### National University of Computer and Emerging Sciences

****

Data Science

**Weather Data Analysis (Rain Prediction)**

**Rafeel Abdul (21L-5448) Saad Hassan (21L-5252)**

**Ahmad Sadeed (21L-7584)**

**8A**

**FAST School of Computing**

**National University of Computer and Emerging Sciences**

**Lahore, Pakistan**

# Problem Statement

The objective of this project is to analyze weather data and perform predictive analysis for rainfall occurrence. This dataset, which includes features such as temperature, humidity, wind speed, and atmospheric pressure, has been collected from various weather stations. By leveraging these meteorological factors, the goal is to develop a model that can accurately predict whether it will rain the next day, helping in weather forecasting and decision-making.

# Source of Data Set

# The dataset was obtained from Kaggle:

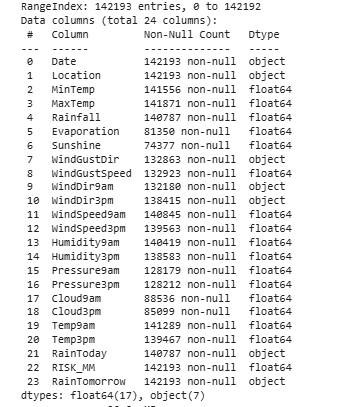
[**https://www.kaggle.com/code/moaazahmed313/data-science-project/input**](https://www.kaggle.com/code/moaazahmed313/data-science-project/input)

### Dataset Overview

The dataset contains 142,145 rows and 22 columns, with each row representing daily weather observations.

The dataset includes meteorological variables such as temperature, humidity, wind speed, cloud cover, and precipitation, which are essential for weather analysis and prediction. Below is a detailed description of each attribute/column:

1. **Location** – Name of the weather station.
2. **MinTemp** – Minimum temperature recorded for the day.
3. **MaxTemp** – Maximum temperature recorded for the day.
4. **Rainfall** – Total daily rainfall in mm.
5. **Evaporation** – Amount of water evaporated during the day.
6. **Sunshine** – Total hours of sunshine recorded in the day.
7. **WindGustDir** – Direction of the strongest wind gust of the day.
8. **WindGustSpeed** – Speed of the strongest wind gust of the day.
9. **WindDir9am** – Wind direction recorded at 9 AM.
10. **WindDir3pm** – Wind direction recorded at 3 PM.
11. **Humidity9am** – Relative humidity at 9 AM.
12. **Humidity3pm** – Relative humidity at 3 PM.
13. **Pressure9am** – Atmospheric pressure at 9 AM.
14. **Pressure3pm** – Atmospheric pressure at 3 PM.
15. **Cloud9am** – Cloud cover at 9 AM.
16. **Cloud3pm** – Cloud cover at 3 PM.
17. **Temp9am** – Temperature recorded at 9 AM.
18. **Temp3pm** – Temperature recorded at 3 PM.
19. **RainToday** – Indicates whether it rained today (Yes or No).
20. **RainTomorrow** – Indicates whether it will rain tomorrow (Yes or No).



### Univariate Analysis:

In univariate analysis, all variables are taken separately to examine their distribution, and also some variables of limited, qualitative, or skewed distribution like mean, median, mode and variation are calculated. For continuous data, it is used either histograms to determine the shape of the distribution of the data set (normal, skewed or uniform distribution type) or for the distribution’s spread, the IQR and outliers using the box plot available type of data. This type of analysis allows us to reveal some sort of periodicity, outliers, or characteristics that need to be addressed or transformed.

### Wind Speed at 9am

The wind speed at 9 AM is heavily skewed toward low values, with most observations near 0.0–0.2 (peak frequency ~14,000) and rare occurrences above 0.6. The median (~0.4–0.5) exceeds the mean (~0.3), confirming right-skewness, while the narrow IQR and small standard deviation (~0.2) indicate consistent, calm conditions. High winds are uncommon, and the data shows no significant outliers, suggesting stable morning wind patterns.

### 

### Wind Speed at 3pm

The 3 PM wind speed distribution shows a right-skewed pattern similar to the 9 AM data but with slightly higher speeds overall. While most observations still cluster at lower values (0.0-0.4), the peak frequency is lower (~12,000 vs. 14,000) and the median appears slightly higher (~0.5-0.6), suggesting afternoon winds are marginally stronger on average. The box plot confirms this shift with a higher median line and slightly longer upper whisker, though the IQR remains narrow and extreme winds (>0.8) remain uncommon, maintaining the pattern of generally light winds throughout the day.

