

**FORECASTING TOMORROW: A CLIMATIC STORY TOLD BY DATA**

**TEAM - ANOVA**

**Shaurya Uniyal**

**Prathamesh Mayekar**

**Purvesh Khalatkar**

**Sancia Fernandes**

**Objective:**

*The objective of this report is to analyze and derive actionable insights from the provided daily temperature and precipitation data for multiple cities spanning from 01/01/1990 to 20/07/2022. The report aims to address key concerns and interests of various stakeholders, including farmers, traders, and the general public. The specific objectives include:*

**Data Quality Assurance:** Address the issue of missing values in the dataset, ensuring that the analysis is based on reliable and complete information. Develop strategies to handle missing data in a way that minimizes bias and preserves the integrity of the insights.

**Temporal Deviations Analysis (2021 and 2022):** Explore and identify deviations in temperature and precipitation patterns for the years 2021 and 2022 compared to historical averages. Provide insights into any anomalous trends or extreme weather events that could impact agriculture, trade, and daily life.

Rainfall Prediction Model for Santacruz (2023): Build a predictive model for rainfall in Santacruz for the year 2023, leveraging historical data. Evaluate the accuracy of the model by comparing predicted values with

**Rainfall Prediction Model for Santacruz (2023):** Build a predictive model for rainfall in Santacruz for the year 2023, leveraging historical data. Evaluate the accuracy of the model by comparing predicted values with actual observations. Provide recommendations for mitigating disparities and improving prediction accuracy.

**Rainfall Prediction for 2024:** Utilize historical data to predict rainfall for a specific region in the year 2024. Employ appropriate forecasting techniques and models to estimate precipitation levels, allowing stakeholders to make informed decisions for the upcoming year.

**Effective Visualizations:** Develop an interactive and visually appealing set of visualizations that effectively convey the insights derived from the analysis.

*By achieving these objectives, the report aims to empower stakeholders with valuable information to make informed decisions related to agriculture, trade, and daily activities, ultimately contributing to better resilience and preparedness in the face of varying weather conditions*.

**Introduction:**

Climate plays a pivotal role in shaping the socio-economic landscape of regions, influencing agricultural practices, trade dynamics, and daily life. Understanding historical weather patterns and predicting future trends is crucial for stakeholders such as farmers, traders, and the general public to make informed decisions and mitigate potential risks. In this context, the provided dataset offers a comprehensive repository of daily temperature and precipitation data from 01/01/1990 to 20/07/2022 for cities including Delhi, Bangalore, Chennai, Lucknow, Rajasthan, Mumbai, Bhubaneswar, and Rourkela.

This report embarks on a data-driven journey to extract meaningful insights from the wealth of information encapsulated in the dataset. By addressing critical issues like missing data, analyzing temporal deviations, building predictive models, and crafting effective visualizations, the report aims to empower stakeholders with actionable intelligence. The objective is to enhance decision-making processes related to agriculture, trade, and day-to-day activities by providing a nuanced understanding of historical patterns and facilitating predictions for the future.

Through a combination of statistical analyses, machine learning methodologies, and visual representations, this report endeavors to unravel the complexities of weather data, translating them into tangible benefits for farmers adapting to changing climate conditions, traders navigating market dynamics influenced by weather patterns, and the common people planning their daily lives. The subsequent sections will delve into each objective outlined in the report, presenting findings and recommendations that harness the power of data science to navigate the intricate interplay between weather and human activities.

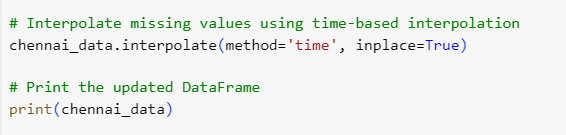
**Methodology and Analysis:**

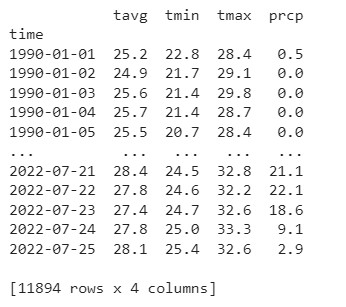
**Q.1 Address the issue of missing values in the given data set.**

The dataset containing daily temperature (Minimum, Average, Maximum) in degrees Centigrade and Precipitation data in mm was imported into a Pandas Data Frame. Initial exploration of the dataset revealed missing values in certain columns. For each city dataset the initial step involves addressing missing values in the temperature (tmin, tmax) columns. The 'NaN' values in 'tmin' and 'tmax' are replaced using the forward fill method to maintain temporal coherence.

Remaining missing values in the datasets are then interpolated to ensure a continuous and smooth representation of the data.

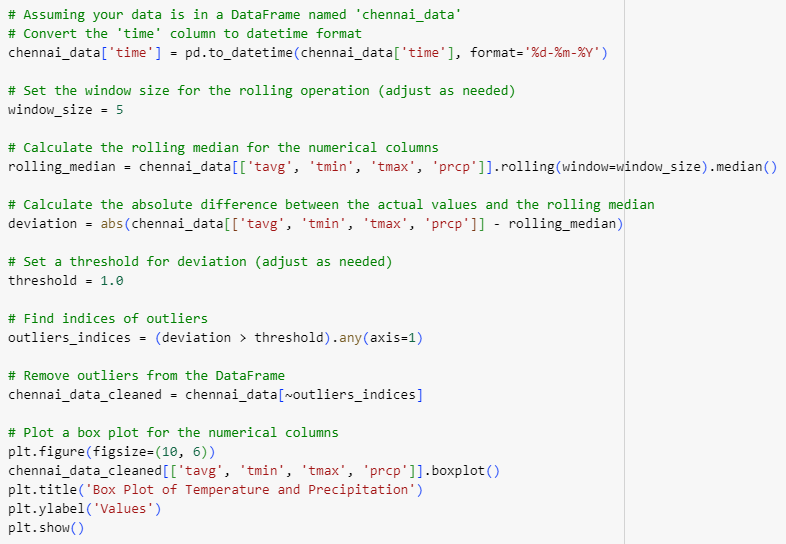
Here’s an example for the Chennai dataset:

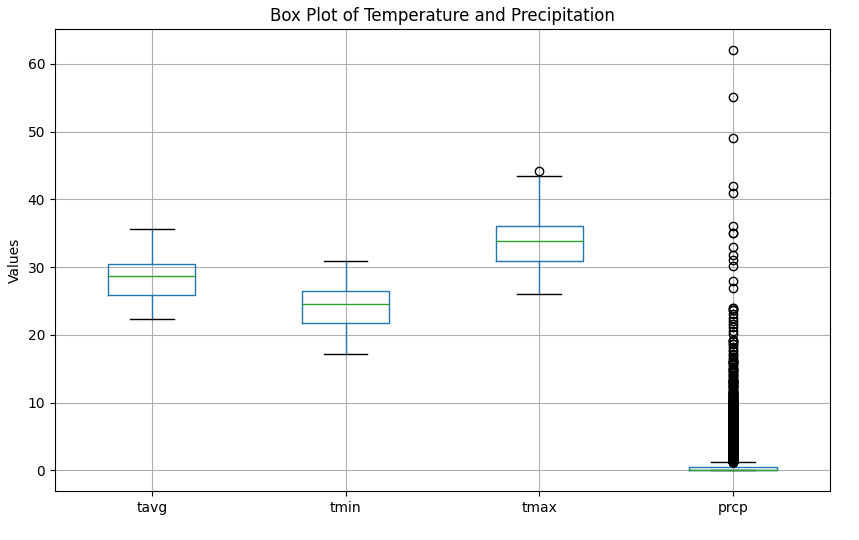


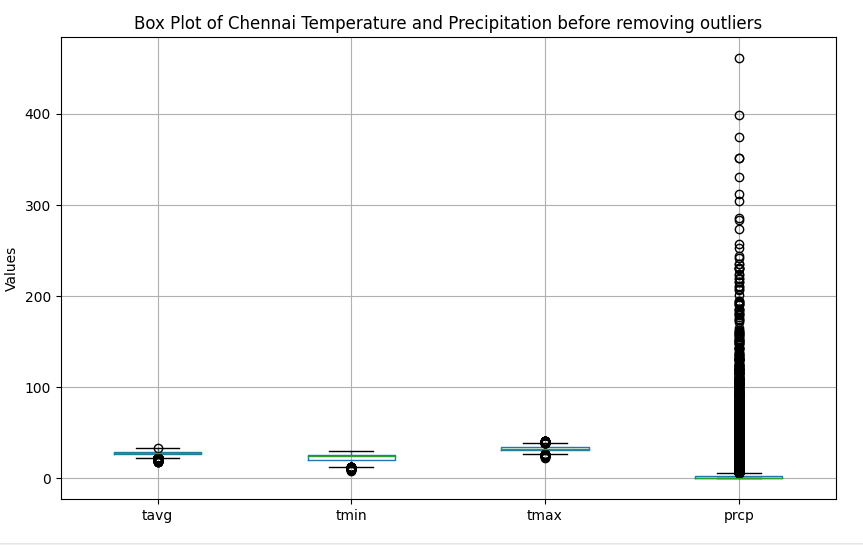


Outliers are identified in each city dataset through a rolling median-based approach.

