IND 522

ADVANCED STATISTICAL MODELLING

FINAL EXAM – FALL 2025

SUMMARY

Dataset Overview

This study focuses on the "Rain in Australia" (weatherAUS) dataset. This dataset includes a collection of daily weather observations from 49 meteorological stations across Australia, spanning approximately 10 years from 2008 to 2017. It contains over 145,460 observations and 23 variables. There are three types of variables in the dataset: numerical, categorical and binary.

- **Numerical Variables** (16): These distinct continuous measurements capture the physical state of the atmosphere.
    i. Temperature: *MinTemp, MaxTemp, Temp9am, Temp3pm.*
    ii. Moisture: *Rainfall, Evaporation, Humidity9am, Humidity3pm.*
    iii. Wind: *WindGustSpeed, WindSpeed9am, WindSpeed3pm*.
    iv. Pressure*: Pressure9am, Pressure3pm.*
    v. Cloud*: Cloud9am, Cloud3pm, Sunshine.*
- **Categorical Variables** (5): These represent qualitative attributes.
    i. Spatial: *Location* (49 unique stations).
    ii. Directional: *WindGustDir, WindDir9am, WindDir3pm* (Compass directions, e.g., 'W', 'NW').
    iii. Temporal: *Date* (Used to derive seasonality, though excluded from direct modeling).
- **Binary Variables** (2): These are the critical "Yes/No" flags.
    i. *RainToday*: Indicates if precipitation exceeded 1mm in the last 24 hours.
    ii. *RainTomorrow* (Target): The prediction target for the classification task.

In classification, we used *RainTomorrow* as the target variable to predict whether it will rain the following day based on current weather patterns. In regression, we used *MaxTemp* as the target variable to model and predict the exact maximum temperature of the day using morning indicators and atmospheric conditions. The primary objective of this study is to apply advanced statistical modelling techniques -ranging from exploratory analysis and dimensionality reduction to predictive modelling and simulation- to comprehend Australian climate patterns and accurately forecast rainfall and extreme temperature events.

Key Findings

This study of statistical analysis showed the following insights about the given dataset:

- **Climatological Patterns & Data Quality**:
    a. Imbalance: The dataset is moderately imbalanced, with "*Rain*" occurring in only 22% of observations. Thus, we need to handle carefully during model training to ensure rare rain events are not overlooked.
    b. The Thermal Normal: The maximum temperature (*MaxTemp*) follows a almost perfect Normal (Gaussian) distribution ($\mu \approx 23.2℃, \sigma \approx 7.1$) , this finding is validated by Kolmogorov-Smirnov tests and Monte Carlo simulations in the study. This suggests temperature fluctuations are stable and predictable processes.
    c. Dimensionality: We reduced the number of 16 numerical variables into 3 Principal Components, that explain 63.5% of the total variance. These components represent "Thermal Conditions", "Atmospheric Pressure" and "Humidity", proving that the weather system in this dataset can be simplified without significant information loss.
- **Classification & Regression:**
    a. Forecasting Rain**:** The **Random Forest Classifier** in the study achieved an AUC of 0.87 and an accuracy of ~85%, while outperforming the **Logistic Regression**. The model identified *Humidity at 3PM, Sunshine duration* and *Atmospheric Pressure* as the strongest precursors to rain.
    b. Predicting Temperature: We developed a robust **Linear Regression** model that predicts daily *Maximum Temperature* with 92.6% accuracy (R^2). The model highlights the strong predictive power of morning temperature (*Temp9am*) and the cooling effect of afternoon humidity.
- **Risk Assessment (Monte Carlo):**
    a. Using Monte Carlo simulation (N=1000), we estimated the probability of an "Extreme Heat Event" ($> 35℃$) to be approximately 5.57%. The narrow 95% confidence interval ($\pm 0.12\%$) indicates a high degree of precision in this estimate, providing a reliable metric for heatwave risk assessment.

Conclusion

This analysis shows that Australian weather patterns can be simplified and identified by statistical tools. We observed that standard meteorological variables can accurately predict both binary outcomes (Rain/No Rain) and continuous metrics (Temperature) with high confidence. These models offer actionable value for sectors such as agriculture, water resource management, and disaster preparedness.

1. EXPLORATORY DATA ANALYSIS

The dataset contains 145,460 observations and 23 features. Upon inspection, several key observations were made regarding the data quality:

- **Missing Data**: Significant missingness was observed in variables such as *Sunshine* (48%), *Evaporation* (43%), *Cloud9am* (38%), and *Cloud3pm* (40%). These features are meteorologically important but will require careful imputation or may be dropped if they introduce too much noise.
- **Target Variable:** The target variable for classification, *RainTomorrow*, has about 3,267 missing entries. For a robust analysis, we will remove rows where the target is unknown.
- The *Date* column needs to be converted to a datetime format to extract seasonal features (Month, Year), which are important for weather prediction.

1.1. Univariate Anaysis

We quantified the distribution of numerical and categorical variables to understand the baseline climate profile.

- **Numerical Variables:**
    a. **Temperature:** *MaxTemp* and *MinTemp* follow roughly normal distributions.
    b. **Rainfall:** This feature is heavily right-skewed (Mean: 2.36mm, Median: 0.0mm). Most days are dry, have 0 rainfall but there are extreme values (up to 371mm), indicating storm events.
    c. **Wind:** *WindGustSpeed* is also right-skewed, suggesting frequent moderate winds interspersed with rare storm gusts.
- **Categorical Variables:**
    a. The distribution of *RainTomorrow* shows a clear class imbalance: 78% of days are "No Rain" vs. 22% "Rain". A naive model predicting "No" would achieve 78% accuracy. Therefore, evaluation must focus on metrics like F1-Score or AUC rather than simple accuracy.
    b. *WindGustDir*: The wind direction is relatively well-distributed, though westerly winds ('W') appear slightly more frequent.

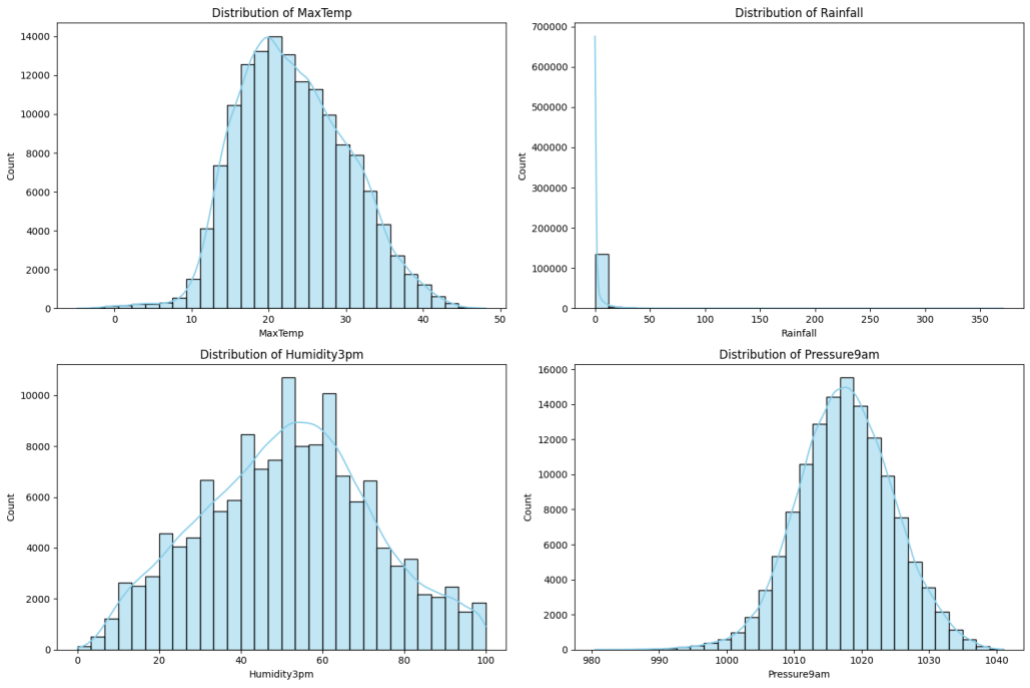
Descriptive statistics for numerical and categorical variables are given below tables:

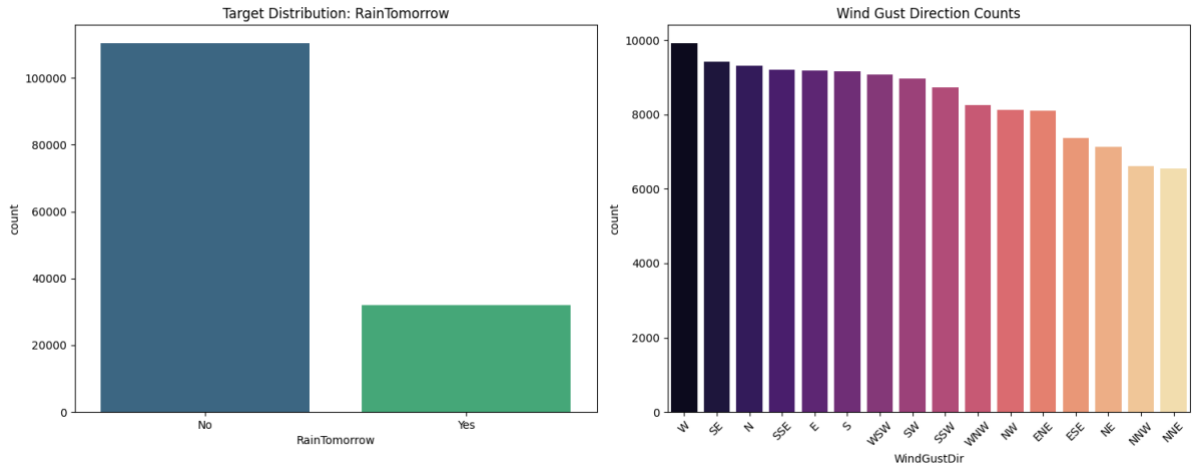
	Descriptive Statistics (Numerical)							
	MinTemp	*MaxTemp*	*Rainfall*	*Evaporation*	*Sunshine*	*WindGustSpeed*	*WindSpeed9am*	*WindSpeed3pm*
count	143975	144199	142199	82670	75625	135197	143693	142398
mean	12.194034	23.221348	2.360918	5.468232	7.611178	40.03523	14.043426	18.662657
std	6.398495	7.119049	8.47806	4.193704	3.785483	13.607062	8.915375	8.8098
min	-8.5	-4.8	0	0	0	6	0	0
25%	7.6	17.9	0	2.6	4.8	31	7	13
50%	12	22.6	0	4.8	8.4	39	13	19
75%	16.9	28.2	0.8	7.4	10.6	48	19	24
max	33.9	48.1	371	145	14.5	135	130	87

	Descriptive Statistics (Numerical)							
	Humidity9am	*Humidity3pm*	*Pressure9am*	*Pressure3pm*	*Cloud9am*	*Cloud3pm*	*Temp9am*	*Temp3pm*
count	142806	140953	130395	130432	89572	86102	143693	141851
mean	68.880831	51.539116	1017.64994	1015.255889	4.447461	4.50993	16.990631	21.68339
std	19.029164	20.795902	7.10653	7.037414	2.887159	2.720357	6.488753	6.93665
min	0	0	980.5	977.1	0	0	-7.2	-5.4
25%	57	37	1012.9	1010.4	1	2	12.3	16.6
50%	70	52	1017.6	1015.2	5	5	16.7	21.1
75%	83	66	1022.4	1020	7	7	21.6	26.4
max	100	100	1041	1039.6	9	9	40.2	46.7

	Descriptive Statistics (Categorical)						
	Date	*Location*	*WindGustDir*	*WindDir9am*	*WindDir3pm*	*RainToday*	RainTomorrow
count	145460	145460	135134	134894	141232	142199	142193
unique	3436	49	16	16	16	2	2
top	12.11.2013	Canberra	W	N	SE	No	No
freq	49	3436	9915	11758	10838	110319	110316

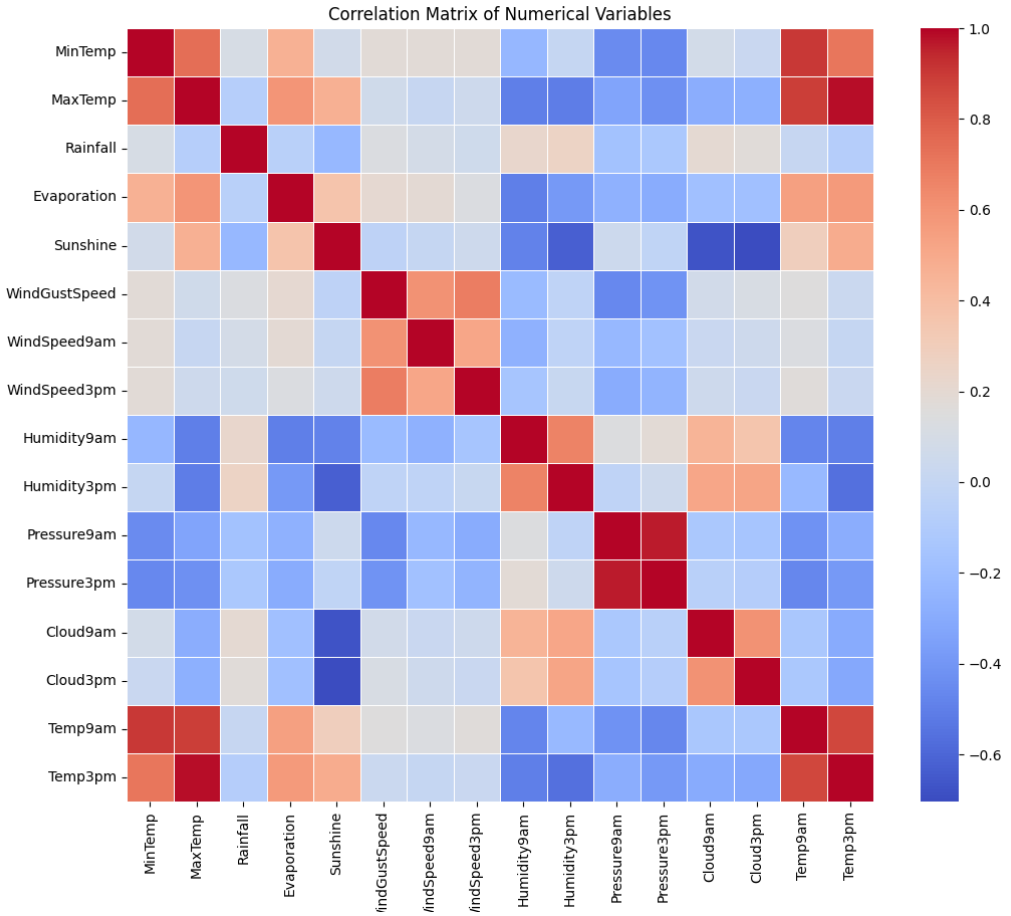
Distribution visualizations of numerical and categorical variables are given below:





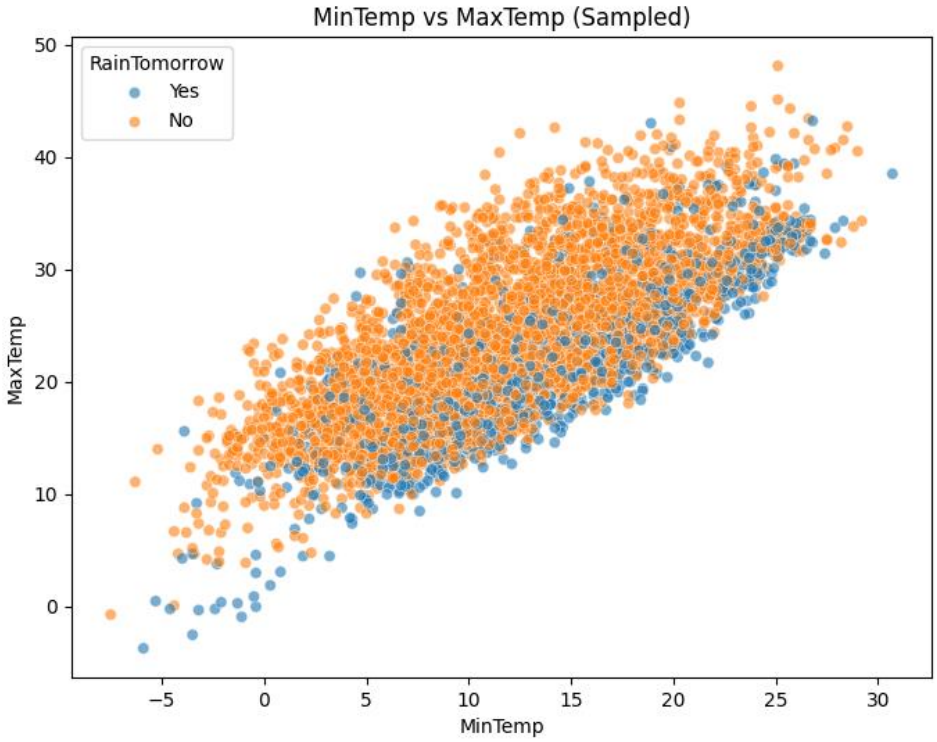
1.2. Bivariate Analysis

The correlation matrix for the variables is given below:

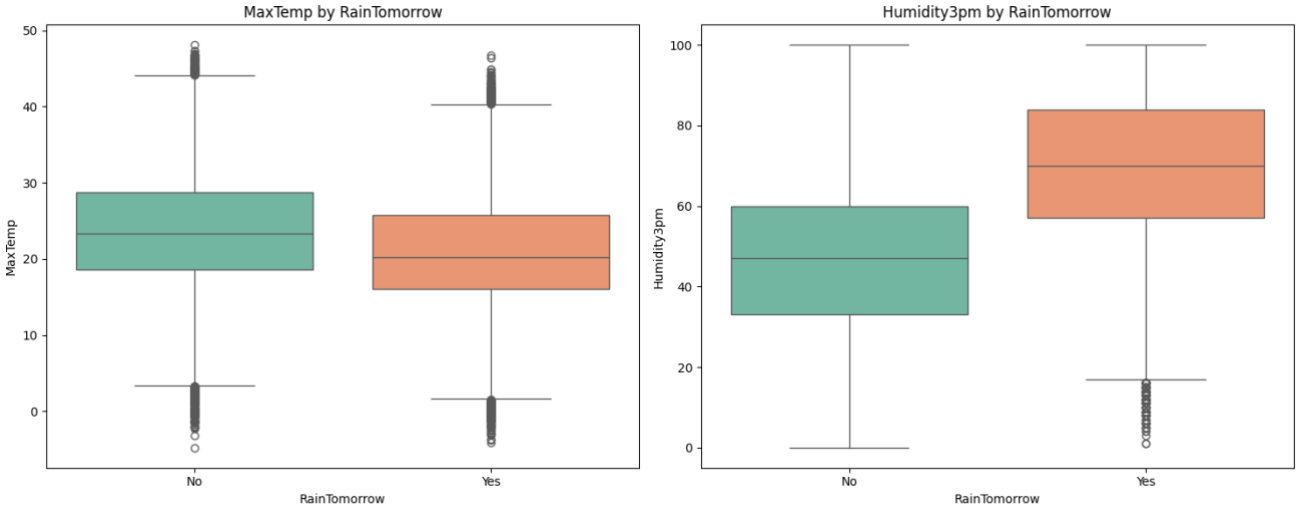


- There is a very strong positive correlation between *MinTemp* and *MaxTemp* and *Temp3pm*, which is expected as warmer nights often precede warmer days.
- *Humidity3pm* has a positive correlation with *Rainfall* and *Cloud9am*, *Cloud3pm*, and a strong negative correlation with *Sunshine* and *MaxTemp*.

The scatter plot of *MinTemp* and *MaxTemp* given below, shows a strong linear relationship. The hue *(RainTomorrow)* suggests that rainy days (orange) tend to have a narrower range between min and max temperatures or occur at specific humidity levels.



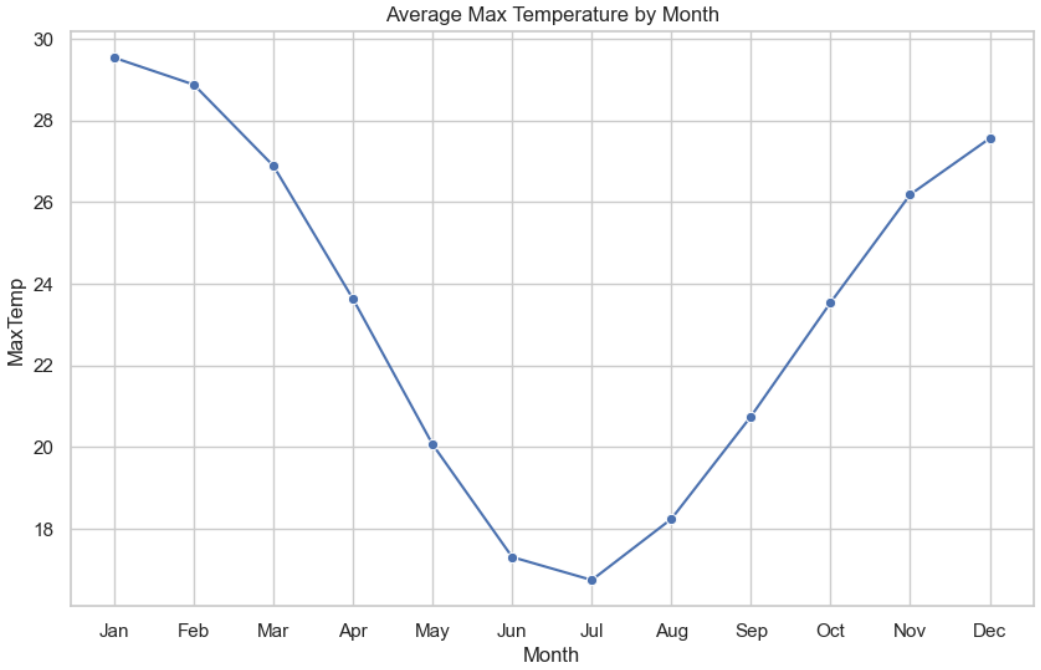
The boxplot of *MaxTemp* vs *RainTomorrow* shows that rainy days tend to have slightly lower maximum temperatures on average. The boxplot of *Humidity3pm* vs *RainTomorrow* shows that days with "*RainTomorrow*=Yes" have significantly higher median humidity at 3pm compared to dry days.



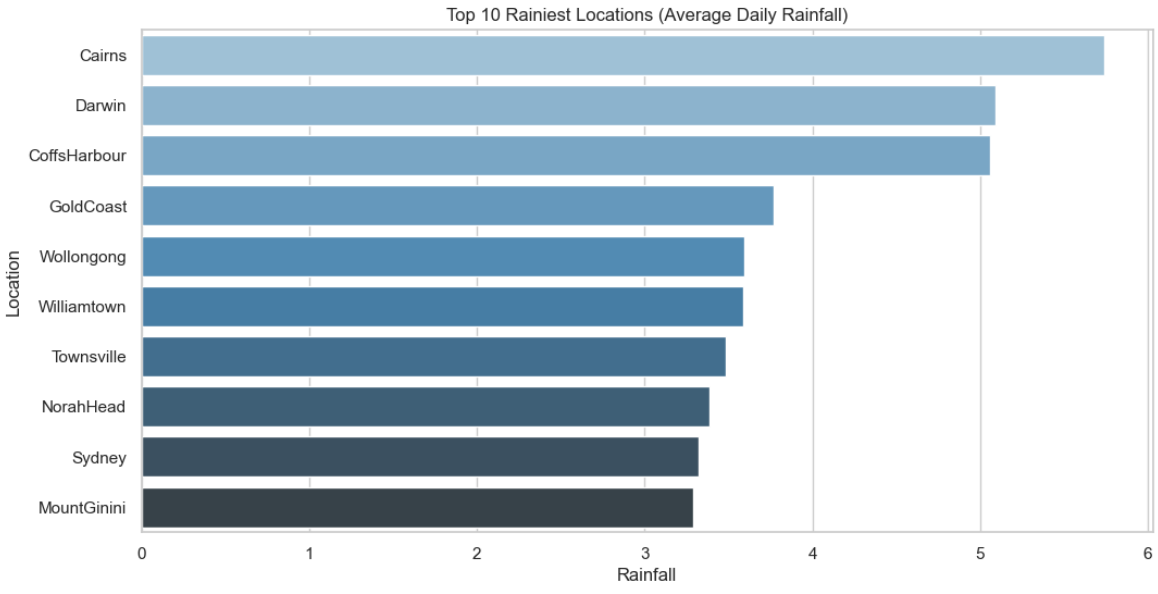
The **Cramer's V score** between *RainToday* and *RainTomorrow*, is calculated as 0.3131, which indicates a moderate positive association, that if it rains today, the probability of it raining tomorrow increases considerably.

