MDSC 102

INFERENTIAL STATISTICS

LAB TEST REPORT

SRI SATHYA SAI INSTITUTE OF HIGHER LEARNING

(Deemed to be University)

Department of Mathematics and Computer Science

Muddenahalli Campus

SUBMITTED BY

BR SRICHARAN

23902

INTRODUCTION: -

Weather data analysis is an indispensable part of understanding and predicting the ever-changing climate conditions. In the era of big data, the availability of vast weather datasets provides valuable insights into temperature variations, precipitation patterns, and air quality across diverse geographical regions. One such comprehensive dataset, the **"Indian Weather Repository,"** sourced from Kaggle, comprises a rich repository of meteorological information. This dataset encompasses 42 columns and a substantial 24,618 rows, covering a wide range of weather-related parameters.

Dataset Overview: The "Indian Weather Repository" dataset is a treasure trove of meteorological information collected from various regions within India. Each row of the dataset represents a unique instance of weather observations, while the 42 columns provide detailed attributes, such as temperature in both Celsius and Fahrenheit, wind speed and direction, humidity levels, air quality measurements, and celestial events like sunrise and sunset times.

Data Cleaning and Completeness: Prior to analysis, a critical aspect of working with any dataset is data cleaning. In this case, it is heartening to note that the dataset emerges clean, devoid of any null values, which speaks to the quality of data collection and curation. However, it is worth mentioning that there are instances where some administrative regions or states have been omitted from the dataset, specifically Telangana, Ladakh, and Delhi (as Union Territories). These minor gaps do not significantly impede the richness of insights that can be derived from the available data.

Objective of Analysis: The primary objective of this analysis is to explore and derive meaningful insights from the "Indian Weather Repository" dataset. By employing data analysis techniques, we aim to uncover patterns, trends, and relationships within the data. This analysis will facilitate a better understanding of India's weather conditions, helping stakeholders make informed decisions

Data :

It consists of the following Data Columns.

country: Country of the weather data

location_name: Name of the location (city)

region: Administrative region of the location

latitude: Latitude coordinate of the location

longitude: Longitude coordinate of the location

timezone: Timezone of the location

last_updated_epoch: Unix timestamp of the last data update

last_updated: Local time of the last data update

temperature_celsius: Temperature in degrees Celsius

temperature_fahrenheit: Temperature in degrees Fahrenheit

condition_text: Weather condition description

wind_mph: Wind speed in miles per hour

wind_kph: Wind speed in kilometers per hour

wind_degree: Wind direction in degrees

wind_direction: Wind direction as 16-point compass

pressure_mb: Pressure in millibars

pressure_in: Pressure in inches

precip_mm: Precipitation amount in millimeters

precip_in: Precipitation amount in inches

humidity: Humidity as a percentage

cloud: Cloud cover as a percentage

feels_like_celsius: Feels-like temperature in Celsius

feels_like_fahrenheit: Feels-like temperature in Fahrenheit

visibility_km: Visibility in kilometers

visibility_miles: Visibility in miles

uv_index: UV Index

gust_mph: Wind gust in miles per hour

gust_kph: Wind gust in kilometers per hour

air_quality_Carbon_Monoxide: Air quality measurement: Carbon Monoxide

air_quality_Ozone: Air quality measurement: Ozone

air_quality_Nitrogen_dioxide: Air quality measurement: Nitrogen Dioxide

air_quality_Sulphur_dioxide: Air quality measurement: Sulphur Dioxide

air_quality_PM2.5: Air quality measurement: PM2.5

air_quality_PM10: Air quality measurement: PM10

air_quality_us-epa-index: Air quality measurement: US EPA Index

air_quality_gb-defra-index: Air quality measurement: GB DEFRA Index

sunrise: Local time of sunrise

sunset: Local time of sunset

moonrise: Local time of moonrise

moonset: Local time of moonset

moon_phase: Current moon phase

moon_illumination: Moon illumination percentage

EDA-Exploratory Data Analysis with Inferences

On Performing Data Analysis the following are the observations.

