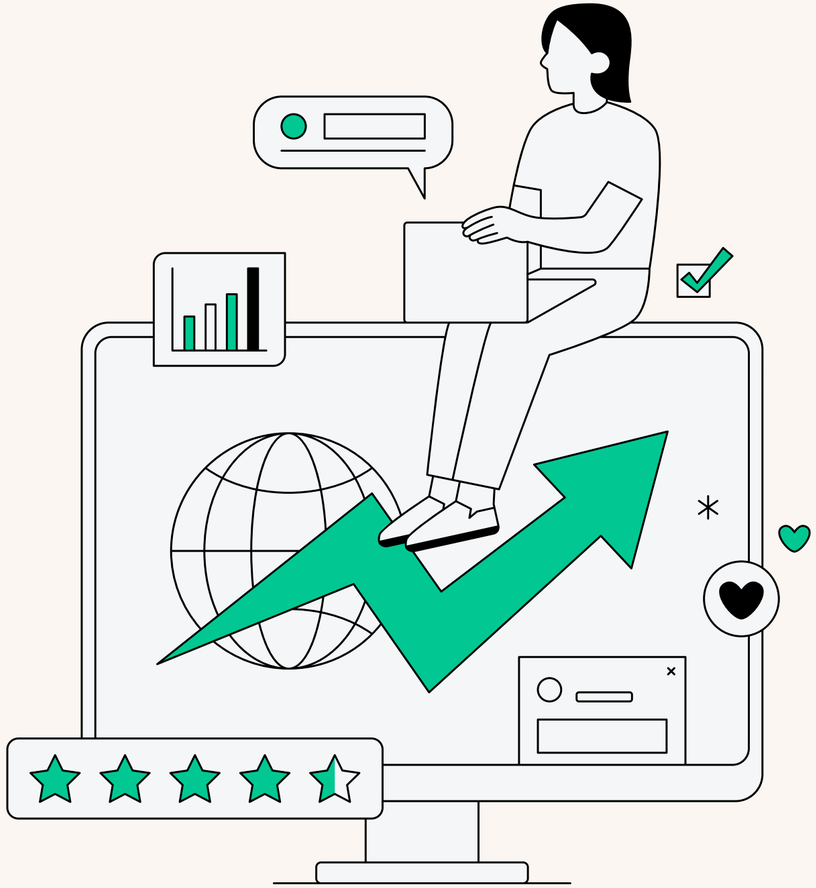
Presented by Mohammad Ehsani

SDA Project

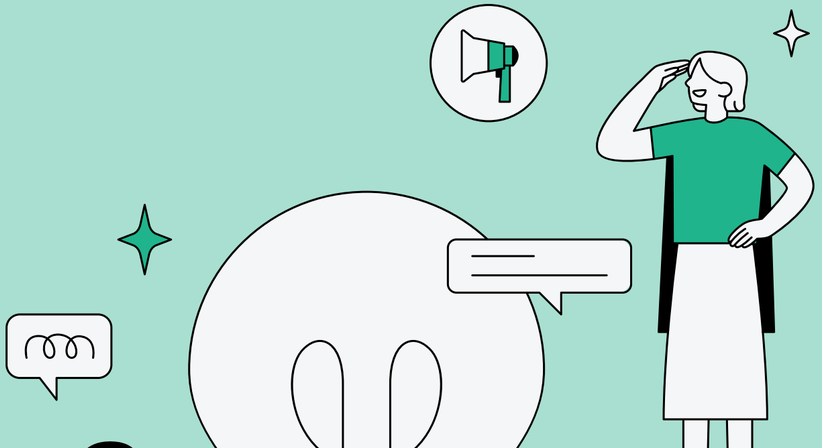
Australia Weather Forecast

Instructor: ROBERTA SICILIANO



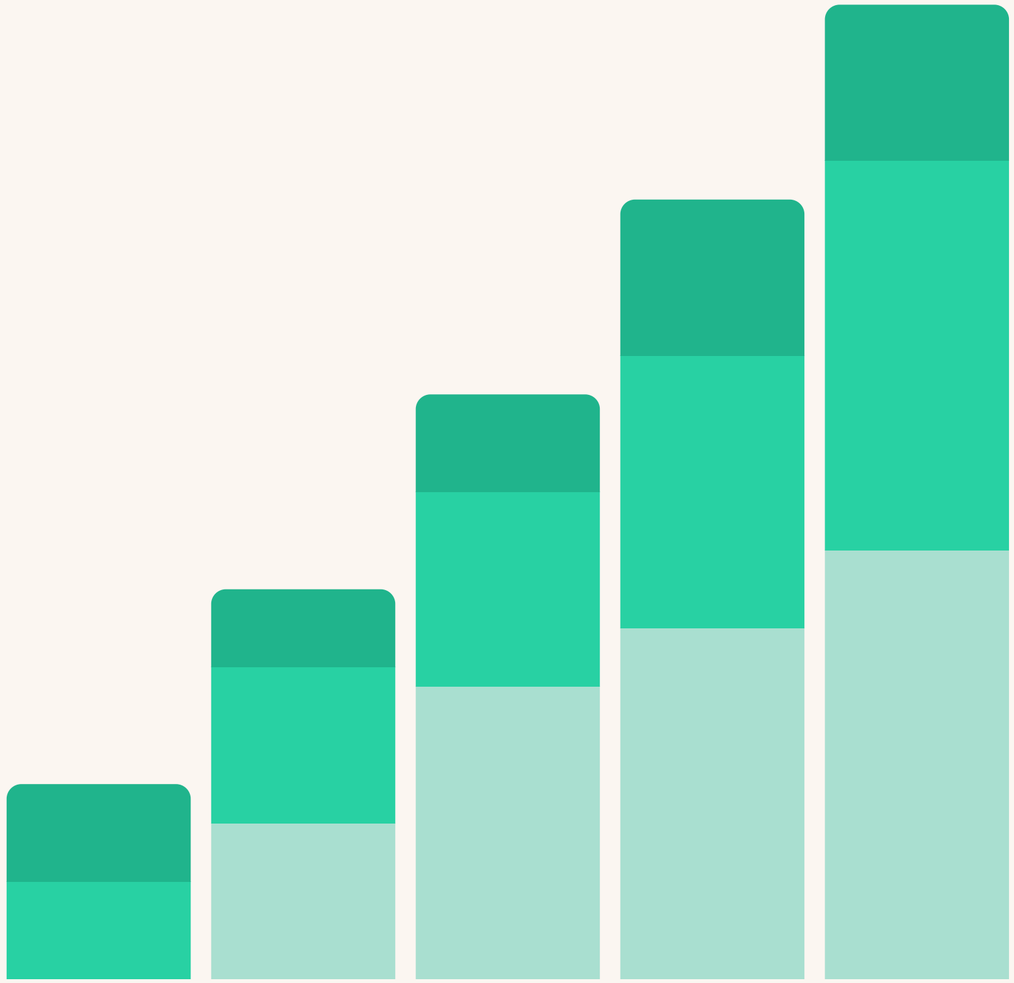
Introduction to results analysis

In this project, my aim is to determine if there will be rainfall in Australia tomorrow. I utilize Python and Scikit-Learn to apply Logistic Regression. My approach involves constructing a classifier to anticipate rainfall occurrences in Australia for the following day. I employ Logistic Regression to train a binary classification model, utilizing the Rain in Australia dataset for this analysis.



