

# Forecasting Rain in Australia

*Silvia Elaluf-Calderwood*

*1/7/2020*

## Introduction

The dataset chosen for this project is *Rain in Australia*, which is a publicly available dataset found on the Kaggle website (see <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>). The dataset contains daily measurements of various weather variables from weather stations at a number of locations throughout Australia, over a period of several years.

The aim of the project is to fit a machine learning binary classification model to the data that will predict whether or not it will rain on the following day.

The approach adopted was to fit several different classes of model to the data and select the one that gave the best performance. Accuracy was used as the measure of performance.

The models chosen were:

- Model 1: Generalized Linear Model
- Model 2: Linear Discriminant Analysis
- Model 3: Quadratic Discriminant Analysis
- Model 4: Multi-Layer Perceptron
- Model 5: k-Nearest Neighbors
- Model 6: Random Forest

The dataset was cleaned and then split into training and test sets, with the training set used to fit the models and the test set used to assess their performance.

While all of the models chosen gave similar performance, the Random Forest model performed best, with an accuracy of 0.86. Although this appears to be a good result, the model has a sensitivity of only 0.50, which means that it results in as many false negatives as true positives. So if the goal of the model is to correctly predict days on which it will rain (as opposed to whether or not it will rain), the model performs about the same as tossing a coin.

## Method

The weather data was downloaded from the Kaggle site in CSV format and saved locally to disk. The CSV file was then loaded into a data frame in memory.

## Initial Exploratory Analysis

The structure and dimensions of the data frame are shown below. This shows that the data contains *RainTomorrow*, the variable to be predicted by the model, along with other variables that can potentially be used as inputs to the model. It also shows that the data is a mixture of numeric and character variables, and that there are some missing values.

```
## 'data.frame':   145463 obs. of  24 variables:
## $ Date          : chr  "2008-12-01" "2008-12-02" "2008-12-03" "2008-12-04" ...
## $ Location      : chr  "Albury" "Albury" "Albury" "Albury" ...
## $ MinTemp       : num  13.4  7.4  12.9  9.2  17.5  14.6  14.3  7.7  9.7  13.1 ...
## $ MaxTemp       : num  22.9  25.1  25.7  28  32.3  29.7  25  26.7  31.9  30.1 ...
## $ Rainfall      : num  0.6  0  0  0  1  0.2  0  0  0  1.4 ...
## $ Evaporation   : num  NA NA NA NA NA NA NA NA NA NA NA ...
## $ Sunshine      : num  NA NA NA NA NA NA NA NA NA NA NA ...
```

```

## $ WindGustDir : chr "W" "WNW" "WSW" "NE" ...
## $ WindGustSpeed: int 44 44 46 24 41 56 50 35 80 28 ...
## $ WindDir9am : chr "W" "NNW" "W" "SE" ...
## $ WindDir3pm : chr "WNW" "WSW" "WSW" "E" ...
## $ WindSpeed9am : int 20 4 19 11 7 19 20 6 7 15 ...
## $ WindSpeed3pm : int 24 22 26 9 20 24 24 17 28 11 ...
## $ Humidity9am : int 71 44 38 45 82 55 49 48 42 58 ...
## $ Humidity3pm : int 22 25 30 16 33 23 19 19 9 27 ...
## $ Pressure9am : num 1008 1011 1008 1018 1011 ...
## $ Pressure3pm : num 1007 1008 1009 1013 1006 ...
## $ Cloud9am : int 8 NA NA NA 7 NA 1 NA NA NA ...
## $ Cloud3pm : int NA NA 2 NA 8 NA NA NA NA NA ...
## $ Temp9am : num 16.9 17.2 21 18.1 17.8 20.6 18.1 16.3 18.3 20.1 ...
## $ Temp3pm : num 21.8 24.3 23.2 26.5 29.7 28.9 24.6 25.5 30.2 28.2 ...
## $ RainToday : chr "No" "No" "No" "No" ...
## $ RISK_MM : num 0 0 0 1 0.2 0 0 0 1.4 0 ...
## $ RainTomorrow : chr "No" "No" "No" "No" ...

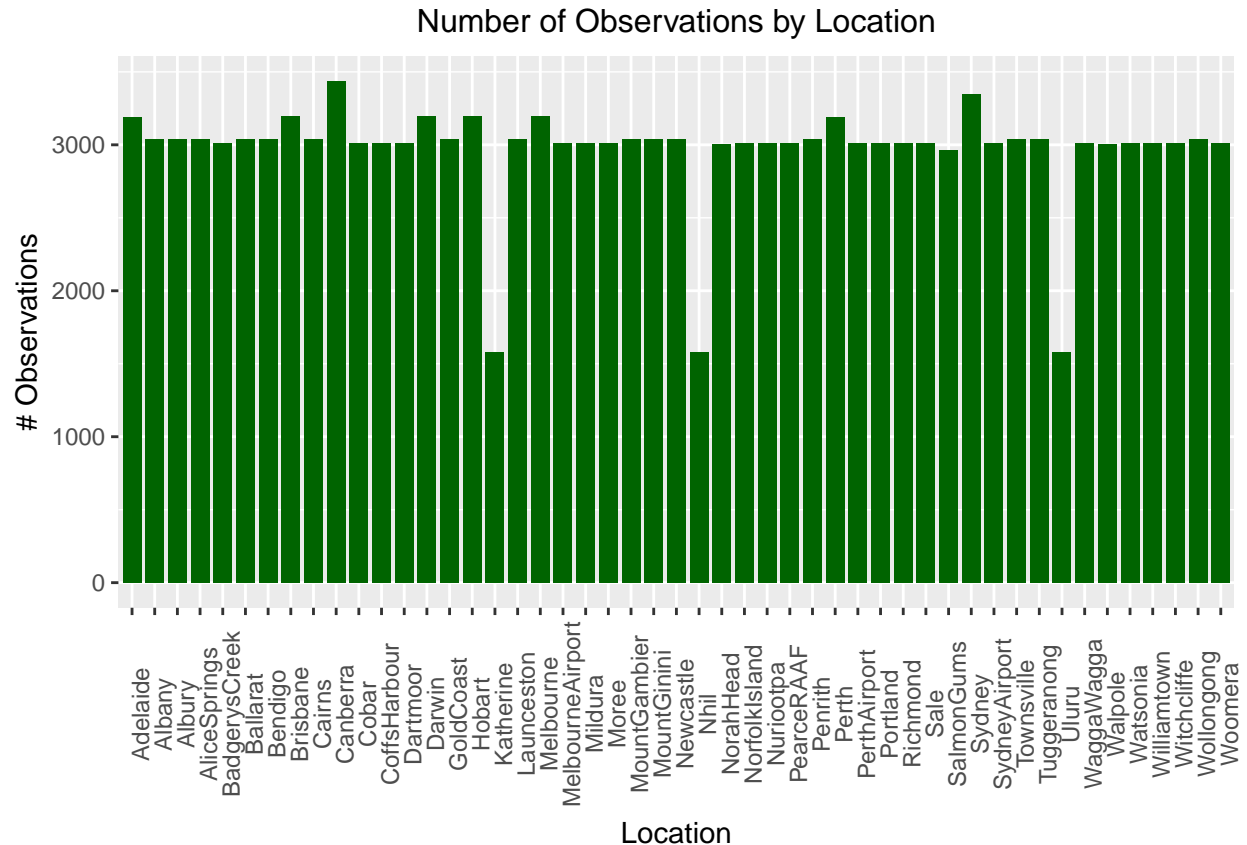
```

The locations at which measurements were recorded and the number of observations at each location are shown in the following table and bar chart.

```

##
##      Adelaide      Albany      Albury      AliceSprings
##      3193          3040          3041          3041
##      BadgerysCreek Ballarat      Bendigo      Brisbane
##      3010          3041          3041          3194
##      Cairns        Canberra      Cobar        CoffsHarbour
##      3041          3437          3010          3010
##      Dartmoor      Darwin        GoldCoast      Hobart
##      3010          3194          3041          3194
##      Katherine     Launceston     Melbourne MelbourneAirport
##      1579          3041          3194          3010
##      Mildura       Moree        MountGambier    MountGinini
##      3010          3010          3041          3041
##      Newcastle     Nhil        NorahHead      NorfolkIsland
##      3041          1579          3005          3010
##      Nuriootpa     PearceRAAF    Penrith        Perth
##      3009          3009          3040          3193
##      PerthAirport  Portland     Richmond      Sale
##      3009          3010          3010          3010
##      SalmonGums    Sydney      SydneyAirport  Townsville
##      2963          3345          3010          3041
##      Tuggeranong    Uluru       WaggaWagga     Walpole
##      3040          1579          3010          3006
##      Watsonia      Williamtown  Witchcliffe    Wollongong
##      3010          3010          3009          3041
##      Woomera
##      3010
## [1] "Number of locations: 49"

```



This shows that most of the locations have approximately 3000 observations.

