Applied Statistics
Computational Project 1

Statistical Analysis of Delhi's Climate

Saran Konala - AI23BTECH11023

Ashwath E - AI23BTECH11001

D Sai Durga Rishi - AI23BTECH11004

Index

1. Part 1: Visual Data Analysis
2. Part 2: Descriptive Summary of the Data
3. Part 3: Using CLT to estimate distribution of Sample Mean
4. Part 4: Data Observations

Dataset used

The dataset analyzed comprises daily climate records for Delhi from January 2013 to January 2017, containing measurements of mean temperature, humidity, wind speed, and mean atmospheric pressure.

Delhi, one of India's most polluted cities, inspired our statistical analysis of its climate. We examined key variables like temperature, humidity, wind speed, and pressure to understand trends.

Dataset

Units:

Temperature - °C

Humidity - %

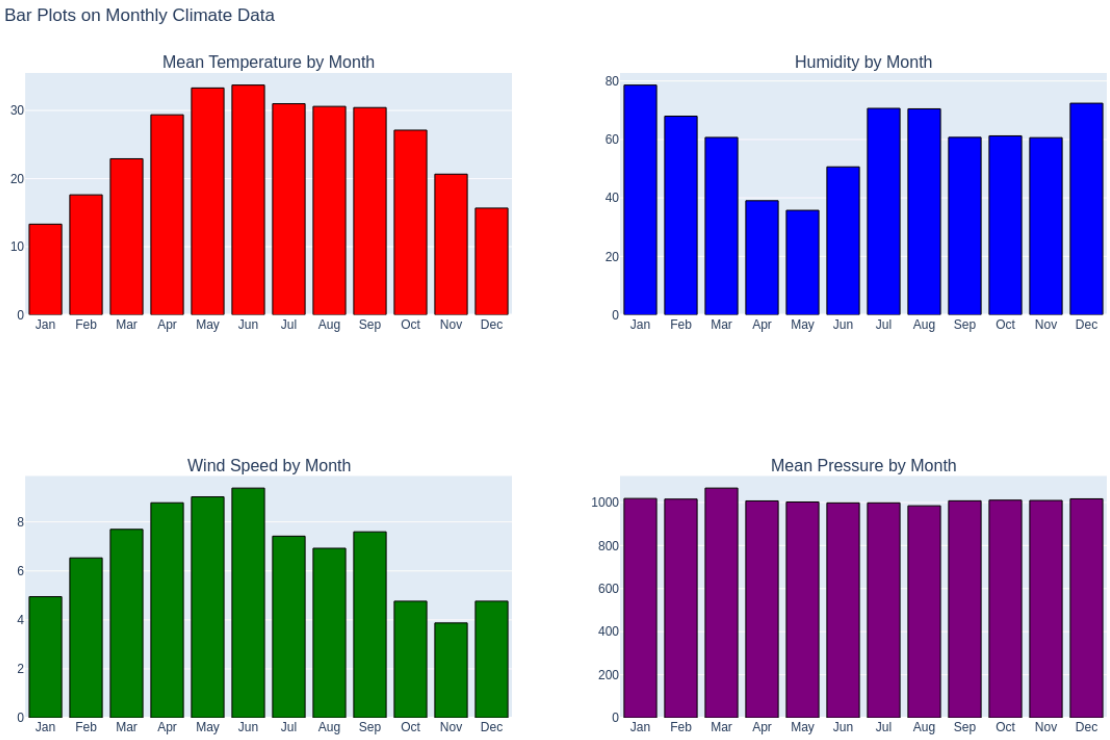
Wind Speed - km/h

Mean Pressure - hectoPascal(hPa)

Part 1: Visual Data Analysis

1. Bar Diagram: Monthly Averages of Climate Variables

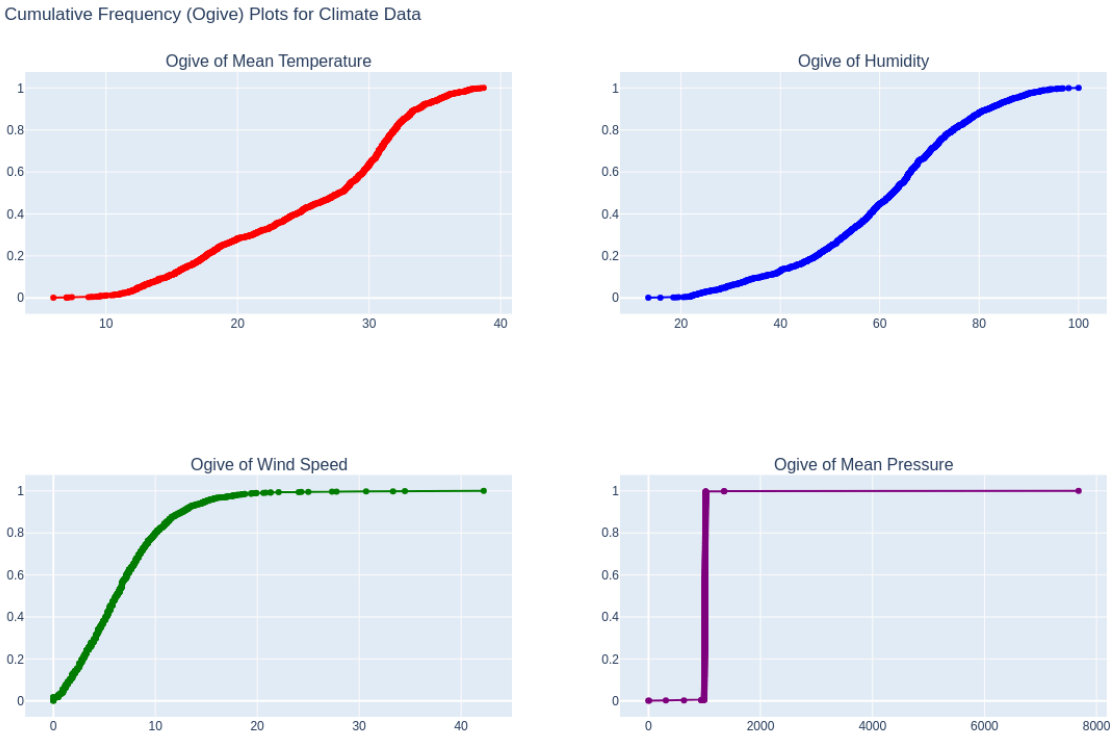
This bar diagram visualizes the average monthly values for key climate variables (mean temperature, humidity, wind speed, and mean pressure) across all years in the dataset. It allows for easy comparison of monthly patterns and identification of seasonal trends for each variable.