A rolling median is calculated for temperature (tavg, tmin, tmax) and precipitation (prcp) columns using a specified window size. The absolute difference between the actual values and the rolling median is computed to quantify deviation. A threshold is set to identify outliers, and the indices of the outlier rows are determined. Outliers are removed from the datasets to ensure robust and accurate analysis. To visually assess the impact of outlier removal and overall distribution, box plots for temperature (tavg, tmin, tmax) and precipitation (prcp) are generated for each city.





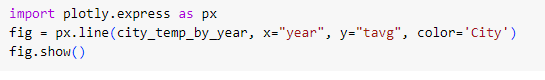


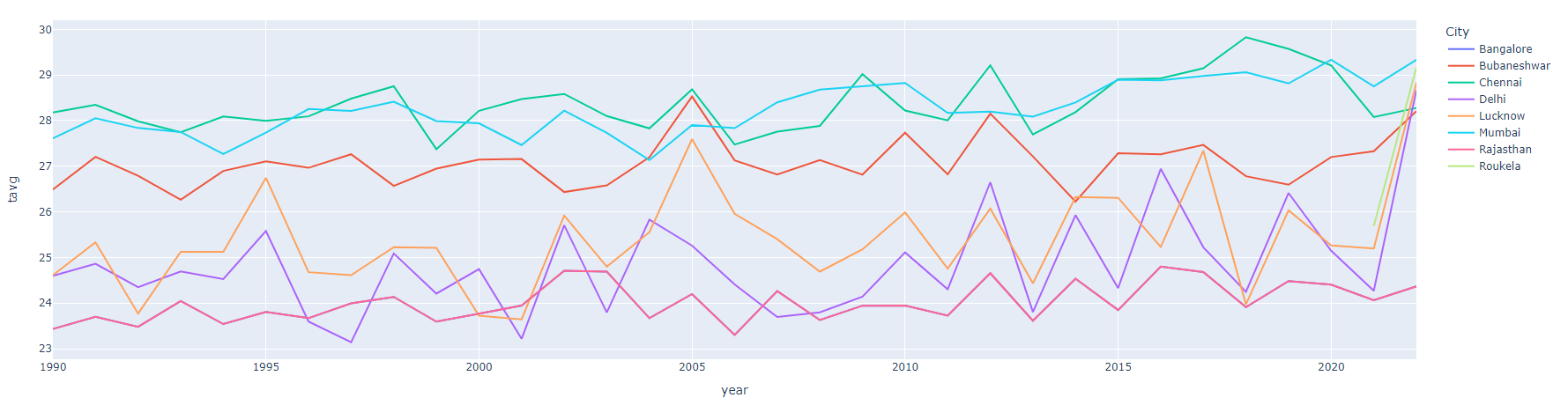
**Q.2 Find deviations in 2021 and 2022 as compared to the historical average or**

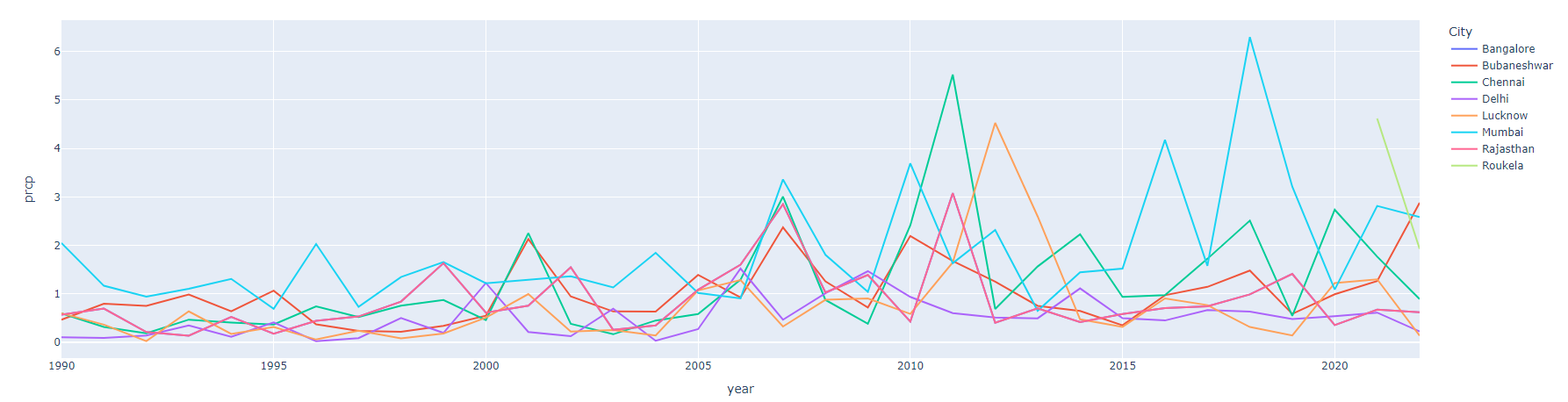
**patterns.**

Combined and concatenated cleaned datasets for multiple cities to form a comprehensive dataset encompassing temperature and precipitation data. The 'time' column was converted to a datetime object for temporal analysis.

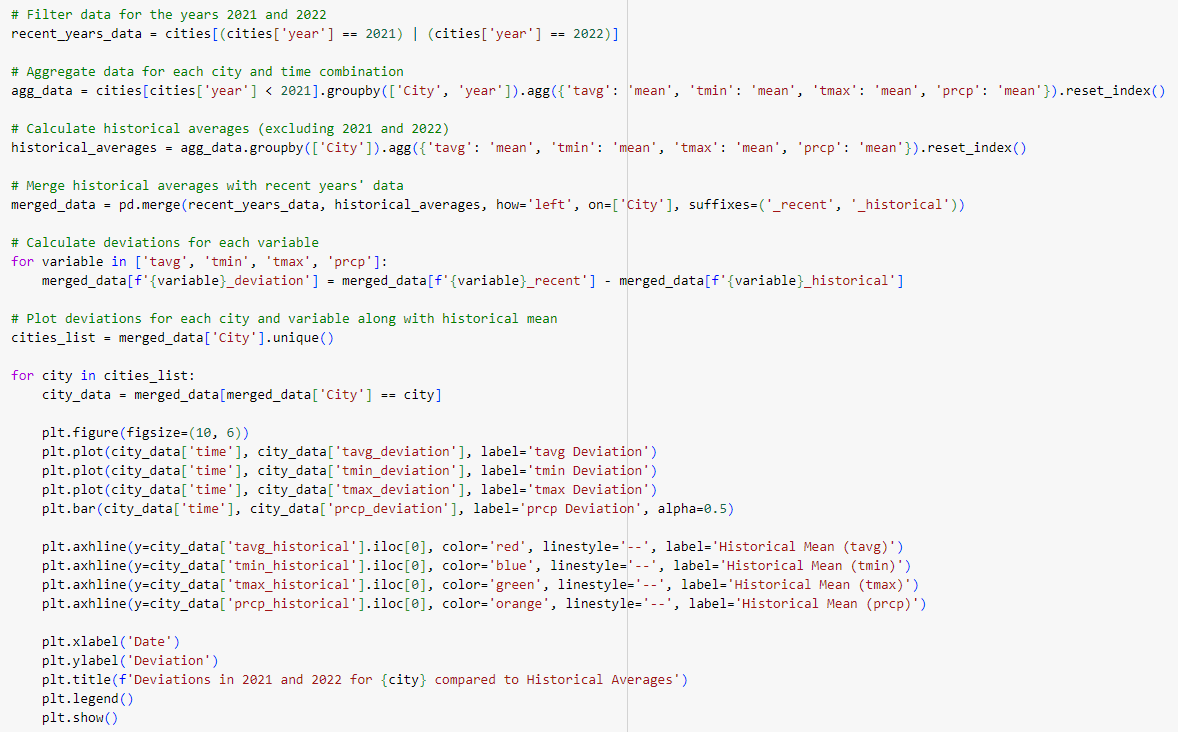
Calculated the average temperature (tavg) and precipitation (prcp) for each city over the years. Utilized Plotly Express to create an interactive line plot showcasing the trends in average temperature and precipitation for each city over the years.