1.3. Patterns, Trends and Anomalies

- **Seasonality:** The "Average Max Temperature by Month" plot confirms the Australian seasonal cycle, with peaks in January/February (Summer) and lows in July (Winter).



- **Geographic Variability:** Areas like Cairns, Darwin and Coffs Harbour experience significantly higher average rainfall than inland regions.

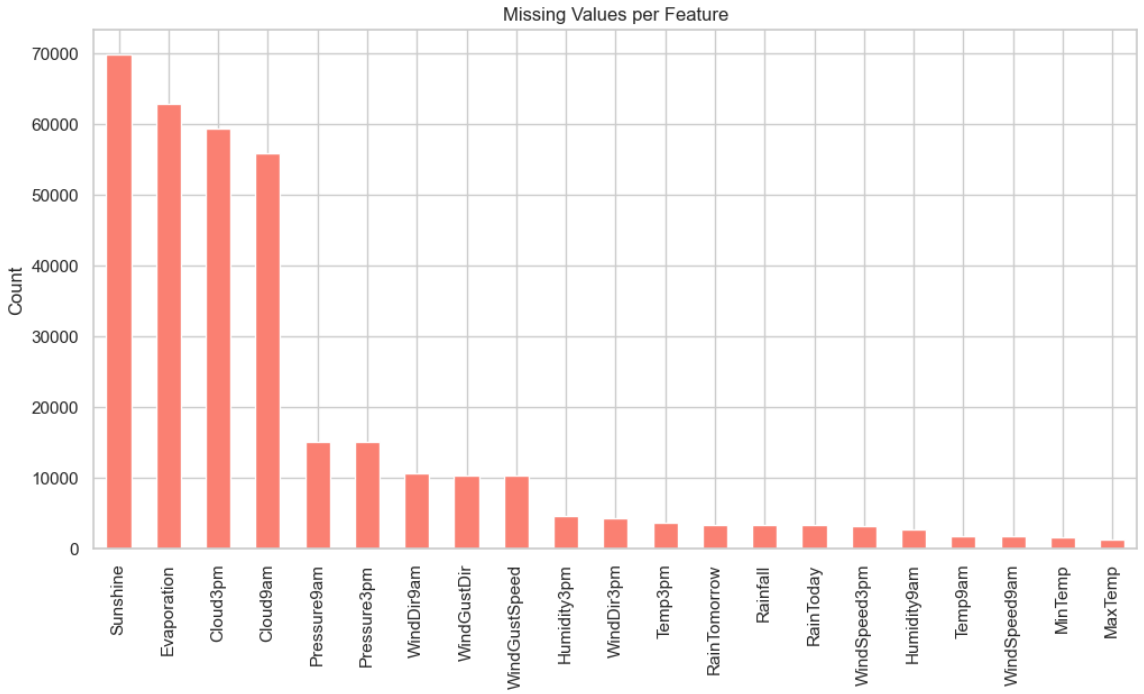


- **Anomalies:** The *Rainfall* variable is extremely zero-inflated (75th percentile is just 0.8mm). However, the maximum is 371mm. This suggests that "*Rainfall*" might need to be transformed for many models to handle it effectively, as the raw values span orders of magnitude.
- **Multicollinearity:** The high correlation between 9am and 3pm measurements (e.g., Temp, Pressure) suggests redundant information. We may use afternoon measurements, which are closer to the target time) or dimensionality reduction to address this problem.

1.4. Missing Data and Outliers

The missing data values per feature given below shows that the variables *Sunshine, Evaporation, Cloud3pm* and *Cloud9am* have very high missing rates. Variables with >40% missingness were dropped to prevent bias. For others, median imputation was used for numerical features (robust to outliers) and mode imputation for categorical features. *RainTomorrow* has ~2.2% missing values. These rows must be dropped for training the classification model.

	Top Missing Values (%)
Sunshine	48.009762
Evaporation	43.166506
Cloud3pm	40.807095
Cloud9am	38.421559
Pressure9am	10.356799
Pressure3pm	10.331363
WindDir9am	7.263853
WindGustDir	7.098859
WindGustSpeed	7.055548
Humidity3pm	3.098446



The IQR method identified outliers as:

```
--- Outlier Detection (IQR Method) ---
Rainfall: 25578 outliers identified
WindGustSpeed: 3092 outliers identified
MaxTemp: 489 outliers identified
```

Rainfall: Over 25,000 outliers detected using the IQR method. These are likely real weather events (storms) rather than errors. Thus, instead of removing them blindly, we applied StandardScaler to normalize the data and selected robust models like Random Forests.

WindGustSpeed: ~3,000 outliers. High winds are rare but significant for predicting storms. Linear models (Regression) might be sensitive to these extremes; Tree-based models (Random Forest, XGBoost) will handle them better.

2. PROBABILITY AND SAMPLING

In this section, we move from descriptive statistics to modelling. We selected Maximum Temperature (*MaxTemp*) as our key variable for probability modeling because the EDA revealed its distribution to be approximately symmetric and bell-shaped, making it an ideal candidate for modelling.

2.1. Estimation and Distribution Fitting

We hypothesized that *MaxTemp* follows a Normal (Gaussian) Distribution $N(\mu, \sigma^2)$. Using the Maximum Likelihood Estimation (MLE) method, we estimated the population parameters from the dataset:

```
MLE Parameters: Mean (mu) = 23.2213, Std Dev (sigma) = 7.1190
```

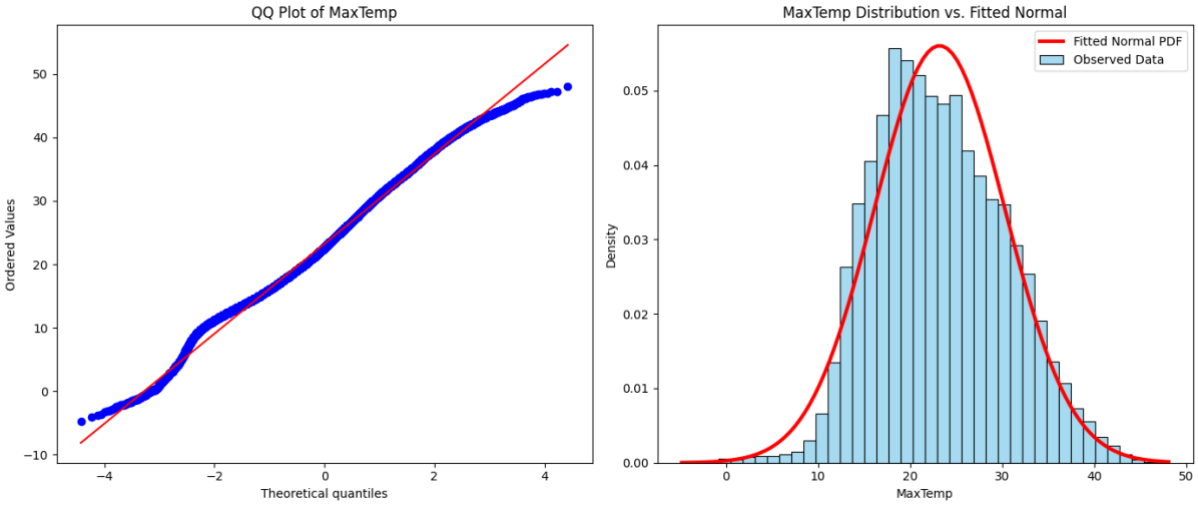
To verify this assumption, we performed a Kolmogorov-Smirnov (KS) Test. The test yielded a statistic of 0.0414 and a p-value close to 0:

```
KS Test: Statistic=0.0414, p-value=6.8922e-215
```

Statistically, this rejects the null hypothesis that the data is perfectly Normal. However, this is a common phenomenon with very large datasets (N > 144,000), where even negligible deviations from the theoretical curve (e.g., such as slight kurtosis in the tails) are flagged as significant.