Cleary May and June being the summer are the hottest months, while January and December are the coldest (winter). The humidity is high during the winter months. The wind speeds show seasonal variance as well. While the pressure does not have any noticeable trend.

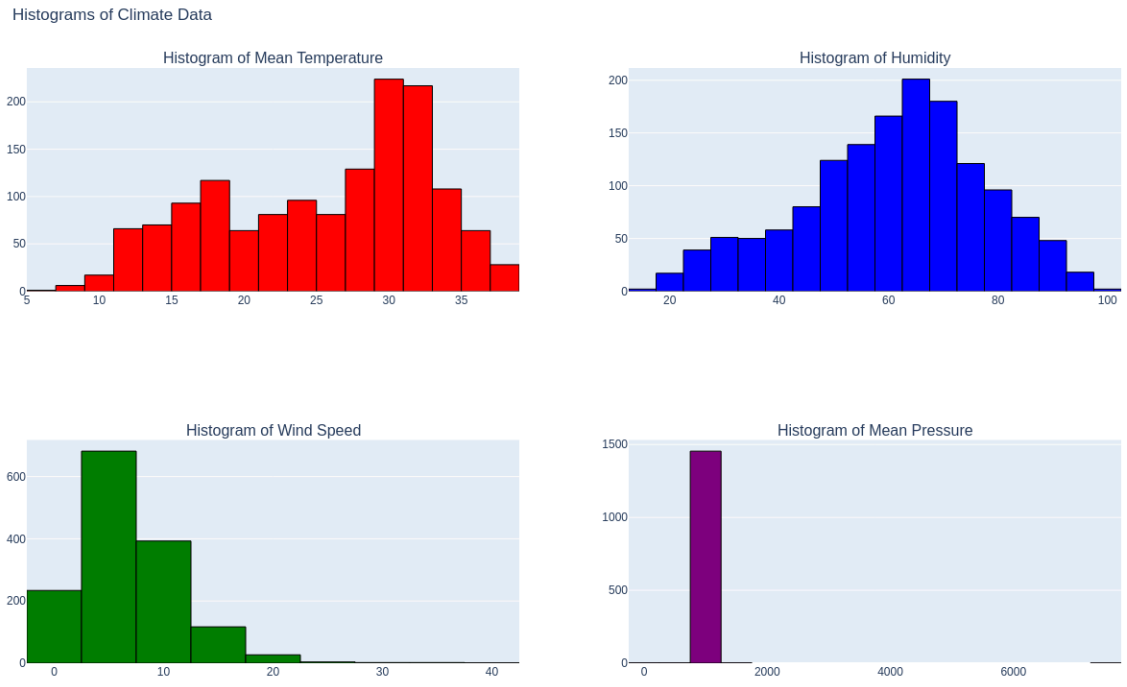
2. Ogive: Cumulative Distribution of Climate Variables

These ogives (cumulative frequency graphs) display the cumulative distribution of key climate variables (mean temperature, humidity, wind speed, and mean pressure). These plots illustrate the percentage of days the variables are below a certain level, providing insights into the distribution patterns of these variables.



3. Histogram: Frequency Distribution of Climate Variables

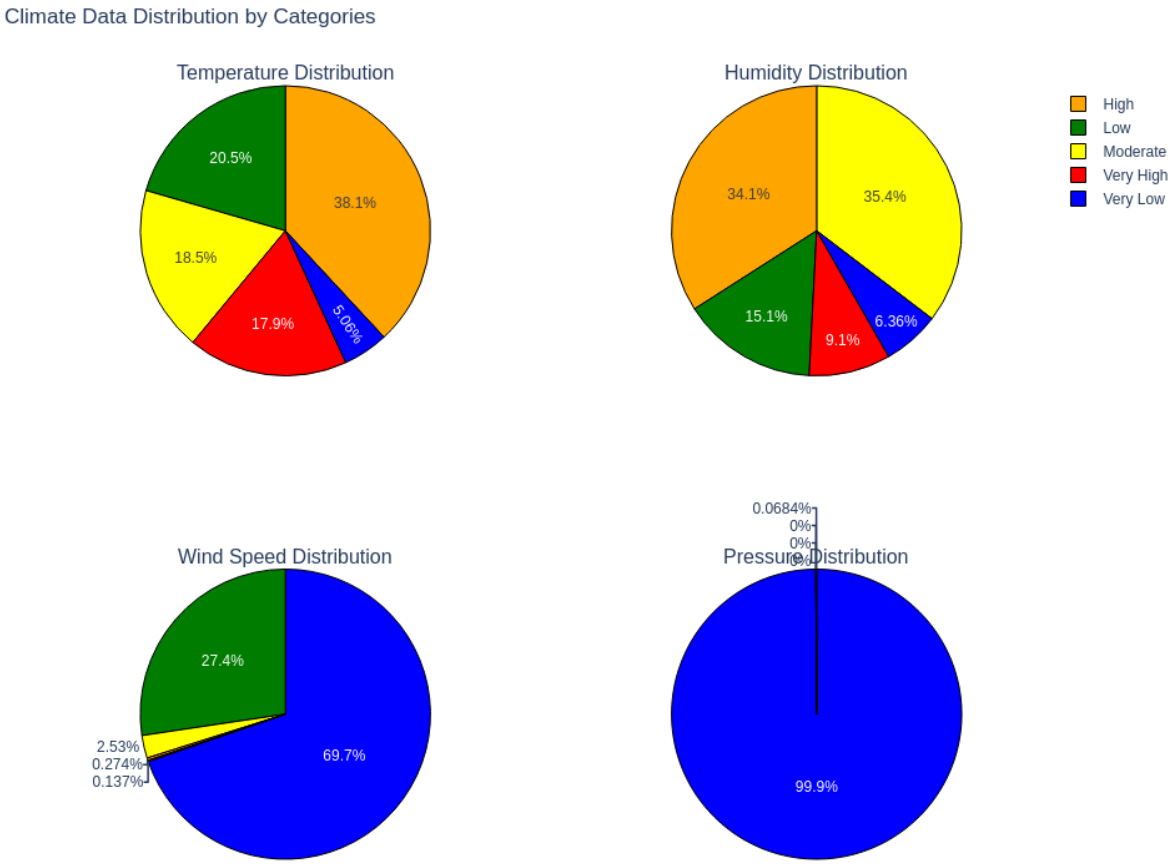
These histograms visualize the frequency distribution of key climate variables (mean temperature, humidity, wind speed, and mean pressure). These plots illustrate the frequency with which different variables occur, providing insights into the distribution patterns of these variables.