### 

### Wind Gust Speed

### The wind gust speed distribution shows a pronounced right-skewed pattern, with the vast majority of observations concentrated at lower speeds (0.0-0.4), evidenced by the peak frequency near 17,500 at the lowest bin. The box plot reveals a median around 0.3-0.4, while the mean is likely higher due to the long tail extending toward 1.0, indicating occasional strong gusts that pull the average upward. Despite these rare higher-speed events, the narrow IQR and clustering of most data points below 0.6 demonstrate that extreme wind gusts are infrequent, with the distribution dominated by mild to moderate speeds.

### 

### Distribution of Sunshine

### The Sunshine distribution shows a bimodal pattern, with two distinct peaks near 0.0 (complete cloud cover) and 1.0 (full sunshine), each reaching frequencies around 40,000–50,000. The box plot confirms this polarization, with the median (~0.5) centered between the extremes, but the wide IQR and long whiskers indicate high variability—days tend to be either very sunny or very cloudy, with few moderate cases. The near-symmetric spread suggests no systematic skew, though the extreme clustering at both ends implies all-or-nothing sunlight conditions dominate the dataset.

### 

### Temperature at 9am

### The 9 AM temperature distribution shows a near-normal distribution, peaking around 0.4–0.6 (frequency ~3500), with symmetrical tails tapering toward extremes (0.0 and 1.0). The box plot confirms this symmetry, with the median near 0.5 and evenly spaced quartiles, indicating balanced variability and no skew. Minimal outliers suggest stable morning temperatures, clustering moderately around the median.

### 

### Temperature at 3pm

### The 3 PM temperature distribution peaks near 0.6–0.8 (frequency ~5000), reflecting warmer afternoon temperatures compared to 9 AM. The box plot shows a higher median (~0.7) and a tighter IQR, indicating less variability and consistent daytime heating. The slight left skew (longer tail toward lower temps) suggests occasional cooler outliers, but most data clusters in the warmer range.

### 

### Rainfall

### The Rainfall distribution is heavily right-skewed, with ~80% of values near 0.0–0.2 (frequency ~80,000), indicating most days have little to no rain. The box plot reveals a median close to 0.0, an extremely compact IQR, and a long upper whisker (to 1.0), highlighting rare but intense rainfall events. This pattern suggests a climate with frequent dry spells punctuated by occasional heavy downpours.

### 

### Pressure at 9am

### The Pressure9am distribution shows a tightly clustered normal distribution, peaking sharply near 0.5 (likely representing standard atmospheric pressure). The box plot confirms minimal variability, with a narrow IQR and median centered at 0.5, indicating remarkably stable morning pressure conditions. The symmetrical tails and absence of outliers suggest a highly consistent pressure regime, typical of calm weather patterns.

### 

### Pressure at 3pm

### The Pressure3pm distribution maintains the same tightly clustered normal distribution as morning readings, peaking sharply near 0.5 with symmetrical tails, indicating stable atmospheric conditions throughout the day. The box plot shows an identical narrow IQR and median at 0.5, confirming minimal diurnal pressure variation. The complete absence of outliers in both plots suggests exceptionally consistent pressure patterns, likely characteristic of a region with low weather volatility.

### 

### Minimum Temperature

### The MinTemp distribution shows a left-skewed pattern, with most values clustered in the 0.4–0.8 range (peak frequency ~3500) and fewer extreme low temperatures. The box plot reveals a median around 0.6, with the lower quartile extending toward 0.4 and the upper quartile near 0.7, indicating slightly warmer minimum temperatures overall. The longer lower whisker (toward 0.0) suggests occasional cold outliers, but the bulk of data demonstrates moderate overnight lows with limited variability.

### 

### Maximum Temperature

### The MaxTemp distribution shows a right-skewed pattern, with most values concentrated between 0.5–0.8 (indicating warm daytime highs) and a frequency peak around 0.6–0.7. The box plot highlights a median near 0.7, with a compact IQR and a long upper whisker extending to 1.0, suggesting consistent warm temperatures punctuated by occasional extreme heat events.

