India Climate and Air Quality Analysis

Evanne Chiang | chian104@purdue.edu

04/18/2025

Code can be found here

Updated Version:

https://nb.anaconda.cloud/jupyterhub/user/a4b75cc1-6683-44a8-8f58-994703d60659/notebooks/indiaproject.ipynb

Introduction

Our planet's climate is constantly monitored across the globe due to its wide-reaching effects on ecosystems, human health, and global sustainability. Understanding a country's climate is an essential part of managing resources, preparing for extreme weather, and developing environmental policies. In this report, we focus on the climate of India, a country with diverse geographical features and weather patterns. By examining trends in temperature, rainfall, and air quality, we aim to gain a clearer understanding of how India's climate has changed over time and what factors influence these changes. Although there are many more elements that can affect climate, this report will focus on these three.

Cleaning and Filtering

In order to utilize the data, all datasets needed to be cleaned. To begin the cleaning process, the .CSV files downloaded from Kaggle are loaded into a Pandas DataFrame using 'pd.read_csv()'. Next, specifically for India's rainfall data, I reshaped the data using the 'pivot_table()' function for better analysis. This allowed me to reorganize the data so that each row represents a different year, and each column represents a specific region in India. The index

parameter was set to ‘YEAR’ to place the years as rows, the columns parameter was set to ‘SUBDIVISION’ to display the regions across the top, and the values parameter was set to ‘ANNUAL’ so that the table would populate with annual rainfall amounts. This step simplified the dataset structure greatly, making it easier to compare rainfall trends across regions time.

```
ReshapedData = InitialData.pivot_table(
        index='YEAR',        #sets years as rows
        columns='SUBDIVISION',  #sets regions as columns
        values='ANNUAL',     #values to populate the table
    )
```

Sorting

Sorting was applied primarily to organize the data chronologically and prepare it for analysis. In the rainfall dataset, the *'Year'* column was used to group and sum monthly rainfall values, and then sorted using sort_values to ensure the data followed a consistent timeline. This was essential for visualizing long-term rainfall trends. Similarly, the temperature dataset was grouped by year to calculate average annual temperatures and then sorted to align with the rainfall data for merging and correlation analysis. By sorting both datasets by year, we ensured accurate comparisons and consistent scatter plot sequencing.

