Precipitation Prediction in Australia

Final Project

Elizabeth Frank

Project Objective

Stay ahead of the weather.
Develop a predictive model, to empower you to plan your day with confidence, by leveraging nearly a decade's worth of daily climate data from across Australia.

Dataset Overview

A comprehensive collection of climate data from the Australian Government Bureau of Meteorology

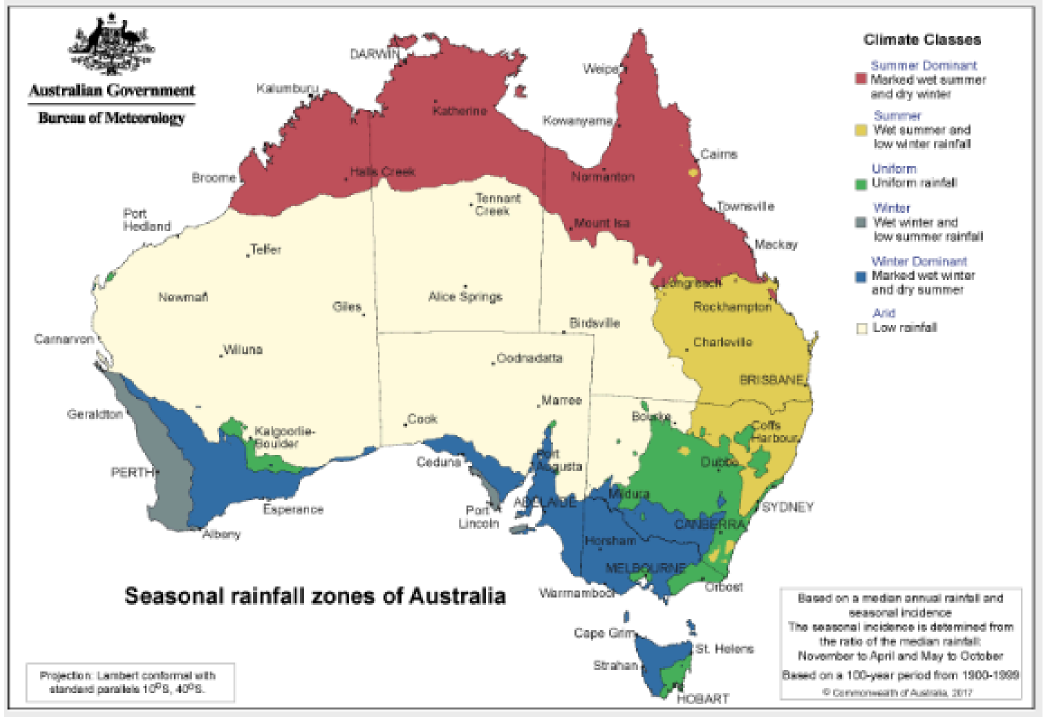
Comprised of 145,460 observations from 49 locations across 6 climate and rainfall regions, with 25 distinct attributes

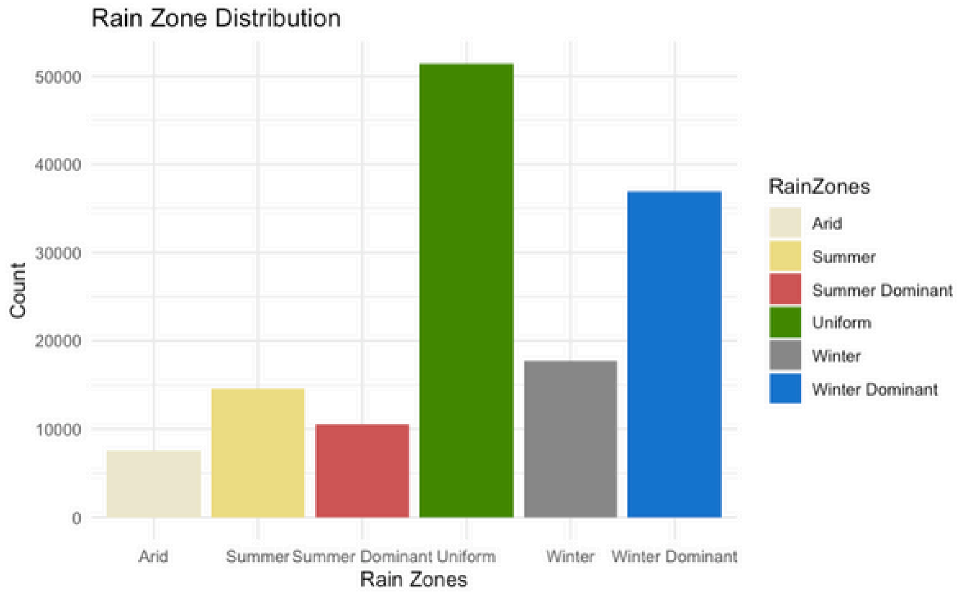
Key Attributes include:

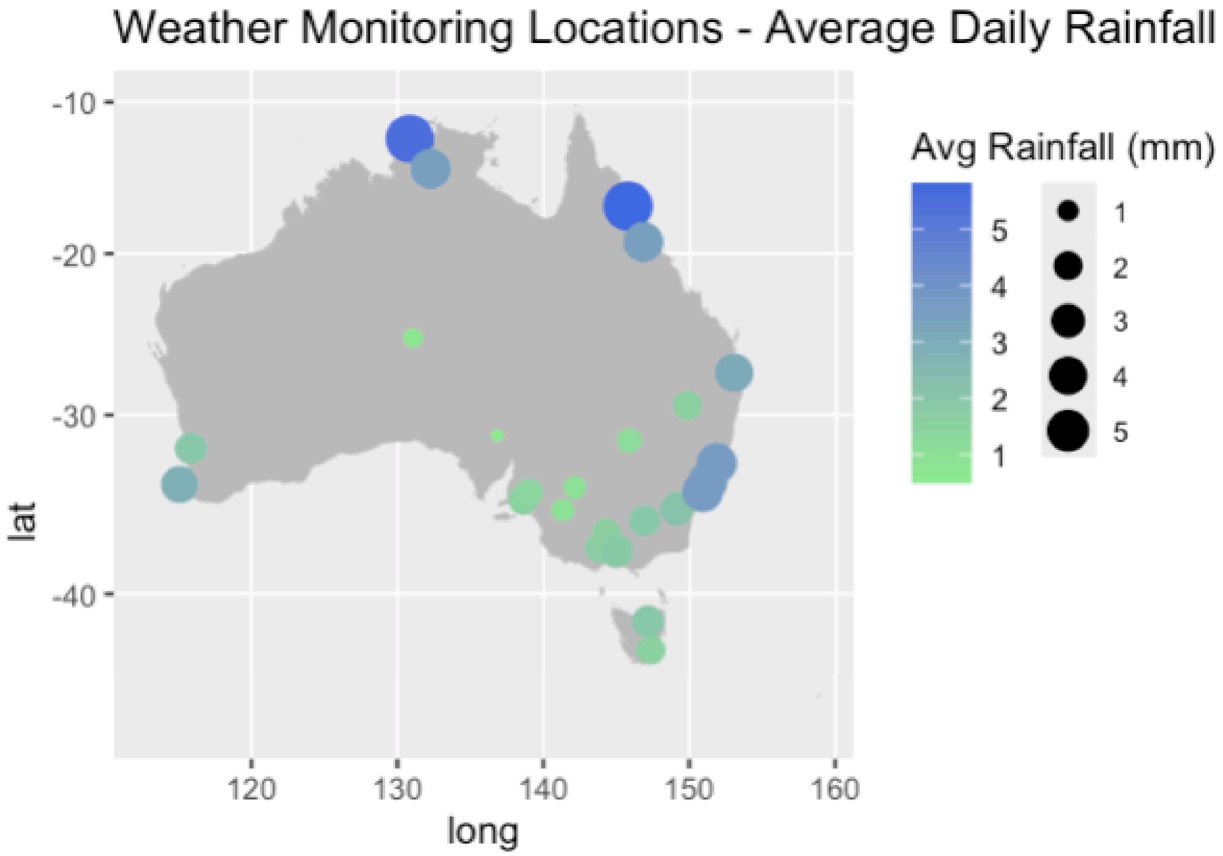
- Temperature
- Rainfall
- Evaporation
- Sunshine
- Wind Gusts and Speed
- Humidity
- Pressure
- Cloud Cover

- **Dependent variable**: Will it rain tomorrow? Yes or No?

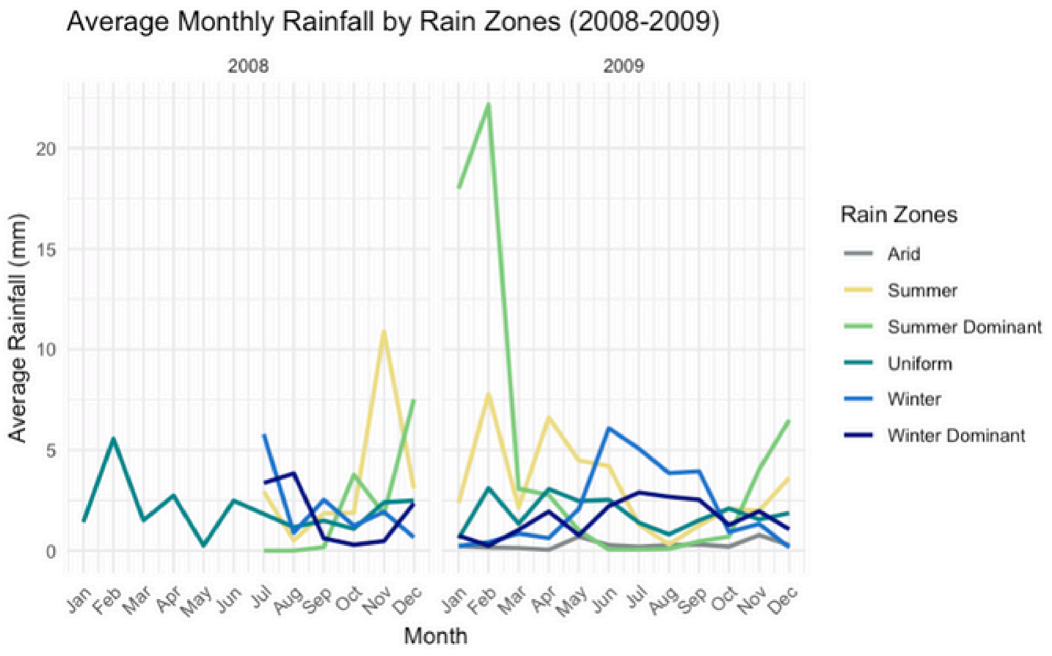
Data Summary







Data Preparation



Missing Values:

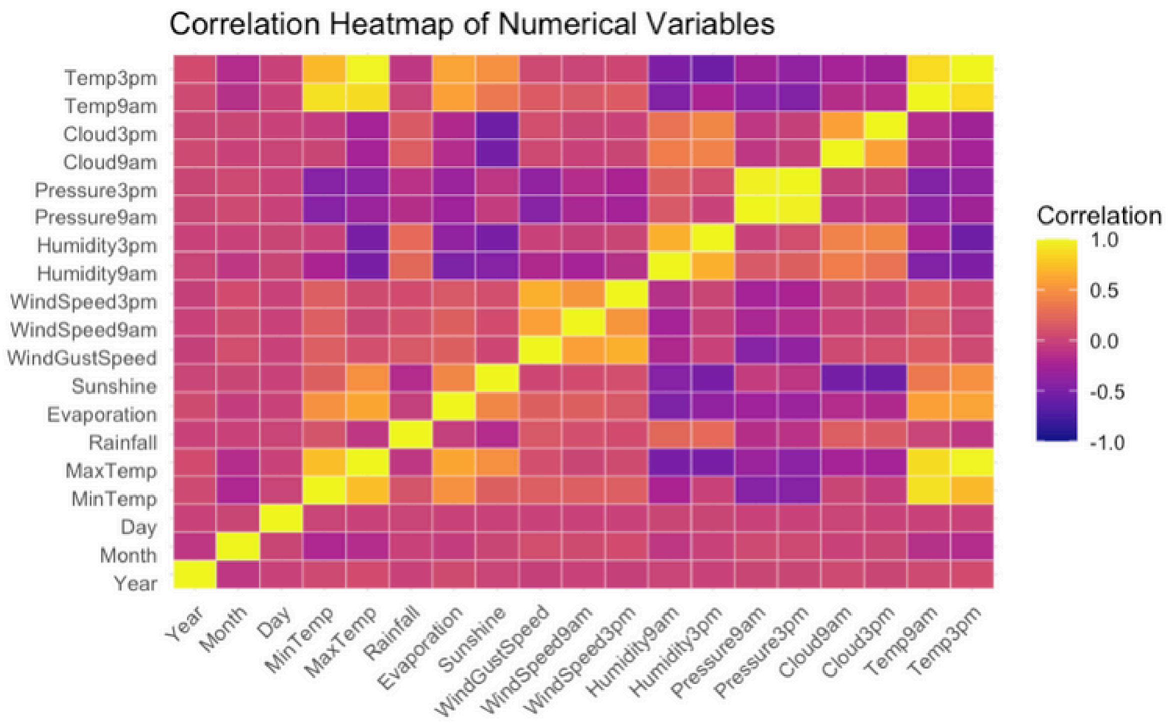
- 9.4% of data missing
- Impute using mean for numerical variables and mode for categorical
- Weather station online differences

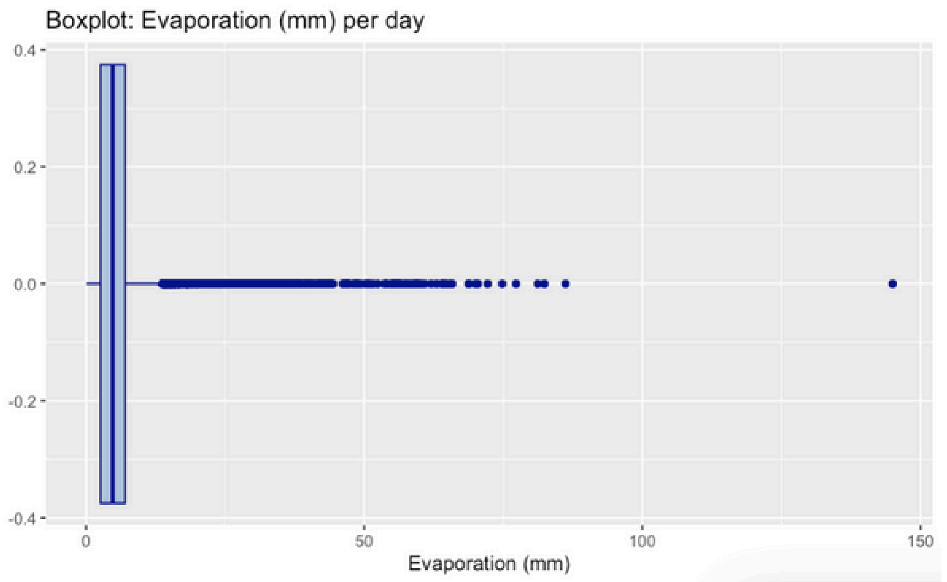
Outliers:

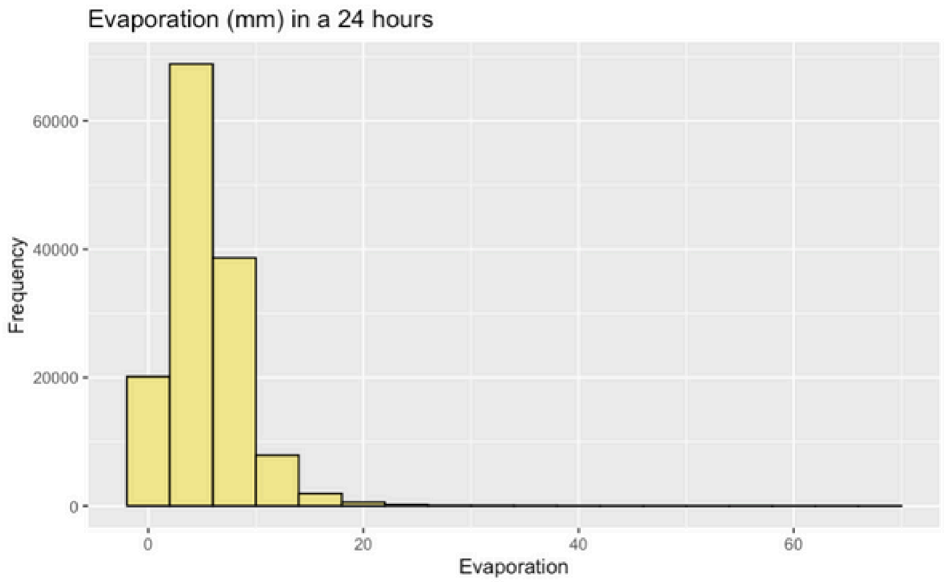
- Rainfall and Evaporation

Feature Engineering:

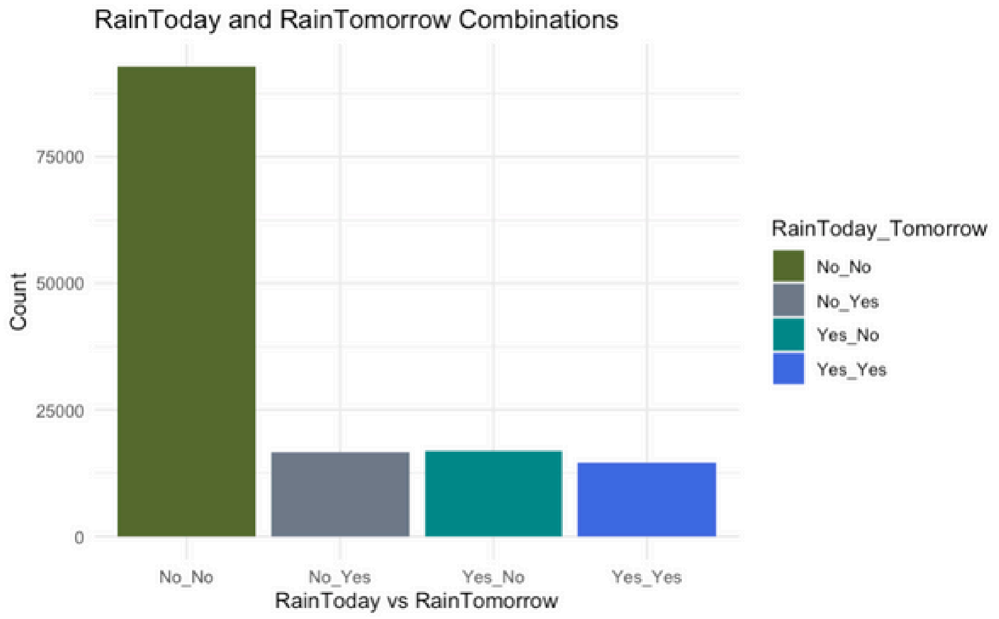
- Collinearity of variables

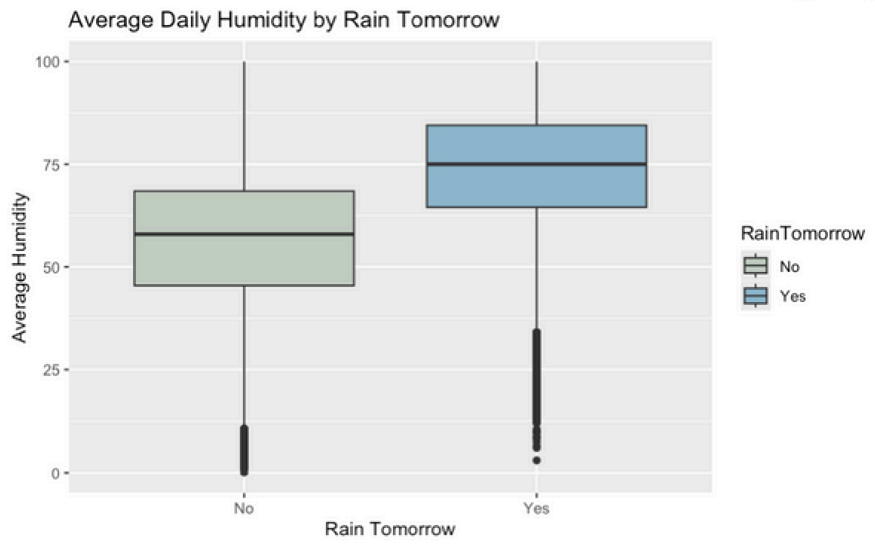


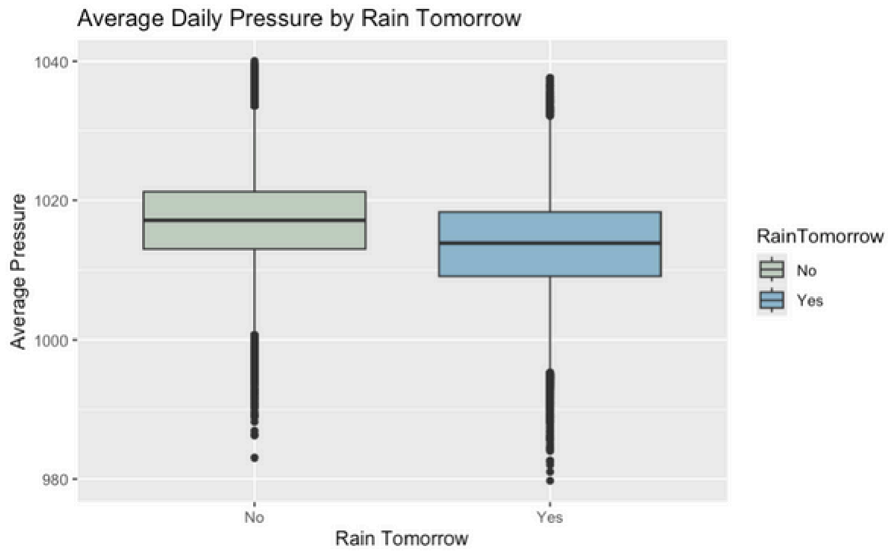


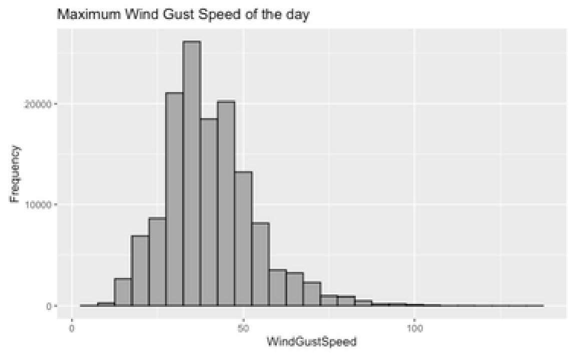


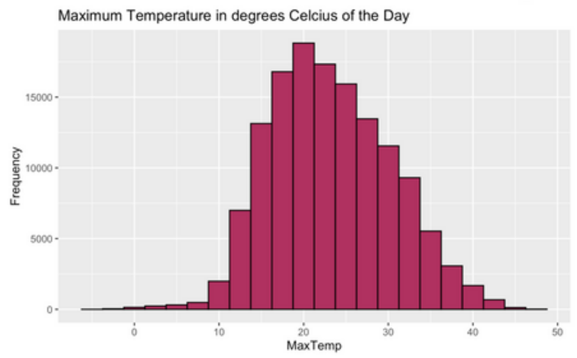
Exploratory Data Analysis

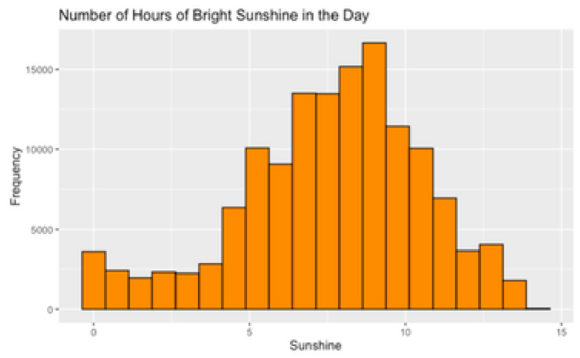

RainToday and RainTomorrow Combinations

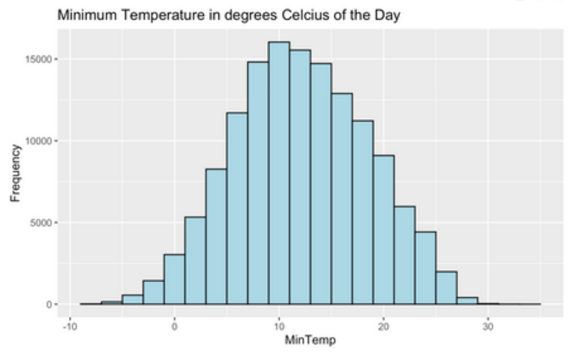

Average Daily Humidity by Rain Tomorrow

