Rain in Australia. Classification Prediction Model

by Sumaira Afzal, Viraja Ketkar, Murlidhar Loka, Vadim Spirkov

Abstract An abstract of less than 150 words.

Introduction

Background

Objective

Data Analisys

The data set we are going to use for our research contains daily weather observations from numerous Australian weather stations from 2007 till 2017. There are over 142000 records. It has been sourced from Kagle

Data Dictionary

We exclude the variable Risk-MM when training your binary classification model. If we don't exclude it, you will leak the answers to our model and reduce its predictability

Column Name	Column Description
Date	Date of observation
Location	Common name of the location of the weather station
MinTemp	Minimum temperature in degrees celsius
MaxTemp	Maximum temperature in degrees celsius
Rainfall	Amount of rainfall recorded for the day in mm
Evaporation	So-called Class A pan evaporation (mm) in the 24 hours to 9am
Sunshine	Number of hours of bright sunshine in the day
WindGustDir	Direction of the strongest wind gust in the 24 hours to midnight
WindGustSpeed	Speed (km/h) of the strongest wind gust in the 24 hours to midnight
WindDir9amDirection	Of the wind at 9am
WindDir3pmDirection	Of the wind at 3pm
WindSpeed9amWind	Wind speed (km/hr) averaged over 10 minutes prior to 9am
WindSpeed3pmWind	Wind Speed (km/hr) averaged over 10 minutes prior to 3pm
Humidity9amHumidity	Humidity (percent) at 9am
Humidity3pmHumidity	Humidity (percent) at 3pm
Pressure9amAtmospheric	Pressure (hpa) reduced to mean sea level at 9am
Pressure3pmAtmospheric	Pressure (hpa) reduced to mean sea level at 3pm
Cloud9amFraction	Area of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast
Cloud3pmFraction	Area of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values

Column Name	Column Description
Temp9amTemperature	Temperature (degrees C) at 9am
Temp3pmTemperature	Temperature (degrees C) at 3pm
RainTodayBoolean	Rainy today. 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
RISK_MM	Amount of rain. A kind of measure of the "risk". This column is redundant and will be dropped
RainTomorrowThe	**Target variable. Will it rain tomorrow?**

Data Exploration

Let's take a close look at the data set. We start with loading weather observations from the file into a data frame. We remove RISK_MM as explained and convert Date column to *date* format

```
weatherData = read.csv("../data/weatherAUS.csv", header = TRUE, na.strings = c("NA","","#NA"),sep=",")
weatherData = subset(weatherData, select = -RISK_MM)
weatherData$Date = as.Date(as.character(weatherData$Date),"%Y-%m-%d")
```

Now let's load coordinates of the weather stations and have a bird-eye view of the weather station locations

