Predicting rain in Australia

Vincent Pluchet

2 June 2021

Contents

1 Executive Summary

1.1 Goal of the study and available data

The goal of this study is to predict **next day rain in 49 locations** in Australia. This challenge was posted by Joe Young on Kaggle at **Rain in Australia**.

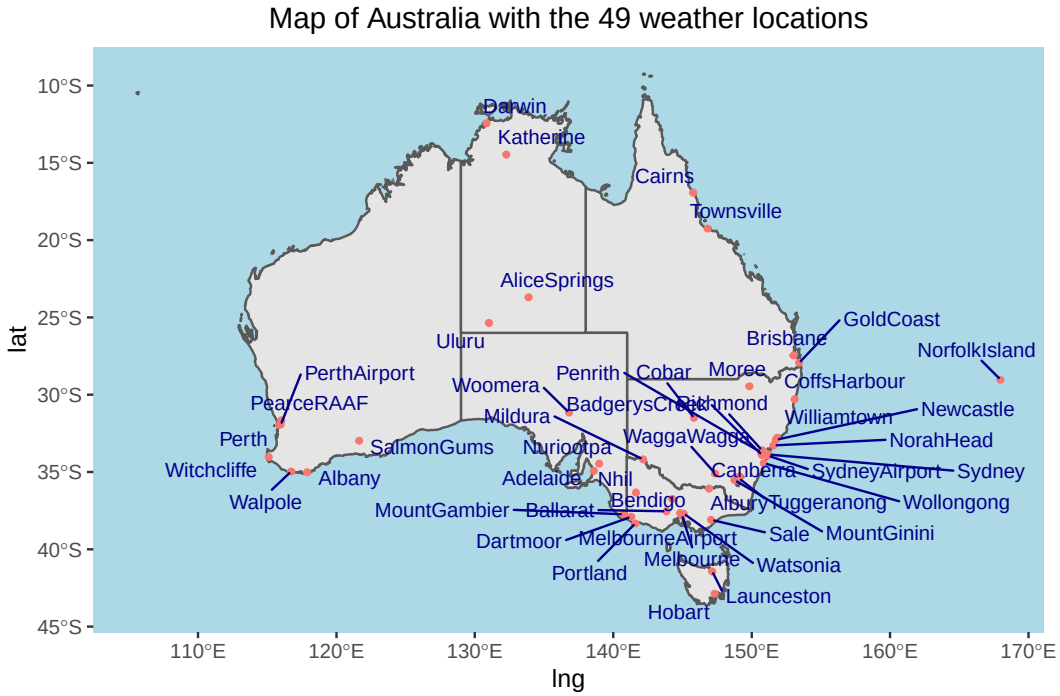
The data set available for the challenge contains about 10 years (of which 8 years are complete, 2009-2016) of daily weather observations from 49 Australian weather stations. RainTomorrow is the target variable to predict. It means: did it rain the next day, Yes or No? The column is Yes if the rain for that day was 1mm or more.

The data set contains daily observations of a number of categorical (like wind direction) and numeric (like temperature, pressure, evaporation) variables for each location. Certain variables are observed twice daily, at 9am and 3pm. The data set also contains the RainTomorrow column for model calibration.

The data set is a csv file called "weatherAUS.csv". It contains overall **145,460** rows and **23** columns.

For the purpose of this report, we also provide a small file containing information about the locations, namely longitude and latitude, which we use for map plots. The file is available on **github**.

The 49 locations are shown on the below map:



1.2 Specific challenges

There are at least three interesting challenges in this study.

- The first challenge is the diversity of the Australian climate. Australia is a huge country with at least 6 main climate zones, as evidenced by the below map posted by the **Bureau of Meteorology of Australia**. Seasonal rainfall behaves very differently according to location, as shown on the **Seasonal Rainfall map**. For instance, climate is very different in Darwin and Sydney. Darwin has significant rainfall in summer and dry winters, whereas Sydney has a relatively uniform rainfall (note: Southern Hemisphere summer is December-February). Any prediction model must therefore take such differences into account.

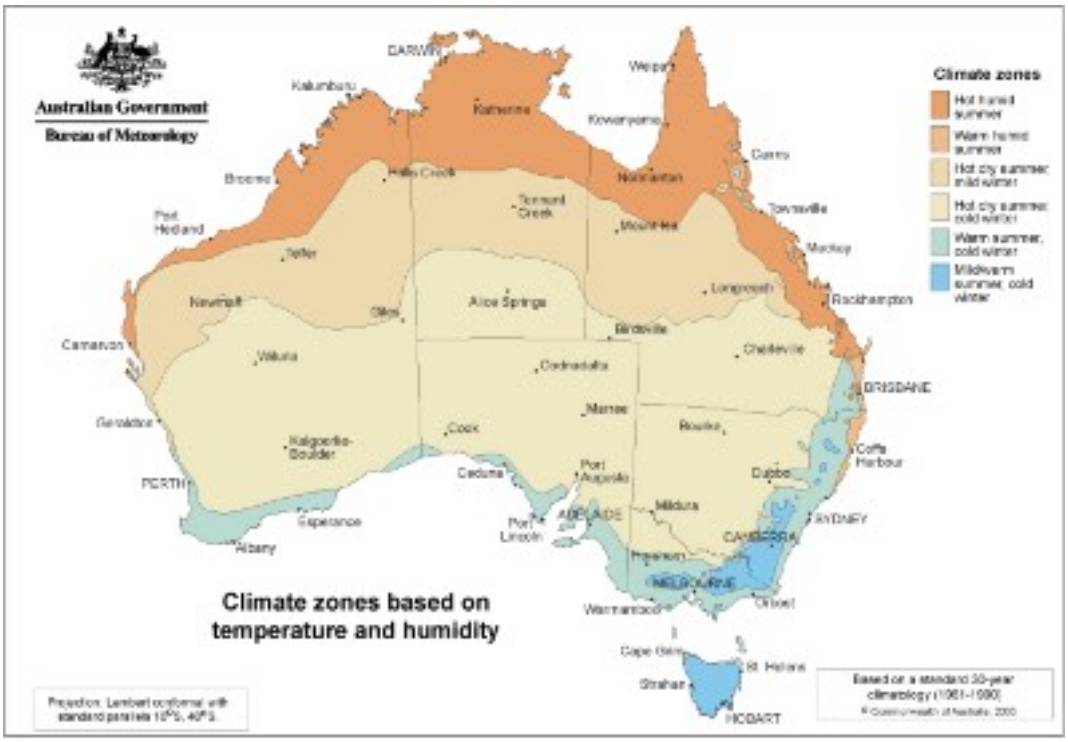


Figure 1: **Australian climates (copyright: Bureau of Meteorology, Australian Government)**

- The second challenge is that rain remains a highly random phenomenon versus the variables measured in the data set. No variable in the set, and no combination of variables, can predict rain with a high degree of certainty. There remains a high random component. Average patterns can be detected but cannot be used in a deterministic way. For instance, we anticipate, on average, that if pressure is low and drops further during the day, this announces a higher chance of next day rain. But, as we will see in the data analysis section, there are many such situations which are not followed by rain, and there are opposite situations followed by rain.
- The third challenge is inherent to the data set itself. Several variables have significant NA levels (Not Available). For instance, Sunshine, a variable which measures the number of hours of bright sunshine in the day, has a non-negligible negative correlation with next day rain: if a day has low Sunshine, the chances of rain are higher. It is therefore a very interesting variable in a prediction model. However in the data set, 48% of the Sunshine data is NA.

"Predict Rain in Australia" is, as a consequence, a very interesting data science challenge.

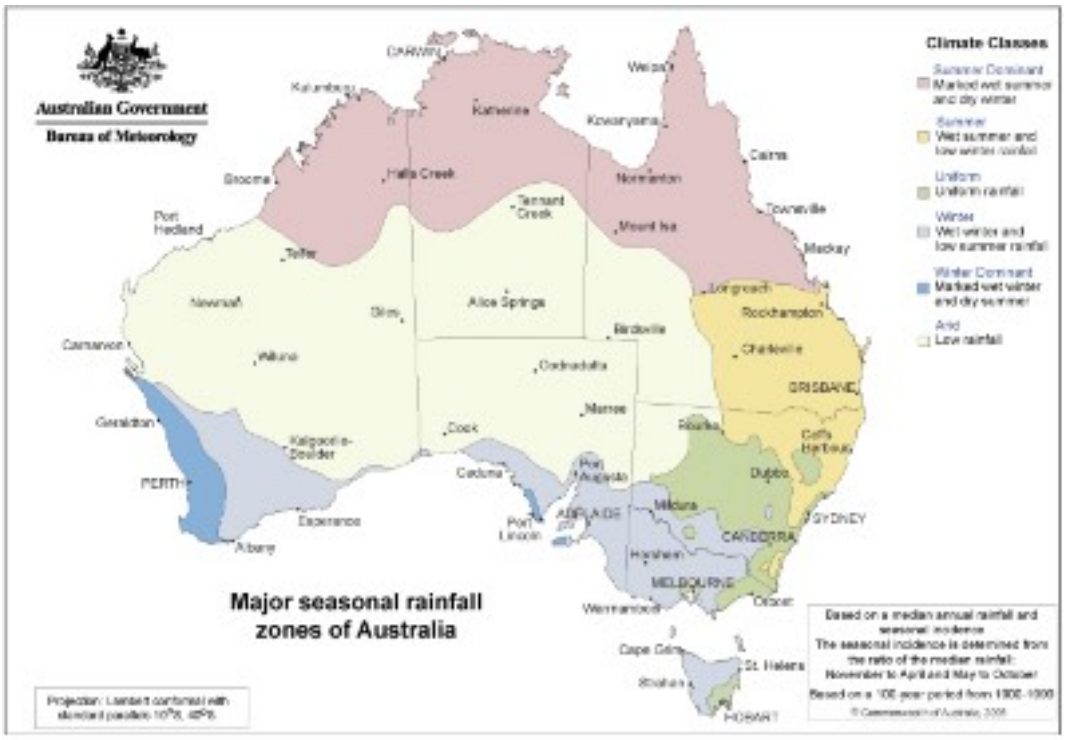


Figure 2: **Australian Seasonal Rainfall (copyright: Bureau of Meteorology, Australian Government)**

1.3 Objective

- weatherAUS is a small data set. It cannot compete with the data sets used by weather specialists worldwide, which use millions of data processed on very powerful servers. The beauty of weatherAUS is that it can be run on a simple laptop and offers a glimpse into the fascinating world of weather forecasting. Weather forecasts nowadays give us very accurate information not only about whether there will be rain or not, but also at what time, how much rainfall should be expected and with what likelihood. We cannot achieve the same performance with a data set which contains, at best, two daily observations.
- Another factor to consider is the high prevalence of "No Rain" in the data. On average, as per the data set, the proportion of days with rain (rainfall $>$ 1mm) is only 22%. This means that a model which would predict "No Rain" would have an accuracy around 78%, which could seem high. This would be a poor model however: it would have a high Specificity (the capacity to correctly predict No Rain Tomorrow) of 100% but a Sensitivity (the capacity to correctly predict Rain Tomorrow) of 0%, since it would always predict No Rain. Therefore its Balanced Accuracy, the average between Sensitivity and Specificity, would be 50%, which is quite low.
- Our objective is therefore to obtain reasonably better Accuracy and Sensitivity levels, without hurting Specificity too much and achieving a relatively good Balanced Accuracy. We consider the Balanced Accuracy rather than an F1 score because we want to give equal weights to Rain predictions and No Rain predictions. We will also assess the predictive values, ie the accuracy of predicting Rain or No Rain.
- We will consider model results per location as well, not just national averages. This is important in order to reflect the diversity of the Australian climate zones and to provide accurate local forecasts.

1.4 Approach and Model Used

- The prediction algorithm for this study was developed using the R programming language.

- For the purpose of model selection and validation, the data set was split between a training data set (called "weather"), containing around 80% of the data, and a validation set (called "validation"), containing around 20% of the data. These proportions were chosen in order to have enough data for model calibration, given the number of locations and variables involved, while still retaining a validation set of significant size. The validation set was kept completely independent from the training set, and was not used at any stage of the model preparation, training or selection. It was only used with the final selected model, in order to measure the performance of the model on an entirely new data set.
- For model training and cross-validation, the "weather" training set was itself partitioned into weather_train (80%) and weather_test sets (20%).
- The NA data was populated using the median for each Location-Month combination. Indeed, given the diversity of the locations, we did not want to populate NAs with global averages, except where necessary due to lack of data. We also did not want to use yearly averages, as there are clear seasonality trends in the data (rainfall in Summer is different than in Winter). For categorical data, the mode was used (the most frequent category value) on a similar basis. We did **not** use any of the "validation" data to populate the NAs, even in the validation set. We performed a small analysis during training to see if the NA replacement method used was distorting the results: no significant distortion was observed.
- New variables were added, replacing some of the provided data. As an example, the data set contains Temp9am and Temp3pm, the temperature measured at 9am and 3pm. These two indicators are highly correlated. It is therefore more interesting to compute, as a new variable, the difference between Temp9am and Temp3pm: thus we can keep Temp3pm and the temperature variation, which provides a trend (is temperature increasing or decreasing). This did improve a bit the performance of the algorithm by removing excessive collinearity.
- Several algorithms were tested and compared in the study, including:
  - simple deterministic models
  - GLM using various predictors or combinations of predictors
  - Random Forest
  - XG Boost, an implementation of gradient boosted decision trees designed for speed and performance
  - QDA, quadratic discriminant analysis
  - KNN, k-Nearest Neighbors
  - An ensemble of the 5 above methods (GLM, RF, XG, QDA, KNN)
  - An ensemble of 3 methods (XG, GLM, QDA)
- Where appropriate, comparisons were made between "global" models (one model for all locations) and "local" models (one separate model for each location). The local models usually perform better than the global models, as they are better capable of handling the specific patterns of each location.
- We assessed 4 primary indicators for each model:
  - Accuracy: accuracy of overall Rain & No-Rain predictions
  - Sensitivity: accuracy of predicting actual rain events
  - Specificity: accuracy of predicting actual no-rain events
  - Balanced Accuracy: average between Sensitivity and Specificity
- For GLM, we used p_values, VIF (variance inflation factor) and ANOVA to analyze the significance of the variables used and the robustness of the model (more details are provided in the model development section).
- Interestingly, advanced algorithms (ie all the above models excluding the simplest ones) all performed within a range of similar results during the training phase: accuracy differences on the training test set were below 4 points, which is not negligible but not huge either and a proof of consistency. Overall results, as well as results by location, still identified very clearly that the best performing models were GLM and XG Boost (before introducing the ensembles). Then the ensembles were created to maximize results by combining the results from the various models.
- **The final selected model is the ensemble of three methods: XG Boost, GLM and QDA**, with predictions based on the majority vote between the three algorithms. This model had the best

performance on Accuracy, Sensitivity and Balanced Accuracy, globally and at location level. GLM and QDA are applied locally (one model per location) following a Principal Component Analysis. XG Boost is applied globally, using longitude and latitude to replace the locations.
- Whilst not strictly necessary, the Principal Component Analysis (PCA) allows to deliver more robust models, with de-correlated and significant predictors. PCA does not improve the overall accuracy (and should not be expected to do so as it is essentially a re-combination of the original predictors).

1.5 Results

- The selected model "Ensemble 3" was tested on the independent validation set
- The detailed results are presented in the Results section, with the key points highlighted here.
- The overall Accuracy is 86.2%, ranging 77%-95% between the various locations. 3 locations out of 49 have an accuracy between 77-80%, 18 are between 80-85%, 18 between 85-90% and 10 above 90%. This is a very good result, showing that the model addressed the diversity of the various locations.
- The Sensitivity is 58.2% (accuracy of predicting actual rain), ranging 30-78% by location (however only 4 locations are below 50%). This is the weak point of the modelization and is not specific to the validation set. All models during the training phase had low sensitivity, QDA having the best result at 60% on the test set but poorer accuracy than other models. Our interpretation is that there remains significant randomness in the rain outcome versus the data measured in the data set: scatterplots in the data analysis section illustrate this point.
- The Specificity is 94.4% (accuracy of predicting actual no-rain), ranging 88%-99% by location. 46 locations are above 90%. This is a strong result.
- The "No Rain" predicting value is 88.6%, meaning that when the model predicts No Rain, the prediction is true in 88.6% of the cases.
- The "Rain" predicting value is 75.1%, meaning that when the model predicts Rain, the prediction is true in 75.1% of the cases.
- The Balanced Accuracy is 76.3%.
- These results are in line with our stated objectives, although Sensitivity remains on the low side. The algorithm performs best in the central locations, where the weather is more stable, less well in the coastal locations of Southern Australia, where the oceanic influence is strong and probably induces significant weather variability. The Model Development and Results sections provide additional information on this.
- In the Conclusion section, we try to highlight additional studies which could further improve the results.

2 Data Analysis

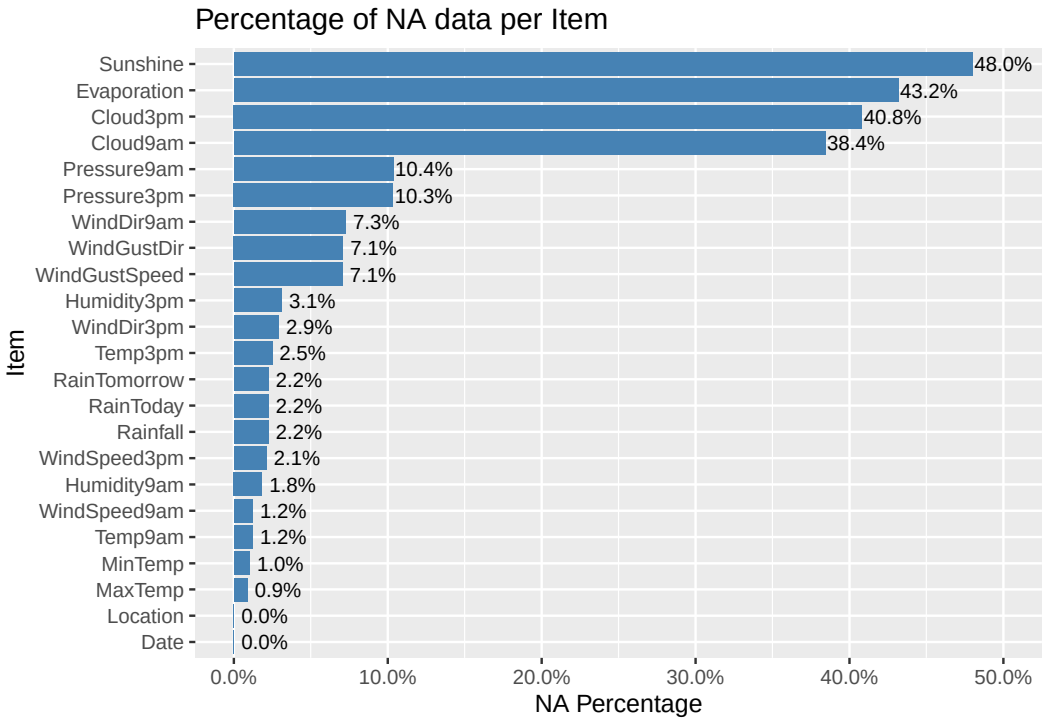
This section contains three main parts:

1. A preliminary review of the data set, to understand where data is missing, if there are specific measurement errors that would need to be taken care of, and if certain elements should be removed from the data set ahead of analysis and modeling. At the end of this part, certain non-useful rows are removed from weatherAUS, and the training set "weather" as well as the "validation" set are created by partitioning weatherAUS. From this point onward, the validation set is set aside and not used, any further analysis is made on the training set only.
2. Further data exploration, to understand better the behavior and influence of the various predictors, namely: the location effect, the time effect, the correlation between the predictors and the capacity of each predictor to provide information about whether there will be rain tomorrow or not.
3. The last part consists in preparing the training set for the modelization part, computing the NA replacement values, computing some additional variables to be used in the analysis, and splitting the training set into weather_train and weather_test for model training and cross-validation purpose.

2.1 Preliminary data review

2.1.1 NA Percentages per predictor

We plot the percentage of NA data for each predictor. The plot shows Sunshine, Evaporation, Cloud3pm, Cloud9am having high NA percentages. There are no NAs for either Date or Location. Other indicators show relatively low percentages of missing data.