The date range of the measurements is a little over 10 years:

```
## [1] "2007-11-01 to 2018-07-30"
```

### Checking for Missing Data

Observations containing missing data should be removed (or replaced with estimates) before attempting to fit any models. This section shows the results of checking columns (variables) and rows (observations) for missing data.

Check columns first:

```
##      Sunshine      Evaporation      Cloud3pm      Cloud9am      Pressure9am
##      70820         64800         60416         57043         15354
##      Pressure3pm      WindDir9am      WindGustDir      WindGustSpeed      WindDir3pm
##      15349         11029         10730         10667         4576
##      Humidity3pm      Temp3pm      WindSpeed3pm      Rainfall      RainToday
##      4412         3571         3407         3218         3218
##      RISK_MM      RainTomorrow      Humidity9am      WindSpeed9am      Temp9am
##      3217         3217         2579         2081         1751
##      MinTemp      MaxTemp      Date      Location
##      1545         1335         0         0
```

These results show that there are four columns with more than 50000 missing values, substantially more than any of the other columns.

Now check how many rows would remain if these four columns were removed and rows containing any missing

values were omitted:

```
## [1] 112658
```

This represents about 77.5% of the observations, which will be regarded as sufficient for fitting and testing models.

## Data Cleaning

The operations involved in cleaning the data are:

1. removing the four columns identified above,
2. omitting rows with any remaining missing data,
3. replacing any character data with numeric values, and
4. converting the column to be predicted, *RainTomorrow*, to a factor for classification purposes.

The *Date* column is also to be removed, along with *RISK\_MM*, which would give the model an unfair advantage. The reason for not including *RISK\_MM* as an input is best explained in the following quote from the creator of the dataset:

*RISK-MM is the amount of rainfall in millimeters for the next day. It includes all forms of precipitation that reach the ground, such as rain, drizzle, hail and snow. And it was the column that was used to actually determine whether or not it rained to create the binary target. For example, if RISK-MM was greater than 0, then the RainTomorrow target variable is equal to Yes.*

After the above operations had been performed, the resulting dataset had the following dimensions:

```
## [1] "No. of rows (observations): 112658"
```

```
## [1] "No. of columns (variables): 18"
```

## Exploratory Analysis of Cleaned Data

The structure of the cleaned data is shown below.

```
## 'data.frame': 112658 obs. of 18 variables:
## $ Location : chr "Albury" "Albury" "Albury" "Albury" ...
## $ MinTemp : num 13.4 7.4 12.9 9.2 17.5 14.6 14.3 7.7 9.7 13.1 ...
## $ MaxTemp : num 22.9 25.1 25.7 28 32.3 29.7 25 26.7 31.9 30.1 ...
## $ Rainfall : num 0.6 0 0 0 1 0.2 0 0 0 1.4 ...
## $ WindGustDir : num 14 15 16 5 14 15 14 14 7 14 ...
## $ WindGustSpeed: int 44 44 46 24 41 56 50 35 80 28 ...
## $ WindDir9am : num 14 7 14 10 2 14 13 11 10 9 ...
## $ WindDir3pm : num 15 16 16 1 8 14 14 14 8 11 ...
## $ WindSpeed9am : int 20 4 19 11 7 19 20 6 7 15 ...
## $ WindSpeed3pm : int 24 22 26 9 20 24 24 17 28 11 ...
## $ Humidity9am : int 71 44 38 45 82 55 49 48 42 58 ...
## $ Humidity3pm : int 22 25 30 16 33 23 19 19 9 27 ...
## $ Pressure9am : num 1008 1011 1008 1018 1011 ...
## $ Pressure3pm : num 1007 1008 1009 1013 1006 ...
## $ Temp9am : num 16.9 17.2 21 18.1 17.8 20.6 18.1 16.3 18.3 20.1 ...
## $ Temp3pm : num 21.8 24.3 23.2 26.5 29.7 28.9 24.6 25.5 30.2 28.2 ...
## $ RainToday : num 0 0 0 0 0 0 0 0 0 1 ...
## $ RainTomorrow : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 ...
```

*Location* remains as a character variable, but it will not be used as a variable in fitting the models.