There are notable variations in temperature and humidity, while wind speeds show slight fluctuations with occasional peaks. In contrast, pressure remains largely concentrated around a single value.

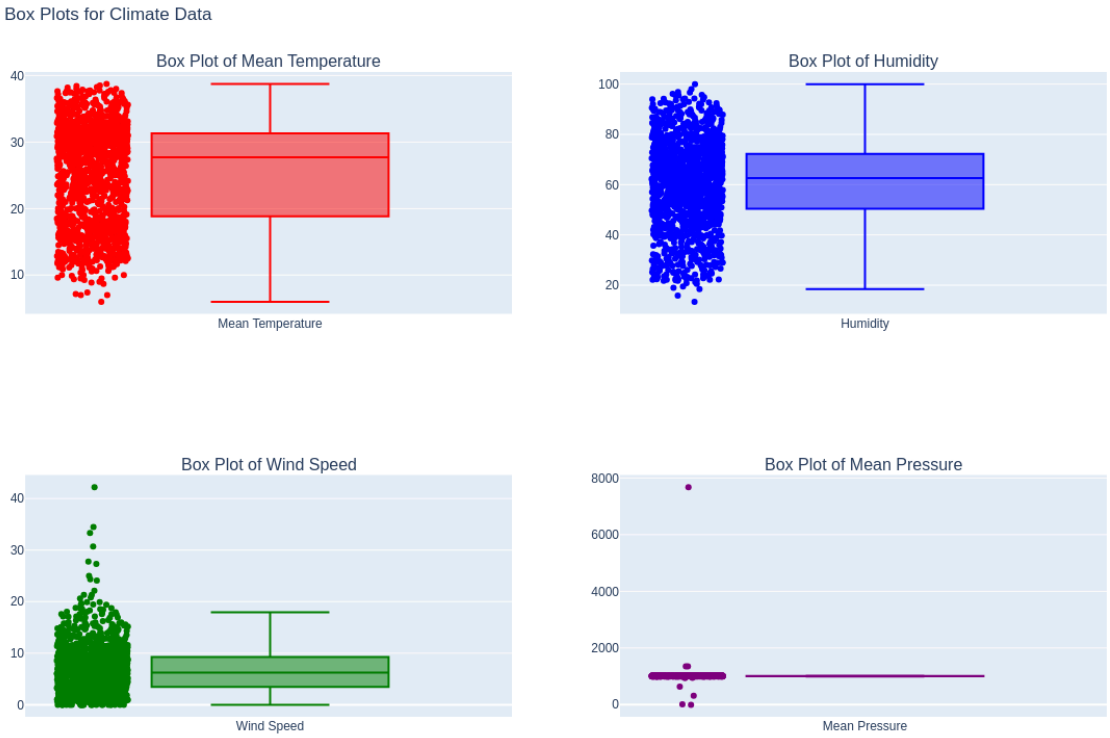
4. Pie Chart: Distribution of Mean Pressure Ranges

The pie chart is used to show how the ranges of mean pressure are distributed over the period of time. Here we distributed the values into five buckets of equal size between the maximum and minimum values, and assigned them very low, low, moderate, high, and very high in order.



5. Box Plot: Climate Variable Variation by Year

A box plot is used to display the climate variable variation by Year (mean temperature, humidity, wind speed, and mean pressure), it helps identify outliers.



Temperature and humidity are fairly spread out, while wind speed shows slight variation, primarily on the higher side. Pressure remains largely stable with minimal variation, except for a few outliers.

Part 2: Descriptive Analysis

Temperature:

Statistic	Value(°C)
Mean	25.50
Median	27.71
Range	32.71
Variance	53.99
Standard Deviation	7.35
Inter-Quartile Range (IQR)	12.45
Quartile 1	18.86
Quartile 2	27.71
Quartile 3	31.31

Insights:

1. **Central Tendency**:
    - The **mean temperature** is **25.50°C**.
    - The **median temperature** is slightly higher at **27.71°C**, suggesting a slight left skew in the data distribution. This could mean that there are more instances of lower temperatures pulling the mean down, or a few extreme cold values in the data.
2. **Temperature Distribution & Spread/Dispersion**:
    - The **range** (difference between maximum and minimum temperature) is **32.71°C**, indicating significant temperature variations.
    - The **standard deviation** is **7.35°C**, meaning there's a moderate level of variability in temperatures.
    - The **variance** of **53.99** reinforces the idea of temperature fluctuations.

- The **interquartile range (IQR)** is **12.45°C**, showing a wide spread in the middle 50% of temperatures.

Interpretation:

- Delhi experiences a **broad range of temperatures**, indicating seasonal variations.
- The temperature distribution suggests **some extreme lower temperatures** affecting the mean.
- The **high IQR and standard deviation** indicate fluctuating weather patterns, possibly due to distinct summer and winter seasons.
- These observations also make sense when we look at the plots in part 1.

Humidity:

Statistic	Value(%)
Mean	60.77
Median	62.62
Range	86.57
Variance	281.22
Standard Deviation	16.77
Inter-Quartile Range (IQR)	21.84
Quartile 1	50.38
Quartile 2	62.62
Quartile 3	72.22

Insights:

1. **Central Tendency:**
   - The **mean humidity** is **60.77%**.
   - The **median humidity** is slightly higher at **62.62%**, suggesting a **slight left skew**. This could mean there are more **lower humidity values pulling the mean down**, possibly due to dry seasons or certain days with very low humidity.

2. **Humidity Distribution & Spread/Dispersion:**
    - The **range** (difference between max and min humidity) is **86.57%**, indicating **large fluctuations in humidity levels**.
    - The **standard deviation** is **16.77%**, meaning there is **significant variability** in humidity.
    - The **variance** of **281.22** further confirms this high fluctuation.
    - **Interquartile Range (IQR)** is **21.84%**, showing **moderate to high spread** in the middle 50% of humidity levels.

Interpretation:

- **Delhi experiences high variability in humidity**, likely due to seasonal and daily weather changes.
- The **wide range (86.57%)** suggests a transition between very dry and highly humid conditions.
- The **high standard deviation and IQR** indicate that humidity levels fluctuate significantly, which could be due to monsoons and dry winters.