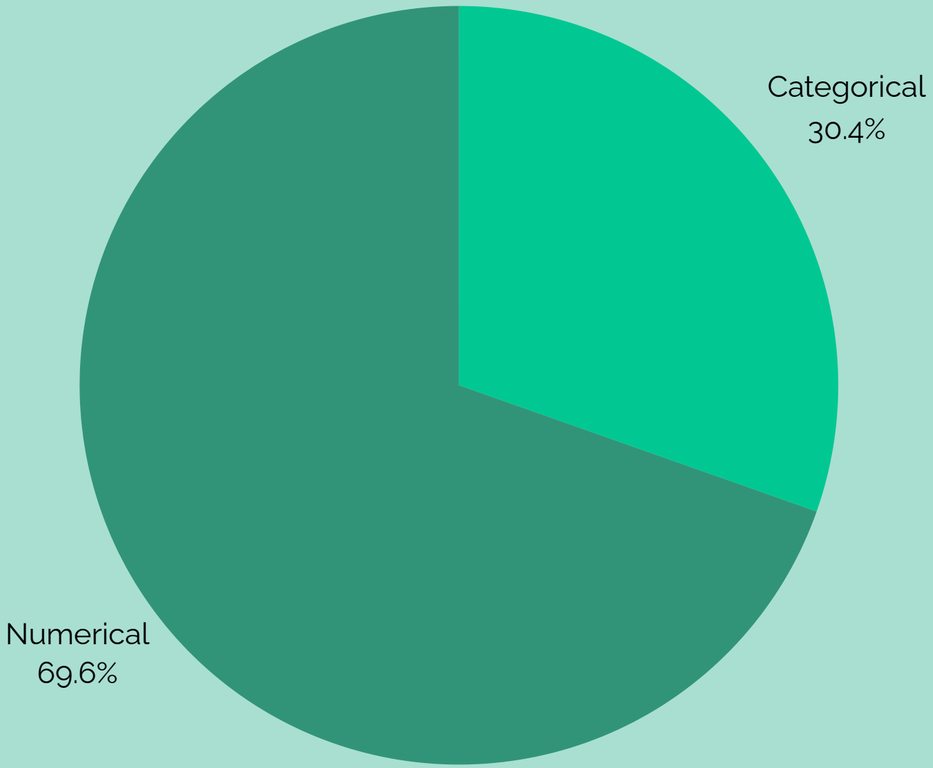
Methodology used in the analysis

This analysis employs a structured approach, beginning with data understanding and **exploratory data analysis (EDA)**, which includes **bivariate analysis**, **data visualization**, and **preprocessing**. **Confirmatory data analysis** follows, encompassing **variable encoding**, **data scaling**, **model training**, and **evaluation**. The methodology concludes with summarizing outcomes from various metrics like **accuracy**, **K-fold cross-validation**, **confusion matrices**, **classification reports**, **probability analyses**, and **ROC-AUC evaluations**, providing comprehensive insights into the case study's findings.



DataSet Structure

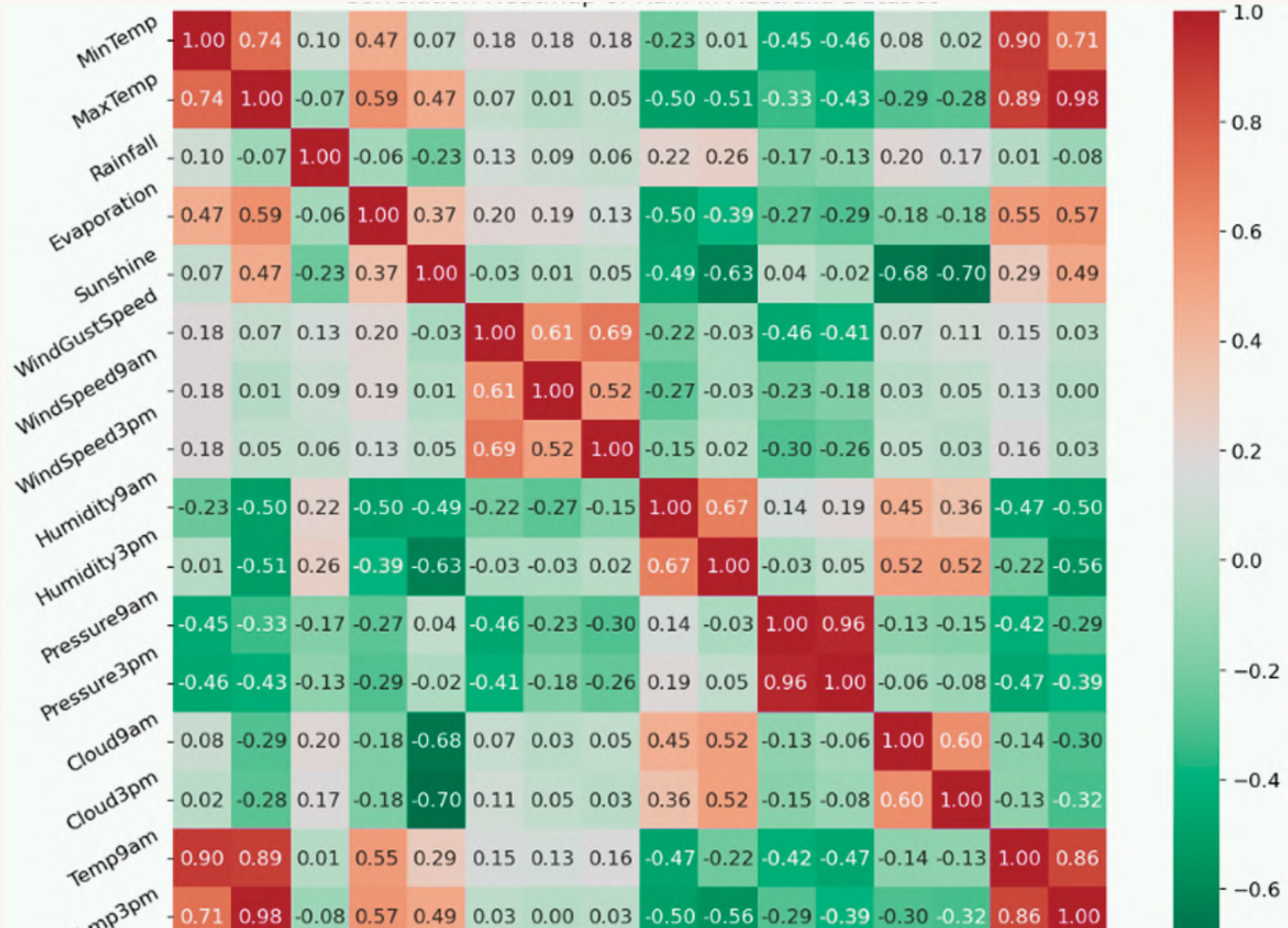
As observed, we have both numerical and categorical variables, each with its respective data type: float for numerical variables and object for categorical variables. Additionally, it's evident that there are some missing values, which I will further address in subsequent slides.



0	Date	145460 non-null	object
1	Location	145460 non-null	object
2	MinTemp	143975 non-null	float64
3	MaxTemp	144199 non-null	float64
4	Rainfall	142199 non-null	float64
5	Evaporation	82670 non-null	float64
6	Sunshine	75625 non-null	float64
7	WindGustDir	135134 non-null	object
8	WindGustSpeed	135197 non-null	float64
9	WindDir9am	134894 non-null	object
10	WindDir3pm	141232 non-null	object
11	WindSpeed9am	143693 non-null	float64
12	WindSpeed3pm	142398 non-null	float64
13	Humidity9am	142806 non-null	float64
14	Humidity3pm	140953 non-null	float64
15	Pressure9am	130395 non-null	float64
16	Pressure3pm	130432 non-null	float64
17	Cloud9am	89572 non-null	float64
18	Cloud3pm	86102 non-null	float64
19	Temp9am	143693 non-null	float64
20	Temp3pm	141851 non-null	float64
21	RainToday	142199 non-null	object
22	RainTomorrow	142193 non-null	object

