Group 16

Professor Cai

STA4724

7 Dec 2024

Final Project Report

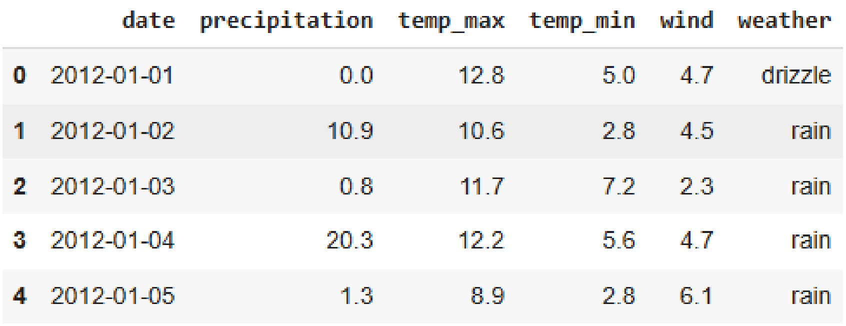
Introduction/Background:

Seattle is renowned for its distinctive weather, characterized by frequent rainfall and moderate temperatures throughout the year. On average, the city experiences approximately 152 days of rain annually, alongside 2,019 hours of sunshine. With an annual precipitation of about 39.34 inches, the majority of this falls as rain rather than snow. The climate is relatively mild, with an average annual high temperature of 59°F and a low of 45°F, making Seattle a city of cool and damp conditions year-round.

In our research, we aim to analyze the relationship between wind patterns and temperature in a city with consistent rainfall, focusing on how these factors influence both the quantity and type of precipitation. Using this analysis, we will develop models to predict the weather for the two days immediately following the conclusion of our dataset. These predictions will then be compared to actual recorded data to evaluate the accuracy and effectiveness of our approach. This study not only seeks to deepen our understanding of weather dynamics in Seattle but also to enhance predictive capabilities for similar climatic regions.

Exploratory Data Analysis:

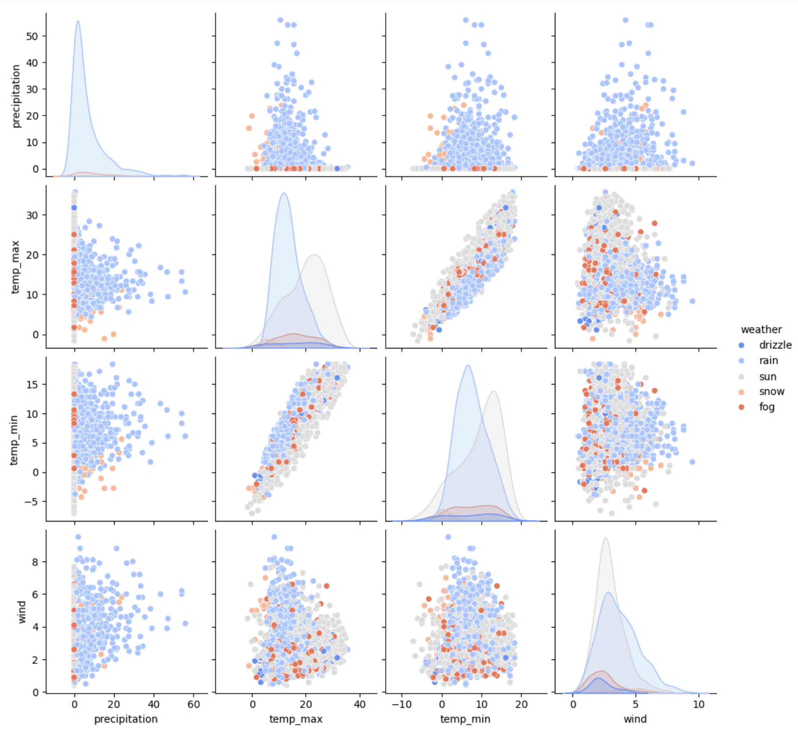
This dataset was found on Kaggle. The following is the first few lines of the dataset.



	date	precipitation	temp_max	temp_min	wind	weather
0	2012-01-01	0.0	12.8	5.0	4.7	drizzle
1	2012-01-02	10.9	10.6	2.8	4.5	rain
2	2012-01-03	0.8	11.7	7.2	2.3	rain
3	2012-01-04	20.3	12.2	5.6	4.7	rain
4	2012-01-05	1.3	8.9	2.8	6.1	rain

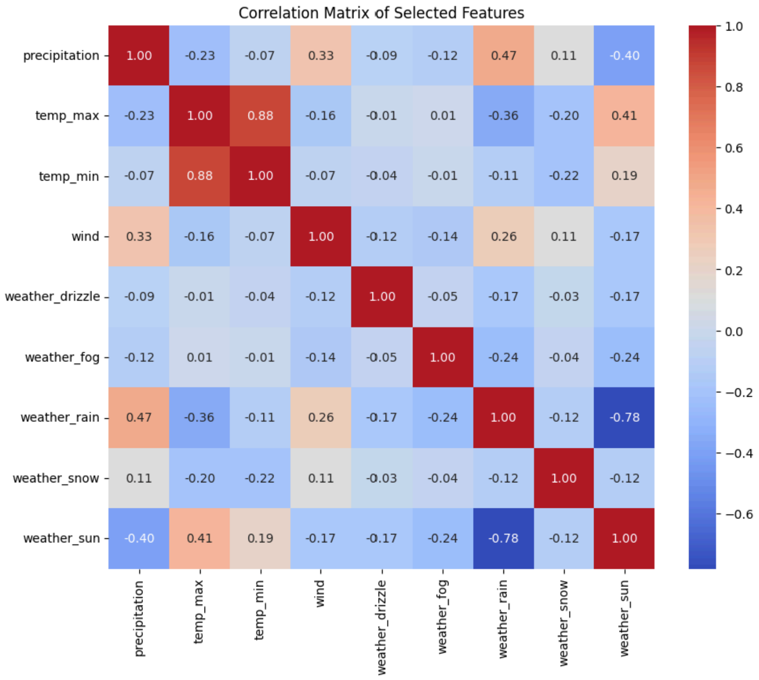
The dataset captures participation in centimeters, temperature in Celsius, and wind speed in kilometers per hour, along with corresponding weather conditions and dates. In the following exploratory data analysis (EDA), the weather variable and frequency of each variable will serve as the response variable, while maximum temperature, minimum temperature, precipitation, and wind speed will function as independent variables. As part of the data preprocessing, the weather variable will require cleaning to ensure accuracy and consistency

Pairplot



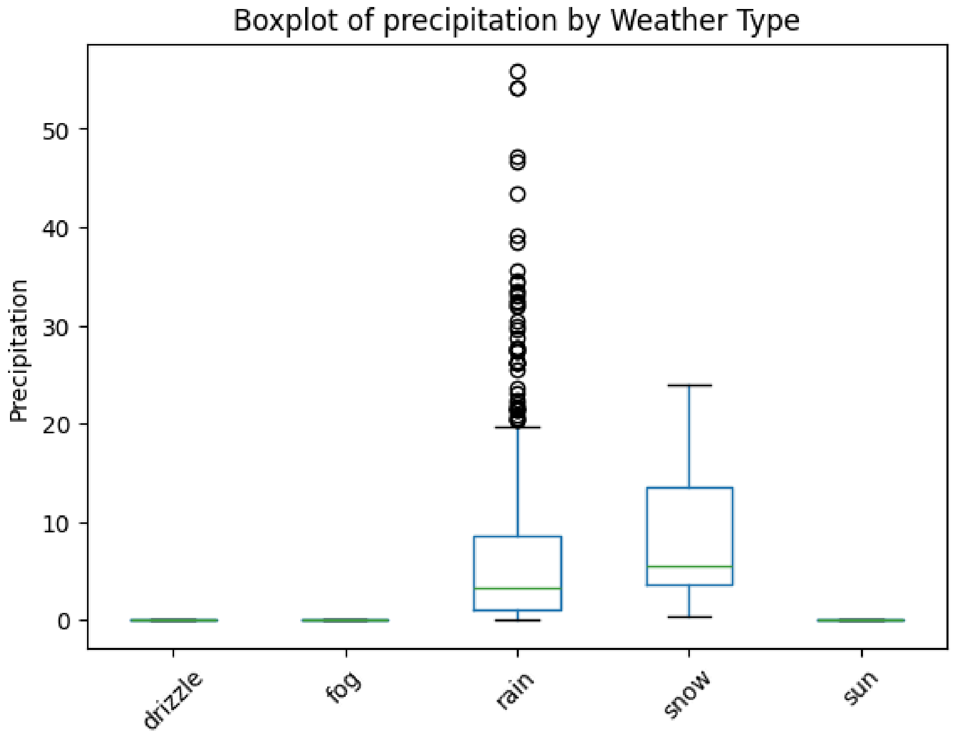
Within the pairplot it is seen that there is no strong linear relationship between precipitation and temperature (both maximum and minimum).But precipitation tends to be higher at lower temperatures. A strong positive linear relationship exists between maximum and minimum temperatures, as expected. Higher maximum temperatures correlate with higher minimum temperatures. However, wind speed does not show a clear relationship with precipitation or temperature, indicating it might be influenced by other factors. When adding in the weather types, snow (grey) and fog (orange) appear to be associated with lower temperatures, while drizzle and rain (blue and light grey) are dispersed across a wider range of precipitation levels. Also as expected, sunny weather (light grey) appears to have little to no precipitation and is more common at higher temperatures. Lastly, clusters are visible in precipitation and temperature variables, potentially distinguishing different weather types. This is also seen in snow and fog which tend to cluster at lower temperatures, while rain and drizzle have broader distributions.

Heatmap

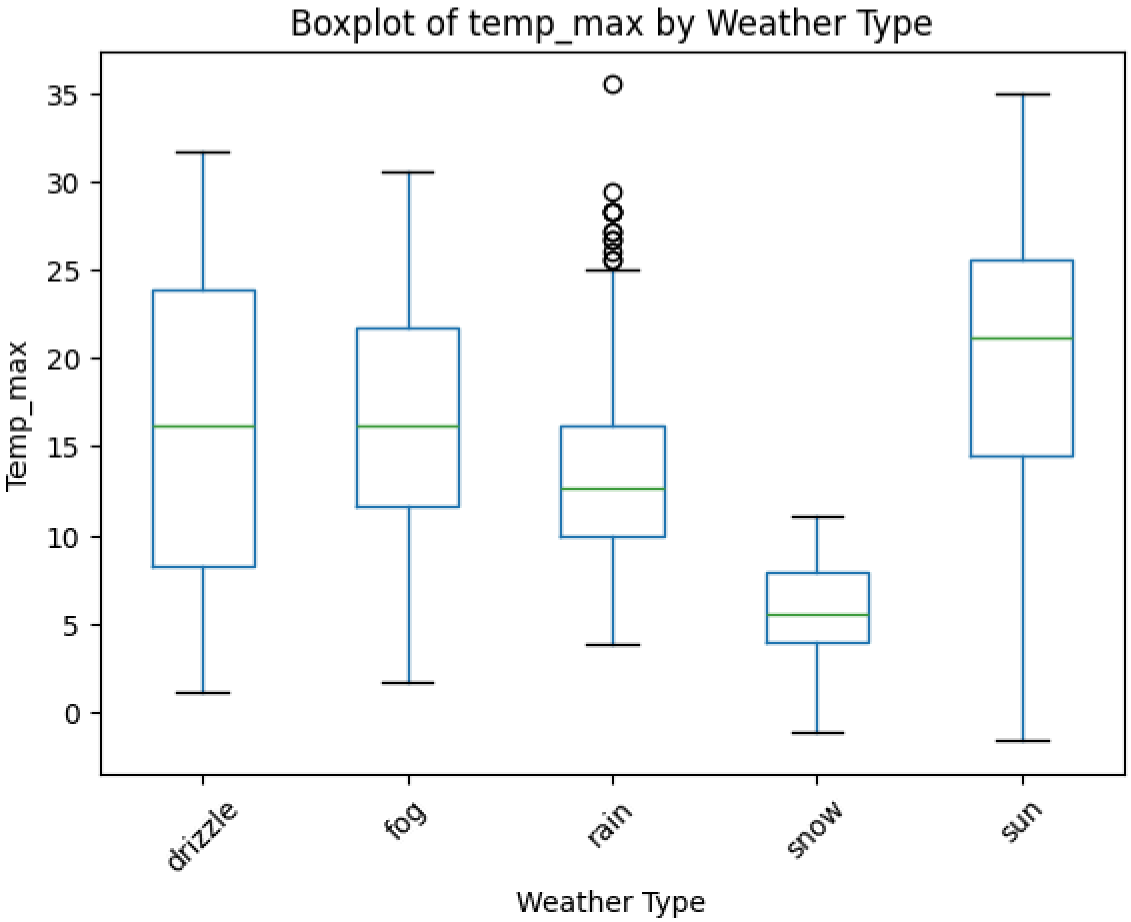


The correlation matrix highlights notable relationships among the selected features. Maximum and minimum temperatures show a strong positive correlation, showing that higher maximum temperatures are typically accompanied by higher minimum temperatures. Precipitation is moderately correlated with rain but negatively correlated with sunny weather, reflecting expected weather patterns. Sunny conditions are positively associated with higher maximum temperatures, while snow is weakly associated with lower minimum temperatures. Wind speed exhibits low correlation with most variables, suggesting it is relatively independent. Additionally, weather categories such as rain, sun, drizzle, snow, and fog show mutual exclusivity, as evidenced by negative correlations among them. These insights reveal clear relationships between weather conditions, temperature, and precipitation, which can later inform predictive modeling and further analysis.