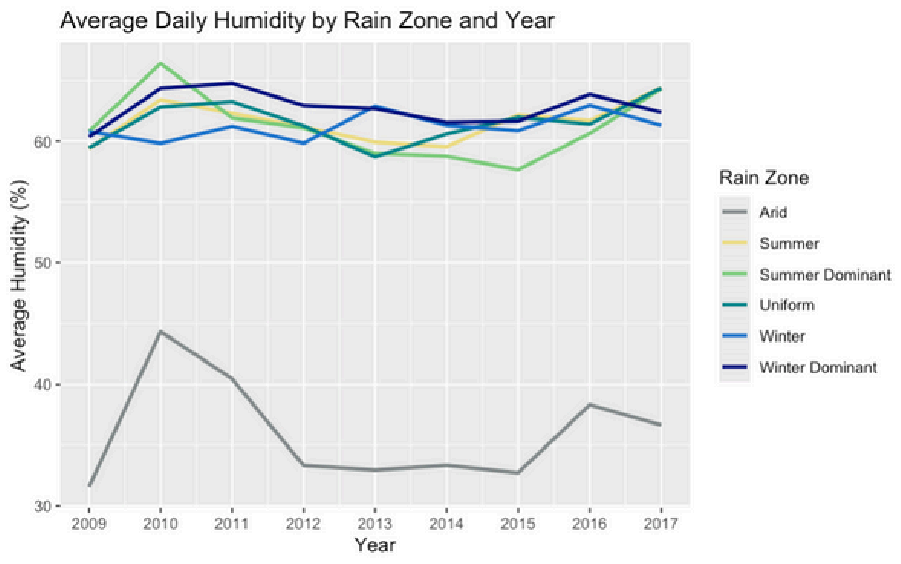

Average Daily Pressure by Rain Tomorrow

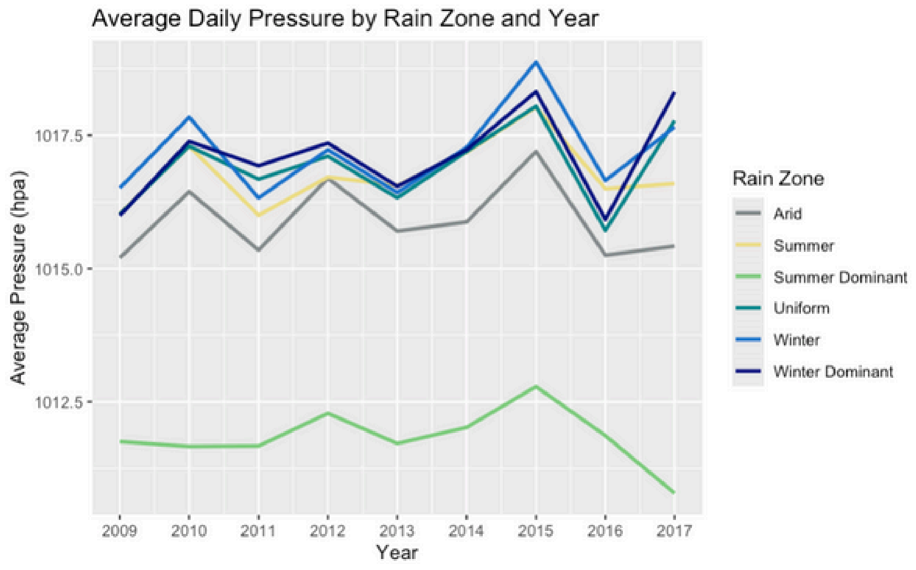

Maximum Wind Gust Speed of the day


Maximum Temperature in degrees Celcius of the Day


Number of Hours of Bright Sunshine in the Day


Minimum Temperature in degrees Celcius of the Day


Average Daily Humidity by Rain Zone and Year


Average Daily Pressure by Rain Zone and Year

Model Selection

Support Vector Machine (SVM):