Correlation for Bivariate Analysis

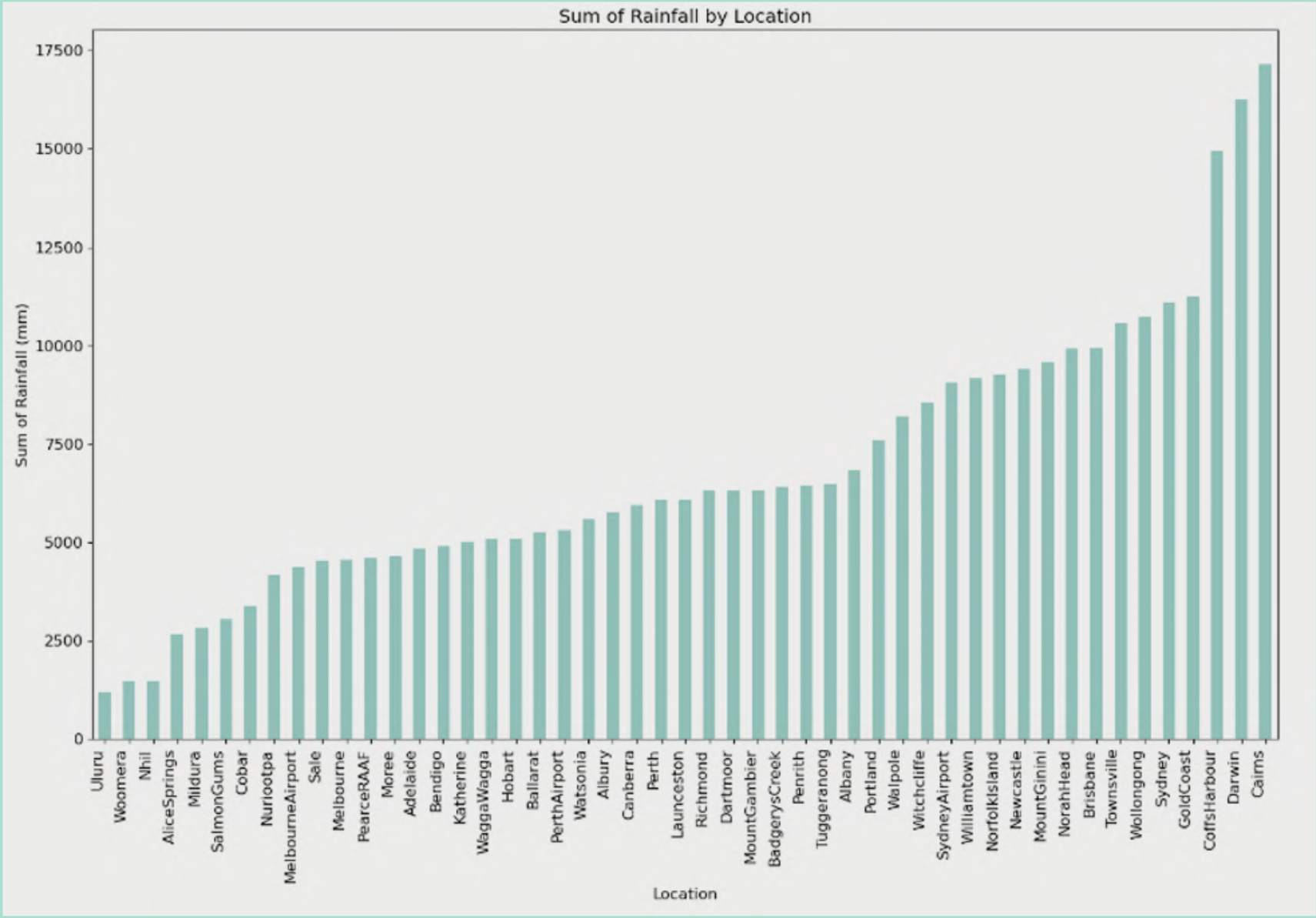
MinTemp & MaxTemp	0.74
MinTemp & Temp3pm	0.71
MinTemp & Temp9am	0.90
MaxTemp & Temp9am	0.89
MaxTemp & Temp3pm	0.98
WindGustSpeed & WindSpeed3pm	0.69
Pressure9am & Pressure3pm	0.96
Temp9am &Temp3pm	0.86



Total Amount Of Rainfall

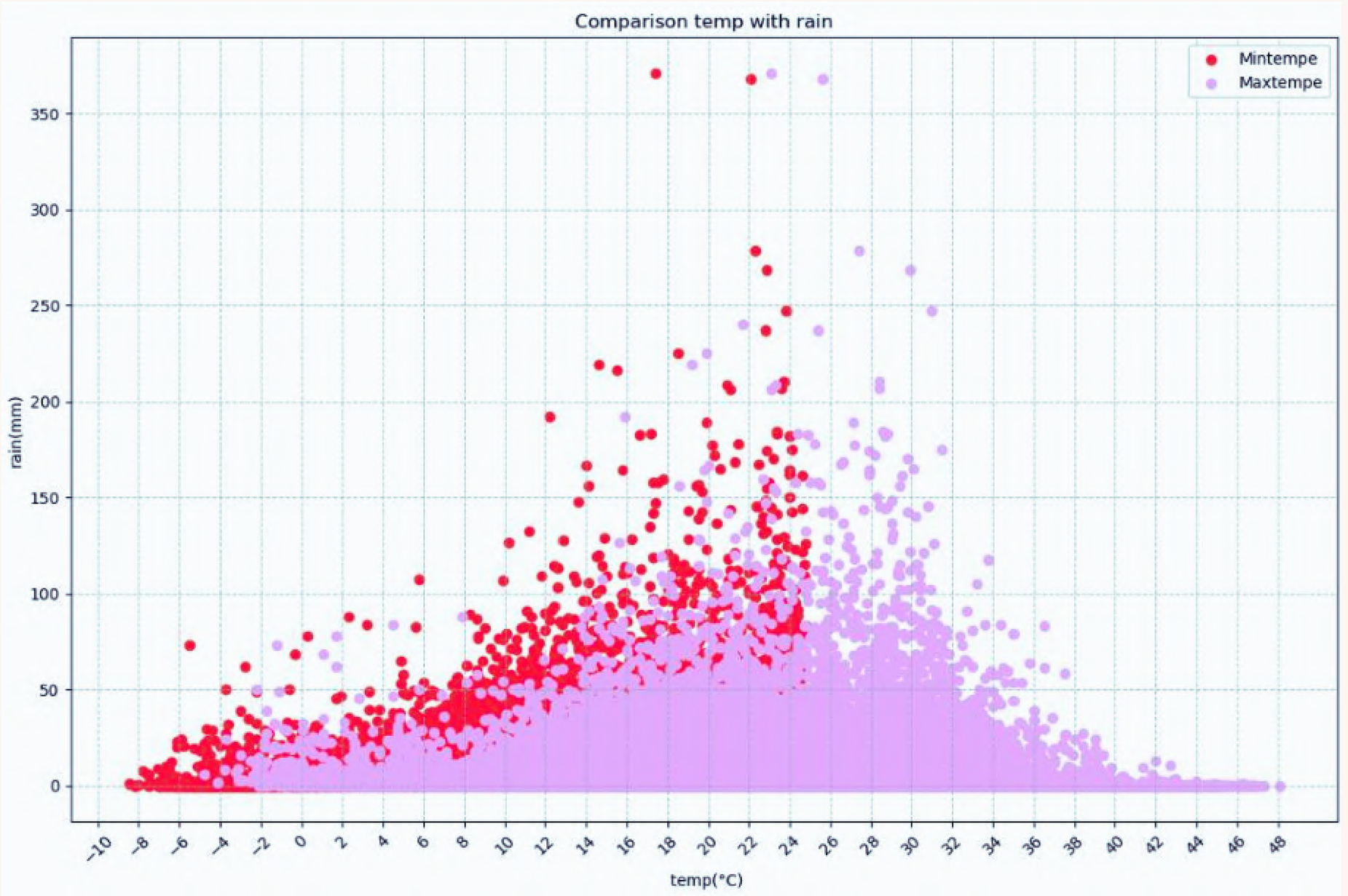
The locations with the highest rainfalls are:
Darwin (2,500 mm)
Cairns (2,500 mm)
Brisbane (2,500 mm)