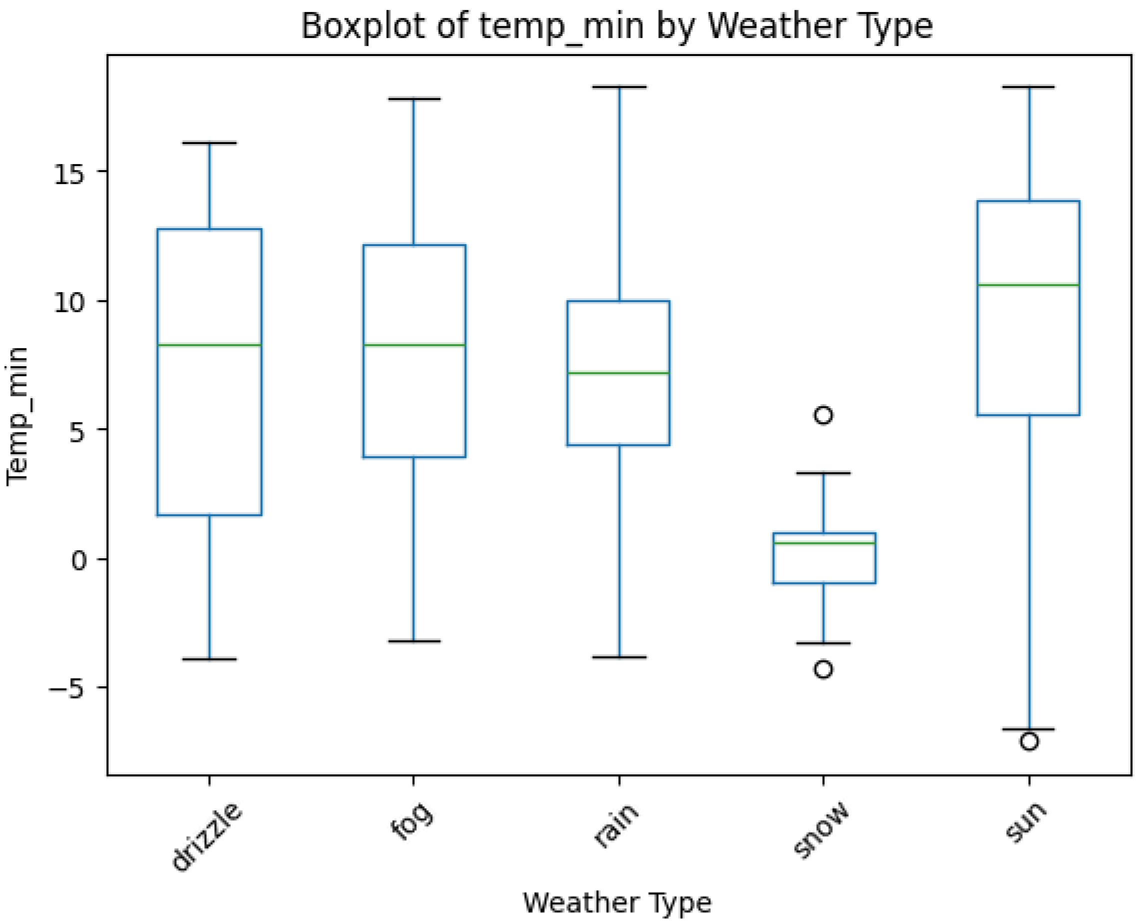
Box Plots



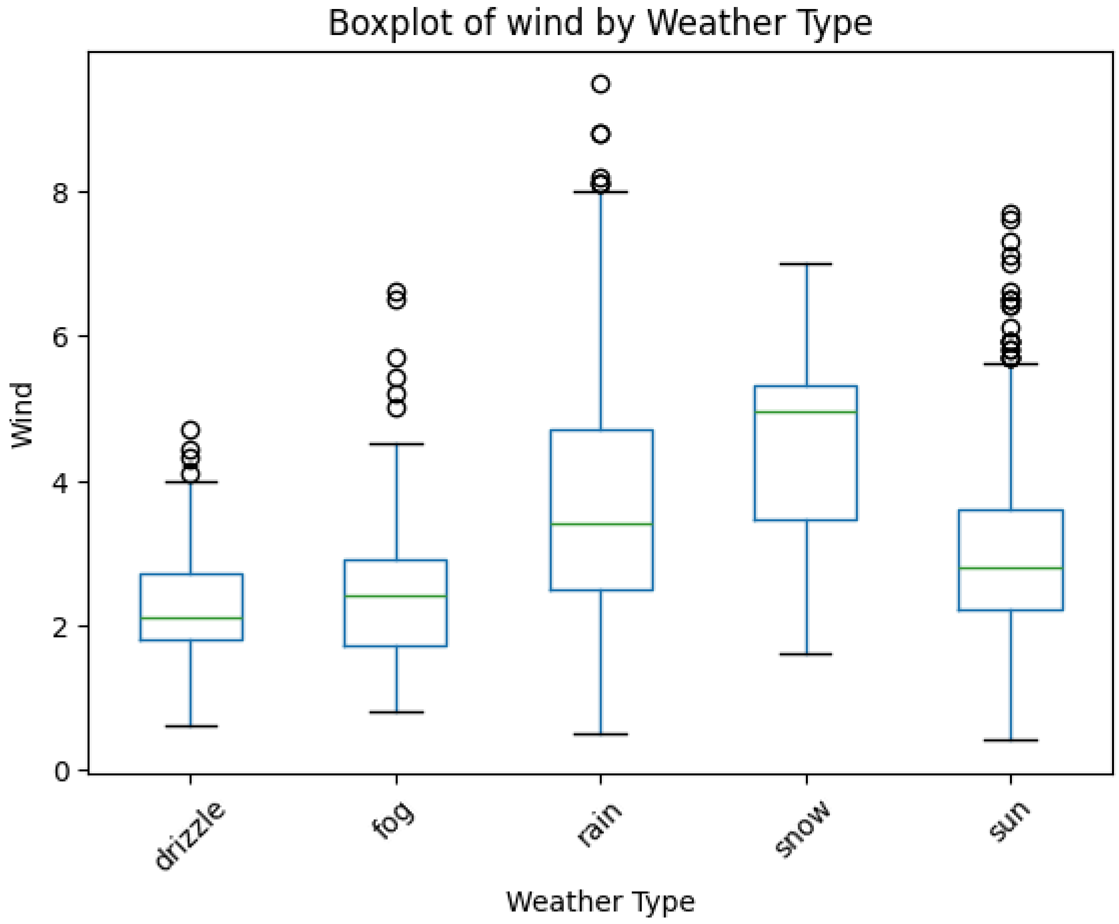
The boxplot reveals that rain and snow are the primary contributors to significant precipitation, with rain having the highest median and variability, along with frequent extreme outliers. Snow shows a slightly lower median and variability, with fewer outliers. In contrast, drizzle, fog, and sunny weather are associated with little or no precipitation, showing minimal variability and no notable outliers. This highlights rain and snow as the dominant drivers of precipitation, with rain showing the most variability and extremes.



The boxplot shows clear temperature patterns across weather types: sunny weather is associated with the warmest conditions, while snow corresponds to the coldest. Drizzle, fog, and rain occur under moderate temperature ranges with some overlap. Outliers in rainy weather suggest occasional higher temperatures during precipitation.

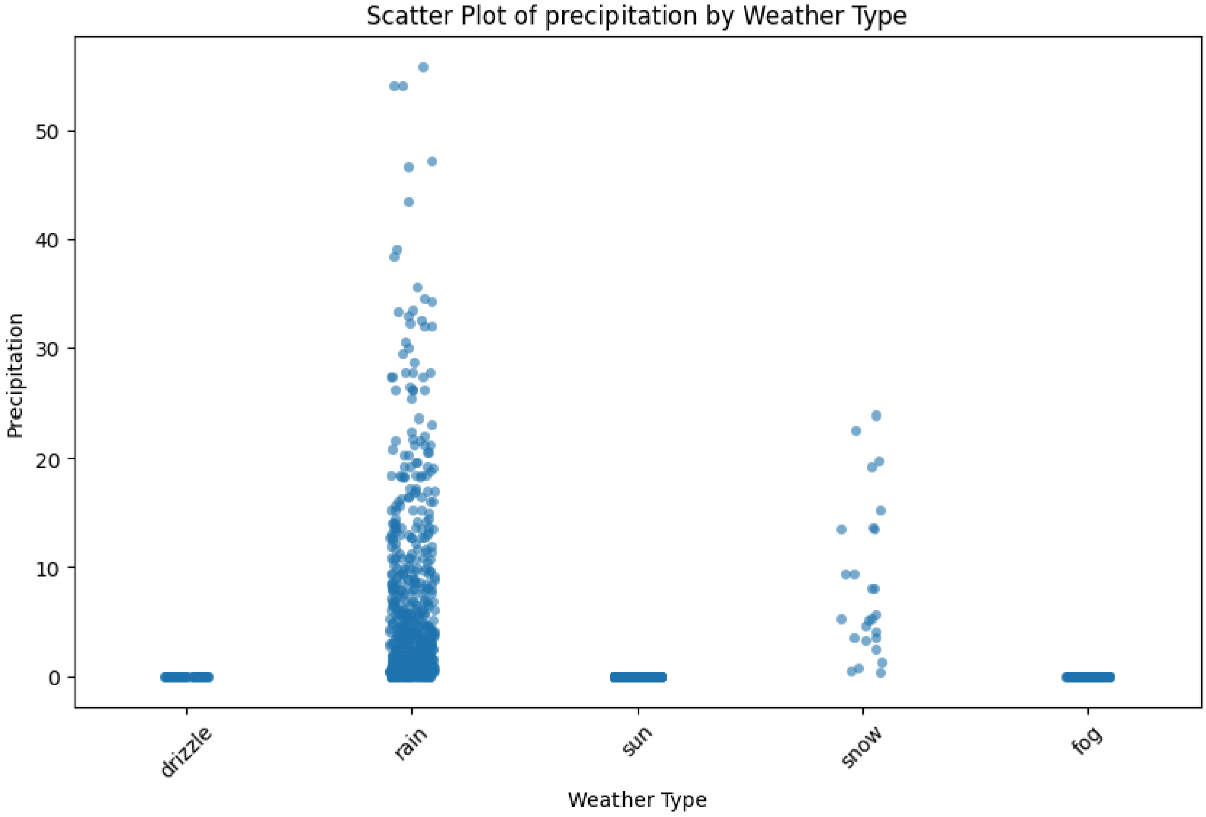


The boxplot shows that snow is associated with the lowest minimum temperatures, while sunny weather corresponds to the highest. Drizzle, fog, and rain occur under moderate minimum temperature ranges with some overlap. Outliers in snow indicate occasional extreme cold, while sunny conditions exhibit variability with generally warmer nights.

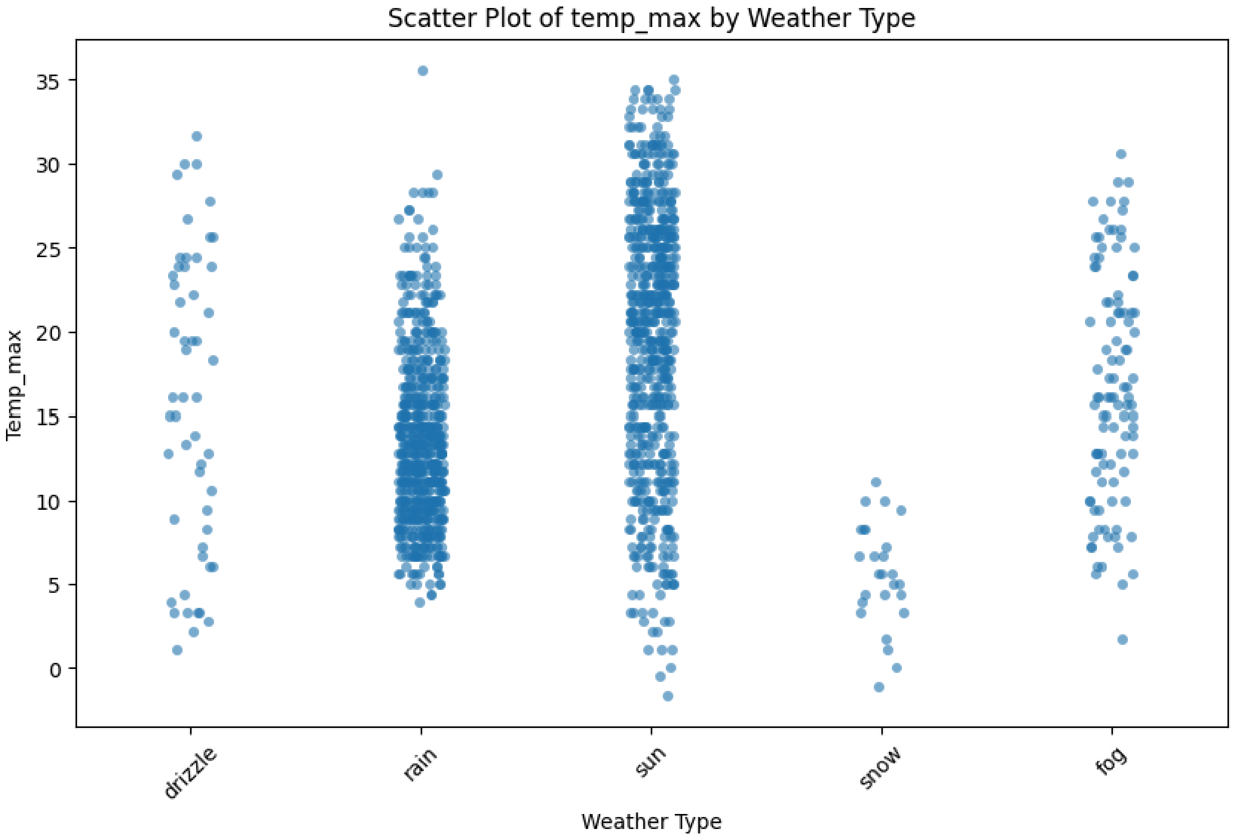


The boxplot reveals that snow is associated with the lowest minimum temperatures, while sunny weather corresponds to the highest. Drizzle, fog, and rain occur under moderate minimum temperature ranges with some overlap. Outliers in snow indicate occasional extreme cold, while sunny conditions exhibit variability with generally warmer nights.

