BM20A6100 Advanced Data Analysis and Machine Learning (Part 2)

Project Work 2 — Predicting Rainfall in Sydney, Australia

Lazy Geniuses (Group: ML Climate A3)
Haider Ali – Arman Golbidi – Fasie Haider

Course period: 1 Sep 2025 – 12 Dec 2025

This report is written in LaTeX

Week ML2 Submission

Contents

1 Introduction

Although forecasting rainfall is a difficult task, it is essential for water planning, agriculture, and public safety. We start this study by examining a historical meteorological record from Sydney, Australia. In order to predict rainfall, our goal is to understand how rainfall behaves over time. In this report, we will give a summary of the work we accomplished in the first phase of the study.

During this stage, we concentrated on comprehending the data and how rainfall varies from month to month and between years rather than developing a predictive model just yet. In order to accomplish that, we focus on four primary analyses that were assigned in the second week's tasks:

- Exploratory visualization of the time series data.
- Decomposition of the series into its constituent components.
- Autocorrelation analysis to identify temporal dependencies.
- The design of a robust data partitioning plan for future model development.

This foundational work is essential for guiding our subsequent feature engineering and model selection processes.

Repository location. All Week 2 materials (code, figures, and this PDF) are hosted publicly at:
`tembooo/ADML_Part2/2.Project Group Work/1.Week2 - Data visualization and project plan`.

2 Data and Methodology

We used the "Rain in Australia" dataset from Kaggle for this research, although we only included recordings from Sydney. After we cleaned it up, the Sydney section had a total of 3,344 daily observations from February 2008 to June 2017.

Since that's the primary item we're attempting to comprehend, we focused primarily on the **Rainfall** column this week (the amount of rain in millimeters). Using the standard Python tools **pandas** for data handling, **Matplotlib** for plots, and a small amount of **Statsmodels** for time-series work, all of the analysis was completed.

3 Exploratory Data Analysis & Visualization

To better comprehend the data, we first showed the daily rainfall across the nine years. As was to be predicted, the daily plot was extremely unpredictable, with lengthy stretches of no rain punctuated with sudden rises. This illustrated the extensive and unpredictable rainfall in Sydney. To lessen some of this noise, we additionally merged the data by month and year. These aggregated perspectives revealed broader seasonal patterns that are not visible in daily data.

