THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Visualization of Complex Data

DATS 6401

Final Term Project (FTP)

Instructor: Dr. Reza Jafari

Date: 26th April, 2024

Project-

WEATHER DASHBOARD

Aravinda Vijayaram Kumar

TABLE OF CONTENTS

TABLE OF FIGURES AND TABLES

ABSTRACT

Weather Insights at Your Fingertips: An Interactive Dashboard for Exploring India's Climate and Global Weather has a profound impact on our daily lives, from planning outdoor activities to managing critical infrastructure. To utilize the wealth of weather data and empower users with actionable insights, I have created an interactive Weather Dashboard that integrates data from India's weather repositories and global repositories obtained from Kaggle. This dashboard offers a dynamic interface to explore, visualize, and analyze a comprehensive range of weather parameters, including temperature, humidity, wind speed, air quality, and cloud cover. By using data preprocessing techniques, outlier management, and statistical analyses like Principal Component Analysis, the dashboard ensures the integrity and usability of the data.

The Weather Dashboard is a tool for presenting different meteorological information in an accessible format, enabling informed decision-making for a diverse audience. Users can delve into country and city-specific weather conditions, access forecasting, and download the data sets for offline analysis. The Weather Dashboard serves as a valuable resource for a variety of real-world applications, including Climate Analysis, Weather Prediction, Environmental Impact Assessment, Tourism Planning, and Geographical Weather Patterns. By leveraging the data and visualizing the information we can develop a deeper appreciation for the intricate interplay of weather phenomena and their impact on the world around us.

INTRODUCTION

This Final Term Project is an interesting and comprehensive exploration of India's weather patterns, complemented by a global weather dataset for a broader perspective. The project began with a dataset search and was followed by a simple yet effective data preprocessing stage, where I cleaned and structured over 100,000 data entries, laying a solid foundation for accurate analysis. Recognizing the potential influence of statistical outliers, I identified them and addressed these anomalies across various scales – globally, nationally, and at the city level. This attention to detail ensured the integrity and reliability of the data and the plots.

The next step involved streamlining the dataset through Principal Component Analysis (PCA), a powerful statistical technique that simplifies the complexity of the weather data while preserving its most informative aspects. Alongside PCA, I also performed a series of normality tests to confirm the data's alignment with the standard normal distribution assumption.

With a refined dataset in hand, I plotted a diverse array of static visual representations, from histograms that outlined distribution curves to box plots that revealed variability and central tendencies within the weather parameters. These visualizations not only illustrated the data's narrative but also set the stage for the project's centerpiece: an interactive dashboard developed using Dash and Plotly.

The interactive dashboard offers users an intuitive interface to engage with and extract meaning from the weather data. Through customizable plots and views, users can delve into the understanding of the weather patterns of temperature, humidity, wind speed, air quality, and other meteorological factors, both at the national and global levels. One of the standout features of the dashboard is the 'Forecast' tab, which integrates with an external API to provide live weather forecasts. This innovative addition transforms the dashboard into a forward-looking tool, empowering users with the ability to anticipate weather changes in real time and plan accordingly. To ensure an engaging user experience, the dashboard features dropdown menus for selecting specific cities or regions, a date range slider for temporal navigation, and interactive maps for a tangible feel of the data. With these tools, users can dive into the depths of weather trends, compare regional variations, and interact with the changing tapestry of India's weather – all at the click of a few buttons.

Overall, this project presents an interesting mixture of plots and interactive dashboards in the area of weather analysis. By combining data preprocessing, statistical techniques, and creative visualization approaches, it presents weather information in an accessible format but also showcases the immense potential of such data to inform and enrich our understanding of the world's meteorological phenomena. In the future, once more data is available this dashboard, can also be updated to predict weather patterns, and weather anomalies such as any natural disasters, etc.

Overall, the dashboard enables climate analysis, aids environmental impact assessments, and serves as a planning companion for travelers. It also offers a lens through which to view geographical weather patterns, highlighting the contrasts and commonalities across different terrains.

DATASET

The "Indian Weather Repository Daily Snapshot" dataset from Kaggle and the " World Weather Repository" datasets are a comprehensive collection of weather data from India and various countries and cities around the world. These datasets meet the project criteria by providing a rich set of meteorological parameters with 42 columns, including temperature, humidity, wind speed, air quality, precipitation, and other relevant weather variables. The Indian dataset contains over 100,000 rows of data, and the global dataset contains over 34000 rows of data, with each row representing a specific location and the corresponding weather conditions observed on a particular day. The key variables in the dataset are:

Dependent Variable:

- temperature_celsius: The temperature in degrees Celsius, which can be considered the primary dependent variable for this project.

Independent Variables:

- humidity: The percentage of moisture in the air.
- wind_mph: The wind speed in miles per hour.
- pressure_mb: The atmospheric pressure in millibars.
- precip_mm: The precipitation in millimeters.
- Various air quality indices.
- sunrise, sunset, moonrise, moonset, moon_phase, moon_illumination: Astronomical data related to the sun and moon.

The inclusion of both numerical and categorical variables, such as weather conditions and moon phases, provides a rich dataset for comprehensive data analysis and visualization.

These datasets are highly valuable for a wide range of real-world applications, particularly in the fields of climate analysis, weather prediction, environmental impact assessment, tourism planning, and geographical weather pattern exploration.

1. Climate Analysis: The dataset allows for the study of long-term climate trends and variations across different regions, enabling researchers and policymakers to better understand the impacts of climate change.

2. Weather Prediction: By analyzing historical weather data, models can be developed to forecast temperature, humidity, precipitation, and other meteorological parameters, which is crucial for industries like agriculture, transportation, and disaster management.

3. Environmental Impact Assessment: The air quality indices in the dataset can be used to analyze the relationship between weather conditions and air pollution, supporting environmental impact assessments and informing policies for sustainable development.

4. Tourism Planning: The weather data can help travelers and tourism operators plan their activities and trips based on the expected weather conditions in different locations, enhancing the overall travel experience.

5. Geographical Weather Patterns: The dataset's comprehensive coverage of various countries and cities allows for the exploration of how weather patterns vary across different geographical regions, providing valuable insights for urban planning, infrastructure development, and disaster preparedness. The datasets are well-structured, with clear variable names and units of measurement. The inclusion of both numerical and categorical

variables, such as weather conditions and astronomical data, enhances the versatility and potential for insightful analysis using these comprehensive weather repositories.

DATA-PREPROCESSING

```
# Loading the data into dataframes and preparing the data
ind_data = pd.read_csv('IndianWeatherRepository.csv')
world_data = pd.read_csv('GlobalWeatherRepository.csv')
# joining the two datasets
data = pd.concat([ind_data, world_data], ignore_index=True)
# dropping 2 columns as not required
data= data.drop(columns=['region','last_updated_epoch'])
# viewing data after dropping 2 columns
print(data.head())
#changing the col to datetime format
data['last_updated'] = pd.to_datetime(data['last_updated'])
# creating a new month col for visualization
data['month'] = data['last_updated'].dt.month
# chking for null values
print(data.isna().sum())
#first 5 rows of data
print(data.head())
# dataset info
print(data.info())
# dadta statistics
print(data.describe())
```

Fig1. Data Loading and Processing

In this Stage of pre-processing the data, I have loaded the data into 2 separate data frames and concatenated them. I did not have to clean much of the data as I had very few null values, especially in the 'region; column only as the global dataset did not have this column and hence had to be dropped. I also dropped the 'last_updated_epoch' as it was a UNIX timestamp of the last_updated which I did not require anywhere in my visualizations. I also created a new 'month' column by extracting the months from the last updated column, to visualize monthly trends for the data. Then the data was checked for any null values which showed that there are no null values in the dataset. The dataset's top rows are displayed as below-

	country	location_name	latitude	longitude	timezone	last_updated	temperature_celsius	temperature_fahrenheit
0	India	Ashoknagar	24.57	77.72	Asia/Kolkata	2023-08-29 10:45:00	27.5	81.5
1	India	Raisen	23.33	77.8	Asia/Kolkata	2023-08-29 10:45:00	27.5	81.5
2	India	Chhindwara	22.07	78.93	Asia/Kolkata	2023-08-29 10:45:00	26.3	79.3
3	India	Betul	21.86	77.93	Asia/Kolkata	2023-08-29 10:45:00	25.6	78.1
4	India	Hoshangabad	22.75	77.72	Asia/Kolkata	2023-08-29 10:45:00	27.2	81

condition_text	wind_mph	wind_kph	wind_degree	wind_direction	pressure_mb	pressure_in	precip_mm	precip_in	humidity	cloud
Partly cloudy	12.8	20.5	281	WNW	1008	29.77	0	0	67	26
Sunny	9.6	15.5	287	WNW	1008	29.78	0	0	70	19
Partly cloudy	11.4	18.4	317	NW	1009	29.78	0	0	70	51
Cloudy	10.5	16.9	297	WNW	1009	29.8	0	0	76	65
Cloudy	10.1	16.2	274	W	1009	29.79	0	0	74	82

visibility_km	visibility_miles	uv_index	gust_mph	gust_kph	air_quality_Carbon_Monoxide	air_quality_Ozone	air_quality_Nitrogen_dioxide
10	6	7	14.8	23.8	243.7	45.8	1.7
10	6	7	11.2	18	240.3	38.3	2.1
10	6	7	13.2	21.2	220.3	57.2	0.6
10	6	6	13	20.9	200.3	25	1.2
10	6	6	11.6	18.7	257	30.8	2.2

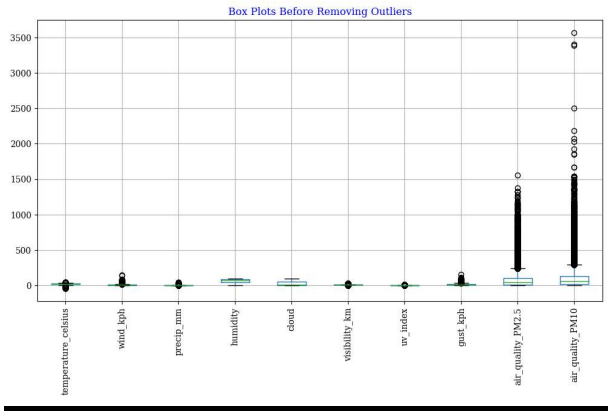
Fig2. First few rows of the dataset

OUTLIER DETECTION

Outlier detection is a method used to find unusual or abnormal data points in a set of information. In data, outliers are points that deviate significantly from the majority, and detecting them helps identify unusual patterns or errors in the information. This method is like finding the odd one out in a group, helping us spot data points that might need special attention or investigation **[1].**

For Skewed distribution such as the data in this dataset, we can use the Inter-Quartile Range (IQR) proximity rule. According to this rule, the data points that fall below ***Q1 – 1.5 IQR*** or above the third quartile ***Q3 + 1.5 IQR*** are outliers, where Q1 and Q3 are the **25th** and **75th percentile** of the dataset, respectively. IQR represents the inter-quartile range and is given by Q3 – Q1 **[1]**.

Also, for this dataset, which is pertaining to weather, we cannot consider a data point as an outlier based on the entire dataset as well. The main reason is that weather and climate are more of a regional or local phenomenon and vary based on geography. Hence 40°C is an outlier in a temperate region, but for a tropical region or equatorial region, it is very natural or common. Hence for outliers in weather pattern, I have done it at 3 levels, one at the global level data, which may or may not give much data about outliers, then next we do it at the specific country level, which is more detailed and is more accurate than the previous one. Finally, I did perform outlier detection for the particular city which shows the real anomalies in the data and the real outlier for that location. This 3-level approach has been included in the interactive dashboard and the global outlier detection has been plotted as a static plot as well. The following is an outlier detection plots for a few important columns-



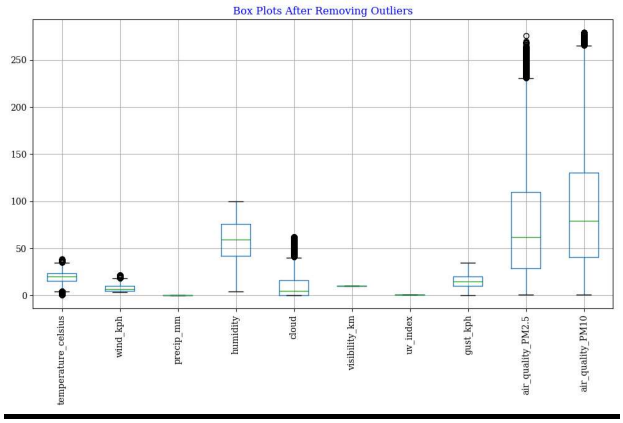


Fig3. Before and after Outlier detection on global data

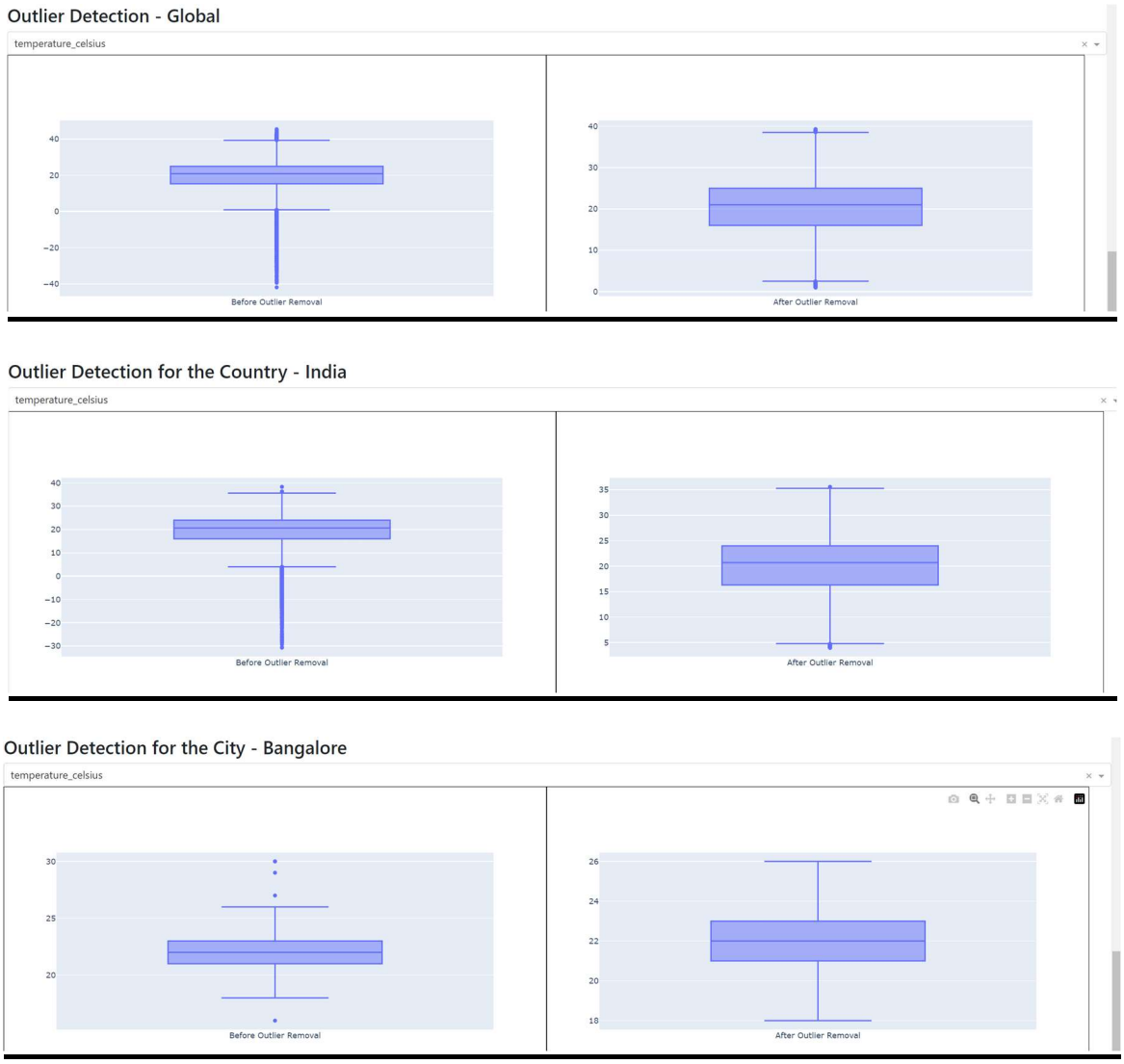


Fig4. Outlier detection in 3 levels- global, country, city

Even after the outlier detection, we cannot discard the values in the weather data as outliers. Outliers in weather data are extreme or unusual weather observations that deviate significantly from the expected norms. These often represent important meteorological events that provide valuable insights into the climate of that region hence discarding these outliers without proper justification may lead to the loss of critical information and the potential to understand the full range of weather patterns.