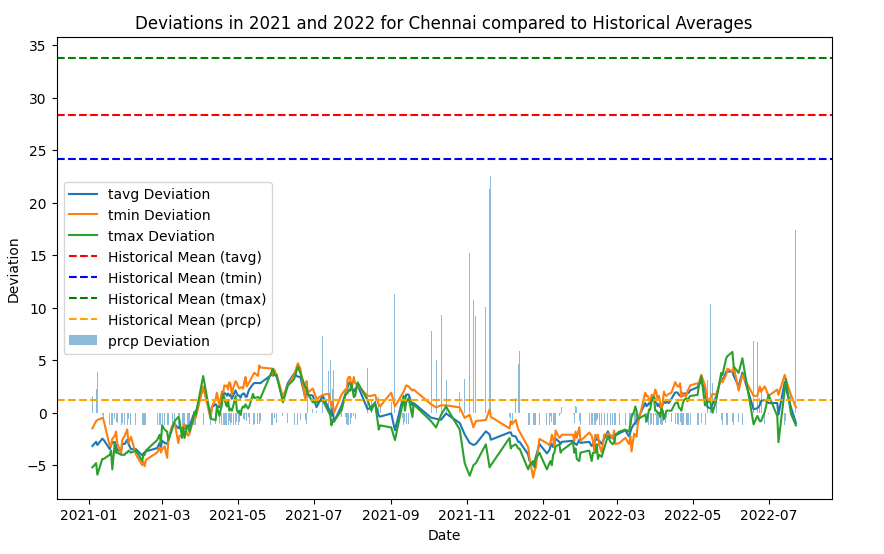


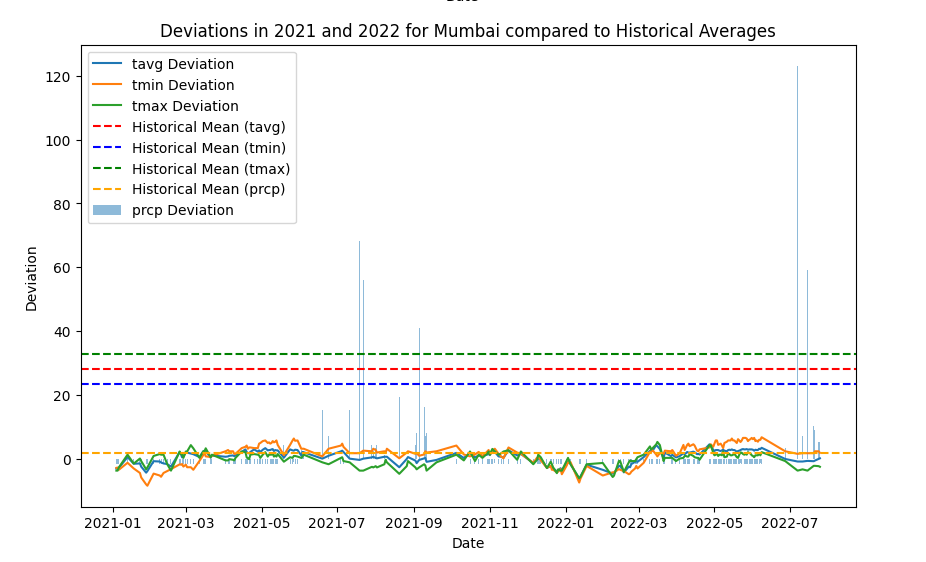


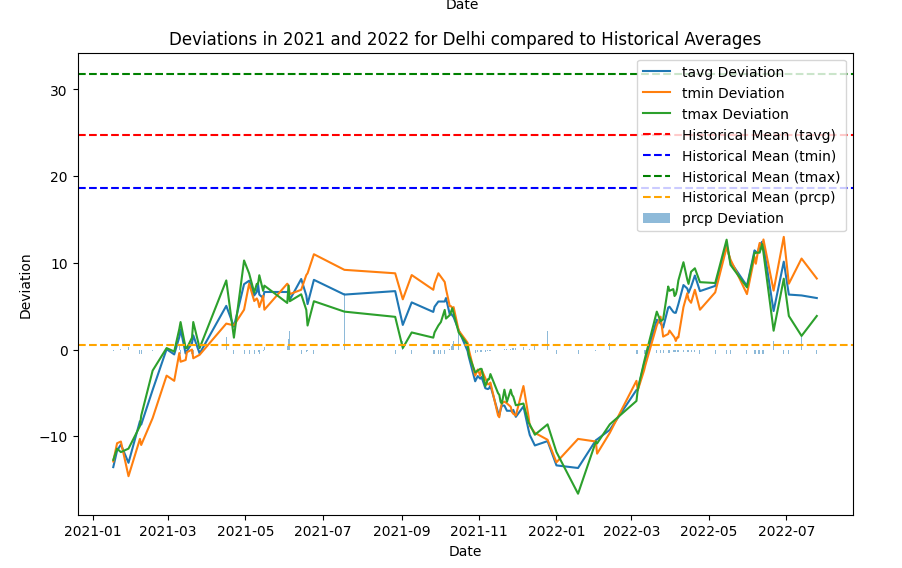


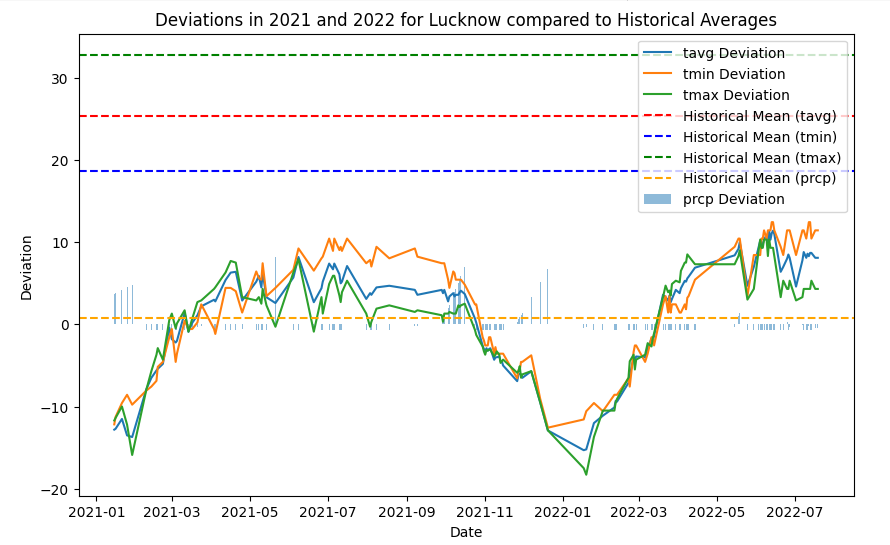
Filtered the dataset for the years 2021 and 2022 to focus on recent trends. Aggregated historical averages for each city, excluding the recent years (2021 and 2022). Calculated deviations in temperature (tavg, tmin, tmax) and precipitation (prcp) for the recent years compared to historical averages. Visualized deviations along with historical means using Matplotlib.