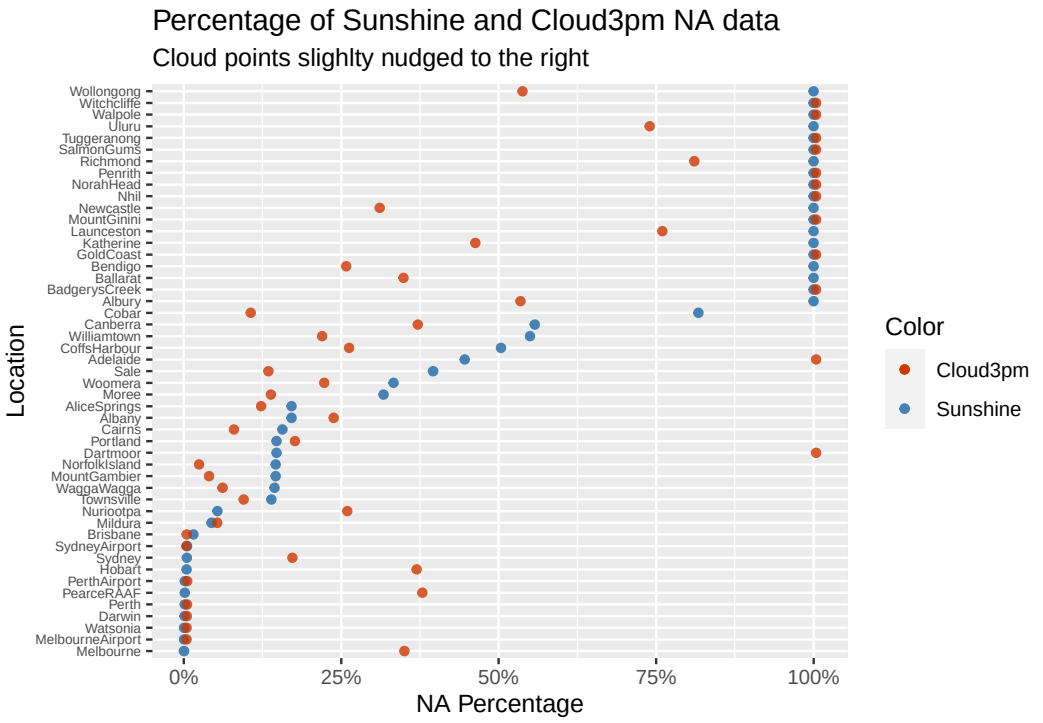
2.1.2 NA Percentages per Location and predictor

We use the nianar package which provides an efficient graph. Whilst certain variables are missing for specific locations, the plot does not evidence any location with a significantly higher proportion of missing

data. Therefore all locations can be used in the analysis. However, we will look at Sunshine and Cloud3pm in a more detailed way to see where they are missing (see next part).

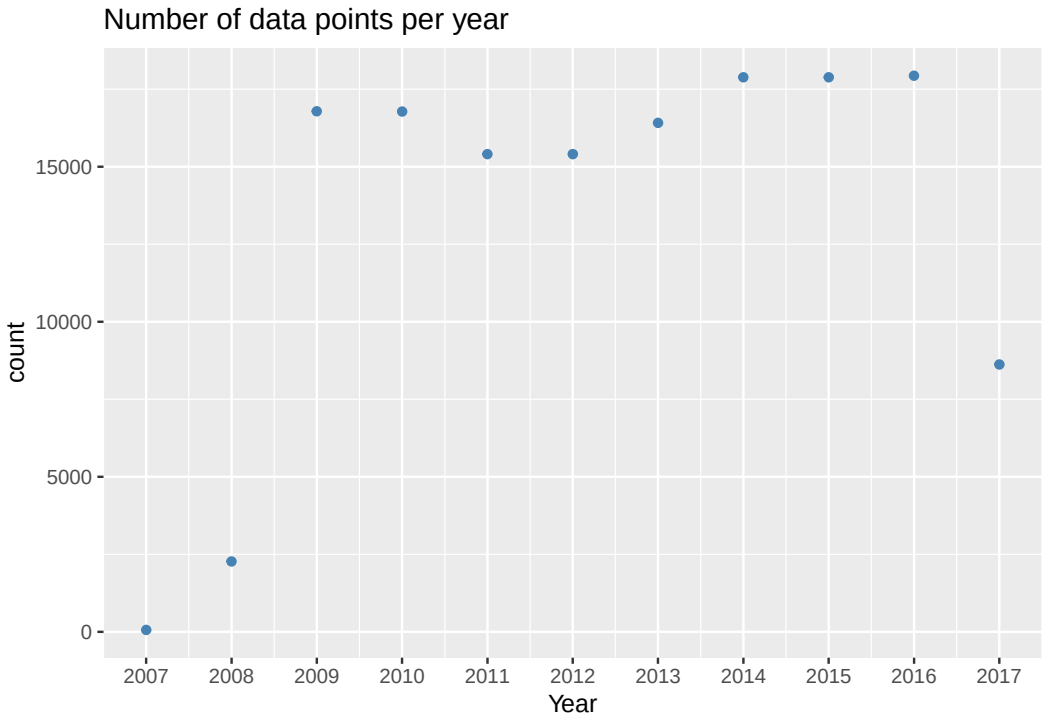
2.1.3 Focusing on Sunshine and Cloud3pm

The graph shows that around 7 locations only have full Sunshine and Cloud3pm data (0% NA). Other locations like Wollongong have partial Cloud3pm data but no Sunshine. Certain locations like Walpole have no data on Sunshine or Cloud3pm. Several locations have partial data on both predictors. So we cannot have a simple split of locations between those which have the data and those which don't. We will need to find a way to handle the NA data.



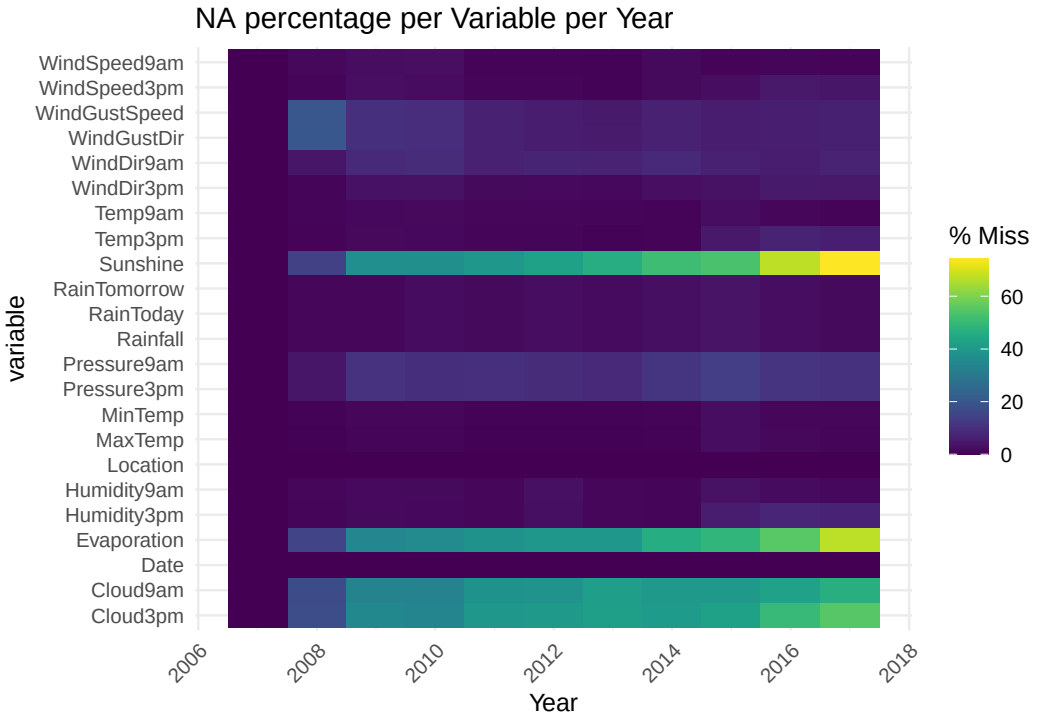
2.1.4 Number of data points by year

The plot shows that years 2007-2008 and 2017 are incomplete. We will therefore remove these years in data pre-processing



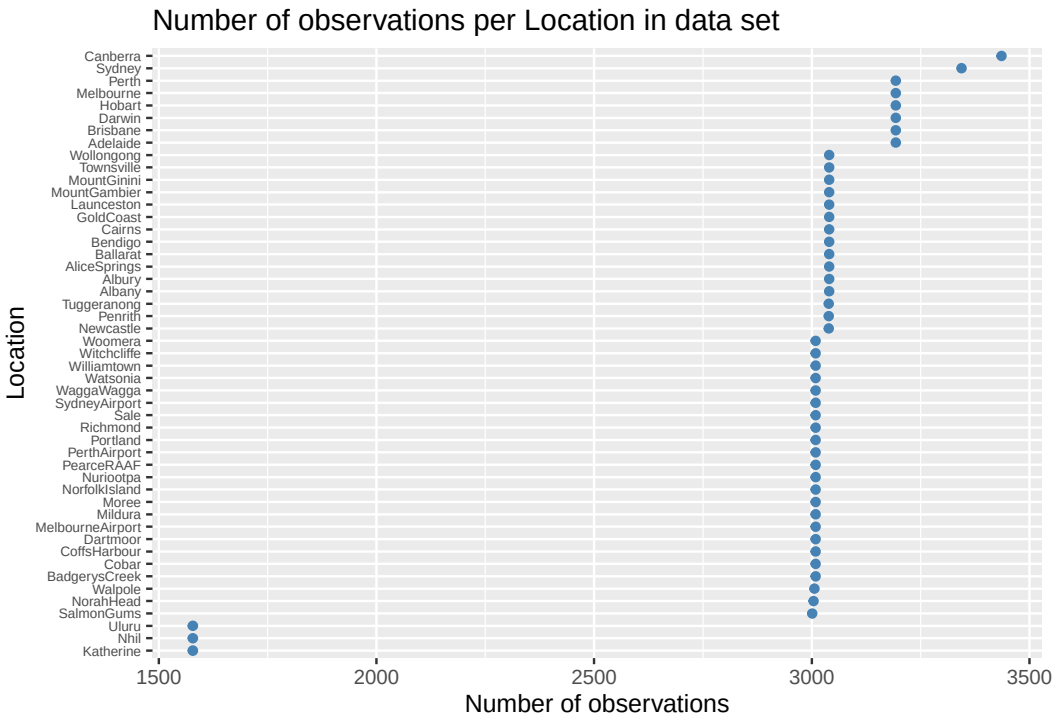
2.1.5 Exploring NAs per descriptor over time

We use the naniar package for easier graphical representation. The plot shows that data collection has been relatively consistent over time



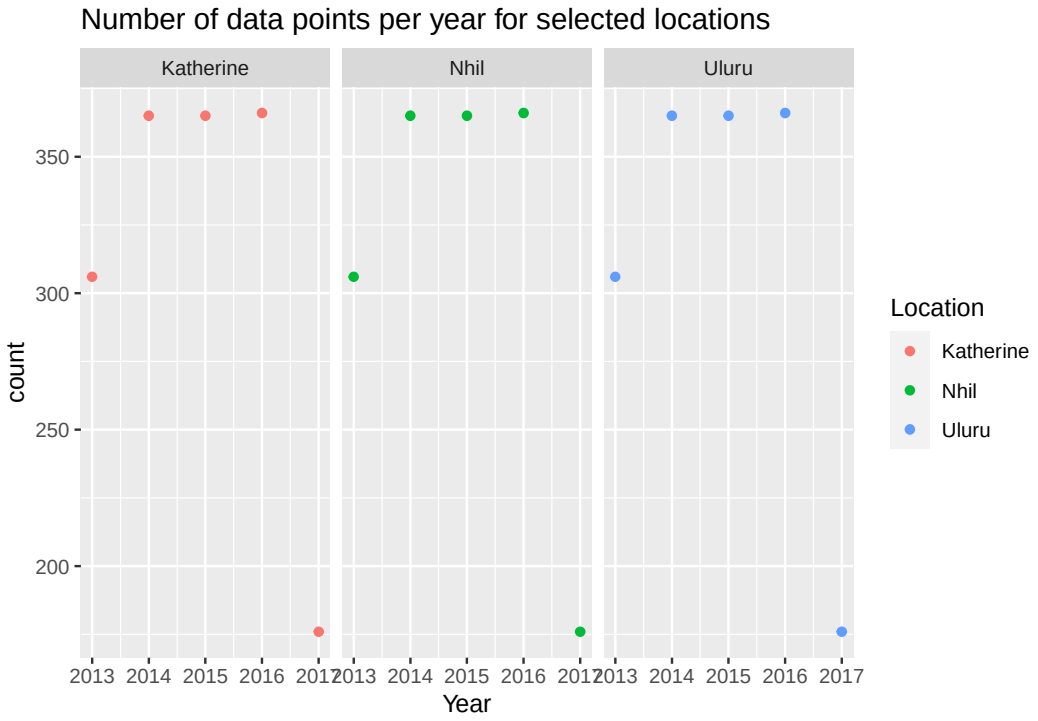
2.1.6 Number of observations per Location

The graph shows three locations that have a much lower number of records. This requires further exploration.



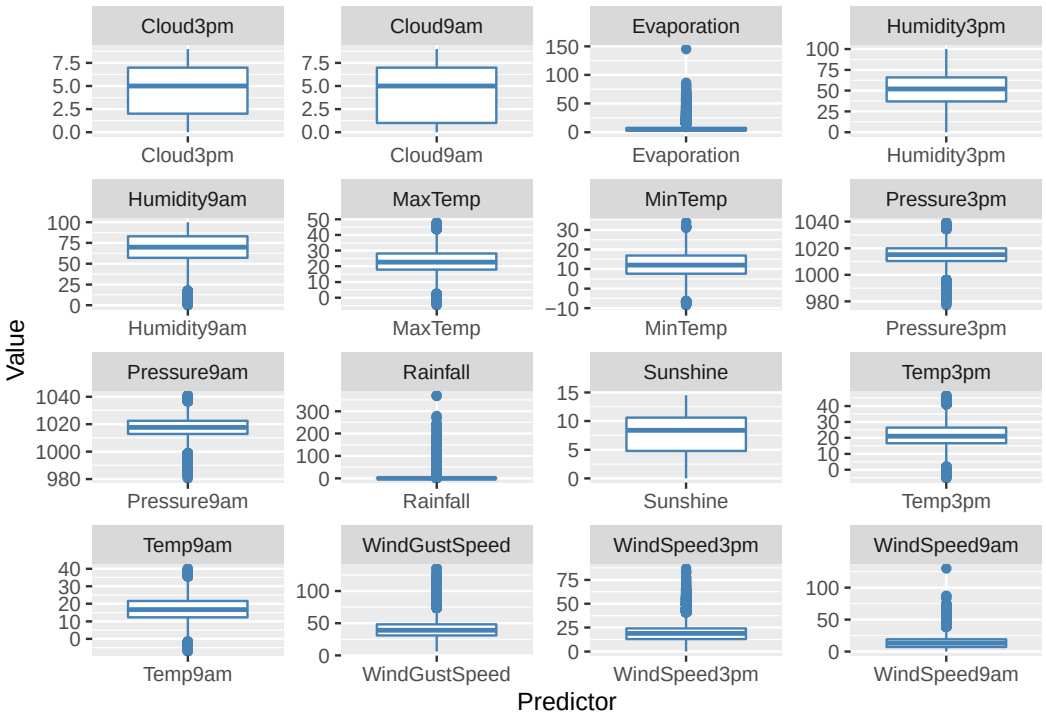
2.1.7 Zooming into the three locations with few records

The graph shows that records started in 2013 for these locations. This is not an issue in itself and therefore the data can be kept. 2017 is incomplete but this year will be removed



2.1.8 Box-Plot of numeric items to detect outliers / unusual data points

The plot shows some significant dispersion for some predictors. A few predictors show a heavy weight of zero values. There are outliers, however no obvious erroneous values. Therefore no specific value restatements appear necessary.



A note on Rainfall and RainToday:

Rainfall is greater than 0 in 36% of the cases only. Rainfall is greater than 1 in 22% of the cases only, indicating that rainfall is generally rare.

We also check that RainToday is identical to Rainfall >1: TRUE.

There is no need for fully correlated predictors, therefore we will remove either Rainfall or RainToday in the study.

2.1.9 Preliminary review conclusion

Based on the above analysis, we will only apply two pre-processing steps before creating the Training and Validation sets:

- We retain only the complete years 2009-2016 (8 years)
- We remove rows with NA in RainTomorrow: as this will be our benchmark to measure the accuracy of the predictions, we cannot keep rows where this indicator is missing

We also convert the categorical variables to factors so that they can be used by the algorithms.

The proportion of retained data for the analysis is 90%.

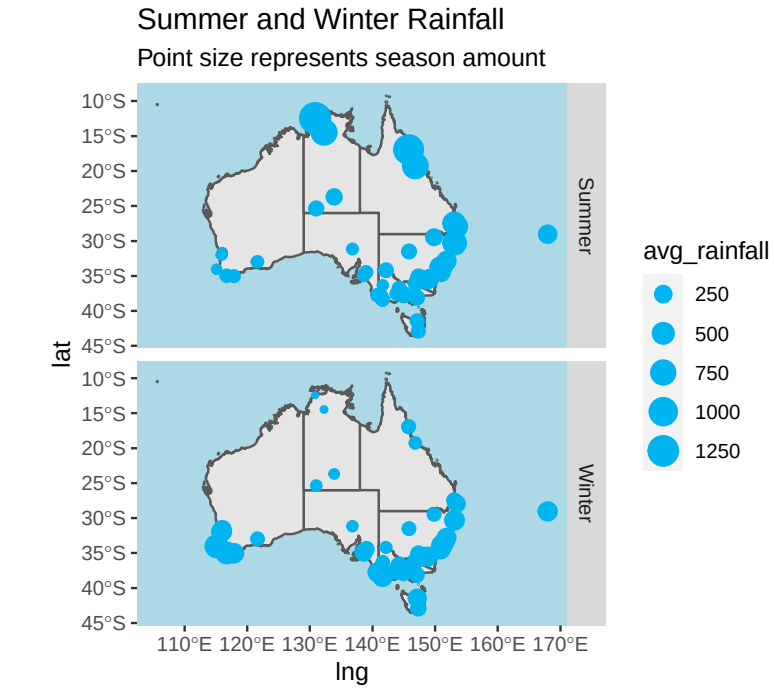
After the above, we randomly split the data set between the training set "weather" and the validation set "validation". Only the weather set is used for further data analysis, NA replacement, model training & cross-validating, and model selection.

2.2 Location and Time effects

After creating the training data set "weather", we try to understand better the behavior and influence of the various predictors, namely: the location effect, the time effect, the correlation between the predictors and the capacity of each predictor to provide information about whether there will be rain tomorrow or not.

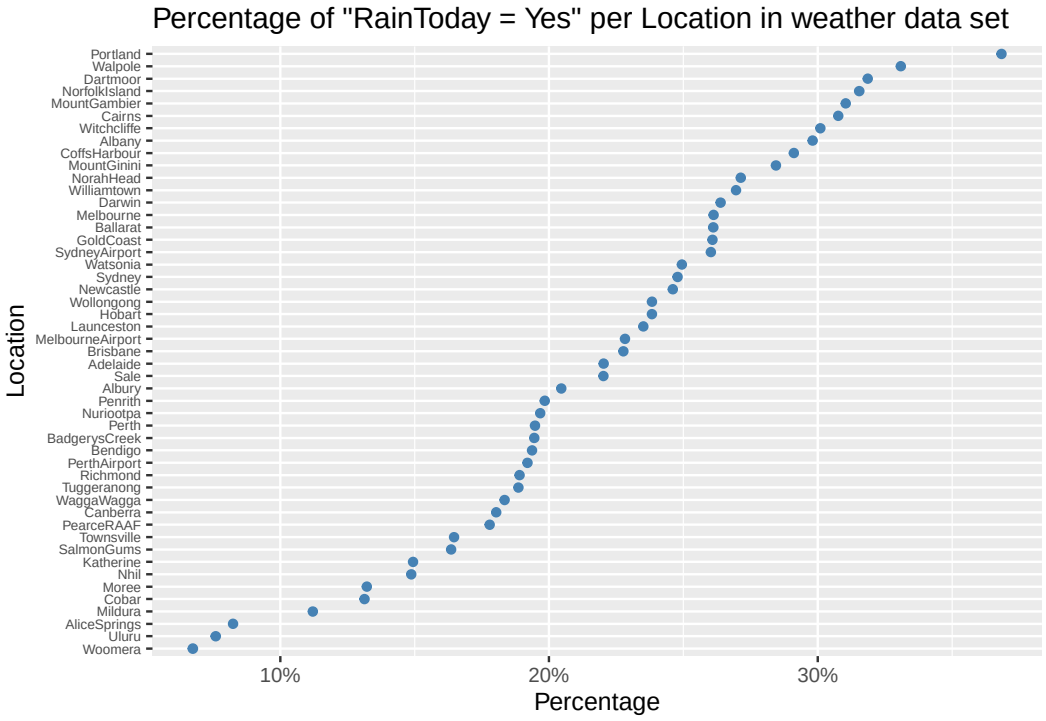
2.2.1 Seasonal rainfall per Location

The map shows the diversity of the Australian rainfall. The north of Australia receives significant rain in Summer and relatively little in Winter. In the south, rainfall is more balanced between both Seasons but it rains more in Winter. Not surprisingly, coastal locations get more rain. Central locations are quite dry.