PRINCIPAL COMPONENT ANALYSIS

I performed Principal Component Analysis (PCA) to address the dimensionality reduction of the weather dataset variables. Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set **[2].** So, these are the key steps performed-

- Data Selection for PCA: Selected a subset of the weather dataset variables to be included in the PCA analysis. This was done so that only unique measurements are included and redundant data, such as temperature, and wind speed in different units are excluded.
- Data Standardization: Before performing the PCA, we standardized the selected columns using the StandardScaler from the sci-kit-learn library.
- Then applied the PCA algorithm to the standardized data.
- The Explained Variance attribute of the PCA model that reveals the proportion of the dataset's total variance that is captured by each principal component is printed.
- Additionally, calculated the condition number of the data matrix, which provides a measure of the sensitivity of the dataset's numerical solution to errors.
- To better understand the PCA results, created two visualizations-
    - A bar chart that displays the explained variance ratio of each principal component, highlighting the importance of the first few components.
    - A line plot that shows the cumulative explained variance, allowing us to determine the number of components needed to capture a desired percentage of the total variance, 95%.

The following are the results of PCA-

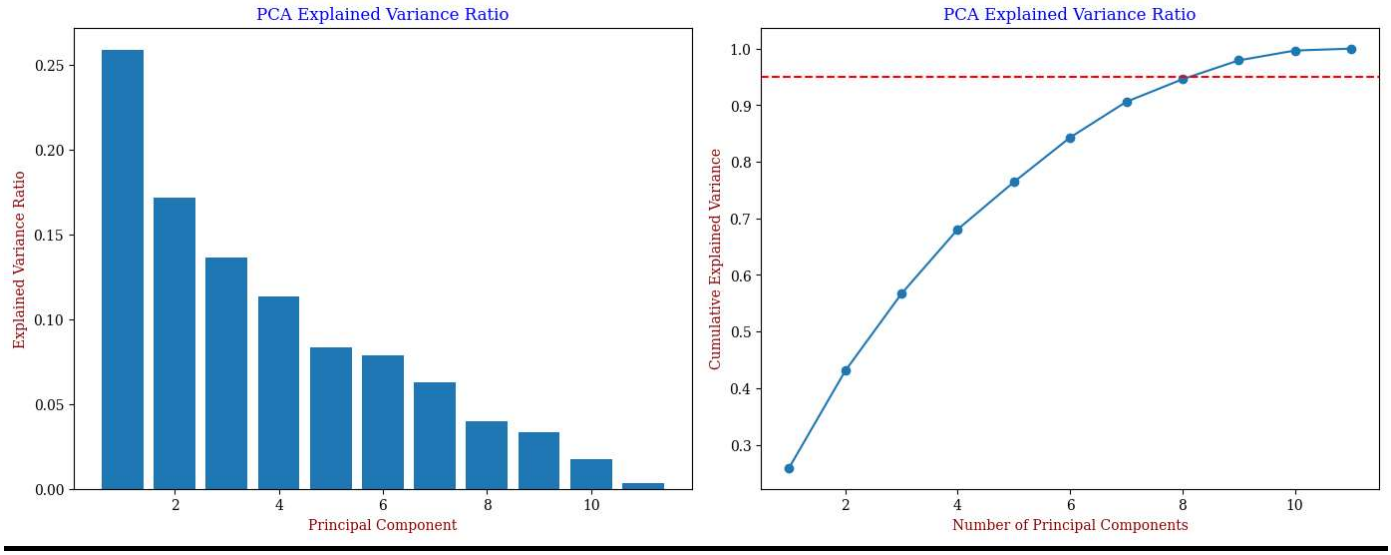


Fig5. PCA Analysis Graphs

Condition Number: 1657.52

The first few components of our PCA analysis highlight the main patterns in our weather data. The first component explains about 25.91% of the overall variability in our data. It's followed by the second and third components, which account for 17.15% and 13.63% of the variability, respectively. Together, the first three components cover around 56.69% of the variation in our data. This suggests they capture the most important weather patterns that are important. After the 8th component, we can see a significant drop in explained variance. This means that while later components do capture some variation, they aren't as important for understanding the big picture of the weather data.

The condition number of 1657.52 is relatively high but not excessively so, particularly in the context of PCA where the scale can vary widely. It suggests some numerical stability issues, but it's not likely to cause significant problems for analysis. However, it should be kept in mind when interpreting results, especially when looking at the smaller components.

The initial principal components likely represent broad weather patterns affecting multiple variables concurrently, such as a hot and dry weather trend influencing temperature and humidity. Later components may capture finer, localized weather phenomena, reflecting the complexity and noise inherent in weather data.

NORMALITY TESTING

One of the most common assumptions for statistical tests is that the data used are normally distributed. Normality tests assess whether a variable follows a normal distribution, which is essential for many statistical analyses. In this case, examining the normality of key weather variables: temperature, humidity, and wind speed.

Here I Conducted three commonly used normality tests: Shapiro-Wilk, Kolmogorov-Smirnov, and D'Agostino's K^2 test.

For each variable, I randomly sampled a subset of data (up to 1000 observations) to ensure computational efficiency while maintaining statistical accuracy. A p-value greater than 0.05 indicates that the variable is normally distributed at a significance level of 5%. Conversely, a p-value less than 0.05 suggests a departure from normality. If all three tests indicate normality ($p > 0.05$), then the variable is approximately normally distributed.

These are the results for the tests –

```
Normality Test Results for Temperature:
Shapiro-Wilk Test: Stat=0.9728066278295141, p=9.762768668233248e-40, Normal=False
Kolmogorov-Smirnov Test: Stat=0.987134186699616, p=0.0, Normal=False
D'Agostino's K^2 Test: Stat=67.53694432091359, p=2.160423429054276e-15, Normal=False

Normality Test Results for Humidity:
Shapiro-Wilk Test: Stat=0.973864014567081, p=3.948928094270841e-39, Normal=False
Kolmogorov-Smirnov Test: Stat=1.0, p=0.0, Normal=False
D'Agostino's K^2 Test: Stat=174.13223240567498, p=1.5405158547252877e-38, Normal=False

Normality Test Results for Wind:
Shapiro-Wilk Test: Stat=0.9179148739706864, p=1.335041050257918e-58, Normal=False
Kolmogorov-Smirnov Test: Stat=0.9860965524865014, p=0.0, Normal=False
D'Agostino's K^2 Test: Stat=111.78240874166343, p=5.330382259142983e-25, Normal=False
```

Fig6. Normality test results

Temperature: All three normality tests (Shapiro-Wilk, Kolmogorov-Smirnov, and D'Agostino's K^2) indicate that the temperature data is not normally distributed ($p < 0.05$).

Humidity: Similarly, for humidity, all three tests reject the assumption of normality ($p < 0.05$).

Wind: The wind speed data also shows a lack of normal distribution according to all three tests ($p < 0.05$).

These findings suggest that the assumptions of normality are violated for all three variables. Also, it is quite common for weather data to not follow a normal distribution. Weather variables often exhibit complex patterns influenced by various factors such as geographical location, time of day, seasonality, and local phenomena. Depending on the analysis objectives and assumptions of the statistical tests being used, it may be necessary to apply appropriate transformations or non-parametric methods to ensure the validity of the results.

DATA TRANSFORMATION

Using the transformation on weather data is not advisable as original weather trends cannot be interpreted. I have tried to perform transformation on the 3 parameters – temperature, humidity, and wind. Boxcox on temperature and humidity while log transform on wind.

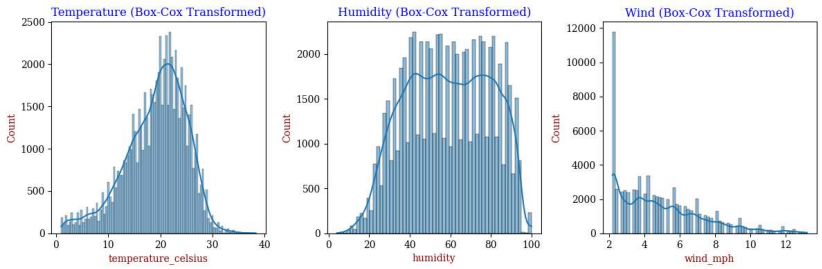


Fig7. Before transformation

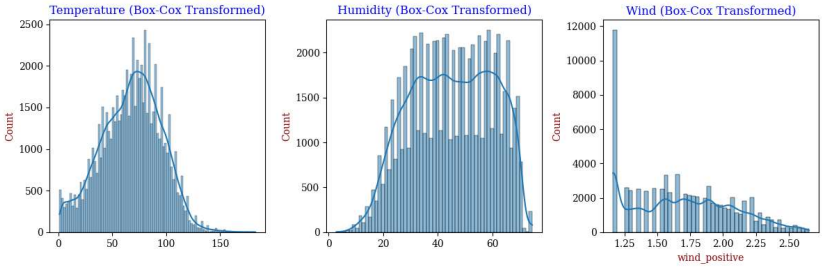


Fig8. After transformation

Before Transformation the temperature appears a little normal with a potential slight right skew, indicating a few days of very high temperatures.. For humidity, the histogram shows a multimodal distribution, with several peaks. It could suggest different weather patterns, like distinct dry and wet seasons, or varying regional climates within the dataset. Wind speed is heavily right-skewed, implying that low wind speeds are far more common than high wind speeds.

After Transformation for temperature, it has stretched the distribution, making it more symmetric and normal-like. In humidity, the multimodal nature is still present, but the peaks appear smoother and less pronounced. For wind the transformation has significantly changed the scale of wind speed, and it remains skewed.

So the usefulness of the transformation is mainly used if any statistical technique has to be applied that assumes normality, else this transformation will obscure some of the natural variability seen in the original distribution.

Ultimately, whether the transformation is useful depends on the context of the analysis. If interpreting the original weather trends is the goal as it is in this project, preserving the original distribution may be preferable.

HEATMAPS AND CORRELLATION

Numerical columns from the data are considered and a Pearson correlation heatmap is plotted and a table is created to show the top correlated values. The heatmap is as shown below-

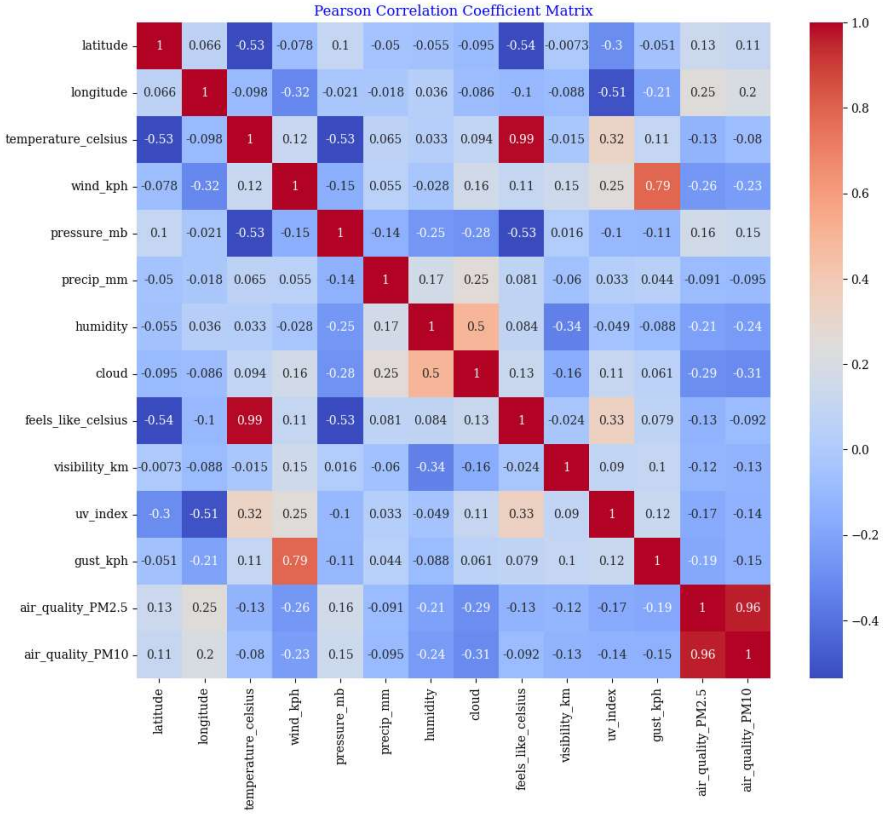


Fig9. Correlation Heatmap

Variable 1	Variable 2	Correlation	Abs Correlation
feels_like_celsius	temperature_celsius	0.99	0.99
temperature_celsius	feels_like_celsius	0.99	0.99
air_quality_PM10	air_quality_PM2.5	0.96	0.96
air_quality_PM2.5	air_quality_PM10	0.96	0.96
wind_kph	gust_kph	0.79	0.79
gust_kph	wind_kph	0.79	0.79
feels_like_celsius	latitude	-0.54	0.54
latitude	feels_like_celsius	-0.54	0.54
latitude	temperature_celsius	-0.53	0.53
temperature_celsius	latitude	-0.53	0.53
pressure_mb	temperature_celsius	-0.53	0.53
temperature_celsius	pressure_mb	-0.53	0.53
pressure_mb	feels_like_celsius	-0.53	0.53
feels_like_celsius	pressure_mb	-0.53	0.53
longitude	uv_index	-0.51	0.51
uv_index	longitude	-0.51	0.51
cloud	humidity	0.50	0.50
humidity	cloud	0.50	0.50
humidity	visibility_km	-0.34	0.34
visibility_km	humidity	-0.34	0.34
uv_index	feels_like_celsius	0.33	0.33
feels_like_celsius	uv_index	0.33	0.33
uv_index	temperature_celsius	0.32	0.32
temperature_celsius	uv_index	0.32	0.32
wind_kph	air_quality_PM2.5	-0.26	0.26
humidity	pressure_mb	-0.25	0.25
pressure_mb	humidity	-0.25	0.25
cloud	precip_mm	0.25	0.25
precip_mm	cloud	0.25	0.25

Fig10. Top correlation table

From the heatmap, there is a strong positive correlation between the actual temperature (temperature_celsius) and the "feels like" temperature (feels_like_celsius), which is expected, as the "feels like" temperature is calculated based on the actual temperature, taking into account factors like humidity and wind chill.

The air quality variables for PM10 and PM2.5 show a strong positive correlation, this also makes sense because PM2.5 is a subset of PM10, so the two measurements are closely related. The average wind speed (wind_kph) and wind gusts (gust_kph) exhibit a substantial positive correlation.

Latitude has a negative correlation with both temperature_celsius and feels_like_celsius. This indicates that as you move away from the equator (towards the poles), the temperatures tend to decrease, reflecting the typical latitudinal temperature patterns. Latitude also has a moderate negative correlation with the UV index (uv_index) which implies that UV levels tend to be higher closer to the equator, likely due to the angle of the sun's rays.

Atmospheric pressure (pressure_mb) is negatively correlated with temperature_celsius and feels_like_celsius and aligns with the general understanding that temperature often drops when pressure increases, and vice versa.

Visibility (visibility_km) shows a moderate negative correlation with humidity and cloud cover which suggests that higher humidity levels and more cloud cover are associated with lower visibility, which is consistent with meteorological principles.