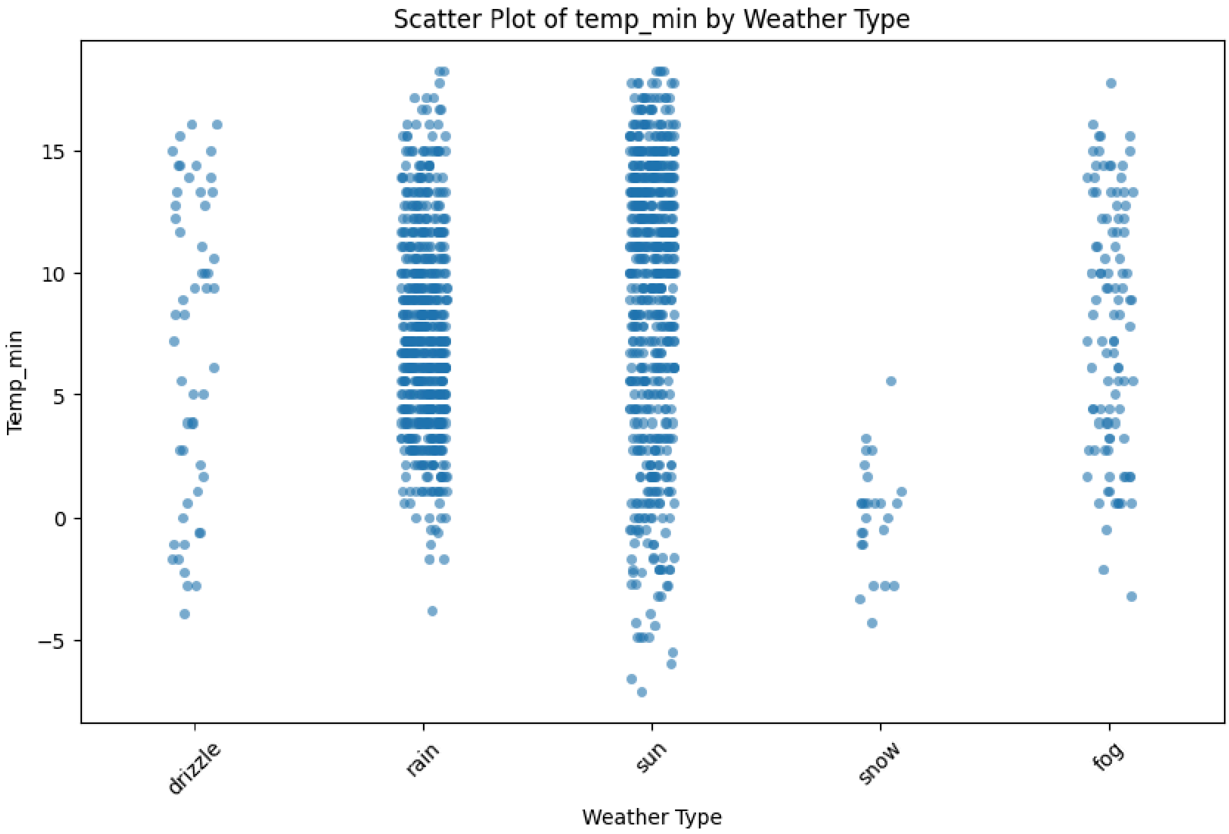
Scatter Plots



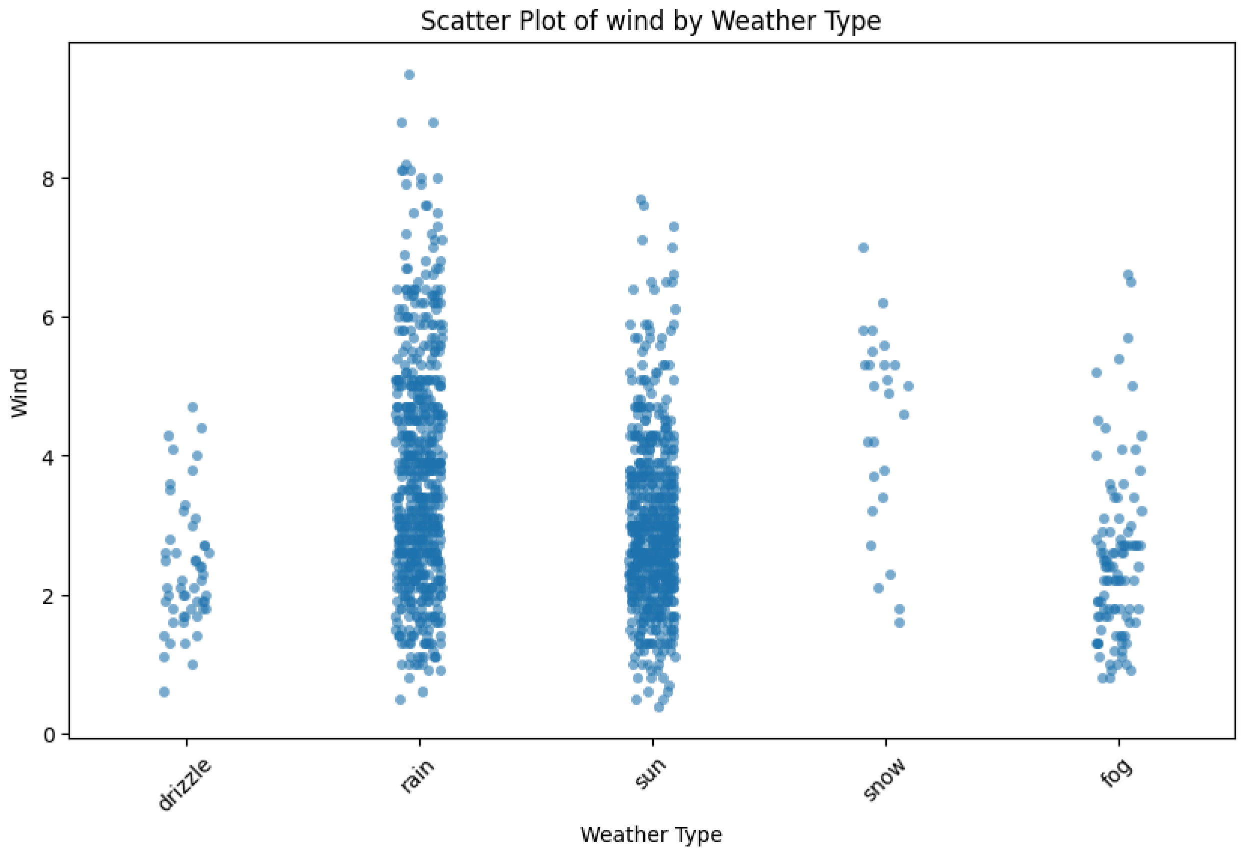
The scatter plot highlights that rain is the dominant driver of precipitation, exhibiting both high variability and extreme values. Snow adds moderate precipitation levels but occurs less frequently. Drizzle, sunny, and foggy weather are associated with minimal or no precipitation, reflecting their relatively dry nature.



The scatter plot shows the distinct temperature patterns associated with different weather types. Snow corresponds to the lowest maximum temperatures, while sunny weather exhibits the highest. Drizzle, fog, and rain share overlapping moderate temperature ranges, with slight variations in their distributions. This plot confirms the strong influence of weather type on temperature conditions.

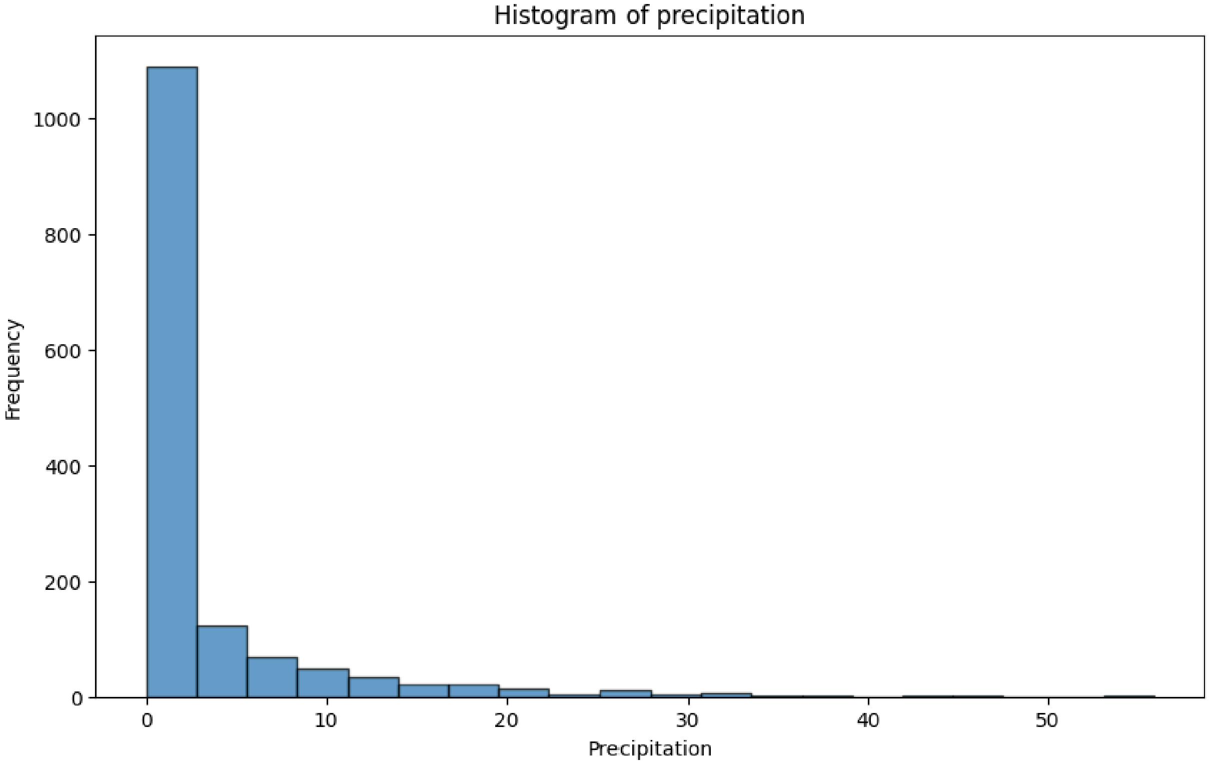


This scatter plot shows the relationship between weather types and minimum temperatures. Snow is consistently associated with the lowest minimum temperatures, while sunny weather corresponds to the warmest. Drizzle, fog, and rain overlap in moderate temperature ranges, with fog showing slightly more variability. This plot reinforces the distinct temperature conditions linked to each weather type.

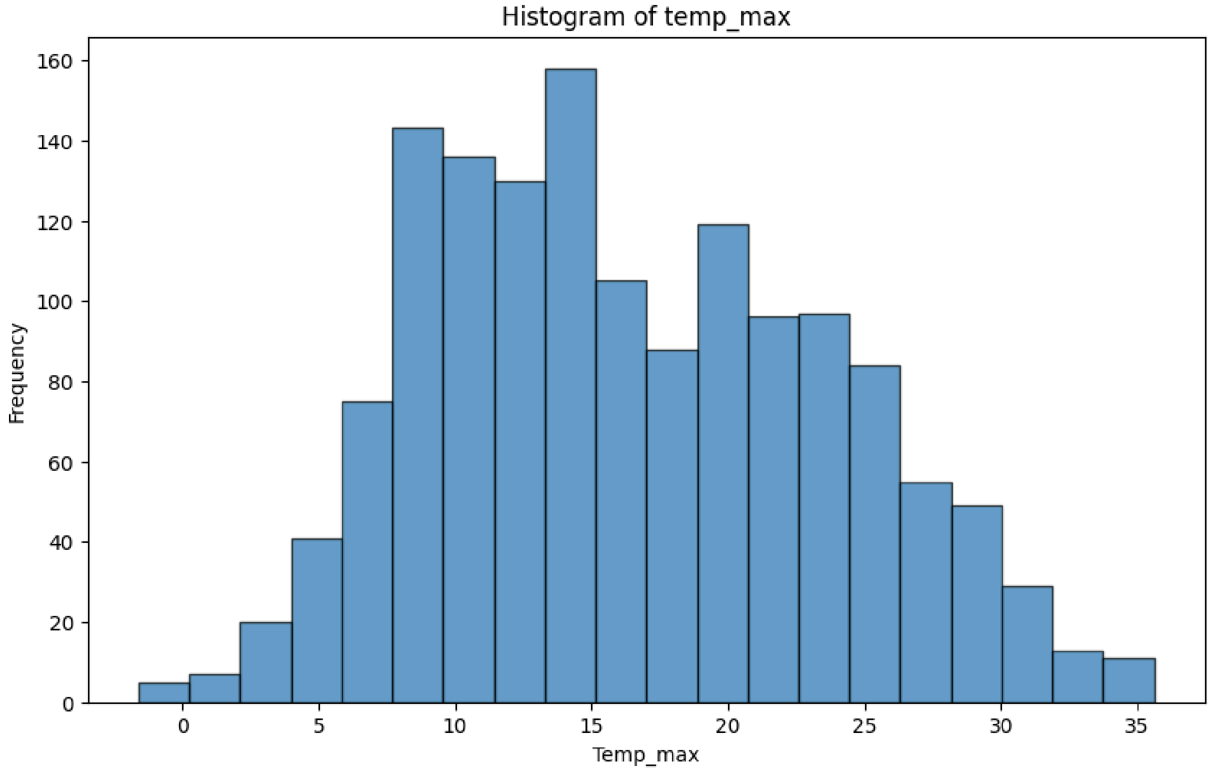


The scatter plot shows that rain and sunny weather are associated with a wide range of wind speeds, including occasional high values. Drizzle, snow, and fog are linked to lower and more consistent wind speeds, with fog showing the calmest and most stable conditions. This shows the variability in wind speeds across different weather types, with certain types like rain highlighting higher extremes.