- Very effective for high dimensional data and both linear and non-linear relationships
- Low interpretability.

SVM with K-fold Cross Validation (SVM with Kfold):

- Hyperparameter tuning by taking average of 10 splits of data and testing on the remainder for best C and sigma values.
- Computationally expensive.

Recursive Partitioning and Regression Trees (Rpart):

- Splits data based on most significant predictors at each node.
- Doesn't handle non-linear data well.

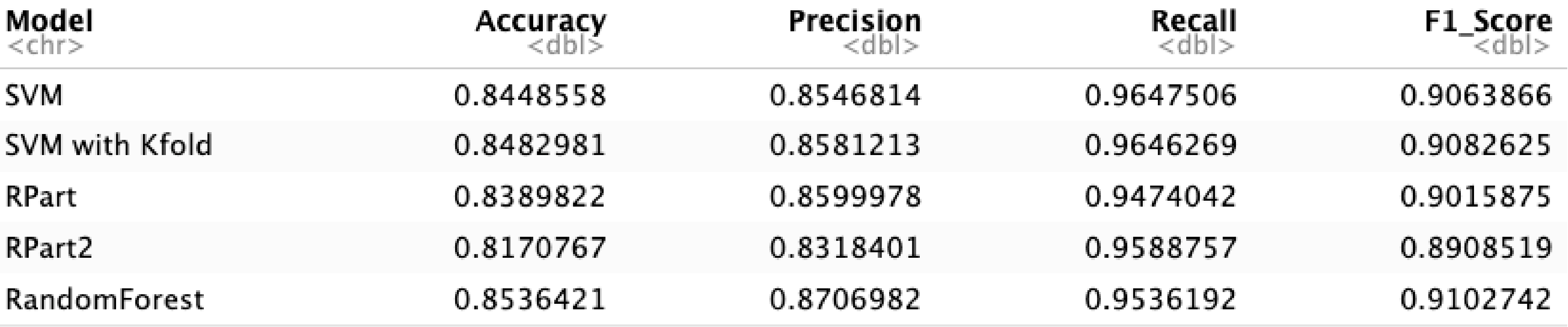
Rpart Simplified (Rpart2):

- Created with the most important variables from Rpart for maximum interpretability.
- Usability is at a cost of performance

Random Forest:

- An ensemble method that uses bagging for multiple decision trees and averages for increased stability and more accurate predictions.
- Low interpretability.

Model Performance



Model <chr>	Accuracy <dbl>	Precision <dbl>	Recall <dbl>	F1_Score <dbl>
SVM	0.8448558	0.8546814	0.9647506	0.9063866
SVM with Kfold	0.8482981	0.8581213	0.9646269	0.9082625
RPart	0.8389822	0.8599978	0.9474042	0.9015875
RPart2	0.8170767	0.8318401	0.9588757	0.8908519
RandomForest	0.8536421	0.8706982	0.9536192	0.9102742

Accuracy: predicts correct rain and no rain most of the time.

- Random Forest has the highest accuracy at 85.4%, followed by SVM with Kfold at 84.8% and SVM at 84.5%.

Precision: predicts rain when it doesn't actually happen.

- Random Forest has the highest precision at 87.1%, followed closely by Rpart 86.0% and SVM with Kfold at 85.8%

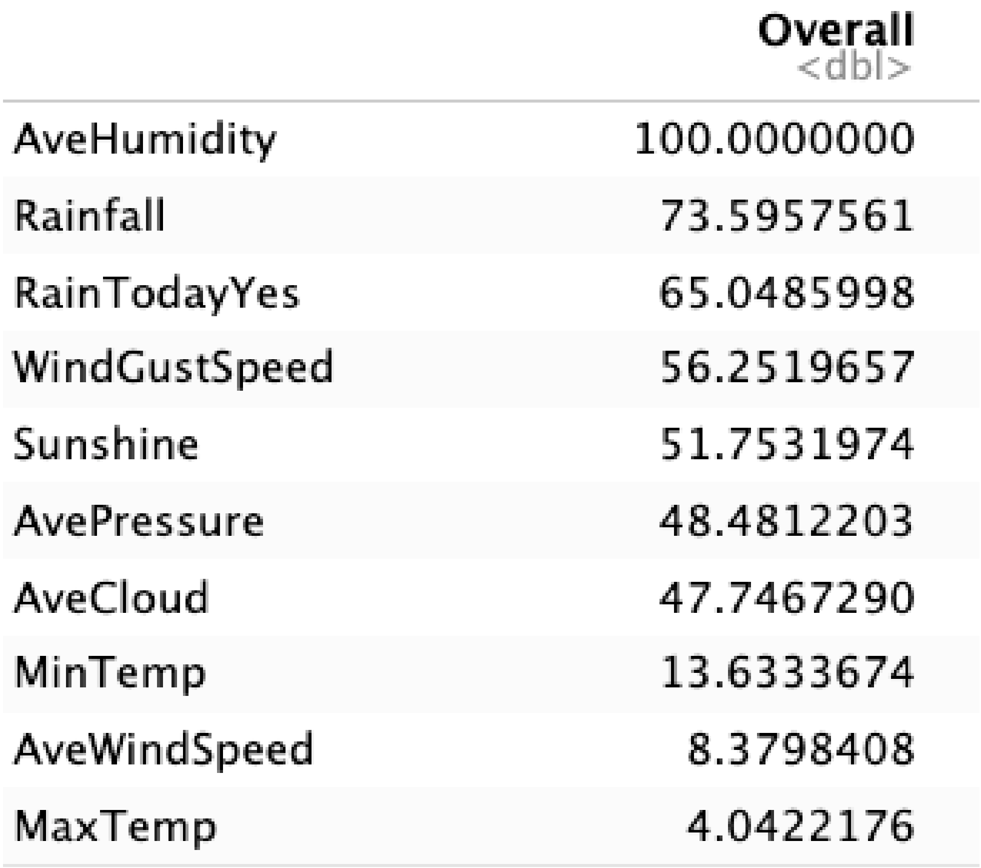
Recall: catchs all rain events at the cost of incorrectly indicating rain when it doesn't actually rain.