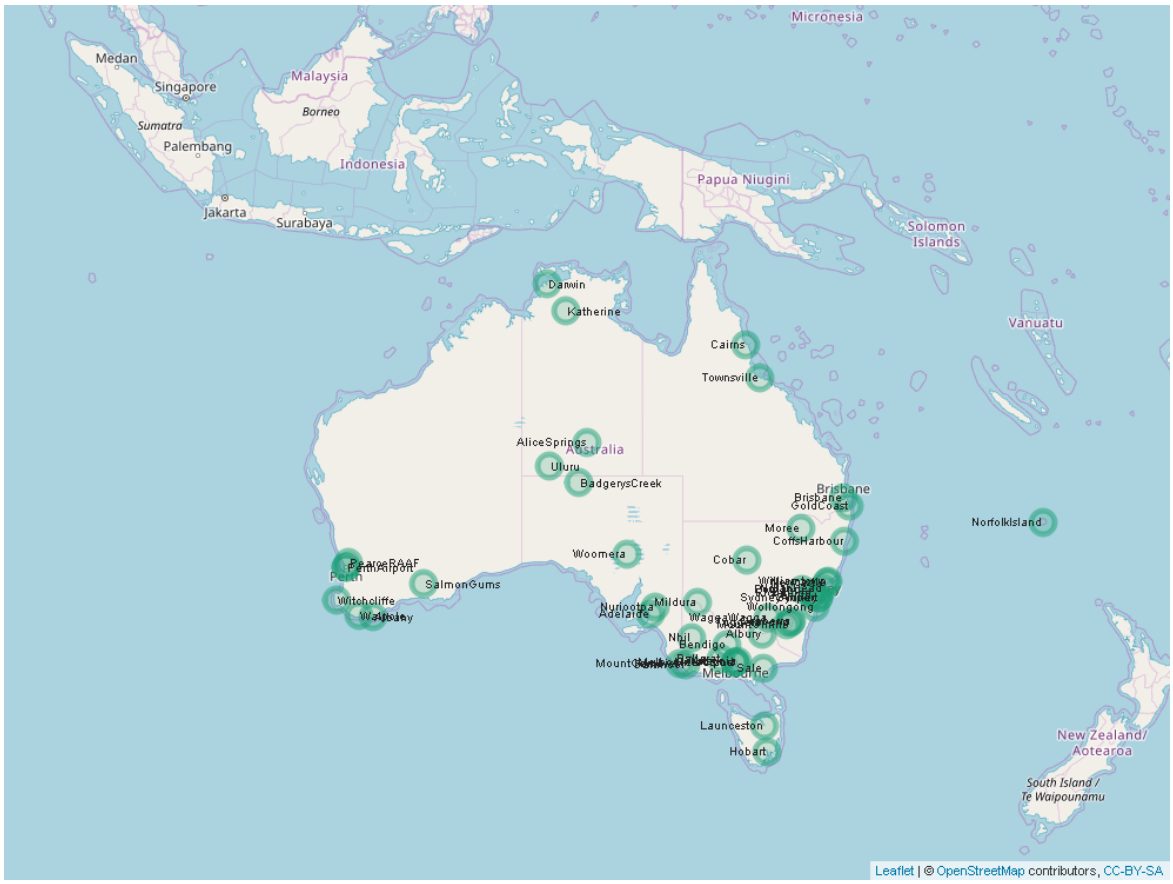

Figure 1: Australian Weather Stations

Let's review data summary

```
summary(weatherData)
```

```
#>       Date               Location         MinTemp         MaxTemp
#>  Min.   :2007-11-01   Canberra:  3418   Min.   :-8.50   Min.   :-4.80
#>  1st Qu.:2011-01-06   Sydney  :  3337   1st Qu.: 7.60   1st Qu.:17.90
#>  Median :2013-05-27   Perth   :  3193   Median :12.00   Median :22.60
#>  Mean   :2013-04-01   Darwin  :  3192   Mean   :12.19   Mean   :23.23
#>  3rd Qu.:2015-06-12   Hobart  :  3188   3rd Qu.:16.80   3rd Qu.:28.20
#>  Max.   :2017-06-25   Brisbane:  3161   Max.   :33.90   Max.   :48.10
#>                       (Other) :122704   NA's   :637     NA's   :322
#>     Rainfall       Evaporation        Sunshine      WindGustDir
#>  Min.   :  0.00   Min.   :  0.00   Min.   : 0.00   W      : 9780
#>  1st Qu.:  0.00   1st Qu.:  2.60   1st Qu.: 4.90   SE     : 9309
#>  Median :  0.00   Median :  4.80   Median : 8.50   E      : 9071
#>  Mean   :  2.35   Mean   :  5.47   Mean   : 7.62   N      : 9033
#>  3rd Qu.:  0.80   3rd Qu.:  7.40   3rd Qu.:10.60   SSE    : 8993
#>  Max.   :371.00   Max.   :145.00   Max.   :14.50   (Other):86677
#>  NA's   :1406     NA's   :60843    NA's   :67816   NA's   : 9330
#>  WindGustSpeed      WindDir9am      WindDir3pm     WindSpeed9am
#>  Min.   :  6.00   N      :11393   SE     :10663   Min.   :  0
#>  1st Qu.: 31.00   SE     : 9162   W      : 9911   1st Qu.:  7
#>  Median : 39.00   E      : 9024   S      : 9598   Median : 13
#>  Mean   : 39.98   SSE    : 8966   WSW    : 9329   Mean   : 14
#>  3rd Qu.: 48.00   NW     : 8552   SW     : 9182   3rd Qu.: 19
#>  Max.   :135.00   (Other):85083   (Other):89732   Max.   :130
#>  NA's   :9270     NA's   :10013   NA's   : 3778   NA's   :1348
#>   WindSpeed3pm    Humidity9am      Humidity3pm      Pressure9am
#>  Min.   : 0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 980.5
#>  1st Qu.:13.00   1st Qu.: 57.00   1st Qu.: 37.00   1st Qu.:1012.9
#>  Median :19.00   Median : 70.00   Median : 52.00   Median :1017.6
#>  Mean   :18.64   Mean   : 68.84   Mean   : 51.48   Mean   :1017.7
#>  3rd Qu.:24.00   3rd Qu.: 83.00   3rd Qu.: 66.00   3rd Qu.:1022.4