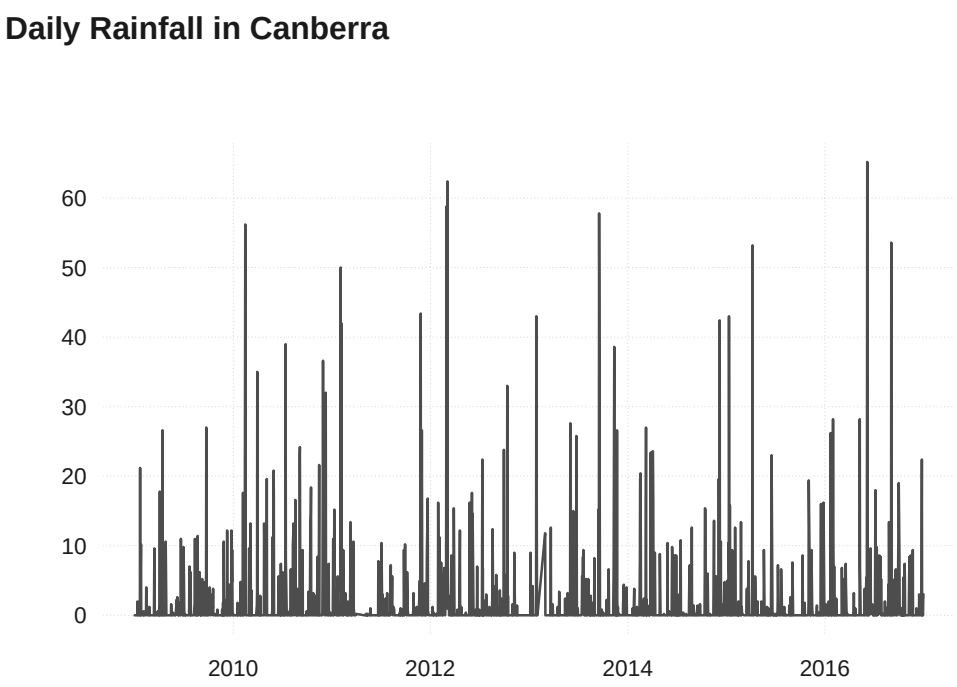
2.2.2 Average percentage of rainy days per Location

The plot shows significant differences in terms of percentage of rainy days per location, as anticipated from the previous map.



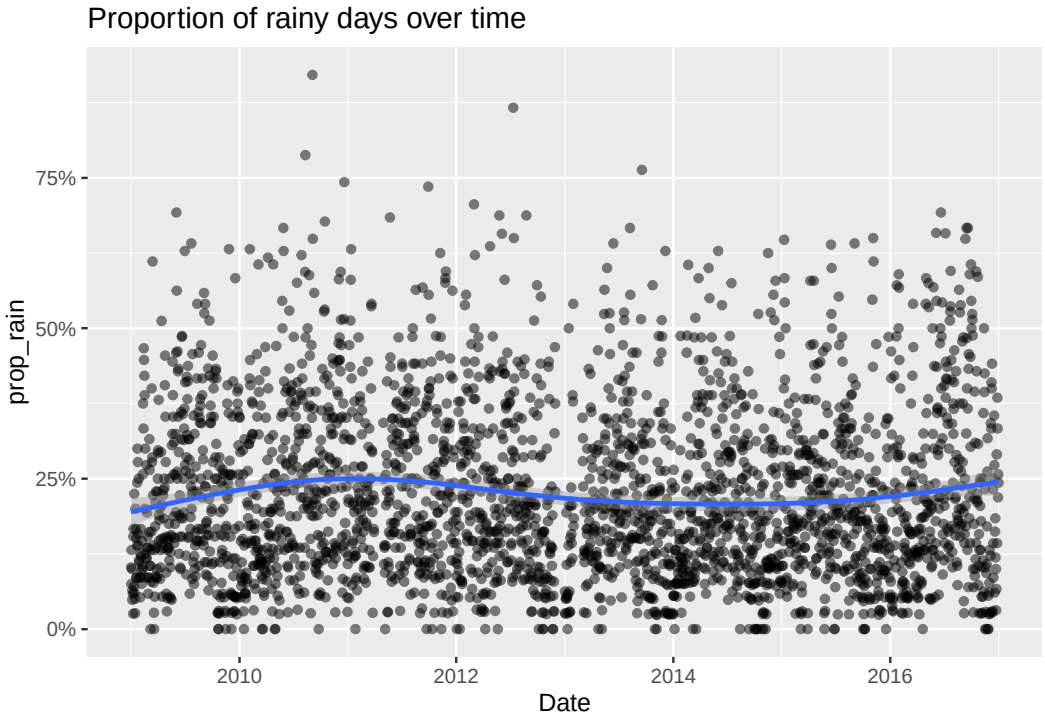
2.2.3 Seasonality component

We start by focusing on rainfall in Canberra. The plot allows to anticipate some seasonal effect. We also used the ts_trend function from the tsbox package, which confirmed that there is some time effect, which needs to be explored further.



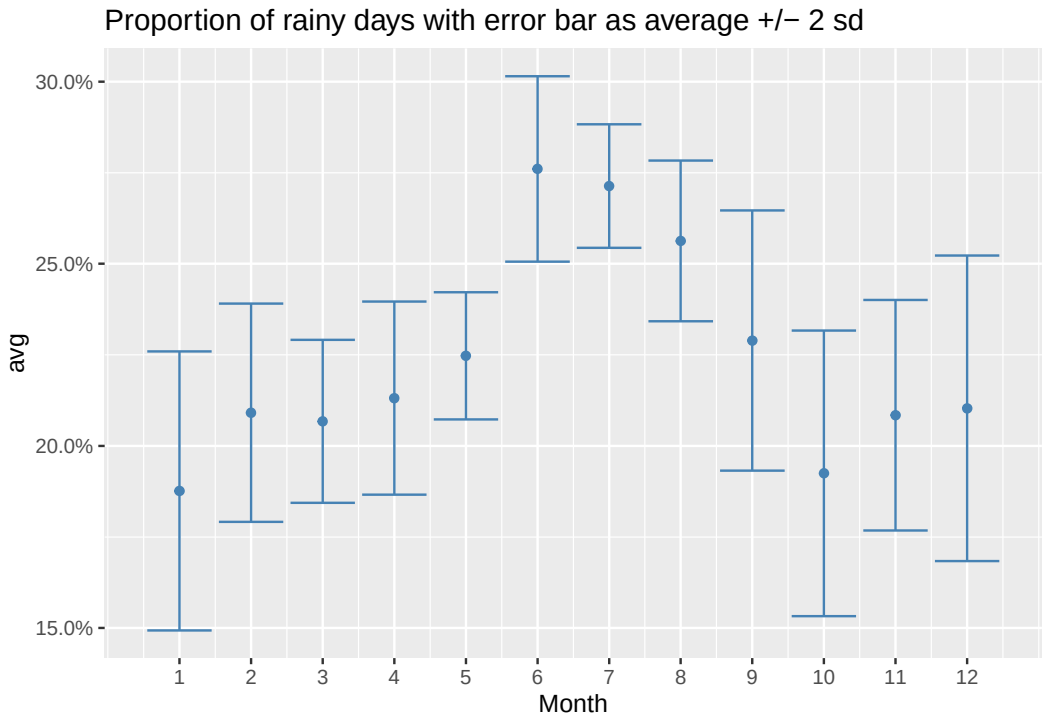
2.2.4 Year effect

We now try to understand if there is a Year effect and a Month effect. We start with the Year effect, by plotting the proportion of rainy days over time. The plot shows a moderate Year effect



2.2.5 Month effect

We now look at the Month effect, by looking at the proportion of rainy days with error bars per month. There is a definite Month effect, June and July are the wetter months. However variations are significant.



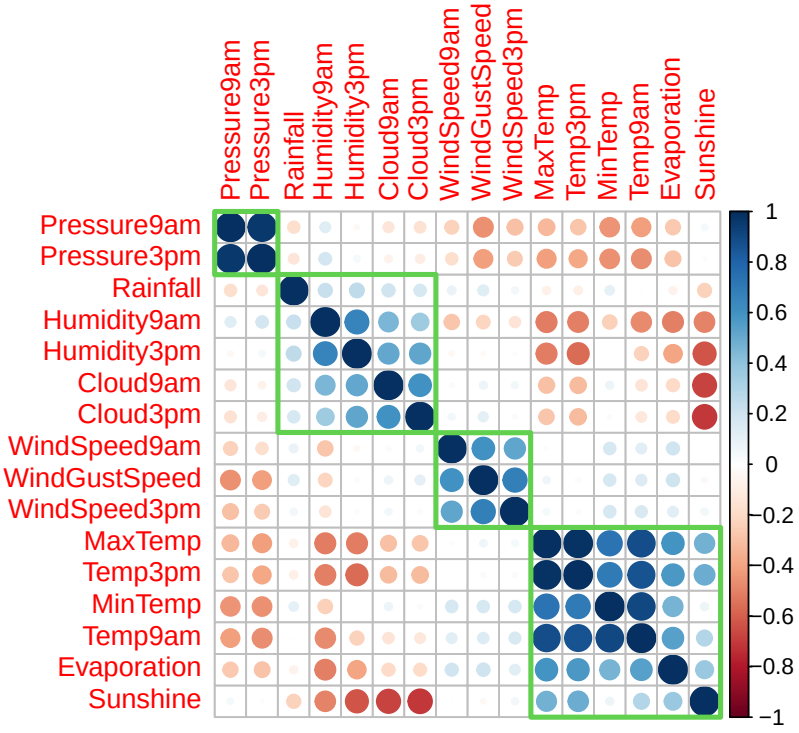
2.2.6 Location and Time conclusion

This section has confirmed that there is a strong Location effect, which will need to be factored in the model. There is also a clear evidence of a Month effect. The Year effect is weaker but will be tested.

2.3 Assessing correlations between predictors

2.3.1 Correlation between numeric variables

The plot shows that a few variables are highly correlated (like Pressure9am and Pressure3pm). In regression, "multicollinearity" refers to predictors that are correlated with other predictors. We will need to be mindful of this when building the model and create predictors with less correlation. High multicollinearity can be a major problem because it increases the variance of the regression coefficients, making them unstable. Having said that, we can also see many factors which are not highly correlated between each other.



2.4 Correlations between predictors and RainTomorrow

2.4.1 Correlation between RainToday and RainTomorrow

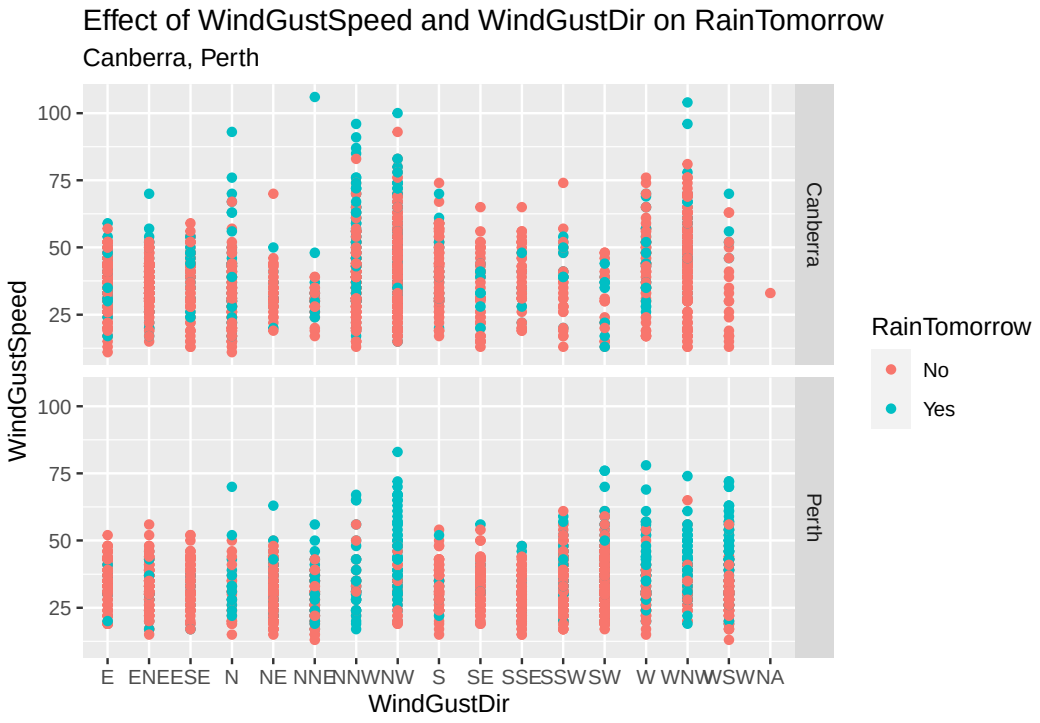
The table below confirms that if it does not rain today, rain tomorrow is unlikely (15% chance), but if it rains today, the chances of rain tomorrow are 46% which is twice the average probability of rain (23%).

Table 1: RainTomorrow vs RainToday

RainToday	Rain_Tomorrow	No_Rain_Tomorrow
No	15%	85%
Yes	46%	54%

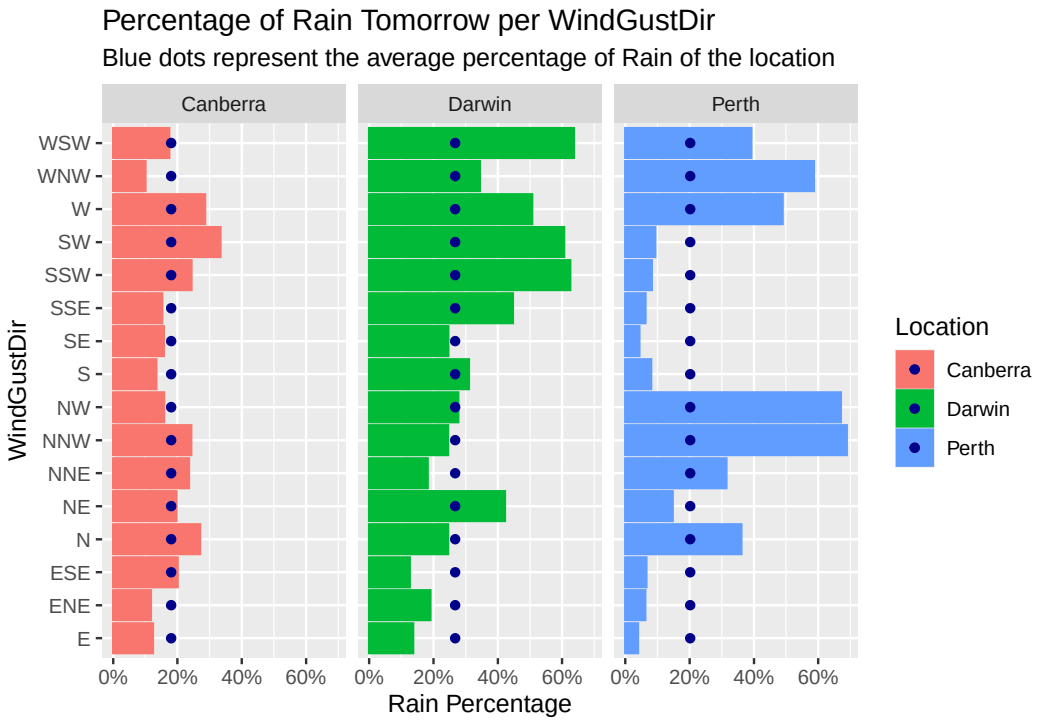
2.4.2 WindGustDir and WindGustSpeed

We will focus on Canberra and Perth. The plot shows that the higher the WingGustSpeed, the higher the chance of rain tomorrow (usually but not always). There is also an indication that WindGustDir plays a role, especially for Perth. WindGustDir is the direction of the strongest wind gust in the 24 hours to midnight and we can understand that it plays a role.



To evidence further the effect of WindGustDir in the chosen locations, we will plot the percentage of RainTomorrow vs WindGustDir. We first compute the average percentage of rain in each location which will also be displayed in the graph. We add Darwin too.

The graph below shows the percentage of RainTomorrow, depending on the WindGustDir. For Darwin and Perth, there is a clear effect of Northern and Western winds, implying greater chances of rain tomorrow. For Canberra, the effect of WindGustDir is less obvious. In any case, this confirms that both WindGustDir and WindGustSpeed are significant variables.



2.4.3 Chi-squared test on categorical data

Pearson's chi-squared test is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. It tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution.

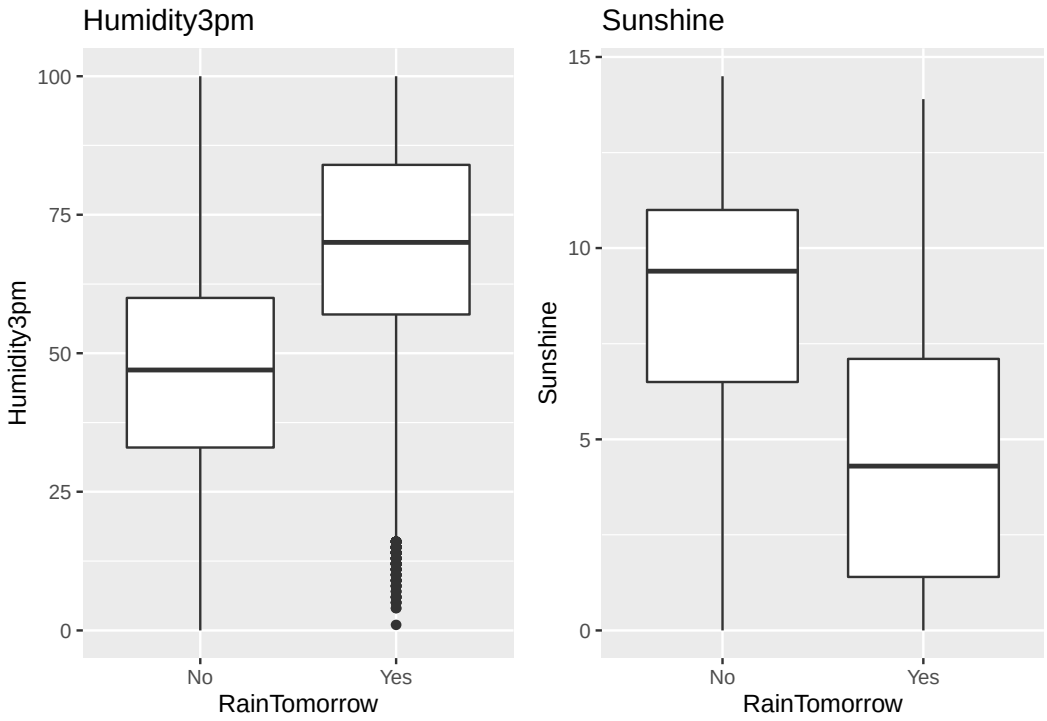
We perform the test versus each categorical variable in the data set. The chi-squared test confirms that we can reject the hypothesis that the categorical variables and RainTomorrow are independent.

