential statistics to forecast how consumers will react to a marketing campaign or a particular investment. The predictive model, which employs statistics to forecast outcomes, is frequently developed by data scientists.

**Models developed**

**Text

Description automatically generated**

We as a team started to discuss and come up with several ideas to ensure which model and what approach would be best suitable for our time series prediction. When we look for several models, SAS has numerous models for the time series forecasting. Any model to be concluded as the best model, we should always consider a few important factors, one of such is whether the model displays any trend or seasonality.

To start with, we first explored our dataset with the SAS Time Series, this helps us know if any trend or seasonality exists. We haven’t encountered any trend or seasonality.

Furthermore, we started developing various models to determine which model is more effective with our dataset. We started with the Seasonal Dummies model, since our forecast is relevant to several seasons and their seasonality. We then started applying Linear Trend, Winters Additive, and many more, however, all of these are giving a very high Mean Absolute Percent Error(MAPE) value, therefore we then started developing ARIMA models. We tried with different Auto-regressive(AR) values for the parameter p, Moving Average(MA) values for the parameter q and also the Number of Differences(d), by this we have observed that the MAPE values are getting lower with different values. After obtaining the lowest MAPE model, we then started adding regressors which are effective with the best model. Of all the regressors, we found that the regressor “humidity” gave us the least MAPE model, followed by the regressors wind\_speed and mean pressure.

Our best model is “Seasonal Dummies + humidity + AR(2)” with the lowest possible MAPE value of 3.9

**Model Forecast**

**Chart

Description automatically generated**

The above figure depicts the forecast for MEANTEMP for the best model. Since the prediction depends on several months in a year, we considered to depict the data in terms of months for the corresponding four years data using python visualization.

In the above forecast, we can observe that the temperature goes down from November,2016 to January,2017 and gradually goes up from February,2017 to May,2017. This means during the rainy and winter months the temperature is low and keeps declining, whereas for the summer months the temperature of Delhi gets increasing and reaches its peak around June and July.

**Model Comparison**

**Chart

Description automatically generated**

By the above graph we compared the top 4 models and that can be indicated four different lines with the four different colors and here in the above graph it indicates the months with the temperature level and here the red most color is the our best model comparison and also we can see that most of the month like june, july and august are the months with equal temperature.

**Best Model**

After successful comparison of models, we can conclude that our best model has the highest R-Square value of 0.967, RMSE of 0.99 and lowest MAPE of 3.90

**Table

Description automatically generated**

**Graphical user interface

Description automatically generated**

**Parameter Estimates for our Best Model**

**Table

Description automatically generated**

From the above table we can observe that the Parameter estimates are high for Autoregressive lag 1 & 2, Seasonal Dummies and for the regressor humidity. Which means these play a vital role in predicting the seasonality of the city Delhi.

# **Findings**

* From the past data of four years, we are able to forecast the temperature levels for the upcoming months of the city Delhi , which can be used to guide any necessary precautions.
* From the visualiations, we can observe that the temperature is extremely high in the months April, May and June which are basically the summer months in Delhi.
* We also observed that the high temperature automatically increases the likelihood of a high humidity level, which leads to a high wind pressure. This observation can be used to make certain precautions by the Delhi government to safeguard the people.
* During the months of November, December, January which are winter months for Delhi, there are high chances for formation of fog which results in bad and dangerous accidents, hence in order to overcome this the Delhi’s government could devise creative ideas such as odd-even rule, which reduces the flow of traffic and can be helpful.
* Since the temperature of Delhi is unpredictable sometimes, we advise the government to make people aware of it and take precautions at least 15 days prior to the start of summers and winters.

# **Responsibilities of Each Team Member**

| Data preprocessing using Python | Vamsi Krishna Malempati/Satish Kumar MNV |
| --- | --- |
| Exploratory data analysis | Vamsi Krishna Malempati/Rounak De |
| Modeling |  |
| Presentation & Report | Rounak De/Vinay Reddy Poreddy/Rushikethan Reddy Dudipala |

# **References**

1. Data source

(<https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data?select=DailyDelhiClimateTrain.csv>)

1. India : Climate, vegetation and wildlife (<https://ncert.nic.in/ncerts/l/fess208.pdf>)
2. Climate of India

(<https://cdn1.byjus.com/wp-content/uploads/2022/02/Climate-of-India.pdf>)

1. Monthly, Seasonal and Annual Rainfall Time Series for All-India, Homogeneous Regions and Meteorological Subdivisions: 1871-2016

(<https://tropmet.res.in/~lip/Publication/RR-pdf/RR-138.pdf>)

1. Distribution of Temperature over the National Capital Territory of Delhi - A Preliminary Report

(<http://imetsociety.org/wp-content/pdf/vayumandal/2014/2014_5.pdf>)