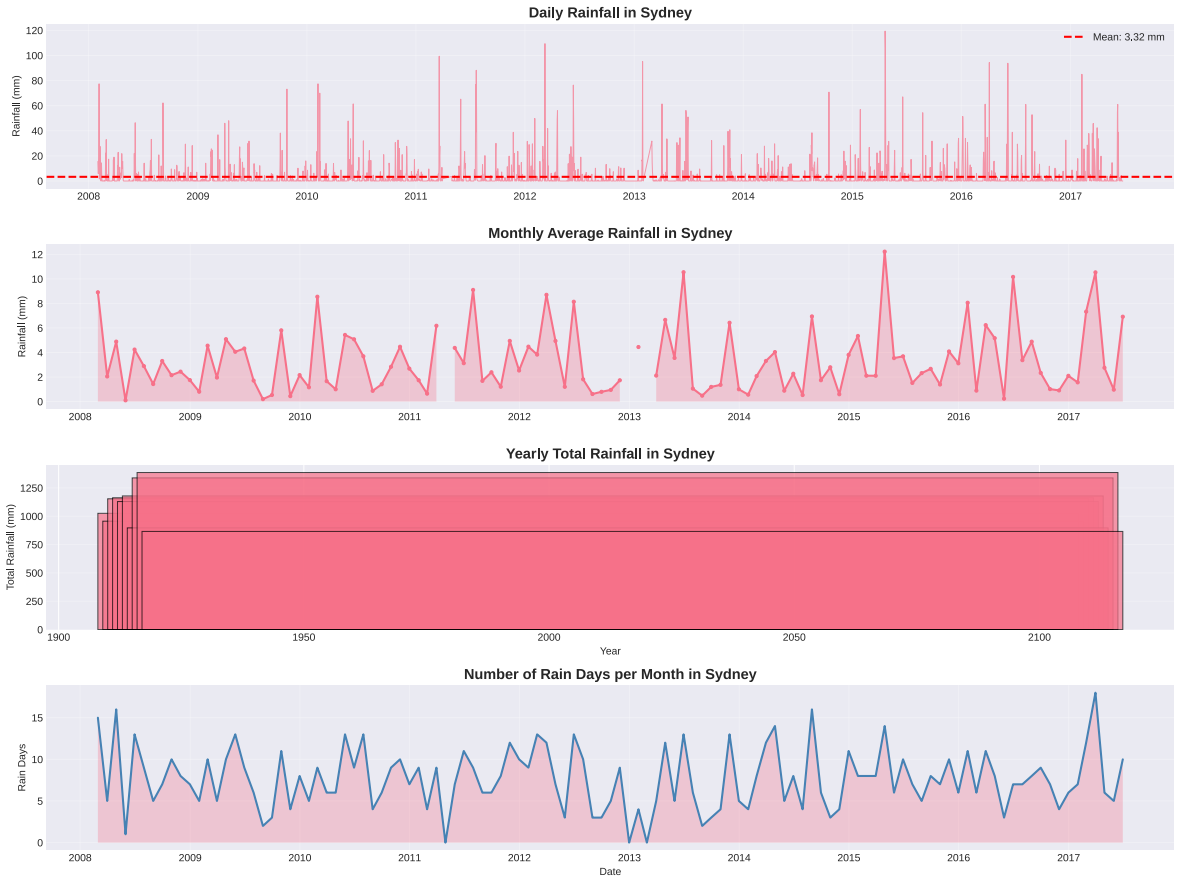

Figure 1: Time Series Overview of Rainfall in Sydney.

Figure 1 analysis. The daily series is highly intermittent and right-skewed with long dry spells and sudden bursts, which motivates treating rain occurrence (binary) separately from rain amount (continuous). Aggregation (monthly/yearly) smooths volatility and reveals seasonal structure used in later sections.

Next, we looked at some descriptive statistics. The average rainfall was 3.32 mm. Based on this disparity, the distribution is obviously skewed to the right. To put it briefly, most days don't have any rain, but when they do, it varies greatly. The standard deviation, which was significantly more than the mean at 9.88 mm, demonstrated the daily variability of rainfall. This was further supported by the seasonal bar chart (Figure 2) and monthly box plots. Heavy rainfall can occur at any moment because outliers were found in nearly every season. One thing was evident from the bar chart, though: late winter and early spring are often the driest months, whereas autumn (March–May) is typically the wettest.