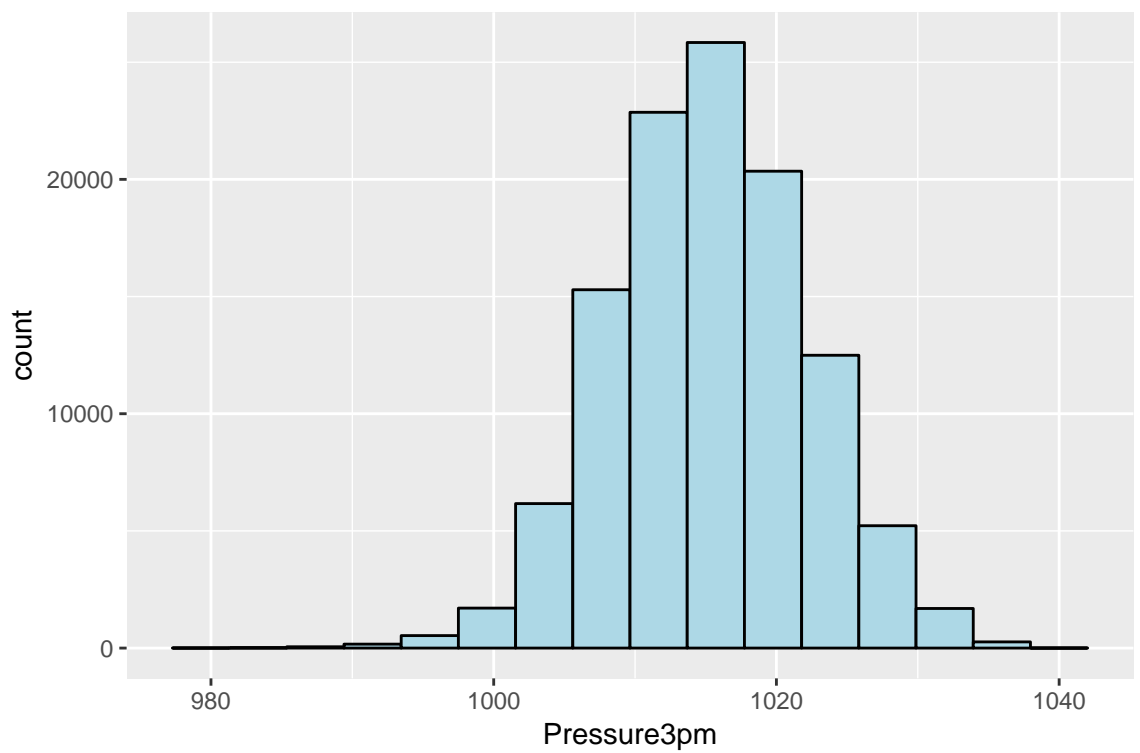
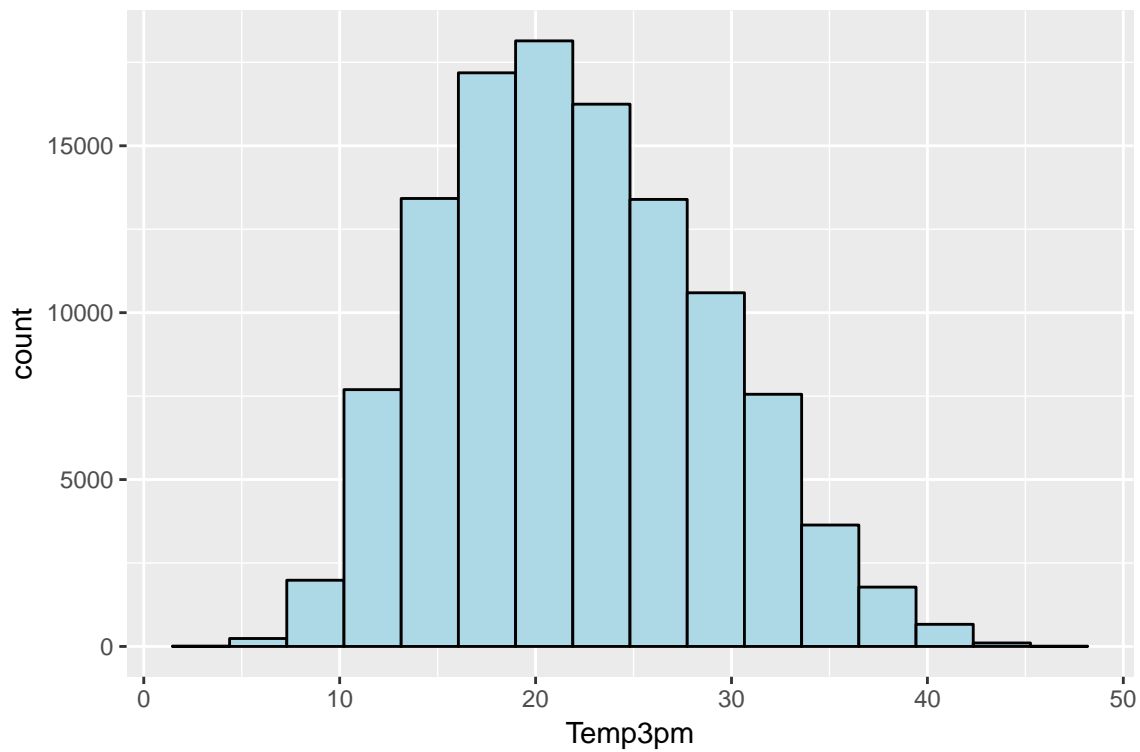
The following table shows summary statistics for each of the variables in the cleaned dataset.

```
##      Location      MinTemp      MaxTemp      Rainfall
## Length:112658   Min.    :-8.70   Min.    : 2.60   Min.    : 0.000
## Class :character 1st Qu.: 7.90   1st Qu.:18.20   1st Qu.: 0.000
## Mode  :character Median :12.30   Median :23.10   Median : 0.000
##                      Mean  :12.54   Mean  :23.61   Mean   : 2.315
##                      3rd Qu.:17.10   3rd Qu.:28.70   3rd Qu.: 0.600
##                      Max.   :33.90   Max.   :48.10   Max.   :367.600
##      WindGustDir   WindGustSpeed   WindDir9am   WindDir3pm
## Min.    : 1.000   Min.    : 7.00   Min.    : 1.000   Min.    : 1.000
## 1st Qu.: 4.000   1st Qu.:31.00   1st Qu.: 4.000   1st Qu.: 5.000
## Median : 9.000   Median :39.00   Median : 8.000   Median : 9.000
## Mean    : 8.713   Mean    :40.77   Mean    : 8.228   Mean    : 8.756
## 3rd Qu.:13.000   3rd Qu.:48.00   3rd Qu.:12.000   3rd Qu.:13.000
## Max.    :16.000   Max.    :135.00   Max.    :16.000   Max.    :16.000
##      WindSpeed9am   WindSpeed3pm   Humidity9am   Humidity3pm
## Min.    : 2.00   Min.    : 2.00   Min.    : 0.00   Min.    : 0.00
## 1st Qu.: 9.00   1st Qu.:13.00   1st Qu.:55.00   1st Qu.:35.00
## Median :13.00   Median :19.00   Median :68.00   Median :51.00
## Mean    :15.17   Mean    :19.51   Mean    :67.15   Mean    :50.27
## 3rd Qu.:20.00   3rd Qu.:24.00   3rd Qu.:81.00   3rd Qu.:65.00
## Max.    :87.00   Max.    :87.00   Max.    :100.00   Max.    :100.00
##      Pressure9am   Pressure3pm   Temp9am      Temp3pm
## Min.    : 980.5   Min.    : 978.9   Min.    : -3.10   Min.    : 2.30
## 1st Qu.:1012.9   1st Qu.:1010.5   1st Qu.:12.60   1st Qu.:16.90
## Median :1017.6   Median :1015.1   Median :17.00   Median :21.60
## Mean    :1017.6   Mean    :1015.2   Mean    :17.37   Mean    :22.08
## 3rd Qu.:1022.4   3rd Qu.:1019.9   3rd Qu.:22.00   3rd Qu.:26.80
## Max.    :1041.0   Max.    :1039.6   Max.    :40.20   Max.    :46.10
##      RainToday     RainTomorrow
## Min.    :0.0000   0:88166
## 1st Qu.:0.0000   1:24492
## Median :0.0000
## Mean    :0.2204
## 3rd Qu.:0.0000
## Max.    :1.0000
```

Apart from *Location* and the binary-valued *RainToday* and *RainTomorrow*, most of the variables are fairly symmetrically distributed, with their means and medians close in value. The exception is *Rainfall*, which is positively skewed, with a minimum of 0, a median of 0, and a maximum of 367.6.

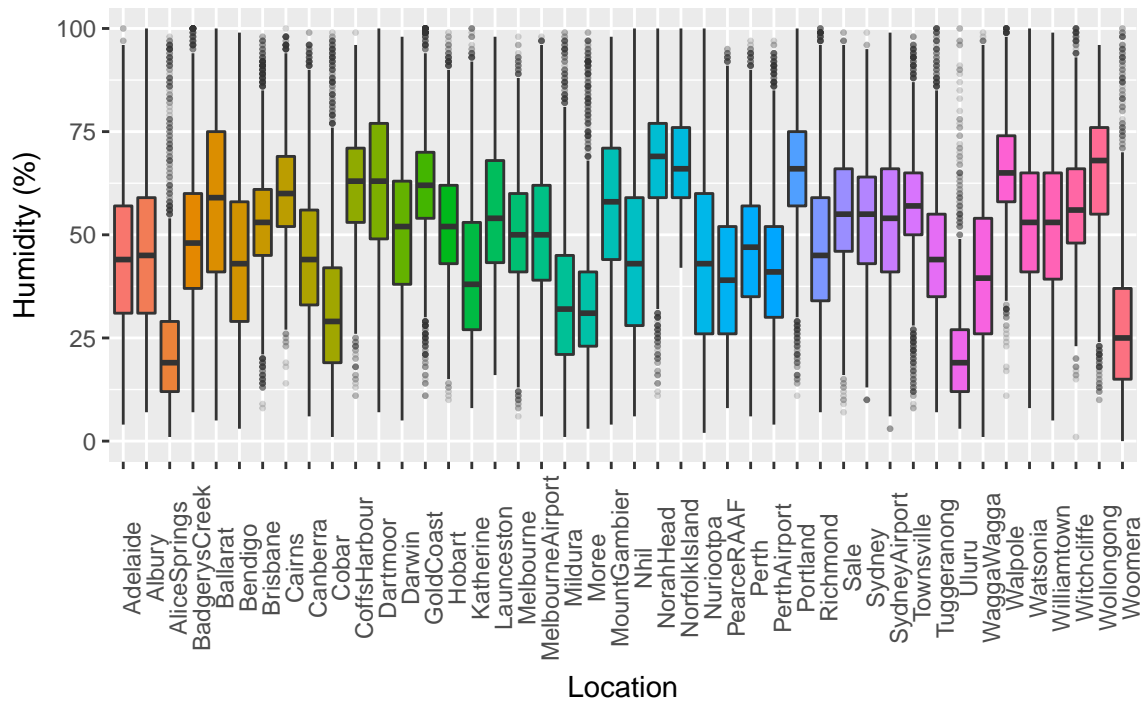
No attempt was made to detect and remove outliers, although removing them from *Rainfall* may have improved the performance of the models.