STATISTICS

	latitude	longitude	last_updated	\
count	137950.00	137950.00	137950	
mean	22.11	64.96	2023-12-02 12:58:10.166727168	
min	-41.30	-175.20	2023-08-29 02:45:00	
25%	17.25	72.93	2023-10-14 02:30:00	
50%	23.70	77.33	2023-12-04 03:30:00	
75%	27.63	83.05	2024-01-19 23:45:00	
max	64.10	179.22	2024-03-06 06:00:00	
std	13.59	42.57	NaN	

	temperature_celsius	temperature_fahrenheit	wind_mph	wind_kph	\
count	137950.00	137950.00	137950.00	137950.00	
mean	19.39	66.89	5.49	8.83	
min	-41.90	-43.40	2.20	3.60	
25%	15.30	59.50	2.70	4.30	
50%	20.80	69.40	4.30	6.80	
75%	24.90	76.70	6.90	11.20	
max	45.40	113.70	91.90	148.00	
std	7.66	13.80	3.68	5.93	

	wind_degree	pressure_mb	pressure_in	...	\
count	137950.00	137950.00	137950.00	...	
mean	154.95	1013.27	29.92	...	
min	1.00	958.00	28.29	...	
25%	55.00	1010.00	29.83	...	
50%	123.00	1014.00	29.94	...	
75%	263.00	1017.00	30.02	...	
max	360.00	1074.00	31.71	...	
std	112.93	5.82	0.17	...	

	air_quality_Carbon_Monoxide	air_quality_Ozone	\
count	137950.00	137950.00	
mean	747.95	40.35	
min	96.80	0.00	
25%	333.80	17.20	
50%	520.70	36.10	
75%	867.80	58.70	
max	41870.10	555.00	
std	937.15	29.76	

	air_quality_Nitrogen_dioxide	air_quality_Sulphur_dioxide	\
count	137950.00	137950.00	
mean	14.89	6.86	
min	0.00	0.00	
25%	4.00	1.20	
50%	9.20	3.00	
75%	19.00	6.50	
max	575.80	557.00	
std	19.46	16.37	

	air_quality_PM2.5	air_quality_PM10	air_quality_us-epa-index	\
count	137950.00	137950.00	137950.00	
mean	80.15	98.21	2.91	
min	0.50	0.50	1.00	
25%	14.10	20.20	1.00	
50%	47.60	62.70	3.00	
75%	106.10	129.70	4.00	
max	1558.80	3566.40	6.00	
std	98.54	117.05	1.56	

	air_quality_gb-defra-index	moon_illumination	month
count	137950.00	137950.00	137950.00
mean	5.72	52.48	7.29
min	1.00	0.00	1.00
25%	2.00	18.00	2.00
50%	5.00	53.00	9.00
75%	10.00	88.00	11.00
max	10.00	100.00	12.00
std	3.84	34.91	4.34

Fig11. Statistics of data

The dataset encompasses a wide range of temperature values, from -41.9°C to 45.4°C, with a mean temperature of approximately 19.39°C. The standard deviation of 7.66°C indicates moderate variability in temperature across observations. Wind speed exhibits significant variability, ranging from 3.6 kph to 148.0 kph, with a mean speed of 8.83 kph. The standard deviation of 5.93 kph suggests considerable dispersion in wind speeds across observations. Air pressure measurements range from 958 mb to 1074 mb, with an average pressure of approximately 1013.27 mb.

We have varying values of air pollutants with the highest being Carbon monoxide, followed by PM2.5 and PM10. Followed by the moon illumination percentage. The Multivariate KDE sampled at a size of 1% is shown below which echoes the correlation that is shown in the heatmap as well.

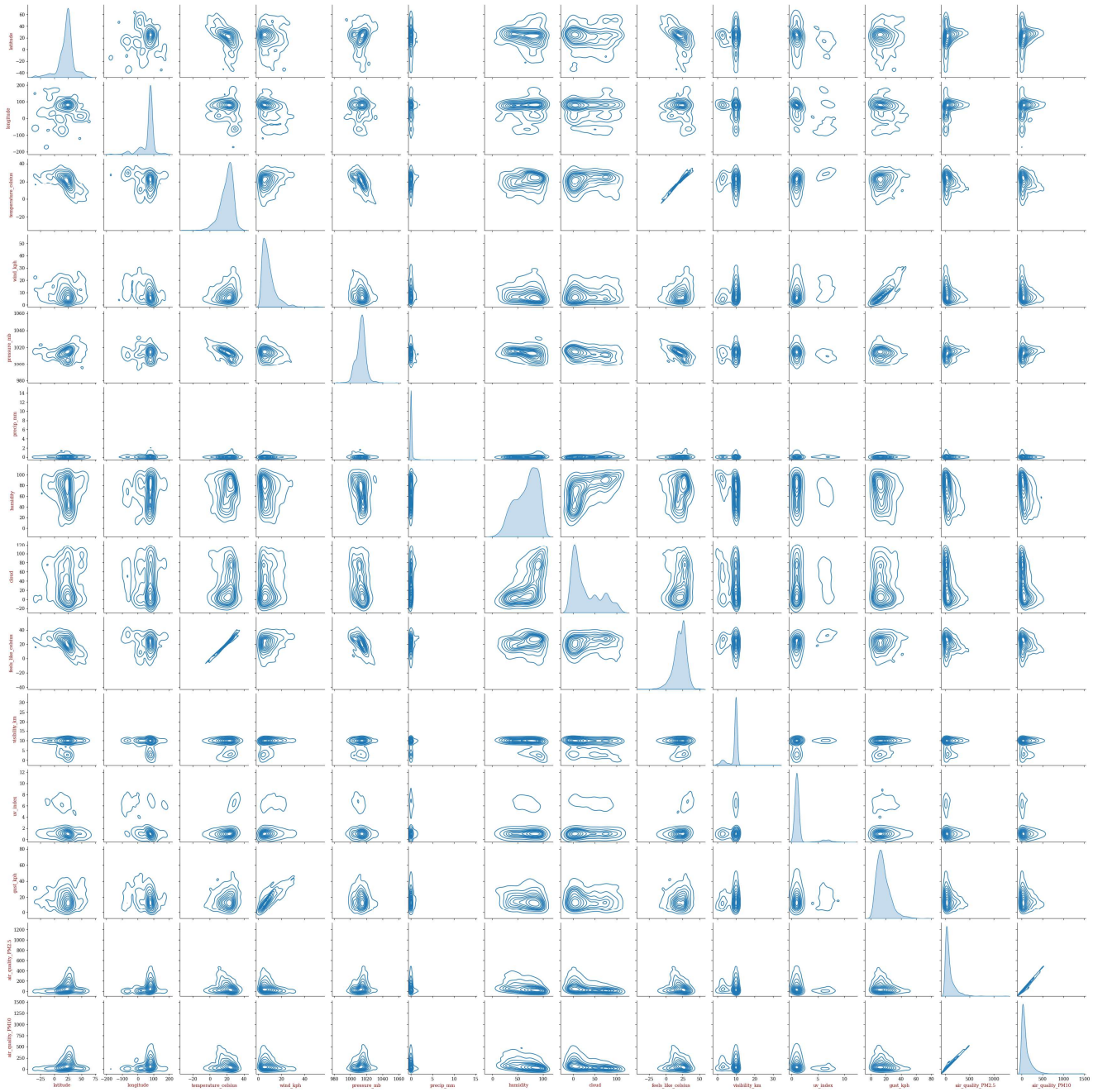


Fig 12. Multivariate KDE

STATIC PLOTS

1. Line Plot-

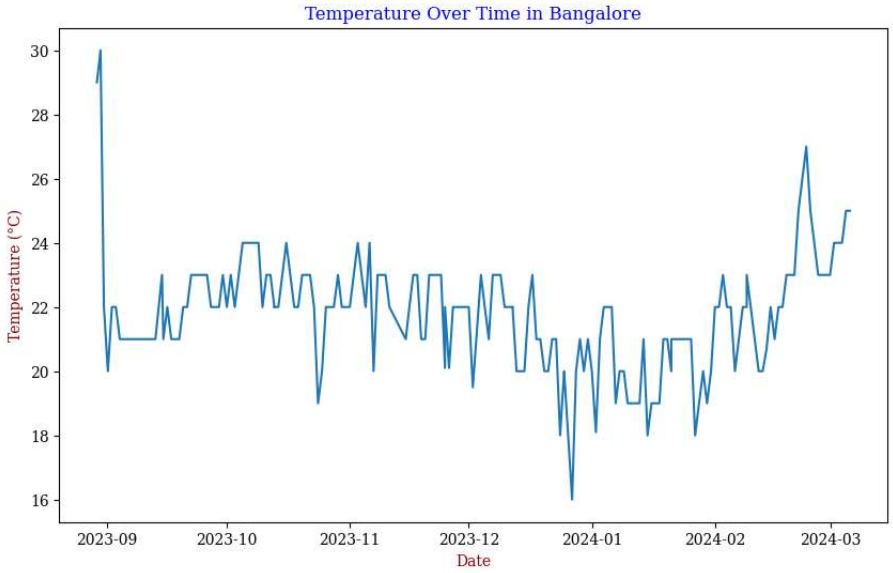


Fig 13. Line Plot

The plot is a time series graph showing the temperature in Bangalore over several months, from September 2023 to March 2024. The temperature varies between approximately 16°C and 30°C. The graph shows an overall increase in temperature over time. Starting in September 2023, temperatures are higher, then they dip — likely reflecting the cooler months — and from around January 2024, there's a significant rise, indicating the onset of warmer weather. This pattern could be indicative of seasonal changes from autumn to winter, followed by spring. There's a notable spike at the beginning of the data in September and a sharp rise towards the end in March. These could be due to sudden weather changes or specific heatwave events. The graph is useful for understanding the climate pattern and temperature trends in Bangalore.

2. Bar Plot-

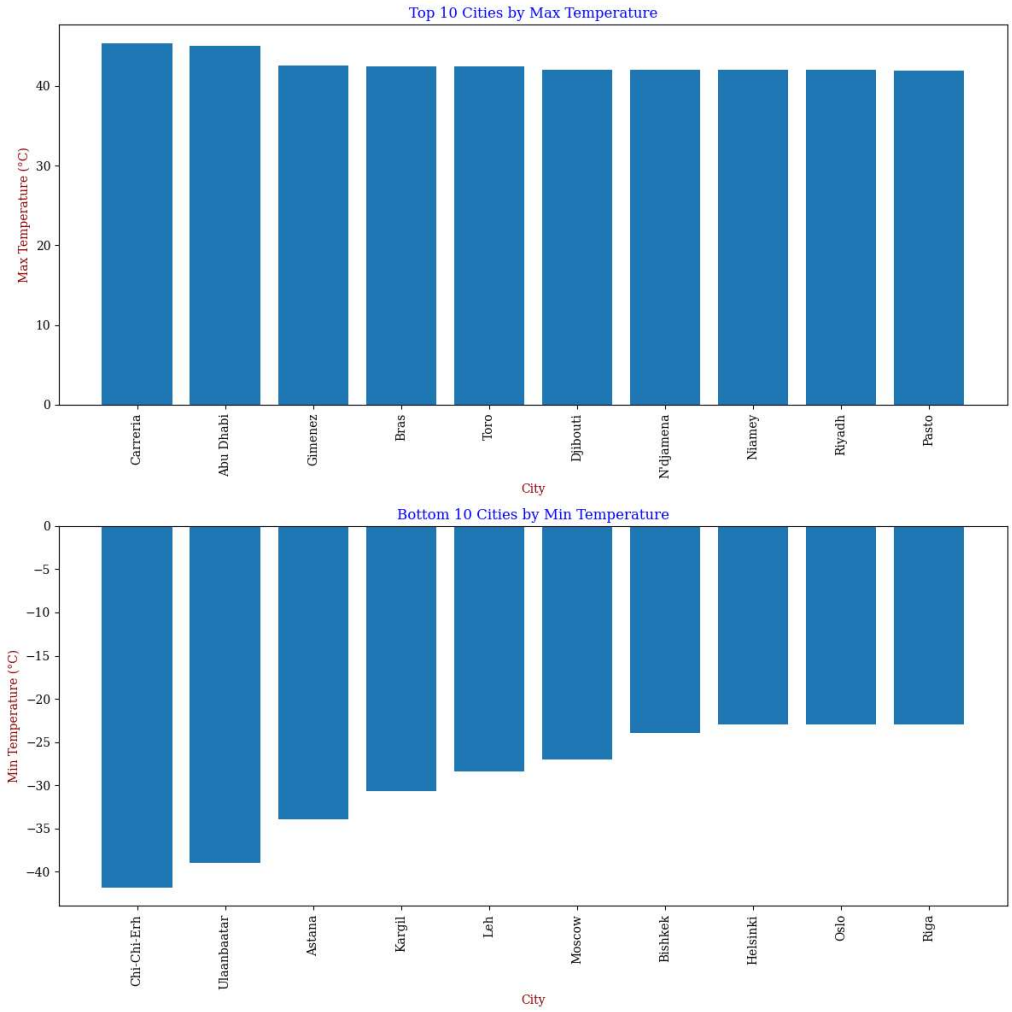


Fig 14. Bar plot 1

This bar chart compares the highest and lowest temperatures ever recorded in different cities, in the dataset.

In Top Cities, shows which cities have experienced the highest temperatures. These places might need more cooling resources and could be at risk for heatwaves.

Bottom Cities shows cities with the lowest temperatures. They might need more heating resources and could face extreme cold weather. These plots are useful in knowing the climatic condition of a particular city and what is peak climate that the city can experience and hence make arrangements according to it. Tourism planners can also know in advance the location's extreme climate and avoid that particular time frame for visits.

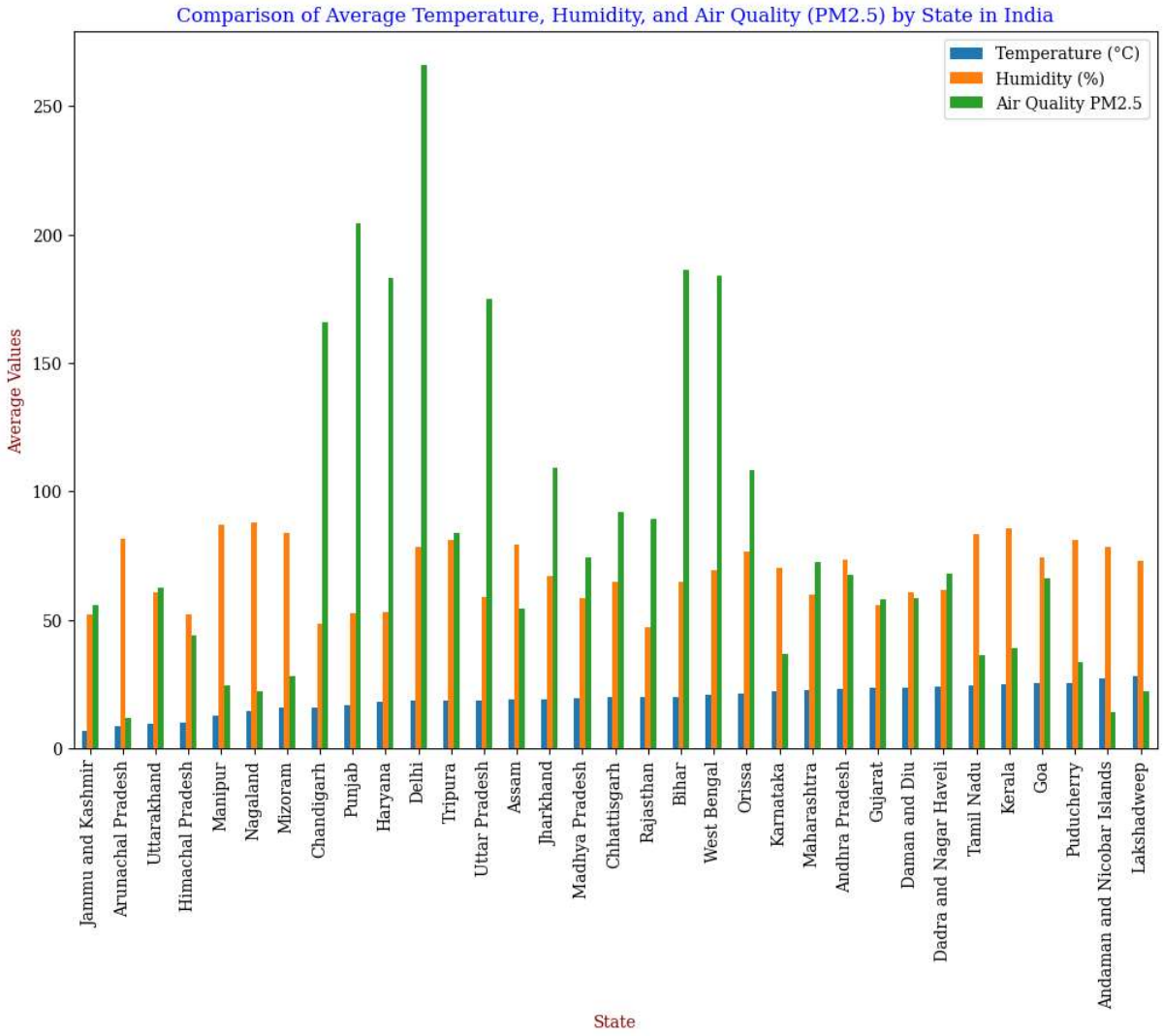


Fig 14. Group Bar plot

This grouped bar chart compares average temperature, humidity, and PM2.5 air quality levels for various states in India. Temperature bars represent the average temperature for each state, useful for identifying warmer and cooler regions. The humidity bars indicate the moisture content in the air, which has implications for comfort, health, and agriculture. Air Quality PM2.5 reflects the average concentration of particulate matter with a diameter of less than 2.5 micrometers, which is an important indicator of air pollution that affects respiratory health. The use of this plot is in identifying regions with poor air quality and implementing mitigation strategies, to provide warnings and advice based on the air quality index and humidity levels, advising travelers about the best time to visit different states based on favorable weather conditions.