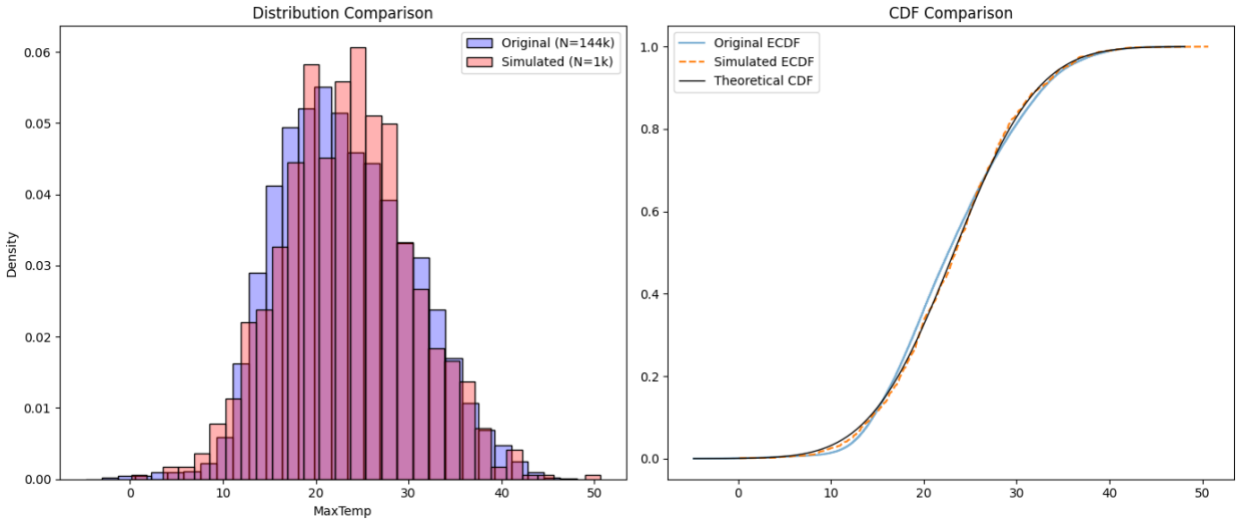
Despite the KS test result, the QQ Plot shows the data points follow the theoretical diagonal line closely, and the Histogram confirms that the theoretical Normal curve (red line) overlaps the observed

data density (blue bars) almost perfectly. We can conclude that the Normal model is a valid approximation for practical purposes, though there is a slight deviation in the tails.



2.2. Simulation and Validation

To further validate our generative model, we performed a Monte Carlo simulation. We generated a synthetic sample of 1,000 data points drawn from $N(23.22, 7.12^2)$. and compared it to the original data.



As seen in above figures, the simulated sample closely mirrors the original data. The Empirical CDF of the simulated data tracks the theoretical CDF almost perfectly, confirming the model captures the data's structure.

Chi-Square Test: We ran a Chi-Square Goodness-of-Fit test on the simulated sample against the theoretical distribution. This yielded a p-value of 0.4793.

Chi-Square Test (Simulated): Statistic=18.6529, p-value=0.4793

Since $p > 0.05$, we fail to reject the null hypothesis. The simulated sample of 1,000 drawn is statistically indistinguishable from a true Normal distribution, which validates that estimated parameters successfully generate realistic weather data.

2.3. Confidence Intervals

Then, we calculated 95% confidence intervals for the mean maximum temperature using two methods:

95% CI (Parametric): (23.184603755107275, 23.258092796186425)
95% CI (Bootstrap): (23.1842, 23.2577)

The two intervals are nearly identical. This convergence occurs because the sample size is massive ($N \approx 144{,}000$). By the Central Limit Theorem, the sampling distribution of the mean converges to Normality regardless of the population's underlying shape, satisfying the assumptions for the parametric t-interval and rendering the computationally expensive bootstrap method redundant in this case.

3. DIMENSIONALITY REDUCTION

In this part, we apply Principal Component Analysis to the numerical features of the weather dataset. The goal is to reduce the 16+ correlated numerical variables (such as multiple temperature and pressure) into a smaller set of uncorrelated components while retaining the maximum amount of information.

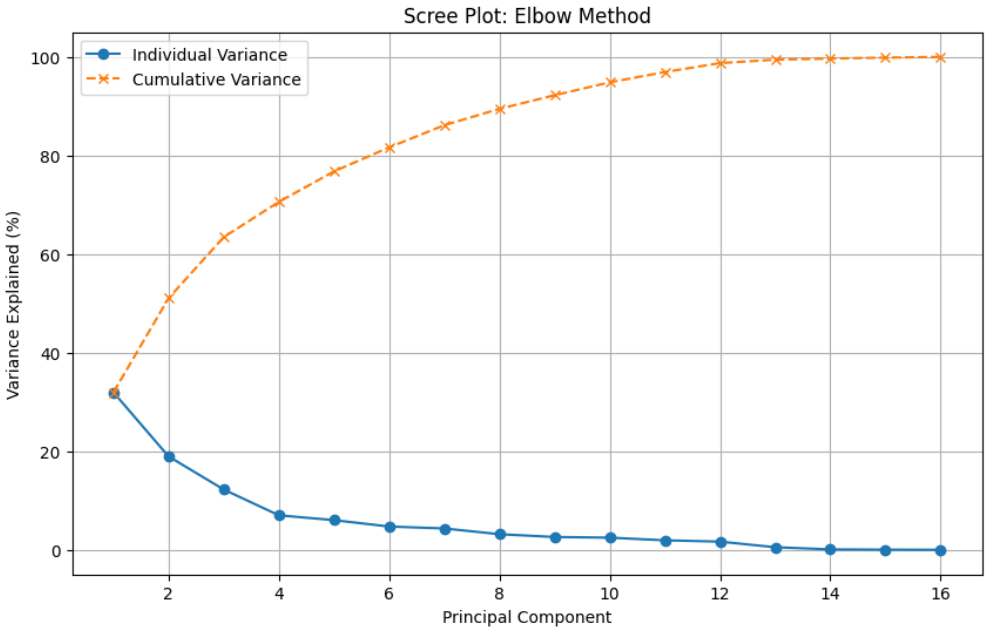
3.1. Variance Analysis

We analyzed the explained variance ratio to determine the optimal number of components. The first three principal components alone capture 63.49% of the total variance in the dataset.

--- Variance Explained by First 3 Components ---
PC1: 32.06% (Cumulative: 32.06%)
PC2: 19.07% (Cumulative: 51.13%)
PC3: 12.37% (Cumulative: 63.49%)

Scree Plot: The scree plot given below displays the variance explained by each component. We observe a steep drop after the first component and an elbow forming around the 4th or 5th component.

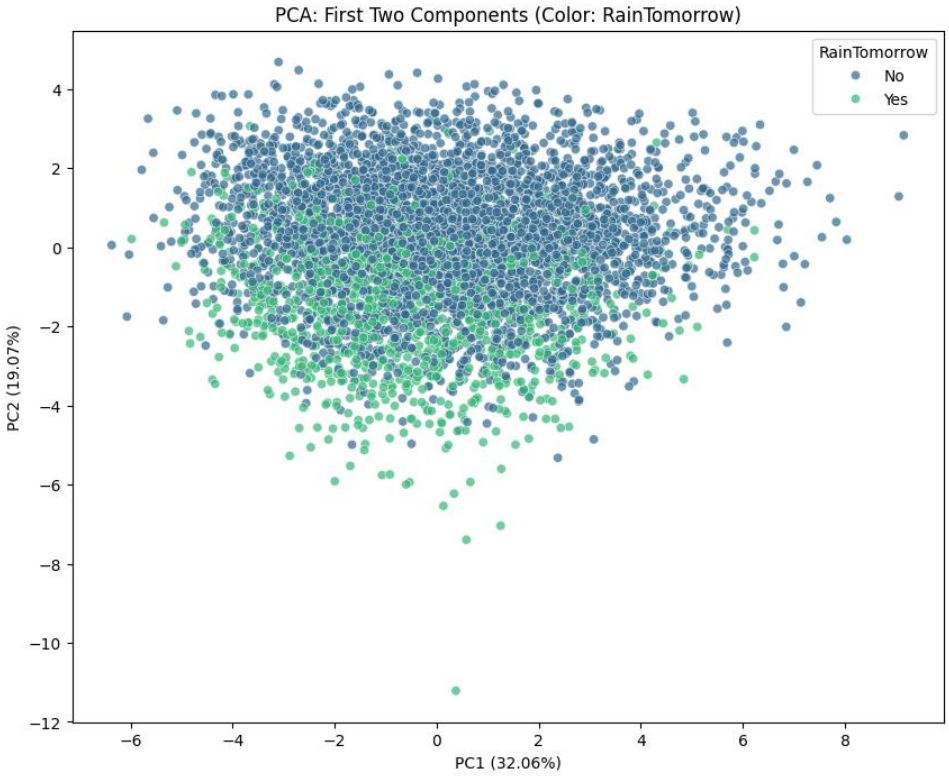
This suggests that retaining the first 3-4 components are likely sufficient to capture most of the patterns while effectively filtering out noise.