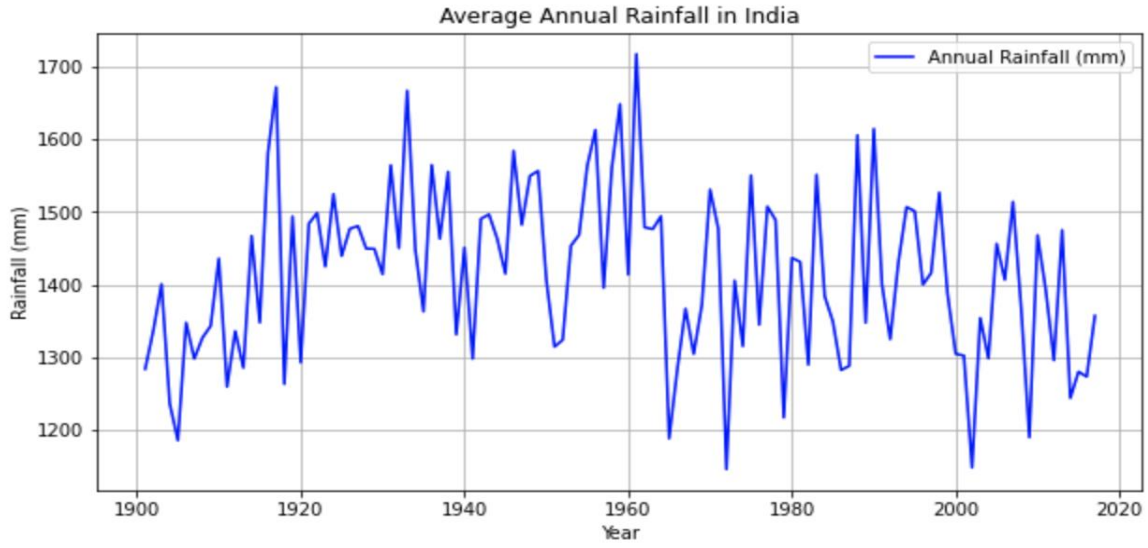
Merging

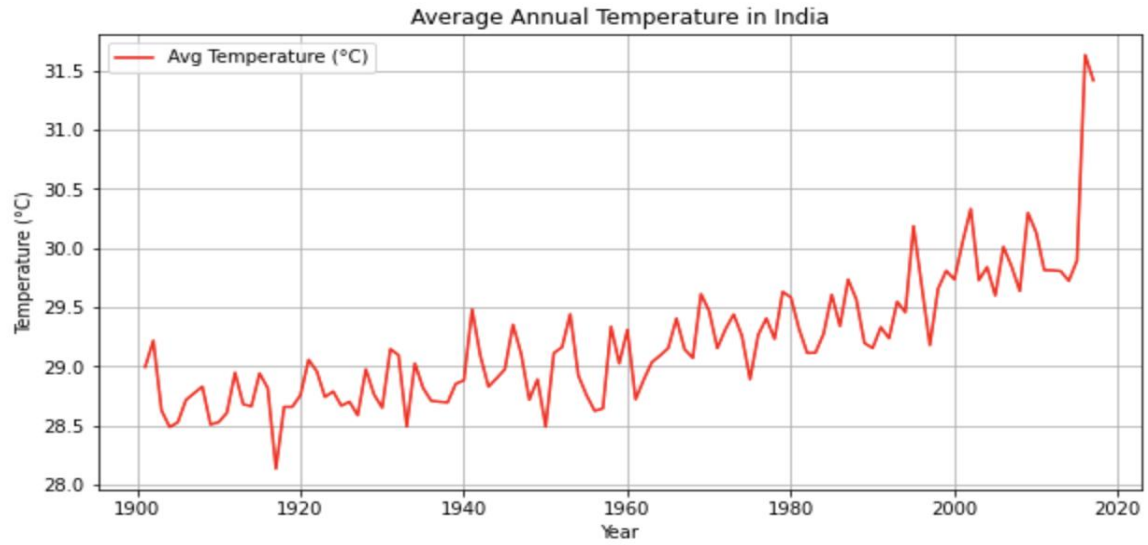
To facilitate integrated analysis, multiple datasets including temperature, rainfall, and air quality were merged based on matching years. The annual average temperature and the total

annual rainfall were first calculated and merged into a new DataFrame using the year as a common key. This allowed for a direct comparison of climate variables within the same time frame. Since the air quality data only covered 2021, it was separately filtered and merged with temperature data from the same year to enable one year comparison. The process ensured that each merged dataset had aligned temporal values, allowing for accurate side by side visualizations and correlation studies.

Visualization

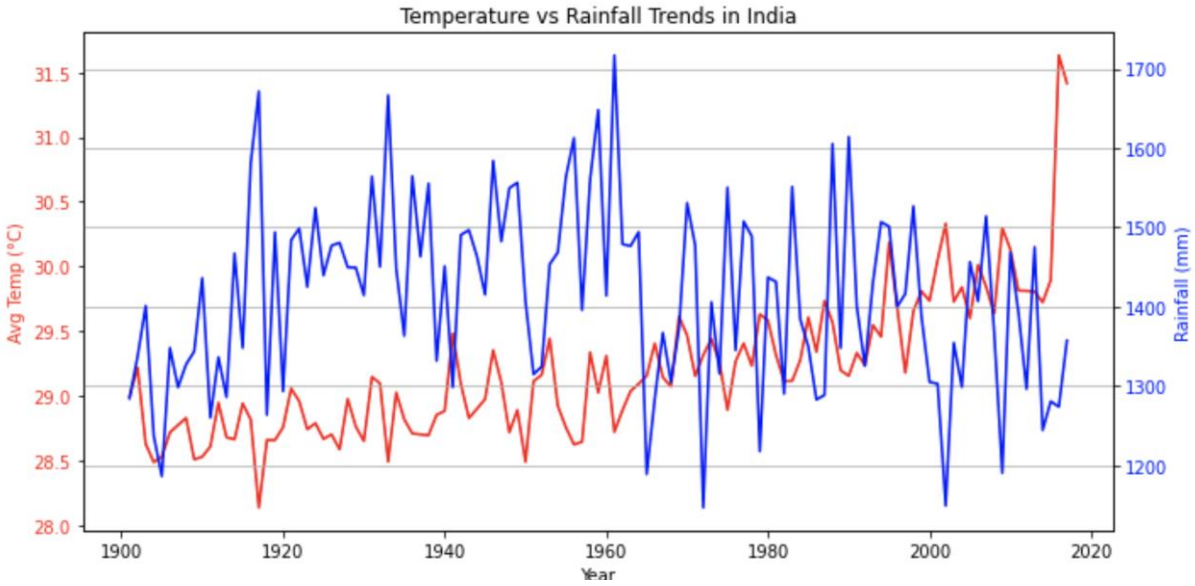
The first visualization we created aimed to explore how average annual temperature and rainfall have changed over time in India. We used two datasets: one on monthly temperatures and the other on monthly rainfall across various regions in India. For each year, we computed the national average temperature by averaging values from January to Decembe