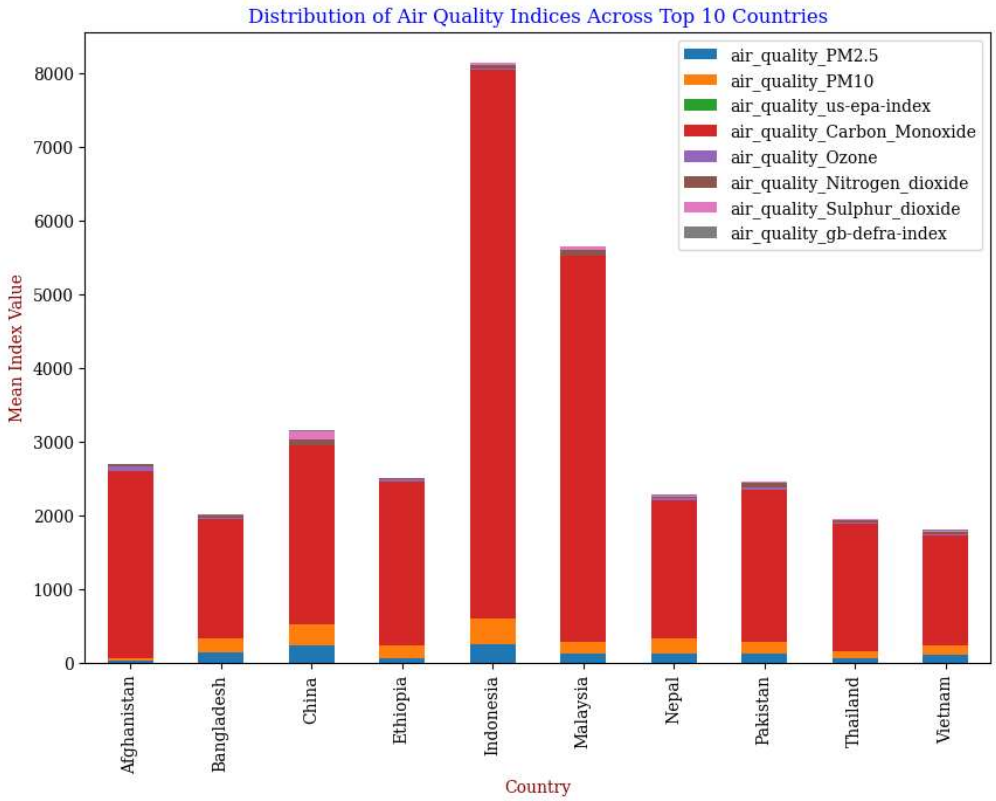


Fig 15. Stack Bar plot

This stacked bar chart shows air quality distribution across the top ten polluted countries. Each bar represents a country and is divided into segments representing different air quality indices. The height of each segment within the bar indicates the average level of a specific air pollutant in that country. We can see that CO is the most abundant pollutant across countries followed by PM2.5 and PM10. This plot can be used to know which is the pollutant that has to be controlled and reduced so as to improve the air quality.

3. Count plot

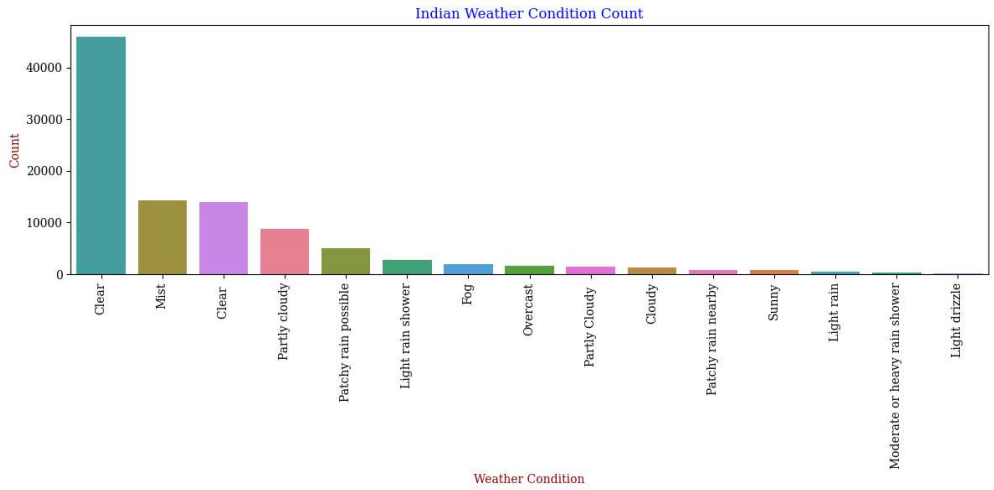


Fig 16. Count plot

This bar chart shows the frequency of different weather conditions in India, with each bar representing a specific weather condition and the height indicating the count of occurrences. The most frequent weather condition is ‘Clear’, significantly higher than any other, indicating that clear skies are common in the dataset's time frame. ‘Mist’ and ‘Fog’ follow showing that they are also common conditions, reflecting the climatic patterns that can affect visibility and air traffic, among other things. The presence of different types of cloudy weather ‘partly cloudy’, and ‘overcast’ suggests variability in cloud cover and potential for precipitation. These kind of plots can be used in public planning by preparing for the most common weather conditions. While the remaining sectors can also plan as per typical weather conditions.

4. Pie Chart

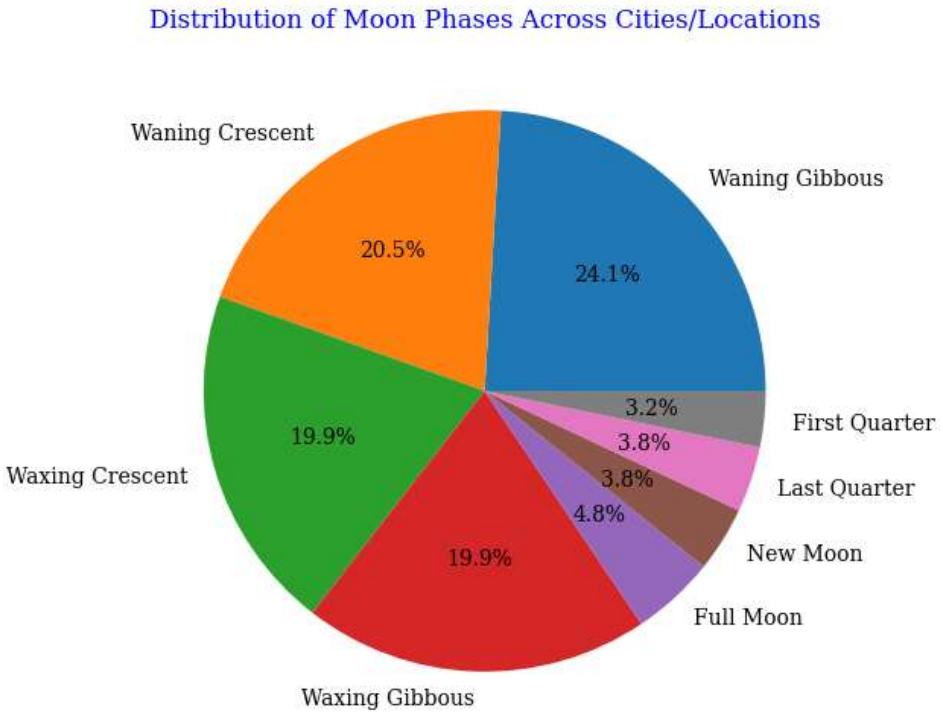


Fig 17. Pie chart

The pie chart provided appears to represent the distribution of moon phases across various cities or locations. Waning Gibbous was represented by the largest segment at 24.1%, suggesting that this moon phase occurred most frequently during the observed period. Waning Crescent is the second-largest segment at 20.5%, indicating that this phase was also common. Waxing Gibbous and Waxing Crescent are both nearly equal at 19.9%, showing that these phases occurred with similar frequency. Full Moon and New Moon would typically be expected to have a relatively low occurrence rate because these phases are transitional and occur for a shorter duration than the crescent and gibbous phases. The first Quarter and Last Quarter are represented by smaller segments at 3.2% and 3.8%, respectively, indicating less frequency. This is consistent with the fact that these phases are also transitional like the full and new moons.

This type of chart could be useful for understanding the prevalence of certain moon phases over a particular timespan, which could have various applications such as planning astronomical events, understanding tidal patterns, or even for cultural or agricultural purposes where moon phases play a role.

5. Dist Plot

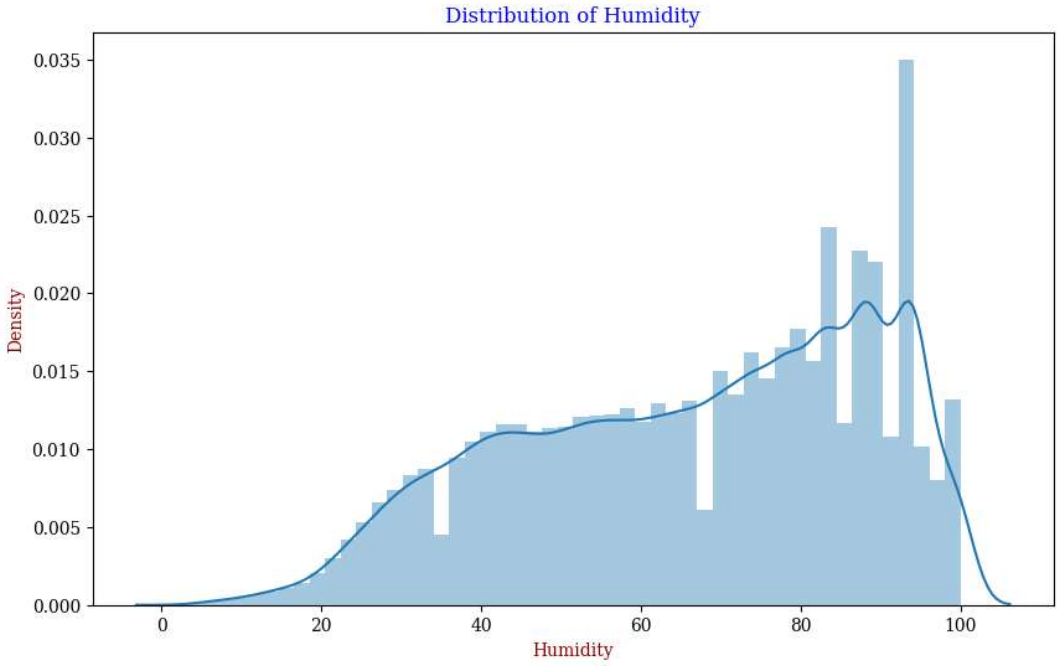


Fig 18. Dist plot

This is a dist plot overlaid with a kernel density estimate (KDE) that depicts the distribution of humidity measurements. The bars show the frequency distribution of humidity data points. The x-axis represents the humidity levels, y-axis represents the density or count of observations within each bin of humidity levels. The distribution seems to be right-skewed, as the tail of the histogram extends further to the right, indicating that there are a number of days with very high humidity levels. he data has a higher frequency of days with moderate to high humidity levels.

There are fewer days with very low or very high humidity levels, as indicated by the tails of the distribution. where most days have a substantial amount of moisture in the air, but occasionally there are extremely humid days, possibly due to seasonal weather patterns or specific weather events. This data could be important for various applications, such as agriculture, where humidity levels can affect crop growth, or urban planning, where it might impact building designs and energy consumption.

6. Pair Plot

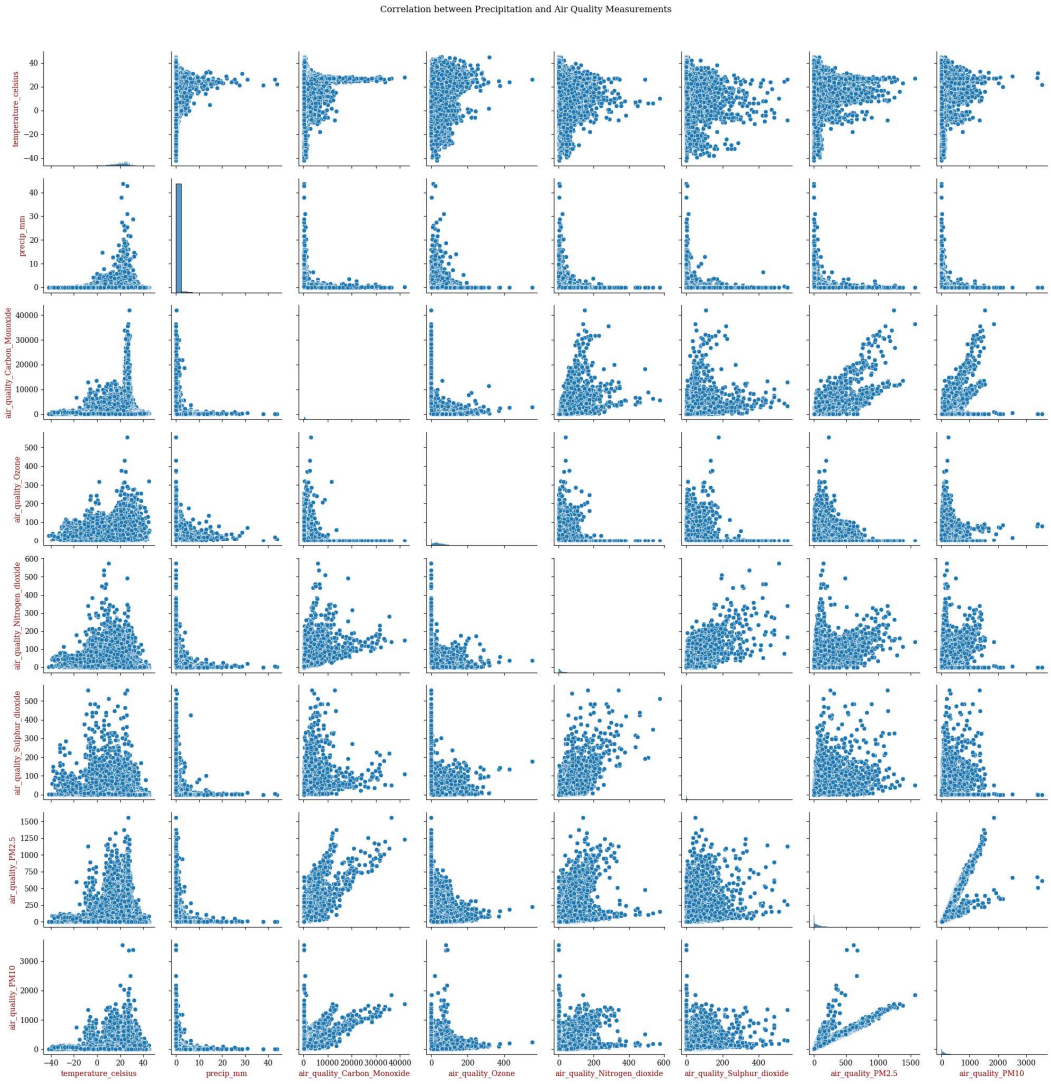


Fig 19. Pair plot

This is a pair plot between temperature, precipitation, and various air quality index values. Here we can see that for higher temperatures the air quality index values are also higher and for precipitation, we can see that they become lower, which means air quality improves when it rains or snows. The remaining pair plots are with each other air quality index and they are usually follow a linear pattern, as any increase in either of them increases the other

7. Histogram with KDE

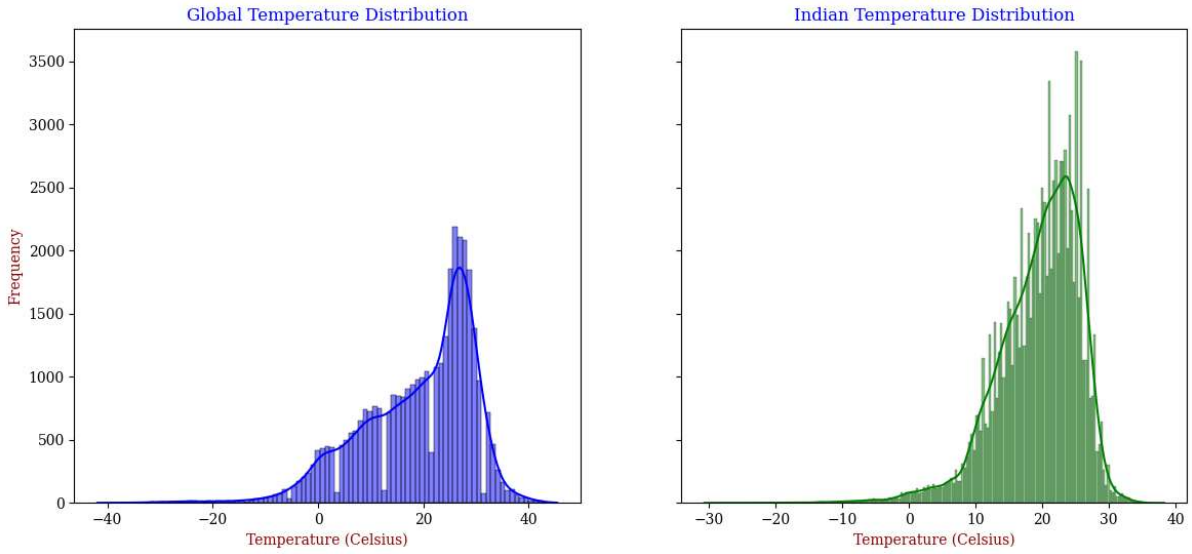


Fig 20. Hist with KDE

These histograms show how temperatures are spread out across the globe and specifically in India.