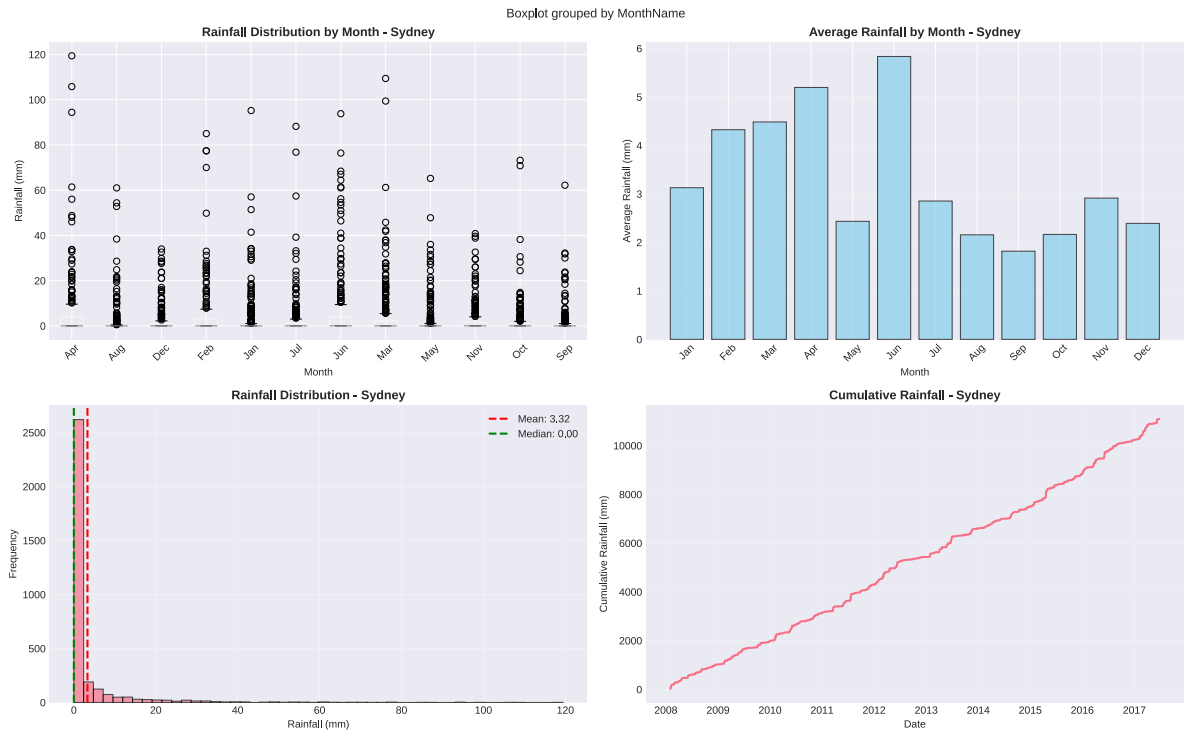

Figure 2: Seasonal and Distributional Analysis of Sydney Rainfall.

Figure 2 analysis. Seasonal bars and monthly box plots indicate relatively wetter conditions in autumn (Mar–May) and drier conditions from late winter to early spring. Wide IQRs and many outliers in most months confirm volatility and the potential for heavy rain events throughout the year.

Our modeling strategy was one of the most important lessons learned from this investigation. It will be challenging to estimate precise rainfall amounts (regression) because the data is highly skewed and comprises a large number of zero-rain days. Predicting whether it will rain at all is more sensible than trying to predict how much.

4 Time Series Decomposition

In this part, we applied an **additive** decomposition model to the monthly averages to separate the long term trend, the seasonal pattern, and the unexplained residual noise.

Figure 3: Additive Time Series Decomposition of Monthly Rainfall in Sydney.

Figure 3 analysis. The trend meanders with multi-year undulations (peaks near 2011 and 2016; dip around 2014), but shows no sustained increase or decrease; the measured strength is weak (≈ 0.0575). Residuals remain large and irregular, indicating substantial variability unexplained by calendar effects.

There was no clear rising or downward pattern in the trend component. Rather, it moved in long waves, with a decrease around 2014 and peaks around 2011 and 2016, respectively. Larger climate factors are probably reflected in these multi-year cycles. In contrast to the amount of the residuals, the measured "strength of trend" was only 0.0575, which is extremely weak. Therefore, even while the curve fluctuates, there isn't a significant long-term tendency in general.

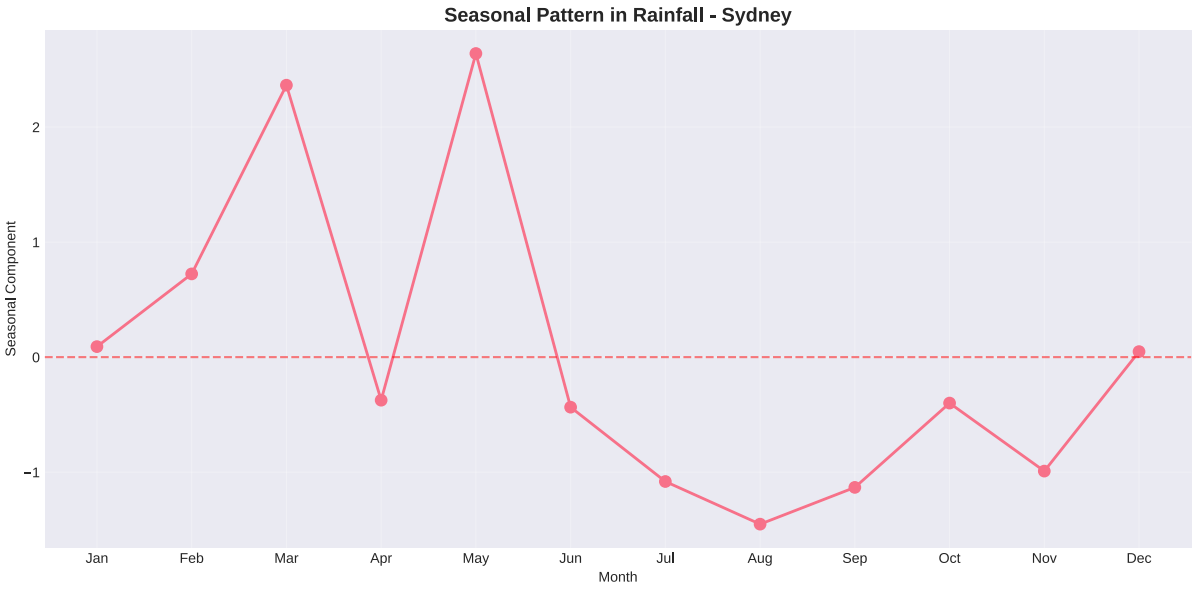

Figure 4: Isolated Seasonal Component of Rainfall.