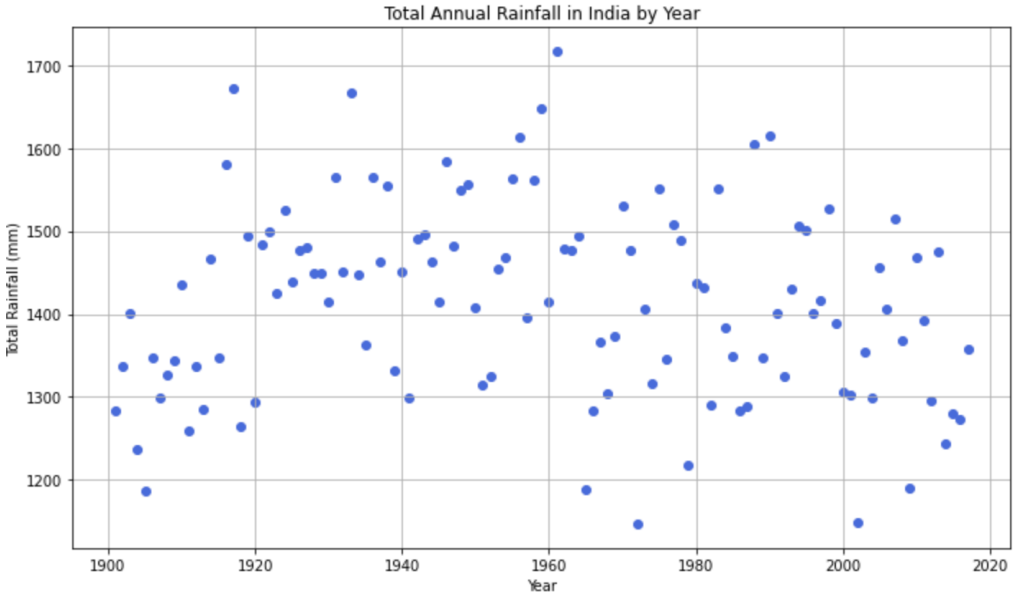


The average temperature in India shows a steady upward rise over the years, whereas rainfall patterns appear more irregular and fluctuating.

The combined chart makes this contrast clearer by displaying both trends on the same timeline, highlighting the lack of a strong visual correlation between rising temperatures and rainfall totals.