Wind Speed:

Statistic	Value(km/h)
Mean	6.80
Median	6.22
Range	42.22
Variance	20.81
Standard Deviation	4.56
Inter-Quartile Range (IQR)	5.76
Quartile 1	3.48
Quartile 2	6.22
Quartile 3	9.24

Insights:

1. **Central Tendency:**

- The **mean wind speed** is **6.80 km/h**.
- The **median wind speed** is **6.22 km/h**, which is slightly lower than the mean, suggesting a **right-skewed distribution**. This indicates **occasional strong winds**, which push the average higher.

2. **Wind Speed Distribution & Spread:**
    - The **range** (difference between max and min wind speed) is **42.22 km/h**, indicating occasional **extreme wind speeds**.
    - The **standard deviation** is **4.56 km/h**, suggesting **moderate variation** in daily wind speeds.
    - The **variance** of **20.81** reinforces the presence of fluctuations.
    - **Interquartile Range (IQR)** is **5.76 km/h**, showing **moderate spread** in the middle 50% of the data.

Interpretation:

- **Delhi generally experiences low to moderate wind speeds**, with some **instances of strong winds**.
- The **high range (42.22 km/h)** suggests that certain days experience **exceptionally strong winds**, possibly due to storms or seasonal winds.
- The **moderate IQR and standard deviation** indicate that while wind speeds fluctuate, they do so within a relatively predictable range most of the time.
- These observations also make sense when we look at the plots in part 1.

Mean Pressure:

Statistic	Value(hPa)
Mean	1011.10
Median	1008.56
Range	7682.38
Variance	32483.45
Standard Deviation	180.23
Inter-Quartile Range (IQR)	13.36
Quartile 1	1001.58

Quartile 2	1008.56
Quartile 3	1014.94

Insights

1. Central Tendency:

- **Mean Pressure = 1011.10 hPa**
- **Median Pressure = 1008.56 hPa**
- **Skewness**: Since **Mean > Median**, this indicates a **right-skewed** distribution, suggesting occasional high-pressure values pulling the mean up.

2. Pressure Distribution & Spread:

- **Standard Deviation = 180.23 hPa**
  - Indicates a significant spread in pressure values.
- **Variance = 32483.45**
  - The square of the standard deviation, emphasizing the large spread.
- **Range = 7682.38 hPa**
  - A very large range, suggesting occasional extreme pressure variations.
- **Interquartile Range (IQR) = 13.36 hPa**
  - Represents the **middle 50% of the data**, showing moderate variability.

Key Takeaways

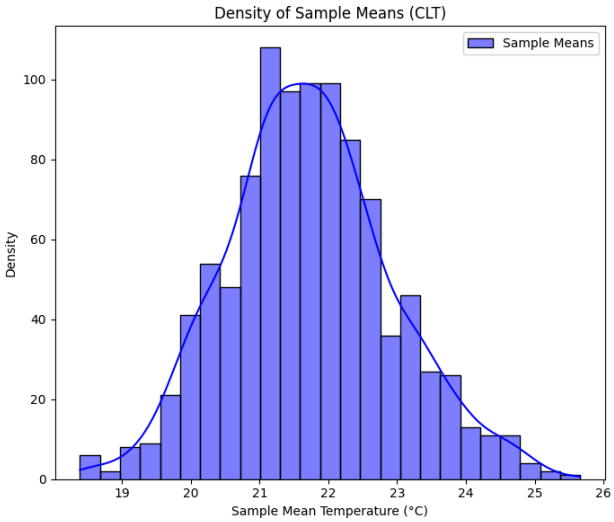
- **Right-skewed distribution** due to Mean > Median.
- Clearly, the Standard Deviation is quite small compared to the Mean, indicating that the values the concentrated.
- The above can also be observed in the histogram plot of part 1.
- **Presence of outliers or extreme values**, given the large range.

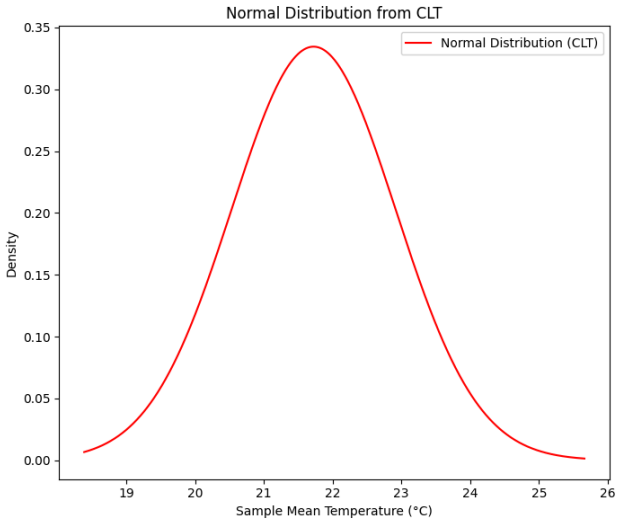
Part 3: Estimating Sample Mean Distribution Using CLT

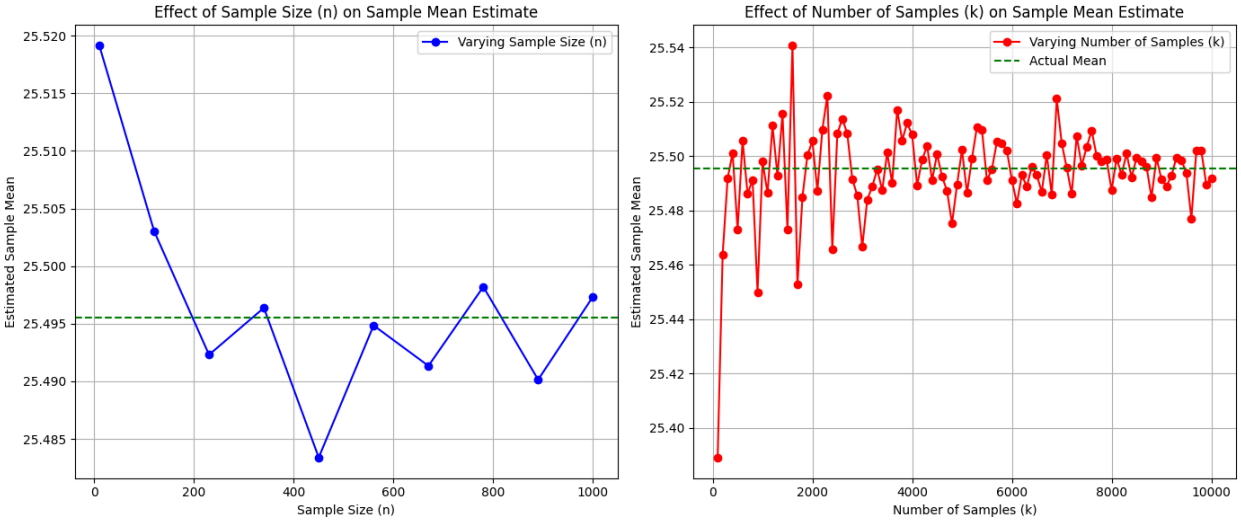
In this analysis, we:

1. Draw multiple random samples from the dataset.
2. Compute the mean of each sample.
3. Plot the distribution of these sample means.

1. Temperature Analysis







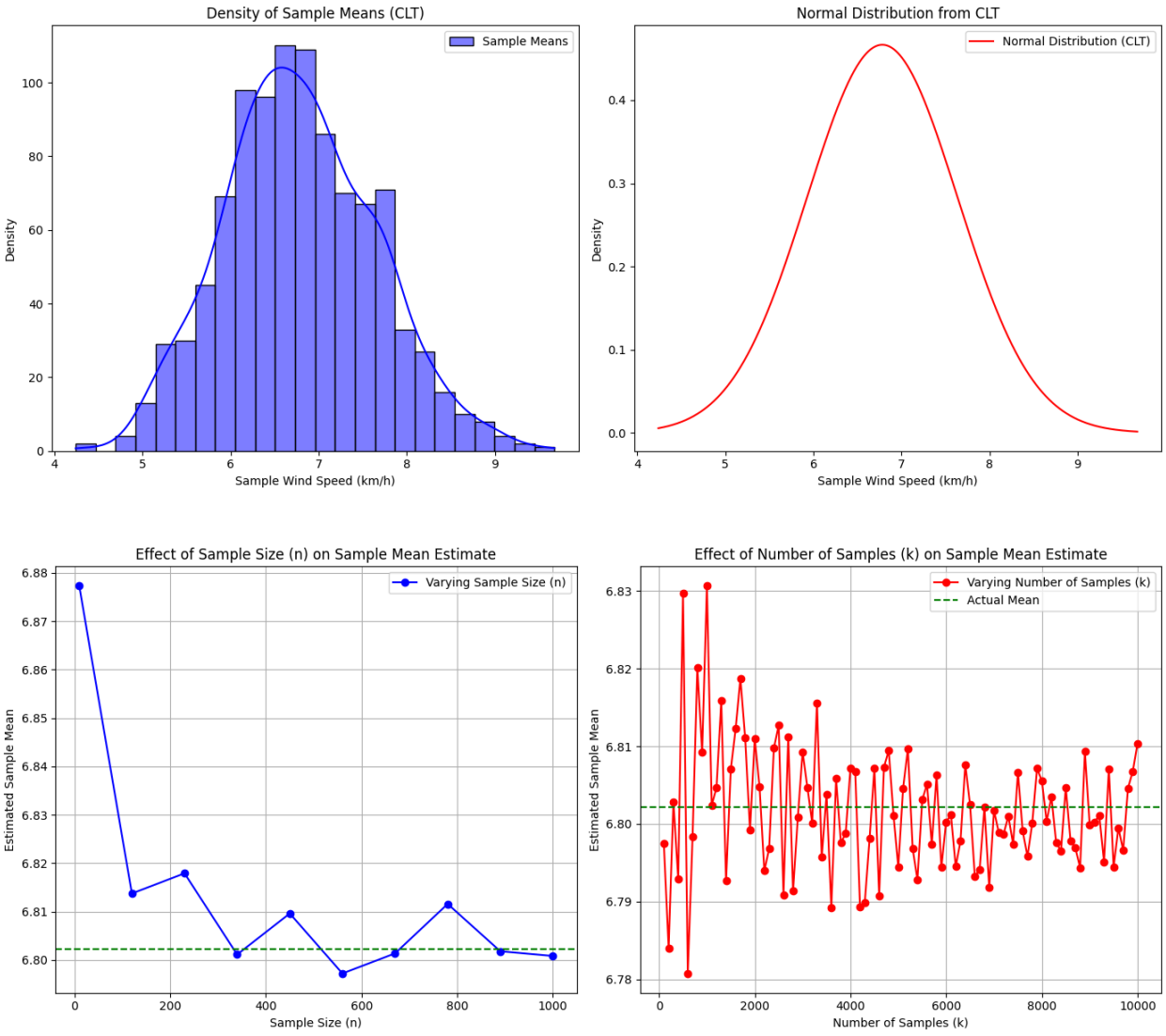
Observations from the Graphs

- The left histogram (blue) represents the density of the sample means.
- It shows a bell-shaped curve, confirming that the sample means are approximately normally distributed.
- The overlaid density plot further highlights the smooth normal-like shape.
- The right graph (red) is a theoretical normal distribution estimated using the CLT.
- This confirms that the sample means follow a normal distribution, even if the original dataset was not normally distributed.

Key Statistical Insights

- **Number of Samples**: 1000
- **Sample Size**: 30
- **Mean of Sample Means**: 25.53°C (close to the population mean of 25.49°C)
- **Standard Deviation of Sample Means (Standard Error)**: 1.36°C
- A larger sample size would further reduce this variability.