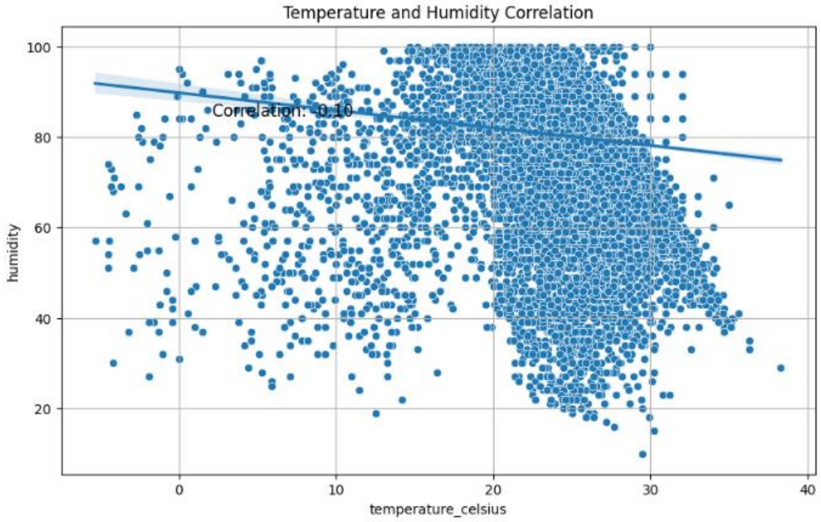
Temperature and Humidity Correlation

The correlation between Temperature and Humidity is -0.10 across India in all the Regions.

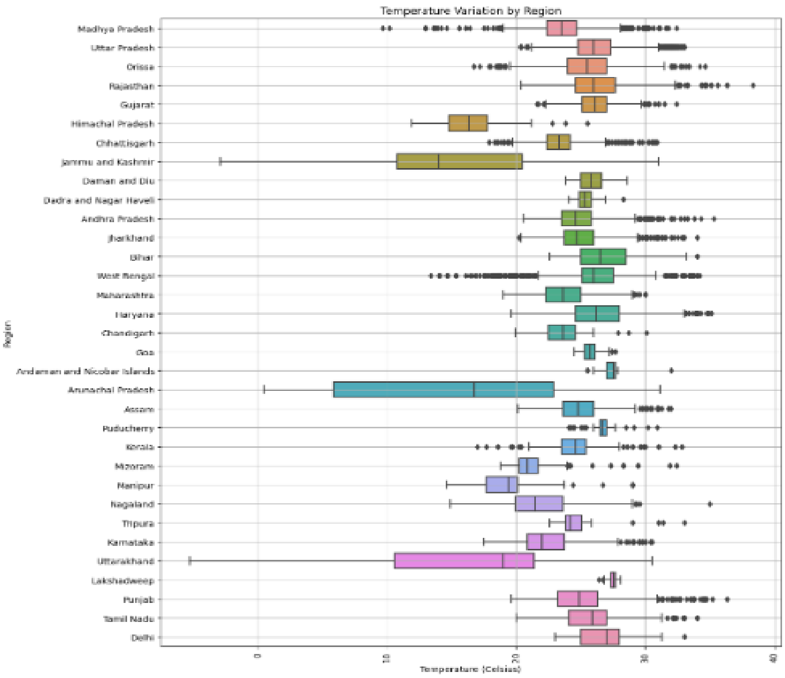
Weak Negative Correlation: This means that as Temperature increases, Humidity tends to decrease slightly, and vice versa. However, the relationship is not very strong.

No Strong Linear Relationship: The correlation of -0.10 is relatively close to zero, indicating that there is no strong linear relationship between Temperature and Humidity. In other words, changes in Temperature do not strongly predict changes in Humidity.

Variability in Weather Conditions: The weak correlation suggests that temperature and humidity can vary independently of each other in different regions and at different times. Weather conditions are influenced by a multitude of factors beyond just temperature and humidity.