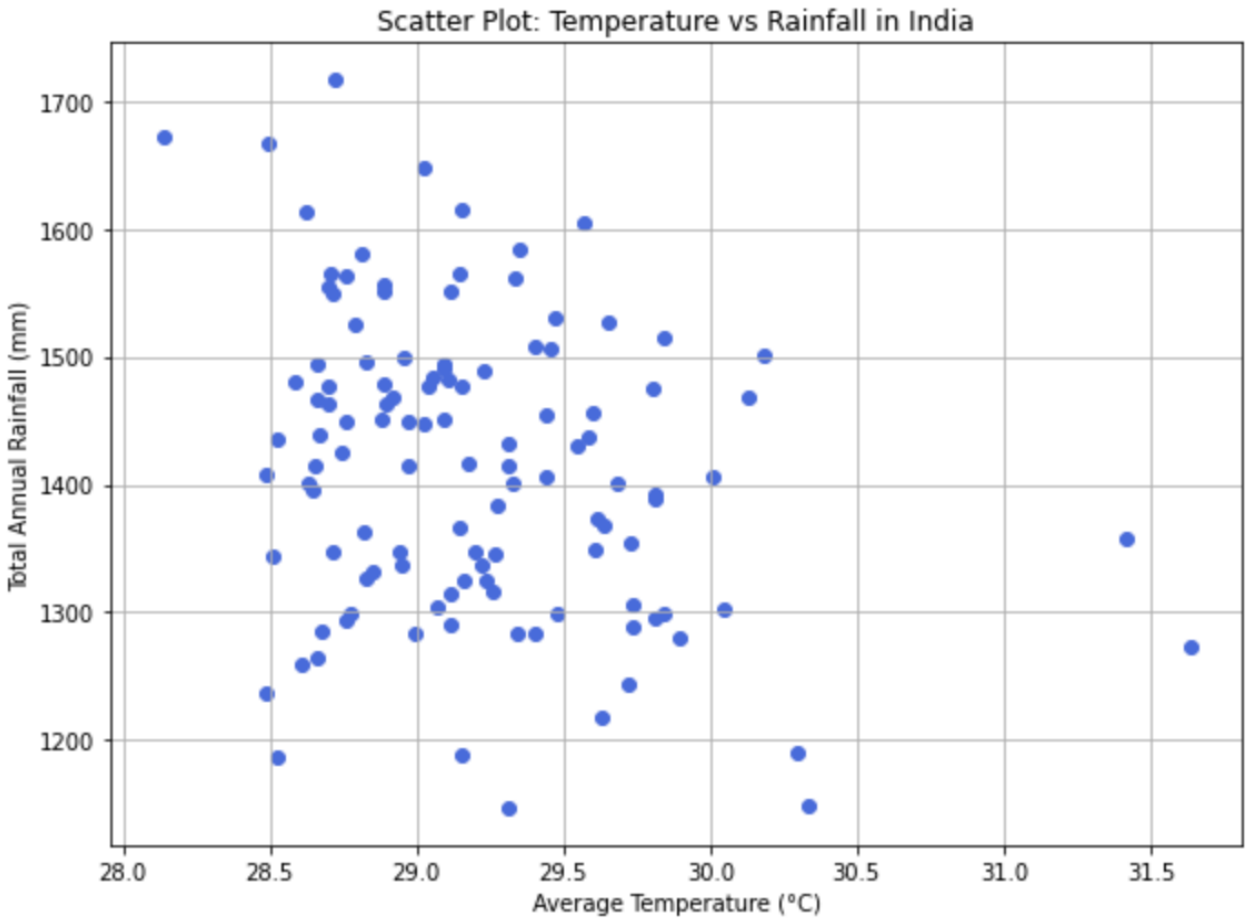


The second visualization focuses solely on India's total annual rainfall trends across the years available in the dataset. For each year, we calculated the national average of total rainfall by summing all monthly values from January through December across various regions in the country. These yearly totals were then visualized using a scatter plot, where each point represents one year's rainfall total.

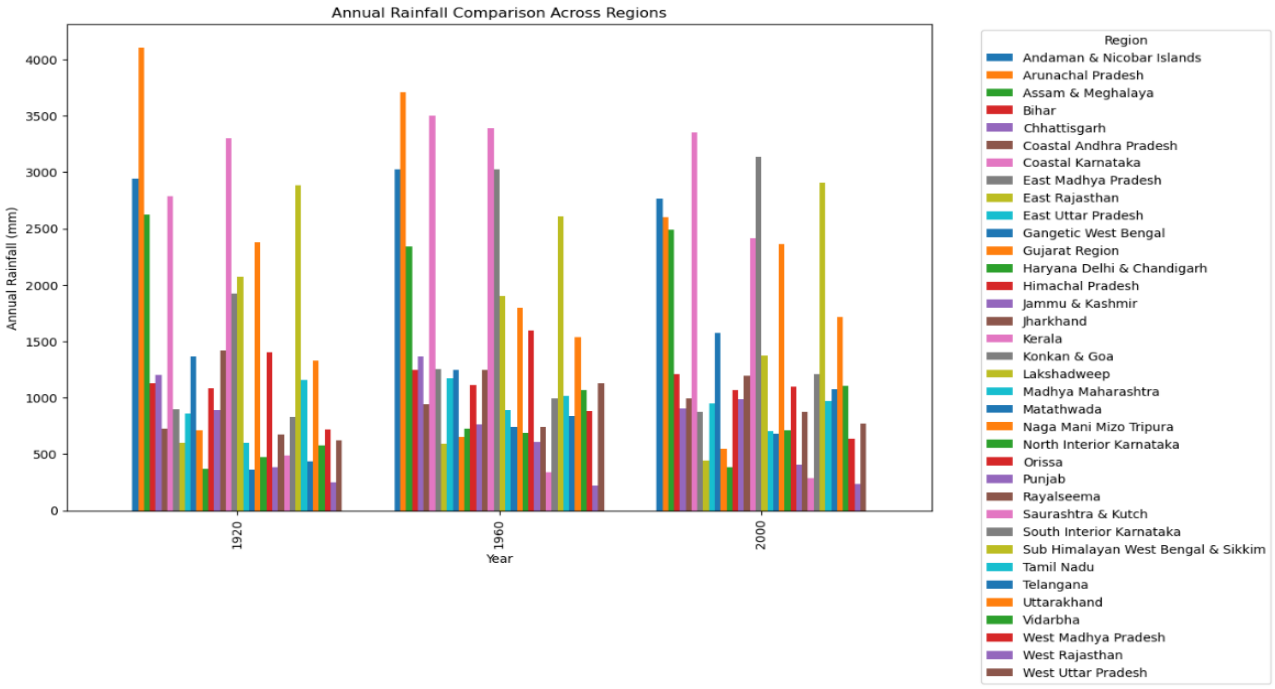


The first chart plots the total annual rainfall in India over the years, revealing natural variation and some fluctuations without a clear long-term trend. The second scatter plot compares total annual rainfall with average annual temperature.

Each point represents a year's data, helping identify if hotter years tend to be wetter or drier. While there's no strong visual correlation, this view allows for deeper exploration of possible climate interactions.