The locations with the lowest rainfalls are:
Uluru (17,500 mm)
Woomera (15,000 mm)
Nhil (12,500 mm)



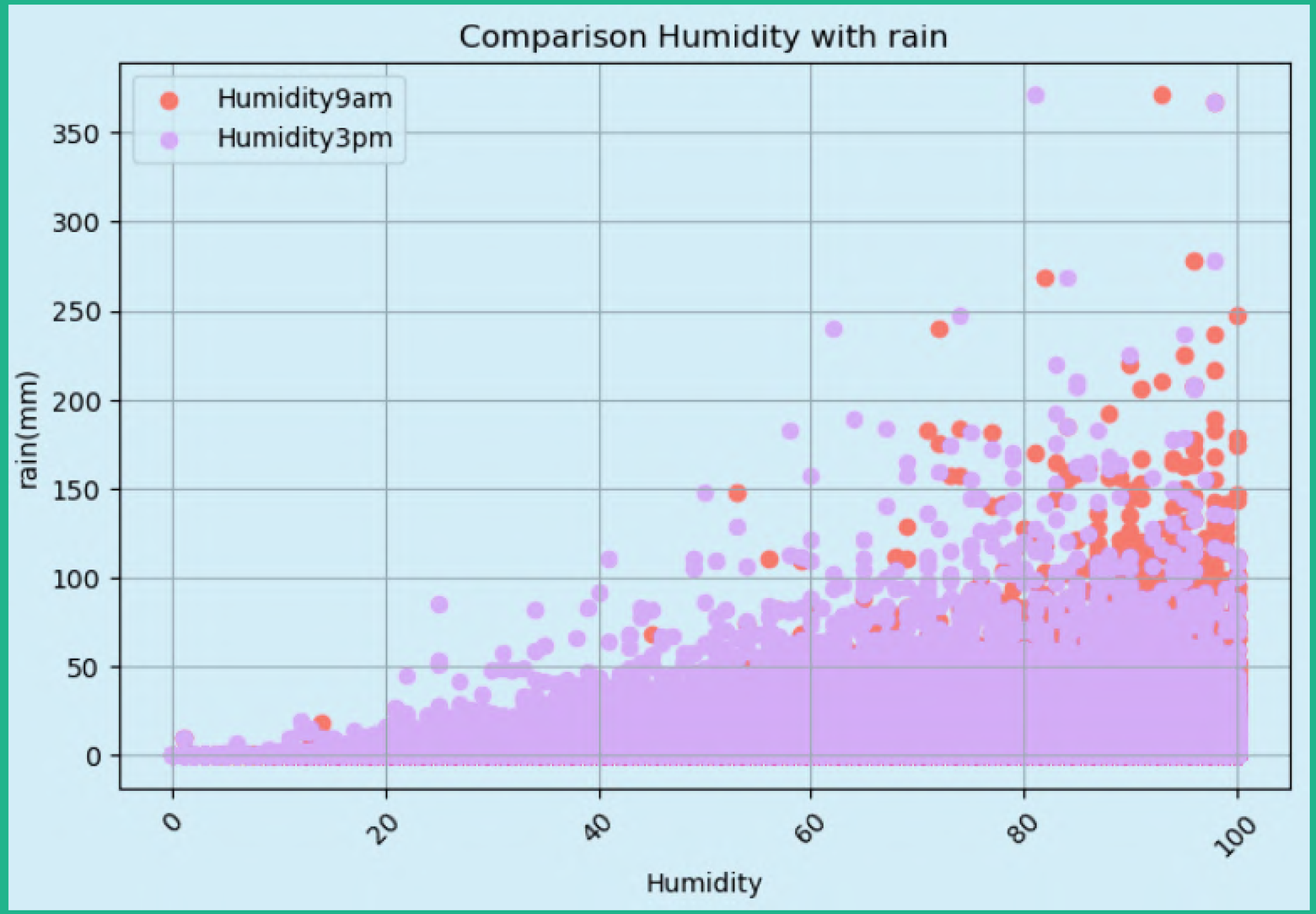
Comparison Temp With Rainfall

Rainfall distribution by temperature: 10-32˚C optimal,
extremes affect precipitation,
and 38> & <9 Least



Comparison Humidity With Rainfall

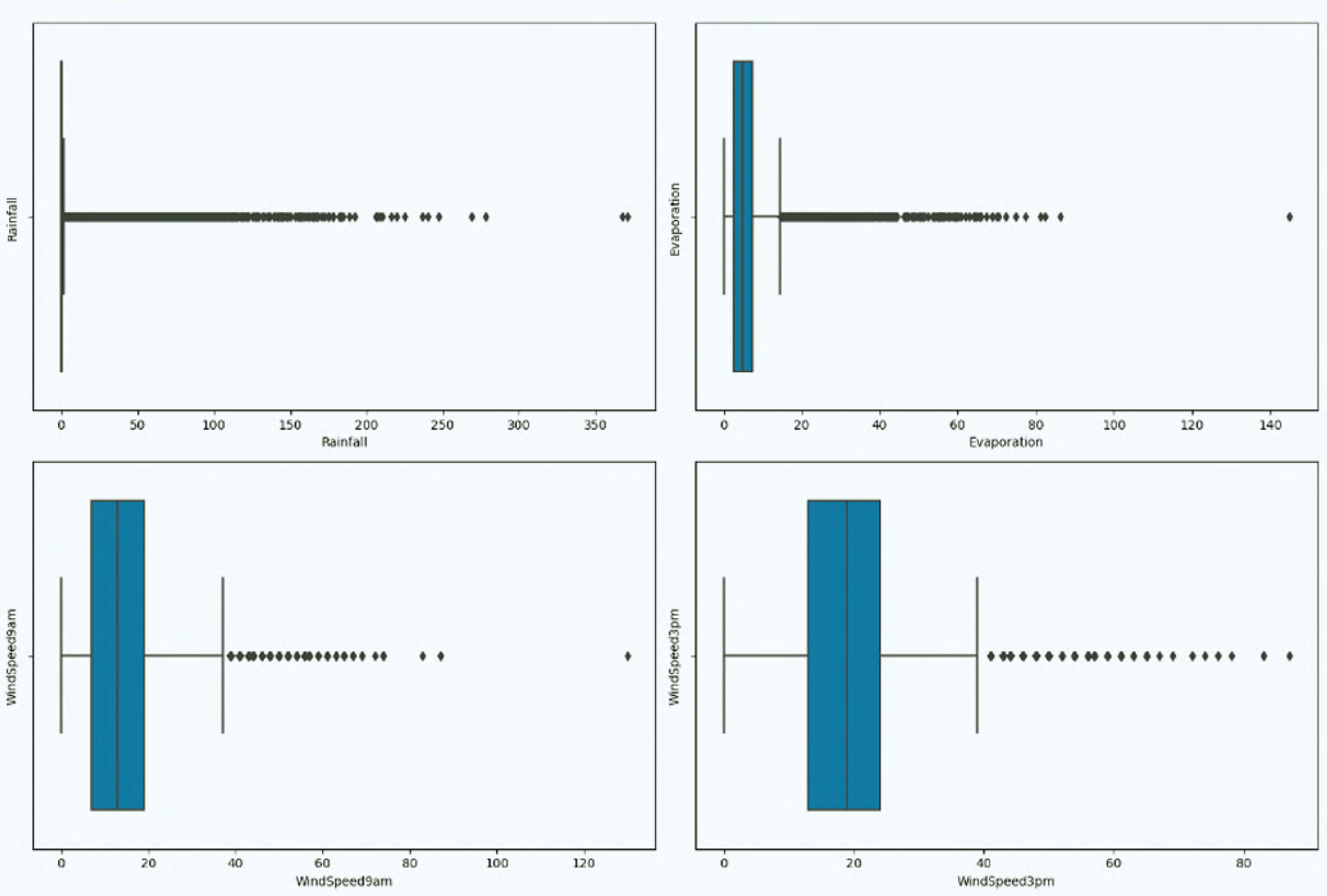
peaks at 70%, drops below 45%. Vital for agriculture, urban planning, managing climate risks, and enhancing resilience.