### 

### Humidity at 9am

### The 9 AM humidity distribution shows most values cluster between 0.6–1.0, peaking near 0.8–0.9, indicating consistently humid mornings. The box plot confirms this with a high median (~0.8), narrow IQR, and rare dry outliers (below 0.4). This suggests stable, predictable humidity levels at this time.

### 

### Humidity at 3pm

### The Humidity3pm distribution shows slightly lower humidity than mornings, with values clustered between 0.4–0.8 and peaking around 0.6–0.7. The box plot reveals a median near 0.6, with a wider spread than 9 AM, reflecting afternoon drying. While still generally humid, the longer lower whisker (to 0.0) shows more frequent dry periods compared to mornings.

### 

### Evaporation

### The Evaporation data shows most values are concentrated at the lower end (0.0–0.4), indicating generally low evaporation rates. The box plot reveals a median near 0.2, with a compact IQR and a long upper whisker extending to 1.0, highlighting rare high-evaporation events. This suggests typically minimal evaporation with occasional spikes, likely tied to hot/dry conditions.

### 

### Cloud at 9am

### The Cloud9am distribution reveals a bimodal pattern, with two clear peaks at 0.0 (clear skies) and 1.0 (fully overcast), each reaching frequencies near 50,000–60,000. The box plot shows a median around 0.5, but the wide IQR and full-range whiskers confirm extreme variability—mornings tend to be either completely clear or fully cloudy, with few intermediate states.

### 

### Cloud at 3pm

### The 3 PM cloud data shows most days are either completely clear (0.0) or fully overcast (1.0), with very few partly cloudy days in between. The pattern is similar to morning conditions, but with slightly fewer cloudy afternoons. The middle value falls at 0.5 simply because days tend to be one extreme or the other, not because partial cloud cover is common.

### 

### Bivariate and Multivariate Analysis:

The bivariate and multivariate analysis provides descriptions of variables in an effort to further understand patterns in these relationships. When used in exploratory analysis, scatter diagrams show the relationship between two continuous variables , and you can use them to detect trends, clustering , or outliers. Correlation matrices depict the degree of relationship between numeral variables and distinguish the positive or negative connections between these variables. For multivariate analysis, pair plots best fit in that the plots offer the analysis of more than one variable at a go with a view of displaying the relationship. These methods are required for analyzing variable associations and direction within feature choosing throughout the prediction approach.

### Minimum Temperature and Maximum Temperature

### The scatter plot shows a strong positive correlation between MinTemp and MaxTemp, meaning higher minimum temperatures typically correspond to higher maximum temperatures. The upward-sloping diagonal pattern of data points clearly demonstrates this direct relationship.

### 

### Minimum Temperature and Humidity at 9am

### The scatter plot of MinTemp vs Humidity9am reveals a weak negative correlation, where mornings with higher minimum temperatures (right side of x-axis) tend to have slightly lower humidity levels (bottom of y-axis). Most data points are scattered widely across the plot, showing no strong linear pattern, though there's a slight concentration of high-humidity readings (0.6–1.0) occurring more often with cooler minimum temps (0.0–0.4).

### 

### Minimum Temperature and Humidity at 3pm

### The scatter plot of MinTemp vs Humidity3pm shows a weak to negligible correlation, with data points scattered widely across the plot without a clear pattern. While there's a slight tendency for higher minimum temperatures (right side) to pair with moderately lower afternoon humidity levels (bottom), the relationship is inconsistent.

### 

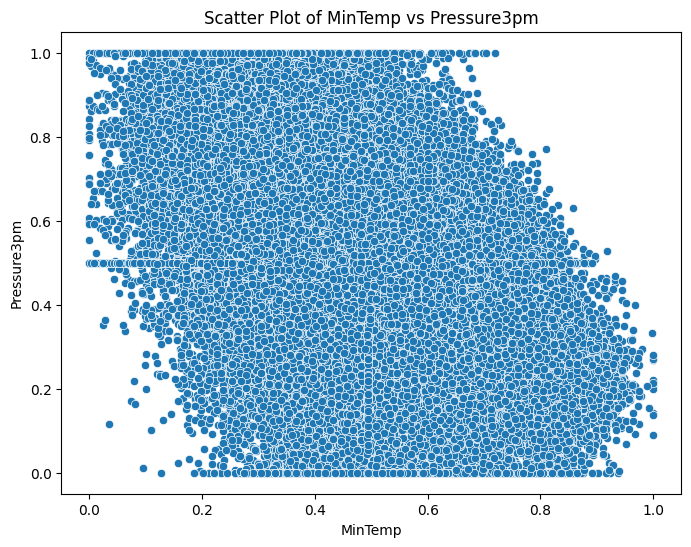
### Minimum Temperature and Pressure at 9am

### The scatter plot of MinTemp vs Pressure9am reveals no clear correlation, as the data points are evenly dispersed across the plot without any distinct pattern. Minimum temperatures (x-axis) show no consistent relationship with morning pressure readings (y-axis), suggesting these variables operate independently in this dataset.