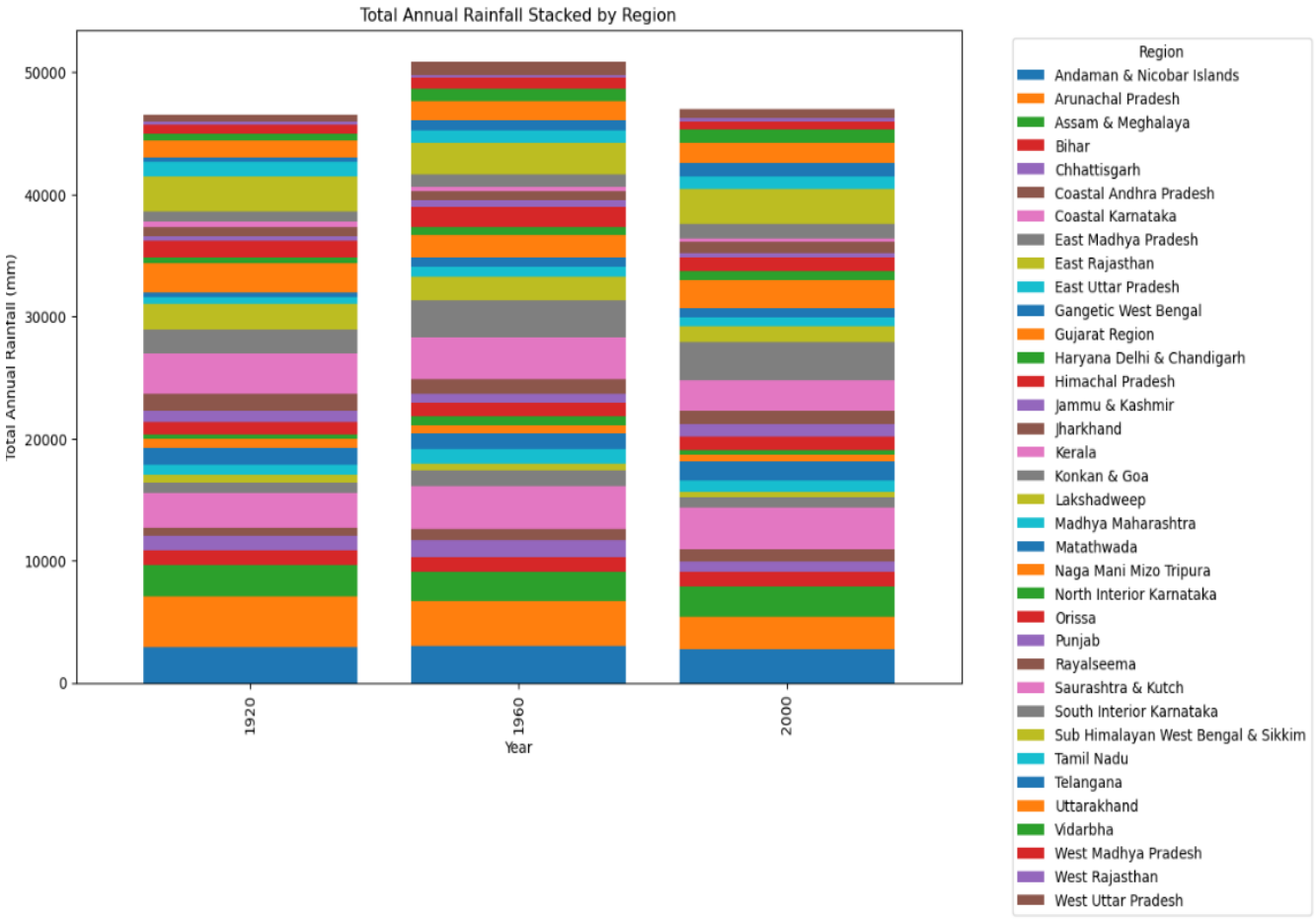


The third visualization compares how annual rainfall is distributed across different regions in India for the years 1920, 1960, and 2000. It includes two charts that together highlight both regional variation and overall national contribution.