Table 2: p-values for corr between categorical variables and RainTomorrow

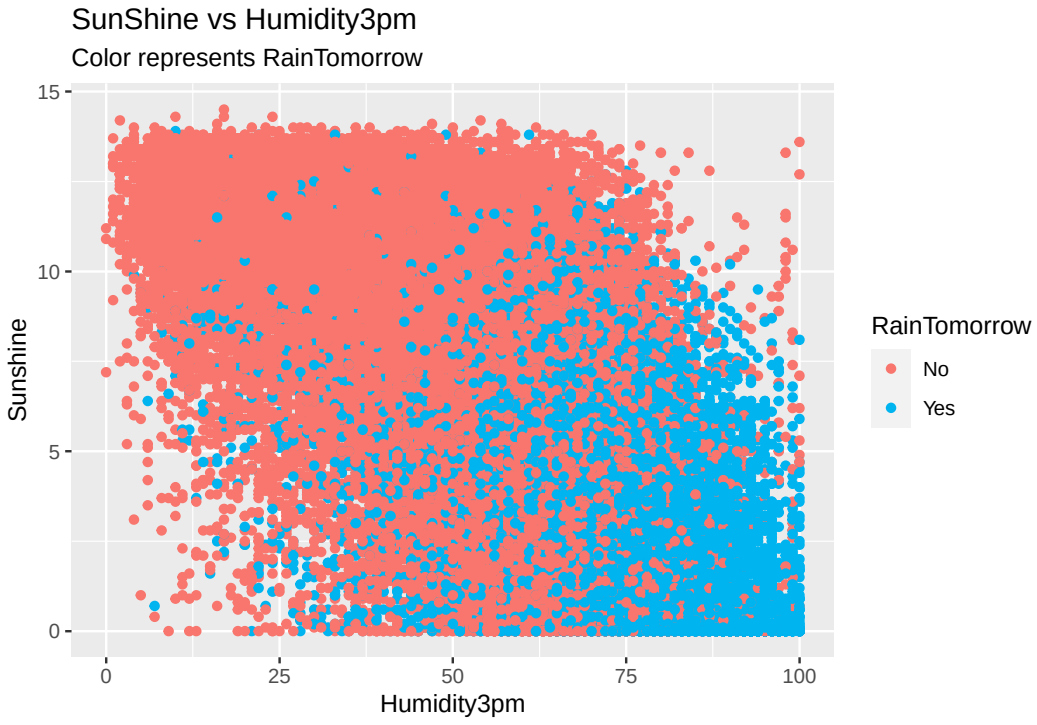
	X_squared	p_value
WindGustDir	1196.34153007083	9.99321947965388e-246
WindDir9am	1739.6930579889	0
WindDir3pm	1012.04984658384	3.52173095591408e-206
RainToday	10169.176770218	0

2.4.4 Humidity3pm and Sunshine versus Rain Tomorrow

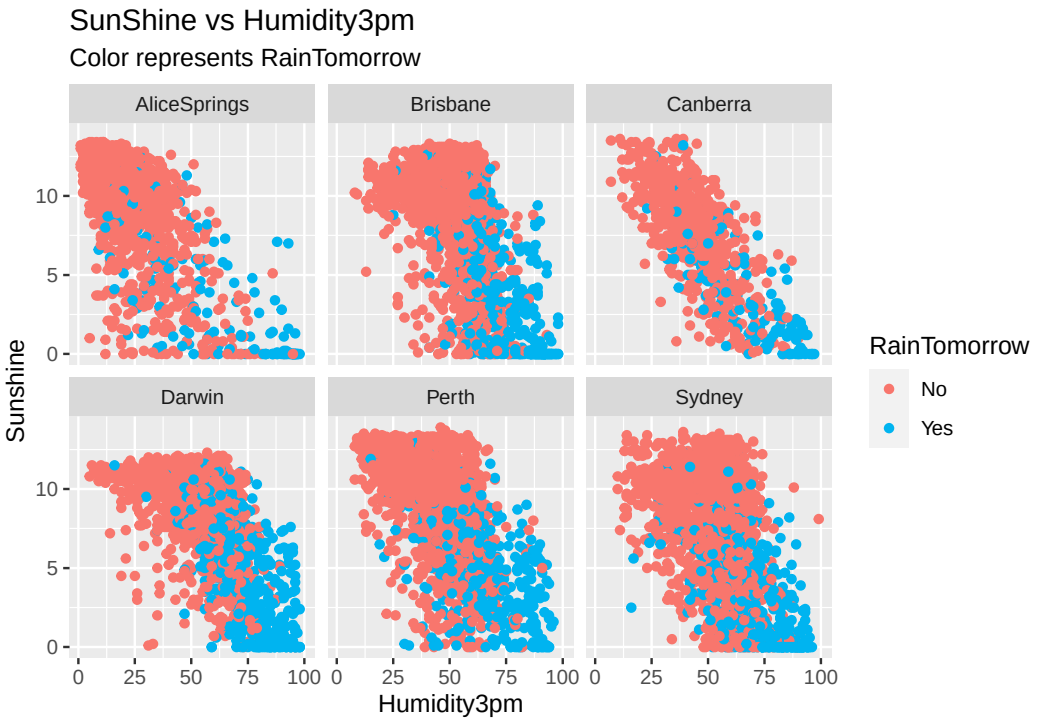
The below boxplots show that Humidity3pm and Sunshine may be good predictors of whether there will be rain tomorrow or not, because the range of values differ reasonably well between both situations.



However a scatter plot reveals that the delimitation between RainTomorrow Yes and No is very imprecise. There are many overlaps between the blue and red colors. This will certainly make predictions difficult: we can have high Humidity, low Sunshine and still no rain.



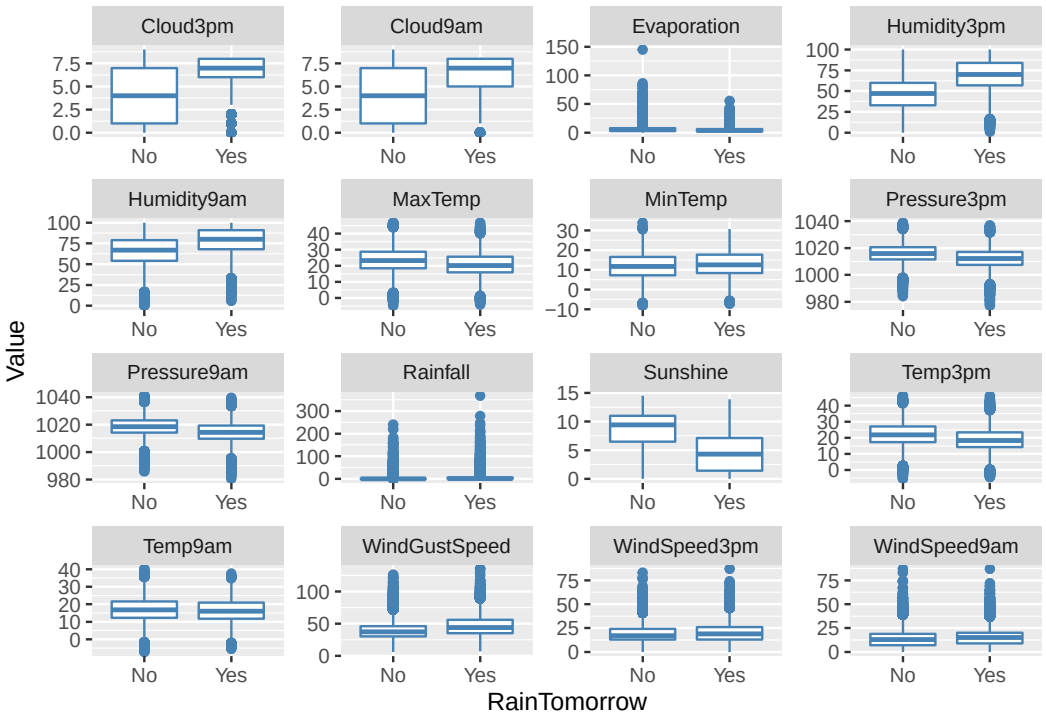
Let's see if this pattern is due to certain locations. We repeat the graph, selecting 6 locations only. The same complex pattern is observed.



This analysis confirms that Humidity3pm and Sunshine are certainly good predictors to use in a prediction model. However, these predictors cannot be used deterministically to predict Rain, given the high level of randomness.

2.4.5 Relationship between all numerical variables and RainTomorrow

We extend the previous study by looking at the relationship between all numeric data versus RainTomorrow. The plot shows that Cloud3pm may also offer some good differentiation between Rain and No Rain (but we know that it is highly correlated with Sunshine). Pressure and WindGustSpeed, and morning indicators for Cloud and Humidity, are also notable. For the other variables, the relationship is less strong.



2.5 Preparing the training set for modelisation

The last part in this section consists in preparing the training set "weather" for the modelization part: computing the NA replacement values, computing some additional variables to be used in the analysis, and splitting the training set into weather_train and weather_test for model training and cross-validation purpose.

2.5.1 NA Replacement values

The method used is as follows:

- NA replacement values are computed on the training "weather" set only. We do not use the validation set data.
- The NA replacement values are populated using the median for each Location-Month combination. Indeed, given the diversity of the location, we do not want to populate NAs with global averages, except where necessary due to lack of data. We also do not want to use yearly averages, as there are clear seasonality trends in the data (rainfall in Summer is different than in Winter).
- For categorical data, the mode is used (the most frequent category value) on a similar basis.
- NA replacement values are stored in a table for each Location, Month and Predictor (actually two tables are used, one for numerical data and one for categorical data). Thus they can be re-used for any new data set containing NAs, such as the validation set at the end of the study.
- Once computed, the NA replacement values are used to populate all NAs in the "weather" training set, for each Location-Month-Predictor combination.
- We did not use specific algorithms to fill in the missing data (such as MissRanger) due to the fact that such algorithms are quite time consuming, it's hard to control what they are doing, and it's not easy to apply them on the validation set without using the validation set data.
- We did, however, a small analysis to see if the NA replacement method used was distorting the results: the result of this analysis is presented in part 3 of the report and it turns out that no significant distortion was observed.

Note on categorical data

We observe that WindGustDir is fully missing for Albany and Newcastle. We cannot use a mode per month. We will therefore fill the WindGustDir NAs for Albany with Walpole data, and Newcastle with NorahHead data, as these locations are geographically close.

Table 3: NA Percentages

Location	WindGustDir	WindDir9am	WindDir3pm	RainToday
Albany	100.00%	0.00%	0.00%	0.00%
Newcastle	100.00%	0.00%	0.00%	0.00%

2.5.2 Creating weather_clean

We create a new training set, "weather_clean" where all NA values have been populated. We can compare the number of NA values in weather: 230,277 and weather_clean: 0.

2.5.3 Creating new variables

- We have seen the need to include seasonality effects. Therefore, we create "Year" and "Month", extracted from the "Date" column, in order to measure yearly and monthly effects. This will be more efficient than using the Date column.
- We have also seen the need to de-correlate certain variables. The 9am and 3pm variables have shown significant correlations. We create new "Var" variables based on the difference in value between 9am and 3pm: these Var variables will be much less correlated with the 3pm indicators, and will measure trends: is pressure increasing or decreasing? Thus, rather than using Pressure9am and Pressure3pm, as an example, we will rather use Pressure3pm and VarPressure.
- We do not include the variation on Cloud because Cloud is an index rather than a quantitative measurement (Cloud measures the fraction of sky obscured by cloud at 9am or 3pm. This is measured in "oktas", which are a unit of eighths).
- So we create the following:
  - VarHumidity = Humidity3pm - Humidity9am,
  - VarTemp = Temp3pm - Temp9am,
  - VarPressure = Pressure3pm - Pressure9am,
  - VarWindSpeed = WindSpeed3pm - WindSpeed9am,
  - Year = year(Date),
  - Month = month(Date)

2.5.4 Selecting variables

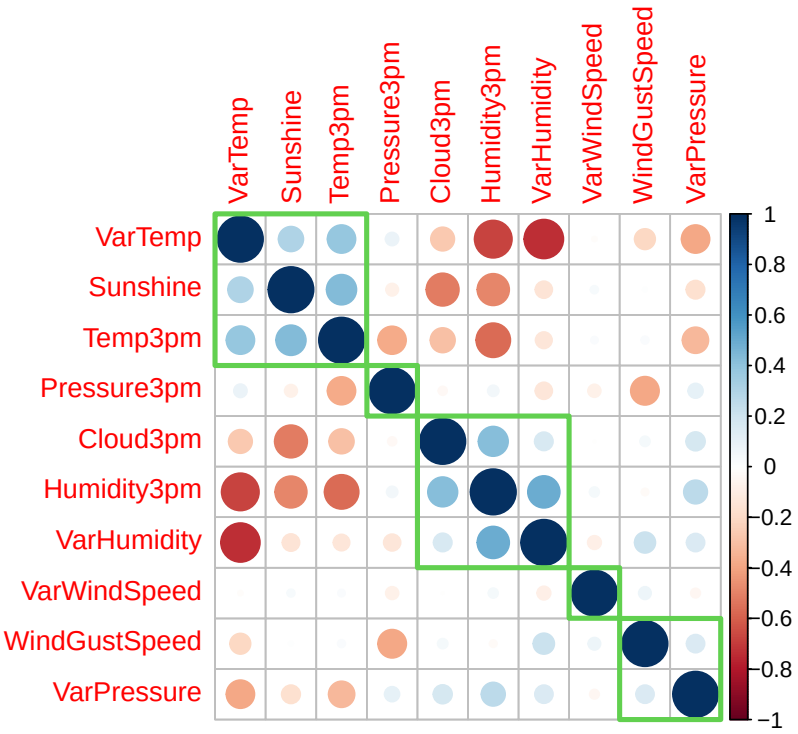
Finally, we create indexes which allow us to easily exclude certain variables from the analysis:

- an index identifying the 9am variables, as these will be replaced by the Var variables
- an index tagging certain variables highly correlated with other predictors, which we feel could be removed without impacting significantly the accuracy of the prediction. These variables are:
  - MinTemp, MaxTemp: highly correlated with Temp9am and Temp3pm
  - WindSpeed3pm: highly correlated with WindGustSpeed
  - WindDir3pm: we will use WindGustDir
  - Evaporation: highly correlated with Sunshine
  - Rainfall: identical to RainToday when Rainfall>1. A small analysis showed that is was more efficient in the GLM to remove Rainfall rather than RainToday.

Note that we do **not** drop these variables from the training set. We are just creating a facility to avoid using them. We will actually test, in the next section, whether their removal is justified.

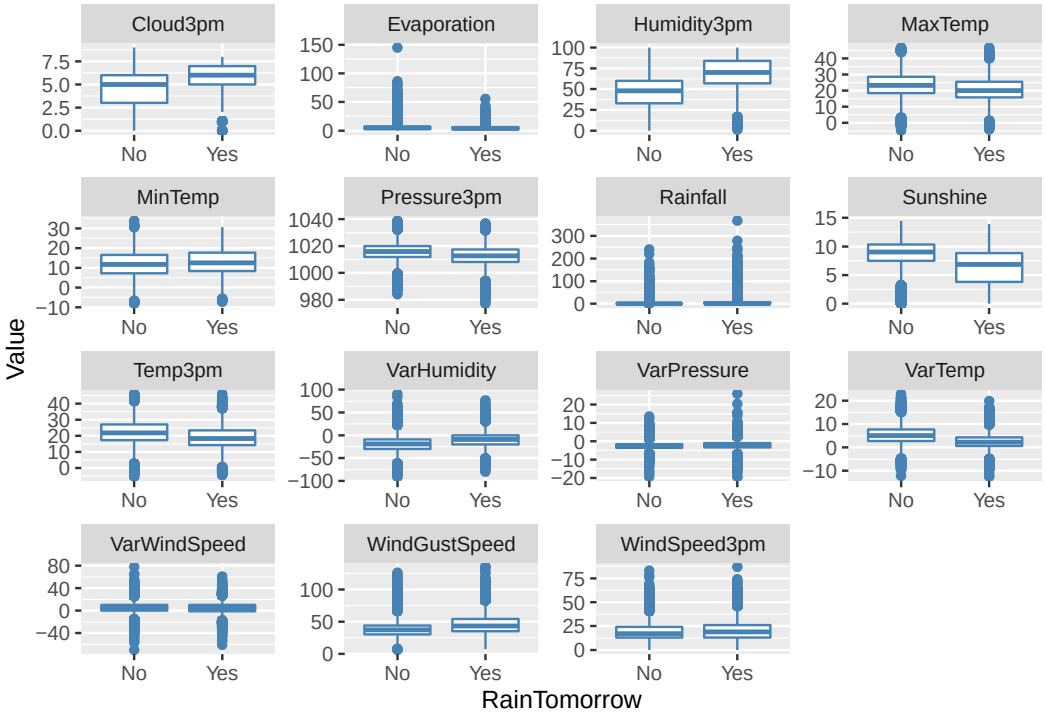
2.5.5 Correlations between the retained significant numeric variables

Now that we have selected a reduced number of significant variables, we can plot their correlations (at least for the numeric ones). Plotting with 5 clusters, we can see that the Var variables have relatively low correlation with the other variables. Some correlation remains between Humidity, SunShine and VarTemp, we will test the significance of these parameters later in the process.



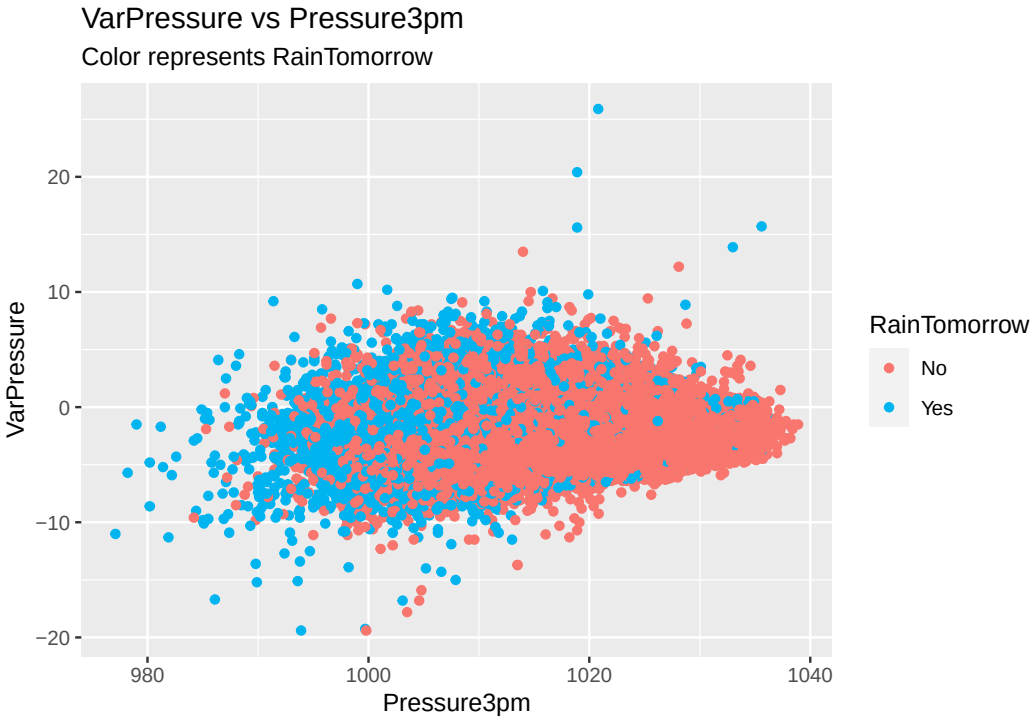
2.5.6 Selected numerical predictors vs RainTomorrow

This graph is a follow-up of the previous relationship graph which was based on weather, before NA replacement. We observe similar trends. Humidity3pm seems the best predictor of Rain Tomorrow at global level (but we know that in practice there are still many overlaps). Among the new variables, VarTemp and VarHumidity show some differentiation between No-Rain and Rain.