Temperature Variation by Region

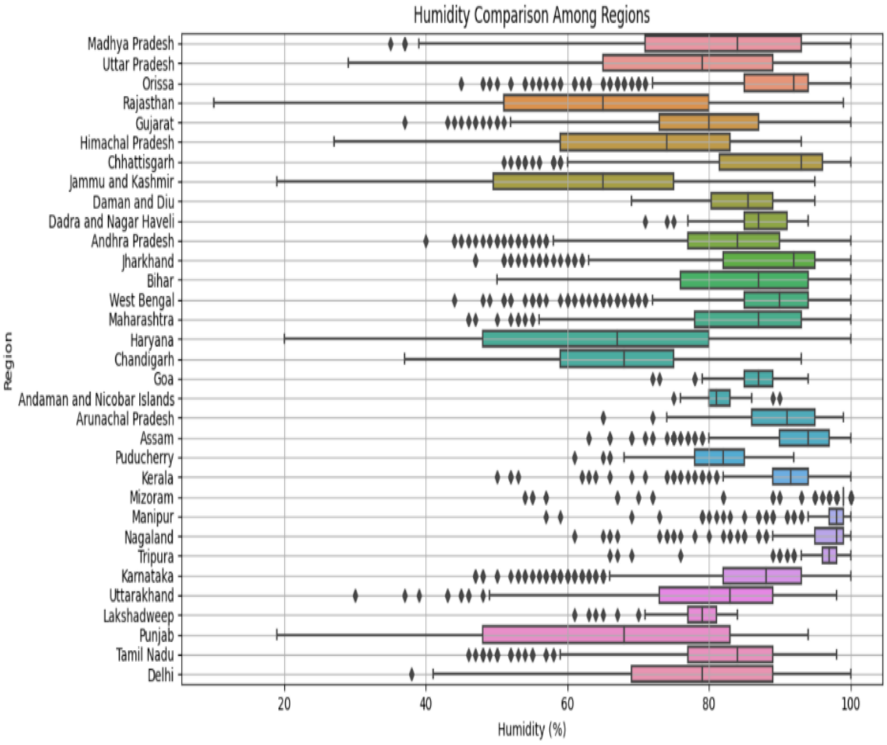


The Temperature Variation among different regions is visualized through Box Plot.

We can observe that there are several Outliers for each region with respect to Temperature. States like Jammu Kashmir ,Arunachal Pradesh, Uttarakhand have no outliers at all. Which means temperature is widely spread along the median.

But for other regions there are outliers which means that the temperature is distributed more than or less than the the actual median for a region. Most the plots are between the temperature 10°C and 38°C

Humidity Comparison among the Regions.

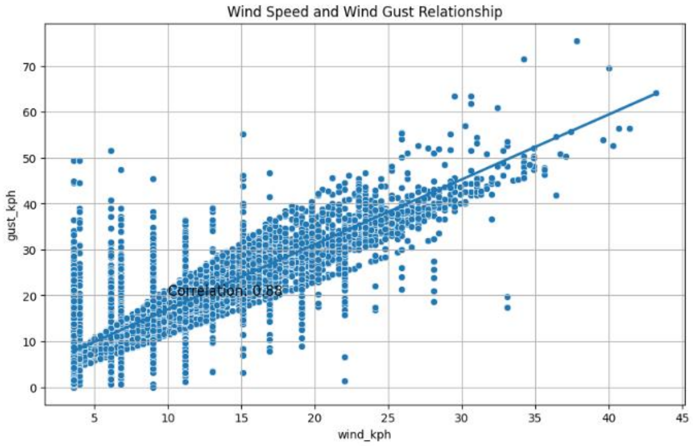


From the given visualization we come to a conclusion that almost all boxplots are left skewed. Rajasthan has humidity range ranging from 8% to 98% and it is left skewed.

We can observe there are a lot of outliers for the plot in the region of West Bengal and it is left skewed.

Although the humidity in Manipur is high, there are several outliers for the data on the left hand side.

Wind Speed and Wind Gust Relationship:



The correlation is -0.88.