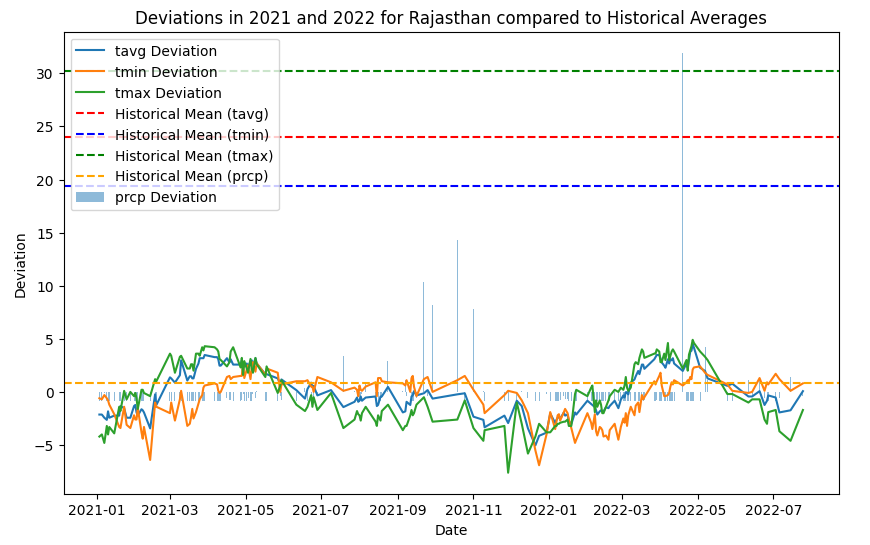


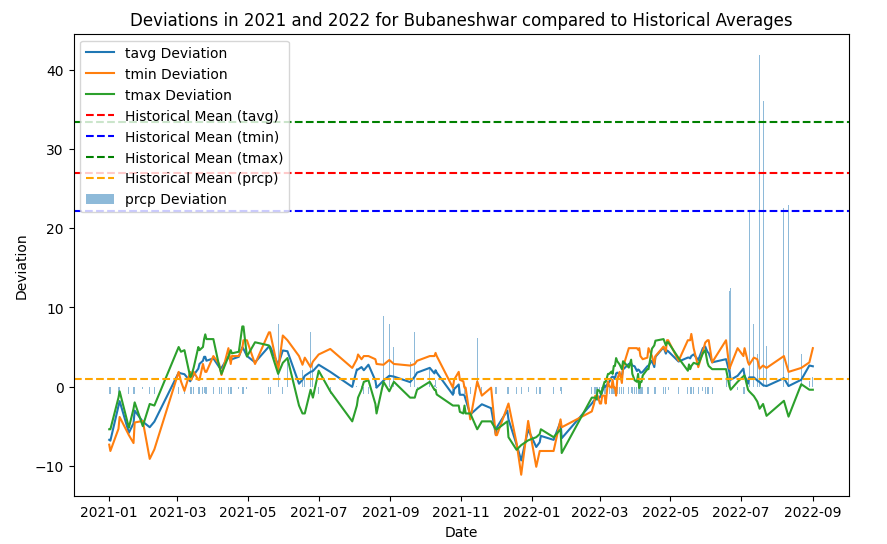


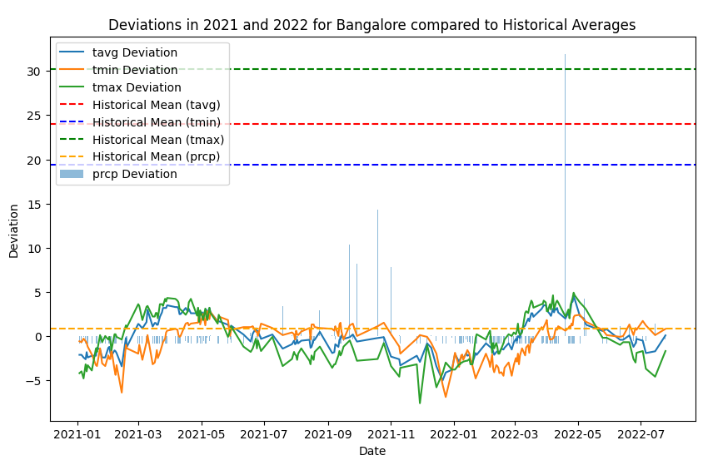




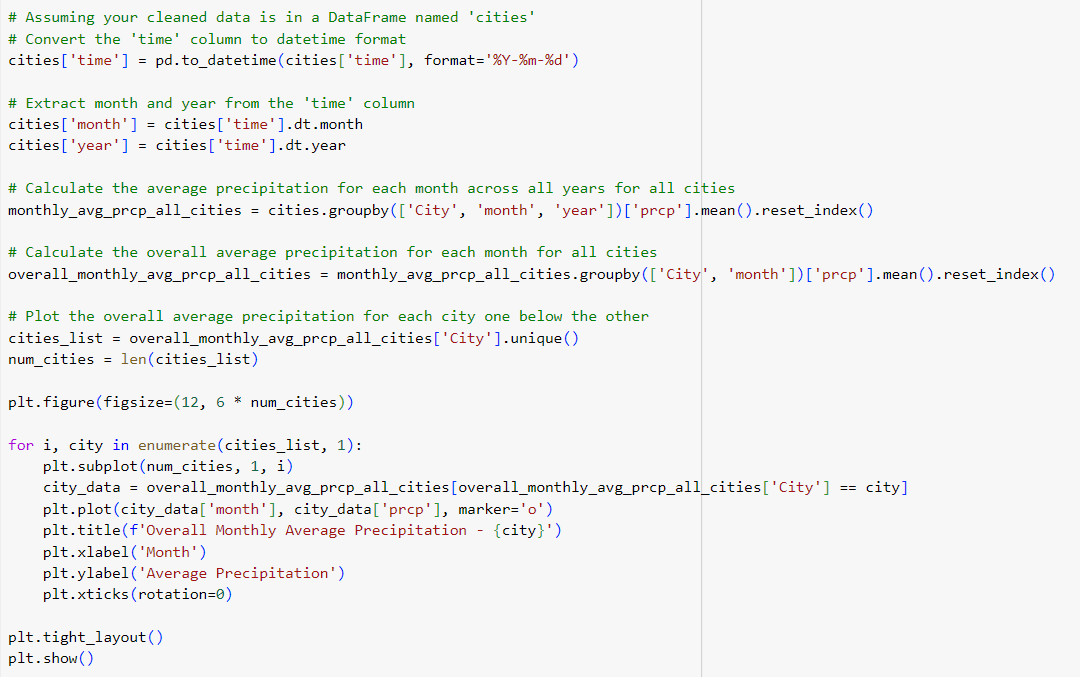


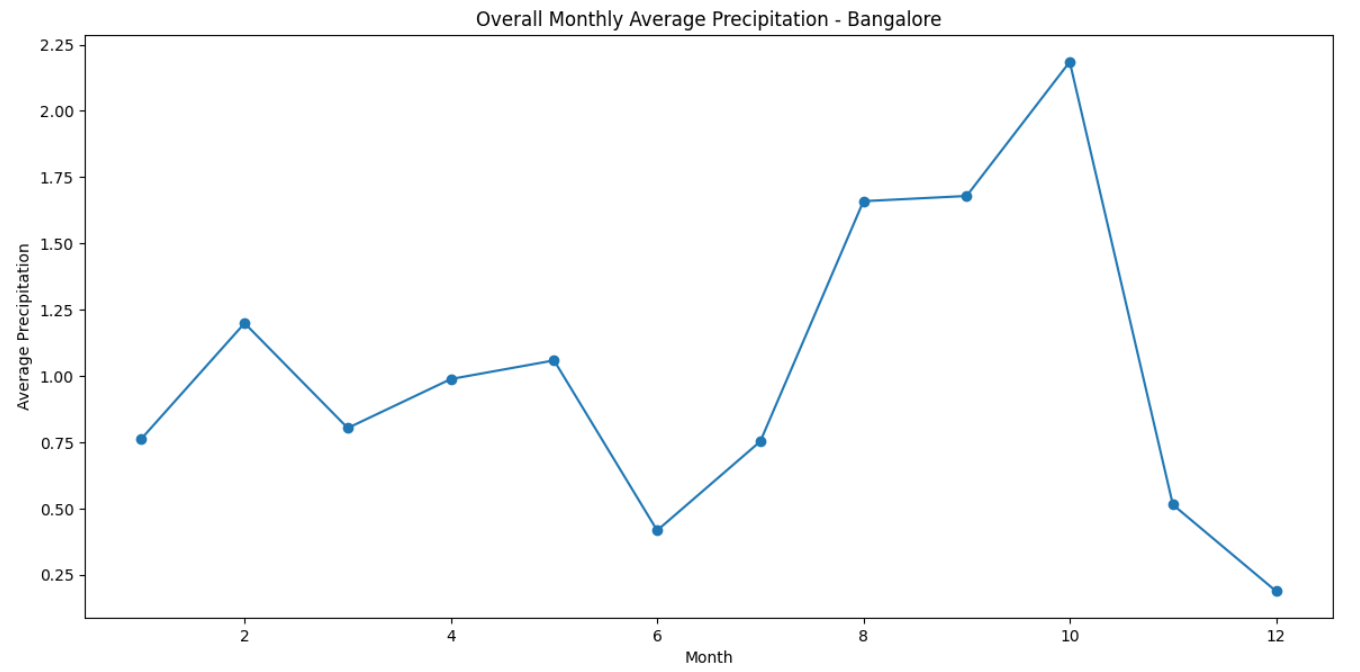


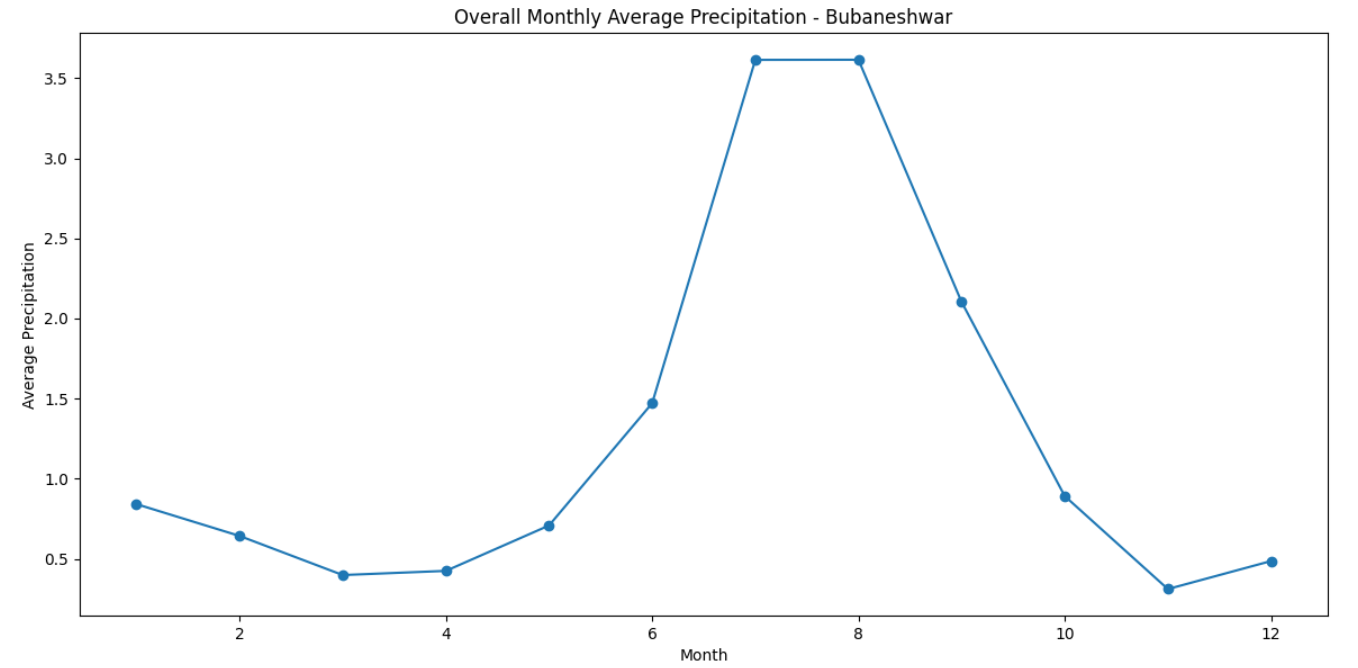


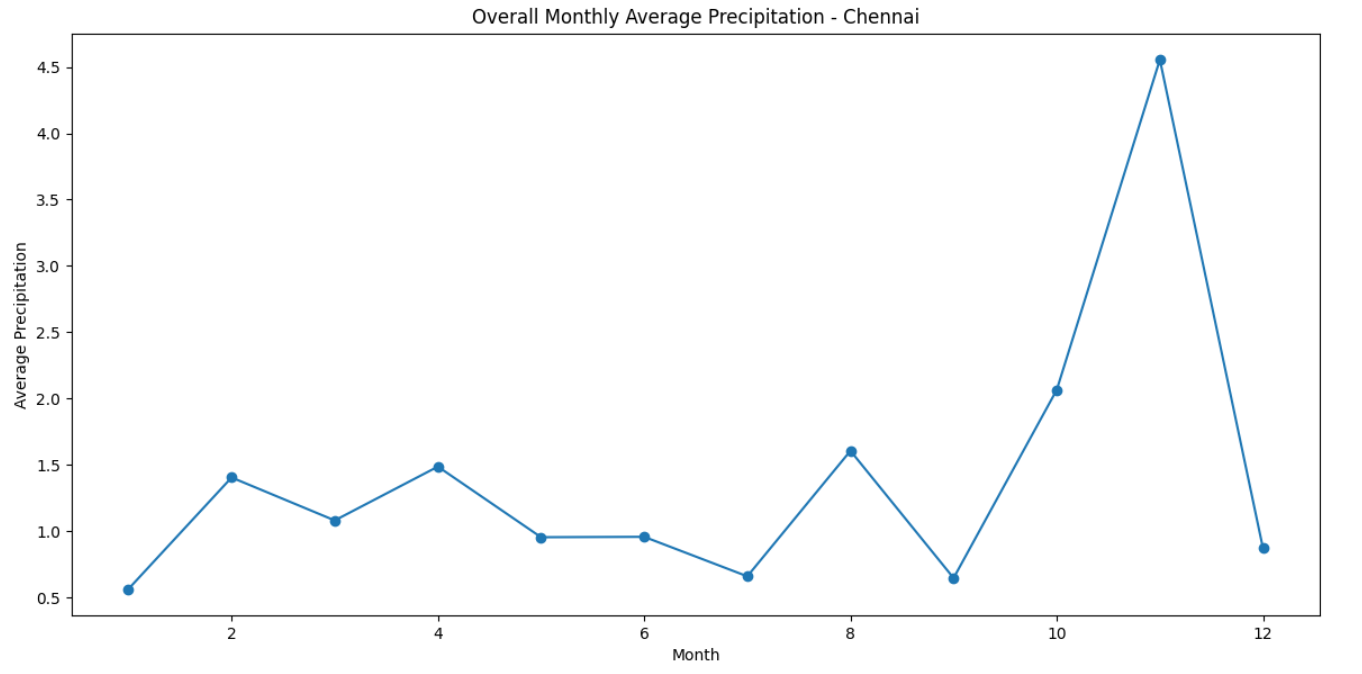


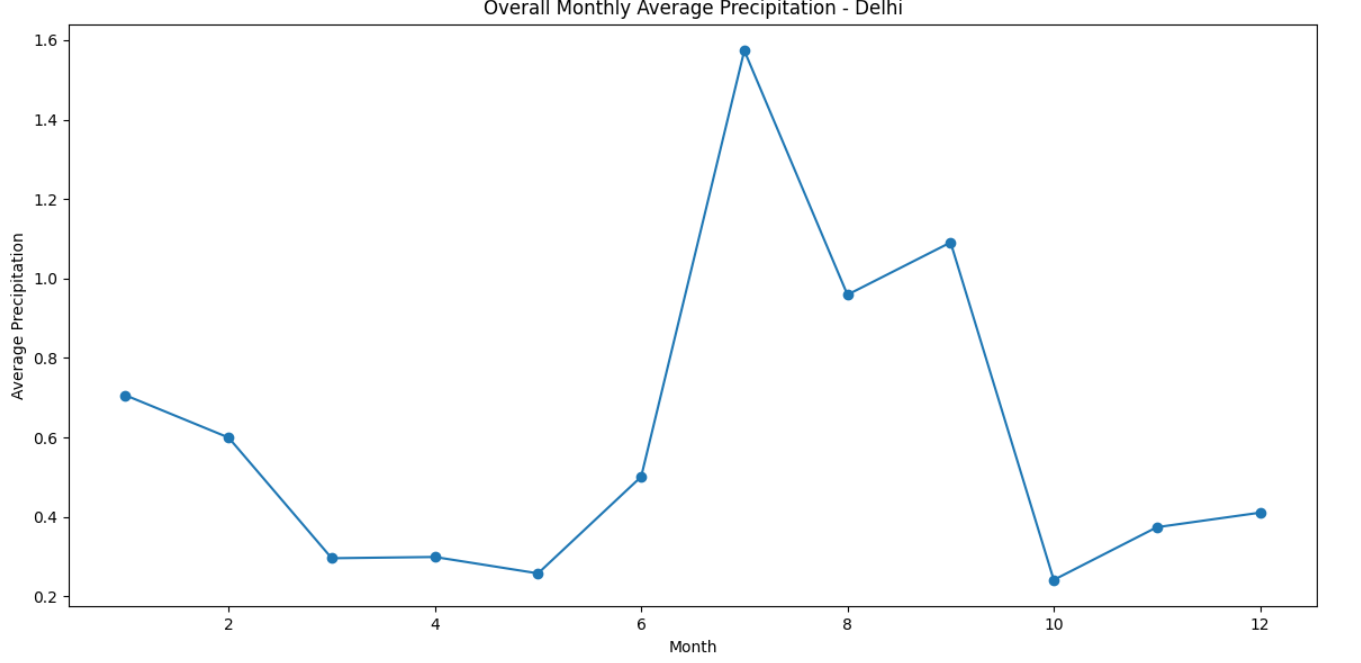
Calculated the average precipitation for each month across all years for all cities. Further aggregated the data to find the overall average precipitation for each month for all cities. Plotted the overall average precipitation for each city, to visually compare monthly trends, providing a comparative view of precipitation trends for each month. The subplot layout aids in clear visualization and comparison of the data.