### 

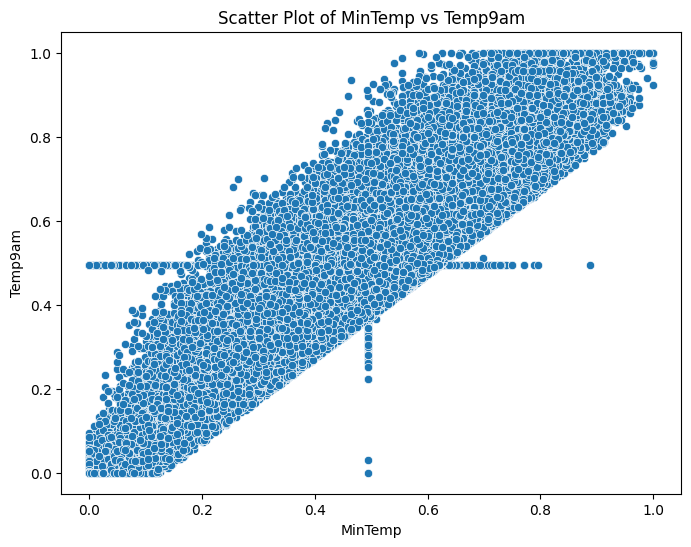
### Minimum Temperature and Pressure at 3pm

### The scatter plot of MinTemp vs Pressure3pm shows no meaningful correlation, with data points randomly scattered across the graph. Morning minimum temperatures (x-axis) appear unrelated to afternoon pressure measurements (y-axis), as the values show no consistent upward or downward trend.



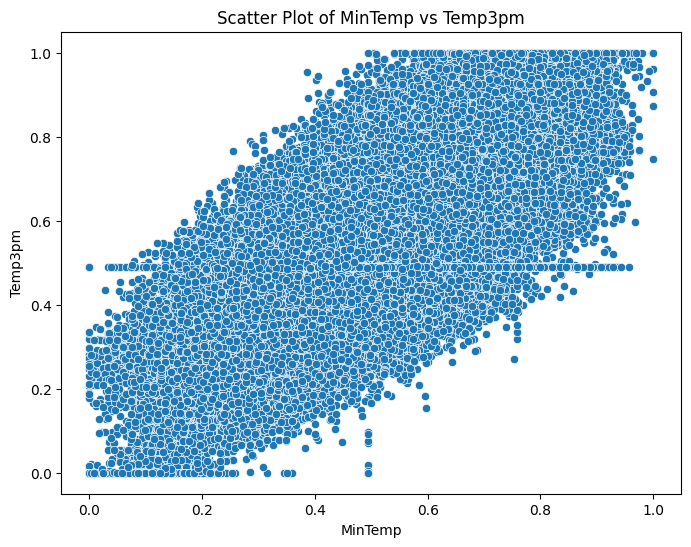
### Minimum Temperature and Temperature at 9am

### The scatter plot of MinTemp vs Pressure9am demonstrates no apparent correlation between the variables, as data points are dispersed uniformly across the plot without forming any discernible pattern. Minimum overnight temperatures (x-axis) show no consistent relationship with 9 AM pressure readings (y-axis), which remain clustered around the 0.4–0.6 range regardless of temperature extremes.