3.2. Visualization and Interpretation

To visualize the high-dimensional data, we plotted the first two principal components against each other.

PC1 vs. PC2 Scatter Plot: As shown in the below, the scatter plot reveals a dense cloud of points with no perfect linear separation between "*Rain*" and "*No Rain*" classes, which is expected for complex weather systems. However, points representing "*Rain*" tend to cluster in specific regions, indicating that PCA has preserved discriminative information useful for classification.



Component Interpretation: By examining the variable loadings in the below plot, we assigned physical meanings to the components:

- **PC1 - "The Thermal Component":**
  - Top Variables: *MaxTemp, Temp3pm, Temp9am, MinTemp*
  - This component is dominated by temperature features. A high value corresponds to a hot day, while a low value indicates a cold day. It effectively summarizes the daily thermal profile into a single index.
- **PC2: "The Atmospheric Pressure & Cloudiness Component":**
  - Top Variables: *Pressure9am, Cloud3pm, Cloud9am, WindGustSpeed, Pressure3pm*
  - This component captures the interaction between pressure, clouds, and wind. Meteorologically, low pressure is associated with storms; thus, this component likely distinguishes between "Clear/High-Pressure" days and "Stormy/Low-Pressure" days.

--- Top Contributing Variables (Loadings) ---

PC1 (Top 5):

MaxTemp 0.397723

Temp3pm 0.391202

Temp9am 0.383953

MinTemp 0.321280

Humidity9am 0.293100

Name: PC1, dtype: float64

PC2 (Top 5):

Pressure9am 0.355823

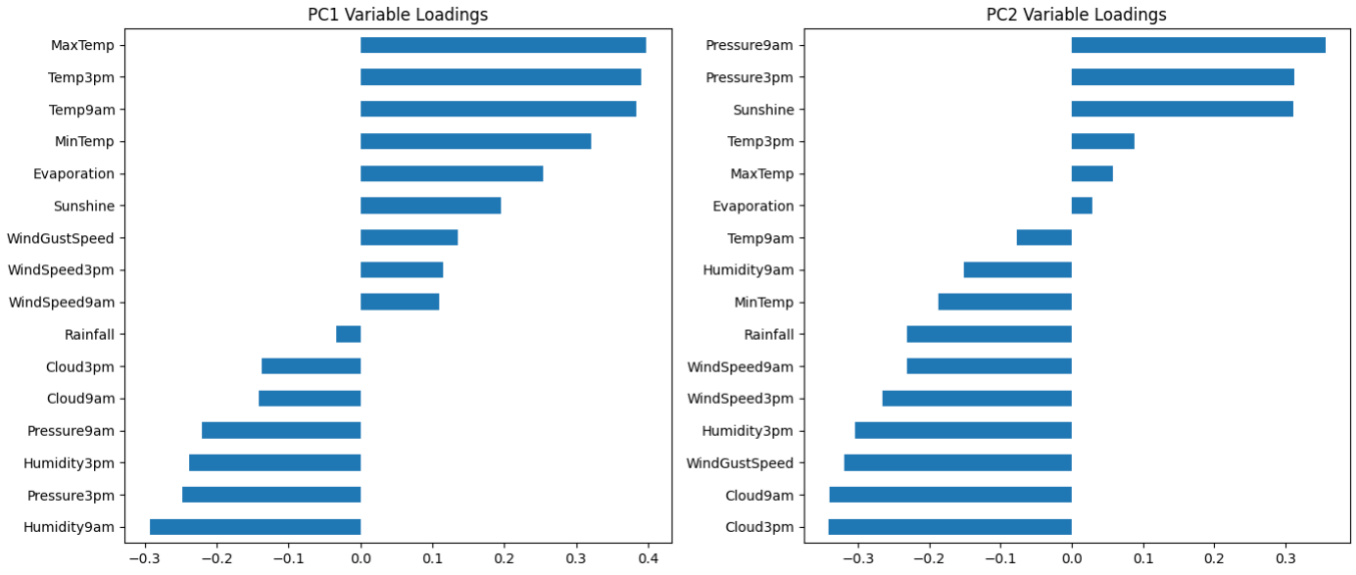
Cloud3pm 0.341121

Cloud9am 0.340226

WindGustSpeed 0.319180

Pressure3pm 0.312108

Name: PC2, dtype: float64



3.3. Discussion: Impact and Methods

Benefits:

- **Multicollinearity**: The original dataset contains pairs with near-perfect correlation (e.g., Temp9am vs. Temp3pm), which can destabilize linear models. PCA resolves this by creating orthogonal components.

- **Noise Reduction**: By focusing on the top components, we discard low-variance signals often attributable to sensor noise.

Drawbacks:

- **Loss of Interpretability**: Models using "PC1" as a feature are harder to explain to stakeholders than those using raw variables like "*Temperature*".

- **Information Loss**: The discarded variance (~36%) may contain subtle signals critical for predicting rare outlier events.

PCA vs. t-SNE: While t-SNE is superior for visualization and revealing local clusters due to its non-linear nature, PCA was chosen here because it preserves global variance and creates a parametric mapping. This makes PCA suitable for feature engineering in a predictive pipeline, whereas t-SNE cannot easily transform new, unseen data.

4. CLASSIFICATION

The primary objective here is to build a predictive model to classify whether it will rain tomorrow (*RainTomorrow* = Yes/No). This provides actionable insights for daily weather forecasting.

4.1. Model Selection and Performance

We trained and evaluated two distinct classifiers to establish a performance baseline and assessing non-linear capabilities:

- **Logistic Regression**: A linear baseline model.
- **Random Forest Classifier**: A non-linear ensemble method robust to outliers and complex interactions.

Performance Summary: Table below summarizes the key metrics on the test set.

Metric	Logistic Regression	Random Forest
Accuracy	84.51%	**85.56%**
AUC Score	0.86	**0.87**
Precision (Rain)	0.73	**0.77**
Recall (Rain)	0.50	0.50
F1-Score (Rain)	0.59	**0.61**

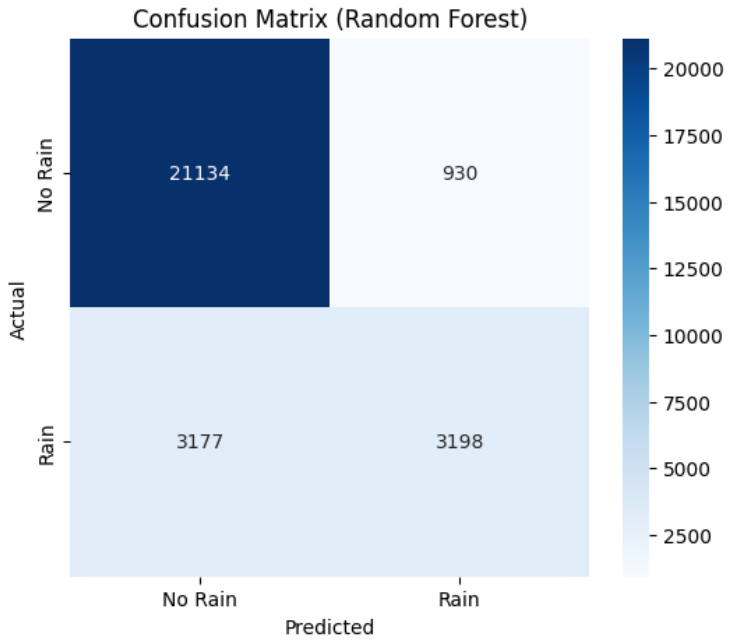
Both models perform well, achieving accuracies around 85%. The Random Forest model holds a slight edge in overall performance, achieving a higher AUC (0.87 vs. 0.86) and better Precision (0.77 vs. 0.73).

A critical insight from the results is the Recall score of approximately 0.50 for the "*Rain*" class. The model currently detects only about half of the actual rainy days. This is a direct consequence of the class imbalance identified in the EDA (only ~22% of days are rainy). The model is biased toward the majority class ("*No Rain*"). For a real-world application, we might need to lower the decision threshold (currently 0.5) to catch more rain events, accepting a trade-off of slightly more false alarms.