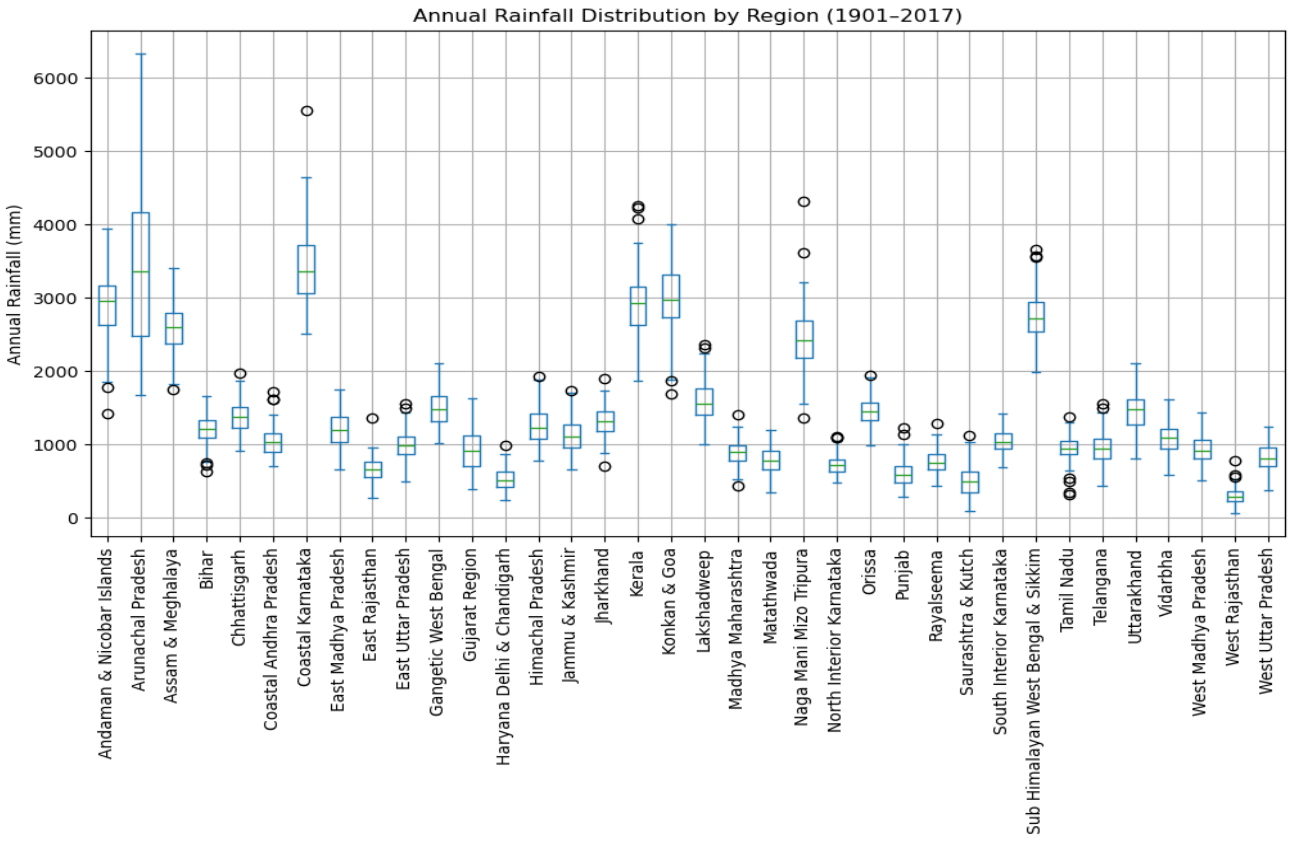
The first chart is a grouped bar chart, where each color represents a different region. This allows for a side-by-side comparison of how much rainfall each region received in the selected years.

The second chart aggregates the total annual rainfall for each of the three years, stacking the contributions of all regions within each bar. This view makes it easy to see which regions contribute the most to India's total rainfall and whether those contributions shift over time.



While the grouped chart emphasizes differences between individual regions, the stacked chart focuses on how each region's share compares within the national total, offering two perspectives on regional rainfall trends across decades.

The fourth visualization examines the distribution of annual rainfall across different regions in India from 1901 to 2017, using box plots to capture variability, medians, and outliers.