Global Histogram -the blue histogram shows that temperatures worldwide are spread out, ranging from very cold to very hot. The curve in the middle suggests that most places have moderate temperatures around 20°C.

Indian Histogram - the green histogram indicates that in India, temperatures are generally warmer. There's less variation in temperature compared to the global data, and higher temperatures are more common.

These distributions are crucial for climate scientists and meteorologists studying regional and global climate patterns. In agricultural planning, understanding temperature distributions helps in crop selection and management practices based on regional climate conditions. Can also be used in energy management as it helps in anticipating energy demands for heating or cooling based on prevailing temperature trends.

8. QQ plot

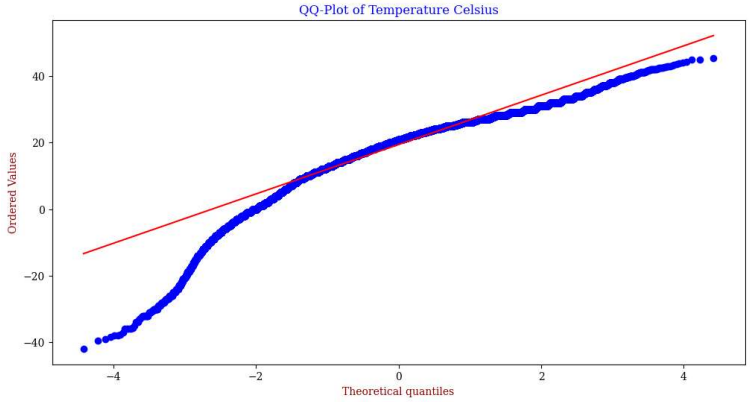


Fig 21. QQ plot

This is a QQ-plot,, for temperature data measured in Celsius. As we know QQ-plot is a graph to assess whether a set of data plausibly came from some theoretical distribution such as a Normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution. The x-axis displays the theoretical quantiles, meaning it does not show actual data, but instead the position a data point would be in if it were in a perfect distribution. The y-axis shows the ordered values from the actual data. In this plot, if the points form a line that closely follows the red reference line, this indicates that the data follows the theoretical distribution. The red line represents where the points would lie if the temperature data were perfectly normally distributed.

The majority of points in the middle of the distribution line up well with the red line, which suggests that temperatures in that range are normally distributed. The lower tail of the plot shows points that deviate below the line, indicating that the lowest temperatures in the dataset are lower than what would be expected in a normal distribution. The upper tail also shows a deviation, with data points lying above the red line, which means the highest temperatures are higher than what would be expected in a normal distribution, or there are more extremely high values than expected. This kind of pattern might reflect occasional extreme weather events that lead to unusually high or low temperatures.

9. KDE plot with Fill

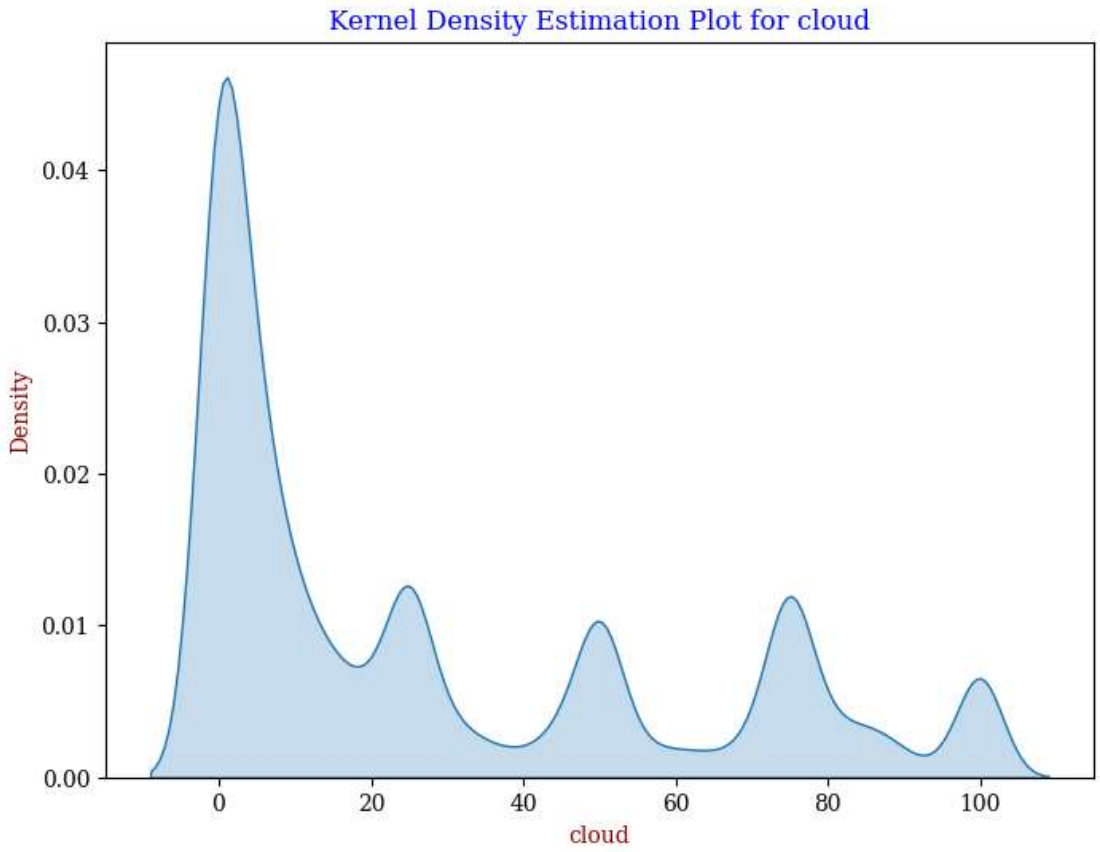


Fig 22. KDE with fill

This is a KDE plot for cloud cover. From this plot, we can observe several peaks of different heights, suggesting that there are several common values for cloud coverage. The highest peak, which is close to zero, suggests that there are a significant numThere are additional peaks at intervals, which could correspond to typical cloud coverage values that are frequently observed. These may represent partly cloudy skies, mostly cloudy skies, overcast conditions, etc.ber of observations with very low cloud coverage, indicating clear skies.he pattern of peaks and valleys suggests that cloud coverage is not evenly distributed; instead, it has several modes. This indicates that the cloud coverage doesn't gradually increase or decrease but instead has preferred states or common conditions.The area under the curve of a KDE plot represents the probability distribution of the data. Where the curve is higher, there is a higher likelihood of observing those values of cloud coverage.

10. Reg plot

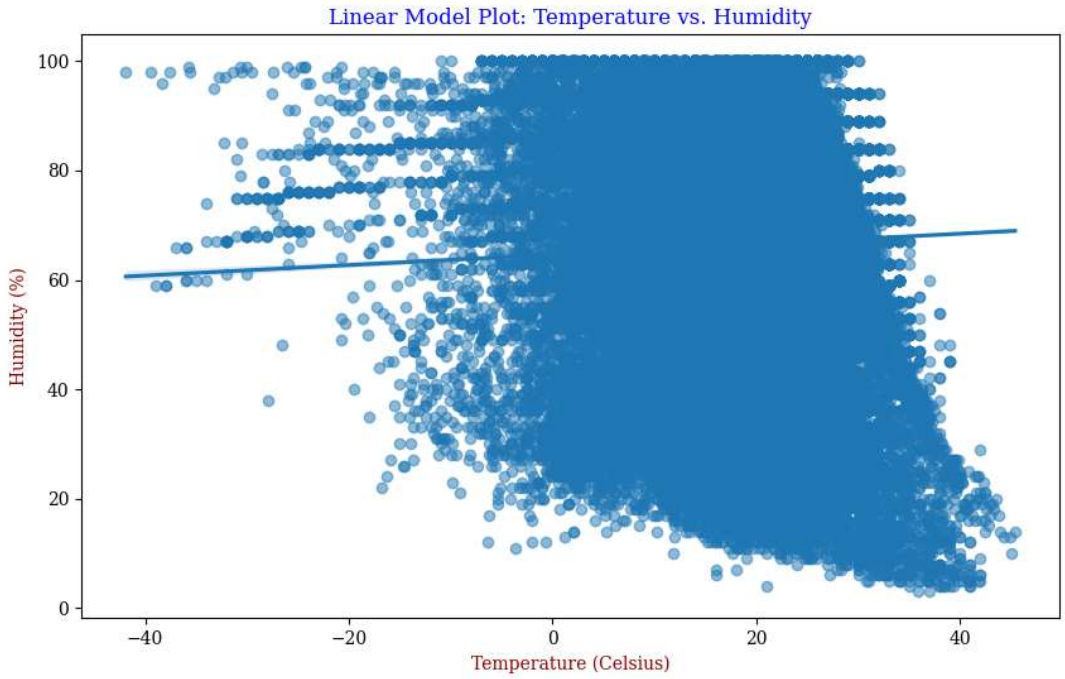


Fig 23. Reg plot

The plot is a scatter plot with a linear model fit, displaying the relationship between temperature on the x-axis and humidity on the y-axis. The temperature values appear to range from below freezing (less than 0°C) to around 40°C, which is typical for a varied climate that experiences both winter and summer seasons. The humidity percentages are spread between 0% and 100%, which is the full possible range for humidity. Lower humidity levels can be seen across the range of temperatures, while higher humidity levels are densely packed at the lower temperatures. The linear model fit line (the straight line through the data points) is relatively flat, which suggests there is not a strong linear relationship between temperature and humidity in this dataset.

11. Multivariate Box plot

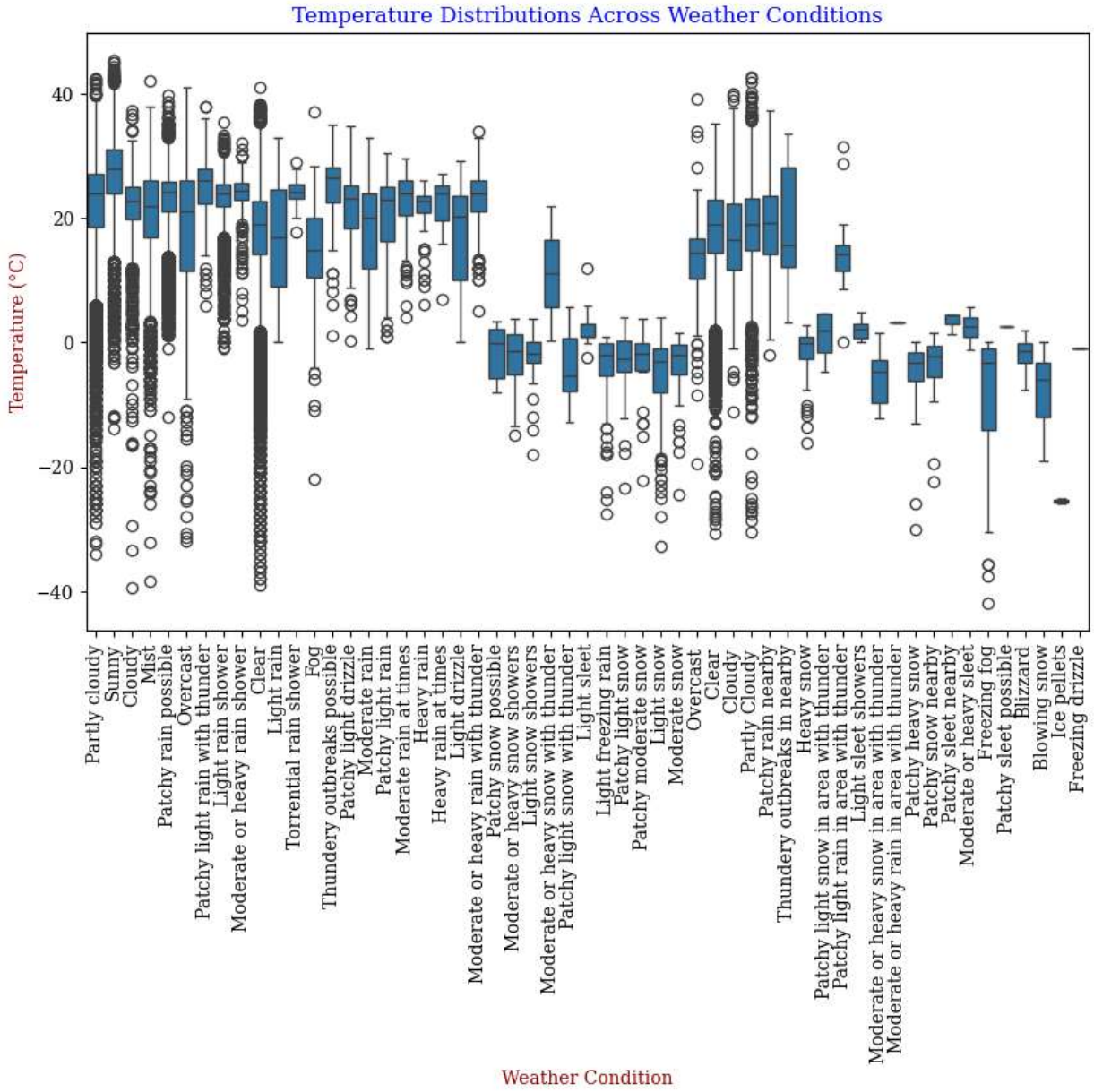


Fig 24. Multivariate Box plot

This is a multivariate box plot showing temperature distributions for different weather conditions. As we can see cloudy, sunny, partly cloudy, clear, etc have higher temperatures, but narrower distribution while others like light rain, fog, thunder mixed rain have a larger distribution of temperatures. We can also observe a lot of outliers in this plot, which may account for sudden climatic change or any disturbance such as a natural disaster or hurricanes etc.