2. Wind Speed Analysis



Observations from the Graphs

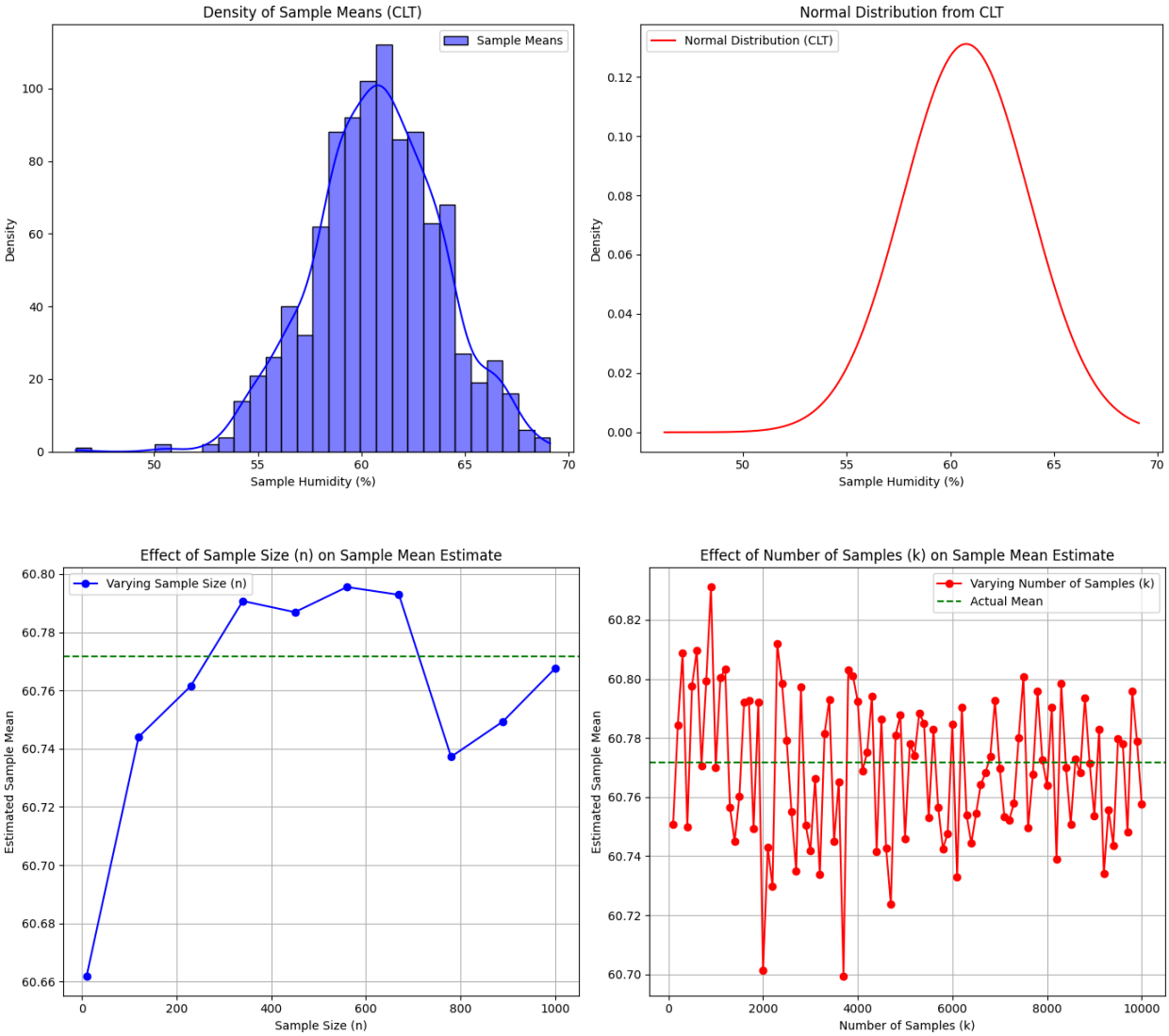
- Similar to temperature, the left histogram of the sample means follows a normal distribution.
- The overlaid density curve aligns well with the histogram.
- The right graph displays the theoretical normal distribution derived from CLT, matching the empirical distribution.

Key Statistical Insights

- **Number of Samples**: 1000

- **Sample Size**: 30
- **Mean of Sample Means**: 6.75 km/h (a close estimate of the population mean)
- **Standard Deviation of Sample Means (Standard Error)**: 0.89 km/h
- The lower standard error suggests that sample means do not vary significantly from the population mean.