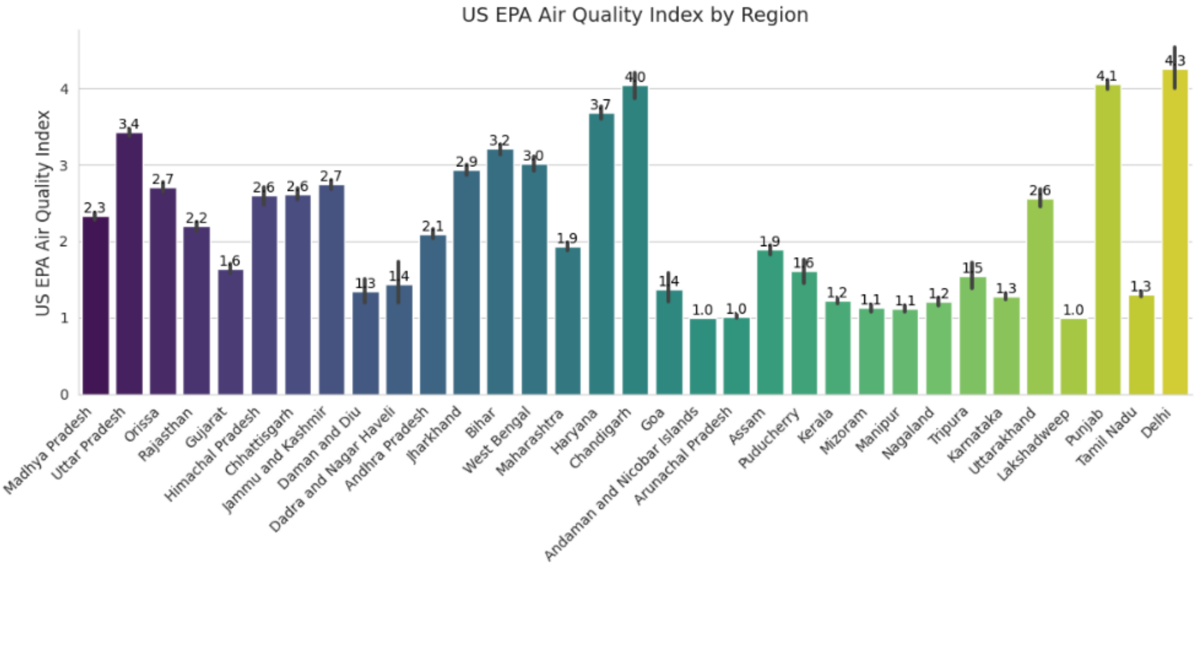
A correlation of -0.88 between wind speed and wind gust indicates a strong negative linear relationship between these two variables.

Inverse Relationship: The negative correlation coefficient (-0.88) suggests that as wind speed increases, wind gust tends to decrease, and vice versa. In other words, when the wind speed is high, wind gusts are likely to be lower, and when wind speed is low, wind gusts tend to be higher.

A high negative correlation implies that there is a degree of consistency in how wind speed and wind gusts vary together. When wind speed is consistently high, wind gusts are consistently low.

US EPA Air Quality Index by Region in India.

The United States Environmental Protection Agency (US EPA) Air Quality Index (AQI) is a standardized system used to communicate air quality conditions to the public. We observe from the plot that the air quality is very poor in Delhi and the quality of air is really good in the region of Andaman and Nicobar Islands,Arunchal Pradesh and Lakshwadeep .

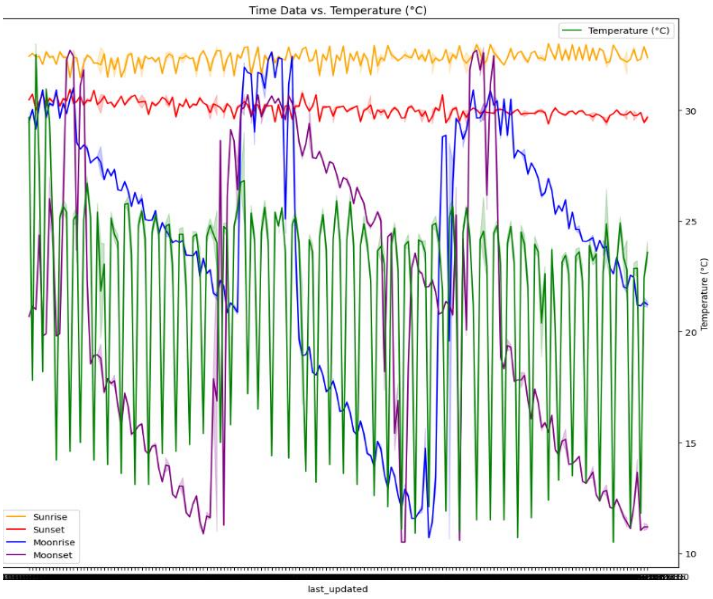


Timedata vs Temperature:

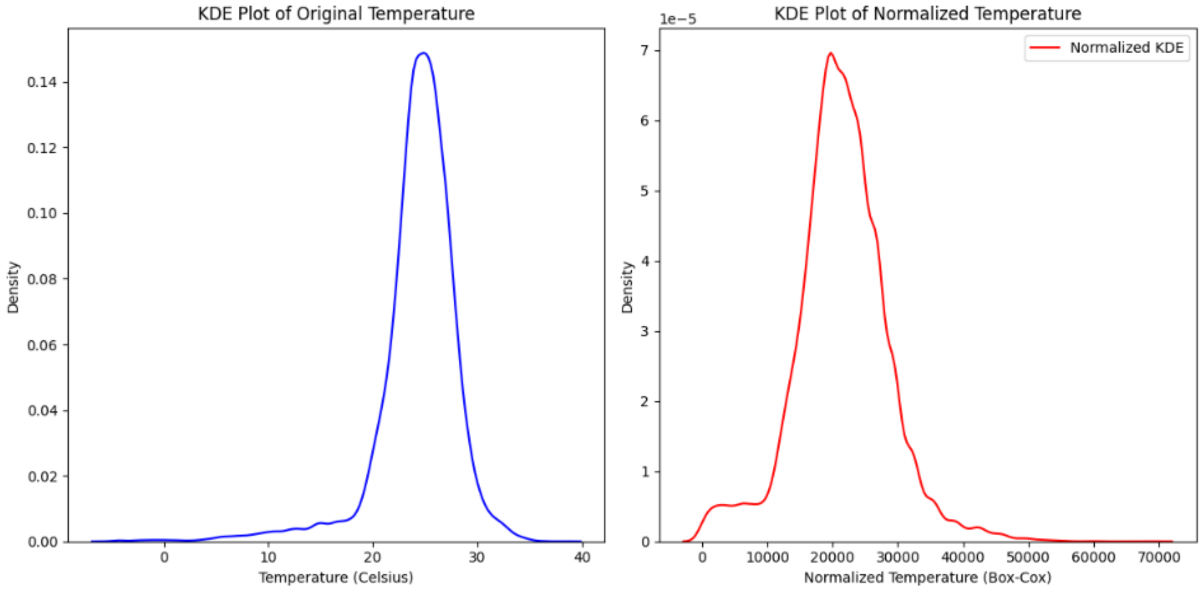
Temperature plays a major role in during sunrise,sunset ,moonrise and moonset.We observe that the temperature varies between 12°C and 28°C.

During sunrise the temperature is always above 31°C.During sunset the temperature the temperature is close to 30°C.

During moonrise and moonset there is significant change in the temperature over the time.