12. Area Plot

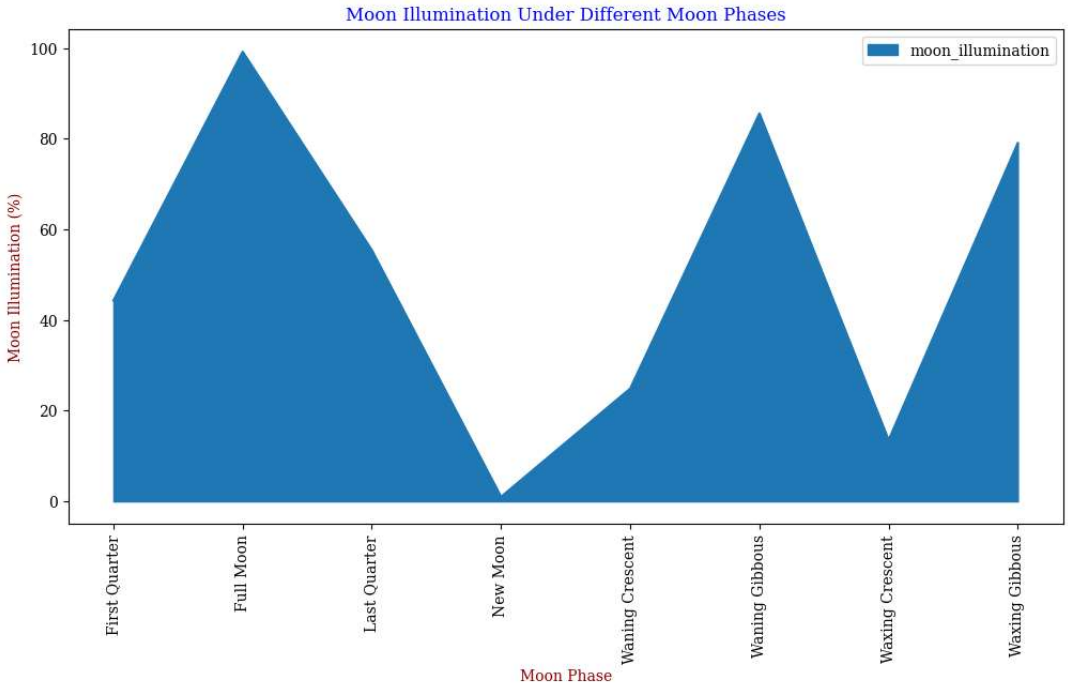


Fig 25. Area plot

This area plot displays the percentage of the Moon illuminated during different phases, such as the new moon, first quarter, full moon, and so on. This data is Crucial for astronomers and navigators for celestial navigation, planning nighttime activities where natural light is significant, and scheduling cultural events according to the lunar calendar.

13. Violin plot

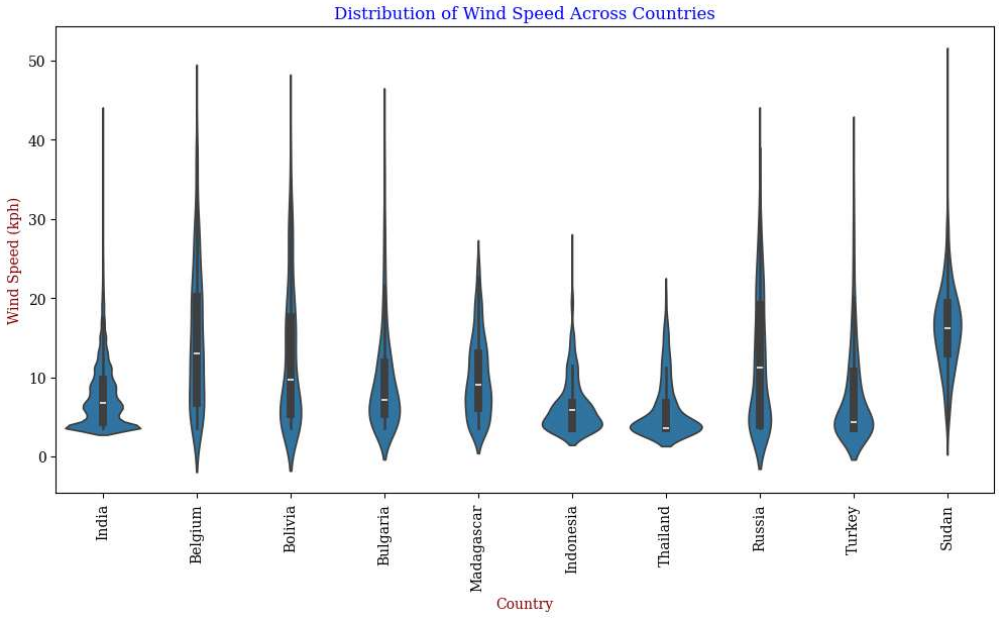


Fig 26. violin plot

This violin chart shows wind speed across different countries. The width of each violin represents the distribution density of wind speeds in a particular country. A wider violin suggests greater variability in wind speeds, while a narrower violin indicates more consistent wind conditions. Countries like Russia, Belgium show a wide distribution of wind speeds, indicating variable wind conditions.

14. Joint Plot

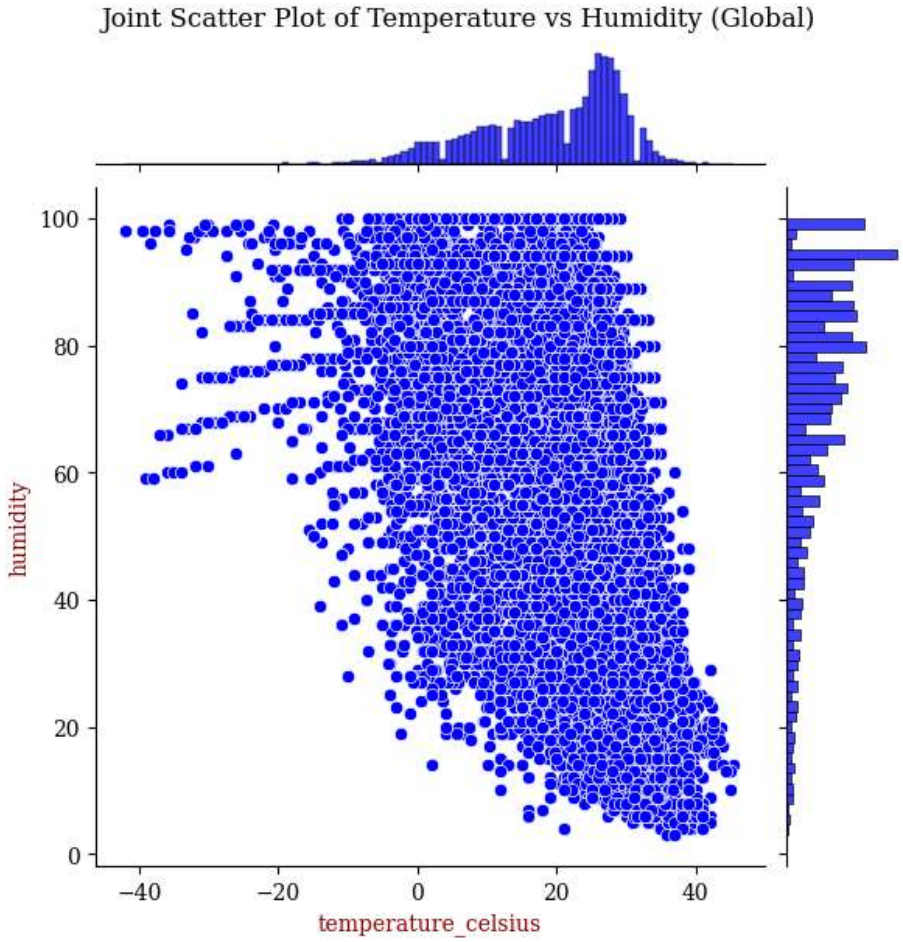


Fig 27. Joint plot

This joint plot shows the relationship between temperature and humidity. Data points are scattered across a wide range of humidity levels at various temperatures. There is no clear linear pattern suggesting that temperature alone does not determine humidity levels. The top histogram represents the distribution of temperature and it appears unimodal with a peak around 20°C, indicating common temperature ranges.

While the right histogram shows the distribution of humidity levels spread across a range but potentially concentrated at higher humidity levels. The lack of a clear linear relationship indicates that factors beyond temperature influence humidity, also humidity levels vary across all temperature ranges, suggesting diverse atmospheric conditions. As temperature increases the humidity decreases and this can be seen slightly.

15. Rug Pot

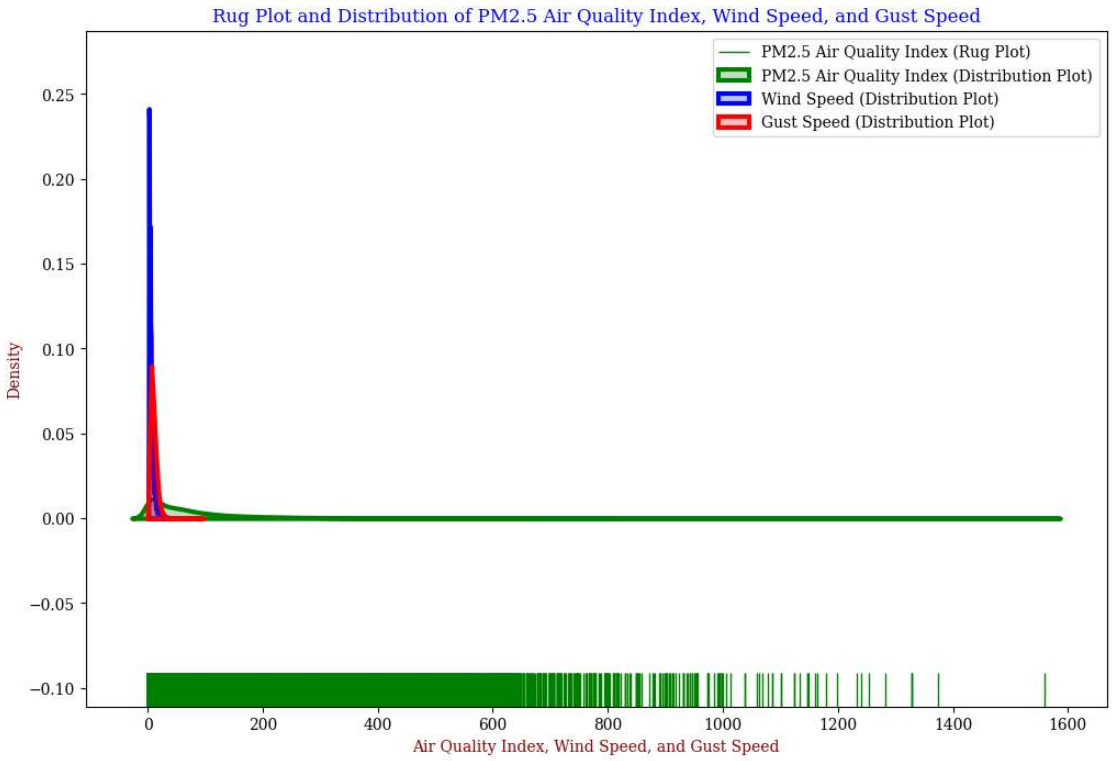


Fig 28. rug plot

This is a rug plot, with KDE. Green lines in the bottom part represent individual data points of the air quality. The concentration of lines indicates frequency at specific values. We are able to see a cluster at lower values, with fewer high values.

The upper graph is a kernel density estimate (KDE), where the green curve is the probability density of PM2.5 values with a peak at the lower end suggesting common occurrence of lower values, with occasional higher values.

The blue curve shows the wind speed density which is also relatively distributed with a single peak at lower values. The red curve shows the gust speed again concentrated at lower values.

A low air quality index suggests that it is positive for public health, but occasional poor air quality episodes occur. This plot can be used to track PM2.5 levels to assess air quality and implement pollution control measures. We can try to understand if air speed and gusts affect pollution and if so how can we use it to our advantage, as some winds are periodical and they become stagnant at a higher level, this is a common cause of air pollution in the north Indian regions as winds die out slowly during winter and mountains trap the pollution.

16. 3D plot

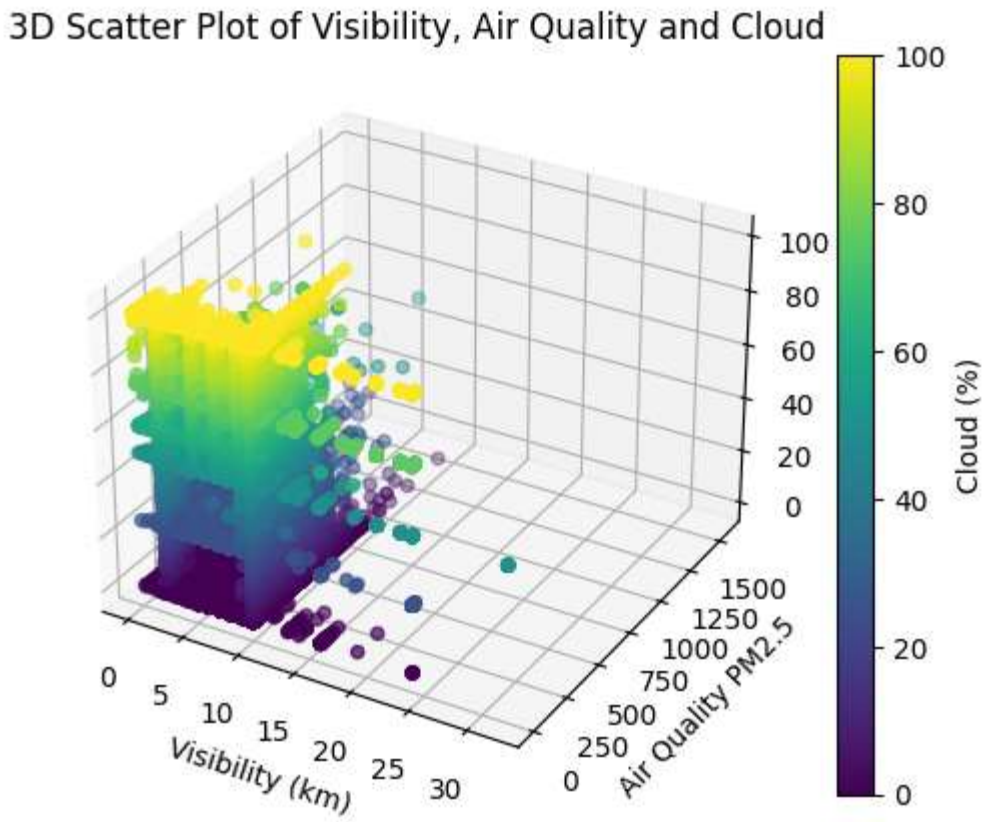


Fig 29. 3D scatter plot

This is a 3D Scatter plot of Visibility, Air Quality, and Cloud Cover. The hue is by the cloud cover and from this plot we can observe that the higher the air quality index, the lower the visibility, and cloud cover also does slightly affect the visibility but not at a higher level. The cloud cover does not affect the air quality to a higher extent as well as higher and lower values of cloud cover are found at lower and higher air quality values.

17. Hexbin plot

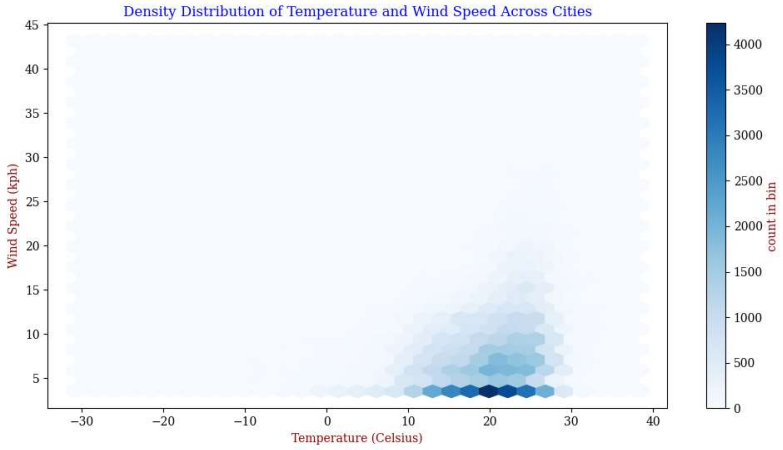


Fig 30. Hexbin plot

This Hexbin plot shows the concentration of temperature and wind speed data points across various cities. The darker areas indicate higher densities of data points, while lighter areas indicate lower densities. This plot can be used to understand the typical weather conditions experienced in different cities and regions. It provides insights into temperature and wind speed distributions, which are essential for various applications such as urban planning and event management. In this plot, we can see that there are very few high wind speed concentrations and that too slightly in higher temperatures.