Figure 4 analysis. The seasonal component is stable across years: relatively higher rainfall in late summer to autumn and lower rainfall in late winter/early spring. This supports including simple calendar features (month, day-of-year) as weak predictors, to be complemented by meteorological covariates.

In this figure the residuals were the most interesting piece. They were large and highly variable, meaning a lot of rainfall behavior cannot be explained by trend or seasonality alone. This makes intuitive sense, storms, sudden downpours, and unpredictable weather events play a huge role in rainfall. This result tells us something important for the next steps: simply relying on date-based features will not be enough. We will need to bring in additional meteorological variables to improve our ability to predict rainfall.

5 Autocorrelation Analysis

We performed an autocorrelation analysis to measure the "memory" of the rainfall series, that is, how a day's rainfall correlates with that of previous days. We plotted the Autocorrelation Function and Partial Autocorrelation Function for daily, weekly, and monthly data, as you can see in Figure 5 below.

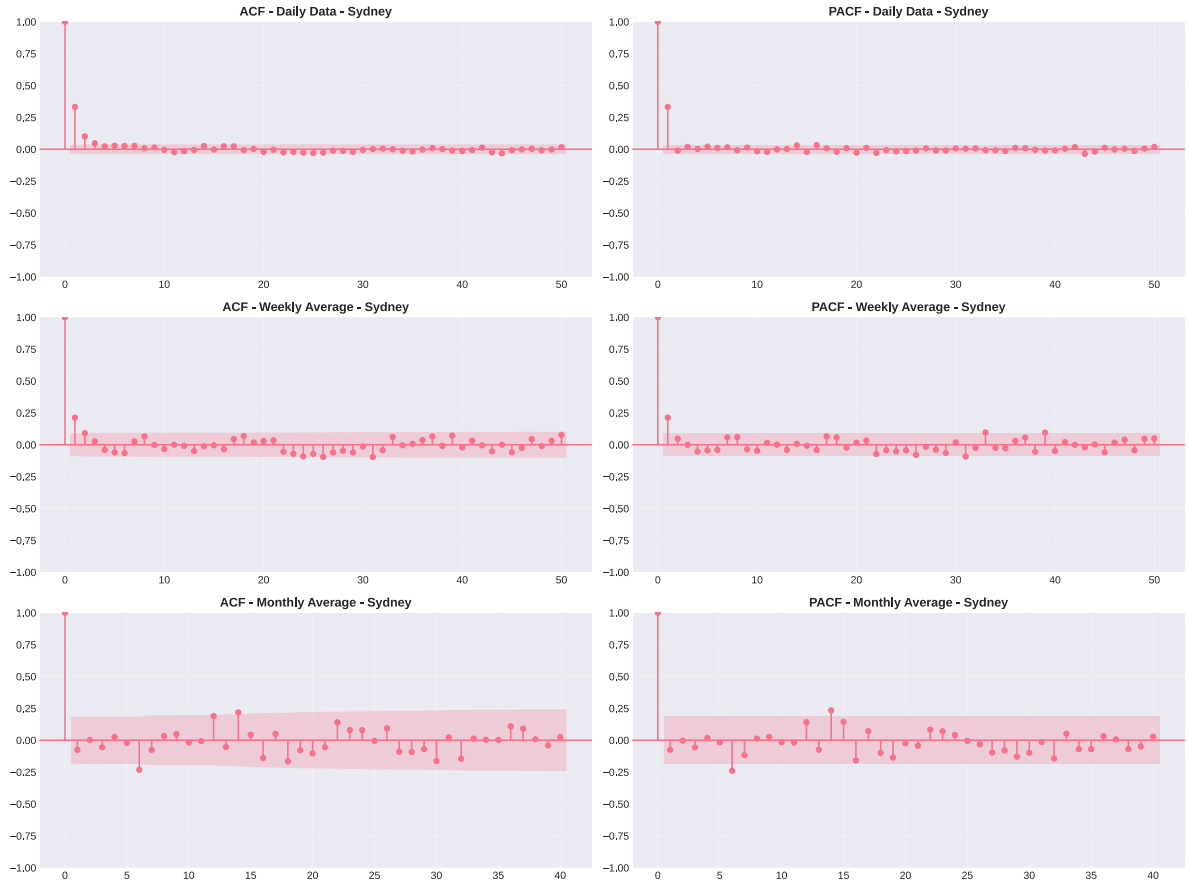

Figure 5: Autocorrelation and Partial Autocorrelation plots.