Samples of the frequency distributions are shown for two of the variables, *Temp3pm* and *Pressure3pm*, in the plots below.

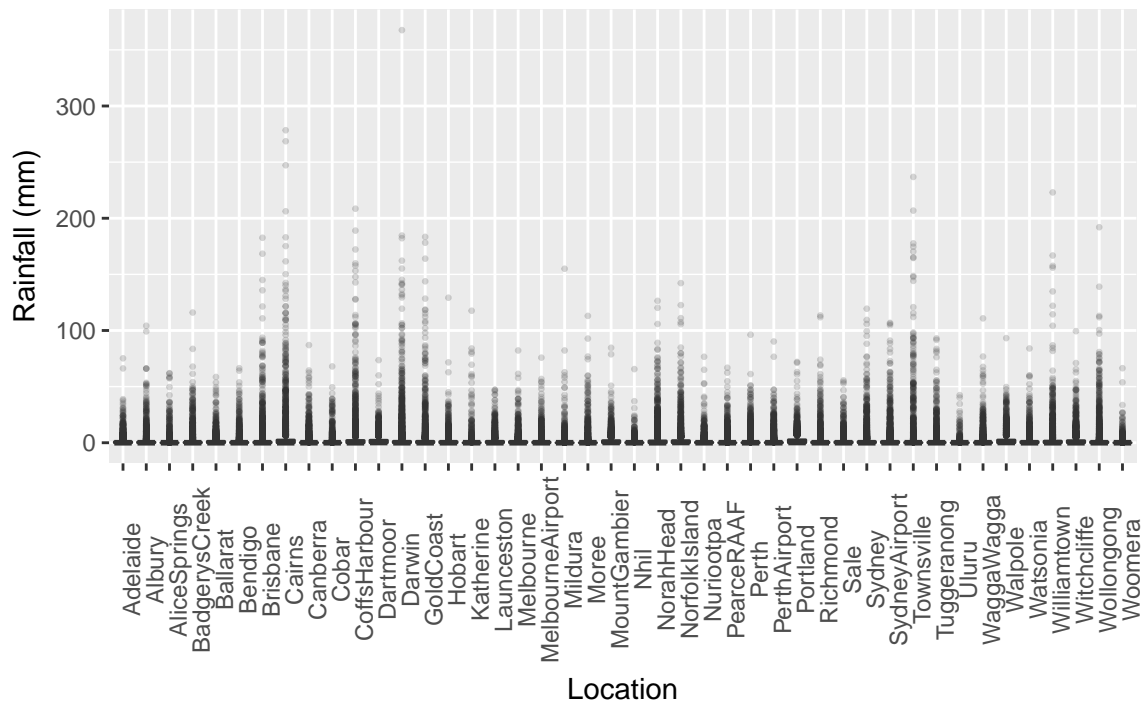


While the above plots show distributions over all locations, the following boxplots show how the distribution varies with location. The first plot shows how *Humidity3pm* varies from location to location, in terms of both median value and spread. The second plot shows how *Rainfall* differs from the other variables, with a median value at or near 0 for all locations but quite different ranges of values.

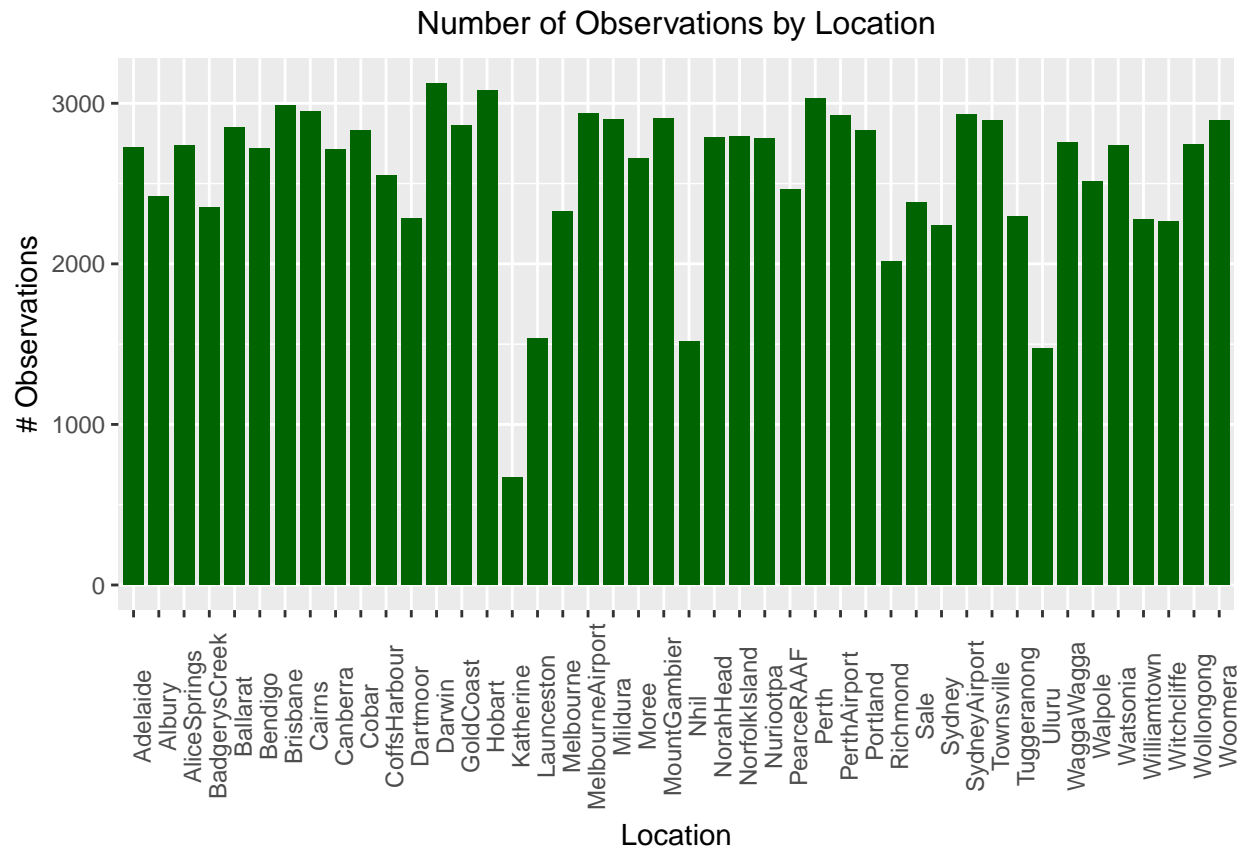
Relative Humidity at 3 pm by Location



Daily Rainfall by Location



The following plot shows how many observations remain for each location.



It can be seen that there is now more variability in the number of observations across locations, due to each location having a different amount of missing data that has been removed.