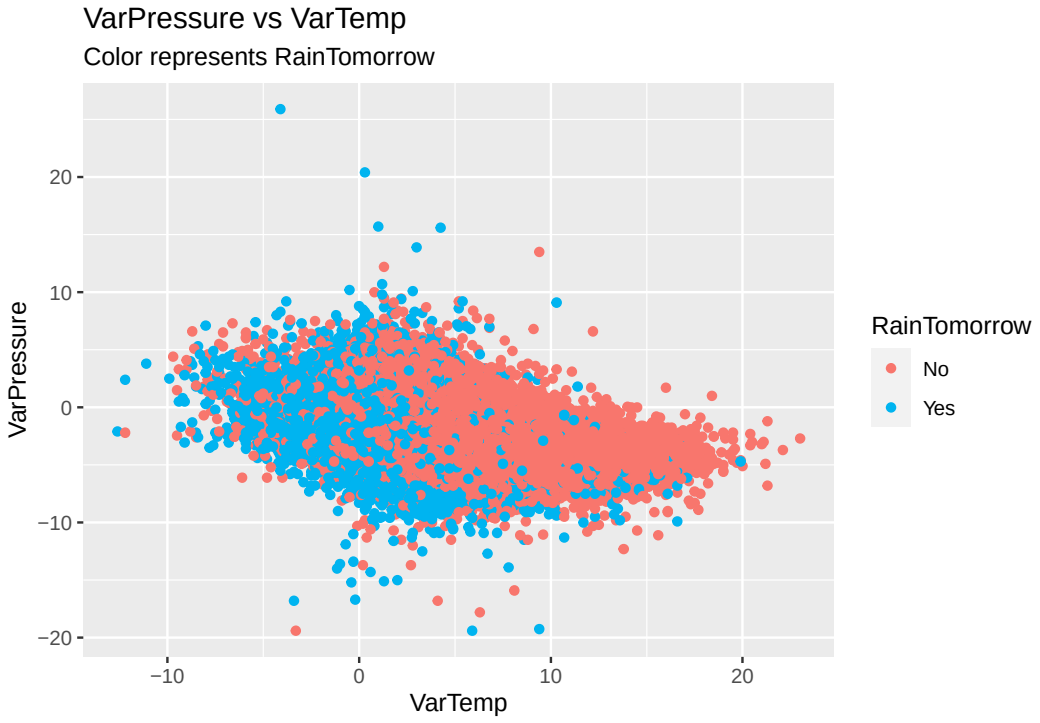


2.5.7 Plotting the new variables vs RainTomorrow

We can plot interesting graphs with the newly created variable. For instance, we expect low pressure and a drop in pressure before rain. So let's plot Pressure3pm and VarPressure. The plot confirms the expected trend, however we see that Rain and No Rain have significant overlaps.



Similarly, we would expect a drop in Temperature ahead of rain. So let's plot VarTemp and VarPressure. Again, some trends can be discerned but many overlaps are observed.



2.5.8 Updating the training and test sets

We now subdivide the training set "weather_clean" into a training set and a test set for model selection and cross-validation in the next section. We therefore have two sets:

- weather_test_clean (20% of weather_clean)
- weather_train_clean (80% of weather_clean)

These two sets are actually identical to the previously created weather_train and weather_test, except that the NA values have now been populated and the new variables created. We retain weather_train and weather_test because we want to test (in the next section) whether GLM results are distorted by the NA replacement.

As a reminder, the "validation" set is completely independent from these sets and has not been used at all. It still contains NA values and no new variables.

3 Model development and selection

3.1 Simple Models

Whilst we know that deterministic models will not have great accuracy, we can still learn from them. We will build two models.

3.1.1 The "No-Rain" model

We begin with a very simple model: since rain is rare, we predict no-rain !

The accuracy is 77.5%. This is quite high and due to the fact that the proportion of rainy days is around 23% (as noted earlier, RainToday is Yes when RainFall >1).

The accuracy is actually equal to the mean of RainTomorrow "No" in the test set: TRUE.

Printing the results of the model, we note that the Specificity is 100% (actual No-Rain predicted with 100% success) and Sensitivity is nil (actual Rain predicted with 0% success since it is never predicted). Balanced accuracy is 50% only. The "No Rain" model is a poor model.

Table 4: Results

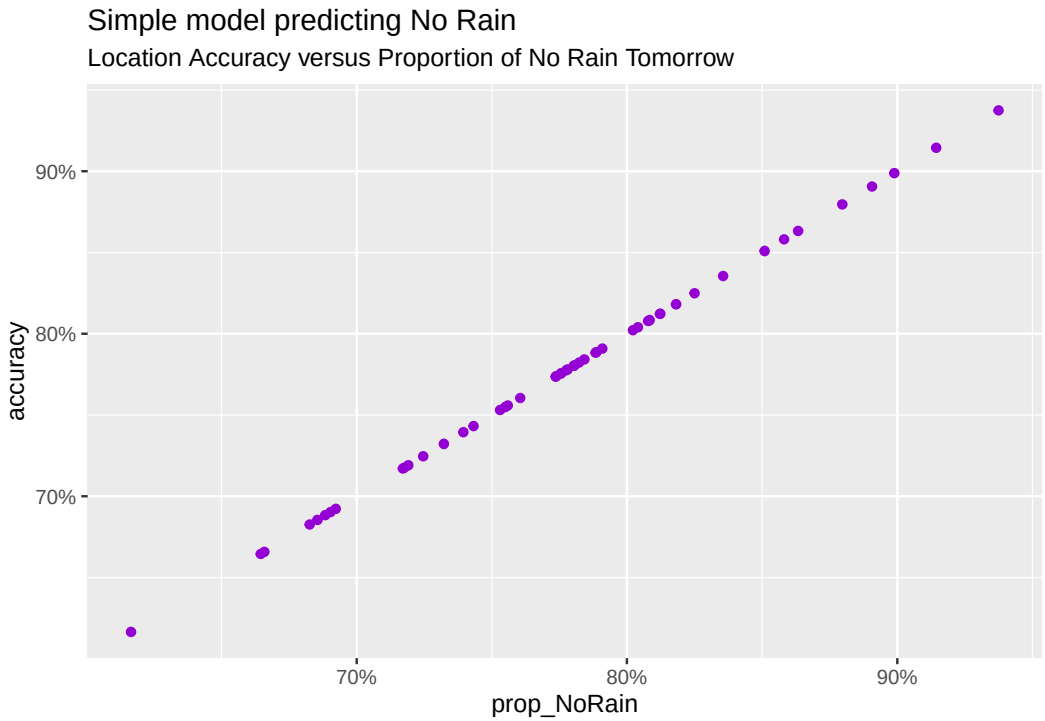
Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Predicting no rain	77.5%	0.0%	100.0%	50.0%

Note on the above table

- Sensitivity = accuracy of predicting actual "Rain Tomorrow"
- Specificity = accuracy of predicting actual "No Rain Tomorrow"
- Balanced_Accuracy = average between Sensitivity and Specificity

Impact of prevalence

The No-Rain model offers the opportunity to visualize the impact of No-Rain prevalence and how it distorts accuracy. In the below chart, we plot the accuracy of the No-Rain model for each location versus the proportion of No-Rain in this location. We see a perfect alignment. If a location has rare rain (high percentage of No-Rain), the model shows high accuracy: but this does not mean that it is a good model, hence the importance of looking at balanced accuracy too.



3.1.2 The Humidity model

We continue with another model based on Humidity3pm. This variable indeed seemed to be a good predictor of RainTomorrow, at least on a global level, based on the box-plots shown earlier.

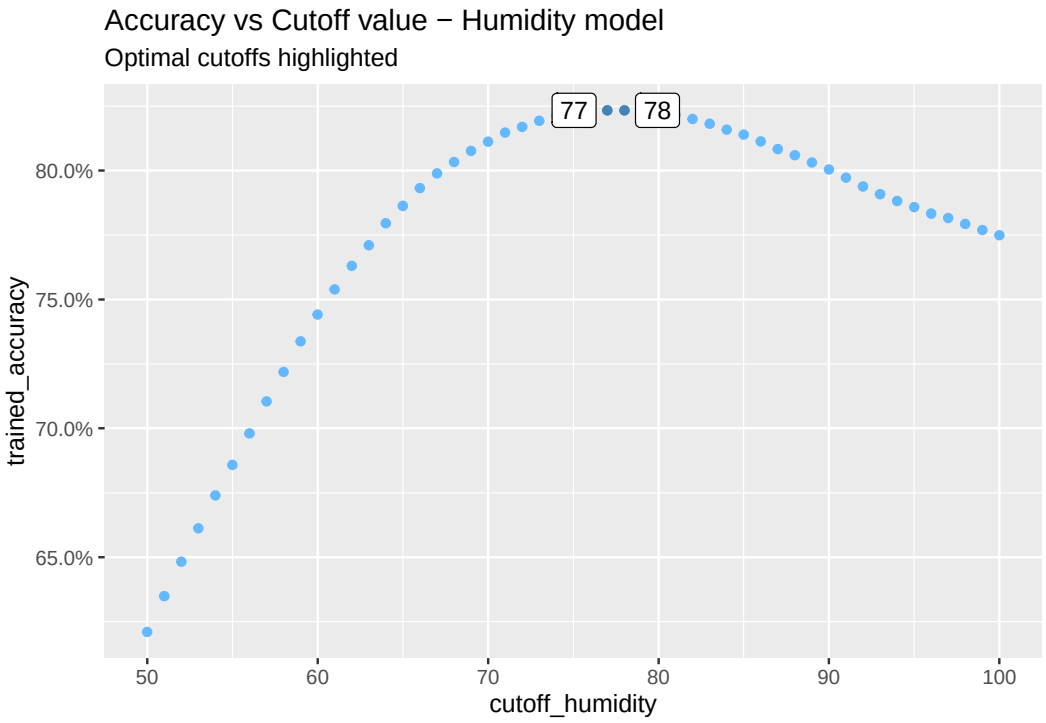
Defining the model

The below table shows that the 3rd Quartile of Humidity3pm (60) with No-RainTomorrow is almost equal to the first Quartile (57) if there is RainTomorrow. Therefore there is limited overlap, which can be used for prediction. We will use this to define a model which predicts RainTomorrow if Humidity3pm is above a certain value.

```
## $No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   33.00   48.00   46.58   60.00  100.00
##
## $Yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   57.00   70.00   68.64   84.00  100.00
```

Training the model

Training the Humidity model on the train set for a number of cutoffs, we can plot the results to identify the optimal cutoff. The optimal cutoff is 77.



Testing the model

We now test the humidity model on the test set. The accuracy is 82.3%

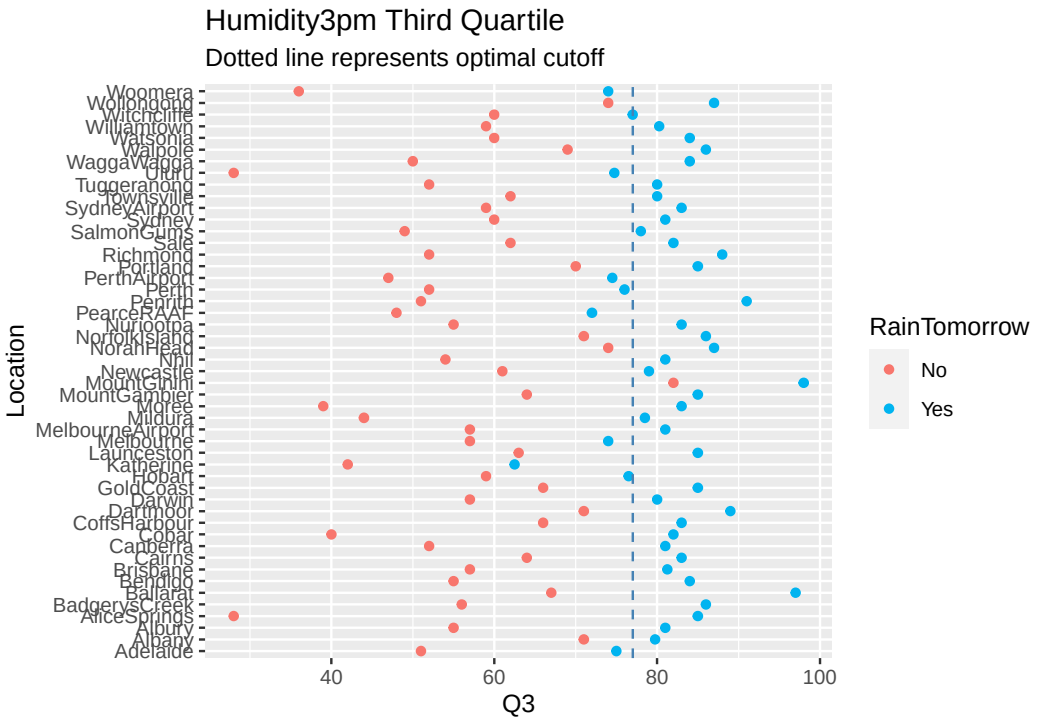
Looking at our four key indicators, Balanced Accuracy has improved to 65% but with still poor Sensitivity 35%:

Table 5: Results

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Predicting no rain	77.5%	0.0%	100.0%	50.0%
Humidity3pm with cutoff	82.3%	35.1%	96.0%	65.5%

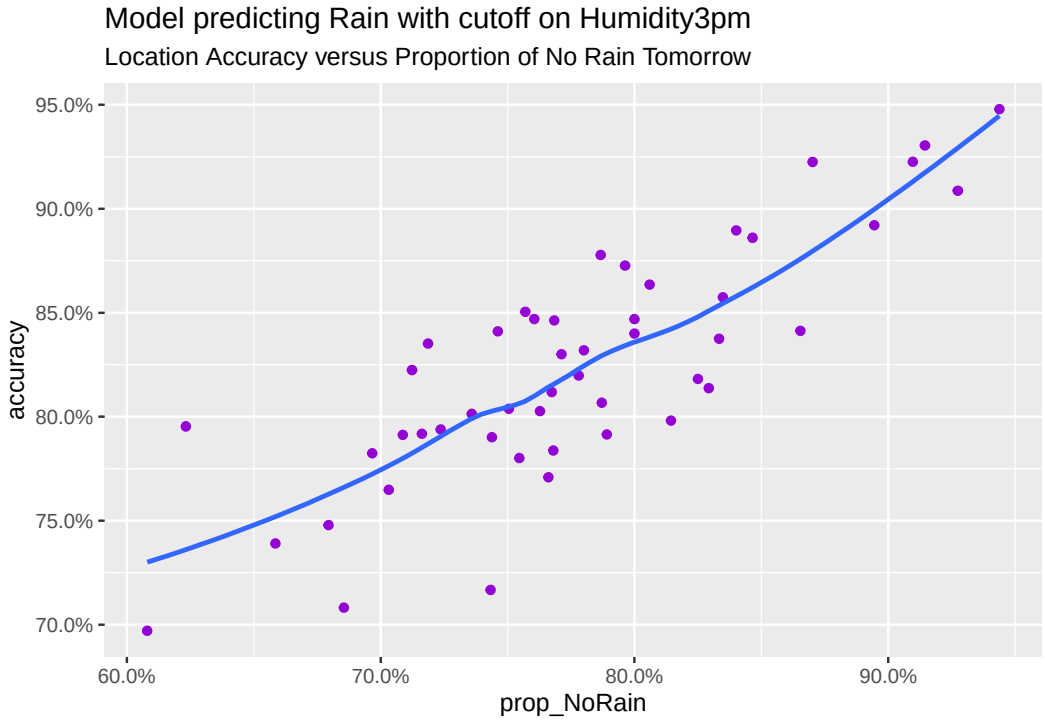
Understanding the cutoff

The cutoff is surprisingly high. Based on the quantile table for Humidity3pm, we were expecting a cutoff around 60. We can understand this by plotting the third quartile of Humidity3pm. The plot shows that humidity levels have very different ranges per location. In particular, there are many locations where Humidity levels are high even in the case of no rain and much higher than the third quartile (60). This pushes the model cutoff higher. This also suggests that using a single cutoff is an over-simplification. In the next section, we will test the Location:Humidity interaction.



Visualizing the impact of prevalence

By plotting accuracy versus the proportion of no Rain in the test set, we can see that Rain prevalence is impacting the accuracy in this model too, to a certain extent.



3.2 GLM

3.2.1 Global GLM using significant variables

We now test a GLM approach based on the significant variables identified in the data analysis section. We define a set of parameters using all the available parameters (this does not include Date but includes Year and Month) but excluding the highly correlated ones. Factors like Location and WindGustDir are included.

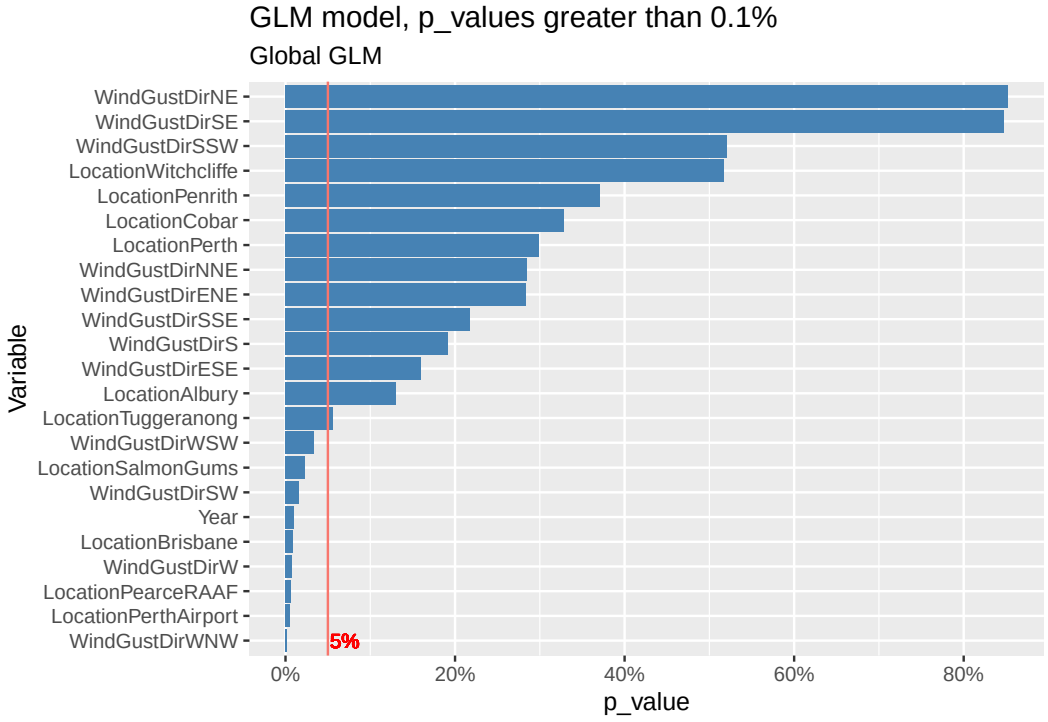
We will call this model Global GLM as we will later compare it with a separate model by location, which we will call Local GLM.

Parameters used

```
## RainTomorrow ~ Location + Sunshine + WindGustDir + WindGustSpeed +
##     Humidity3pm + Pressure3pm + Cloud3pm + Temp3pm + RainToday +
##     VarHumidity + VarTemp + VarPressure + VarWindSpeed + Year +
##     Month
```

Parameters significance The GLM summary provides many details, however we do not print it in this report because it is quite lengthy. One of the key points is that most variables are significant. We can confirm this by plotting p_values exceeding 1%. Predictors not appearing on the graph have low p_values. All retained variables have low p-values with three exceptions:

- certain values of WindGustDir have high p_values however N and W combinations have high significance, as anticipated in the data exploration, indicating that we should retain this variable
- certain locations are not significant but we will ignore this as location will be handled separately in the next step
- Year is not highly significant (p_value 1%) (however we tested separately that removing it reduces accuracy)



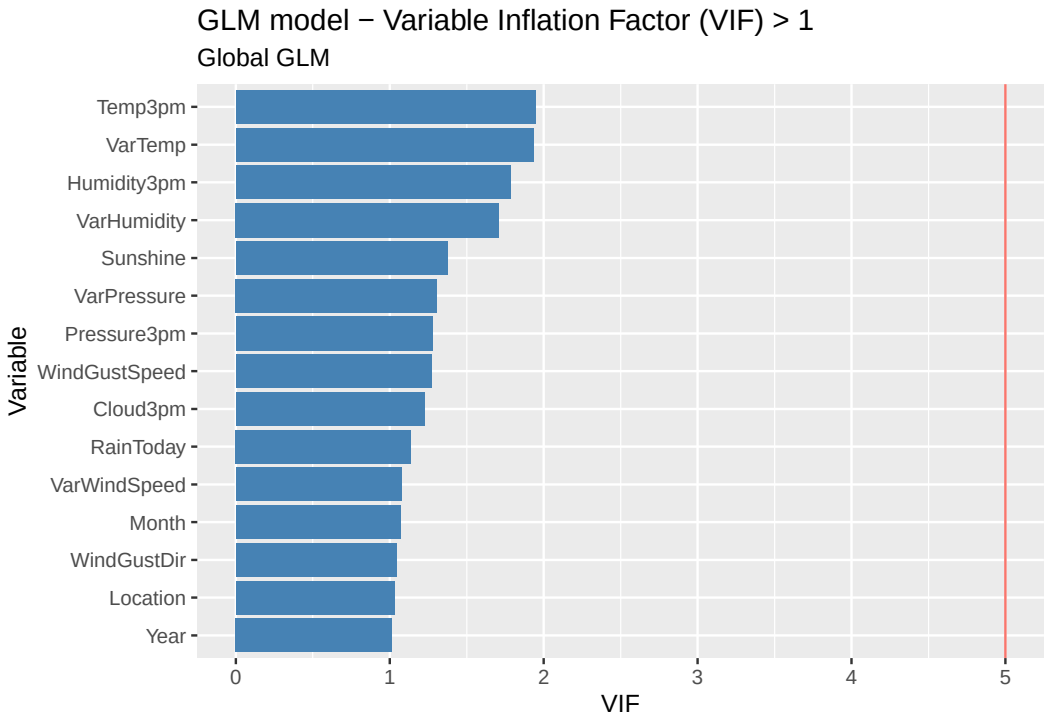
VIF

Looking at p_values only, however, is not sufficient. Multicollinearity can distort p-values therefore we need to analyze it. The most common way to detect multicollinearity is by using the variance inflation factor (VIF), which measures the correlation and strength of correlation between the predictor variables in a regression model.