18. Strip plot

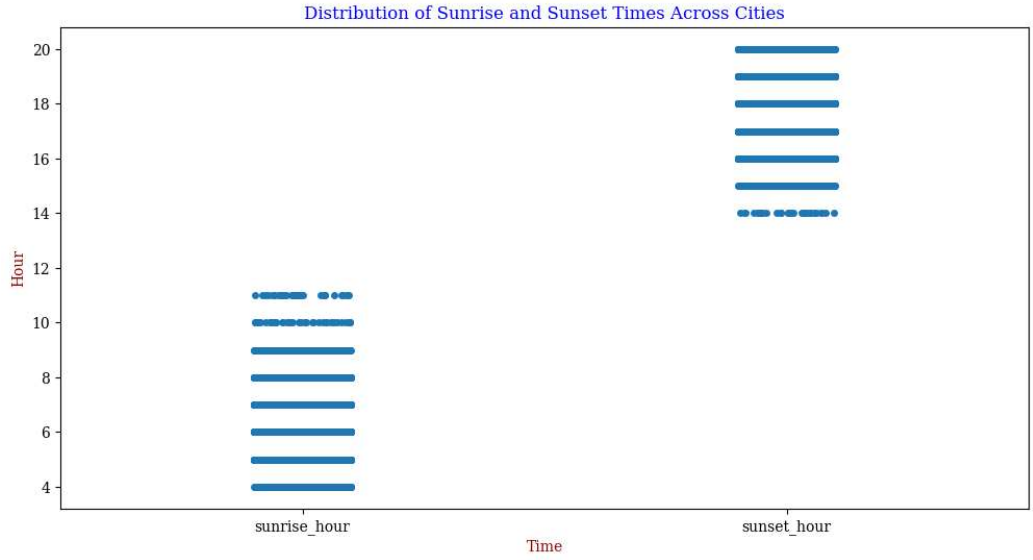


Fig 31. strip plot

This strip plot represents the range of sunrise and sunset times across different cities, with each horizontal bar indicating the times of these events. This plot is valuable for travel planning, Daylight schedule planning, photography enthusiasts seeking optimal lighting conditions, individuals observing religious practices tied to solar events, and researchers studying diurnal patterns in various regions. These plots are also useful in astronomy observations.

19. Subplot-

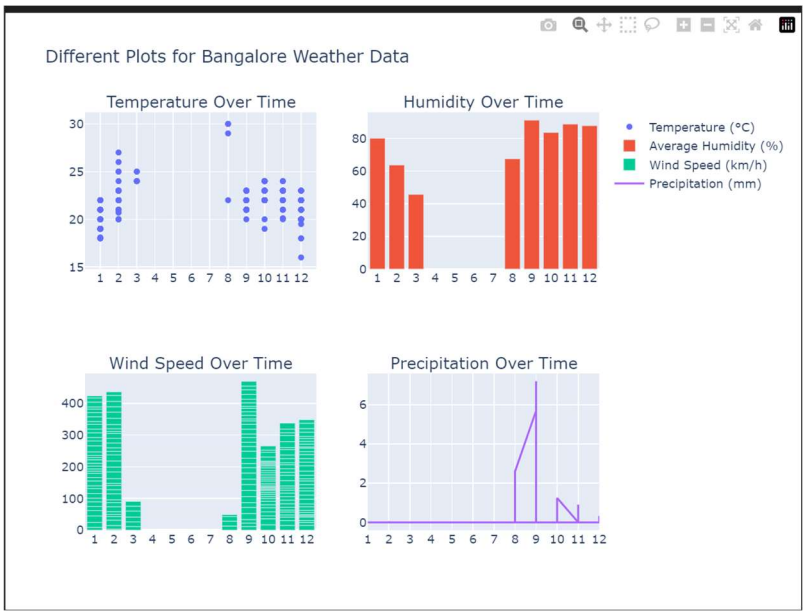
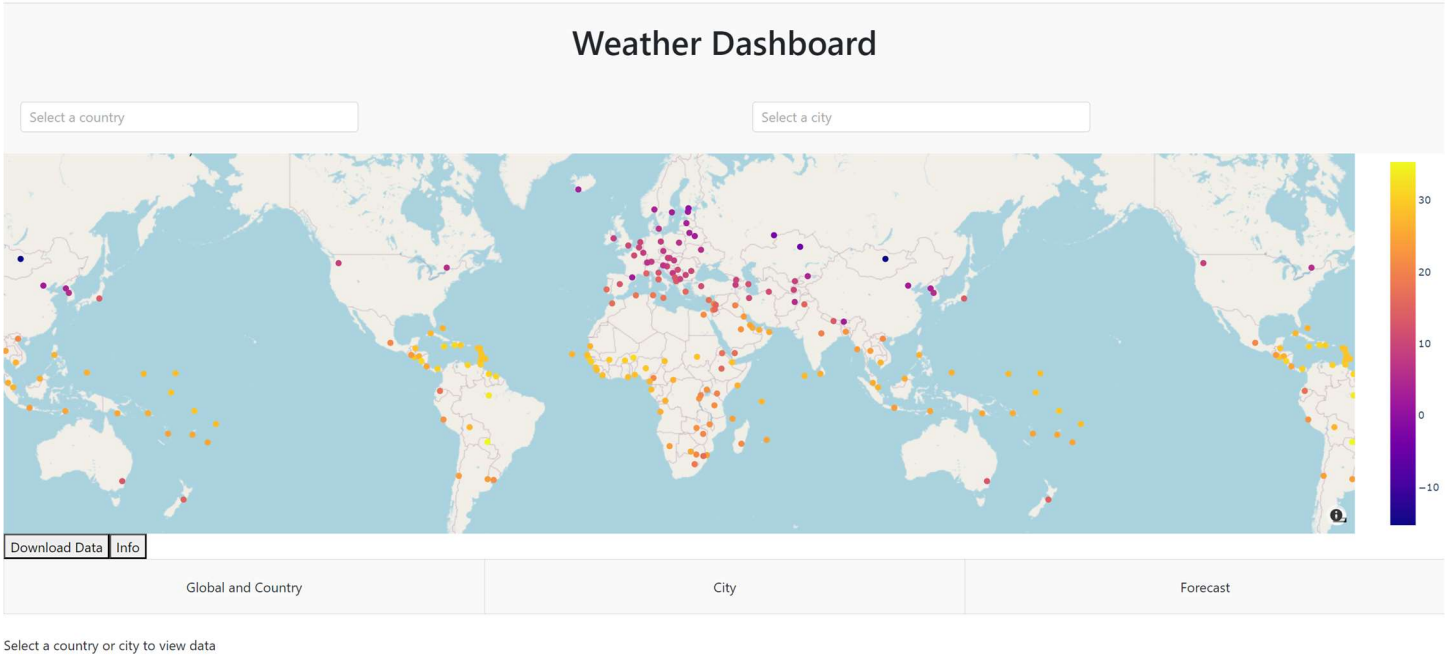


Fig. 32. Subplot of a city

This plot tells the climatic conditions of a city. There is a temperature over time, Humidity over time, wind speed, and precipitation over time plots. From these plots we can see that during September, there was high humidity, similarly, there was high precipitation suggesting a monsoon, also we have high wind speeds and lower temperatures, which shows that there were thunderstorms in the city during the month or heavy rains with high-speed winds lashed through the city. This is just one example of how these plots can be used. We can get much more information regarding the season, weather disasters, excess rain, etc. from these plots.

DASHBOARD

URL-https: https://dashapp-a53uxqxvkq-ue.a.run.app/

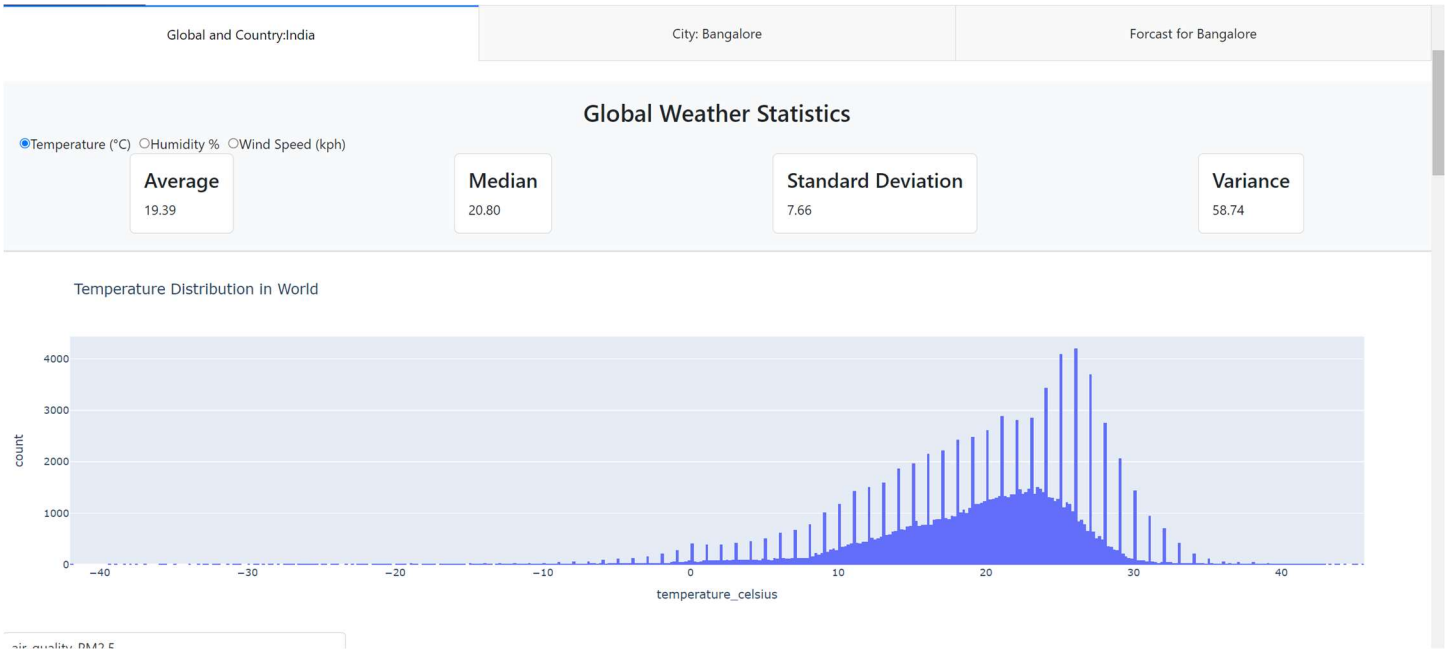


This is the Home Page with 2 dropdowns to select the country a city, and map, a download button to download the data file in a CSV format and an info button to view info on data columns and parameters.

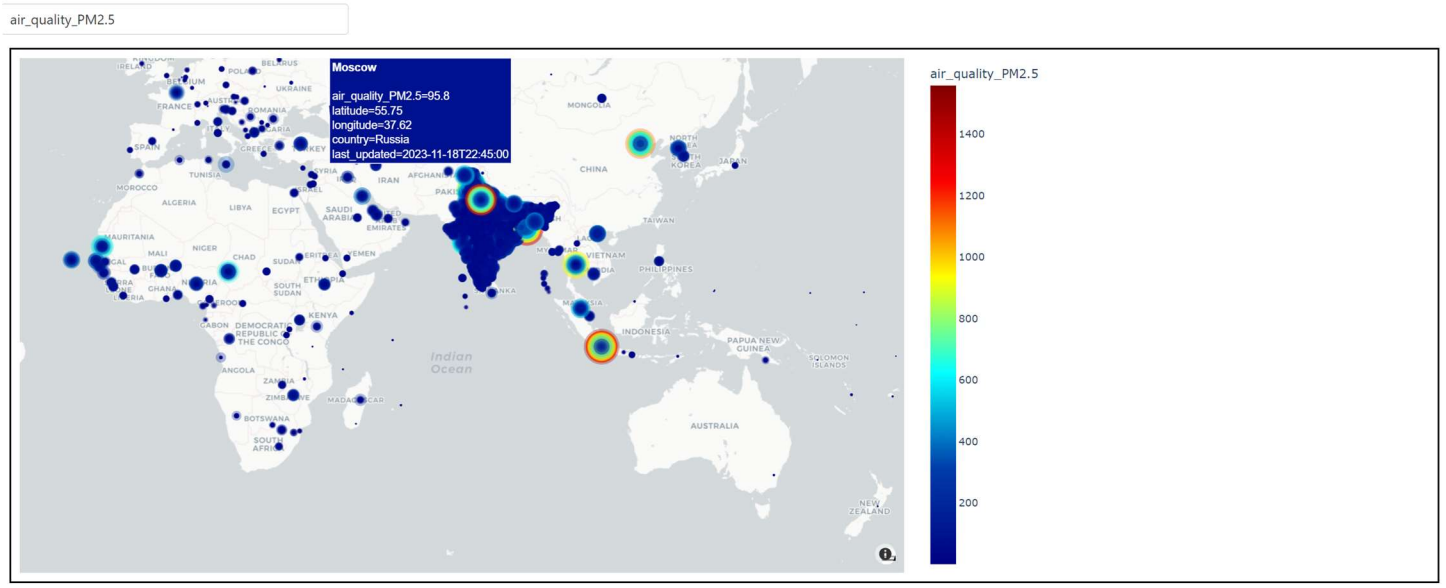
I have 3 tabs for Global and country-related plots and data. City tab for city-related plots and data and finally the Forecast tab, which is the live forecast for selected cities for that day.

Once the country and city are selected, the tabs are updated and the graphs are generated.

The Country tab has the following data-



There is a global weather statistics, where mean, median, deviation and variance of temp, wind and humidity can be checked. There is a temperature distribution of the world.



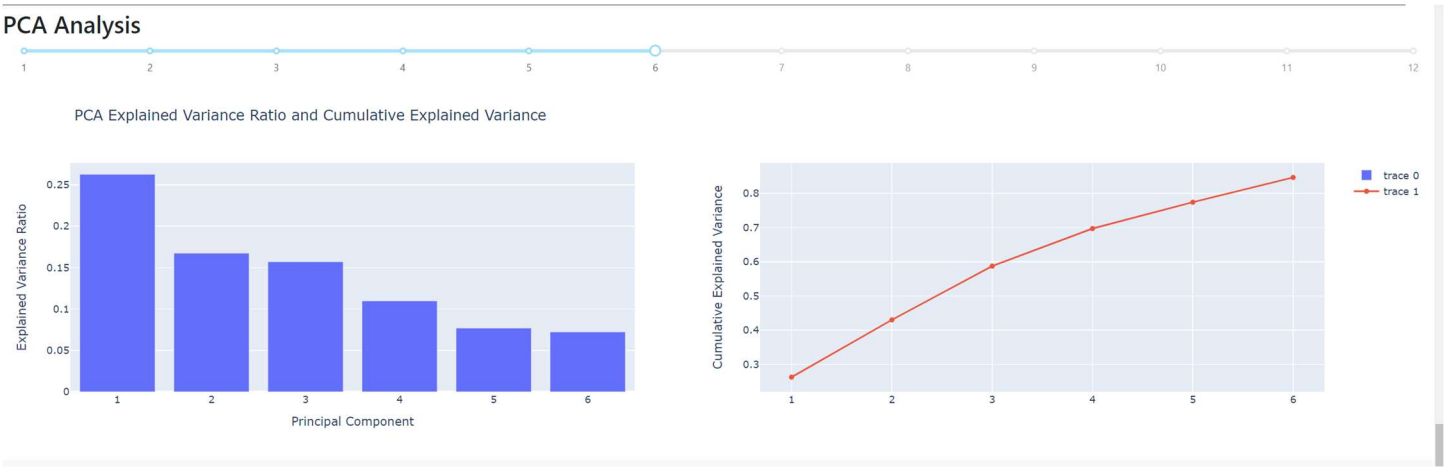
Following this we have an interactive scatterplot where we can change the air quality parameter to view the air quality index in a particular part of the world, what is the concentration, and how varied it is. This also has a loading icon, as the graphs take a few seconds to load.