Outliers Engineering

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm
count	143975.000000	144199.000000	142199.000000	82670.000000	75625.000000	135197.000000	143693.000000	142398.000000
mean	12.194034	23.221348	2.360918	5.468232	7.611178	40.035230	14.043426	18.662657
std	6.398495	7.119049	8.478060	4.193704	3.785483	13.607062	8.915375	8.809800
min	-8.500000	-4.800000	0.000000	0.000000	0.000000	6.000000	0.000000	0.000000
25%	7.600000	17.900000	0.000000	2.600000	4.800000	31.000000	7.000000	13.000000
50%	12.000000	22.600000	0.000000	4.800000	8.400000	39.000000	13.000000	19.000000
75%	16.900000	28.200000	0.800000	7.400000	10.600000	48.000000	19.000000	24.000000
max	33.900000	48.100000	371.000000	145.000000	14.500000	135.000000	130.000000	87.000000

With closer look, we can see Rainfall, Evaporation, WindSpeed9am, and WindSpeed3pm columns may contain outliers. This suspicion arises from the significant disparity between their 75th percentile values and their respective maximum values. To visualize and identify these outliers effectively, I will draw boxplots for the aforementioned variables.

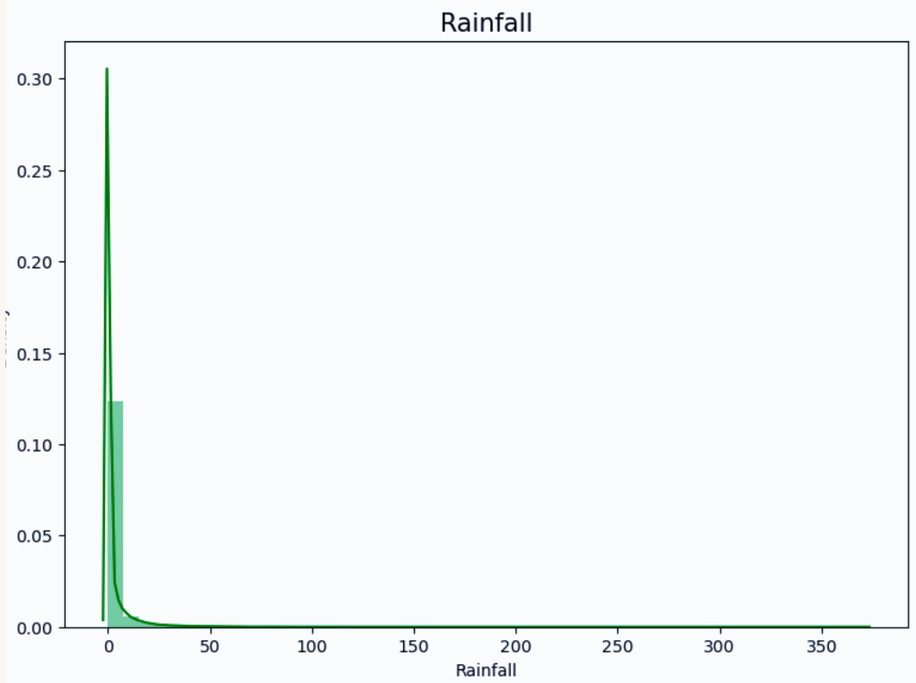

Rainfall
0
50
100
150
200
250
300
350
Rainfall
Evaporation
0
20
40
60
80
100
120
140
Evaporation
WindSpeed9am
0
20
40
60
80
100
120
WindSpeed9am
WindSpeed3pm
0
20
40
60
80
WindSpeed3pm