The value for VIF starts at 1 and has no upper limit. A general rule of thumb for interpreting VIFs is as follows:

- A value of 1 indicates there is no correlation between a given predictor variable and any other predictor variables in the model.
- A value between 1 and 5 indicates moderate correlation but this is often not severe enough to require attention.
- A value greater than 5 indicates potentially severe correlation between a given predictor variable and other predictor variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.

We use the VIF function from DescTools. The below plot shows that all variables are well below 5. This confirms the significance of the chosen variables.



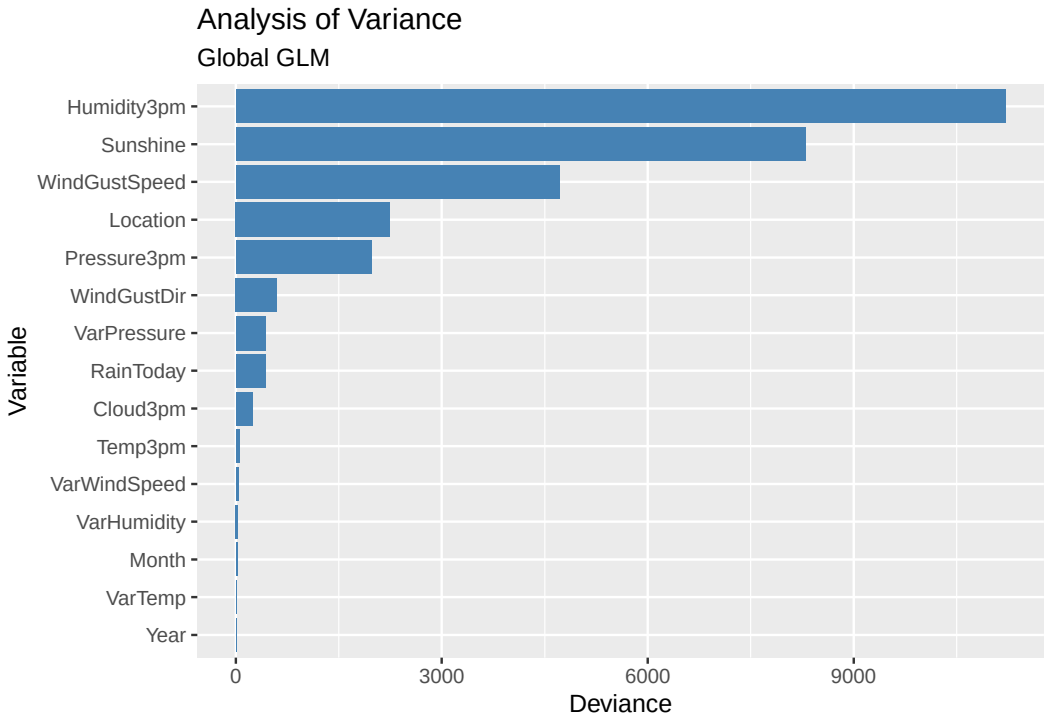
Global GLM results

The overall accuracy of the global GLM model is 84.9%. The key results are as follows:

Table 6: Results

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Global GLM	84.9%	52.3%	94.3%	73.3%

Analysis of Variance (ANOVA) We compute and plot the ANOVA to analyze the importance of each variable. # The ANOVA plot shows the most important variables at the top of the graph, mainly Humidity3pm, Sunshine, WindGustSpeed, Location. However we will not discard the other variables, which still contribute and have low VIF and low p-values as we have seen.



3.2.2 GLM based on Location and Location:Humidity only

As a quick follow-up to the Humidity model, and in order to assess how much humidity contributes to the overall accuracy of the global GLM model, we perform a quick glm based on Location and the interaction between Location and Humidity (because one of the problems observed in the Humidity3pm model in the earlier section was that one cutoff for all locations was an over-simplication).

The parameter used is therefore:

```
## RainTomorrow ~ Location + Location:Humidity3pm
```

The results are as follows and confirm that the global GLM with significant variables has the best Accuracy, Sensitivity (ability to correctly predict Rain) and Balanced Accuracy, at the price of a slightly reduced Specificity. The Humidity GLM model explains a good part of the accuracy, confirming that other predictors are less important for the overall accuracy, however the Humidity GLM model is well below the Global GLM model in terms of Sensitivity.

Table 7: Results

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Predicting no rain	77.5%	0.0%	100.0%	50.0%
Humidity3pm with cutoff	82.3%	35.1%	96.0%	65.5%
GLM Location & Humidity	83.1%	39.5%	95.8%	67.6%

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Global GLM	84.9%	52.3%	94.3%	73.3%

3.2.3 Checking impact of NA replacement

It is important to check whether NA replacement has impacted GLM results. We will therefore compare the current results with a GLM model applied on the data excluding NAs.

We notice a significant reduction of the size of the data sets when we remove all rows containing NA values. In particular, we have only 26 locations remaining instead of 49 in the full data set.

The table below shows a better accuracy when NAs are fully excluded:

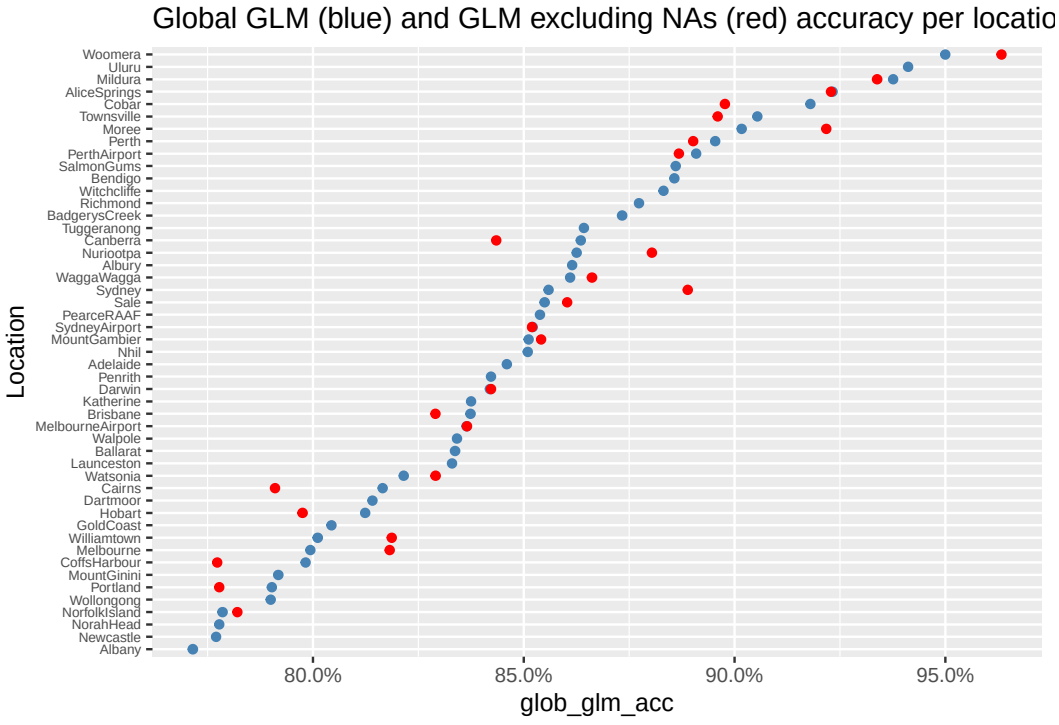
Table 8: Results with and without NAs

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Global GLM	84.9%	52.3%	94.3%	73.3%
GLM excluding NAs	85.7%	56.1%	94.3%	75.2%

However, this is misleading, as we are not comparing the same number of locations. Indeed, if we plot the accuracy per location for the global GLM model and the model based on data where NAs have been removed, we observe that:

- in 14 cases, no-NA accuracy is better (red point to the right of the blue point)
- however in 12 cases, no-NA accuracy is worse (red point to the left)

It is therefore a mixed picture. We conclude that there is no evidence that NA replacement have significantly distorted the GLM results.



This can be further confirmed by re-running the Global GLM on the 26 same locations only. Although there is still a small difference, especially on the Sensitivity, the results are overall very close to the no-NA GLM. So obviously, having no NAs in the data would help, but NA replacement has not impaired results significantly.

Table 9: Results with and without NAs

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Global GLM	84.9%	52.3%	94.3%	73.3%
GLM excluding NAs	85.7%	56.1%	94.3%	75.2%
Global GLM same locations	85.6%	54.6%	94.3%	74.5%

3.2.4 Checking the parameter selection

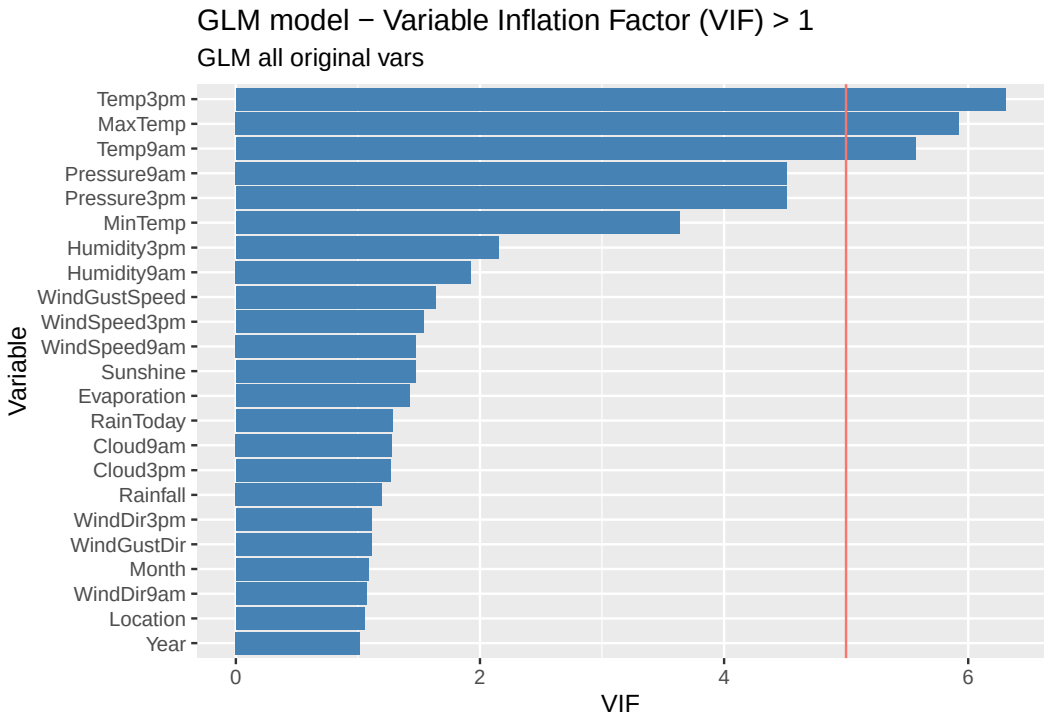
Before moving to the next step and exploring interactions, we need to check whether selecting only the significant parameters was the right choice. We will therefore compute two other models:

- a model based on the original parameters (excluding the created Var parameters)
- a model based on the significant parameters without removing the highly correlated parameters

A model based on all original parameters would have the following parameters:

```
## RainTomorrow ~ Location + MinTemp + MaxTemp + Rainfall + Evaporation +
##     Sunshine + WindGustDir + WindGustSpeed + WindDir9am + WindDir3pm +
##     WindSpeed9am + WindSpeed3pm + Humidity9am + Humidity3pm +
##     Pressure9am + Pressure3pm + Cloud9am + Cloud3pm + Temp9am +
##     Temp3pm + RainToday + Year + Month
```

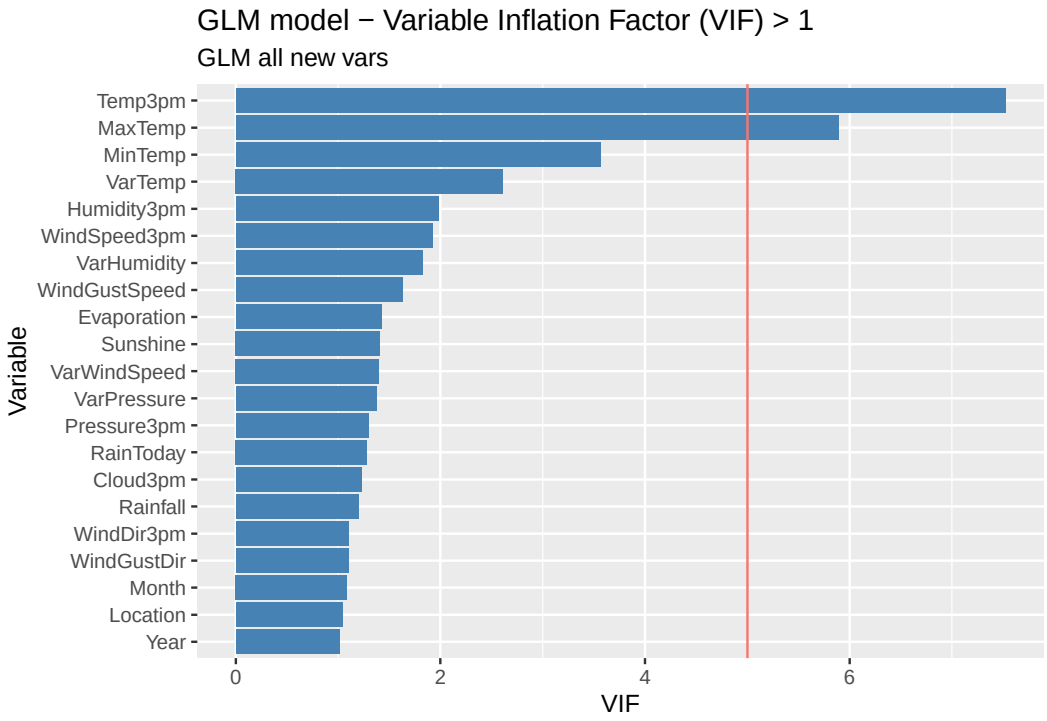
Its accuracy would be 84.8%, however with Temperature indicators having a high VIF and several indicators having high p_values, therefore producing a less reliable model than the "Significant Variables" model.



A model based on the new parameters without removing any of the highly correlated parameters, would have the following parameters:

```
## RainTomorrow ~ Location + MinTemp + MaxTemp + Rainfall + Evaporation +
##     Sunshine + WindGustDir + WindGustSpeed + WindDir3pm + WindSpeed3pm +
##     Humidity3pm + Pressure3pm + Cloud3pm + Temp3pm + RainToday +
##     VarHumidity + VarTemp + VarPressure + VarWindSpeed + Year +
##     Month
```

Its accuracy would be 84.9%, however with Temperature however with two Temperature indicators having a high VIF.



The comparison of the various models is summarized in the below table:

Table 10: Model accuracy on Test set

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Predicting no rain	77.5%	0.0%	100.0%	50.0%
Humidity3pm with cutoff	82.3%	35.1%	96.0%	65.5%
GLM Location & Humidity	83.1%	39.5%	95.8%	67.6%
Global GLM	84.9%	52.3%	94.3%	73.3%
GLM all original vars	84.8%	52.3%	94.3%	73.3%
GLM all new vars	84.9%	52.2%	94.3%	73.3%

Conclusion Although accuracies are similar, the GLM model with significant variables ("Global GLM" in the above table) offers the advantage of having variables with low VIF. It is therefore the preferred model in this section.

3.2.5 Assessing the interactions between significant variables

So far, we have not included any interactions between the variables. We now run a model with two-by-two interactions to see which interactions are significant. We will however not include interactions for Location (will be handled separately in the next stage), some of the categorical predictors and Year, in order to avoid too much complexity.

The parameters used are as follows (the ()^2 in the formula creates two by two interactions between the parameters):

```
## RainTomorrow ~ Location + WindGustDir + Year + (Sunshine + WindGustSpeed +
##     Humidity3pm + Pressure3pm + Cloud3pm + Temp3pm + RainToday +
##     VarHumidity + VarTemp + VarPressure + VarWindSpeed + Month)^2
```

The accuracy improves slightly to 85.2%.

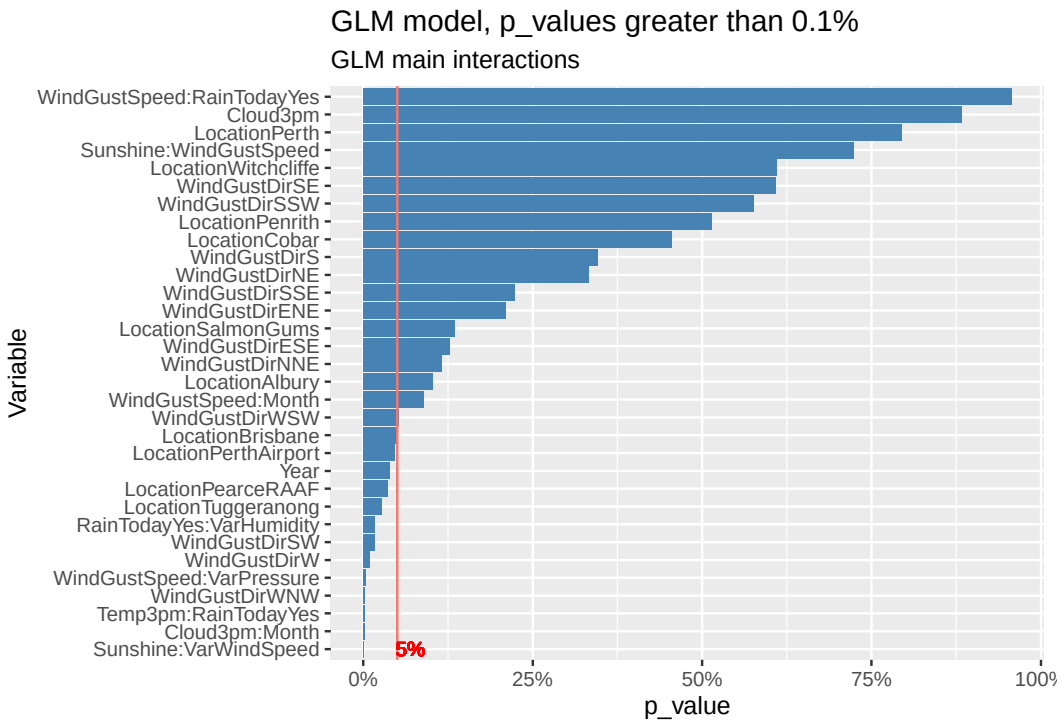
Many interactions are not significant when looking at the GLM summary and we have high multicollinearity (we do not plot the VIF graph as it is very cluttered).

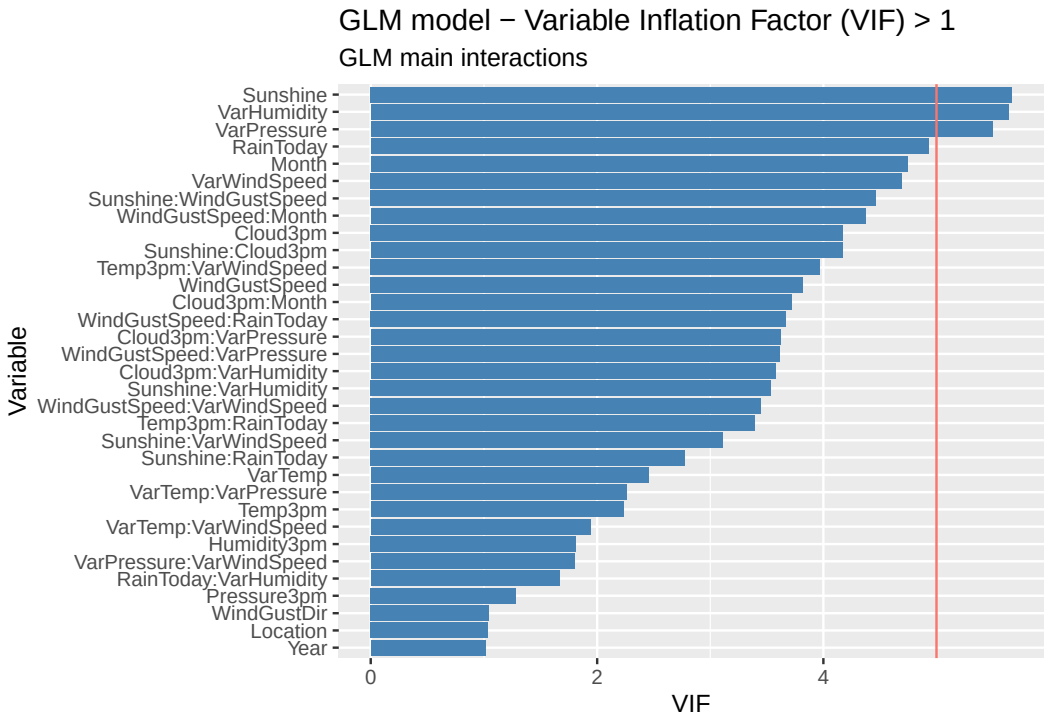
We will therefore filter down the number of interactions and re-run a model. We keep only interactions with a p-value less than 5% and VIF < 6.