Figure 5 analysis. The daily ACF shows a clear spike at lag 1 (≈ 0.332) with rapid decay; the PACF confirms a dominant lag-1 effect, consistent with an AR(1)-like process. This motivates including at least $\texttt{Rainfall}_{t-1}$ in baseline models.

For the daily series, the ACF showed a clear spike at lag 1 (value: 0.332), which means rainfall on a given day has a moderate relationship with rainfall the day before. This effect disappears quickly after two days, and the PACF confirmed a strong lag-1 influence only. This is typical of an AR(1) process.

Lag plots (Figure 6) visually confirmed this: only the lag-1 plot showed a clear positive pattern.

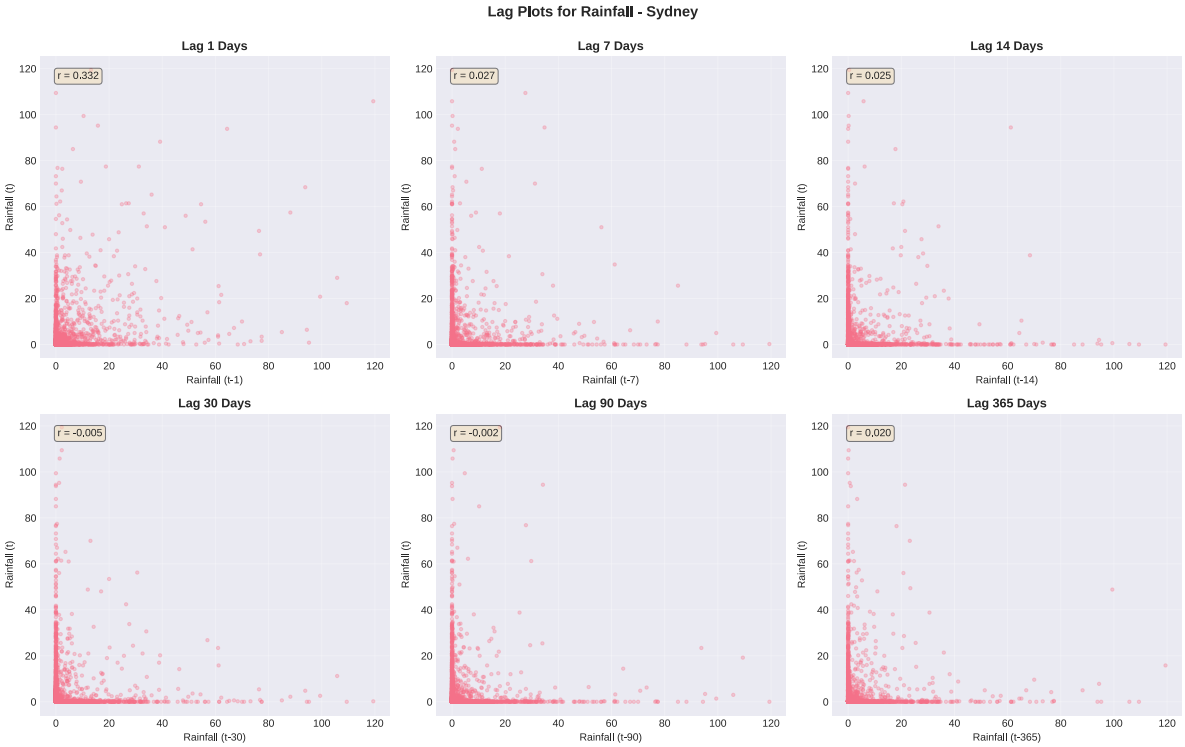

Figure 6: Lag Plots for Daily Rainfall.

Figure 6 analysis. The lag-1 panel shows a visible positive association; higher-lag panels are diffuse, reinforcing the short memory of the series. This directly shapes our feature engineering: include at least the previous day's rainfall as an input feature.

This finding will directly shape our feature engineering. At minimum, we should include Rainfall from the previous day as one of the input features for the model.