Next we have some static plots displayed for global data that are already explained above

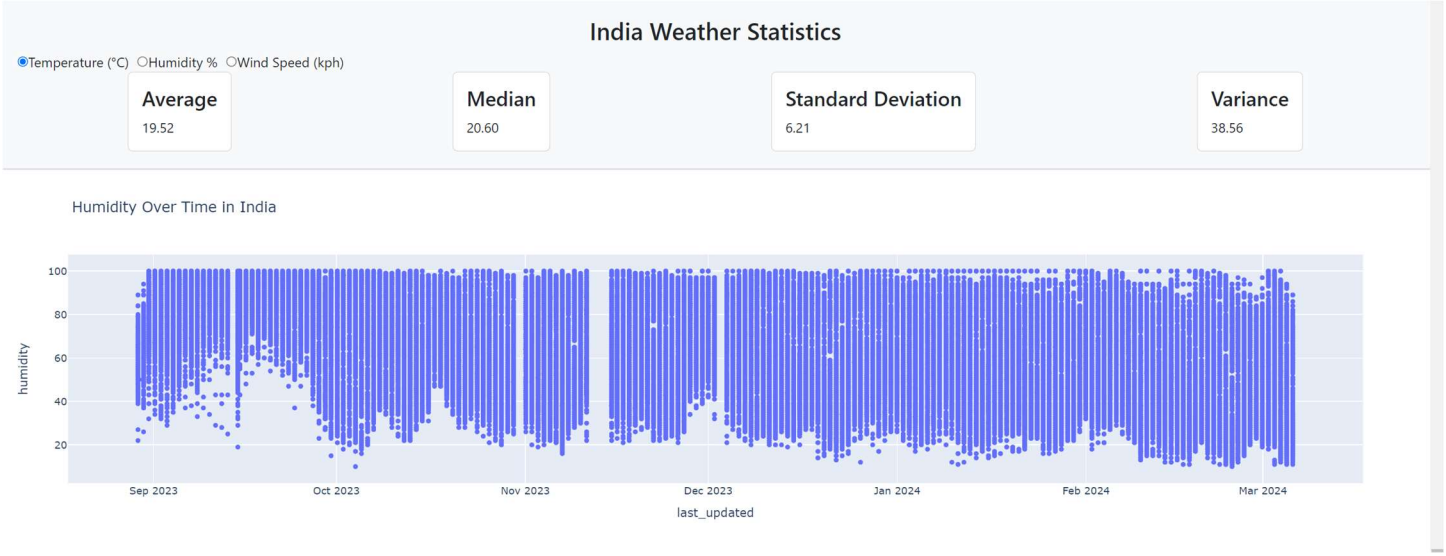
Then we have normality test and global outlier detection-

Here we can select the parameter in the dropdown view if it is normal or not and similarly for outlier detection as well we can see before and after outlier detection boxplots.

Next we have the PCA in the dashboard-



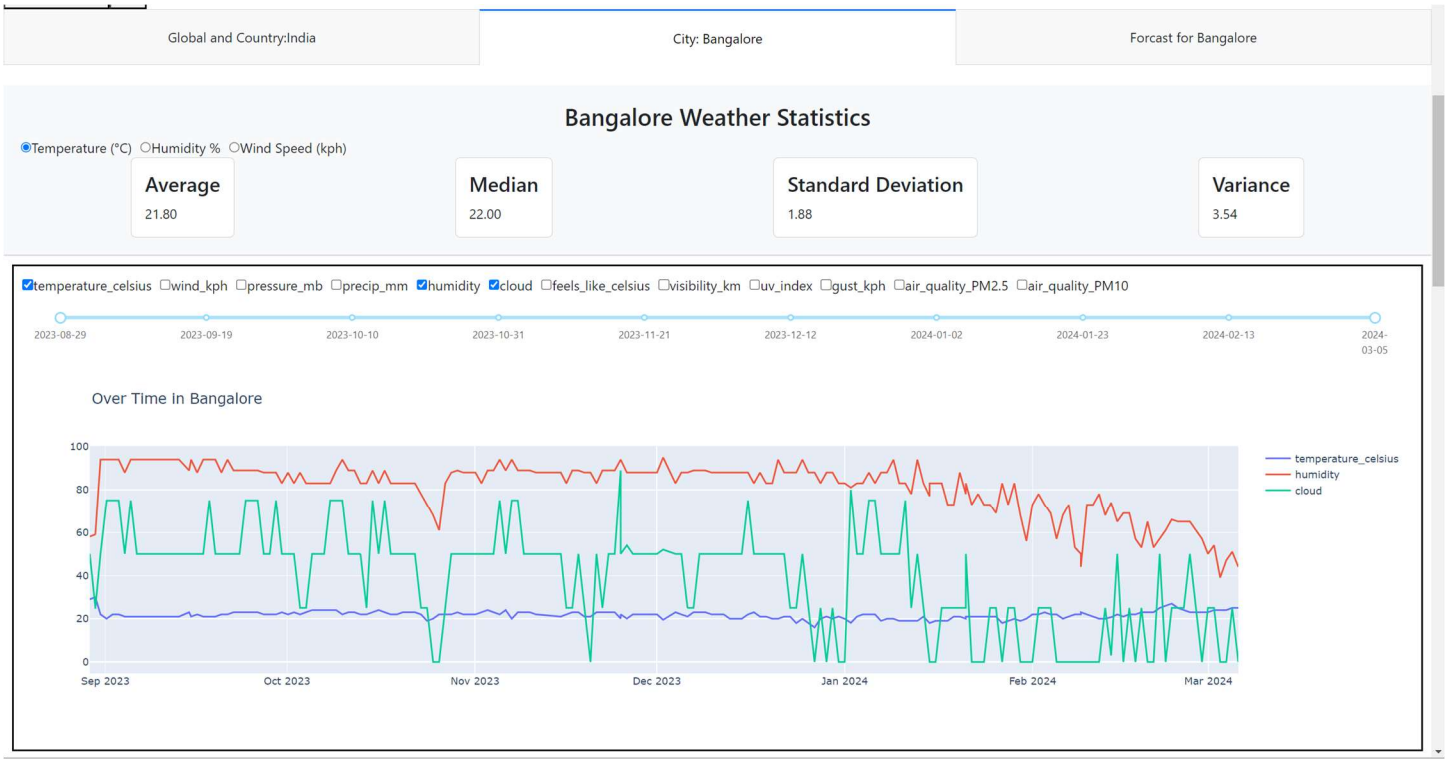
Here we have the slider to select the number of components and we can view how much of the explained variance is each component has and the cumulative explained variance as well.



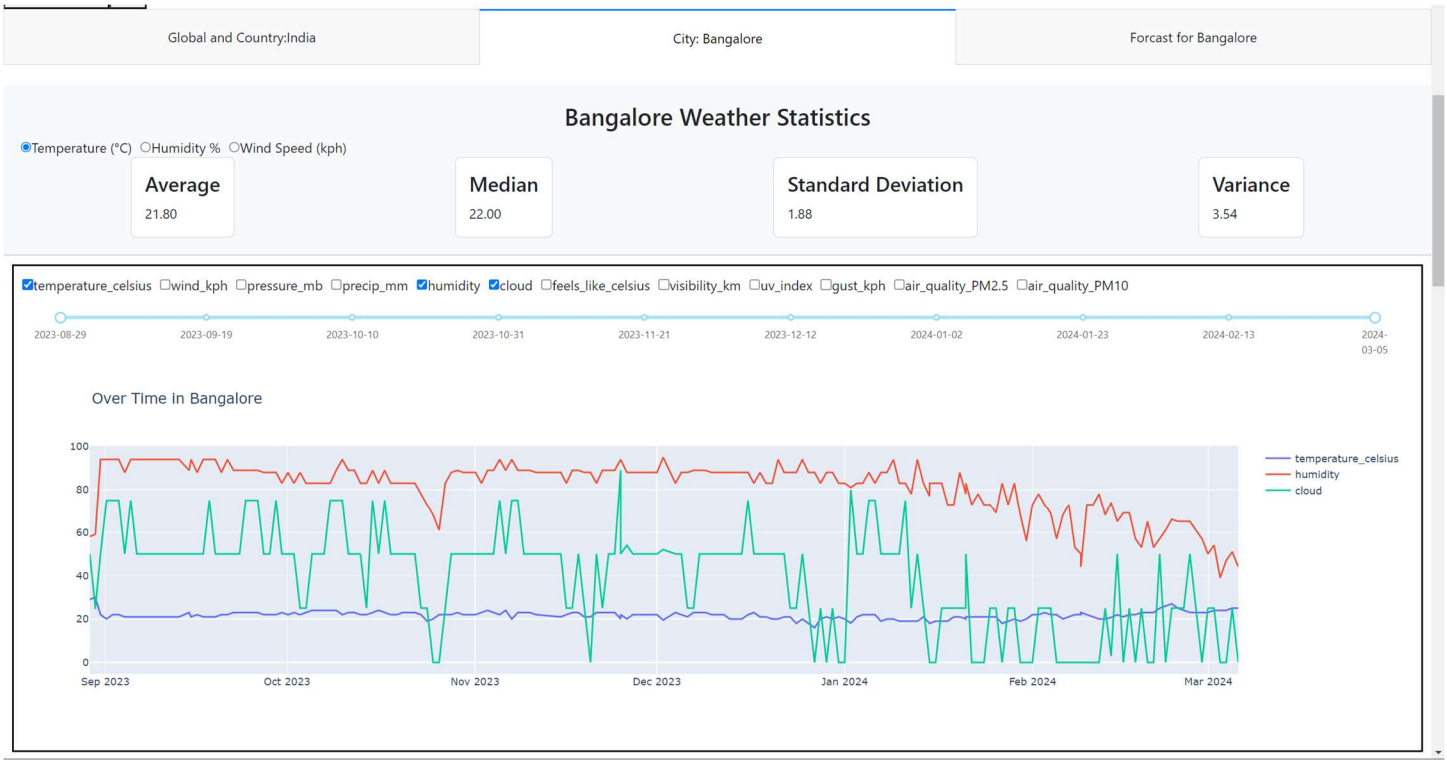
Further there country country-related statistics and humidity distribution in the particular country. This graph is important ot know what type of climate a country has, usually higher humidity, then there is a higher chance of rainfall as well. We can also get to know if it is a tropical country or a temperate just by viewing this plot.

Then we have country-level outlier detection for various parameters-

Moving on to the City dashboard-

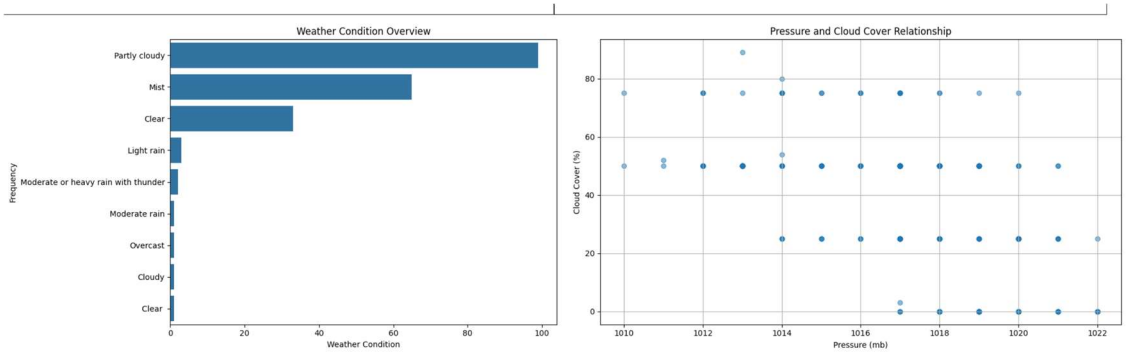


We have the city-related statistics and following that is a dynamic line plot with date slider and checkboxes for various parameters to plots. These plots are useful in understanding the relationship between different parameters in shaping the weather pattern of the city. W can view a particular time frame and understand, why there is higher temperature, what is the cause, or why did it rain heavily etc.



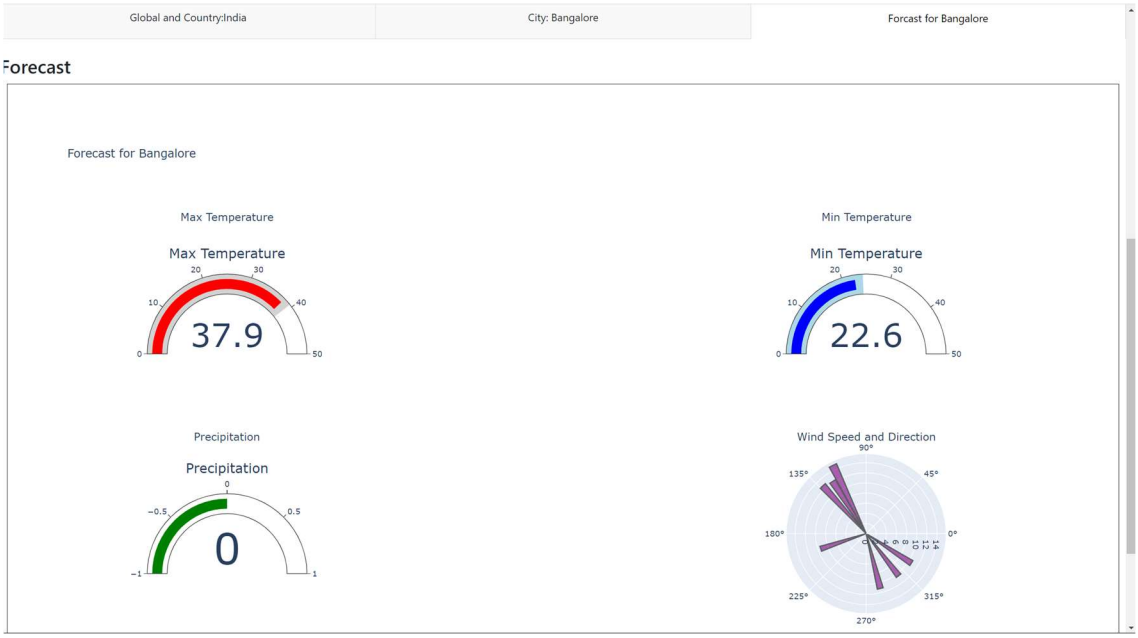
Next, we have a city-level outlier detection. These 3 levels of outlier detection give us the accurate outliers for the respective locations, rather than removing real weather change data.

Then we have 2 static plots showing the common weather conditions in the city and pressure and cloud cover relationship in the city-



The first plot can let us know about the common weather of the city, also help us plan out trips to visit, while the second plot serves the scientific purpose of showing the pressure and cloud cover relation. When there is high pressure, the cloud cover is less, which is a common phenomenon. Following this we have the subplot that was explained previously.

In the third tab, the Forcast Tab-



We have the live forecast of the weather pattern of temperature, precipitation, and wind speed along with the direction for that particular day in the city. These data are collected from the openmeteo API and then displayed

as plots. Integrating live forecast data weather analysis dashboard enhances its functionality by providing real-time insights into temperature, precipitation, and wind speed patterns for specific cities. This provides access to the most recent weather forecasts, ensuring access to the latest data to plan activities or make decisions, Live forecasts enhance the accuracy of weather predictions by incorporating the latest meteorological data, Combining historical weather analysis with live forecasts provides a comprehensive view of weather patterns over time.

CONCLUSION

The weather dashboard project offers a holistic and enriching learning experience, involving the depths of data visualization, user interaction, and real-time data analysis. The array of graphs created in this project—from heatmaps to time series and kernel density estimations—has reinforced the understanding of weather dynamics. For instance, the heatmaps and Pearson correlation matrices uncovered the relationships between different weather parameters, revealing how factors such as humidity and temperature are interrelated. Time series analysis of temperature over time provided insights into seasonal patterns and temperature trends. The kernel density plots and box plots unveiled the distribution characteristics of weather phenomena, allowing for a in depth understanding of variability within climatic data.

The dashboard in this project is an easy tool to access and operate. It allows users to filter by country, city, and various weather conditions, the dashboard facilitates a user-directed exploration of the data, catering to both broad and specific informational needs. The interactive elements, like sliders and dropdowns, make it convenient for users to customize their queries, ensuring that the information gleaned is pertinent to their interests. This interactivity enhances user engagement with the data, making the experience both informative and compelling. The app's functionality is evidenced by its robust performance, integrating multiple levels of data analysis—from statistical testing to live forecasting. The normality tests and outlier detection features ensure that users are viewing quality-controlled data, while the PCA offers an advanced level of analysis for those interested in the underlying structure of the weather data. Additionally, the inclusion of live forecasting signifies the app's functionality extending beyond static historical data analysis, offering users valuable insights into current and forthcoming weather conditions. This blend of historical and real-time data serves not only to inform but also to prepare users for future conditions.

In conclusion, this weather dashboard project encapsulates the essence of data science's potential to transform raw data into interactive, insightful, and user-centric applications. It demonstrates how well-designed tools can make data accessible and actionable, providing users with powerful resources to understand and interact with the world around them.

REFERENCES

[1] Outlier Detection & Removal, CHIRAG GOYAL, https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/

[2] A Step-by-Step Explanation of Principal Component Analysis (PCA), Zakaria Jaadi, https://builtin.com/data-science/step-step-explanation-principal-component-analysis