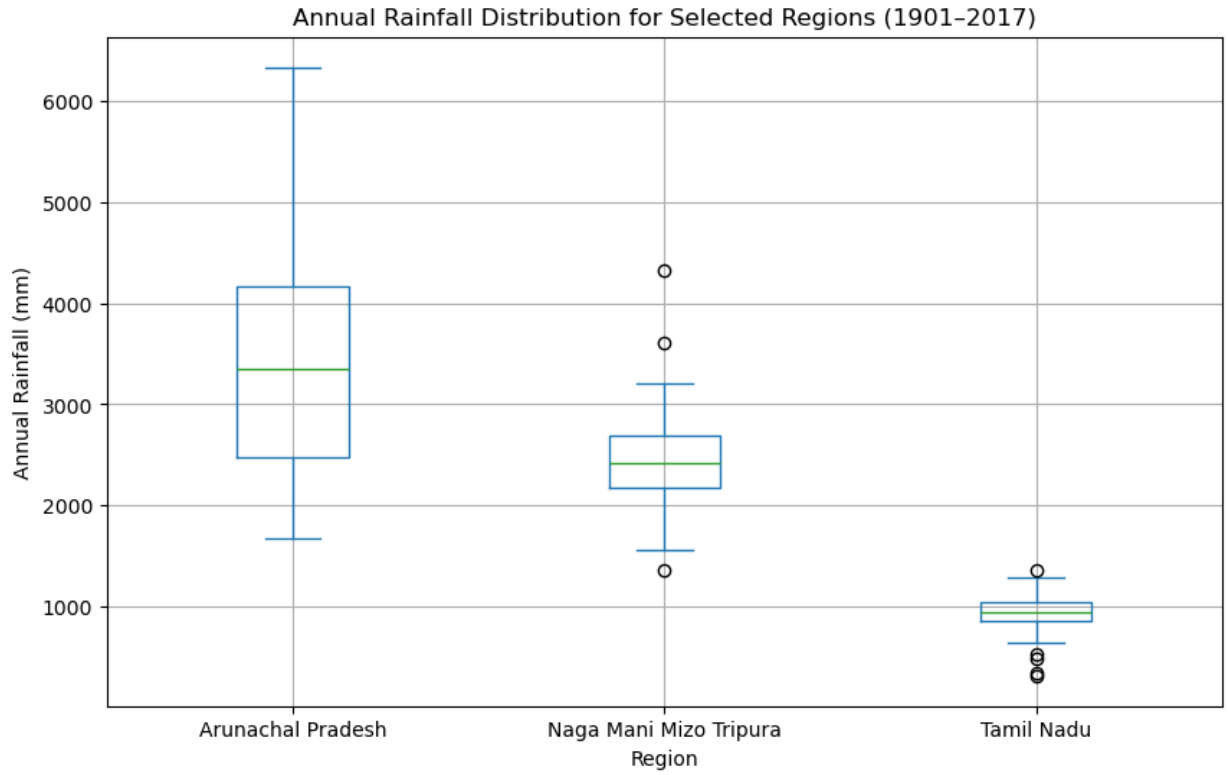


The first chart displays box plots for all regions, offering a side-by-side comparison of rainfall distribution across the country. Each box represents the interquartile range for that region, while the lines and dots show the full spread of data and outliers.

The second chart zooms in on three selected regions: Arunachal Pradesh, Naga Mani Mizo Tripura, and Tamil Nadu, to allow for a more detailed comparison.

Arunachal Pradesh shows both the highest median and the widest spread, while Tamil Nadu displays a much lower and tighter distribution. Naga Mani Mizo

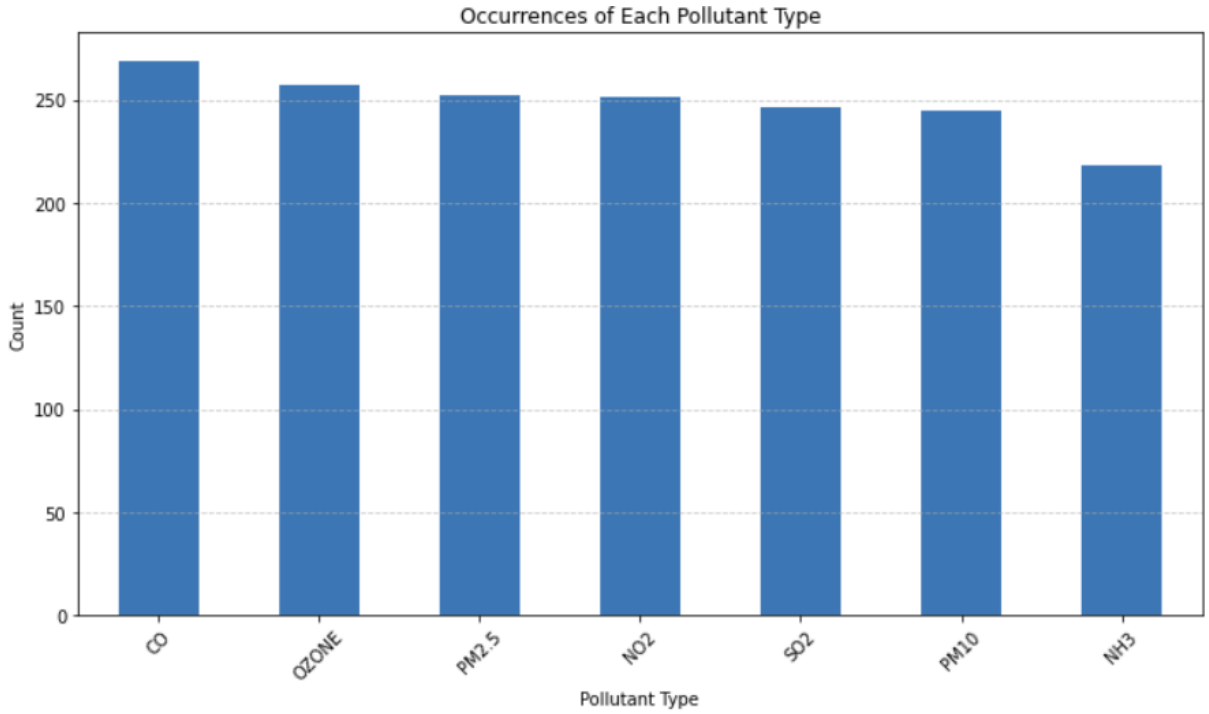
Tripura lies somewhere in between, showing moderate rainfall with a few notable outliers.