#>  Max.   :87.00   Max.   :100.00   Max.   :100.00   Max.   :1041.0
#>  NA's   :2630    NA's   :1774     NA's   :3610     NA's   :14014
#>   Pressure3pm        Cloud9am        Cloud3pm        Temp9am
#>  Min.   : 977.1   Min.   :0.00    Min.   :0.0     Min.   :-7.20
#>  1st Qu.:1010.4   1st Qu.:1.00    1st Qu.:2.0     1st Qu.:12.30
#>  Median :1015.2   Median :5.00    Median :5.0     Median :16.70
#>  Mean   :1015.3   Mean   :4.44    Mean   :4.5     Mean   :16.99
#>  3rd Qu.:1020.0   3rd Qu.:7.00    3rd Qu.:7.0     3rd Qu.:21.60
#>  Max.   :1039.6   Max.   :9.00    Max.   :9.0     Max.   :40.20
#>  NA's   :13981    NA's   :53657   NA's   :57094   NA's   :904
#>     Temp3pm      RainToday     RainTomorrow
#>  Min.   :-5.40   No  :109332   No :110316
#>  1st Qu.:16.60   Yes : 31455   Yes: 31877
#>  Median :21.10   NA's:  1406
#>  Mean   :21.69
#>  3rd Qu.:26.40
#>  Max.   :46.70
#>  NA's   :2726
```

Missing Data

Further analisys of data shows that many features are missing. Some data losses are very significant. We are going to identify what data is missing and if it is feasible to recover the data.

```
print(sort(colSums(is.na(weatherData)), decreasing = T))
```

```
#>      Sunshine   Evaporation      Cloud3pm      Cloud9am   Pressure9am
#>         67816         60843         57094         53657         14014
#>   Pressure3pm    WindDir9am   WindGustDir WindGustSpeed    WindDir3pm
#>         13981         10013          9330          9270          3778
#>   Humidity3pm       Temp3pm  WindSpeed3pm   Humidity9am      Rainfall
```

```
#>          3610          2726          2630          1774          1406
#>     RainToday  WindSpeed9am       Temp9am       MinTemp       MaxTemp
#>          1406          1348           904           637           322
#>          Date      Location  RainTomorrow
#>             0             0             0
```

To speed up data processing and plot rendering we are going to use a data sample. For population of 142K obesrvations, 20K sample size would be sufficient for 99% confidence level with the condience interval 1

```
weatherSample = sample_n(weatherData, 20000)
aggr(weatherSample, numbers = F, prop = T, col = mainPalette, sortVars = T, bars = F, varheight = T)
```