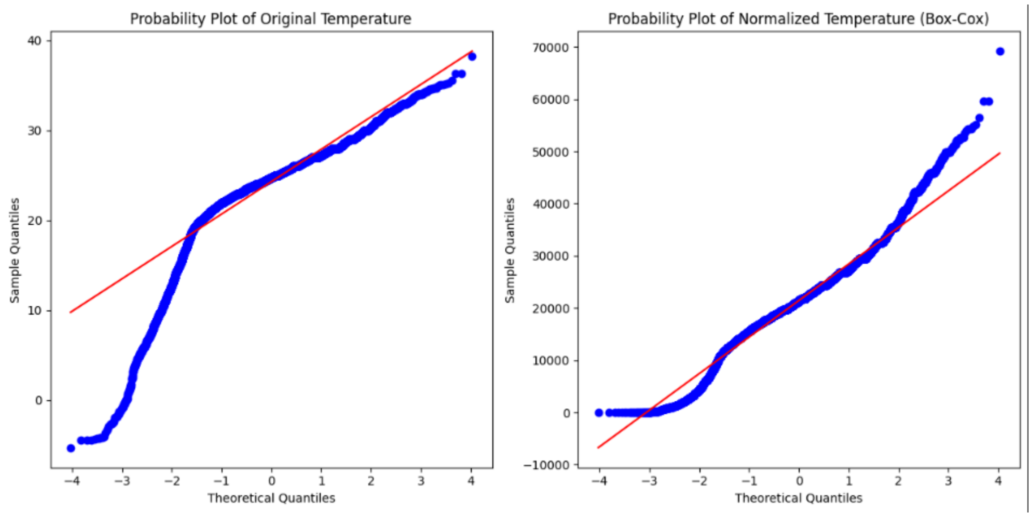
Normalizing Temperature using BoxCox



Original Temperature Distribution: The KDE plot of the original "temperature_celsius" data reveals that it is right-skewed, with a peak around a specific temperature range.

Normalized Temperature Distribution: After applying the Box-Cox transformation, the normalized data shows a more symmetric and bell-shaped distribution.

Effect of the Transformation: The transformation significantly reduces the skewness of the data, making it more suitable for statistical analysis that assumes normality.



Hypothesis Testing

Hypothesis testing for Temperature(Z-test):-

Hypothesis testing is a statistical method used to determine whether there is enough evidence to support or reject a specific claim about a population parameter, typically the population mean. In this report, we will perform a hypothesis test to examine whether the population mean temperature in Celsius is different from 25°C using a **significance level (alpha) of 0.05.**

Hypotheses:

Null Hypothesis (H0): The population mean temperature (μ) is equal to 25°C.

Alternative Hypothesis (Ha): The population mean temperature (μ) is not equal to 25°C.

Definitions:

Null Hypothesis (H0): The null hypothesis is a statement that suggests there is no significant difference or effect. In this case, H0 suggests that the population mean temperature is equal to 25°C.

Alternative Hypothesis (Ha): The alternative hypothesis is a statement that contradicts the null hypothesis and suggests that there is a significant difference or effect. Here Ha suggests that the population mean temperature is not equal to 25°C.

Significance Level (α): The significance level (alpha) is the probability of making a Type I error (rejecting the null hypothesis when it is true). A common choice is α = 0.05, indicating a 5% chance of making a Type I error.

Sample Mean (x̄): The sample mean is the average value of the observations in a sample. In this test, the sample mean is calculated as approximately 24.29°C.

Sample Size (n): The sample size is the number of observations in the sample. In this test, the sample size is 24618.

<u>Sample Standard Deviation (s):</u> The sample standard deviation is a measure of the variability or spread of the sample data. In this test, the sample standard deviation is approximately 3.91°C.

<u>Population Mean (μ):</u> The population mean is the average value of the entire population. In this test, the null hypothesis suggests that the population mean is 25°C.

<u>Z-Score (z):</u> The Z-score is a measure of how many standard deviations a sample mean is away from the population mean. In this test, the Z-score is approximately -28.43.

<u>P-Value:</u> The p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the one observed in the sample, assuming the null hypothesis is true. In this test, the p-value is 0.0.

Hypothesis Test:

To test the null hypothesis (H0), we calculated the Z-score using the formula:

$$\mathbf{Z = (\bar{x} - \mu)/(s/\sqrt{n})}$$

Where:

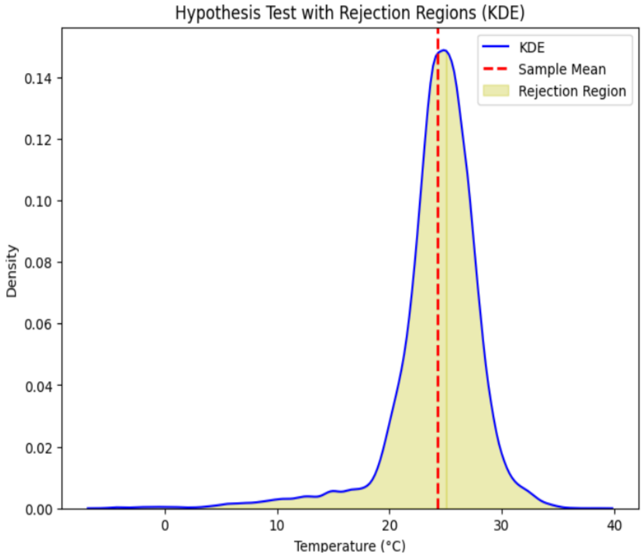
$\bar{x}$ is the sample mean (approximately 24.29)