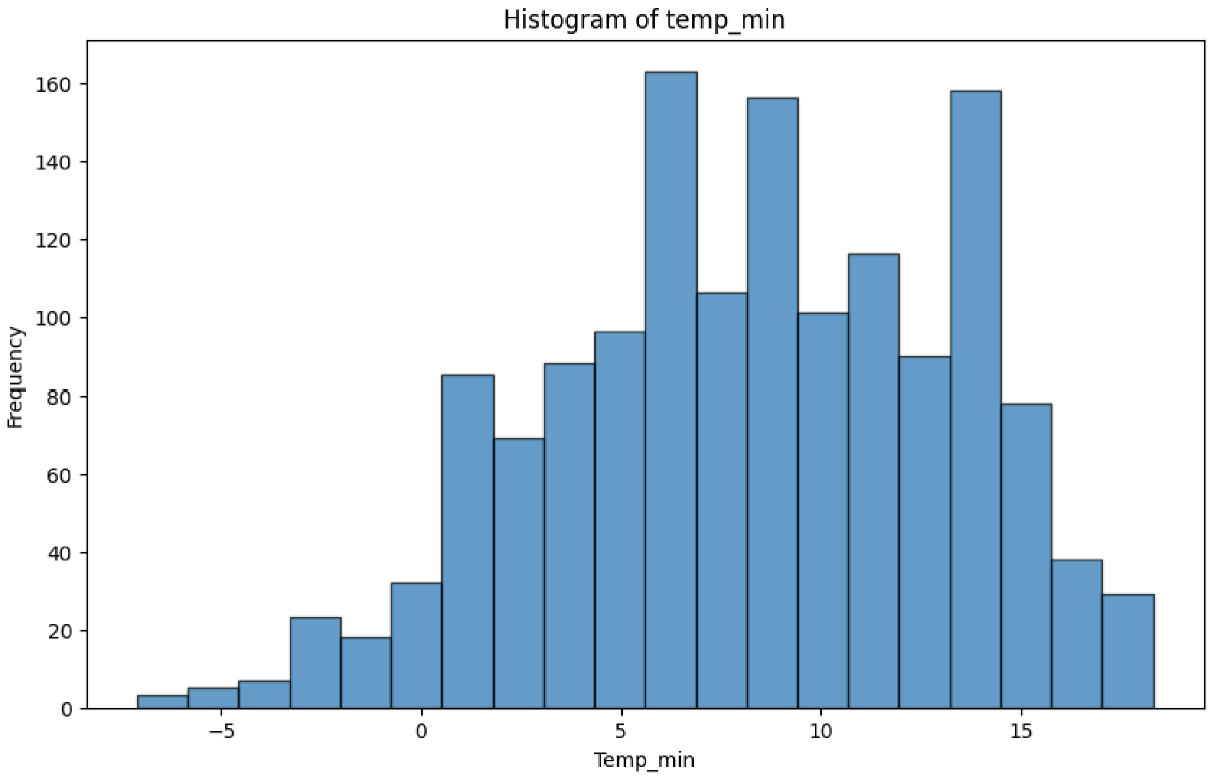
Histograms



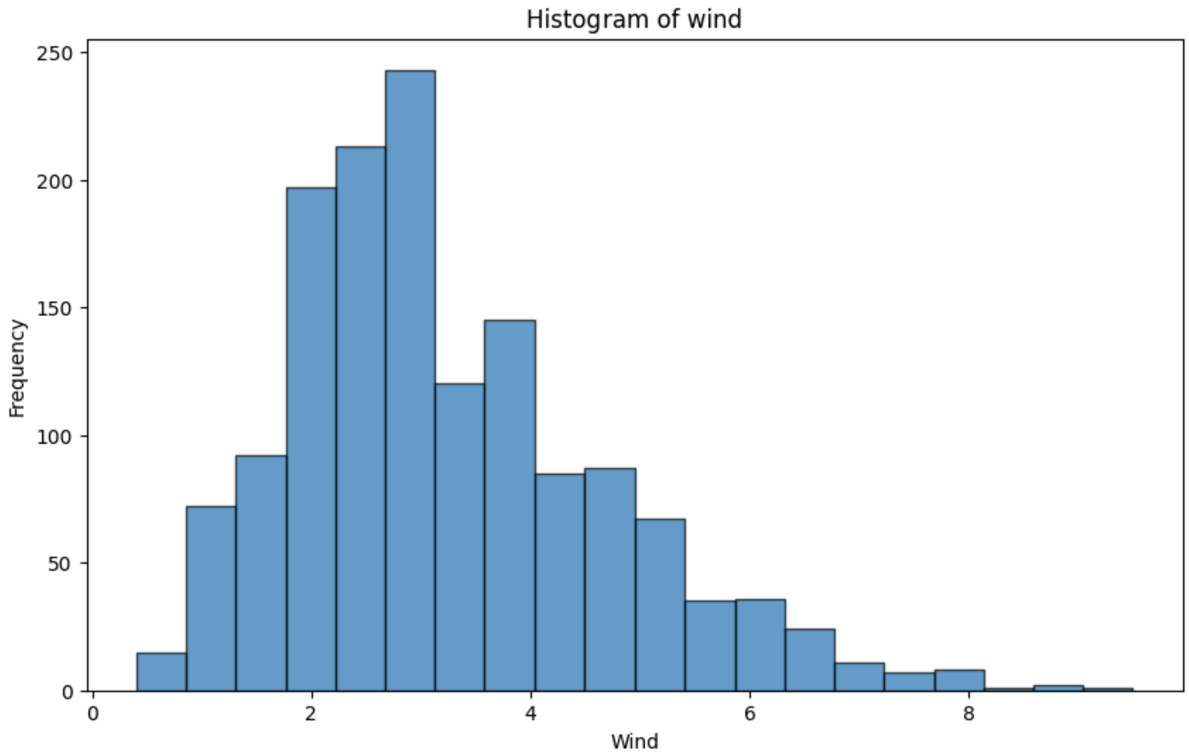
This histogram shows that precipitation events are predominantly light or nonexistent, with a few instances of moderate precipitation and very rare cases of extreme precipitation. The right-skewed distribution emphasizes the rarity of high precipitation levels in the dataset.



This histogram shows that maximum temperatures are mostly distributed around 10 degrees celsius to 20 degrees celsius with 15 degrees celsius being the most frequent. While extreme maximum temperatures (both very low and very high) are relatively rare, the dataset spans a wide range of values. The distribution is nearly normal, with a slight skew toward higher temperatures.

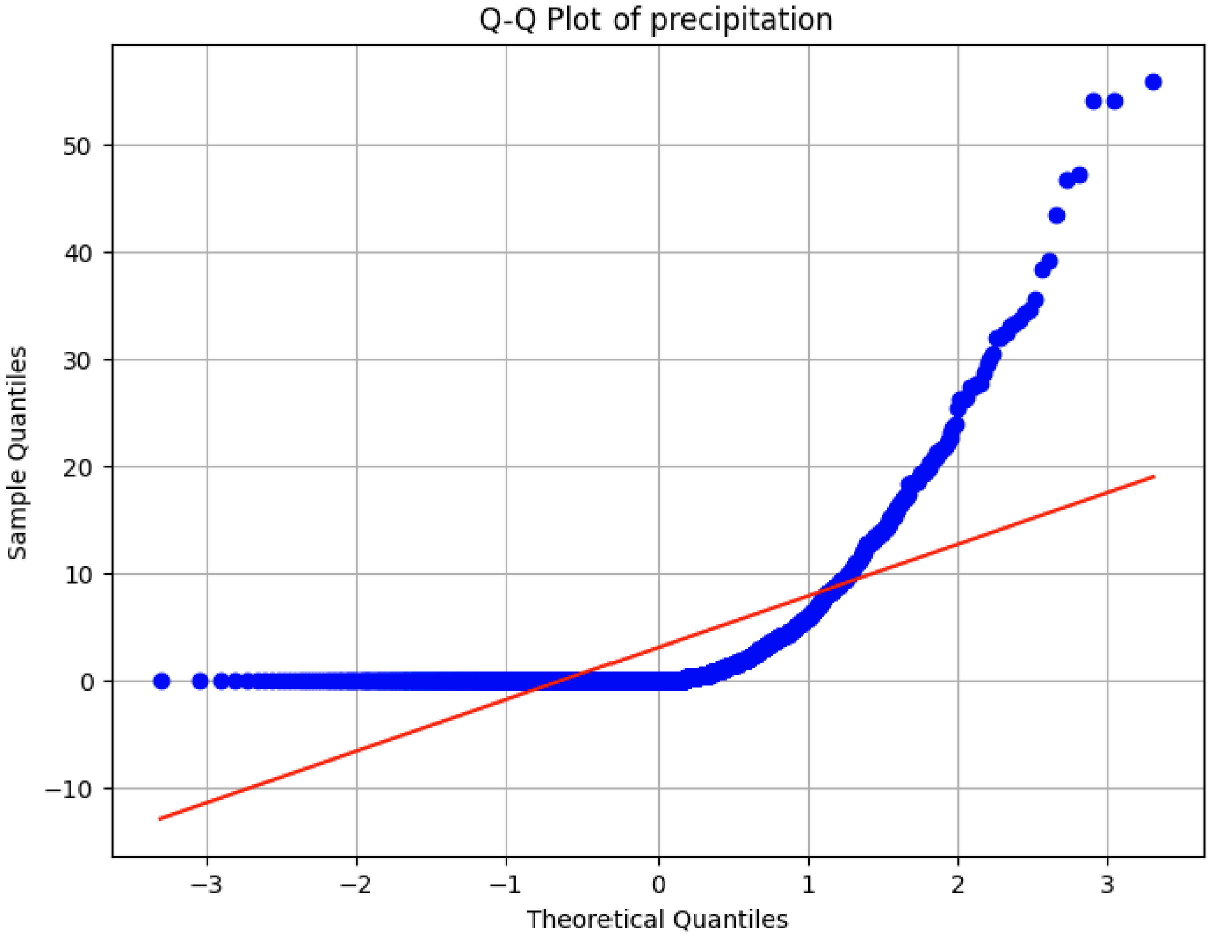


The histogram shows that minimum temperatures are most frequently in the range of 5 degrees celsius to 10 degrees celsius, with a moderate spread from below -5 degrees celsius to 15 degrees celsius. Sub-zero minimum temperatures are relatively rare, while the dataset exhibits a roughly symmetrical distribution centered around 8 degrees celsius. This suggests that moderate minimum temperatures are the norm, with occasional colder conditions.

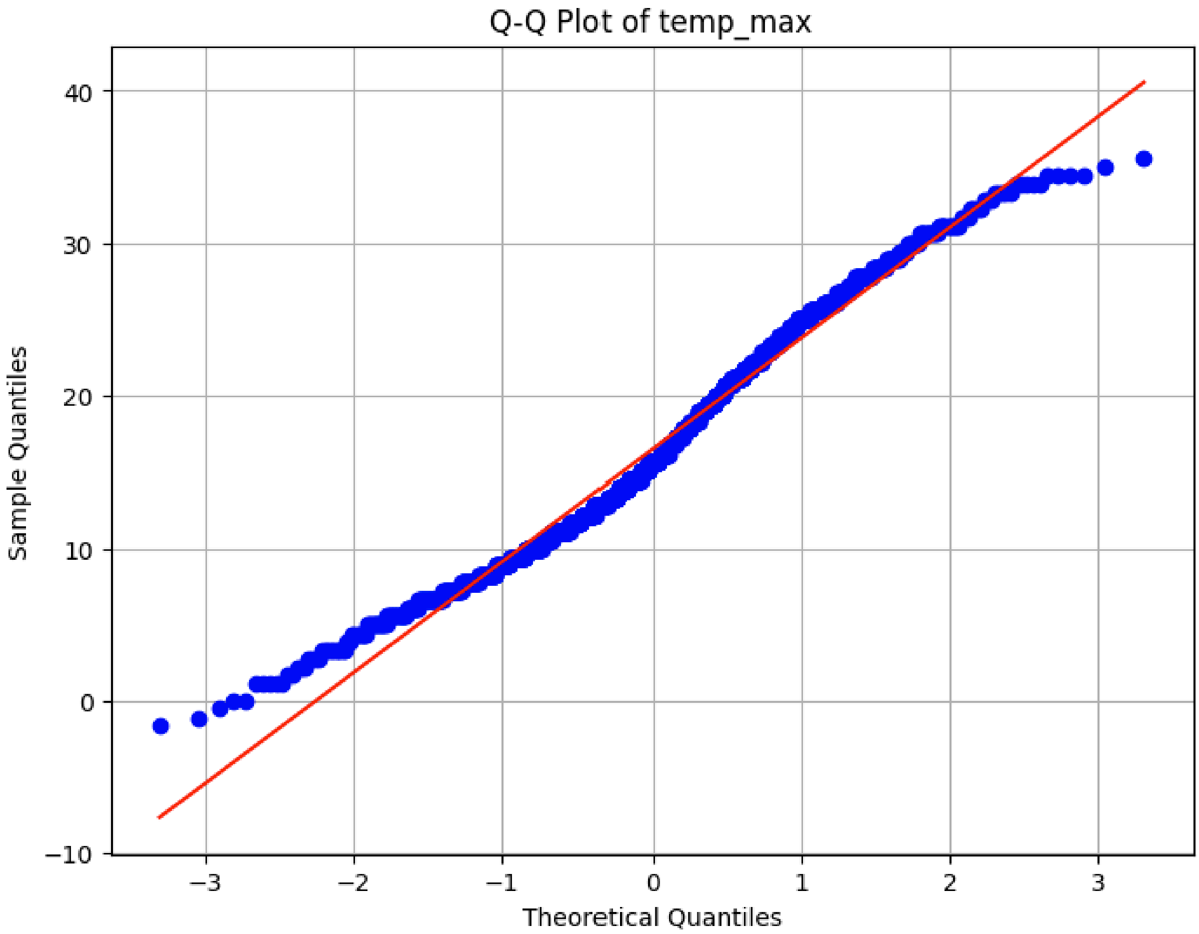


This histogram shows that moderate wind speeds (2–4 km/h) are the most common, with a peak around 3 km/h. Wind speeds exhibit a slight right-skew, with higher speeds above 5 km/h occurring less frequently and very low speeds near 0 km/h being rare. This distribution highlights a predominance of calm to moderate wind conditions in the dataset.