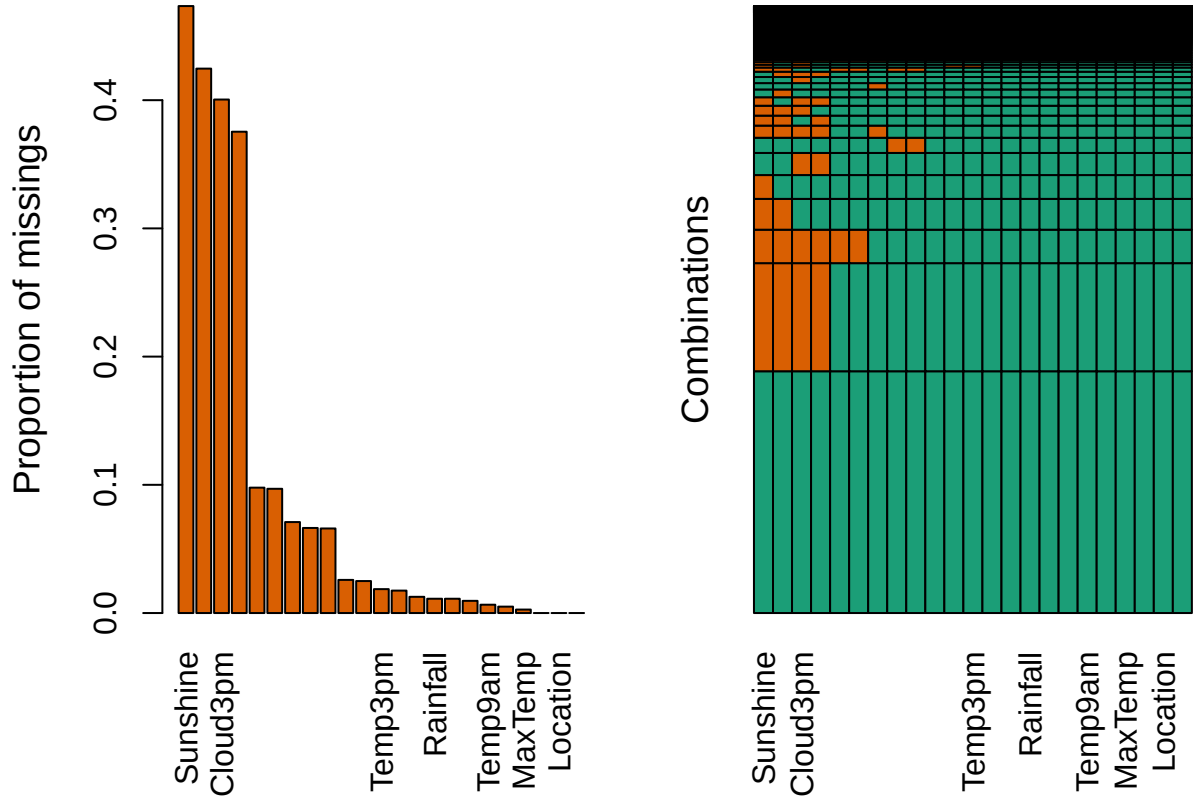

Figure 2: Missing Data Summary

```
#>
#>  Variables sorted by number of missings:
#>       Variable   Count
#>       Sunshine 0.47340
#>    Evaporation 0.42460
#>       Cloud3pm 0.40050
#>       Cloud9am 0.37540
#>    Pressure9am 0.09780
#>    Pressure3pm 0.09690
#>     WindDir9am 0.07095
#>    WindGustDir 0.06635
#>  WindGustSpeed 0.06595
#>     WindDir3pm 0.02585
#>    Humidity3pm 0.02495
#>        Temp3pm 0.01865
#>   WindSpeed3pm 0.01750
#>    Humidity9am 0.01265
#>       Rainfall 0.01115
```

```
#>      RainToday 0.01115
#>   WindSpeed9am 0.00950
#>        Temp9am 0.00650
#>        MinTemp 0.00500
#>        MaxTemp 0.00270
#>           Date 0.00000
#>       Location 0.00000
#>   RainTomorrow 0.00000
```

As demonstrated in Figure 2 *Sunshine*, *Evaporation* and *Clouds* columns saffer the loss of data between **48%** and **38%**. This is significant! Sinse we are dealing with the weather patterns we should be observing cyclical data pattenrs. Let's review data stribution of features that damaged the most.