μ is the population mean (25)

s is the sample standard deviation (approximately 3.91)

n is the sample size (24618)

Z is the Z-score (approximately -28.43)



We then calculated the p-value for a two-tailed test, which was 0.0. This p-value represents the probability of observing a sample mean as extreme as the one obtained (or more extreme) under the assumption that the null hypothesis is true.

Conclusion:

Comparing the p-value (0.0) to the significance level (α = 0.05), we found that the p-value is much smaller than α. Therefore, we reject the null hypothesis (H0). This means that there is enough evidence to conclude that the population mean temperature is different from 25°C. The Z-score of -28.43 suggests that the sample mean is significantly lower than the hypothesized population mean.

In summary, the data provides strong evidence to suggest that the population mean temperature in Celsius is different from 25°C.

Hypothesis testing for Temperature(T-test):-

T-Statistic: The t-statistic quantifies the difference between the sample mean and the null hypothesis mean in terms of standard errors. It is calculated using the formula:

t_statistic = (sample_mean - population_mean) / (sample_std / √n)

In this analysis, the t-statistic is -28.43.

Degrees of Freedom (df): The degrees of freedom are determined by the sample size and are used in t-distribution calculations. For this analysis, df = n - 1, so df = 24,617.

P-Value: The p-value represents the probability of observing a t-statistic as extreme as the one calculated under the null hypothesis. In this case, it is a two-tailed test, so the p-value is calculated as 2 * (1 - stats.t.cdf(abs(t_statistic), df=degrees_of_freedom)). The p-value is found to be extremely close to 0 (P-value ≈ 0).

Analysis:

The null hypothesis (H0) assumes that the population mean temperature is 25°C.

The t-statistic is -28.43, indicating a substantial difference between the sample mean (24.29°C) and the null hypothesis mean (25°C).

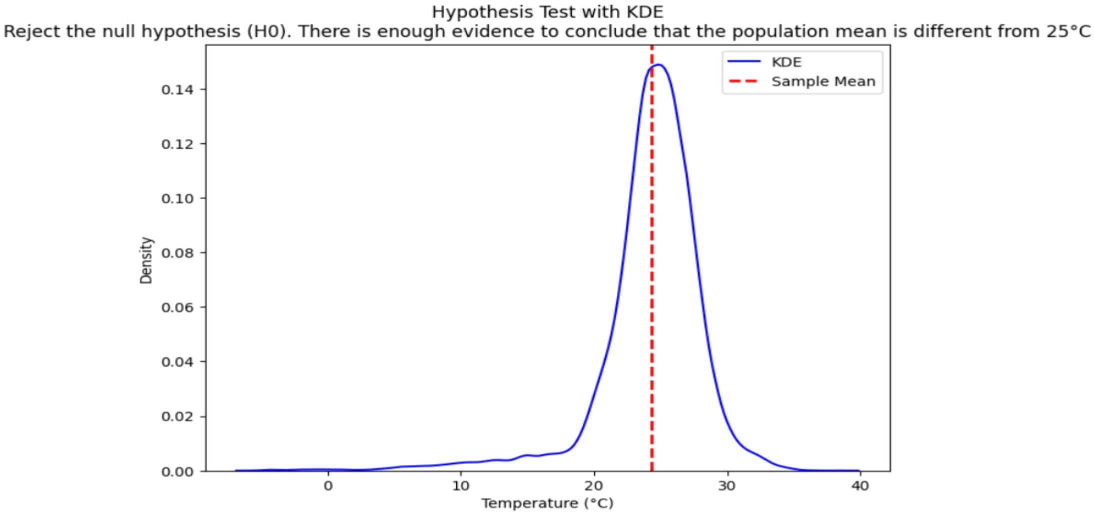
The degrees of freedom are 24,617, reflecting the sample size.

The p-value is calculated to be nearly 0, significantly below the 0.05 significance level.