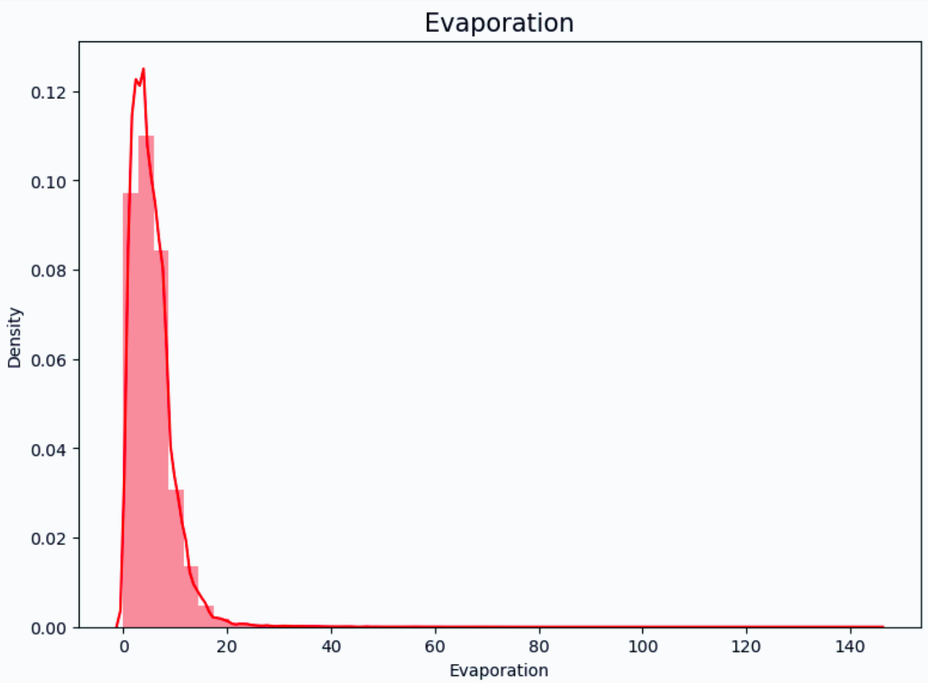
(IQR) method

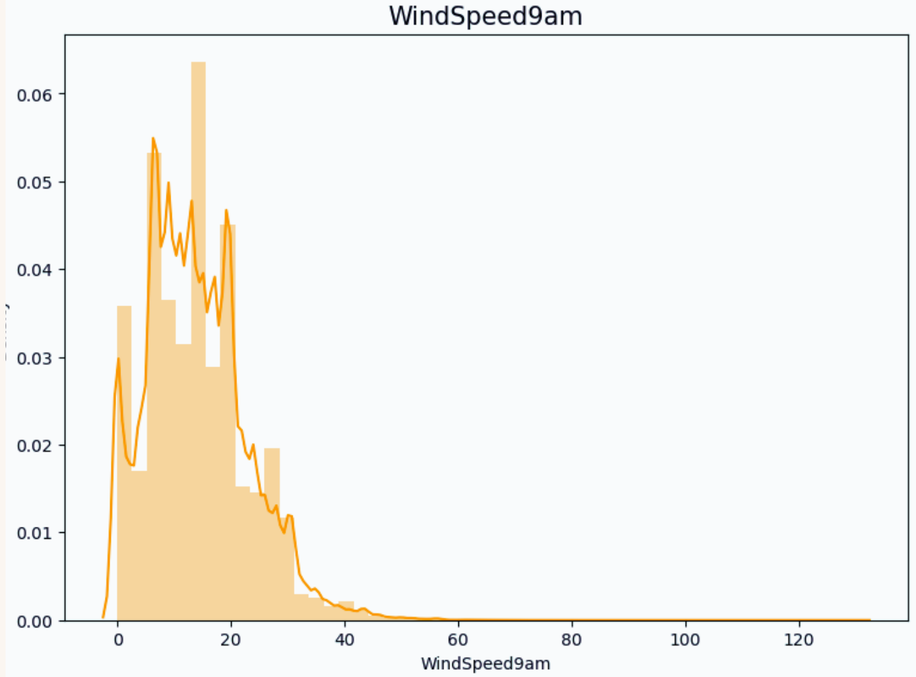
Lower Fence

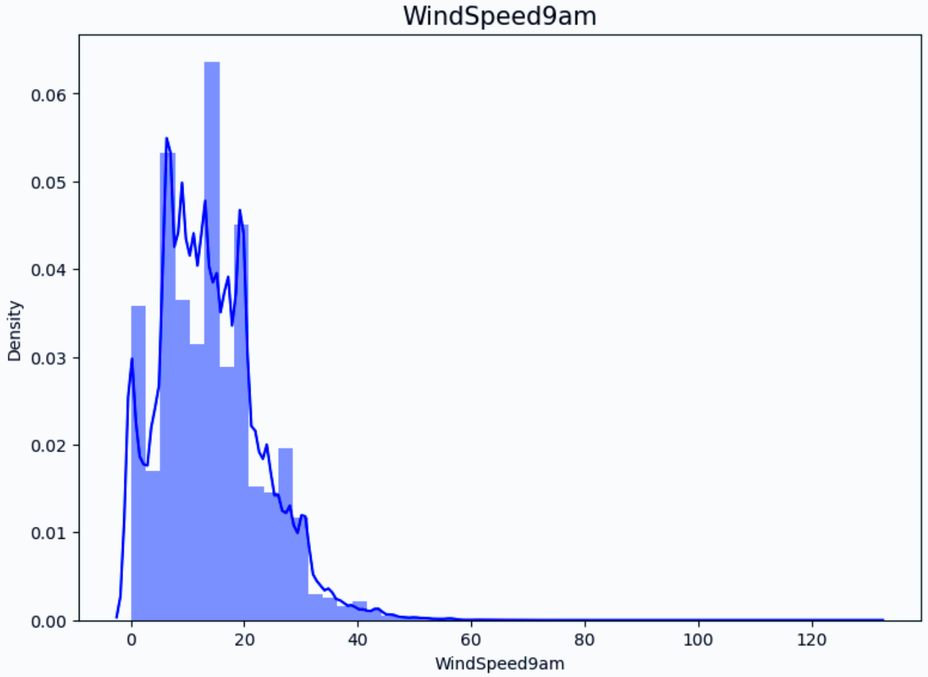
Upper Fence

Top-Coding



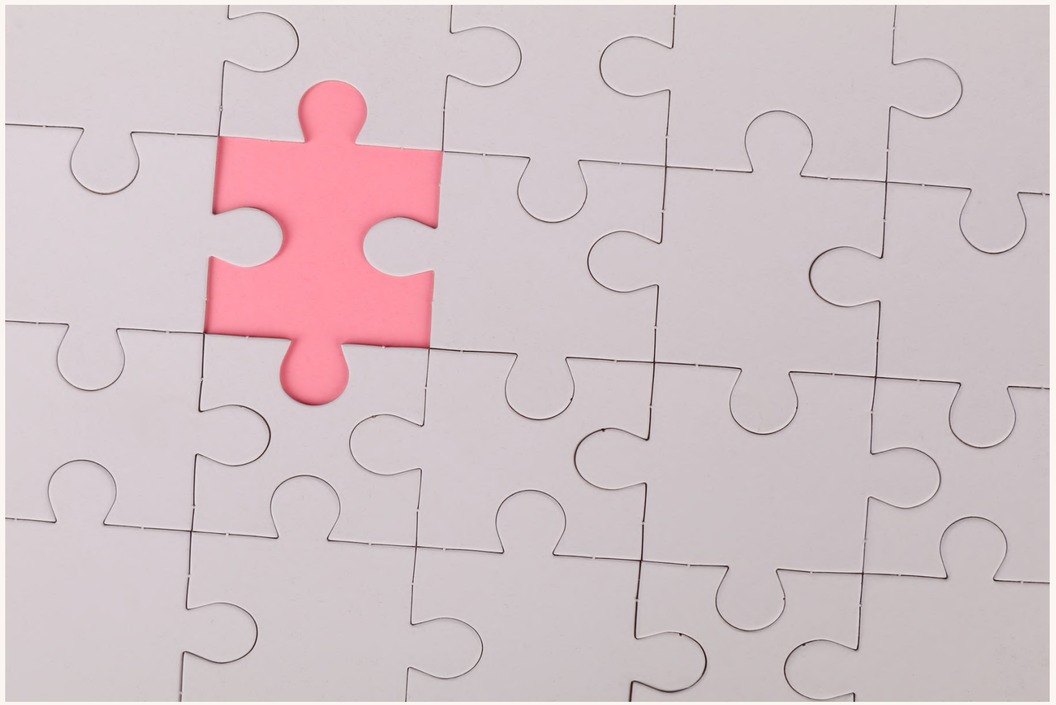






Mode for Categorical

Median for Numerical



Column Name	Data Type	Null-Values
Location	object	0.00%
MinTemp	float64	1.02%
MaxTemp	float64	0.87%
Rainfall	float64	2.24%
Evaporation	float64	43.17%
Sunshine	float64	48.01%
WindGustDir	object	7.10%
WindGust Speed	float64	7.06%
Wind Dir9am	object	7.26%
Wind Dir3pm	object	2.91%
Wind Speed9am	float64	1.21%
Wind Speed3pm	float64	2.11%
Humidity9am	float64	1.82%
Humidity 3pm	float64	3.10%
Pressure9am	float64	10.36%
Pressure3pm	float64	10.33%
Cloud9am	float64	38.42%
Cloud3pm	float64	40.81%
Temp9am	float64	1.21%
Temp3pm	float64	2.48%
Rain Today	object	2.24%
RainTomorrow	object	2.25%