- SVM has the highest recall at 96.48%, but nearly same is SVM with Kfold at 96.46%.

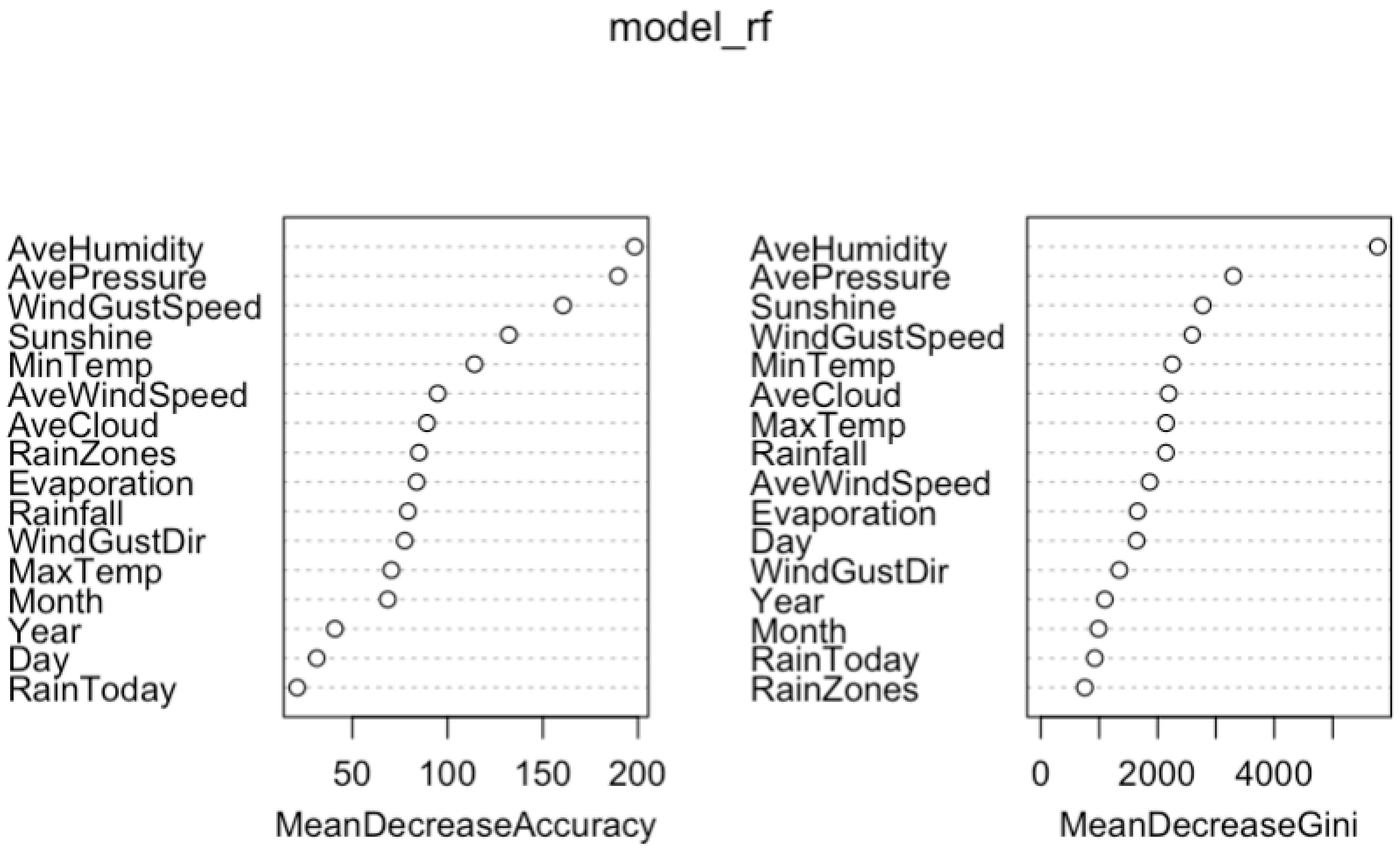
F1-Score: balance of precision and recall, minimizes both false positives and negatives and has higher power predicting rain events.

- Random Forest again yields the highest F1-score at 91.0%, followed by SVM with Kfold at 90.8% and SVM at 90.7%.