It also appears that there are now fewer locations in the plot than there were originally:

```
## [1] "Number of locations originally: 49"
```

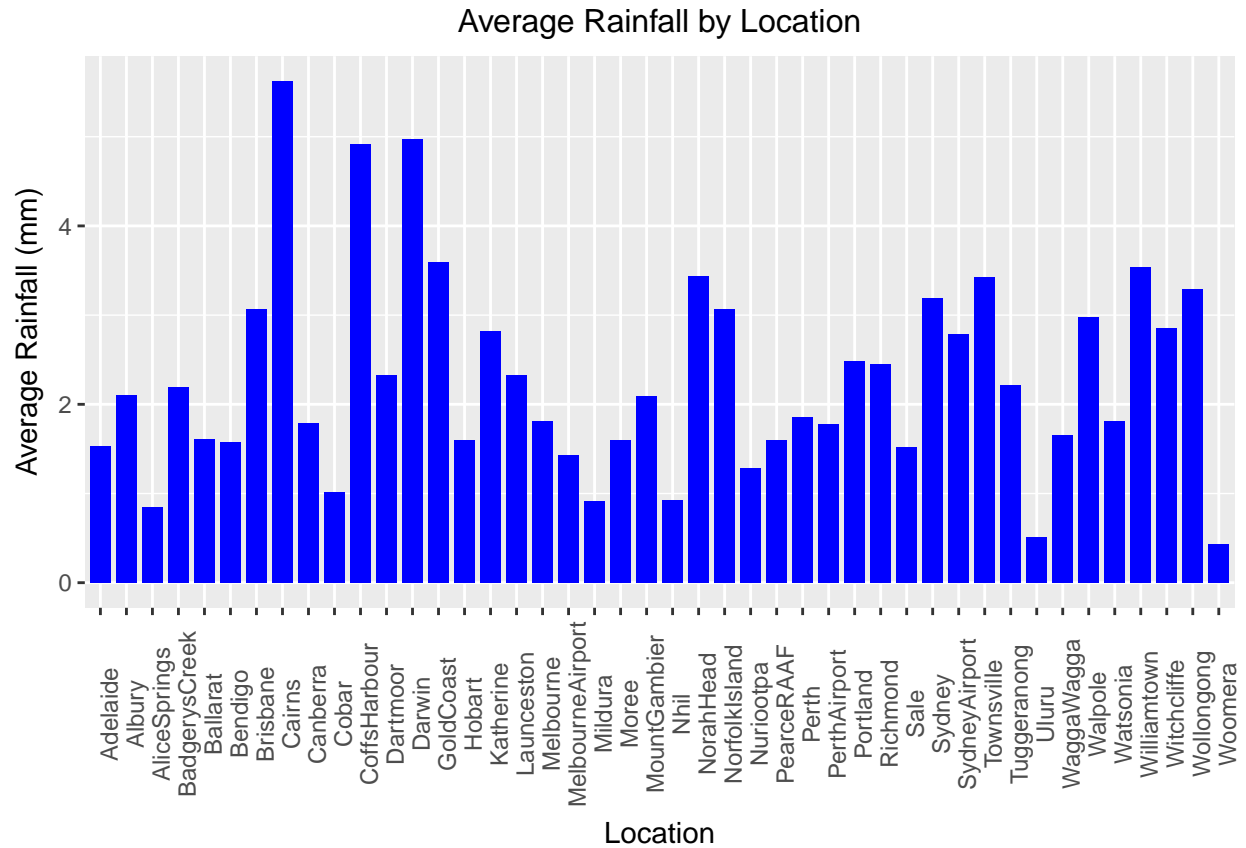
```
## [1] "Number of locations remaining : 44"
```

The reason that some locations are missing from the cleaned data is that these locations did not have any measurements for one or more of the variables.



As the model will be predicting whether or not it will rain, it is worth investigating how rainfall varies across locations, in terms of both average rainfall and days with rain.

The average rainfall for each location, based on the observations in the dataset, is shown in the plot below.



It can be seen from the plot that places like Uluru (Ayres Rock) and Woomera have a very low average rainfall, as expected; while Cairns and Darwin have a much higher average rainfall, also as expected.

The percentage of days with rain at these locations follows a similar trend, as revealed by the following table and plot.

Overall, the number of days without rain is much greater than the number of days with rain.

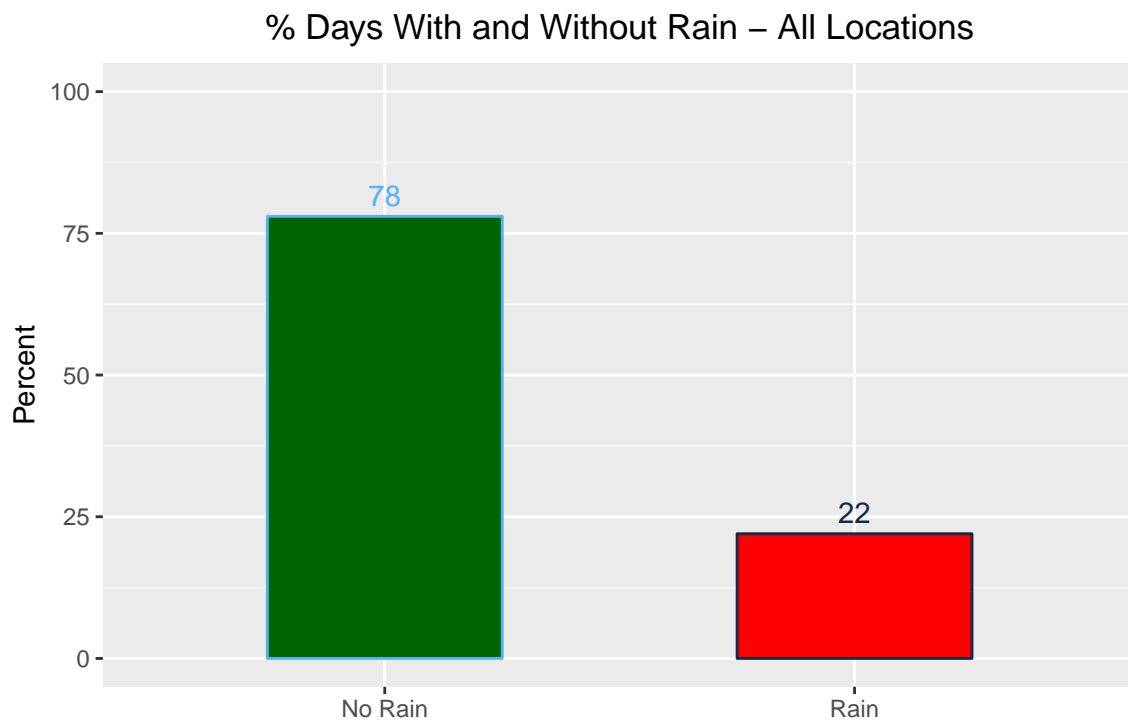
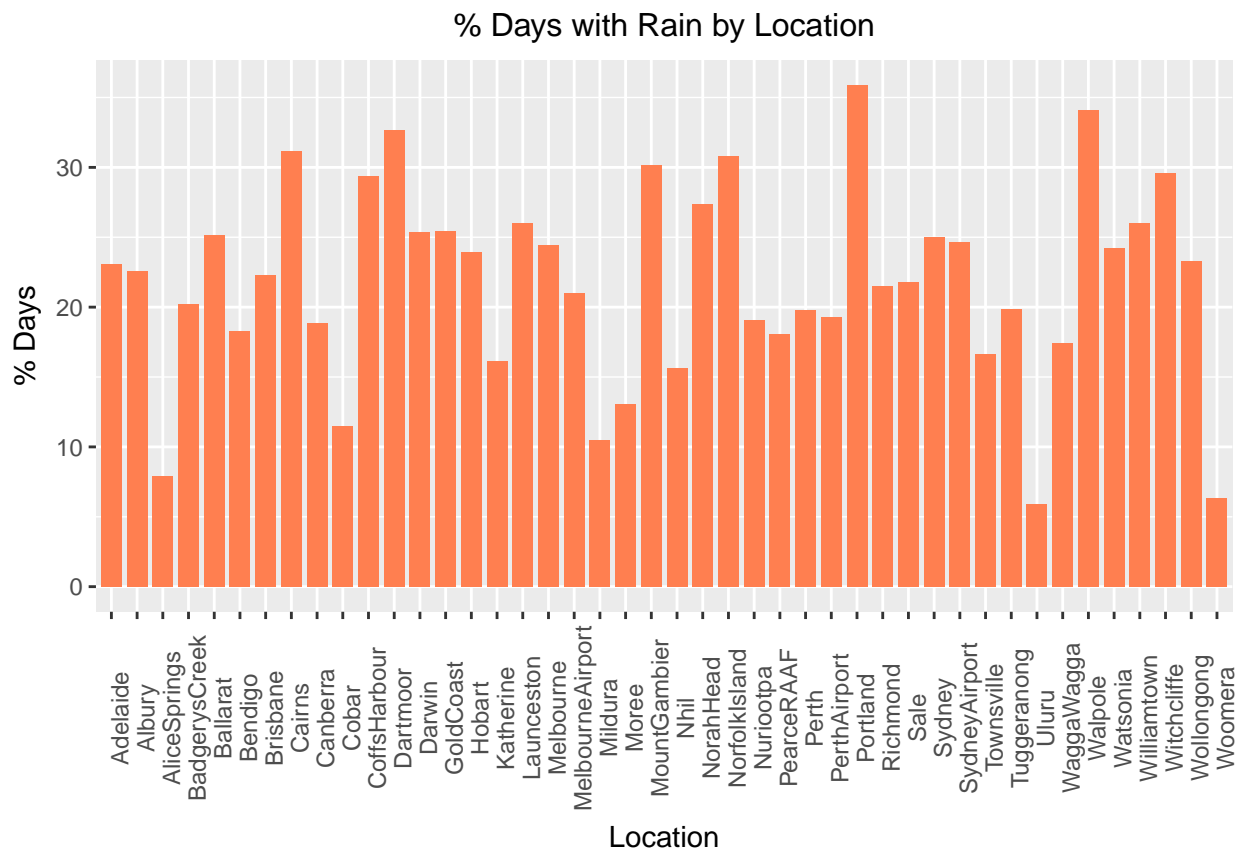
##	Location	Rainy Days	Total Days	% Rainy Days
## 1	Adelaide	629	2724	23.1
## 2	Albury	546	2422	22.5
## 3	AliceSprings	216	2735	7.9
## 4	BadgerysCreek	475	2350	20.2
## 5	Ballarat	717	2849	25.2
## 6	Bendigo	497	2718	18.3
## 7	Brisbane	667	2989	22.3
## 8	Cairns	920	2952	31.2
## 9	Canberra	511	2714	18.8
## 10	Cobar	326	2833	11.5
## 11	CoffsHarbour	749	2548	29.4
## 12	Dartmoor	746	2284	32.7
## 13	Darwin	792	3124	25.4
## 14	GoldCoast	728	2864	25.4
## 15	Hobart	737	3082	23.9