The parameters are as follows:

```
## RainTomorrow ~ Location + Sunshine + WindGustDir + WindGustSpeed +
##     Humidity3pm + Pressure3pm + Cloud3pm + Temp3pm + RainToday +
##     VarHumidity + VarTemp + VarPressure + VarWindSpeed + Year +
##     Month + Sunshine:WindGustSpeed + Sunshine:Cloud3pm + Sunshine:RainToday +
##     Sunshine:VarHumidity + Sunshine:VarWindSpeed + WindGustSpeed:RainToday +
##     WindGustSpeed:VarPressure + WindGustSpeed:VarWindSpeed +
##     WindGustSpeed:Month + Cloud3pm:VarHumidity + Cloud3pm:VarPressure +
##     Cloud3pm:Month + Temp3pm:RainToday + Temp3pm:VarWindSpeed +
##     RainToday:VarHumidity + VarTemp:VarPressure + VarTemp:VarWindSpeed +
##     VarPressure:VarWindSpeed
```

The accuracy is now 84.8%. Some variables have high VIF, due to the inter-correlations and some have high p-values.





We summarize the results of this section

Table 11: Model accuracy on Test set

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Predicting no rain	77.5%	0.0%	100.0%	50.0%
Humidity3pm with cutoff	82.3%	35.1%	96.0%	65.5%
GLM Location & Humidity	83.1%	39.5%	95.8%	67.6%
Global GLM	84.9%	52.3%	94.3%	73.3%
GLM all original vars	84.8%	52.3%	94.3%	73.3%
GLM all new vars	84.9%	52.2%	94.3%	73.3%
GLM with interactions	85.2%	53.8%	94.3%	74.1%
GLM main interactions	84.8%	52.1%	94.3%	73.2%

We could continue by computing the ANOVA and selecting variables further. However, this would take us to a lower accuracy than was obtained without interactions. Therefore we conclude that **adding interactions does not materially improve the GLM model**. It can improve accuracy marginally, at the cost of higher model complexity. Going forward, we retain the significant parameters only.

3.2.6 Separate GLM per location

We will now compare the global model with a model that treats each location as a separate data subset. Indeed there are reasons to believe that parameters may behave differently based on the location. This should allow to better cater for the specificities of each location.

We use the significant parameters and remove the location only. The parameters are as follows:

```
## RainTomorrow ~ Sunshine + WindGustDir + WindGustSpeed + Humidity3pm +
```

```
##     Pressure3pm + Cloud3pm + Temp3pm + RainToday + VarHumidity +
##     VarTemp + VarPressure + VarWindSpeed + Year + Month
```

Using the lapply and Map functions, we split the data set by location and create 49 GLM models, one per location. We apply these models to the test set also split by location.

Note that we can still access the model for each location, for instance the command "summary(m_loc[["Canberra"]])" would provide the summary of the Canberra model (m_loc being a list containing the local models).

The overall accuracy improves to 85.8%. The table shows that the separate GLM model by location has better Sensitivity and Balanced Accuracy :

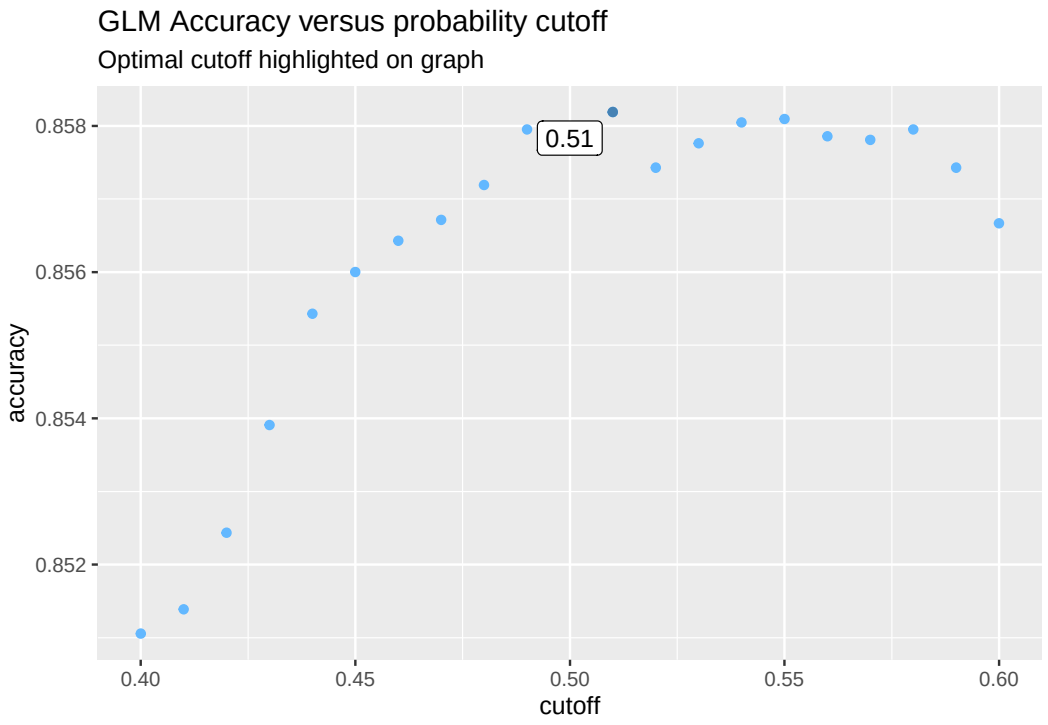
Table 12: Accuracy on test set

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Predicting no rain	77.5%	0.0%	100.0%	50.0%
Humidity3pm with cutoff	82.3%	35.1%	96.0%	65.5%
GLM Location & Humidity	83.1%	39.5%	95.8%	67.6%
Global GLM	84.9%	52.3%	94.3%	73.3%
Local GLM	85.8%	56.9%	94.2%	75.5%

3.2.7 Testing the probability cutoff

In extracting the predictions from the GLM models, we are using a probability cutoff at 0.5: if the probability is higher than 0.5, "Rain Tomorrow Yes" is predicted. We analyze the impact of this cutoff.

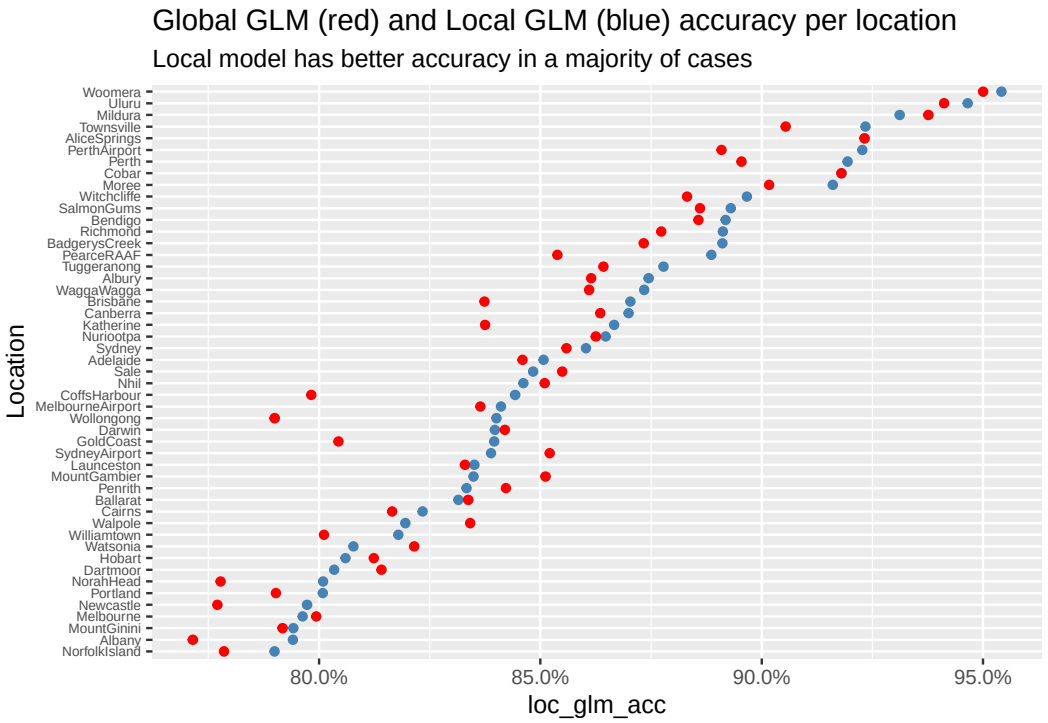
The below graph shows the model accuracy based on various cutoffs.



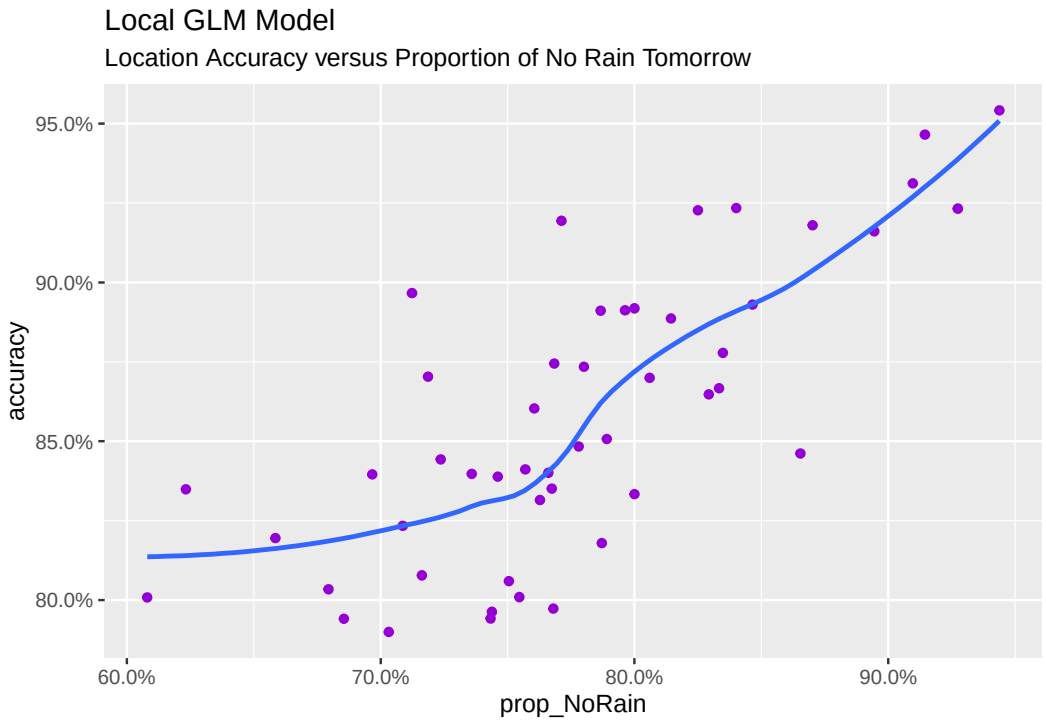
The optimal probability cutoff could be 0.51 however the improvement is not significant: the maximum accuracy is 85.8%. Such a cutoff would also decrease Specificity by reducing Yes predictions. For these two reasons, we will retain the usual cutoff.

3.2.8 Plotting the accuracy of the GLM models

The below plot compares the local model (one GLM model per location) with the global model (one global GLM model). We can see that the local model has better accuracy in a majority of cases (blue dots to the right of the red dots). Therefore this model has not only better overall accuracy, it offers better accuracy for most locations as well.



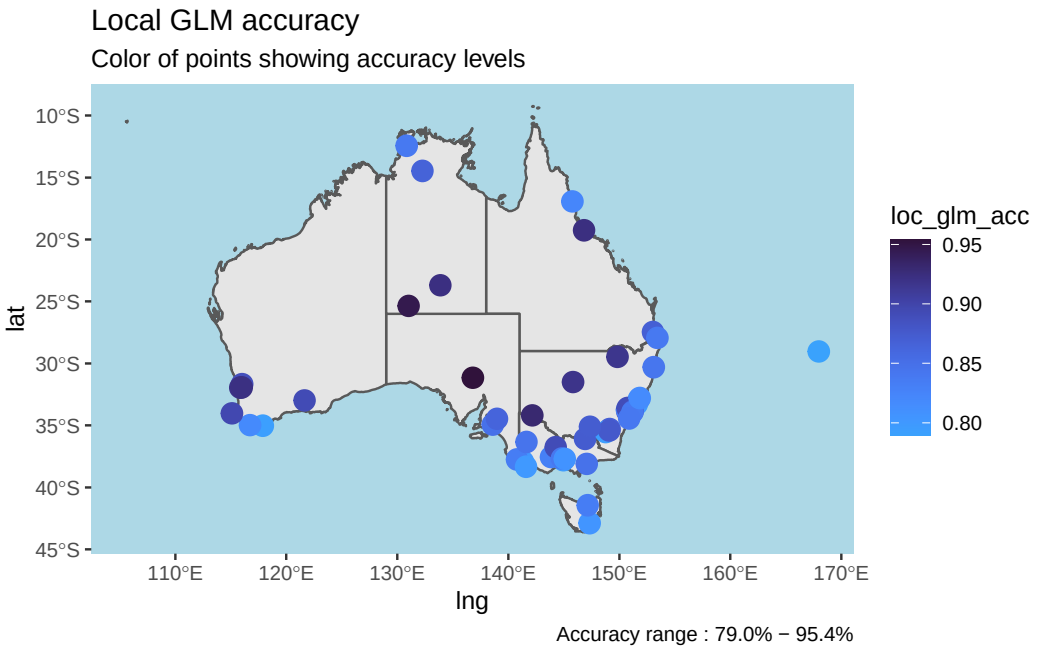
Like we did for some previous models, we plot the accuracy per location versus the No-Rain proportion in that location, to visualize the impact of prevalence. We can see that whilst there is still a link between the No Rain prevalence and model accuracy, the link is not very strong. This is consistent with the better balanced accuracy observed.



3.2.9 Trying to understand where the model fails

In this section, we try to understand what factors prevent the model from performing better, especially in terms of Sensitivity.

We start by plotting the local GLM accuracy per location on a map. This shows that the accuracy is lower in coastal locations. These are indeed areas where we expect the weather to change frequently, due to the oceanic influence. This is an indication that there is a limitation to what the model can do, due to the randomness introduced by this influence.

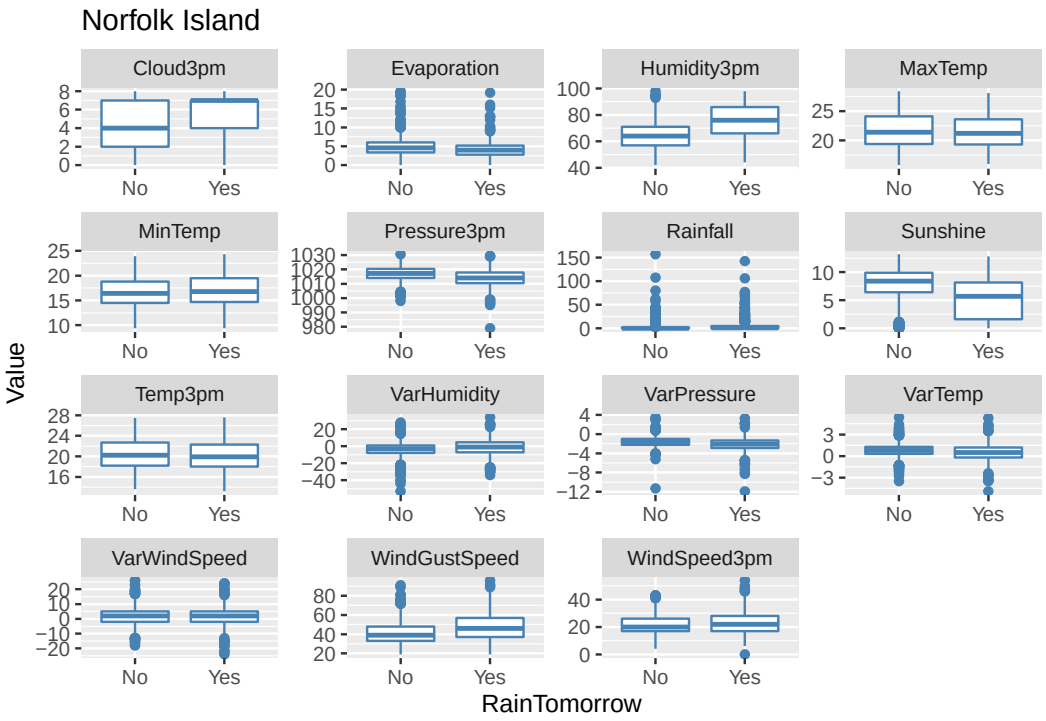


Let's zoom into Norfolk's results, which is the location where the model performs worst. We can compare the general results with the results for Norfolk. All indicators are lower in Norfolk:

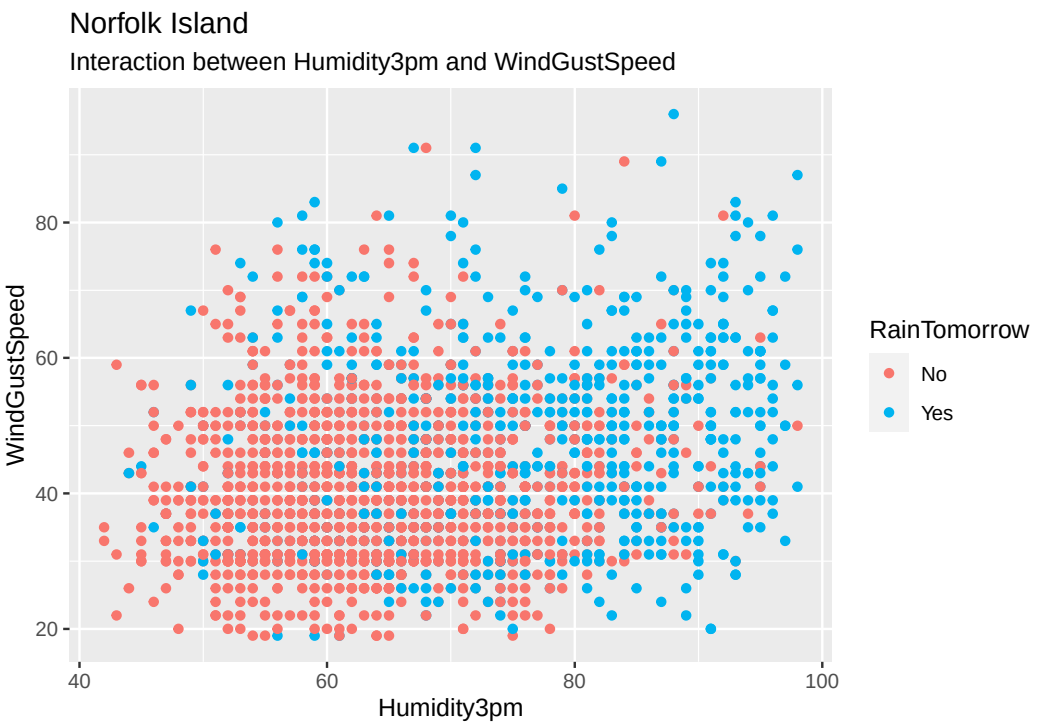
Table 13: Local GLM Accuracy: overall vs Norfolk

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Local GLM	85.8%	56.9%	94.2%	75.5%
Norfolk Island	79.0%	53.2%	91.0%	72.1%

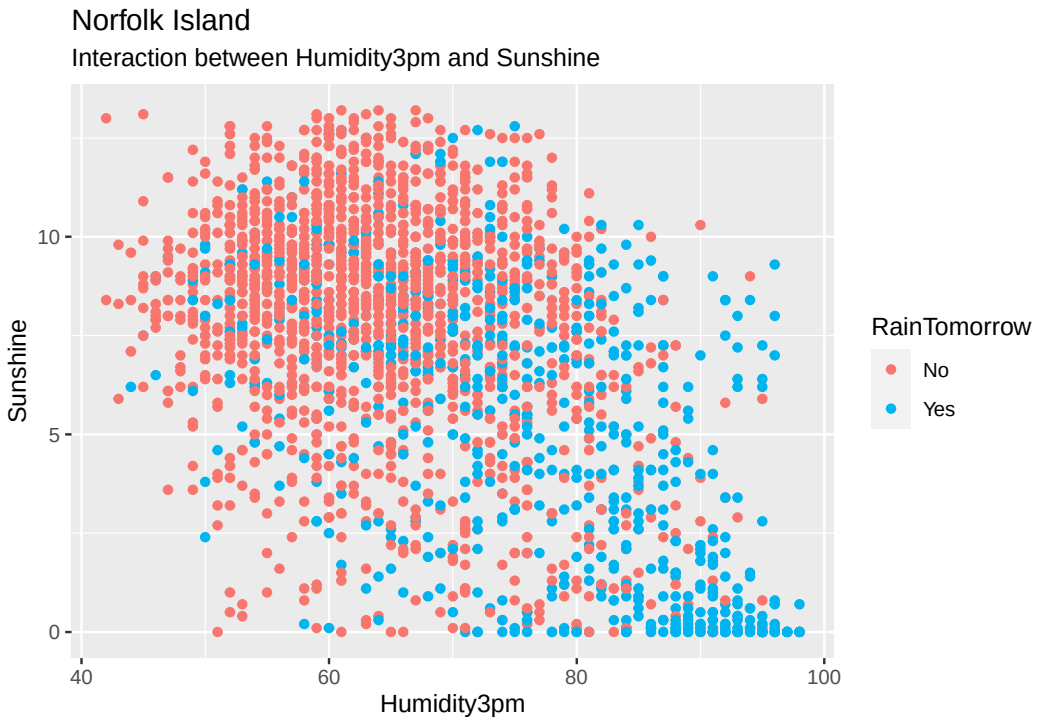
The below graph helps to understand why the model has lower accuracy in Norfolk: there are significant overlaps for all predictors between Rain and No-Rain. There are trends, of course, like high Humidity and low Sunshine indicate a higher chance of rain, however there is no clear-cut separation.