6 Data Partitioning Plan for Model Development

Since this is time-series data, the dataset must be split chronologically to avoid leakage. We created a three-way split:

- **Training Set (70%)**: 2008-02-01 to 2014-09-25
- **Validation Set (15%)**: 2014-09-26 to 2016-02-09
- **Test Set (15%)**: 2016-02-10 to 2017-06-25

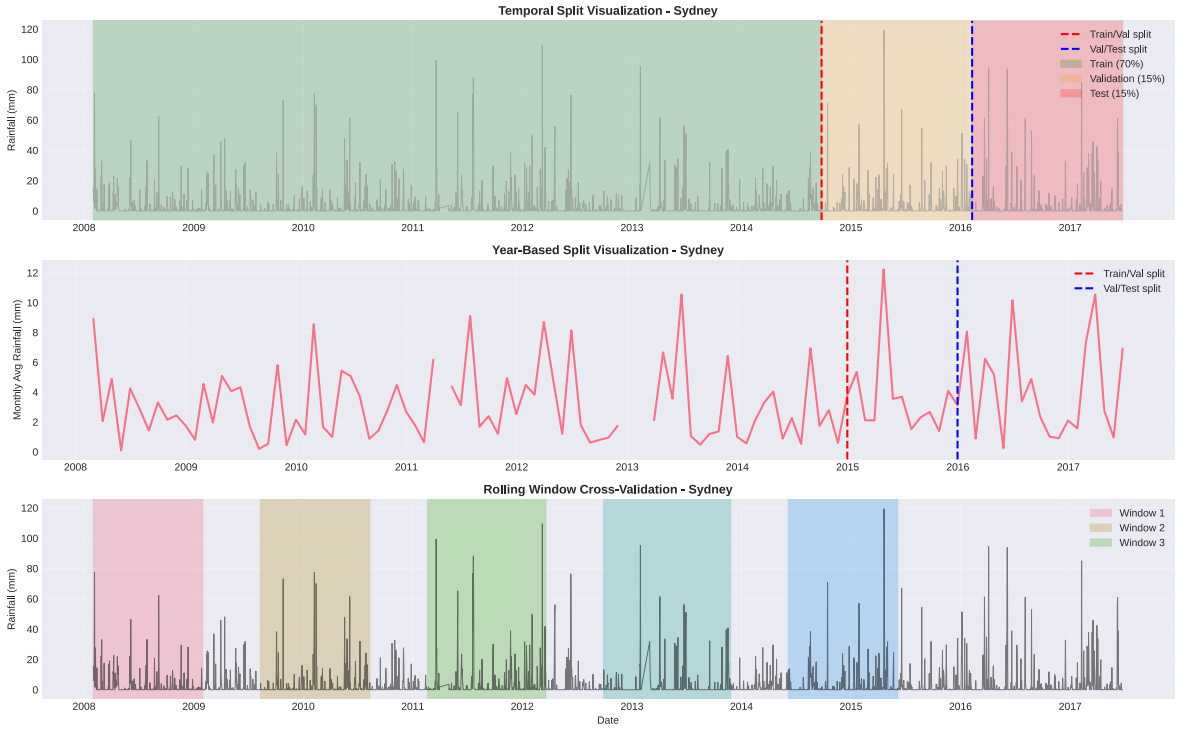

Figure 7: Visualization of the Data Partitioning Strategy.

Figure 7 analysis. The training set spans multiple full seasonal cycles, the validation window supports tuning and model selection, and the hold-out test window is reserved for the final unbiased evaluation.

The training set covers enough years to capture seasonal patterns. The validation set will help tune models, and the test set remains untouched for a final unbiased evaluation.

7 Conclusion and Next Steps

From this first phase of the project, we have learned several important things which is given below:

- Sydney rainfall is extremely skewed and unpredictable, so a binary classification approach is more practical at the start.
- The data exhibits weak to moderate seasonality and no significant long term trend.
- There is a strong short term temporal dependence, with the previous day's rainfall being a significant predictor.

Based on the work which we have done, our next tasks will be:

- **Feature Engineering:** Creating lagged rainfall variables and temporal features (e.g., month, day of year).
- **Data Preprocessing:** Handling missing values in other predictive columns and scaling numerical features.
- **Baseline Modeling:** Developing a simple baseline model.

References

- Kaggle: *Rain in Australia* (`https://www.kaggle.com/datasets/jsphyg/weather-dataset`
- Statsmodels: Time-series decomposition and ACF/PACF documentation.