Together, these charts highlight how rainfall variability differs greatly across regions, both in terms of overall amount and consistency, emphasizing the importance of region-specific analysis when studying climate trends in India.

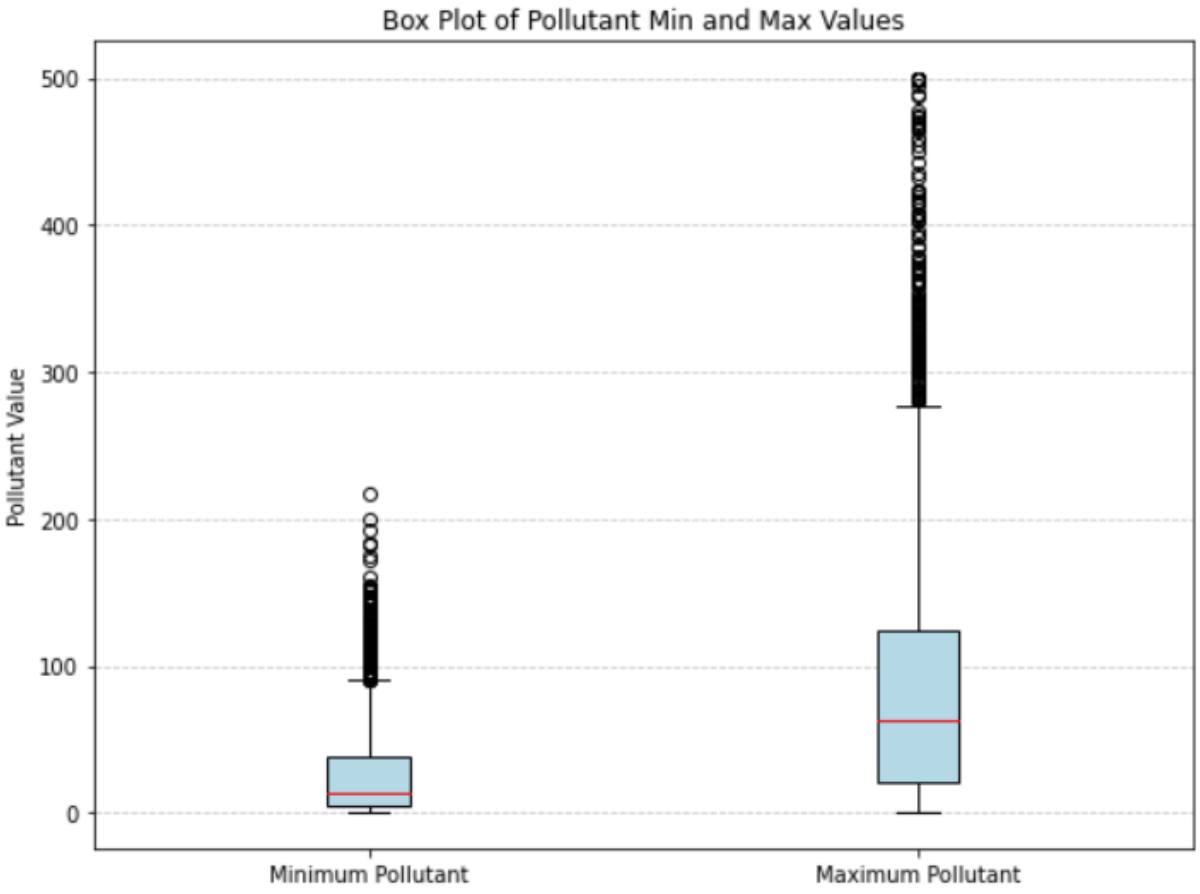
The fifth visualization shifts focus from climate patterns to air quality by examining the distribution and variability of major air pollutants recorded in India.

The first chart is a bar graph showing the total number of occurrences for each pollutant type across the dataset. It includes seven key pollutants: CO, Ozone, PM2.5, NO2, SO2, PM10, and NH3. The frequencies are fairly balanced, with carbon monoxide (CO) being the most frequently detected and ammonia (NH3) the least. This gives a sense of how commonly each pollutant appears in monitoring records.



The second chart is a box plot of the minimum and maximum number of pollutants detected, offering insight into how many pollutants were typically present in the air, while also highlighting any outliers.

The distribution of maximum values shows significant variation, indicating occasional spikes in pollution levels, whereas minimum values remain more stable, reflecting baseline air quality conditions.



Stacking and Pivoting

Stacking and pivoting both played a key role in filtering the data. For example, we had to use the “pivot_table()” function to help utilize the data for the rainfall in India. In addition, when cleaning the data involving the air quality, we also used the “stack()” function to sort the data. This made the data easier to understand and manipulate because

it allowed us to convert a long data structure into a wide data structure. These resources helped us comprehend the data much easier, and they improved the efficiency of the data wrangling and visualization process.