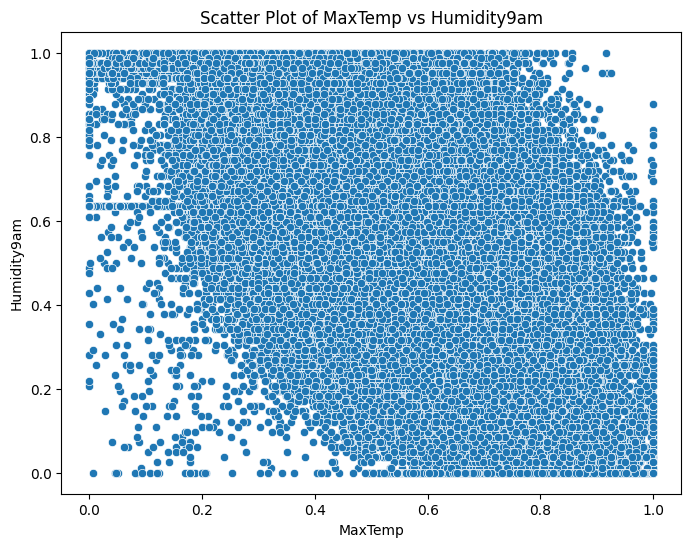
### Minimum Temperature and Temperature at 3pm

### The scatter plot of MinTemp vs Temp3pm reveals a strong positive correlation, where higher minimum overnight temperatures consistently correspond to higher afternoon temperatures. Data points form a clear upward diagonal trend from the lower-left (cool nights and cool afternoons) to the upper-right (warm nights and hot afternoons), indicating that overnight lows strongly influence daytime heating patterns.



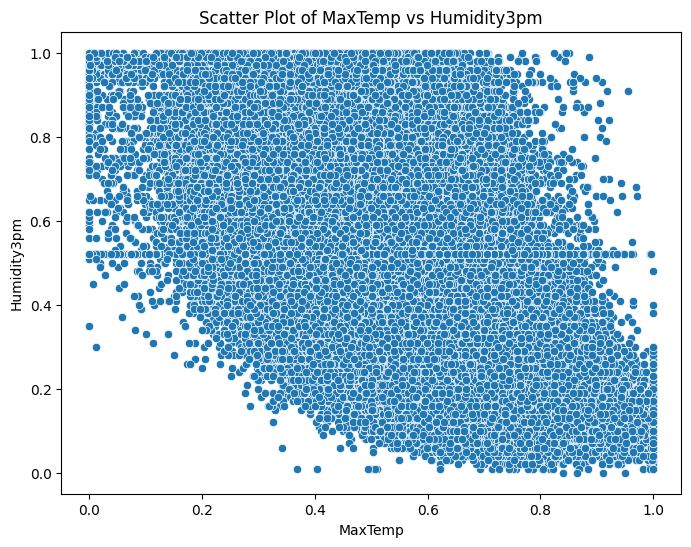
### Maximum Temperature and Humidity at 9am

### The scatter plot of MaxTemp vs Humidity9am shows a moderate negative correlation, where days with higher maximum temperatures (right side of x-axis) tend to have lower morning humidity levels (bottom of y-axis). While the trend is visible, the data points are somewhat scattered, indicating that other factors like overnight cooling or local breezes also influence morning humidity.



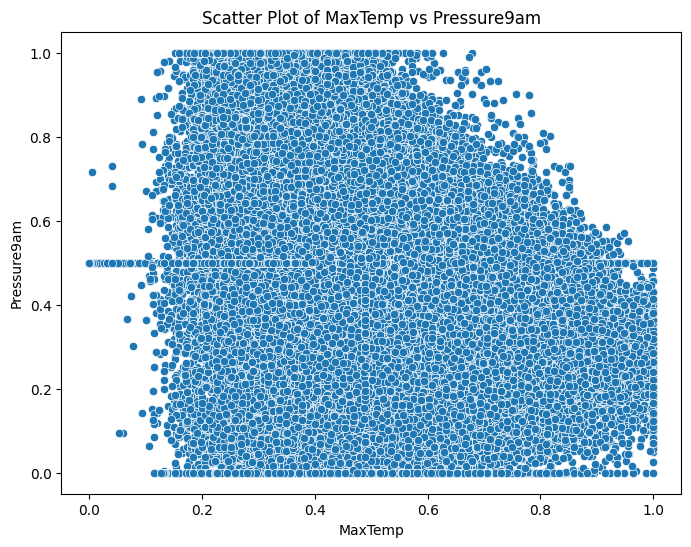
### Maximum Temperature and Humidity at 3pm

### The scatter plot of MaxTemp vs Humidity3pm reveals a stronger negative correlation than the morning data, with most high-temperature days (>0.6) showing notably lower afternoon humidity (<0.4). The points form a clearer downward trend compared to the 9 AM plot, as daytime heating intensifies moisture evaporation.