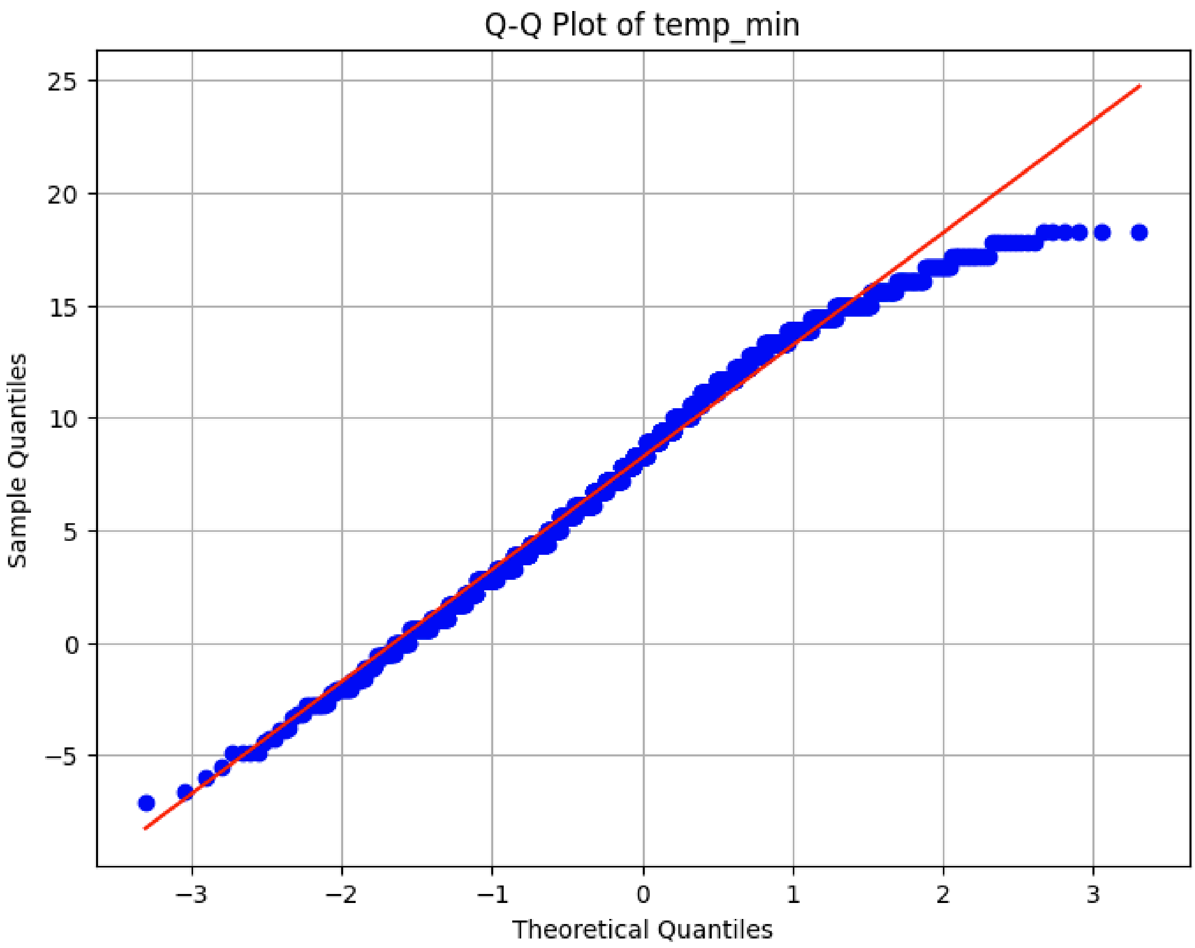
Q-Q Plots



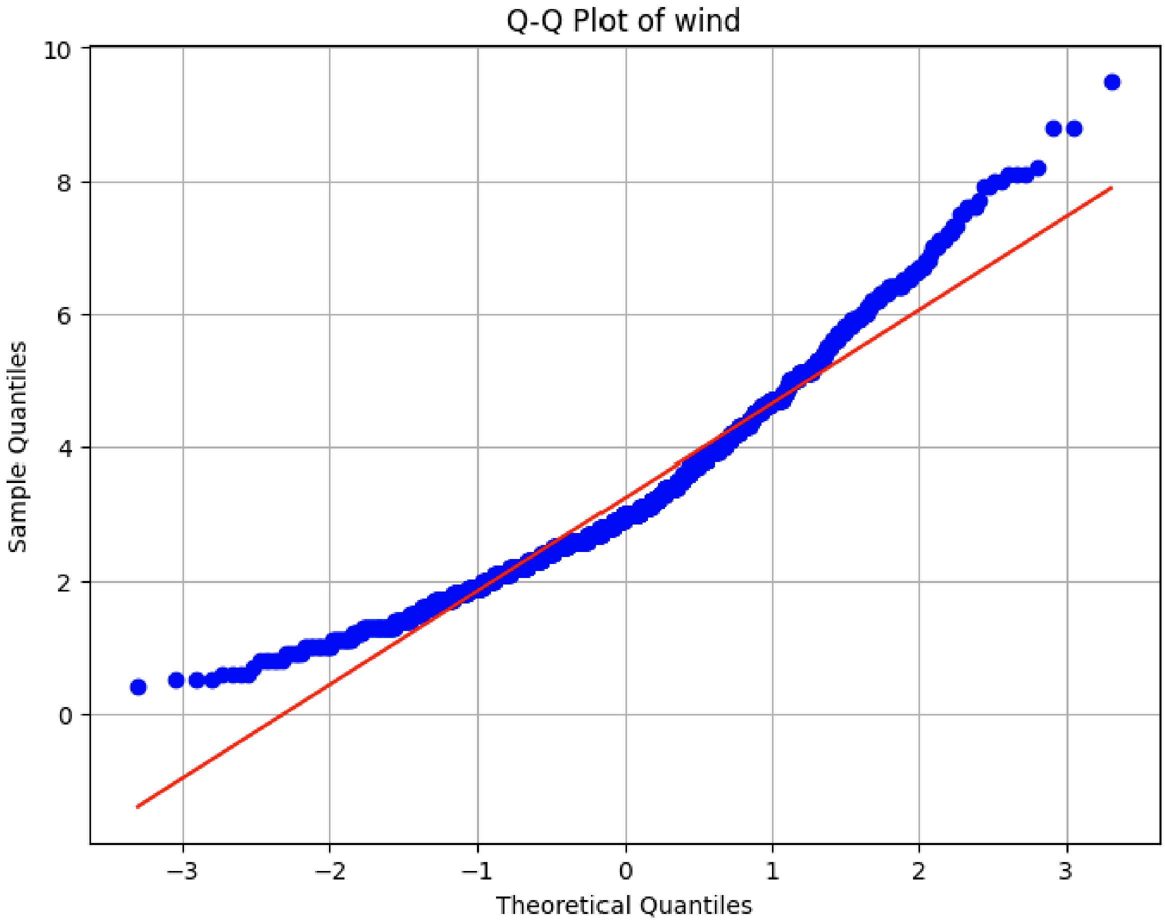
This Q-Q plot confirms that the precipitation data is strongly right-skewed and not normally distributed. Extreme precipitation values are much more frequent than expected under a normal distribution, while lower values (including zero precipitation) are relatively common and align more closely with a normal distribution. This skewness suggests the need for transformations or non-parametric approaches for analysis.



This Q-Q plot shows that the maximum temperature data is approximately normally distributed, with minor deviations in the tails. There are slightly more extreme values at both ends of the distribution than a perfect normal model would predict. These deviations suggest that while the data is mostly normal, careful consideration of the extreme values may be necessary for further analysis.

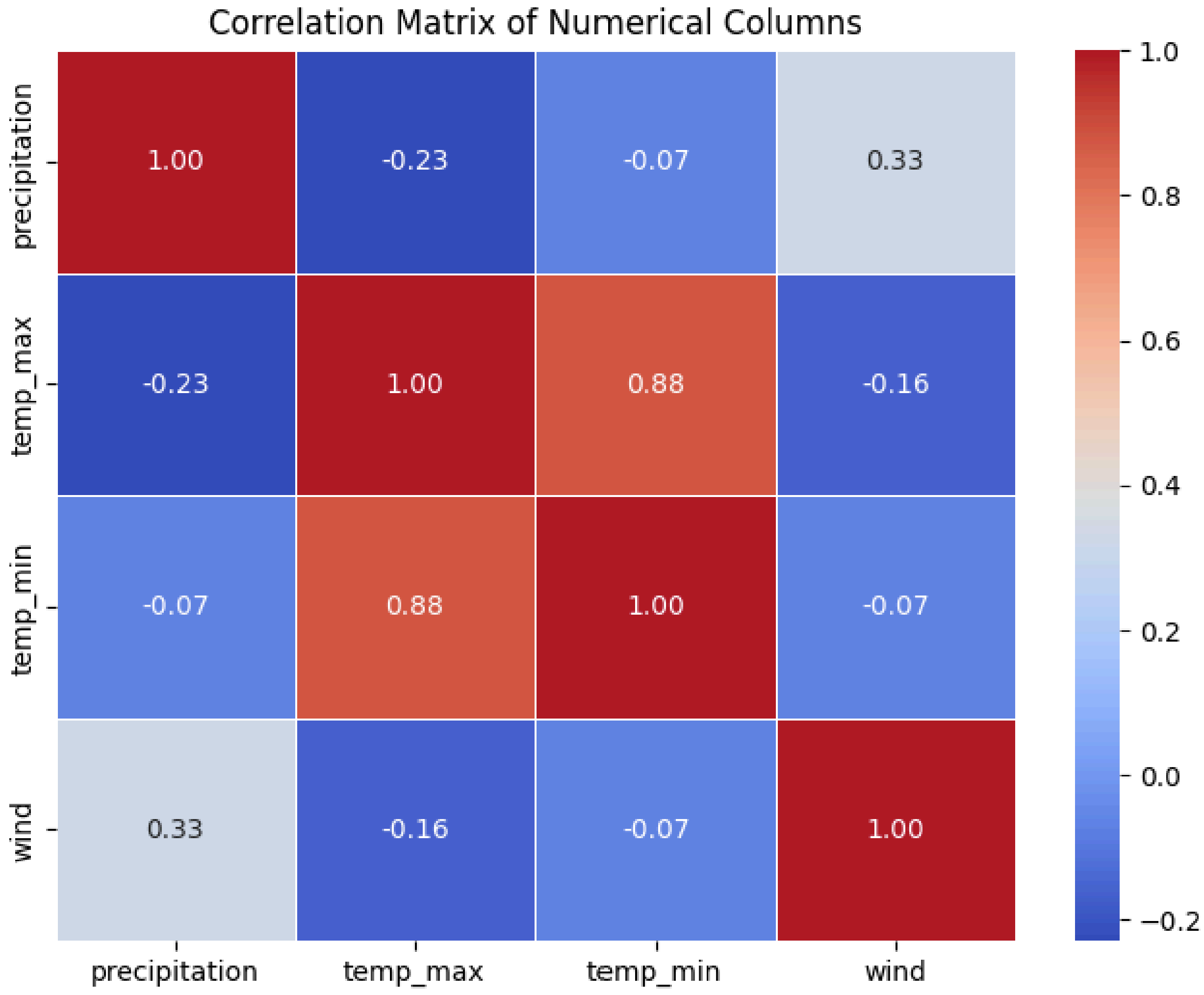


This Q-Q plot indicates that the minimum temperature data is approximately normally distributed, with minor deviations at the tails. Higher and lower extremes occur slightly more frequently than expected under a normal distribution, but the central range aligns closely with the theoretical model. These findings suggest that the data can mostly be treated as normal, with some attention given to the extremes.



This Q-Q plot suggests that wind speed data is approximately normal in the central range but shows moderate deviations in the tails. Higher-than-expected extreme wind speeds are observed in the upper tail, while slightly more frequent low wind speeds are seen in the lower tail. Overall, the data largely follows a normal distribution, with minor tail deviations.

Correlation Matrices



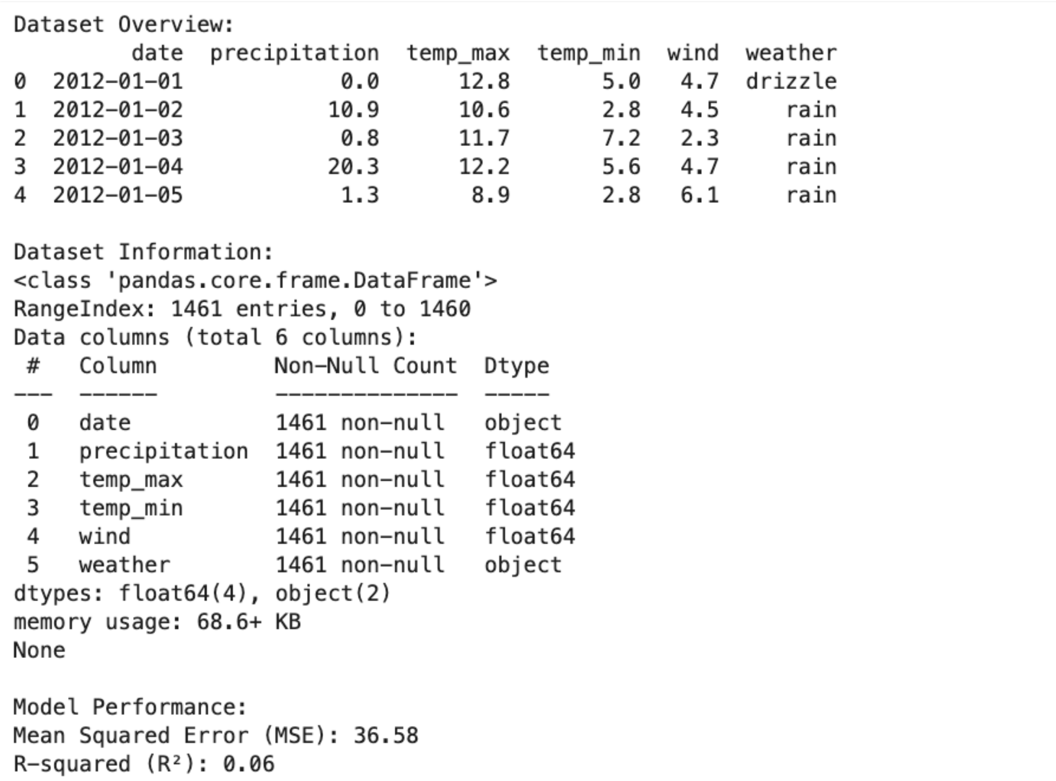
The correlation matrix highlights a strong relationship between maximum and minimum temperatures and a moderate positive relationship between precipitation and wind. Other variables, such as precipitation and temperatures, exhibit weak or negligible correlations, suggesting limited linear associations. These insights provide a basis for identifying key variable interactions in the dataset.

Data Processing:

During the data processing stage, the dataset was first examined for missing values, and none were identified. The weather column was then encoded to represent different weather categories in separate columns, with each column indicating the presence of a specific weather condition using True or False. These values were subsequently transformed into numerical representations, with True replaced by 1 and False replaced by 0, ensuring all columns were numerical.

Based on the exploratory data analysis (EDA), it was determined that the majority of the data followed a normal distribution, and there was no need to address outliers or inconsistencies. This result was anticipated, as the dataset was sourced from a reputable website that had already performed a thorough data cleaning process.