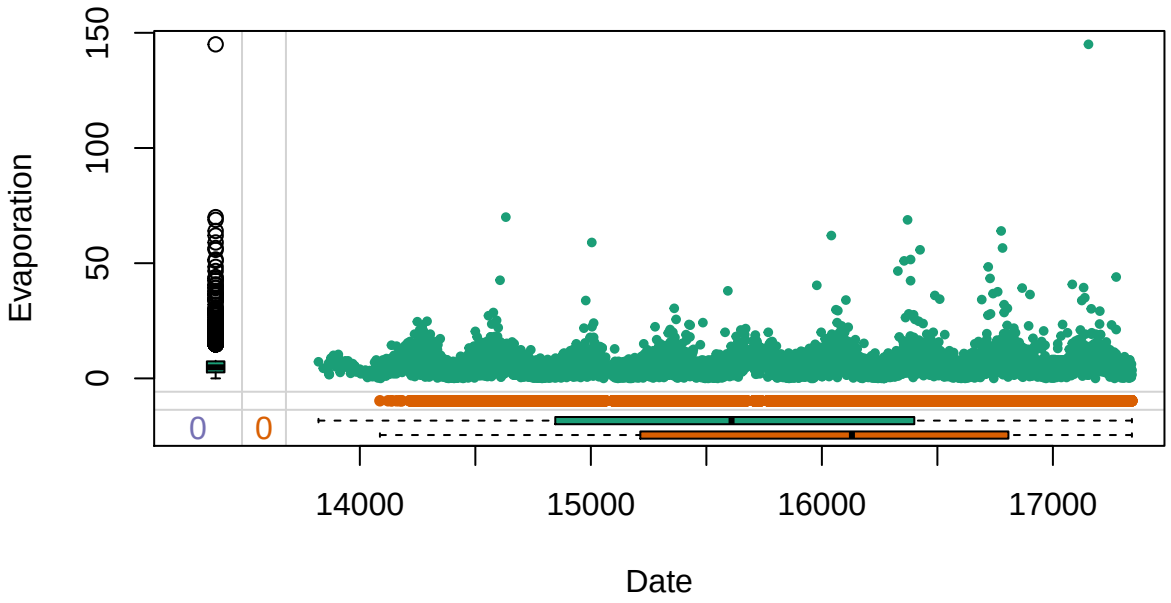

Figure 3: Date/Evaporation Margin Plot

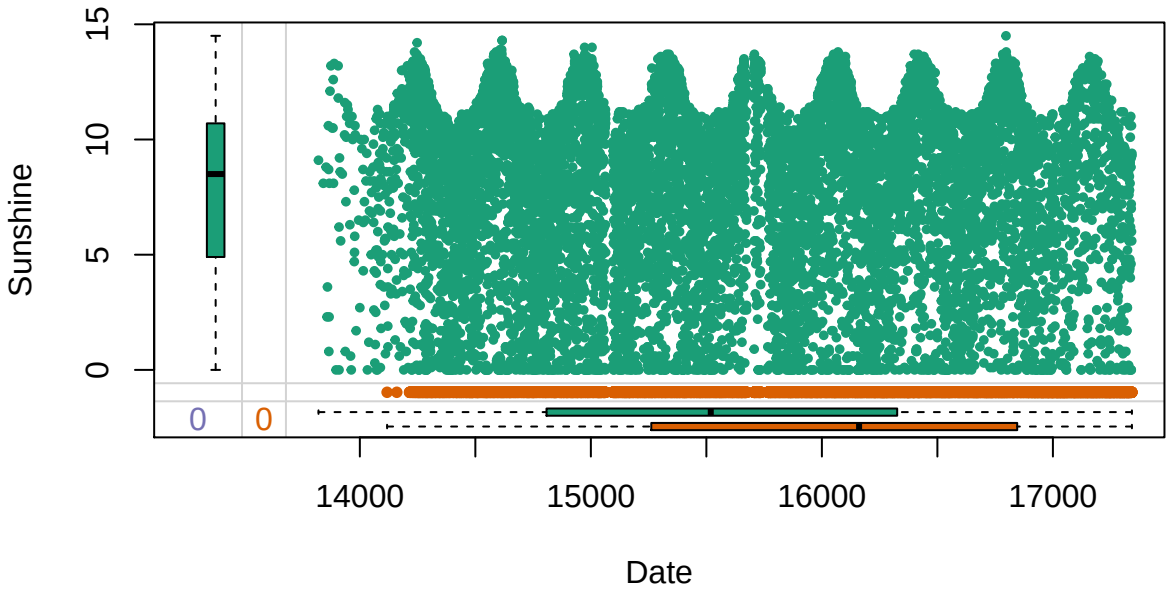

Figure 4: Date/ Sunshine Margin Plot

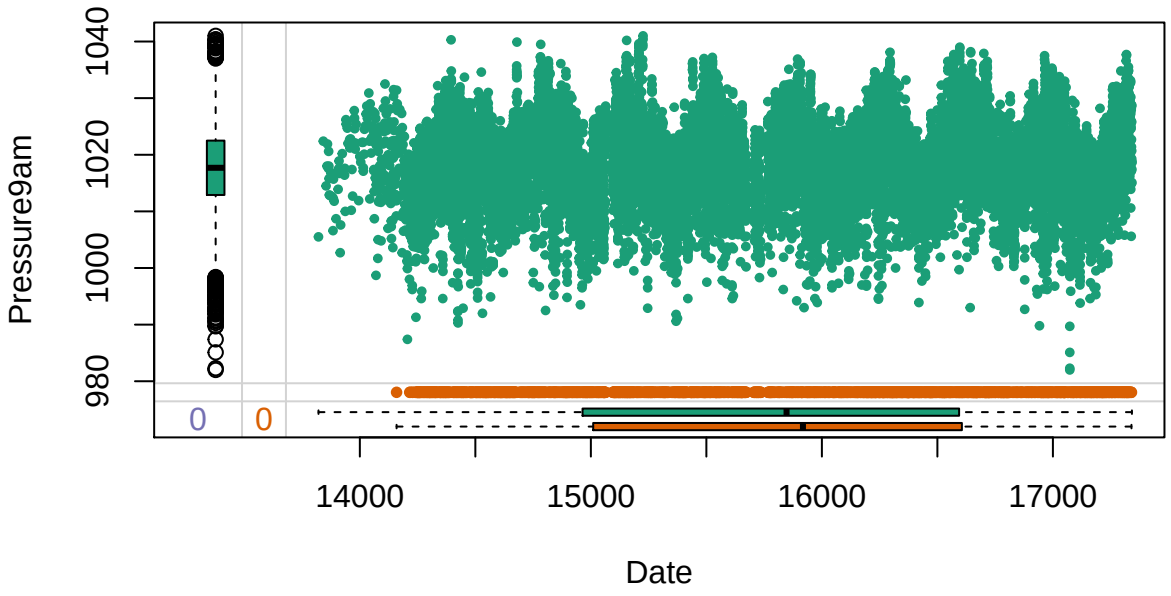

Figure 5: Date/ Pressure3pm Margin Plot

So what do the margin plots tell us? First of all let's take a look at *Date* axis. The *Date* has been converted to number to ensure continuous flow of the data . All features we picked exhibit cyclical pattern as expected. Along the vertical axis we obesrve the box plot of the respective feature. *Evaporaton* data is quite remarkable (Figure **??**); it has very narrow distribution and a lot of so-called outliers. Though forces of nature follow sesonal patters they often exhibit wdie ranhe of seasonal

anomalies, which the plots highlight. The disribution of the missing data of a given feature is depicted along the horizontal axis. In all three cases the missing data is randomly distirubted along observed date range. Along the horizontal axis we may see box plots of the date and a given feature. *Presure9am* ((Figure **??**)) distributed evenly across the observed date frame. *Evaporation* and *Sunshine* exibit more data lossess towards the end of the observed period