## 16	Katherine	108	670	16.1
## 17	Launceston	400	1538	26.0
## 18	Melbourne	567	2323	24.4
## 19	MelbourneAirport	616	2936	21.0
## 20	Mildura	303	2897	10.5
## 21	Moree	347	2659	13.1
## 22	MountGambier	878	2908	30.2
## 23	Nhil	237	1519	15.6
## 24	NorahHead	763	2785	27.4
## 25	NorfolkIsland	860	2795	30.8
## 26	Nuriootpa	531	2783	19.1
## 27	PearceRAAF	446	2466	18.1
## 28	Perth	599	3031	19.8
## 29	PerthAirport	563	2923	19.3
## 30	Portland	1015	2828	35.9
## 31	Richmond	433	2012	21.5
## 32	Sale	519	2382	21.8
## 33	Sydney	559	2236	25.0
## 34	SydneyAirport	723	2933	24.7
## 35	Townsville	482	2894	16.7
## 36	Tuggeranong	456	2293	19.9
## 37	Uluru	87	1475	5.9
## 38	WaggaWagga	481	2758	17.4
## 39	Walpole	855	2510	34.1
## 40	Watsonia	663	2739	24.2
## 41	Williamstown	593	2278	26.0
## 42	Witchcliffe	670	2267	29.6
## 43	Wollongong	639	2742	23.3
## 44	Woomera	183	2890	6.3

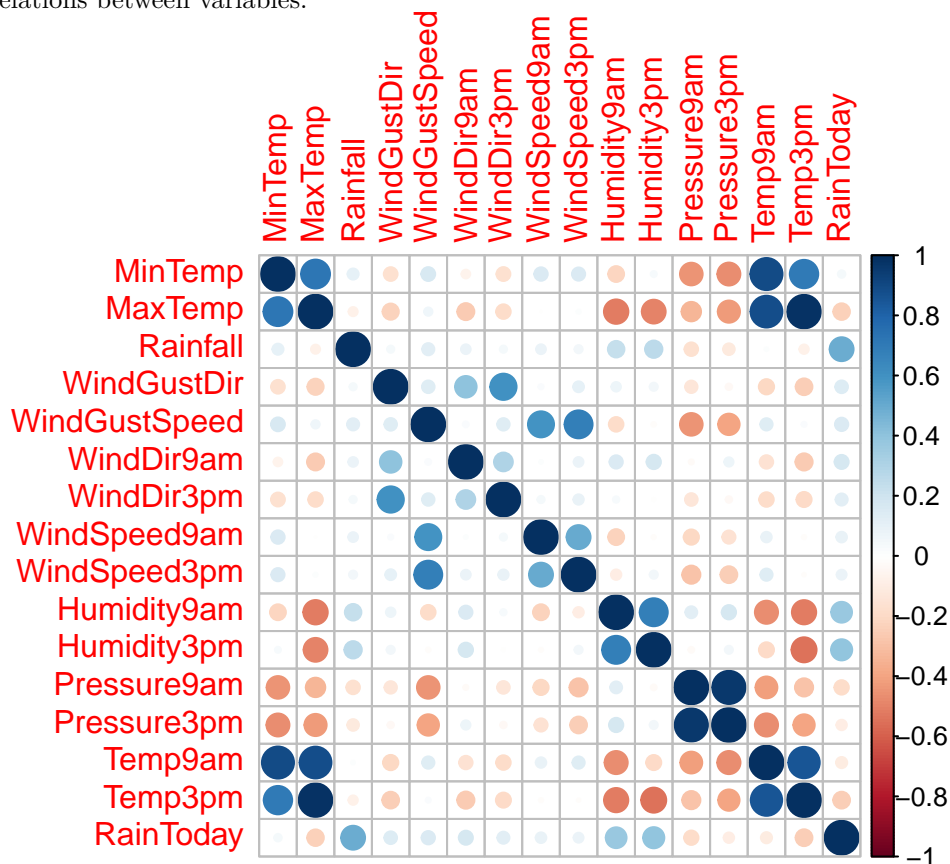
## [1] "Total days in the dataset: 112658"

## [1] "Number of rainy days: 24832"

## [1] "Number of fine days: 87826"



Check for correlations between variables:

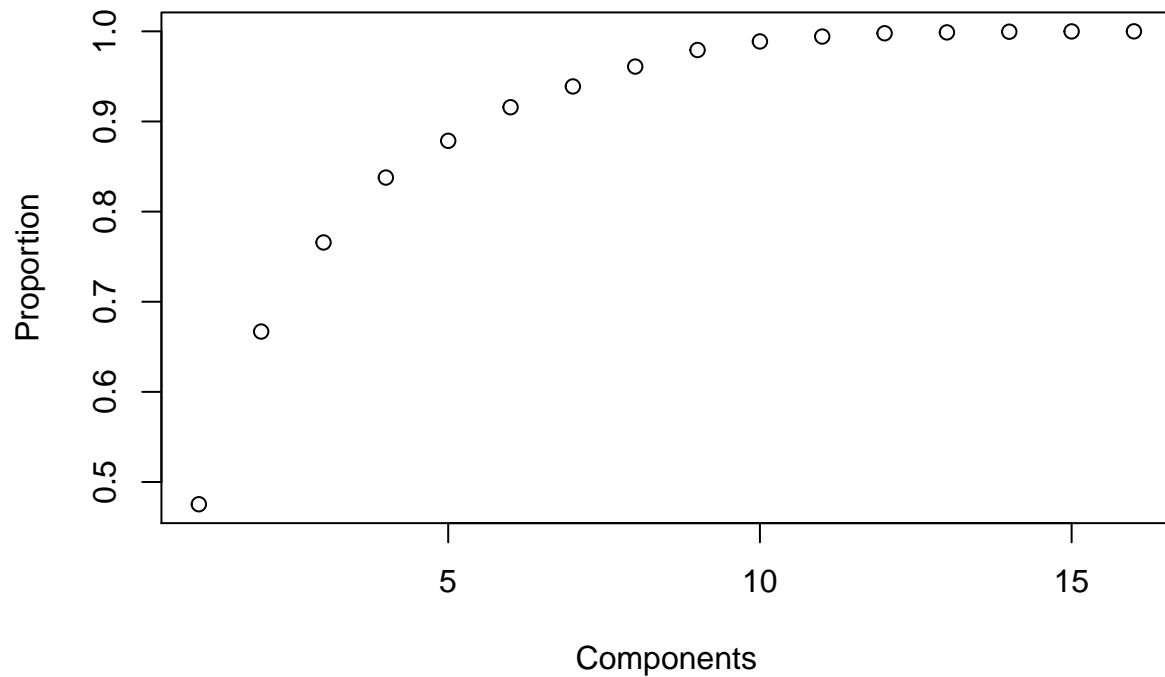


The plot shows that there is a strong positive correlation between all the temperature measurements on a given day at a given location. There is also strong positive correlation between the morning and afternoon atmospheric pressure measurements and a slightly weaker correlation between the morning and afternoon humidity measurements. A negative correlation exists between maximum temperature and humidity, and also between minimum temperature and atmospheric pressure.

These correlations suggest that it may be possible to reduce the dimensionality of the data. Principal component analysis indicated that eight principal components are required to explain 95% of the variation in the data - a reduction in the dimensionality of 50%. As the dimensionality is already fairly small and reducing it would probably result in lower prediction accuracy, this was not investigated any further.

```
## Importance of components:
##          PC1      PC2      PC3      PC4      PC5      PC6
## Standard deviation 26.9110 17.0855 12.27306 10.47337 7.88318 7.53333
## Proportion of Variance 0.4753 0.1916 0.09886 0.07199 0.04079 0.03725
## Cumulative Proportion 0.4753 0.6669 0.76578 0.83777 0.87856 0.91581
##          PC7      PC8      PC9      PC10     PC11     PC12
## Standard deviation  5.92865 5.79690 5.29169 3.80213 2.88068 2.35244
## Proportion of Variance 0.02307 0.02206 0.01838 0.00949 0.00545 0.00363
## Cumulative Proportion 0.93888 0.96093 0.97931 0.98880 0.99425 0.99788
##          PC13     PC14     PC15     PC16
## Standard deviation 1.21193 1.06775 0.71920 0.32236
## Proportion of Variance 0.00096 0.00075 0.00034 0.00007
## Cumulative Proportion 0.99884 0.99959 0.99993 1.00000
```

## Cumulative Proportion of Variance from Principal Components



To see whether *RainToday* is a good predictor of *RainTomorrow*, the correlation coefficient for the two variables was calculated:

```
## [1] "Correlation coefficient: 0.33"
```

This value indicates that rain today is a poor predictor of rain tomorrow.

From the above results, it appears that the dataset is now in a suitable form for fitting models.

## Choosing and Fitting the Models

The approach adopted was to choose six different classification models, fit a model based on each, and see which model gave the best performance.

The models chosen and the relevant R libraries are shown in the table below.

Model Type	caret Method	Library
Generalized Linear Model	glm	stats (system)
Linear Discriminant Analysis	lda	MASS (system)
Quadratic Discriminant Analysis	qda	MASS (system)
Multi-Layer Perceptron	mlp	RSNNS
k-Nearest Neighbors	knn	class (system)
Random Forest	rf	randomForest

As the functions for training and prediction with these models have different syntax, the *caret* package was chosen so that a uniform syntax could be used. This allows the training of all the models to be performed within a loop, or with a single call to the *lapply()* function.

Prior to fitting the models, the data was standardized so that all values were in the range 0 to 1, and then split into training and test sets:

```
# Scale the data
maxs <- apply(dfWeather[,2:16], 2, max)
mins <- apply(dfWeather[,2:16], 2, min)
dfScaled <- as.data.frame(scale(dfWeather[,2:16], center = mins, scale = maxs-mins))
dfScaled <- cbind(dfScaled, dfWeather[,17:18])

# Create datasets for training and testing
set.seed(1)
test_index <- createDataPartition(y=dfScaled$RainTomorrow, times=1, p=0.2, list=FALSE)
dfTrain <- dfScaled[-test_index,]
dfTest <- dfScaled[test_index,]
rm(dfScaled)
```

The training and predicting could then be performed, and the accuracy determined for each of the models:

```
# Train the models
models <- c("glm", "lda", "qda", "mlp", "knn", "rf")
t0 <- proc.time()
fits <- lapply(models, function(model) {
  caret::train(RainTomorrow ~ ., data = dfTrain, method = model)
})

t1 <- proc.time()
names(fits) <- models

# Perform prediction with each of the models
y_hats <- sapply(fits, function(fit) {
  y_hat <- predict(fit, dfTest)
})
t2 <- proc.time()

# Calculate the accuracy of the models and format as a table
accuracy <- colMeans(y_hats == dfTest$RainTomorrow)
accuracy <- sort(round(accuracy, 4), decreasing = TRUE)
```

```

accuracy <- cbind(names(accuracy), accuracy)
row.names(accuracy) <- NULL
acc_tbl <- kable(accuracy, col.names = c("Model", "Accuracy"))

## [1] "Elapsed time for training: 15642 seconds"
## [1] "Elapsed time for predicting: 34 seconds"

```

## Results

All six models were fitted successfully, with accuracies in the range 0.83 to 0.86. The accuracies for each of the models, based on predictions for the test set, are tabulated below.

Model	Accuracy
rf	0.8587
mlp	0.8528
glm	0.8517
lda	0.8503
knn	0.8472
qda	0.8359

The table shows that the Random Forest model gave marginally better performance than the other models. This model was chosen as the one to investigate further.

The Variable Importance table below shows which of the input variables have the greatest predictive power.