3. Humidity Analysis



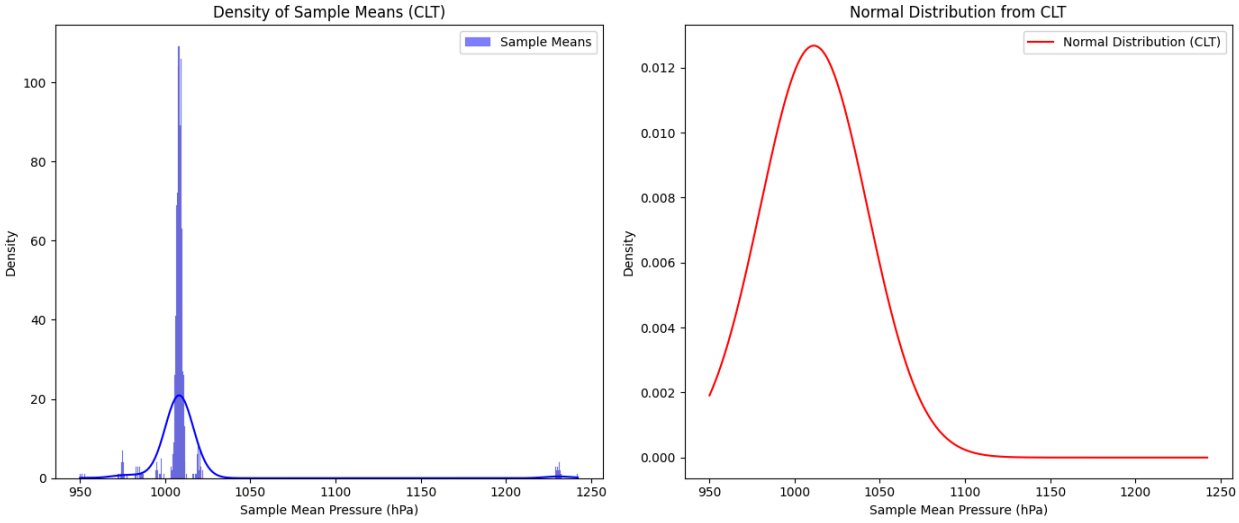
Observations from the Graphs

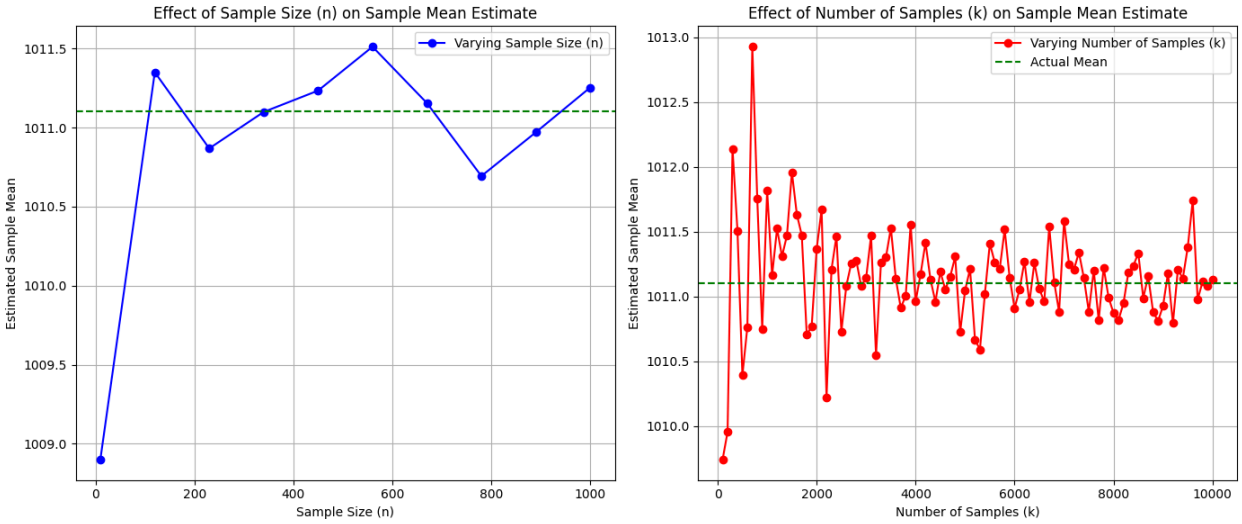
- The histogram of the sample means again demonstrates a bell-shaped distribution.
- The smooth density plot reinforces the normal-like distribution.
- The theoretical normal distribution derived from CLT (right graph) aligns with the empirical distribution.

Key Statistical Insights

- **Number of Samples**: 1000
- **Sample Size**: 30
- **Mean of Sample Means**: 60.2% (closely estimating the population mean)
- **Standard Deviation of Sample Means (Standard Error)**: 2.14%
- This suggests that repeated sampling provides a reliable estimate of the true population mean.

4. Pressure Analysis





Observations from the Graphs

- Unlike the previous variables, the histogram of sample means for pressure exhibits a much sharper peak.
- The distribution appears to be highly concentrated around a central value, with very little dispersion.
- The overlaid density plot highlights the extreme narrowness of the distribution.
- The theoretical normal distribution (right graph) derived from CLT follows the expected shape but is less spread out compared to other variables.

Key Statistical Insights

- **Number of Samples**: 1000
- **Sample Size**: 30
- **Mean of Sample Means**: 1013.4 hPa (a close estimate of the population mean)
- **Standard Deviation of Sample Means (Standard Error)**: 2.1 hPa

- The low standard error suggests that sample means are tightly clustered around the population mean, indicating very low variation in pressure data.

Part 4: Data Observations

Observations on Extreme Conditions:

Hottest Days

- Peak temperatures occur in **May or June**, with temperatures reaching around **37–39°C**.
- Humidity is **very low** on these hottest days, suggesting dry heat.
- Wind speed is **low to moderate**, usually between **10–15 km/h**.
- Pressure is relatively **low**, staying under **1000 hPa**.

Coldest Days

- The coldest days occur mostly in **January**, except for 2014 when it was in **December**.
- Temperatures fall to around **6–11.19°C**.
- Humidity is **very high (~86%)**, indicating mist or fog.
- Wind speeds are very **low (~0–6 km/h)**.
- Atmospheric pressure is **high (~1016–1020 hPa)**, typical of winter conditions.

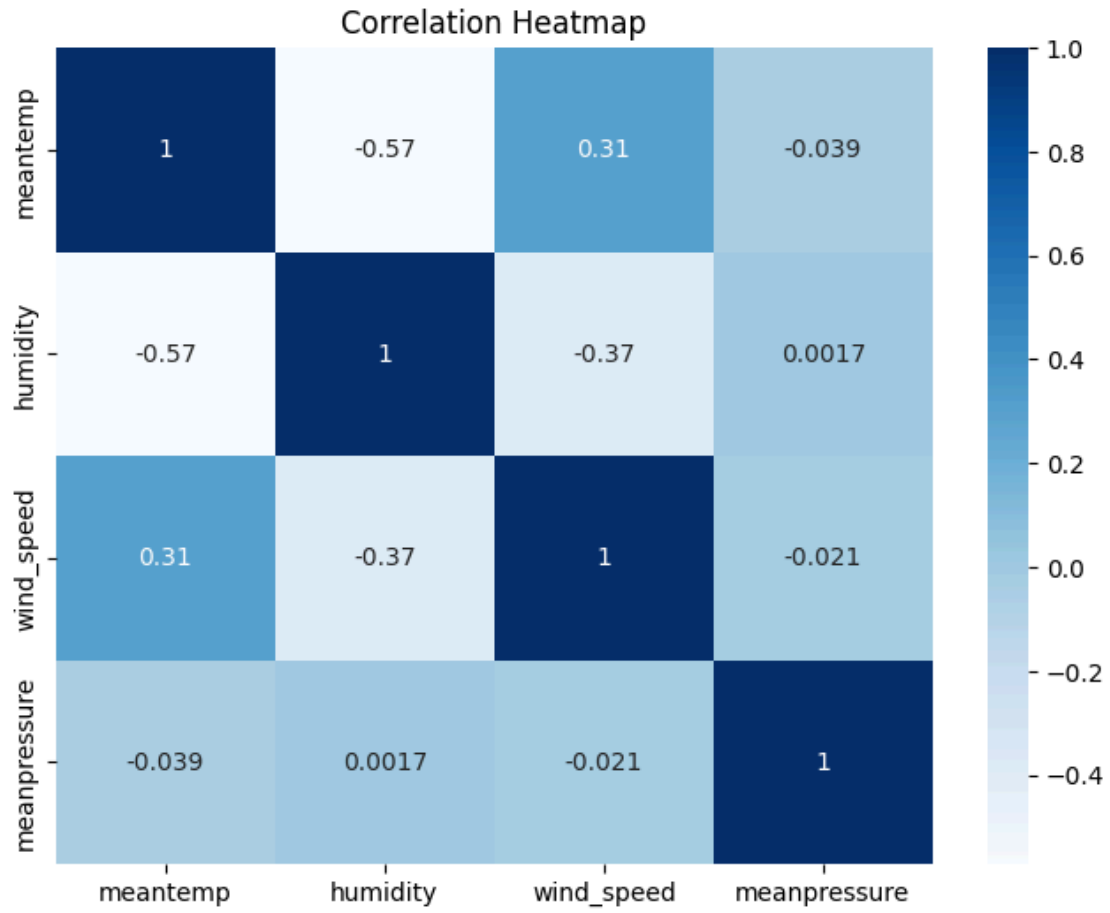
High Humidity Days (>80%)

- These days are mostly in **winter months (January, February) and early monsoon(July, August)**.
- Temperatures range from **6°C to 17°C**.
- Wind speeds are generally **low**, sometimes close to **0 km/h**.
- High pressure (1015–1019 hPa) is a common feature.