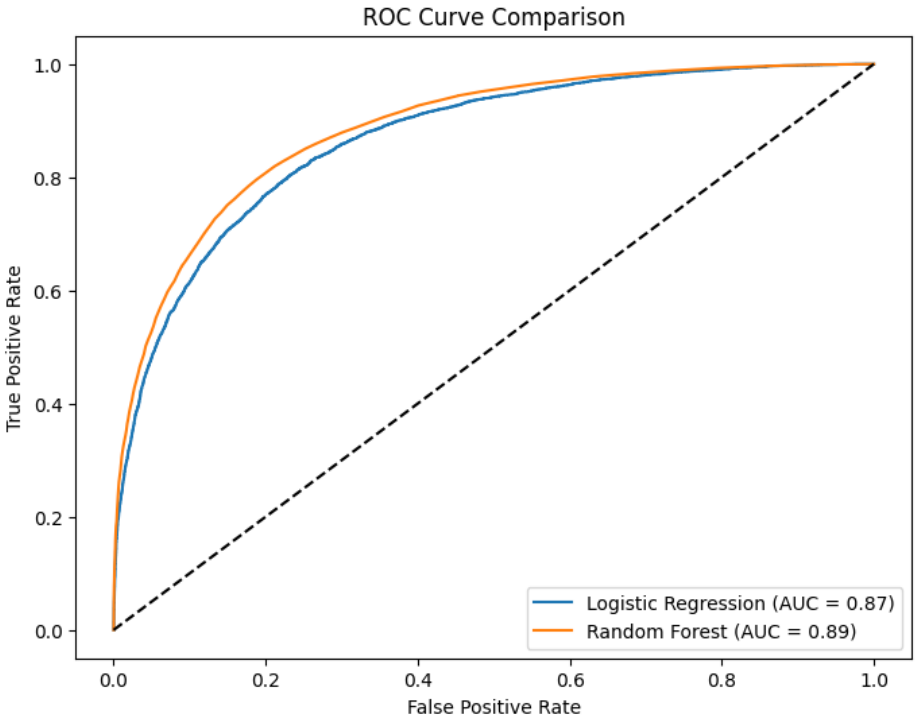
4.2. Visualization of Results

Confusion Matrix Analysis: The confusion matrix given below visualizes the model's biases.

- **True Negatives**: The model is extremely effective at predicting "*No Rain*," correctly identifying the vast majority of dry days.
- **False Negatives**: The relatively high number of false negatives (actual rain predicted as dry) confirms the recall issue discussed above.

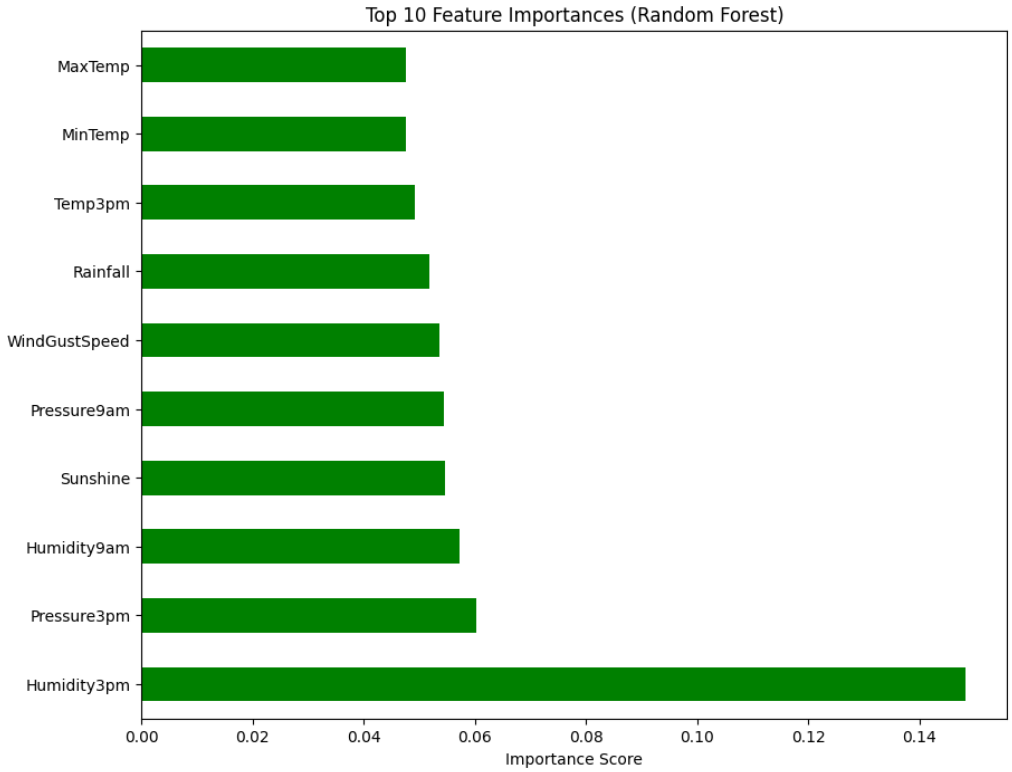


ROC Curve Comparison: ROC curve given below illustrates the trade-off between sensitivity and specificity. The curves for both models arch well above the diagonal random guess line, indicating strong predictive power. The Random Forest curve stays consistently above the Logistic Regression curve, confirming it as the superior model for this task across different thresholds.



4.3. Feature Importance

To understand why the model predicts rain, we analyzed the feature importance scores from the Random Forest model.



Top Predictors of Rain:

i. *Humidity3pm*: This is the single most critical predictor. High humidity in the afternoon is a massive physical signal for impending precipitation.

ii. *Humidity9am*: The model identifies morning humidity as the second most vital indicator. A moist atmosphere at the start of the day sets the stage for saturation and rain later on.
iii. *Sunshine*: Shows a strong inverse relationship; lower sunshine hours strongly predict rain.
iv. *Pressure3pm*: Low atmospheric pressure is a classic meteorological indicator of incoming storms.

These findings align perfectly with earlier EDA and PCA results, confirming that "Atmospheric" variables (Moisture and Pressure) are the key drivers of rainfall patterns.

5. REGRESSION ANALYSIS

In this part, we developed a regression model to predict a numerical outcome: *MaxTemp*. By modeling *MaxTemp* as a function of morning observations and other atmospheric variables, we aim to understand the drivers of daily temperature peaks.

5.1. Model Performance

We trained two models to establish a baseline and test for stability: an OLS Linear Regression and a Ridge Regression (regularized). The models achieved nearly identical and good performance on the test set:

--- OLS Performance ---
RMSE: 1.9471
R2: 0.9265

--- Ridge Performance ---
RMSE: 1.9471
R2: 0.9265

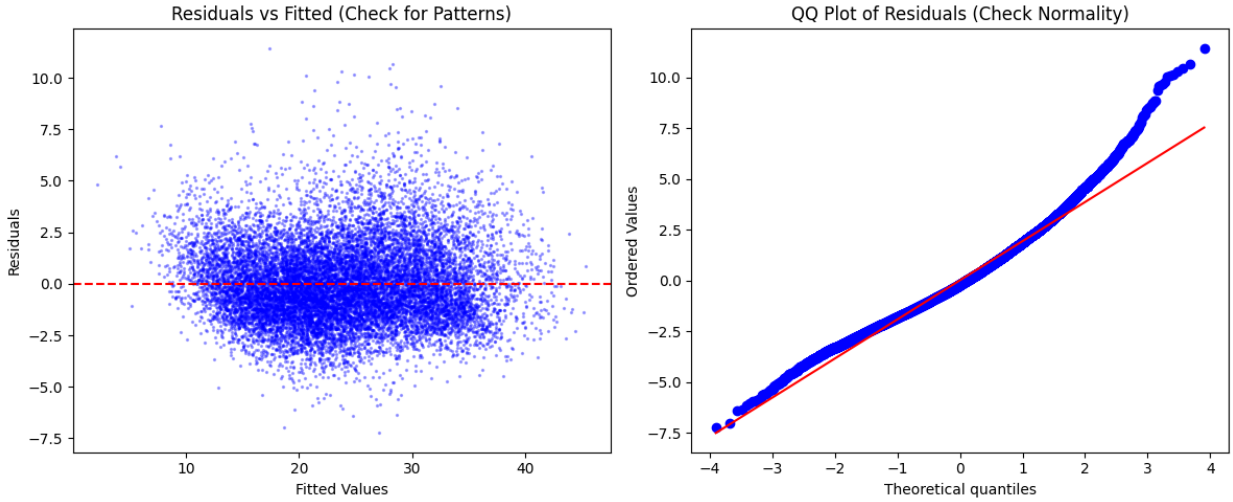
The R^2 value of 0.9265 indicates that our model explains approximately 92.7% of the variance in maximum daily temperatures. The RMSE of 1.95 implies that the model's predictions are typically within ~2°C of the actual temperature. The fact that Ridge Regression produced identical results suggests that while multicollinearity is present, it has not severely destabilized the OLS coefficients for prediction purposes.