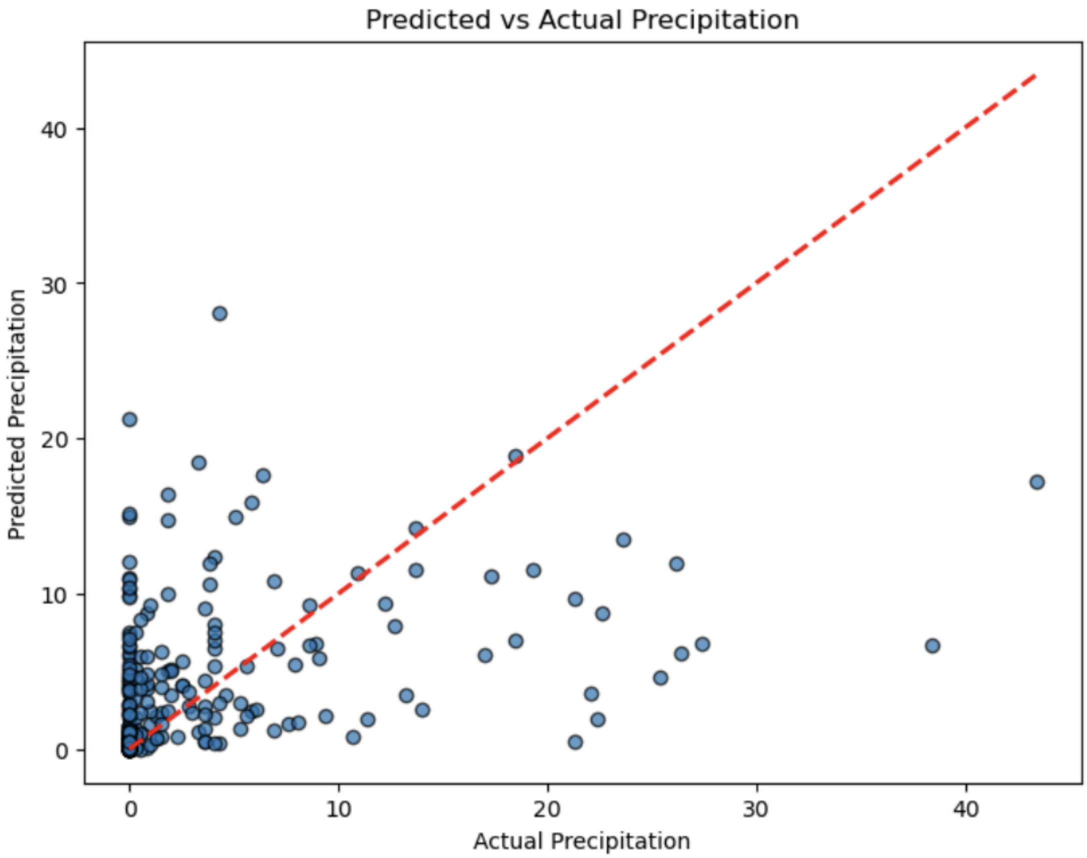
Methodology:



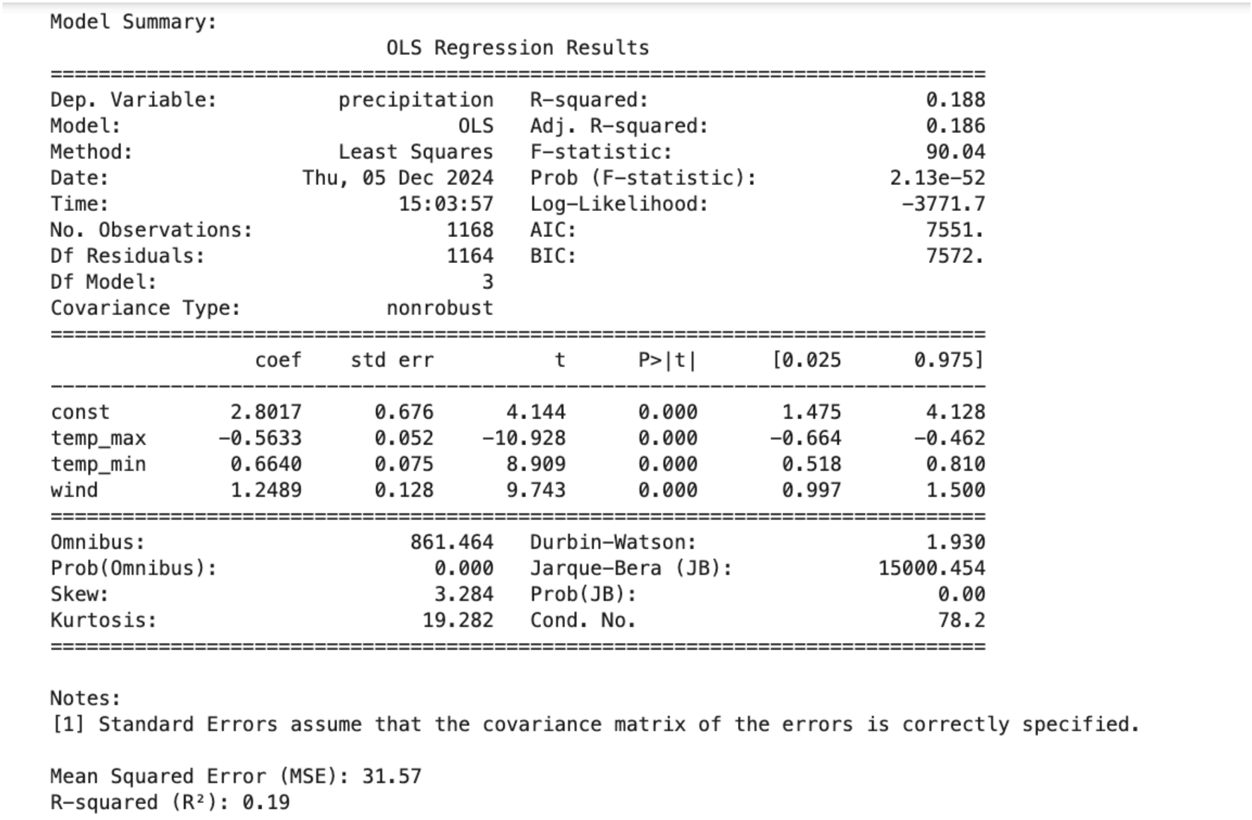
```
Dataset Overview:
         date  precipitation  temp_max  temp_min  wind  weather
0  2012-01-01            0.0      12.8       5.0   4.7  drizzle
1  2012-01-02           10.9      10.6       2.8   4.5     rain
2  2012-01-03            0.8      11.7       7.2   2.3     rain
3  2012-01-04           20.3      12.2       5.6   4.7     rain
4  2012-01-05            1.3       8.9       2.8   6.1     rain

Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1461 entries, 0 to 1460
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   date           1461 non-null   object 
 1   precipitation  1461 non-null   float64
 2   temp_max       1461 non-null   float64
 3   temp_min       1461 non-null   float64
 4   wind           1461 non-null   float64
 5   weather        1461 non-null   object 
dtypes: float64(4), object(2)
memory usage: 68.6+ KB
None

Model Performance:
Mean Squared Error (MSE): 36.58
R-squared (R²): 0.06
```

The dataset contains 1,461 rows and six columns: date, precipitation, temp_max, temp_min, wind, and weather. The date column shows the day of observation, and precipitation records the amount of rainfall in millimeters. Temp_max and temp_min represent the maximum and minimum temperatures in degrees Celsius, while wind shows the wind speed in meters per second. The weather column describes the weather conditions, such as drizzle or rain. All columns have complete data with no missing values.

A model was created to predict precipitation, but it did not perform well. The average prediction error, measured by Mean Squared Error (MSE), was 36.58, and the R-squared score, which indicates how much of the data's variability the model explains, was only 0.06. This suggests the model had limited accuracy in predicting precipitation.

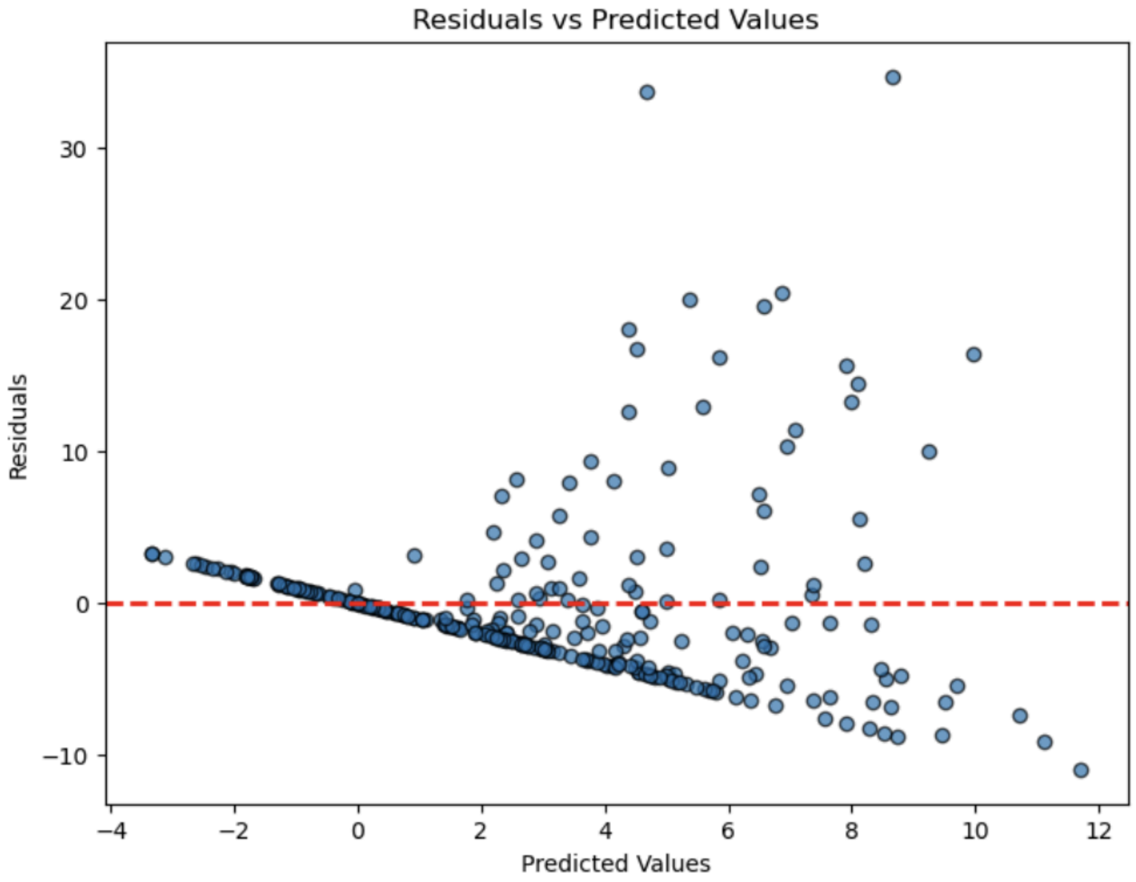


This graph shows that the model predicts low to moderate precipitation levels fairly well, with many points close to the red dashed line. However, as actual precipitation increases, the predictions become less accurate, with points spreading further from the line. This suggests the model struggles to predict higher precipitation levels and could benefit from more data or refinement to improve accuracy for extreme cases.

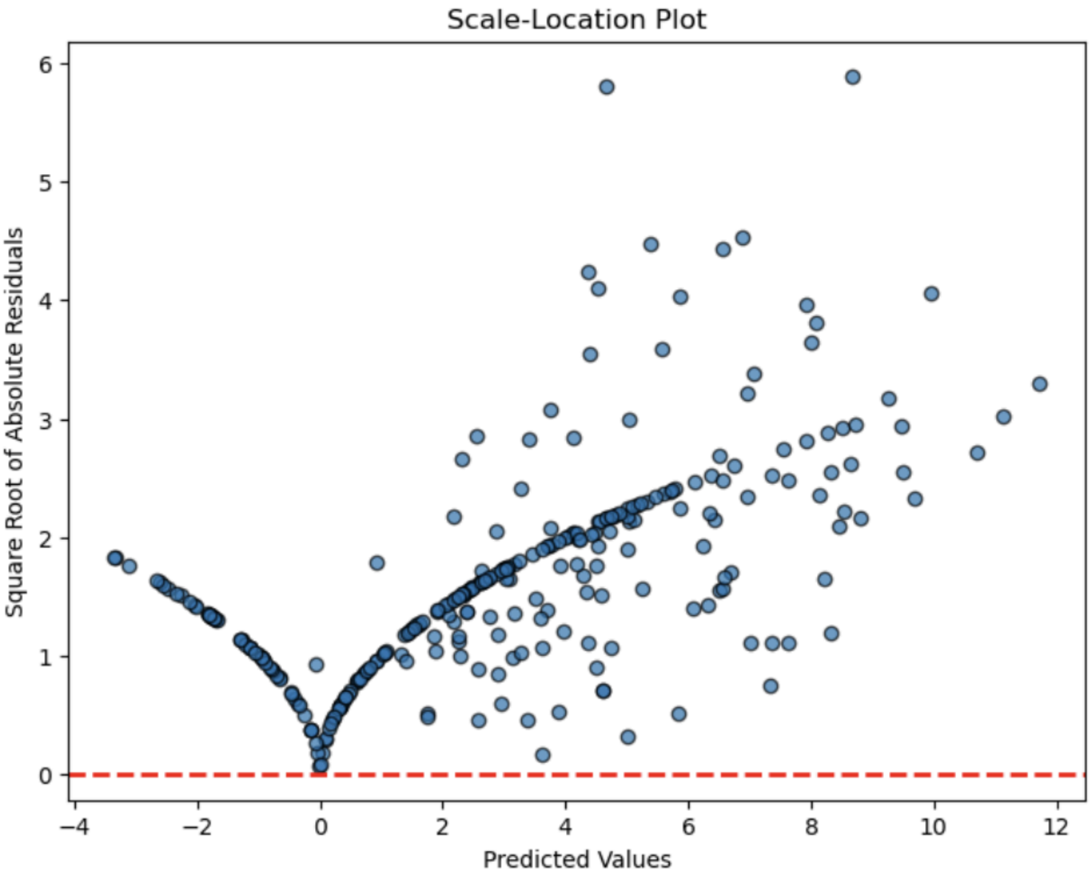


```
Model Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:          precipitation   R-squared:                       0.188
Model:                            OLS   Adj. R-squared:                  0.186
Method:                 Least Squares   F-statistic:                     90.04
Date:                Thu, 05 Dec 2024   Prob (F-statistic):           2.13e-52
Time:                        15:03:57   Log-Likelihood:                -3771.7
No. Observations:                1168   AIC:                             7551.
Df Residuals:                    1164   BIC:                             7572.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.8017      0.676      4.144      0.000       1.475       4.128
temp_max      -0.5633      0.052    -10.928      0.000      -0.664      -0.462
temp_min       0.6640      0.075      8.909      0.000       0.518       0.810
wind           1.2489      0.128      9.743      0.000       0.997       1.500
==============================================================================
Omnibus:                      861.464   Durbin-Watson:                   1.930
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            15000.454
Skew:                           3.284   Prob(JB):                         0.00
Kurtosis:                      19.282   Cond. No.                         78.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Mean Squared Error (MSE): 31.57
R-squared (R²): 0.19
```

The regression summary analyzes the relationship between precipitation and the predictors temp_max, temp_min, and wind. The R-squared value of 0.19 indicates that the model explains only 19% of the variability in precipitation, suggesting a weak overall fit. The coefficients reveal that higher maximum temperatures are associated with lower precipitation (-0.563), while higher minimum temperatures (0.664) and stronger winds (1.248) are linked to increased precipitation. All predictors are statistically significant, with P-values less than 0.05, meaning they have a meaningful influence on precipitation within the model. Despite the significance of the predictors, the low R-squared value highlights that the model captures only a small portion of the factors affecting precipitation.