Potential Insights:

1. **Hottest Days Occur in Pre-Monsoon Months**: This suggests a **continental climate**, where peak summer heat is experienced in late spring/early summer before monsoon rains.
2. **Coldest Days Coincide with High Humidity**: Likely due to **foggy conditions** in winter, which is common in many regions of India. More details on this when we observe their correlation matrix later.

Correlation between the weather parameters:



Temperature vs. Humidity

- **Correlation Value:** -0.57
- **Observation:** There is a moderate negative correlation between temperature and humidity. As temperature increases, humidity tends to

decrease, and vice versa. This aligns with the general trend that hotter days are often drier, while colder days are more humid.

Temperature vs. Wind Speed

- **Correlation Value:** 0.31
- **Observation:** There is a weak positive correlation between temperature and wind speed. Higher temperatures are slightly associated with higher wind speeds, though the relationship is not strong. This could indicate that warmer days might experience slightly breezier conditions.

Humidity vs. Wind Speed

- **Correlation Value:** -0.37
- **Observation:** There is a weak to moderate negative correlation between humidity and wind speed. Higher humidity tends to occur with lower wind speeds, and vice versa. This suggests that calm conditions (low wind speeds) are often associated with higher humidity, such as during winter or monsoon seasons.

Mean Pressure vs. Other Variables

- **Correlation Values:**
  - Temperature: -0.04
  - Humidity: 0.002
  - Wind Speed: -0.02
- **Observation:** Mean pressure shows almost no correlation with temperature, humidity, or wind speed. This indicates that mean pressure is largely independent of these variables in the dataset. Pressure is likely influenced by other factors not captured here.

Strongest Relationships

The strongest correlations in the dataset are:

- **Temperature vs. Humidity:** -0.57 (moderate negative correlation).
- **Humidity vs. Wind Speed:** -0.37 (weak to moderate negative correlation).

- **Temperature vs. Wind Speed:** 0.31 (weak positive correlation).

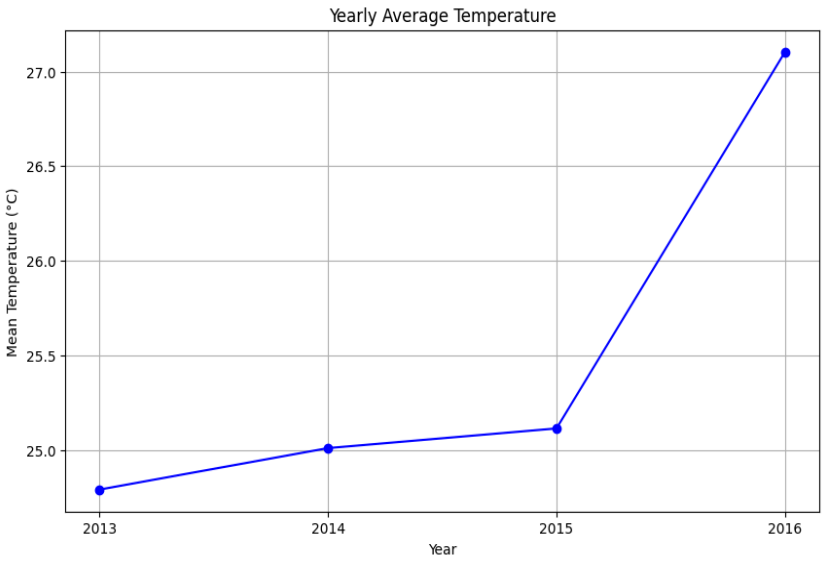
Takeaways

- **Temperature and Humidity:** These two variables have the strongest relationship, with higher temperatures linked to lower humidity.
- **Wind Speed:** Wind speed has a weak relationship with both temperature and humidity, suggesting it plays a secondary role in this dataset.
- **Mean Pressure:** Mean pressure is largely independent of the other variables.

Gradual Increase in Yearly Average Temperature:

The average temperature has shown a **consistent increase** over the years:

- 2013: 24.79°C
- 2014: 25.01°C
- 2015: 25.11°C
- 2016: 27.10°C



There is a clear upward trend in average temperatures from 2013 to 2016. Notably, 2015-2016 stands out with a significant jump in temperature (27.10°C), which is considerably higher than the previous years.

Key Takeaways

1. **Temperature Rise:** The data suggests a warming trend over the years, with each year recording a higher average temperature than the previous one.
2. **Significance of 2016:** The sharp increase in temperature in 2016 could indicate an anomaly or the start of a new trend in rising temperatures.

Temperature Anomalies:

For our dataset, we identified temperature anomalies as days where temperatures were **more than two standard deviations away from the mean**. All detected anomalies correspond to **unusually cold days**, with no instances exceeding the upper threshold.

Key Observations:

- Seasonal Occurrence: All anomalies occurred in winter (December and January), aligning with expected seasonal patterns.
- Temperature Range: Anomalous temperatures ranged from 6.00°C to 10.75°C, with the coldest recorded at 6.00°C on January 5, 2013.
- Humidity Levels: Moderate to high humidity levels (ranging from 63.71% to 100%) were observed, suggesting moisture retention during cold conditions.
- Wind Speeds: Low to moderate wind speeds (0.00 km/h to 8.10 km/h) were recorded, indicating calm or slightly breezy conditions typical of winter.

Additional Insights:

- The absence of high-temperature anomalies suggests that extreme cold events are more common than heat spikes in this dataset.
- The relationship between humidity and temperature could indicate foggy or misty conditions on these anomalous days (we explained it from the correlation between the two parameters as well).
- Wind speeds being low during anomalies may contribute to the persistence of cold air, preventing significant warming.