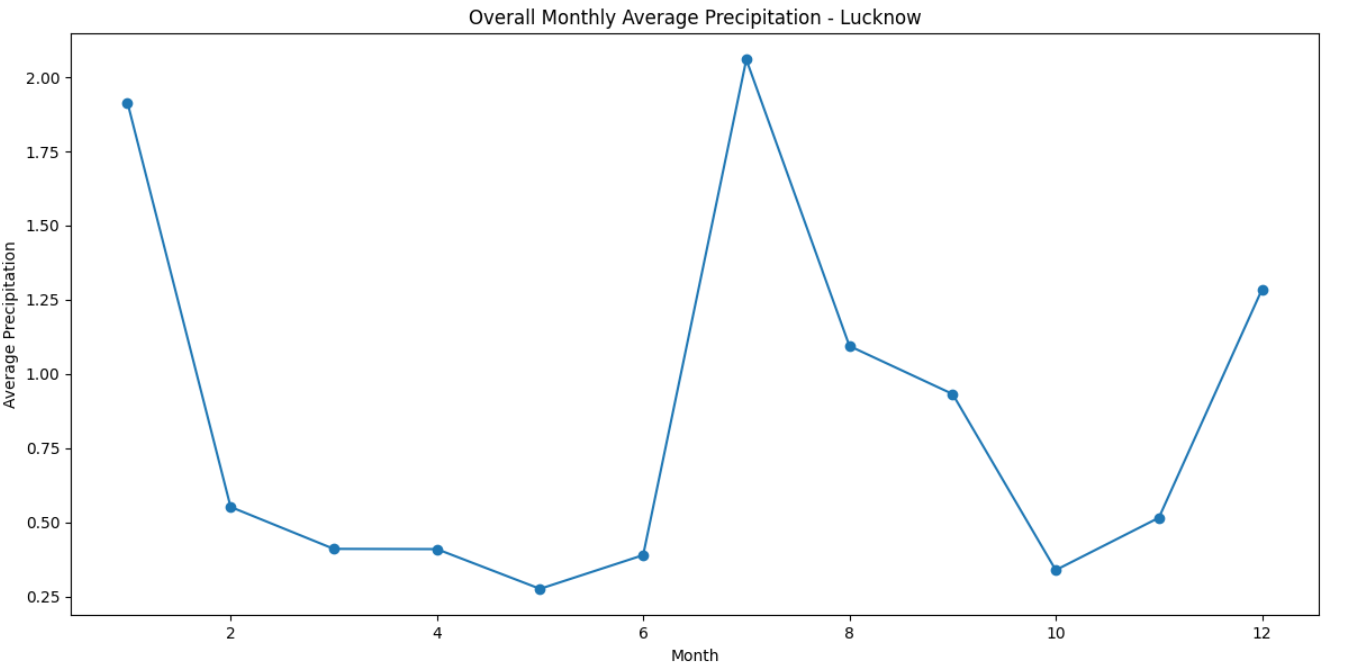


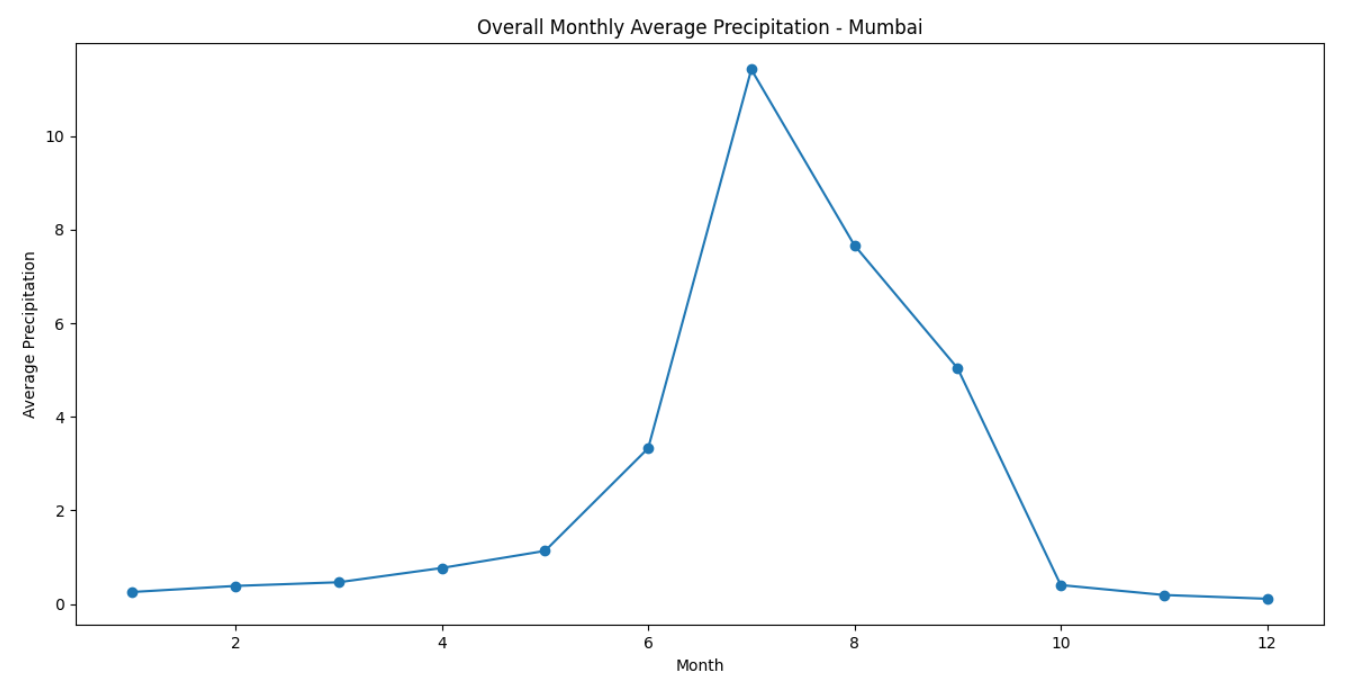


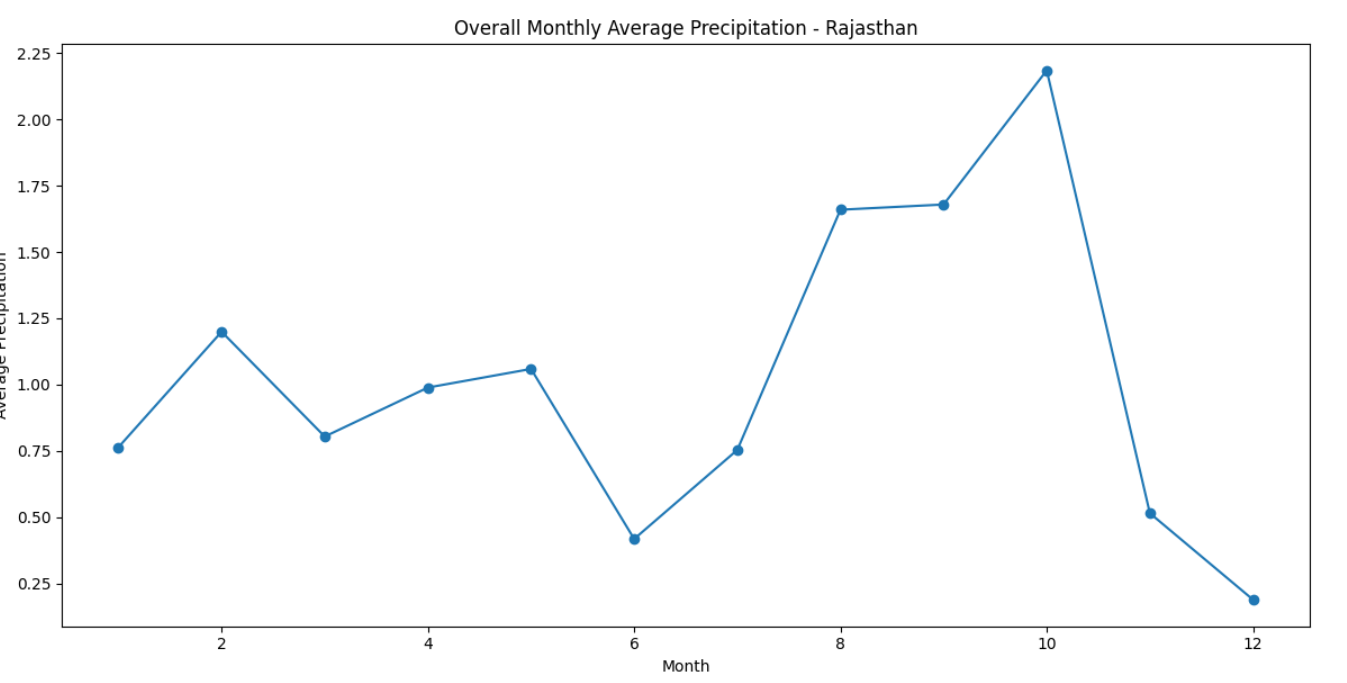


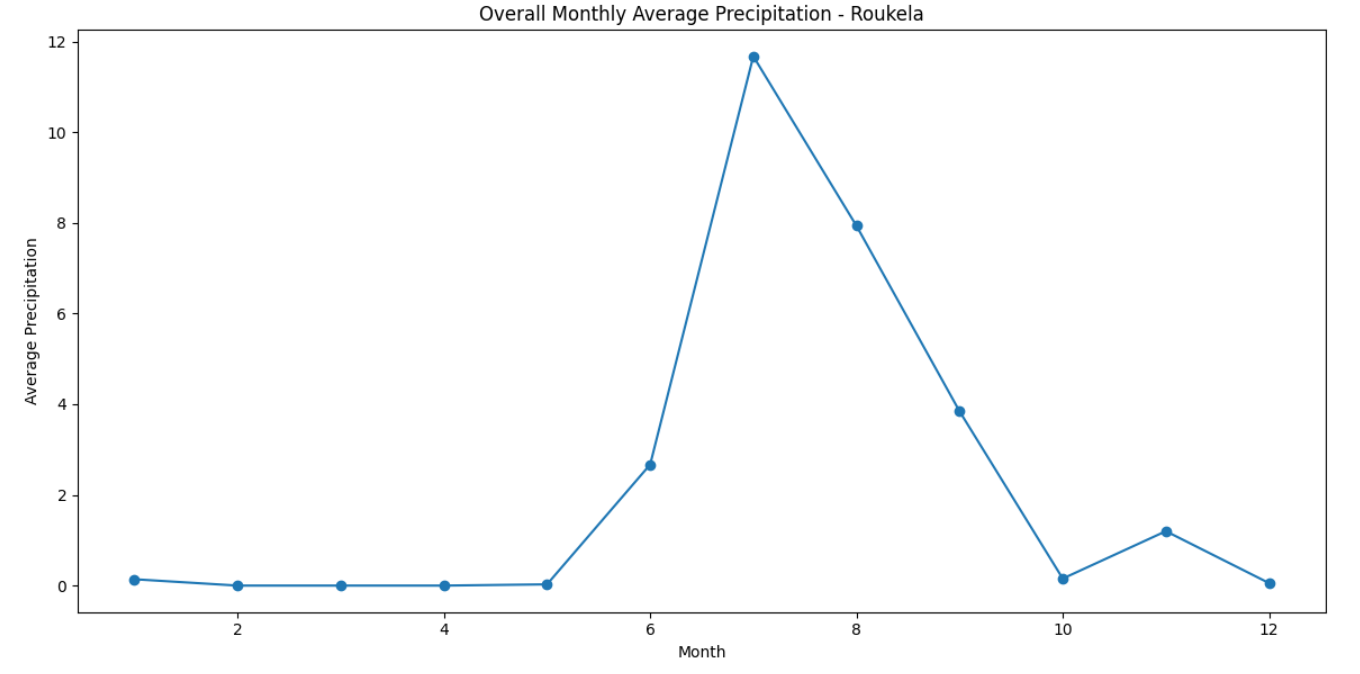










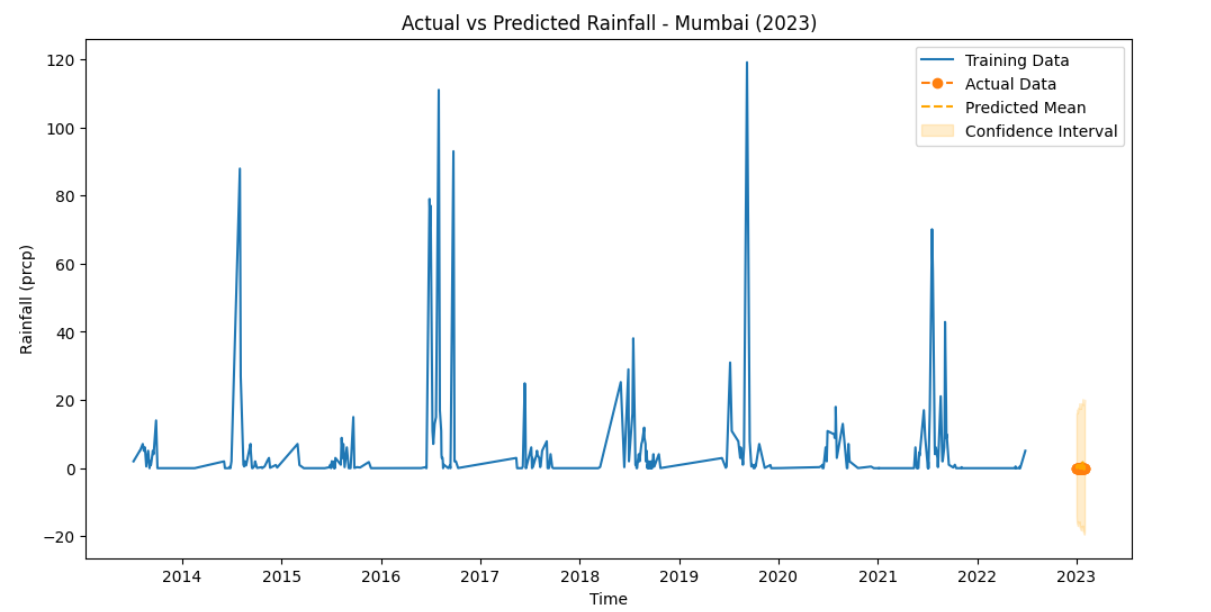


**Q.3. Build a model to predict the rainfall for Santacruz for the year 2023 and**

**compare it with actual values. Comment on the disparity if any.**

Imported necessary libraries - pandas, numpy, matplotlib, statsmodels, and sklearn.