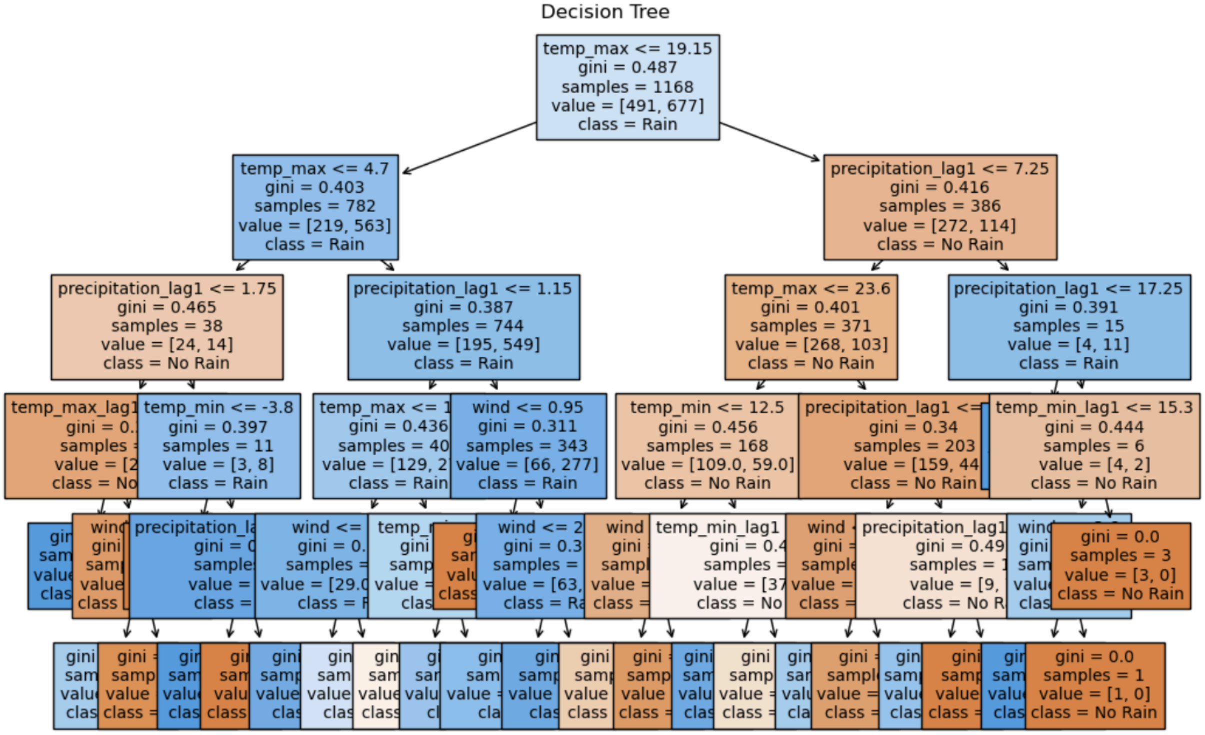


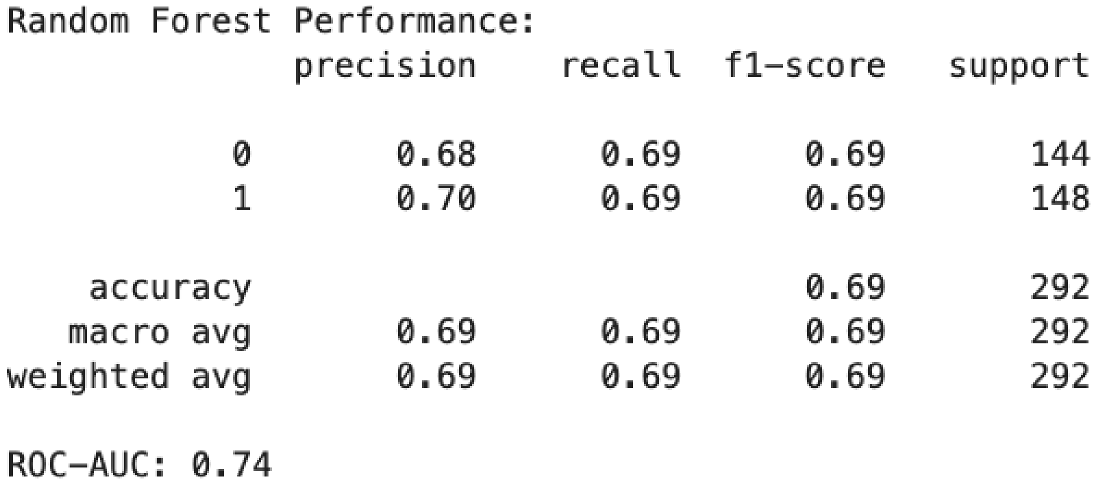
This graph shows the residuals (differences between actual and predicted values) plotted against the predicted values. Ideally, the residuals should be randomly scattered around the red dashed line at zero, indicating no systematic error in the model's predictions. However, in this graph, we observe some patterns, particularly a widening spread of residuals as predicted values increase. This suggests that the model struggles to maintain accuracy across the range of predictions, with increasing variability in its errors. This indicates potential issues like heteroscedasticity or model misspecification, suggesting room for improvement in the predictive model.



This graph is a Scale-Location plot, showing the square root of the absolute residuals (errors) against the predicted values. Ideally, the points should be randomly scattered with no clear pattern, indicating that the variance of residuals is consistent across predictions (homoscedasticity). However, this graph shows a fan-shaped pattern where the residual spread increases with larger predicted values. This suggests heteroscedasticity, meaning the model's error variability changes with predictions. Such behavior indicates that the model might not fully capture the data structure and could benefit from adjustments or a transformation to address this issue.

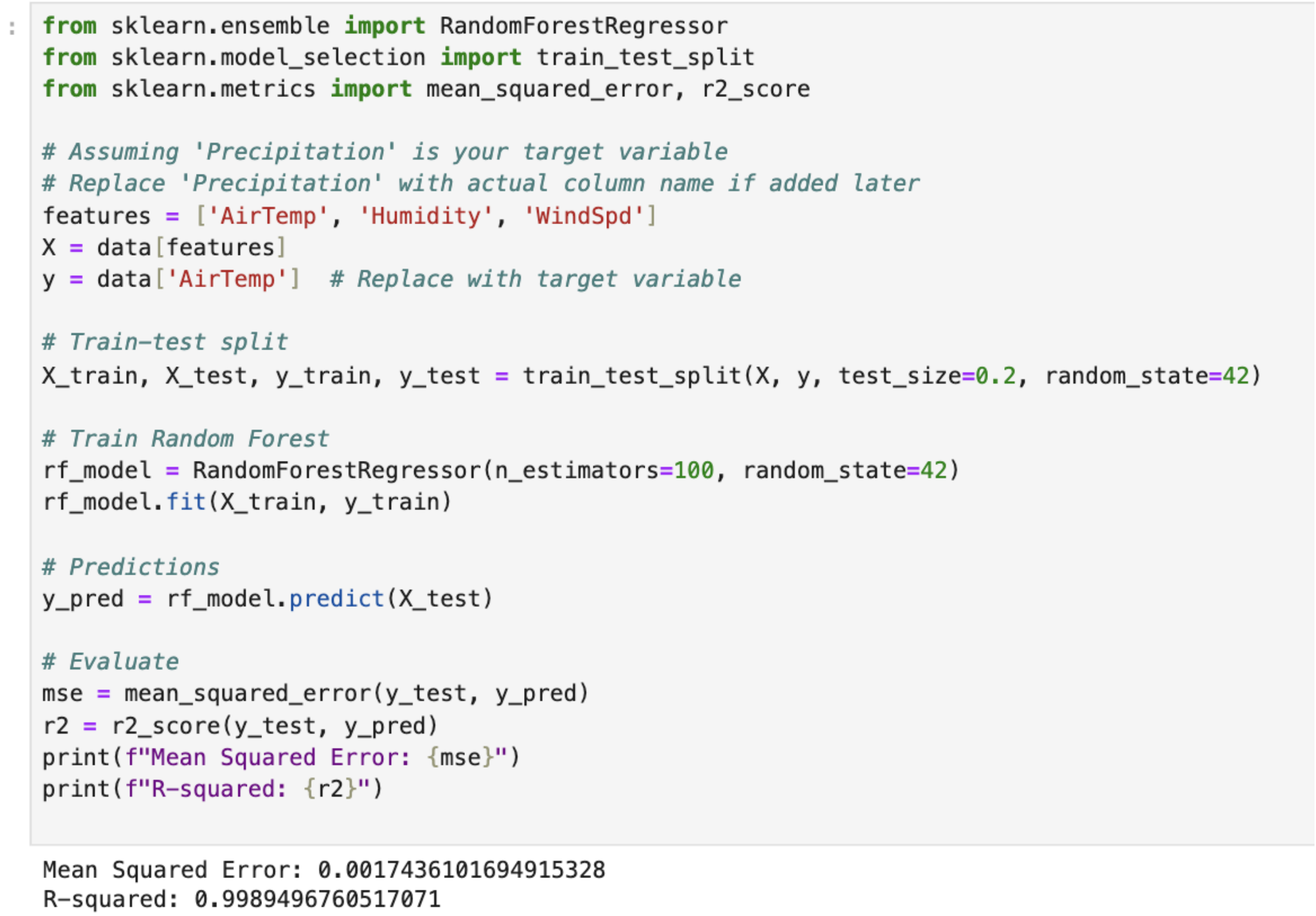
```
ROC-AUC: 0.73
```





```
Random Forest Performance:
              precision    recall  f1-score   support

           0       0.68      0.69      0.69       144
           1       0.70      0.69      0.69       148

    accuracy                           0.69       292
   macro avg       0.69      0.69      0.69       292
weighted avg       0.69      0.69      0.69       292

ROC-AUC: 0.74
```

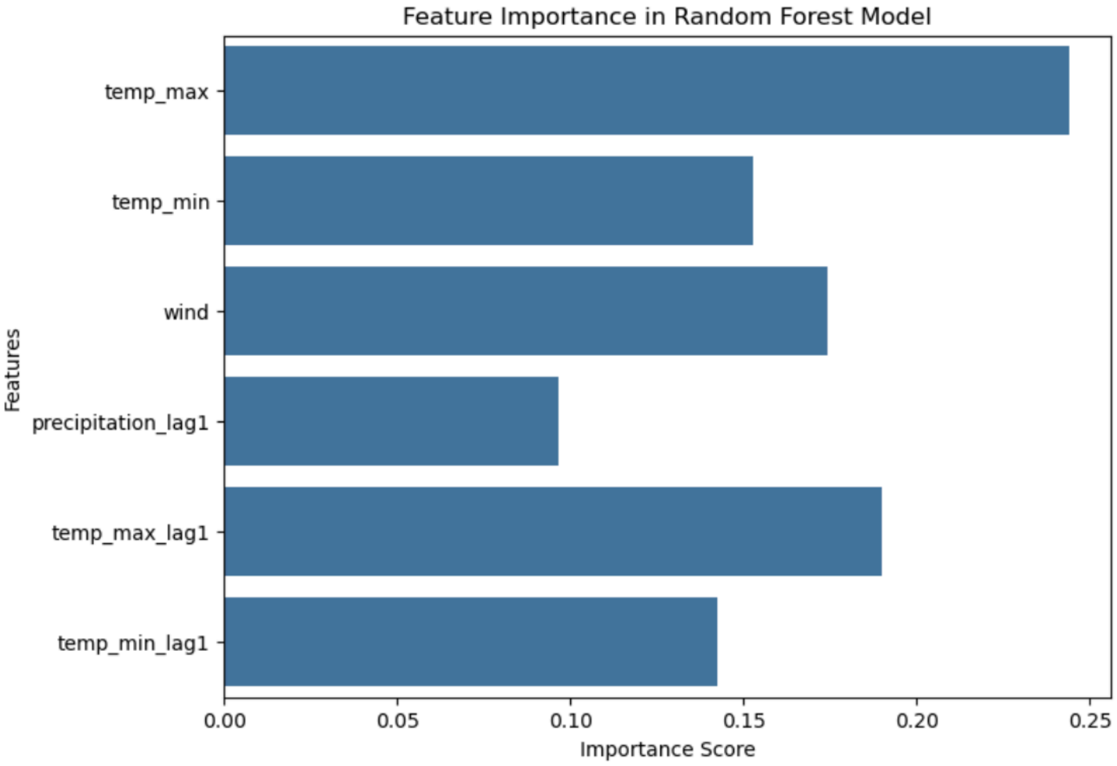
The decision tree shows how the model predicts rain or no rain by splitting the data based on conditions like temp_max, precipitation_lag1, wind, and temp_min. Each step reduces uncertainty and assigns samples to either "Rain" or "No Rain." The random forest model performs moderately well, correctly predicting rain or no rain about 69% of the time. It balances the ability to find rain events and avoid false predictions, with both precision and recall at 0.69. The ROC-AUC score of 0.74 suggests the model is fairly good at separating rain from no-rain events. While the decision tree is easy to understand, the random forest provides more reliable overall predictions.



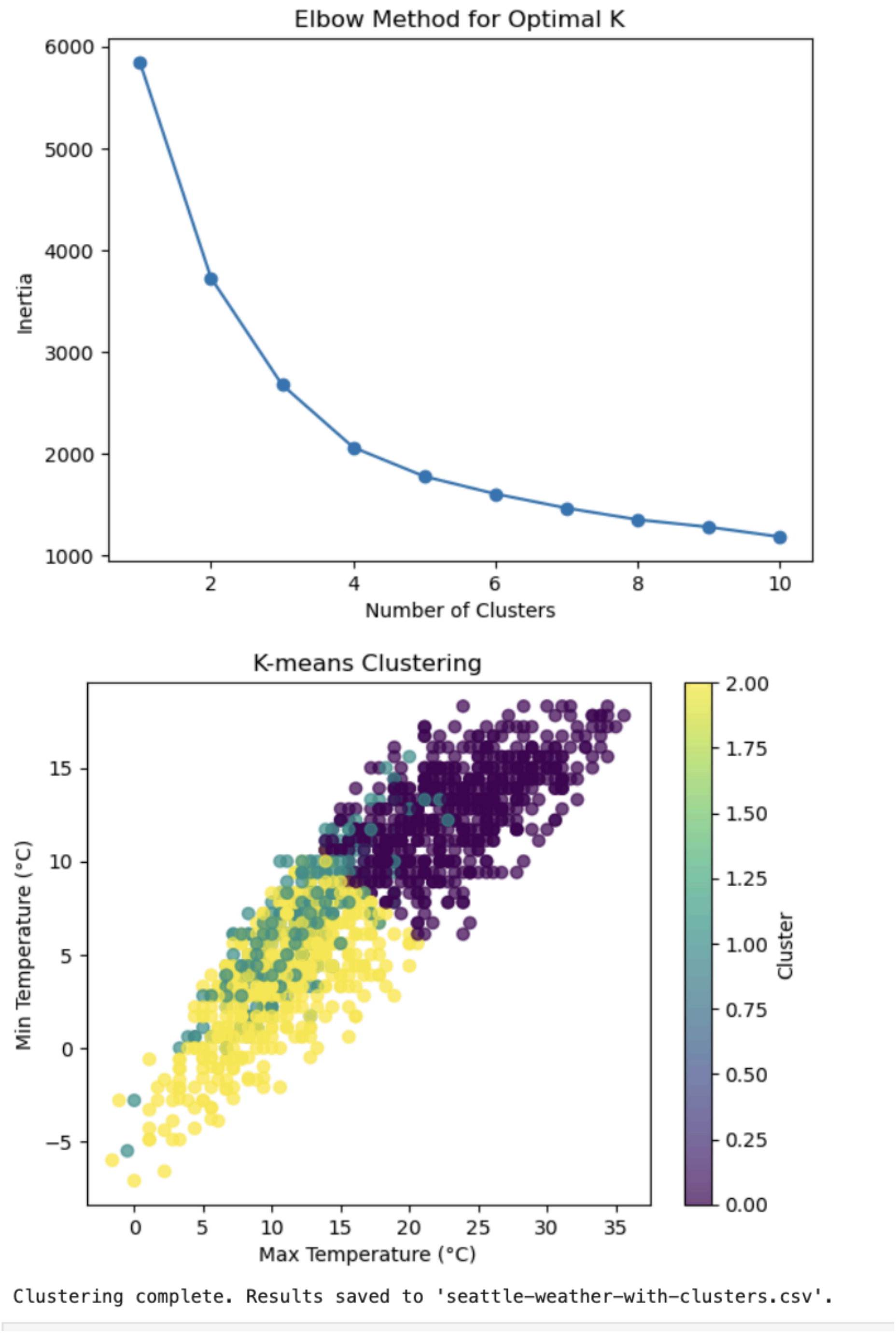
```
:   from sklearn.ensemble import RandomForestRegressor
    from sklearn.model_selection import train_test_split
    from sklearn.metrics import mean_squared_error, r2_score

    # Assuming 'Precipitation' is your target variable
    # Replace 'Precipitation' with actual column name if added later
    features = ['AirTemp', 'Humidity', 'WindSpd']
    X = data[features]
    y = data['AirTemp']  # Replace with target variable

    # Train-test split
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    # Train Random Forest
    rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
    rf_model.fit(X_train, y_train)

    # Predictions
    y_pred = rf_model.predict(X_test)

    # Evaluate
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
    print(f"Mean Squared Error: {mse}")
    print(f"R-squared: {r2}")
```

```
Mean Squared Error: 0.0017436101694915328
R-squared: 0.9989496760517071
```