5.2. Diagnostic Checks

To ensure the validity of our regression inferences, we performed standard residual diagnostics:

- **Linearity & Homoscedasticity**: The "Residuals vs. Fitted" plot given below shows a random scatter of points around zero with no obvious curvature. This confirms that the linear assumption holds well and the variance of errors is constant.

- **Normality:** The Q-Q Plot given below shows the residuals following closely to the diagonal red line, confirming that the errors are normally distributed.



5.3. Interpretation of Coefficients

We analysed the coefficients to understand the physical indicators of *MaxTemp*.

--- OLS Model Summary (Coefficients) ---

	coef	std err	t	P>|t|	[0.025	0.975]
const	23.7273	0.008	2973.550	0.000	23.712	23.743
MinTemp	-0.0343	0.025	-1.375	0.169	-0.083	0.015
Rainfall	-0.0116	0.009	-1.359	0.174	-0.028	0.005
WindGustSpeed	-0.1230	0.009	-13.058	0.000	-0.142	-0.105
Humidity9am	1.2893	0.014	91.622	0.000	1.262	1.317
Humidity3pm	-2.8635	0.013	-221.021	0.000	-2.889	-2.838
Pressure9am	2.1150	0.032	65.084	0.000	2.051	2.179
Pressure3pm	-2.3593	0.032	-73.104	0.000	-2.423	-2.296
Temp9am	6.2434	0.027	227.911	0.000	6.190	6.297
Cloud9am	-0.2066	0.011	-18.431	0.000	-0.229	-0.185
Cloud3pm	0.1059	0.011	9.854	0.000	0.085	0.127

Significant Predictors:

i. *Temp9am* (+6.24): This is by far the strongest positive predictor. A higher temperature at 9 AM naturally sets a higher baseline for the day's peak.

ii. *Humidity3pm* (-2.86): This variable has a strong negative effect. Higher afternoon humidity typically correlates with cloud cover or evaporative cooling, which suppresses the temperature rise.

iii. Pressure: Interestingly, *Pressure9am* has a positive coefficient (+2.12) while *Pressure3pm* has a negative one (-2.36). This opposing sign pattern suggests that the change or gradient in pressure throughout the day is a more significant physical driver than the absolute pressure value itself.

Multicollinearity Analysis: The Variance Inflation Factor (VIF) analysis detected significant multicollinearity, particularly for *Pressure9am* (VIF=16.5) and *Pressure3pm* (VIF=16.3), which exceed the standard threshold of 10. While high VIF increases the variance of individual coefficient estimates, it does not degrade the predictive power (R^2)of the model. Since the primary goal is prediction, retaining these variables is justified.

--- Variance Inflation Factors (VIF) ---

	Feature	VIF
5	Pressure9am	16.503952
6	Pressure3pm	16.294229
7	Temp9am	11.782138
0	MinTemp	9.763563
3	Humidity9am	3.099324
4	Humidity3pm	2.628993
8	Cloud9am	1.969224
9	Cloud3pm	1.818399
2	WindGustSpeed	1.391638
1	Rainfall	1.144169

5.4. Model Robustness

To assess stability, we performed 10-Fold Cross-Validation:

--- Cross-Validation (10-Fold) ---
Average RMSE: 1.9421
Std RMSE: 0.0137

The extremely low standard deviation confirms that the model is robust and performs consistently across different subsets of the data, with no signs of overfitting.

6. MONTE CARLO METHODS

In this part, we employed Monte Carlo simulation to solve an inferential problem: estimating the probability of an "Extreme Heat Event." We defined an extreme heat event as a day where the Maximum Temperature exceeds 35°C (*MaxTemp* > 35).

6.1. Methodology

We treated the observed dataset as our population and used non-parametric Bootstrap Resampling to estimate the uncertainty of our probability estimate. We simulated 1,000 alternative history scenarios (N=1,000) by resampling with replacement from the original temperature data.

6.2. Results

The simulation yielded highly precise estimates for the risk of extreme heat:

- **Empirical Probability**: 5.57%. This is the baseline probability observed in the raw data.
- **Monte Carlo Mean Estimate**: 5.57%. The simulation average aligns almost perfectly with the empirical truth, confirming the estimator is unbiased.
- **Uncertainty**:
    - Standard Error: 0.00062.
    - 95% Confidence Interval: [5.45%, 5.69%].

We are 95% confident that the true long-term probability of an extreme heat day lies between 5.45% and 5.69%. The narrowness of this interval indicates a high degree of precision, attributable to the large sample size (~144,000 records).

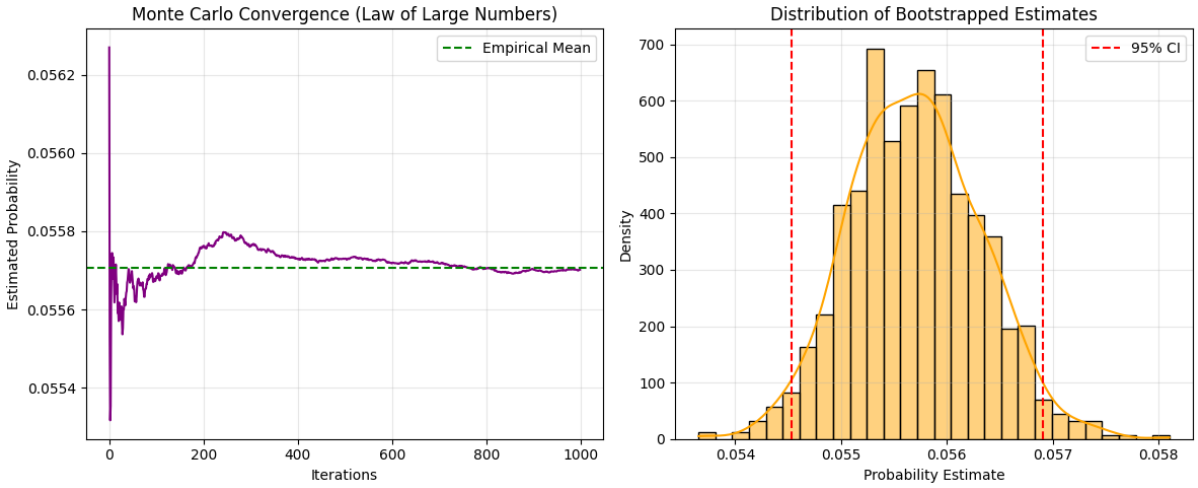
```
--- Monte Carlo Results (N=1000) ---
Empirical Probability: 0.05571 (approx. 5.57%)
MC Mean Estimate:      0.05570
Standard Error:        0.00062

95% Confidence Interval: [0.05454, 0.05691]
```

6.3. Visualization

We generated two key plots to validate the statistical properties of our simulation:

- **Convergence**: The below figure on the left shows the running mean of our probability estimate. The purple line fluctuates initially but quickly stabilizes and converges to the green empirical mean line. This visually demonstrates the Law of Large Numbers—as we run more simulations, our approximation converges to the true expected value.
- **Distribution (Central Limit Theorem):** The below figure on the right displays the histogram of the 1,000 bootstrapped probability estimates. Despite the underlying data being binary (Yes/No for >35°C), the distribution of the estimates forms a perfect bell curve (Normal distribution). This validates the Central Limit Theorem, allowing us to trust the confidence intervals derived from this method.



7. CONCLUSIONS

This study applied a rigorous statistical framework to ten years of Australian weather data. We successfully established that while weather is inherently chaotic, it adheres to predictable statistical laws. Our analysis confirmed that Maximum Temperature follows a stable Normal distribution, while *Rainfall* is driven by complex interactions between atmospheric moisture and pressure. Using dimensionality reduction, we simplified 16 complex variables into three core components—Thermal, Atmospheric, and Humidity—without losing significant information.

Our predictive modelling demonstrated high efficacy:

- **Forecasting**: The Random Forest classifier predicted rain with 85% accuracy (AUC 0.87), identifying 3 PM Humidity as the critical tipping point for precipitation.
- **Estimation**: The Regression model predicted daily maximum temperatures with 92% accuracy R^2, proving that afternoon cooling effects are quantifiable.

- **Risk**: Monte Carlo simulations provided a precise risk assessment for extreme heat events (5.57% probability), offering a valuable metric for long-term planning.