Variable Importance



	Overall <dbl>
AveHumidity	100.0000000
Rainfall	73.5957561
RainTodayYes	65.0485998
WindGustSpeed	56.2519657
Sunshine	51.7531974
AvePressure	48.4812203
AveCloud	47.7467290
MinTemp	13.6333674
AveWindSpeed	8.3798408
MaxTemp	4.0422176

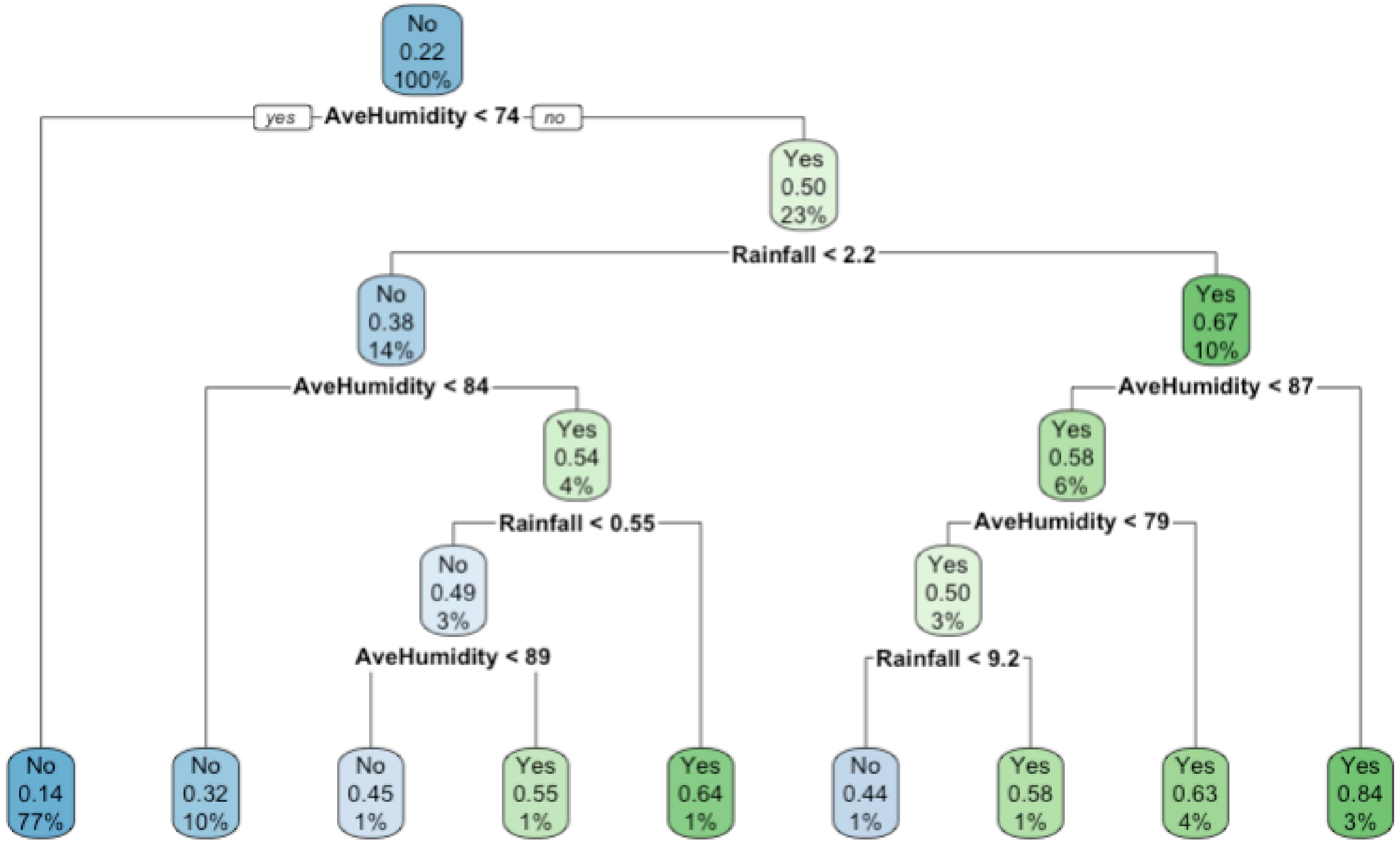


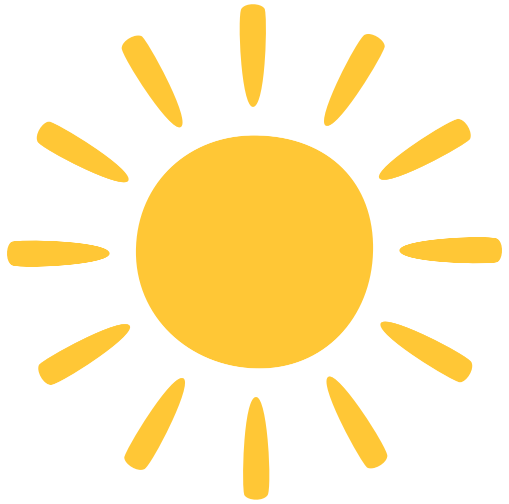
Recursive Partitioning and Regression Trees (Rpart)

Random Forest

Model Deployment

- If average **humidity** is **< 74%** the model predicts no rain with a probability of 22%
- If average **humidity** is **> 74%** the model moves on to rainfall.
  - If **rainfall** is **< 2.2 mm** in the last 24 hours, the model predicts no rain at 38%
  - The average **humidity** is **> 84%** and **rainfall** is **< .55 mm** and the model predicts rain tomorrow with 54% probability
- If the **rainfall** is **> 2.2 mm**, the model predicts rain tomorrow with a probability of 50%
  - And the average **humidity** is **> 87%** the model predicts rain tomorrow at an 84% probability
- If the **rainfall** is **< 9.2 mm**, and the average **humidity** is **< 79%** there is a 44% percent probability of no rain tomorrow.





Insights

- What is the average humidity today?
- Has it rained and how much in the last 24 hours?
- Has there been a drop in atmospheric pressure ?

Thank you!