One-Hot Encoding

Original

RainToday
Yes
No
Yes
No

Encoded

RainToday Yes	RainToday No
1	0
0	1
1	0
0	1

Scaling

Original

RainFall
50
36
24
42

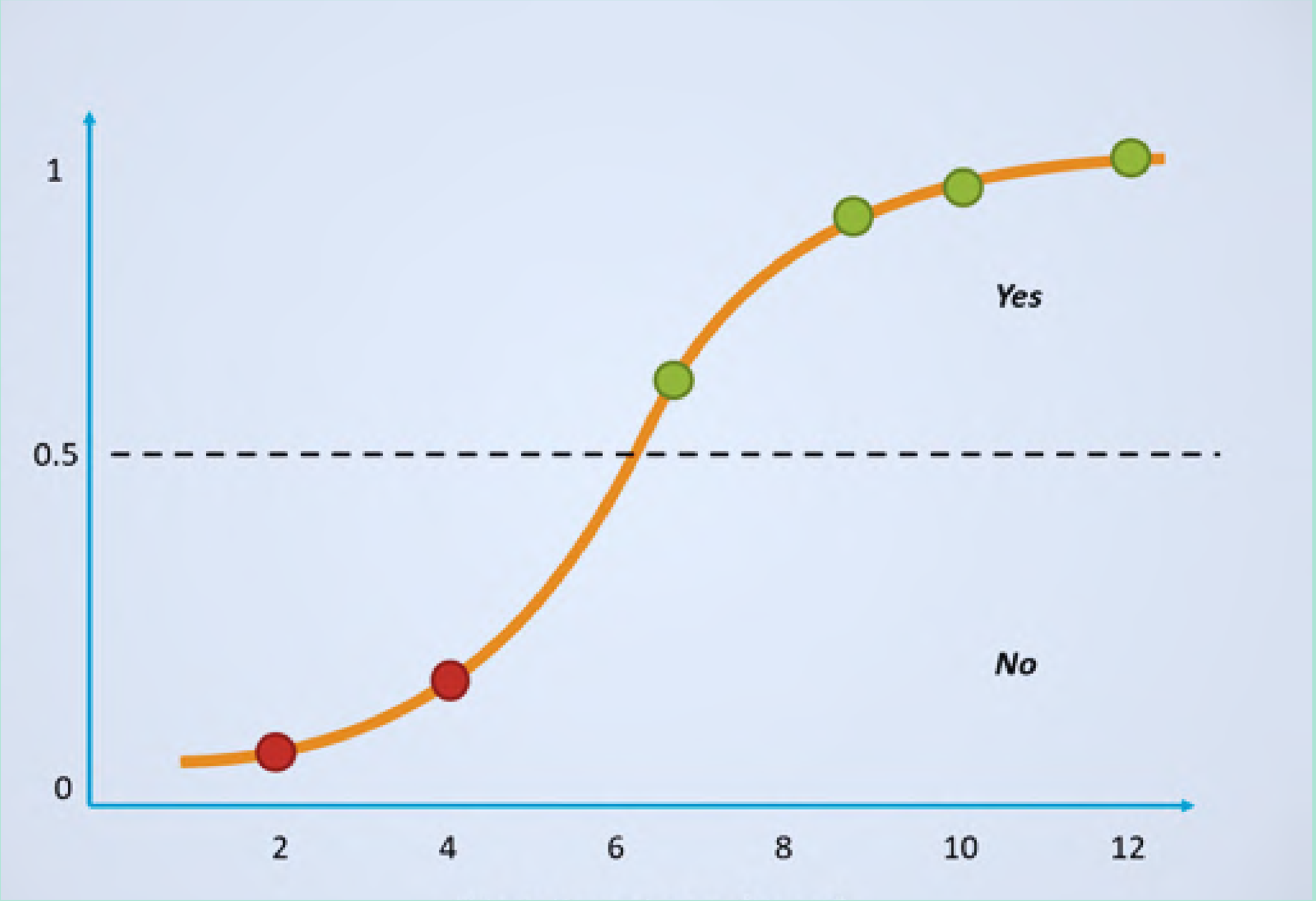
Divide all by 50, which is the maximum number of column to scale it.

Scaled

RainFall
1
0.72
0.48
0.84

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Logistic Regression

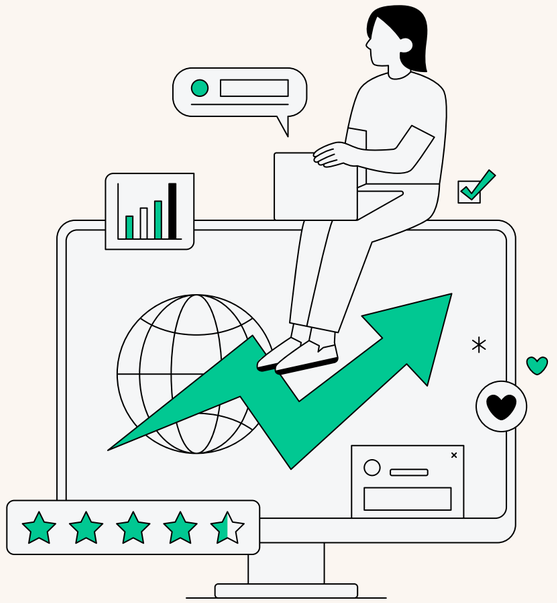


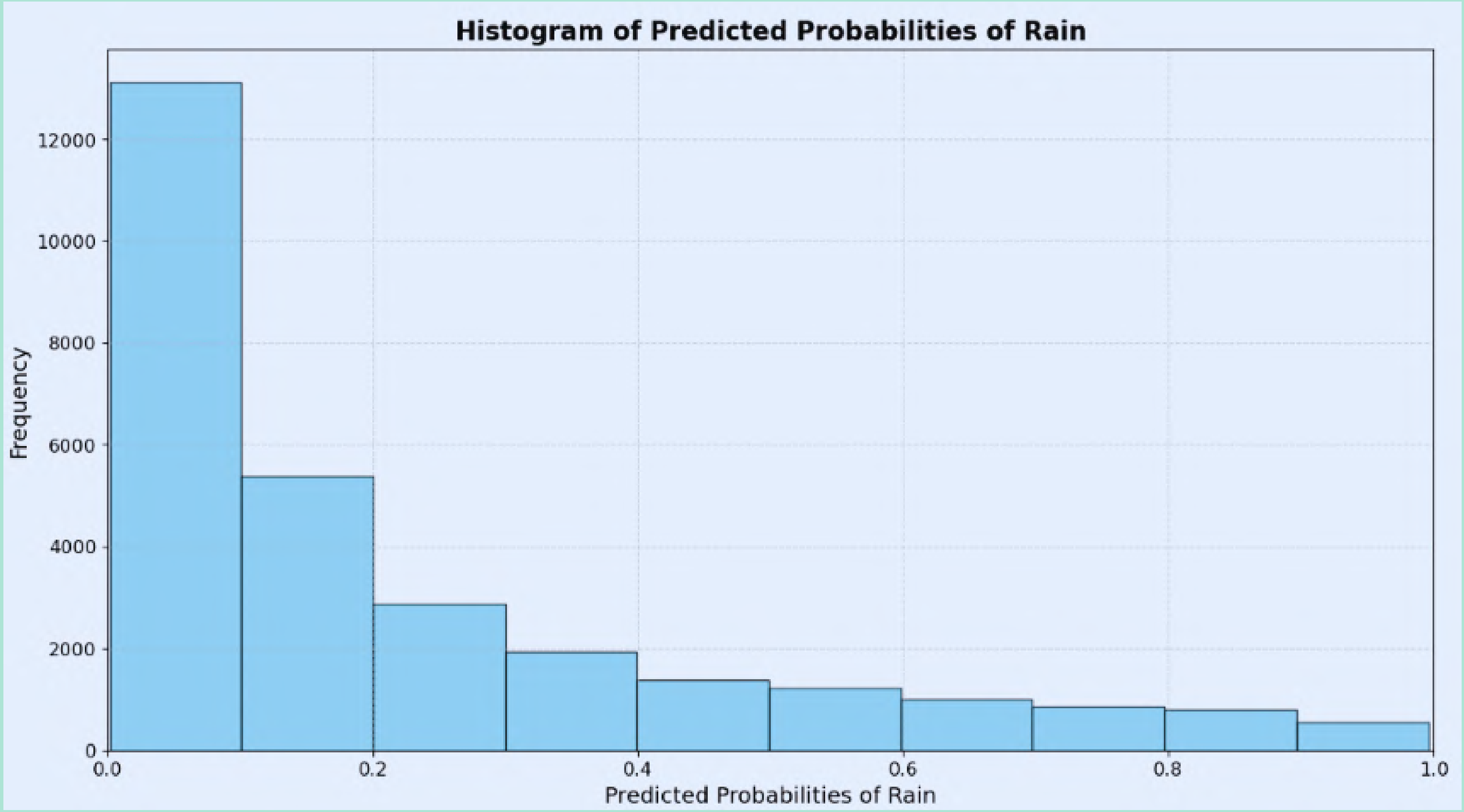
Probability of Rain

	Prob of - No rain tomorrow (0)	Prob of - Rain tomorrow (1)
0	0.749256	0.250744
1	0.833476	0.166524
2	0.827682	0.172318
3	0.624588	0.375412
4	0.884235	0.115765
5	0.974373	0.025627
6	0.756804	0.243196
7	0.292372	0.707628
8	0.819927	0.180073
9	0.690680	0.309320

Model accuracy Score:

0.8388




Histogram of Predicted Probabilities of Rain
Frequency
12000
10000
8000
6000
4000
2000
0
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Probabilities of Rain

Overfitting and Underfitting

Training-Set Score: 0.8401

Testing-Set Score: 0.8388

K-Fold Cross validation

Cross-validated scores: [0.8385 0.8392 0.8433 0.8360 0.8394]