Let's examine one more dimention of the missing data, namely featues vs feature vs location

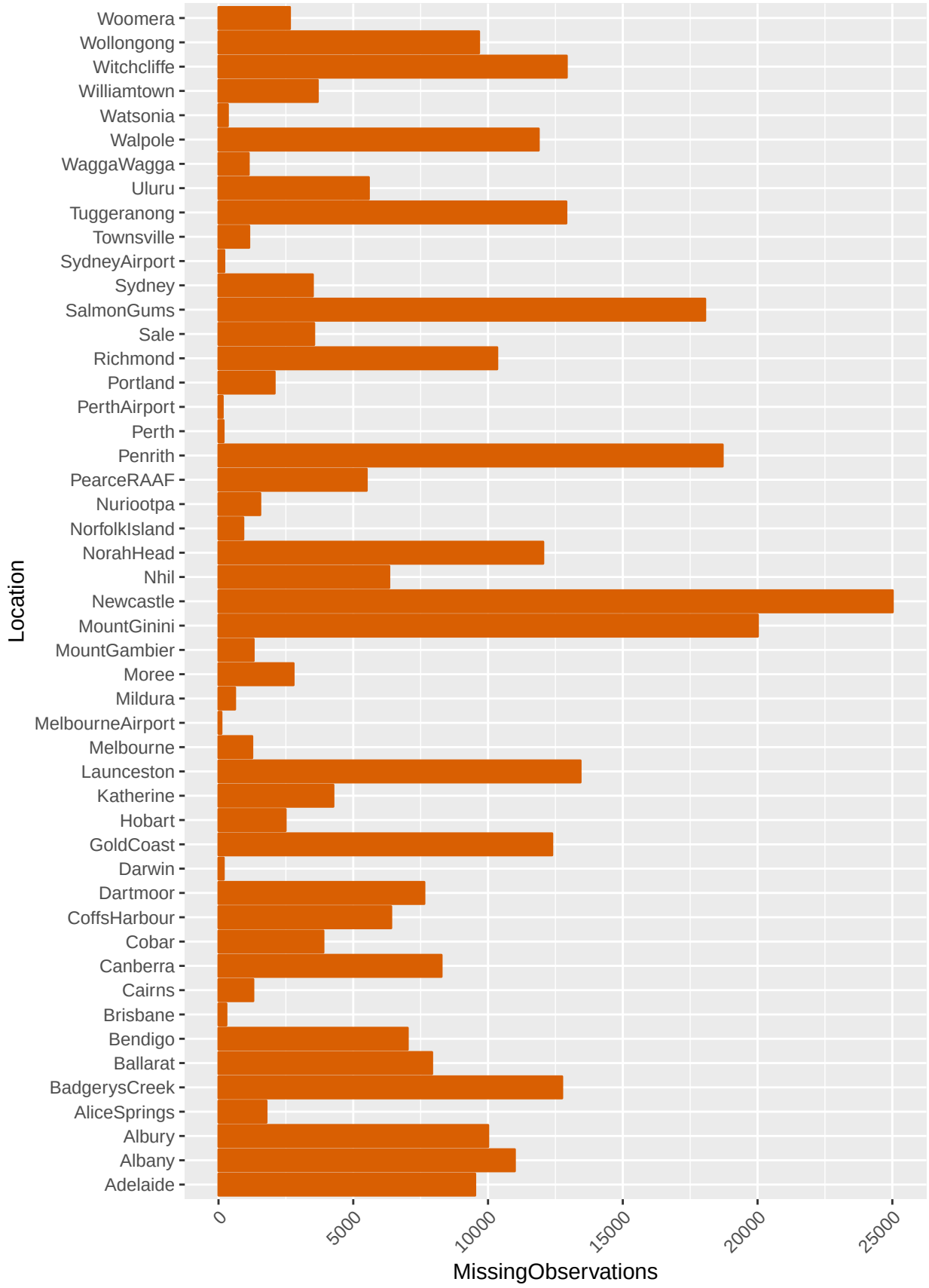

Figure 6: Missing Data By Location

Remarkably Firure 6 shows that **NA** observations are missing on average per location. Though if we take a second look at the weather station map 1 we would see that Mount Gini (the station that

miss the most data), Bendigo and Ballarat are close to Melbrun, where the staff has kept observing data on regular basis. Newcastle to Sydney and so on…