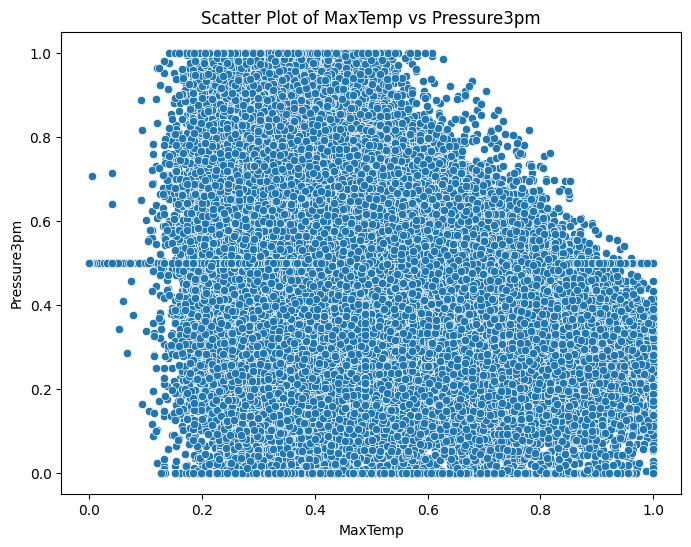
### Maximum Temperature and Pressure at 9am

### The scatter plot of MaxTemp vs Temp9am shows a strong positive correlation, with warmer morning temperatures (x-axis) consistently leading to higher daytime maximums (y-axis). Data points form a clear upward diagonal trend, though with slightly more spread than MinTemp comparisons, suggesting morning readings are a reliable (but not perfect) predictor of afternoon heat potential.



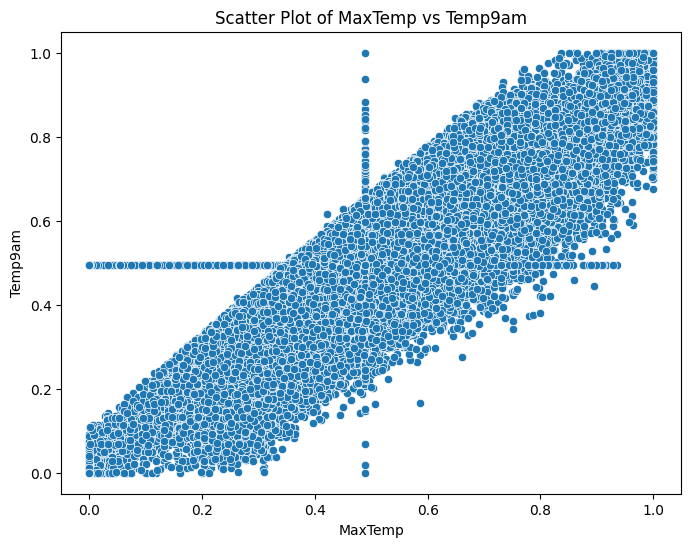
### Maximum Temperature and Pressure at 3pm

### The scatter plot of MaxTemp vs Pressure3pm reveals no meaningful correlation, with data points scattered uniformly across the plot without any discernible pattern. Maximum daytime temperatures (x-axis) show no consistent relationship with afternoon pressure measurements (y-axis), which remain tightly clustered around the 0.4–0.6 range regardless of whether temperatures were cool (near 0.0) or extremely hot (near 1.0).



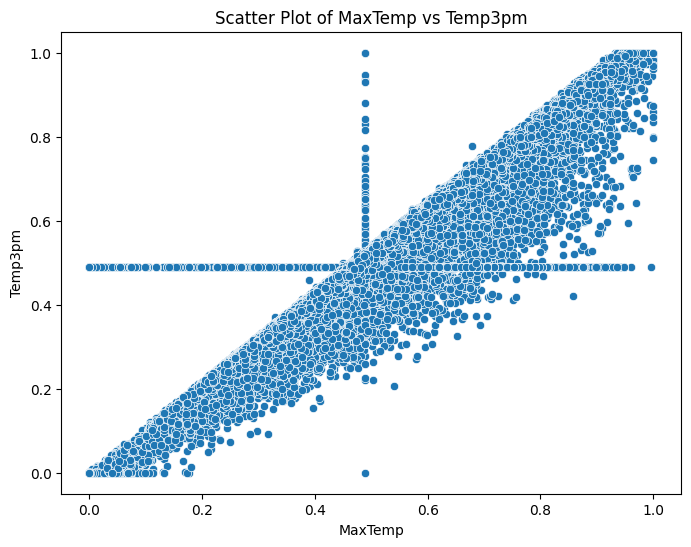
### Maximum Temperature and Temperature at 9am