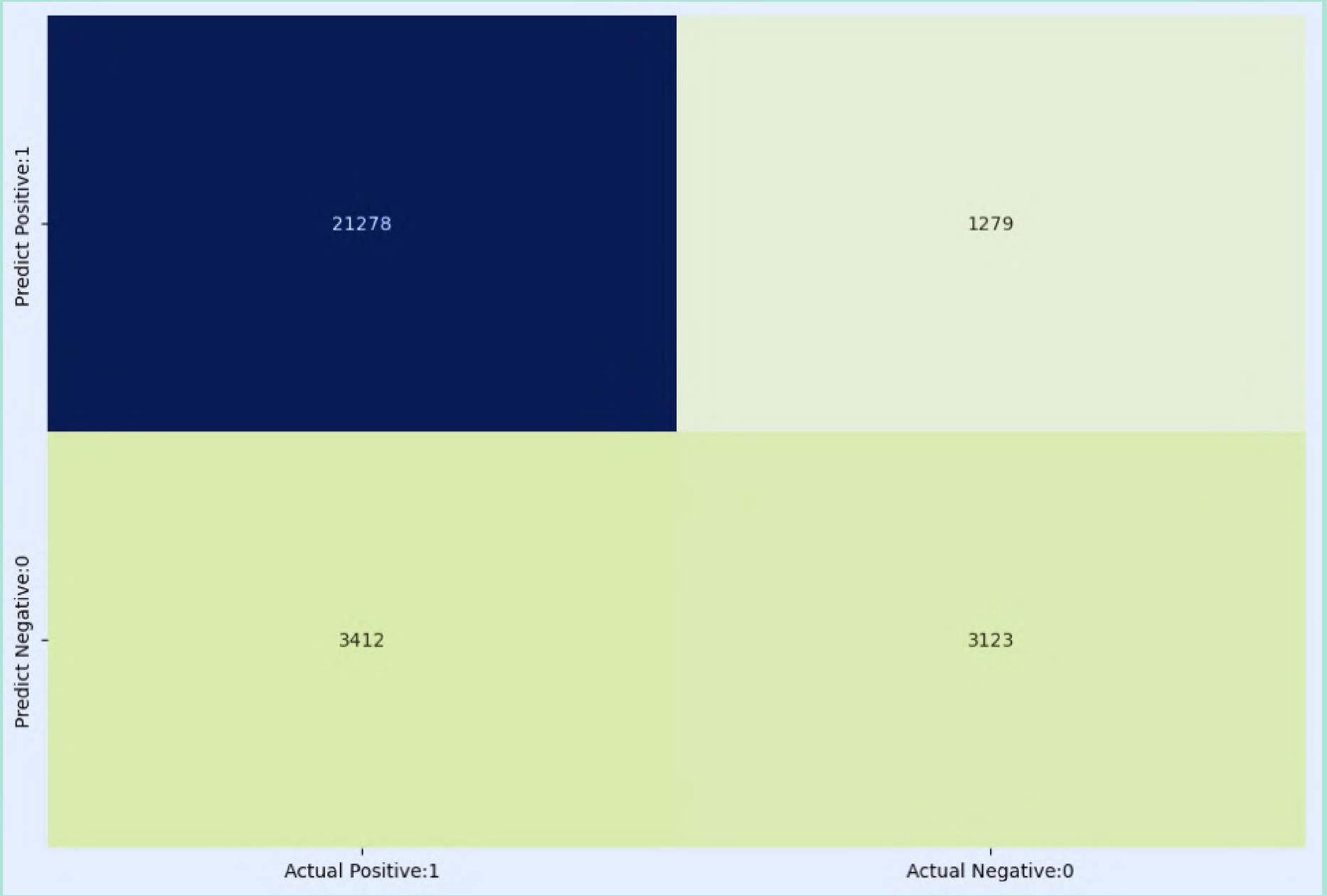
Mean accuracy: 0.8393200610449659

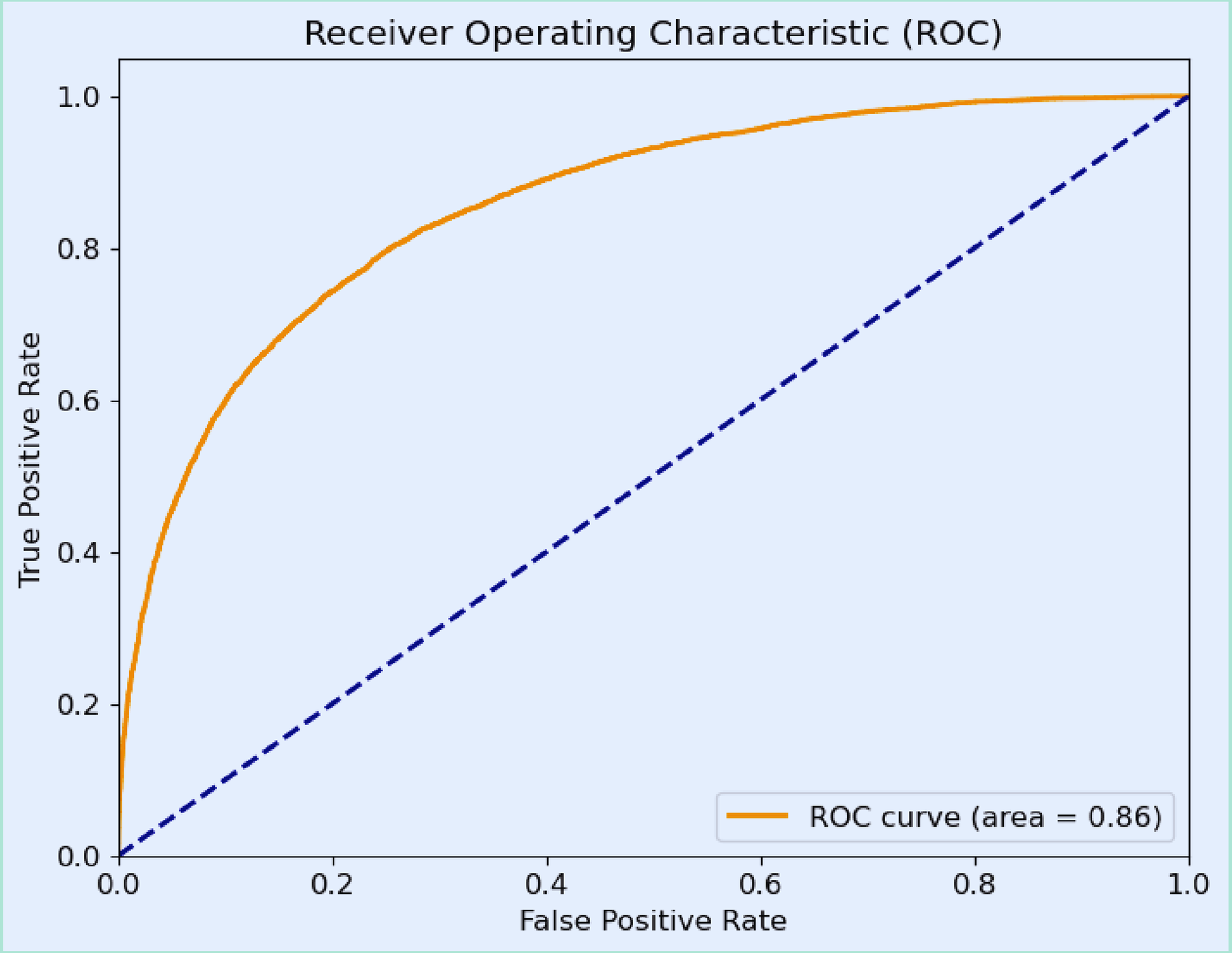
Precision : 0.9433

Recall : 0.8618

Specificity : 0.7095

F1-score : 0.9007




Receiver Operating Characteristic (ROC)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.86)

Presented by Mohammad Ehsani

Thank you very much!

Australia Weather Forecast

Instructor: ROBERTA SICILIANO