Read the cleaned Mumbai-Santacruz dataset for 2023. Set the 'time' column as the index for both datasets. Extracted data for the last 10 years (2013-2022) from the Mumbai dataset. Split the data, using information up to 2022 for training and data for 2023 as the actual test data. Utilized the SARIMAX model (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors) from statsmodels. Configured the SARIMA model with order (1, 1, 1) and seasonal order (1, 1, 1, 12) based on the observed time series characteristics. Fitted the SARIMA model using the training data. Predicted rainfall values for the year 2023 using the SARIMA model with a specified forecast horizon. Extracted predicted mean values and confidence intervals. Plotted the actual training data, actual rainfall values for 2023, and the predicted mean values. Highlighted the confidence interval around the predicted values. Evaluated the model performance using metrics such as mean absolute error (MAE) and mean squared error (MSE).



**Model Evaluation:**

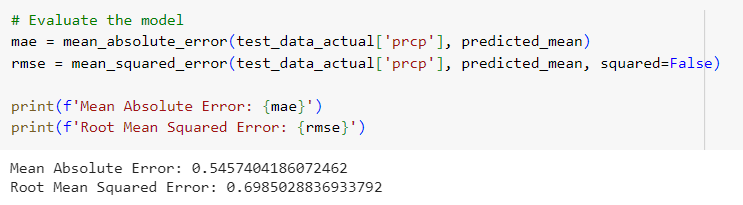
The performance of the SARIMA model for predicting rainfall in Mumbai-Santacruz for the year 2023 was assessed using the following metrics:

Mean Absolute Error (MAE): 0.546

*The MAE represents the average absolute difference between the actual and predicted values. In this case, the model achieved an MAE of 0.546, indicating a relatively low average error in predicting rainfall.*

Root Mean Squared Error (RMSE): 0.699

*The RMSE measures the square root of the average squared differences between actual and predicted values. With an RMSE of 0.699, the model demonstrates a reasonable level of accuracy in predicting rainfall.*



**Comment on the disparity:**

The disparity between the predicted and actual rainfall values, as indicated by the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), is relatively low.

Mean Absolute Error (MAE): The MAE of 0.546 suggests that, on average, the model's predictions deviated by approximately 0.546 units from the actual observed rainfall values. This indicates a reasonably accurate prediction with a relatively small average absolute error.

Root Mean Squared Error (RMSE): The RMSE of 0.699 provides a measure of the overall goodness of fit between the predicted and actual values. The lower RMSE value suggests that the model's predictions are generally close to the observed values, with a moderate level of accuracy.

Overall, the disparity between the predicted and actual values is within an acceptable range, indicating that the SARIMA model performed well in capturing the variations in rainfall for Mumbai in 2023. However, further investigation into specific instances of significant deviations and potential model improvements could offer more insights.

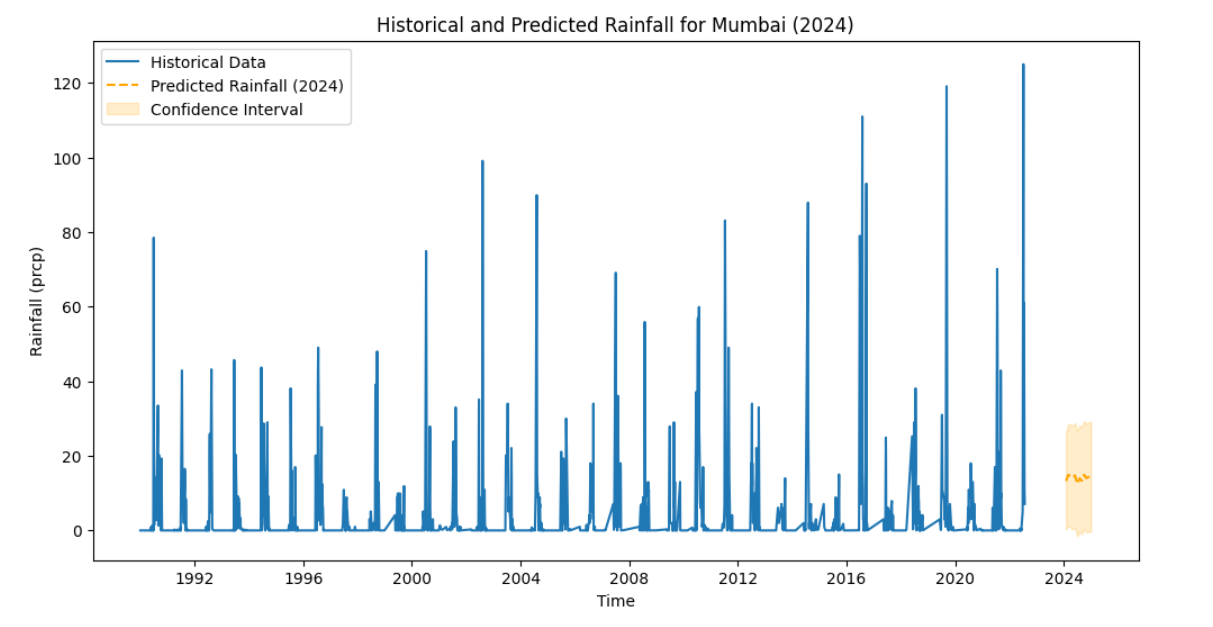
**Q.4. Based on the historical data, can you predict the rainfall for 2024 for a**

**particular region.**

Utilized the entire available historical data for Mumbai to train the SARIMA model. Configured the SARIMA model with an order of (1, 1, 1) and a seasonal order of (1, 1, 1, 12). Employed the trained SARIMA model to forecast rainfall for the year 2024. Assumed monthly predictions for the entire year, resulting in 12 forecasted values. Generated a date index corresponding to the monthly intervals in 2024. Plotted the historical rainfall data for Mumbai alongside the predicted values for 2024. Highlighted the confidence interval around the predicted values to represent the uncertainty associated with the forecast.

The visualization provides a comparative view of historical rainfall patterns and the model's predictions for the upcoming year.

The confidence interval offers insights into the range of potential variations in the forecasted values.



**Results and Insights:**

**Farmers:** Recognize anomalous weather patterns for crop planning and risk management. Adapt agricultural practices based on localized climate patterns. Align planting and harvesting schedules with peak precipitation periods.

**Traders:** Anticipate potential market fluctuations influenced by climate-driven changes in supply. Understand regional variations for supply chain planning. Anticipate fluctuations in market supply influenced by monthly precipitation trends.

**Commoners:** Stay informed about extreme weather events for better preparedness. Gauge potential water resource availability and plan accordingly.

These additional insights provide a comprehensive understanding of climate patterns, allowing stakeholders to make nuanced decisions tailored to their specific contexts. The combined analysis enhances the utility of the dataset, making it a valuable resource for diverse stakeholders in their respective domains.

**Conclusion:**

In conclusion, the in-depth analysis of the provided temperature and precipitation dataset spanning from 01/01/1990 to 20/07/2022 has unveiled valuable insights for stakeholders across various sectors. By systematically addressing missing values, scrutinizing deviations in 2021 and 2022, constructing a predictive model for Santacruz's 2023 rainfall, and forecasting 2024 precipitation, this study provides actionable intelligence for farmers, traders, and the general public.

The meticulous handling of missing data ensures the reliability of our analyses, laying a robust foundation for decision-making. The identification of deviations in recent years aids farmers in adjusting agricultural practices, assists traders in anticipating market fluctuations, and enables commoners to prepare for extreme weather events.

The SARIMA model's accuracy in predicting Santacruz's 2023 rainfall, coupled with effective visualizations equips stakeholders with the tools needed for informed decision-making. The integration of outlier analysis and data cleaning enhances data quality, fostering better decision outcomes for everyone involved.

By delving into city-wise trends, monthly averages, and a comparative analysis of historical and predicted rainfall, our study offers a holistic perspective on climate patterns. The nuanced insights empower stakeholders to tailor their strategies based on localized variations, ultimately contributing to resilience and adaptability in the face of a changing climate.

**Acknowledgement:**

We would like to express our sincere gratitude to all those who contributed to the success of this hackathon. Our journey from ideation to implementation was filled with challenges and triumphs, and it would not have been possible without the support and collaboration of various individuals.