### The scatter plot of MaxTemp vs Temp9am reveals a strong positive correlation, where warmer morning temperatures consistently lead to higher daytime maximums. The data points form a clear upward trend from the lower-left (cool mornings and cool afternoons) to the upper-right (warm mornings and hot afternoons), though with slightly more variability than MinTemp comparisons.



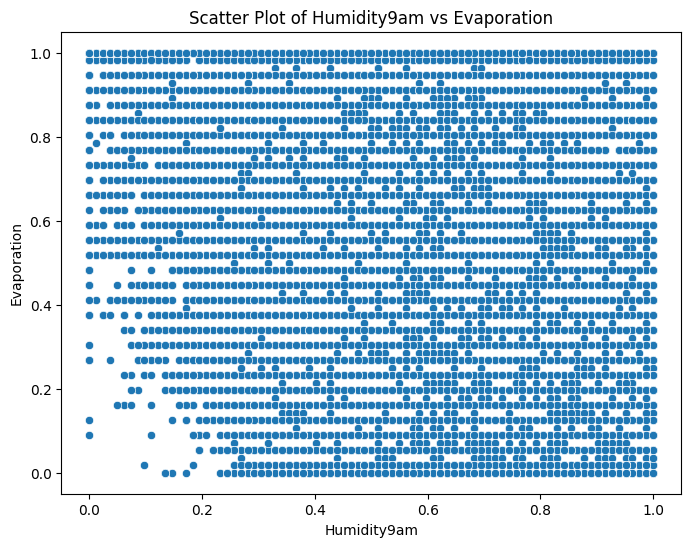
### Maximum Temperature and Temperature at 3pm

### The scatter plot of MaxTemp vs Temp3pm reveals an almost perfect positive correlation, with data points forming a tight diagonal line from the lower-left to the upper-right. This near-linear relationship indicates that the recorded maximum temperature for the day typically matches the 3 PM reading exactly, suggesting that 3 PM temperatures essentially represent the daily maximum in this dataset.



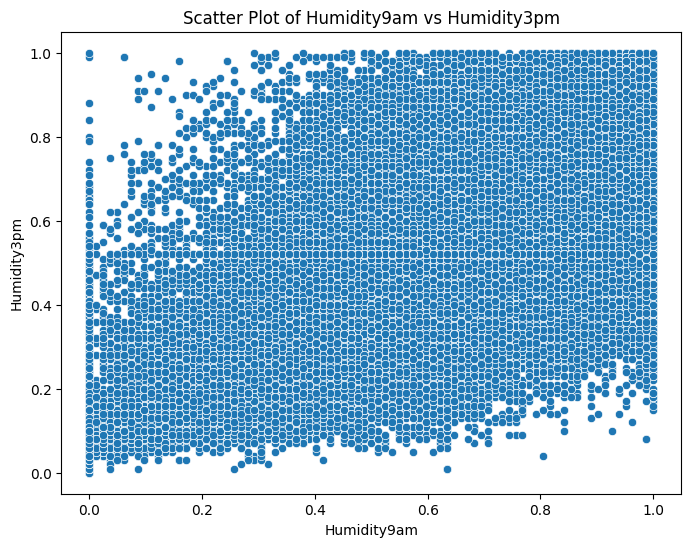
### Humidity at 9am and Evaporation

### The scatter plot of Humidity9am vs Evaporation reveals a moderate negative correlation, where higher morning humidity levels (x-axis) generally correspond to lower evaporation rates (y-axis). Data points trend downward from the upper-left (high humidity with low evaporation) to the lower-right (low humidity with high evaporation), though with considerable scatter.



### Humidity at 9am and Humidity at 3pm

### The scatter plot of Humidity9am vs Humidity3pm shows a moderate positive correlation, where higher morning humidity levels generally lead to higher afternoon readings, though with significant variability. Data points form a loosely clustered upward trend.