Conclusion:

The p-value is well below the 0.05 significance level, leading to the rejection of the null hypothesis. This suggests that there is strong evidence to conclude that the population mean temperature is different from 25°C.

In summary, based on this statistical analysis, the data provides compelling evidence that the population mean temperature in the dataset is not 25°C. The sample mean temperature of 24.29°C, along with a small p-value, indicates a significant difference. Further investigations or analyses may be needed to explore the nature of this difference and its implications.



Hypothesis testing for Rainfall(Z-test):-

Hypothesis Testing

The variable of interest in this analysis is "precip_mm," representing the daily precipitation in millimetres.

Hypotheses:

Null Hypothesis (H0): Population mean of daily precipitation (μ) is equal to 0.1 millimetres

Alternative Hypothesis (Ha):Population mean of daily precipitation (μ) is not equal to 0.1 millimetres

The significance level (alpha) is set at 0.05, which is the probability of making a Type I error (rejecting the null hypothesis when it is true).

Results and Conclusion

The calculated statistics and the outcome of the hypothesis test are as follows:

Sample Mean: 0.2789987

Sample Size: 24,618

Sample Standard Deviation: 1.1877268

Z-Score Value: 23.6461104

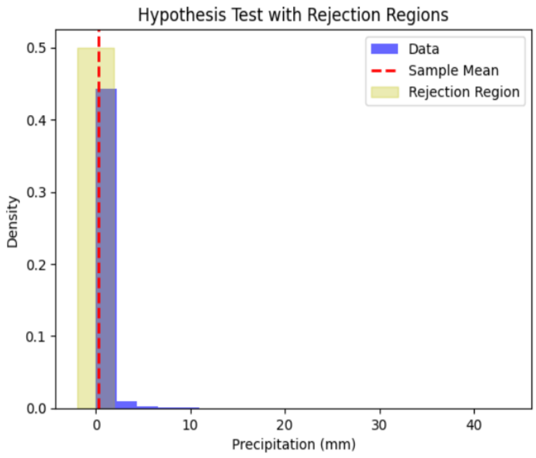
CDF of the Standard Normal Distribution at the Absolute Value of z_score: 1.0

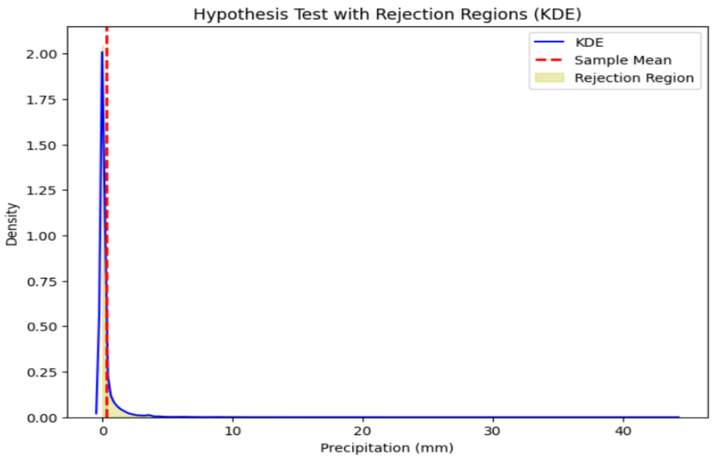
P-Value: 0.0

Based on the results, the p-value is much smaller than the significance level (alpha). Specifically, the p-value is 0.0, which is less than 0.05. This means that there is strong evidence to reject the null hypothesis (H0).

Conclusion: We reject the null hypothesis. There is enough statistical evidence to conclude that the population mean of daily precipitation in the specific region is different from 0.1 millimetres.

In other words, the data provides strong support that the daily precipitation in this region is not, on average, equal to 0.1 millimetres. The sample mean of 0.2789987 suggests that the average daily precipitation is significantly higher than the value stated in the null hypothesis.


Hypothesis Test with Rejection Regions
Data
Sample Mean
Rejection Region
Density
Precipitation (mm)


Hypothesis Test with Rejection Regions (KDE)
KDE
Sample Mean
Rejection Region
Density
Precipitation (mm)