Data correlation and other observations

Let's examine how the features are correlated to each other. Knowing weather we can make an accurate prediction that the temparature features should be highly correlated, as well as pressure, wind speed, clouds and humidity groups

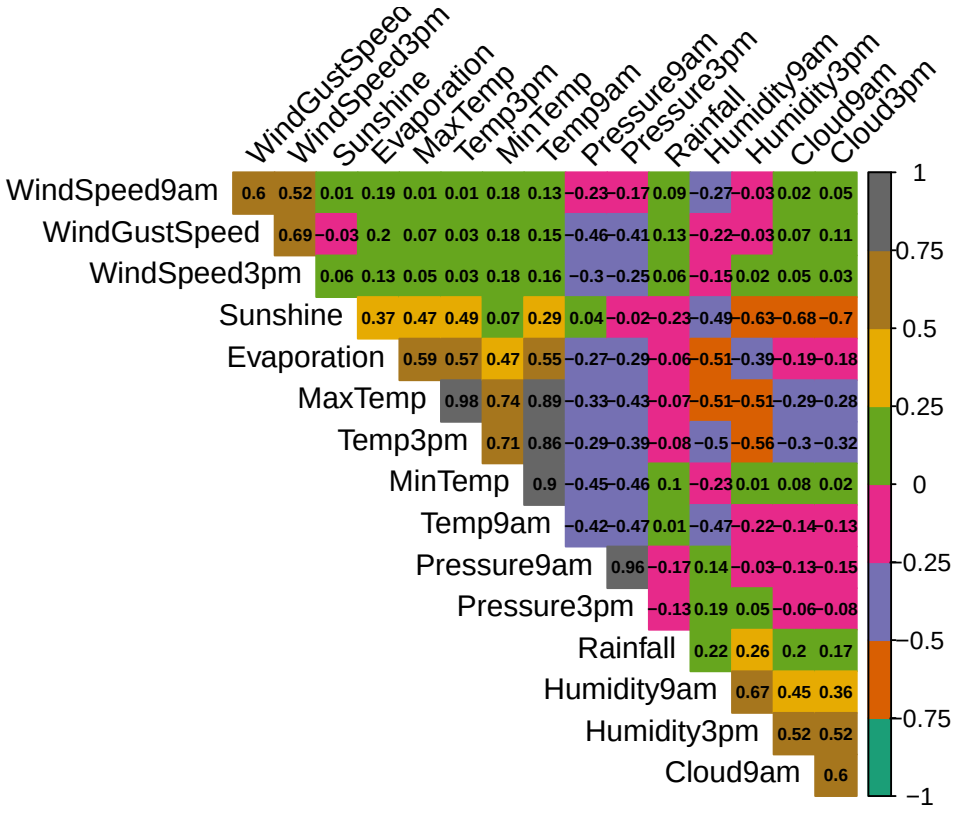

Figure 7: Data Correlation

Figure 7 confirms our intial guess. This observation will help us to elminate redundant features later when we get to the point of selecting useful predictors for our model

Data Preparation

Modeling and Evalutation

Decision Tree Model

Naive Bayes Model

Random Forest Model

Logistic Regression Model

Model Comparison

Model Deployment

Conclusion

Bibliography

Sumaira Afzal
York University School of Continuing Studies

Viraja Ketkar
York University School of Continuing Studies

Murlidhar Loka
York University School of Continuing Studies

Vadim Spirkov
York University School of Continuing Studies