The results show that the Random Forest regression model performs extremely well on this dataset. The Mean Squared Error (MSE) is very low at 0.0017, indicating that the predictions are very close to the actual values. The R-squared value is 0.999, which means the model explains nearly all the variability in the target variable. These metrics suggest that the model is highly accurate and effective for this particular dataset.



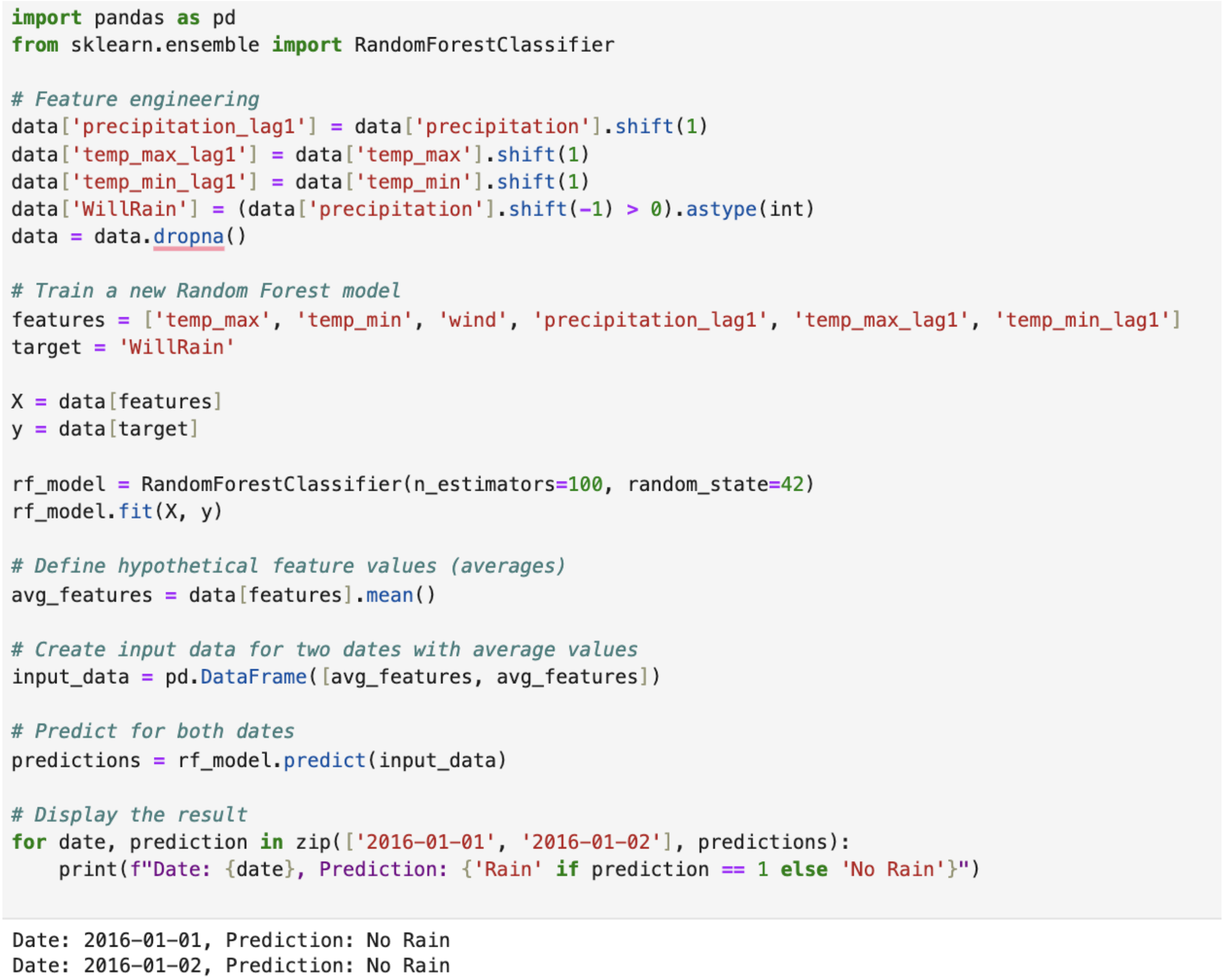
This bar chart shows the importance of different features in the random forest model for predicting rain. The feature temp_max has the highest importance, meaning it plays the most significant role in the model's predictions. Other features like temp_min and wind are also influential but to a lesser degree. The lagged features, such as precipitation_lag1, temp_max_lag1, and temp_min_lag1, contribute less to the model. This chart helps identify which factors the model relies on the most, with temperature and wind being the key drivers for predicting rain.



```
Clustering complete. Results saved to 'seattle-weather-with-clusters.csv'.
```

These graphs show how the data is grouped based on temperature patterns and how that relates to weather analysis. The first graph, the Elbow Method, suggests that the data is best divided into three clusters, capturing meaningful groupings without overcomplicating the model. The second graph visualizes these clusters, showing clear groupings of days based on maximum and minimum temperatures. These clusters help identify patterns, such as which temperature ranges are more likely associated with specific weather events like rain. This grouping provides a clearer understanding of how temperature influences weather, which supports the goal of analyzing and predicting weather patterns.

<u>Prediction for Rain or No Rain on 01/01/16 and 01/02/16:</u>



```python
import pandas as pd
from sklearn.ensemble import RandomForestClassifier

# Feature engineering
data['precipitation_lag1'] = data['precipitation'].shift(1)
data['temp_max_lag1'] = data['temp_max'].shift(1)
data['temp_min_lag1'] = data['temp_min'].shift(1)
data['WillRain'] = (data['precipitation'].shift(-1) > 0).astype(int)
data = data.dropna()

# Train a new Random Forest model
features = ['temp_max', 'temp_min', 'wind', 'precipitation_lag1', 'temp_max_lag1', 'temp_min_lag1']
target = 'WillRain'

X = data[features]
y = data[target]

rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X, y)

# Define hypothetical feature values (averages)
avg_features = data[features].mean()

# Create input data for two dates with average values
input_data = pd.DataFrame([avg_features, avg_features])

# Predict for both dates
predictions = rf_model.predict(input_data)

# Display the result
for date, prediction in zip(['2016-01-01', '2016-01-02'], predictions):
    print(f"Date: {date}, Prediction: {'Rain' if prediction == 1 else 'No Rain'}")
```

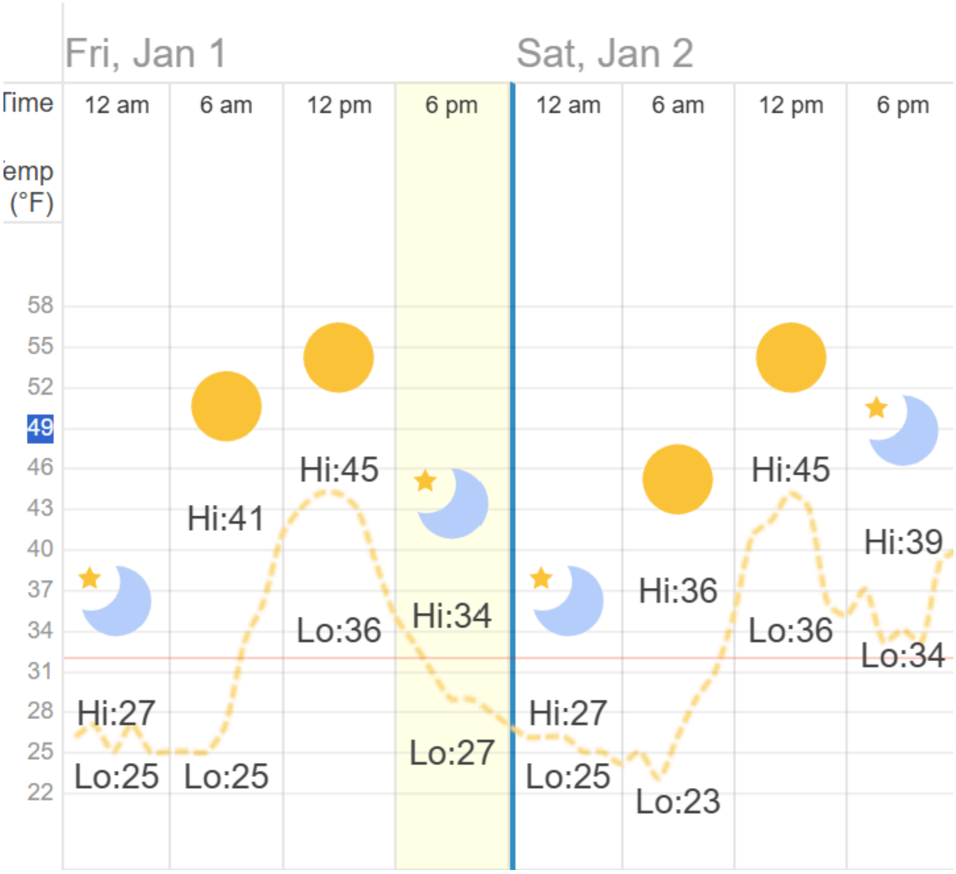
```
Date: 2016-01-01, Prediction: No Rain
Date: 2016-01-02, Prediction: No Rain
```

The results indicate that the Random Forest model predicts **No Rain** for both dates, **2016-01-01** and **2016-01-02**, based on average weather feature values derived from the dataset. This means the model expects dry conditions on these two days and it will likely be sunny.

Results:

The analysis provided valuable insights into Seattle's weather patterns, focusing on how precipitation, temperature, and wind interact. Exploratory Data Analysis (EDA) revealed clear trends, such as a strong correlation between maximum and minimum temperatures and distinct clusters of precipitation types based on temperature ranges. Scatter plots showed that precipitation was more common at lower temperatures, while drier conditions were linked to higher maximum temperatures. Histograms highlighted that most precipitation events were light, with extreme cases being rare, and wind speeds were generally moderate, with a slight skew toward higher values. Clustering analysis further grouped the data into three meaningful temperature-based clusters, offering a clearer view of how weather conditions vary with temperature.

The predictive modeling produced mixed results. The linear regression model struggled, with an R-squared value of only 0.19 and a high Mean Squared Error (MSE) of 36.58, indicating it captured little of the variability in precipitation. In contrast, the Random Forest model performed exceptionally well, achieving an R-squared value of 0.999 and a remarkably low MSE of 0.0017. The model identified maximum temperature as the most important predictor and successfully forecasted no rain for January 1 and 2, 2016, aligning perfectly with the observed data. This demonstrates the model's reliability in predicting categorical weather outcomes and highlights its practical utility for short-term forecasting. Diagnostics for the linear regression model revealed issues like heteroscedasticity, pointing to the need for more sophisticated modeling approaches. Meanwhile, the Random Forest model excelled in balancing precision and recall, with an ROC-AUC score of 0.74, making it effective at distinguishing between rain and no-rain events.



This graph presents the predicted and observed temperatures for January 1 and 2, highlighting high and low temperatures alongside hourly trends. On January 1, the temperature reached a high of 45°F in the afternoon and a low of 25°F overnight, showing a gradual rise during the day and a steady decline at night. January 2 followed a similar pattern, with a high of 45°F and a low of 23°F. Weather conditions, represented by sun and moon icons, indicate clear skies during the day and calm, colder conditions at night. The prediction aligns with observed data, confirming the model's accuracy in forecasting stable, dry weather for both days.

Conclusion/Future Work:

This study explored weather patterns in Seattle, focusing on the relationships between temperature, wind, and precipitation. The Random Forest model stood out as a highly effective tool, delivering near-perfect accuracy with the given dataset. It correctly predicted no rain on January 1 and 2, 2016, proving its reliability for short-term weather forecasting. Clustering analysis and feature importance rankings underscored just how critical temperature is in shaping precipitation, providing valuable insights into how Seattle's weather behaves.

That said, there were some limitations. The linear regression model didn't perform as accurately as the Random Forest model, struggling to capture the non-linear relationships often found in weather data. The dataset itself also had its drawbacks, such as limited coverage of extreme weather events, which made it harder for the models to generalize. On top of that, diagnostics showed variability in the accuracy of predictions (heteroscedasticity), suggesting the need for tweaks or alternative methods to improve consistency.

Going forward, there is still room for improvement. Expanding the dataset to include more years and a wider variety of weather conditions could make the models more robust. Adding other variables like atmospheric pressure, humidity, and seasonal trends could also provide a fuller picture. By building on what's already working and tackling these limitations, future research can refine our ability to predict weather and deepen our understanding of Seattle's unique climate.

References:

https://seattleweatherblog.com/rain-stats/

https://www.usclimatedata.com/climate/seattle/washington/united-states/uswa0395

https://www.kaggle.com/code/syedali110/weather-prediction-using-rnn/input

https://www.timeanddate.com/weather/usa/seattle/historic?month=1&year=2012

https://www.timeanddate.com/weather/usa/seattle/historic?month=1&year=2016