### Correlation Matrix:

### The correlation matrix reveals several strong relationships between weather variables. Temperature measurements show the highest correlations, with MinTemp and MaxTemp strongly linked to Temp9am (r=0.90 and r=0.88 respectively) and Temp3pm (r=0.70 and r=0.97), confirming that daily temperatures are highly consistent. Humidity displays expected negative correlations with temperature (Humidity9am vs MaxTemp: r=-0.50) and positive correlations with rainfall (r=0.39). Pressure measurements show near-perfect correlation between morning and afternoon readings (r=0.96) but minimal relationships with other variables. Wind speeds demonstrate moderate internal consistency (WindGustSpeed vs WindSpeed3pm: r=0.66), while cloud cover correlates negatively with sunshine (Cloud3pm vs Sunshine: r=-0.53). The strongest inverse relationship appears between MaxTemp and Humidity3pm (r=-0.56), highlighting how afternoon heat reduces moisture in the air. Most surprising is the weak correlation between rainfall and other variables (all |r|<0.41), suggesting precipitation events are relatively independent of other measured conditions.

### 

## Summary of Feature Importance Analysis

### Correlation Analysis with the Target Variable

The target variable, Rain Tomorrow, was assessed for its correlation with various numerical features.

**Numerical Features Correlation with Target Variable**

|  |  |  |
| --- | --- | --- |
| Numerical Features | Correlation | Inference |
| Humidity3pm | 0.43 | Strong Positive Correlation |
| Rainfall | 0.32 | Moderate Positive Correlation |
| RainToday | 0.30 | Moderate Positive Correlation |
| Cloud3pm | 0.29 | Moderate Positive Correlation |
| Humidity9am | 0.25 | Moderate Positive Correlation |
| Cloud9am | 0.24 | Moderate Positive Correlation |
| WindGustSpeed | 0.22 | Moderate Positive Correlation |
| WindSpeed9am | 0.08 | Weak/Neutral Correlation |
| WindSpeed3pm | 0.08 | Weak/Neutral Correlation |
| MinTemp | 0.08 | Weak/Neutral Correlation |
| Temp9am | -0.02 | Weak/Neutral Correlation |
| Evaporation | -0.10 | Weak/Neutral Correlation |
| MaxTemp | -0.15 | Weak/Neutral Correlation |
| Temp3pm | -0.19 | Weak/Neutral Correlation |
| Pressure3pm | -0.21 | Moderate Negative Correlation |

**Key Observations from Correlation Analysis:**

**Highly Significant Features:**

* **Humidity3pm (0.44)**: The strongest predictor of rain tomorrow, indicating that high afternoon humidity significantly increases precipitation likelihood.
* **Rainfall (0.32) & RainToday (0.31)**: Current/preceding rain events are strongly tied to future rainfall, reflecting weather persistence.

**Moderately Significant Features:**

* **Cloud Cover (Cloud3pm: 0.30, Cloud9am: 0.25)**: Overcast skies correlate with higher rain chances, though less decisively than humidity.
* **Humidity9am (0.26)**: Morning moisture levels contribute moderately to rain prediction.
* **WindGustSpeed (0.22)**: Strong gusts weakly signal stormy conditions.

**Significantly Weak Features:**

* **Wind Speeds (9am/3pm: 0.08–0.09) & MinTemp (0.08)**: Minimal impact on rain prediction.
* **Temp9am (-0.03)**: Nearly neutral, suggesting morning heat alone doesn’t dictate rain.

**Irrelevant/Negative Predictors:**

* **Sunshine (-0.29) & High Temperatures (MaxTemp: -0.16, Temp3pm: -0.19)**: Dry, sunny conditions reduce rain probability.
* **Pressure (-0.21 to -0.23)**: Higher pressure systems typically suppress rainfall.

## Conclusion:

The analysis underscores the critical role of humidity-based features in predicting rainfall, with Humidity3pm emerging as the strongest predictor, reflecting how afternoon moisture levels directly influence precipitation likelihood. Rainfall and RainToday further enhance the model by capturing immediate weather persistence, while cloud cover, Cloud3pm, provides secondary contextual signals.

Moderately useful features like Humidity9am and WindGustSpeed contribute auxiliary insights but are less decisive. In contrast, temperature variables MaxTemp, Temp3pm, and pressure exhibit inverse relationships with rain, highlighting their utility in identifying dry conditions rather than wet ones. Notably, wind speeds and Sunshine proved minimally informative or counter-predictive.