Variable	Importance
Humidity3pm	5533.3762
Pressure3pm	2246.5936
Humidity9am	2237.0872
Pressure9am	2180.1202
WindGustSpeed	2031.0359
Temp3pm	2008.0924
Rainfall	1950.7387
MinTemp	1820.6301
MaxTemp	1790.1789
Temp9am	1712.6539
WindSpeed3pm	1278.1451
WindSpeed9am	1211.3622
WindDir9am	1145.8934
WindDir3pm	1112.7313
WindGustDir	1104.4871
RainToday	914.6302

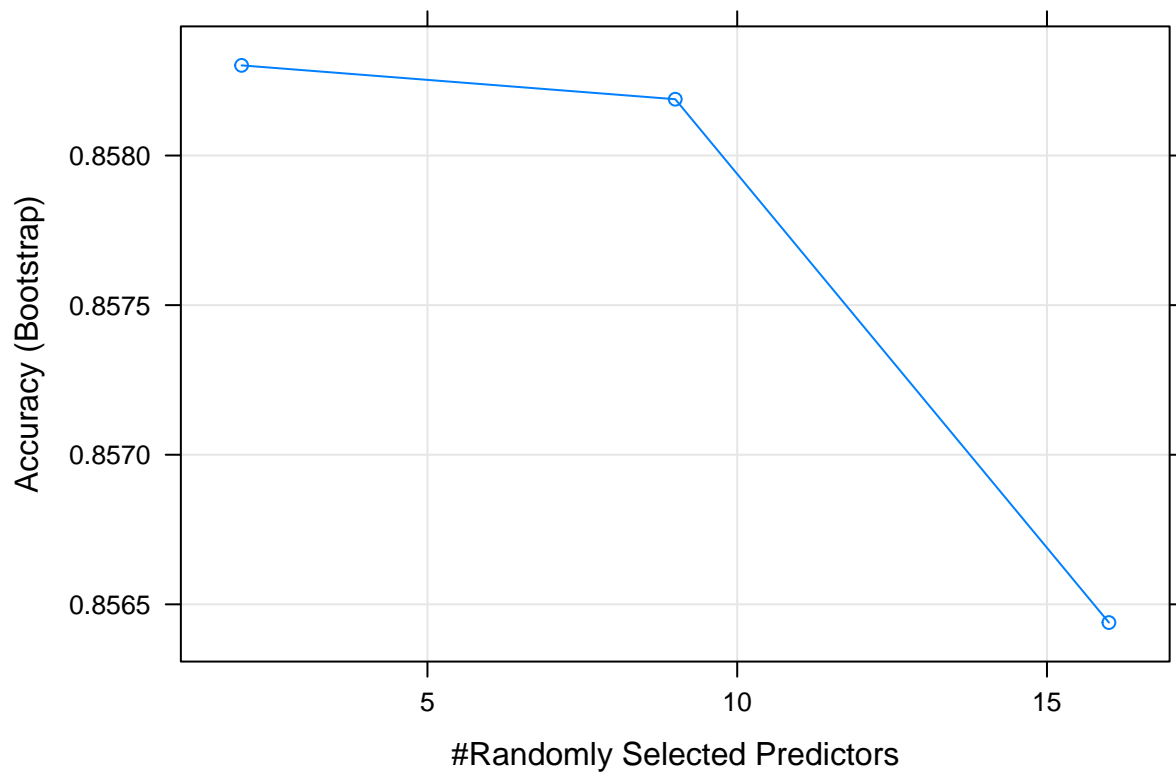
The table shows that relative humidity and atmospheric pressure on the day are the most important predictors of rain on the following day. It also shows that rain on the day is a poor predictor of rain on the following day, which is consistent with the low correlation between the two that was noted above.

Additional details of the random forest model fitted are presented below.

```

## Random Forest
##
## 90125 samples
##    16 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 90125, 90125, 90125, 90125, 90125, 90125, ...
## Resampling results across tuning parameters:
##
##  mtry  Accuracy  Kappa
##    2    0.8583013 0.5251468
##    9    0.8581882 0.5330717
##   16    0.8564392 0.5296578
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.

```



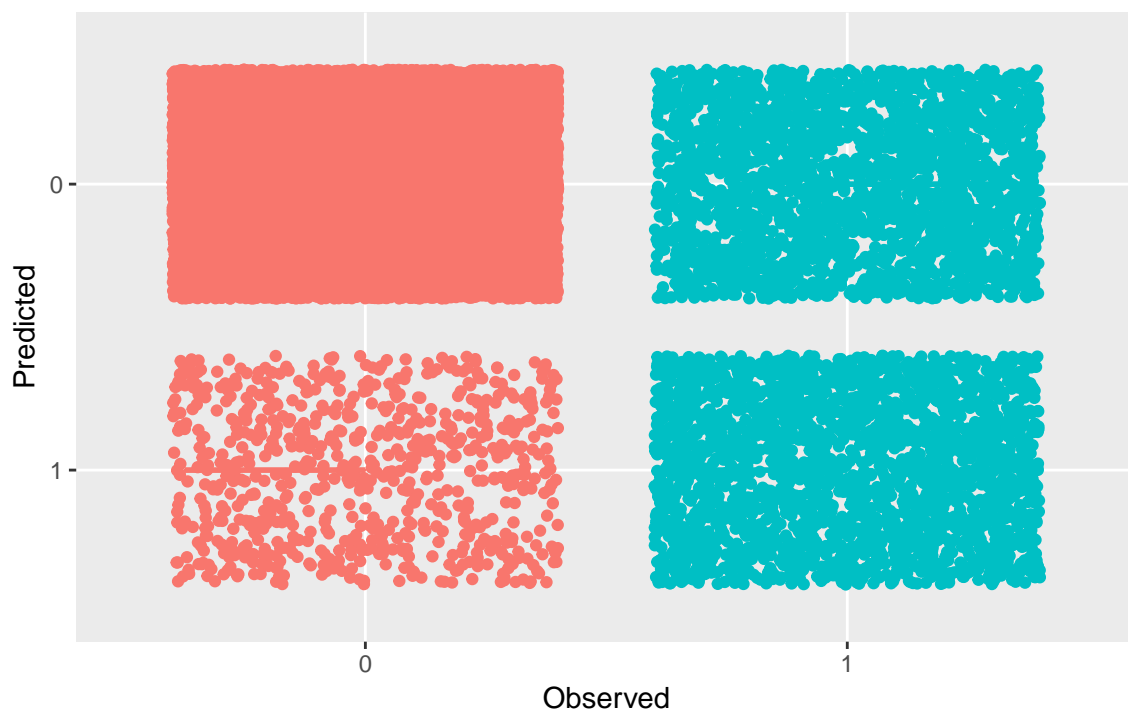
Although an accuracy of approximately 0.86 was obtained for the test set predictions, the model's ability to predict rain is considerably lower. This can be seen by examining the confusion matrix:



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##           0 16902 2451
##           1   732 2448
##
##           Accuracy : 0.8587
##           95% CI : (0.8541, 0.8633)
##           No Information Rate : 0.7826
##           P-Value [Acc > NIR] : < 2.2e-16
##
##           Kappa : 0.5247
##           McNemar's Test P-Value : < 2.2e-16
##
##           Sensitivity : 0.4997
##           Specificity : 0.9585
##           Pos Pred Value : 0.7698
##           Neg Pred Value : 0.8734
##           Prevalence : 0.2174
##           Detection Rate : 0.1086
##           Detection Prevalence : 0.1411
##           Balanced Accuracy : 0.7291
##
##           'Positive' Class : 1
##
```

The confusion matrix shows that the number of false negatives (2451) slightly exceeds the number of true positives (2448). This means that on days that it rained, the model predicted correctly only 50% of the time. This is reflected in the sensitivity value of 0.4997, and can be seen in a visualization of the confusion matrix.

Confusion Matrix Plot for 'Rain Tomorrow'



The density of the dots in the two boxes on the right, representing true positives and false negatives, is approximately equal, indicating that on days that it rains the model is just as likely to have predicted that it will be fine.

## Conclusion

This project has demonstrated that it is possible to construct models to predict whether or not it will rain tomorrow, based on various weather measurements made today. Six different classification models were fitted, and all had a similar accuracy. A Random Forest model had the highest accuracy, so it was examined in more detail.

The model identified relative humidity and atmospheric pressure on the previous day as the best predictors of whether or not it will rain, and rain on the previous day as the worst.

The results also highlighted the need to consider other model statistics, such as sensitivity and specificity, when assessing the performance of a model. In this case, although the accuracy was fairly high at 0.86, the sensitivity was low, with a value of 0.50. Specificity was good, with a value of 0.96.

The high specificity value means that the false positive rate is low. The low sensitivity value means that the false negative rate is relatively high, so that on days that it rains, there is a 50-50 chance that the model predicted correctly. This is in contrast with the 86% chance that the model makes a correct prediction on average.