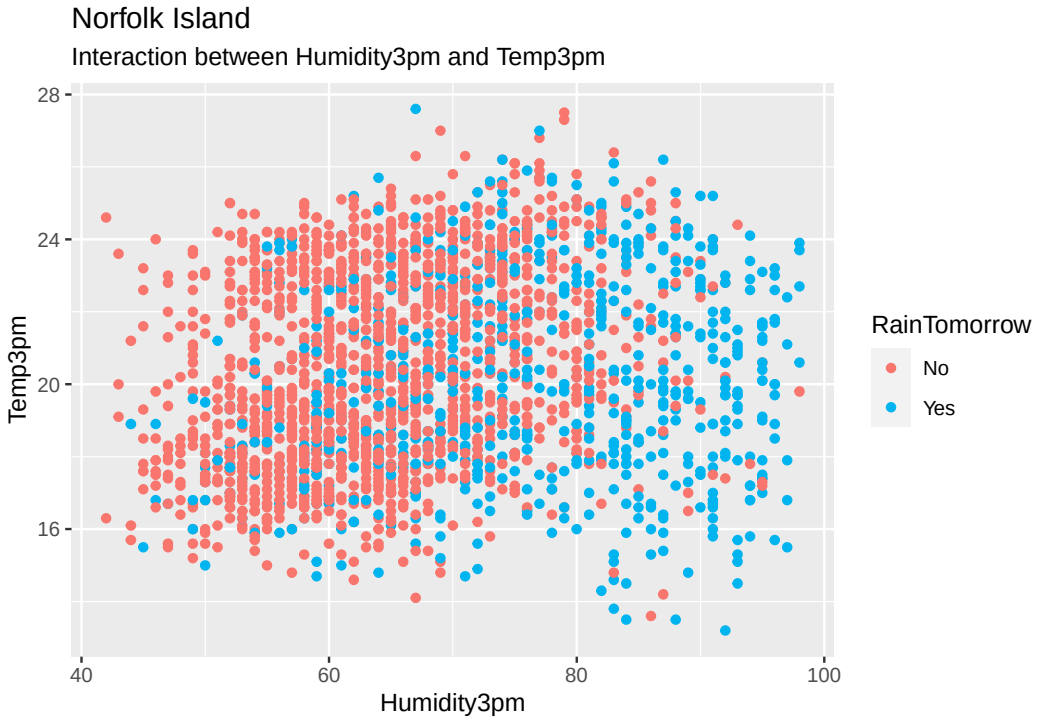


Another way to understand why predictions are difficult is to look at scatterplots. The plot below for instance shows that there are no clear-cut separations on a combination of Humidity3pm and WindGustSpeed, between No Rain Tomorrow and Rain Tomorrow:



The same is true for Humidity3pm:Sunshine and Humidity3pm:Temp3pm.





The above shows that it will be difficult for any model to do better predictions. We will try other methods however.

3.3 Random Forest model

Based on the earlier data analysis, there are reasons to believe that applying thresholds to each predictor could be a way to separate the No and Yes cases for RainTomorrow. Rather than using a simple decision tree, we use a Random Forest approach, as Random Forests correct for decision trees' habit of overfitting to their training set and generally outperform decision trees.

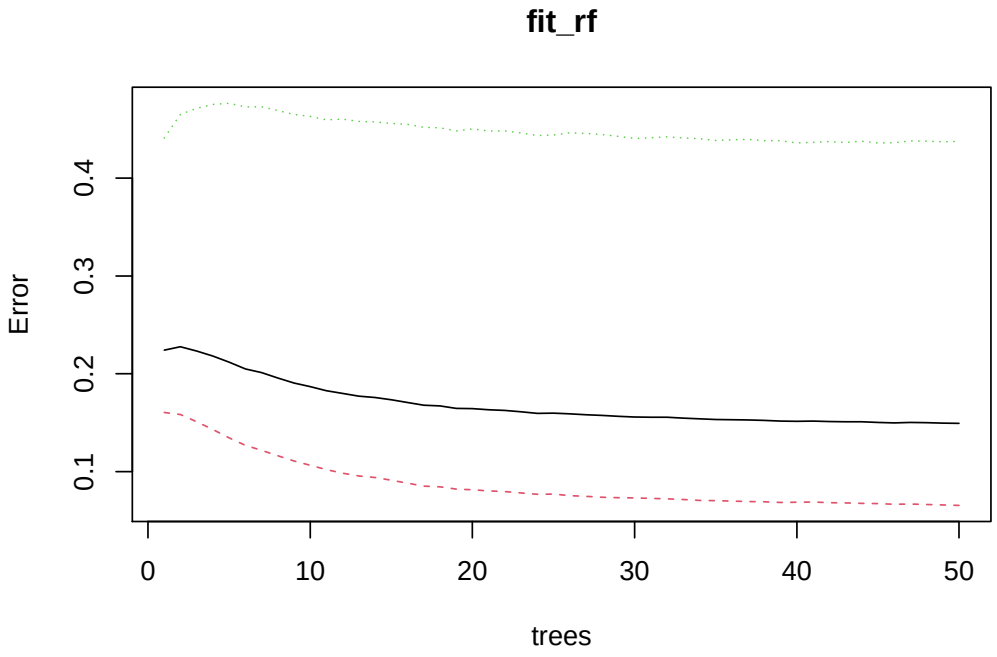
Similar to our GLM approach, we will start with a global model and then compare it with a local model. We use the same significant variables used in the GLM approach.

There is no need for data pre-processing (centering and scaling) before a Random Forest analysis.

3.3.1 Global Random Forest

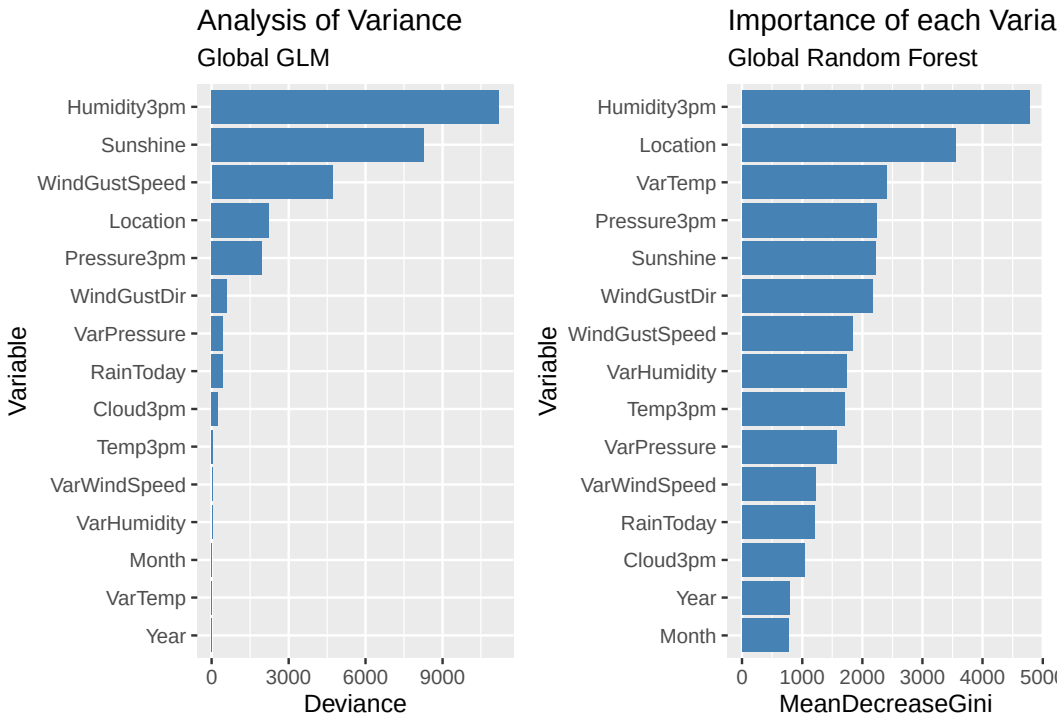
Simulations, not displayed here for the sake of space, show that training the model for *nodesize* and *mtry* optimization (two key parameters of a Random Forest model) is not really necessary, having only marginal effects on the performance of the model.

In terms of trees, we chose 50 trees based on the below plot which shows that beyond 40 trees there is little improvement in the overall accuracy.



The accuracy of the global Random Forest model is 85.2%.

The comparison with the global GLM shows that GLM and RandomForest attribute different importance to various parameters. For instance Year is not negligible in Random Forest:



3.3.2 Local Random Forest

We then build a separate RandomForest per location.

The accuracy 85.5%, just slightly higher than the global Random Forest model.

3.3.3 Random Forest conclusion

The below table shows that GLM and RF models have similar accuracies, but GLM performs better overall, especially on Sensitivity:

Table 14: Model accuracy on Test set

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Local GLM	85.8%	56.9%	94.2%	75.5%
Global Random Forest	85.2%	56.0%	93.8%	74.9%
Local Random Forest	85.5%	55.4%	94.3%	74.9%

3.4 XG Boost model

Gradient-boosted trees are generally expected to perform better than Random Forest. We will therefore try an XG Boost model. XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.

XG Boost works only with numerical data so:

- for Location, we will use latitude and longitude (this is actually more efficient than converting the Location factor to numeric)

- for the other categorical data, we convert factors to numeric.

The model was trained with simulations on *max.depth*, *eta* and *nrounds*. These simulations are not detailed here for the sake of space. The best parameters were found to be:

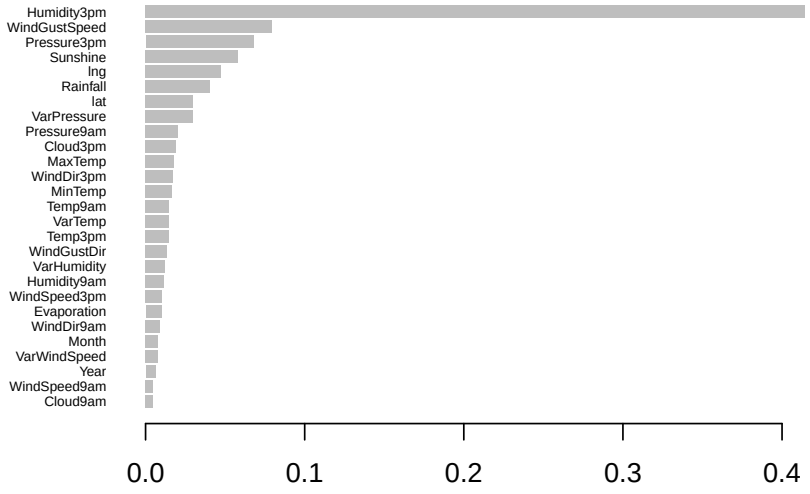
- max.depth = 6 (maximum tree depth)
- eta = 0.5 (learning rate)
- nrounds = 50 (max number of boosting iterations)

The accuracy of the model is 86.1%. XG Boost achieves the best results so far, with a significant increase in Sensitivity even versus GLM.

Table 15: Model accuracy on test set

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Global GLM	84.9%	52.3%	94.3%	73.3%
Local GLM	85.8%	56.9%	94.2%	75.5%
Global Random Forest	85.2%	56.0%	93.8%	74.9%
Local Random Forest	85.5%	55.4%	94.3%	74.9%
Global XG Boost	86.1%	58.1%	94.3%	76.2%

We can visualize the importance of each predictor in XG Boost:



3.5 PCA + GLM

We have seen that certain predictors remain correlated, even after the replacement of morning (9am) values by var. We now try to address this issue by doing a principal component analysis of the predictors, followed by a glm study.

We do not expect an improvement in accuracy from this step, as the theory predicts that results will be identical. However, some improvement could be observed if PCA allows to factor specific effects which disappeared when we removed certain non significant predictors in the previous GLM. In any case, PCA can contribute to more stable models because its components are de-correlated, and as such, is worth investigating.

We perform PCA on the numeric variables in the training set, and then perform either a Global GLM or a Local GLM, using the PCA components as well as Location & WindGustDir as predictors.

3.5.1 PCA + Global GLM

By plotting the PCA results, we see that 15 components are needed to cover 99% of the variance in the data set:

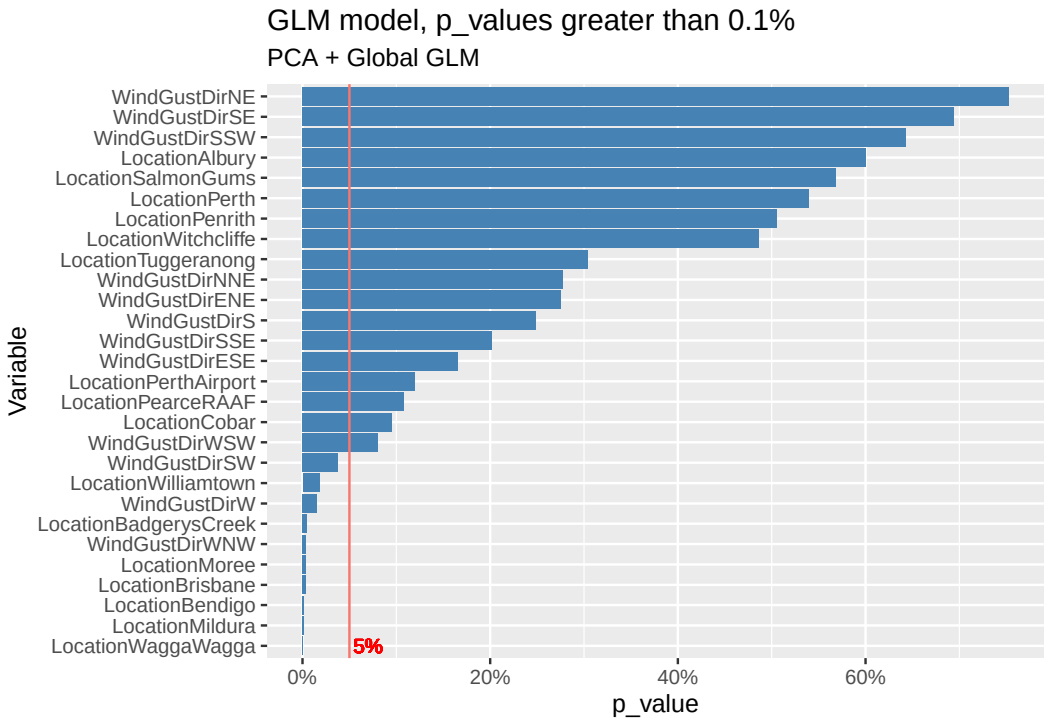
Table 16: PCA Variance

	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	2.063	0.250	0.250
PC2	1.709	0.172	0.422
PC3	1.293	0.098	0.520
PC4	1.120	0.074	0.594
PC5	1.023	0.062	0.656
PC6	1.010	0.060	0.716
PC7	0.998	0.059	0.774
PC8	0.926	0.050	0.825
PC9	0.858	0.043	0.868
PC10	0.739	0.032	0.900
PC11	0.715	0.030	0.930
PC12	0.653	0.025	0.955
PC13	0.542	0.017	0.973
PC14	0.462	0.013	0.985
PC15	0.424	0.011	0.996
PC16	0.246	0.004	0.999
PC17	0.102	0.001	1.000

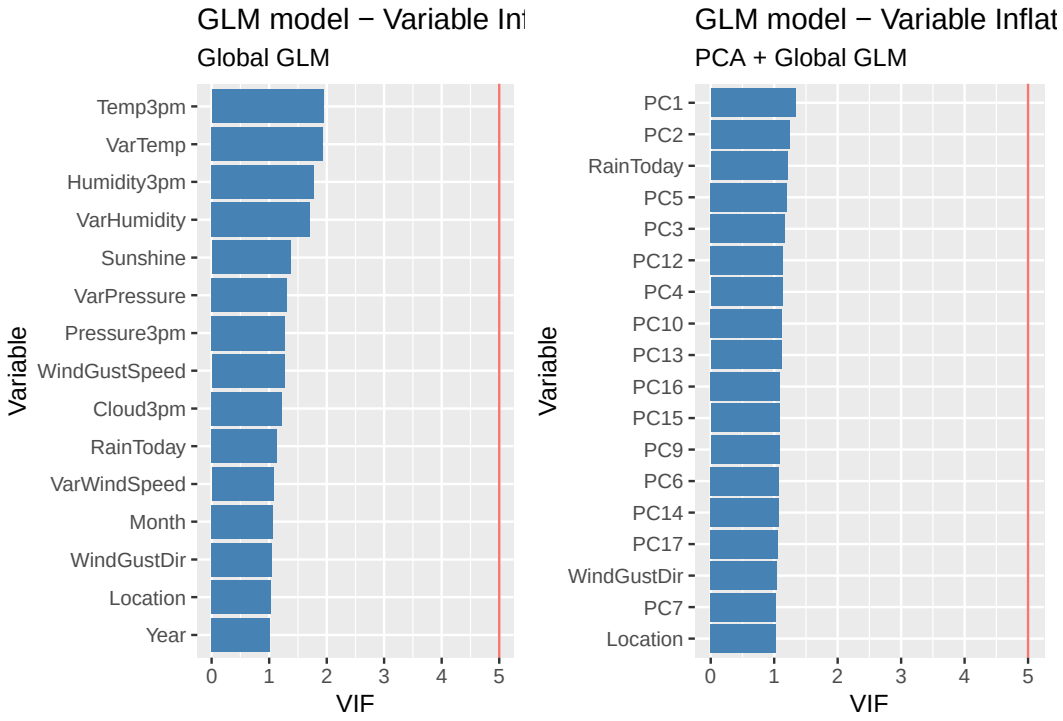
We will retain all the components, except numbers 8 and 11 which, from experience, have limited significance in the glm. The parameters used in GLM will therefore be as follows:

```
## RainTomorrow ~ Location + RainToday + WindGustDir + PC1 + PC2 +
##     PC3 + PC4 + PC5 + PC6 + PC7 + PC9 + PC10 + PC12 + PC13 +
##     PC14 + PC15 + PC16 + PC17
```

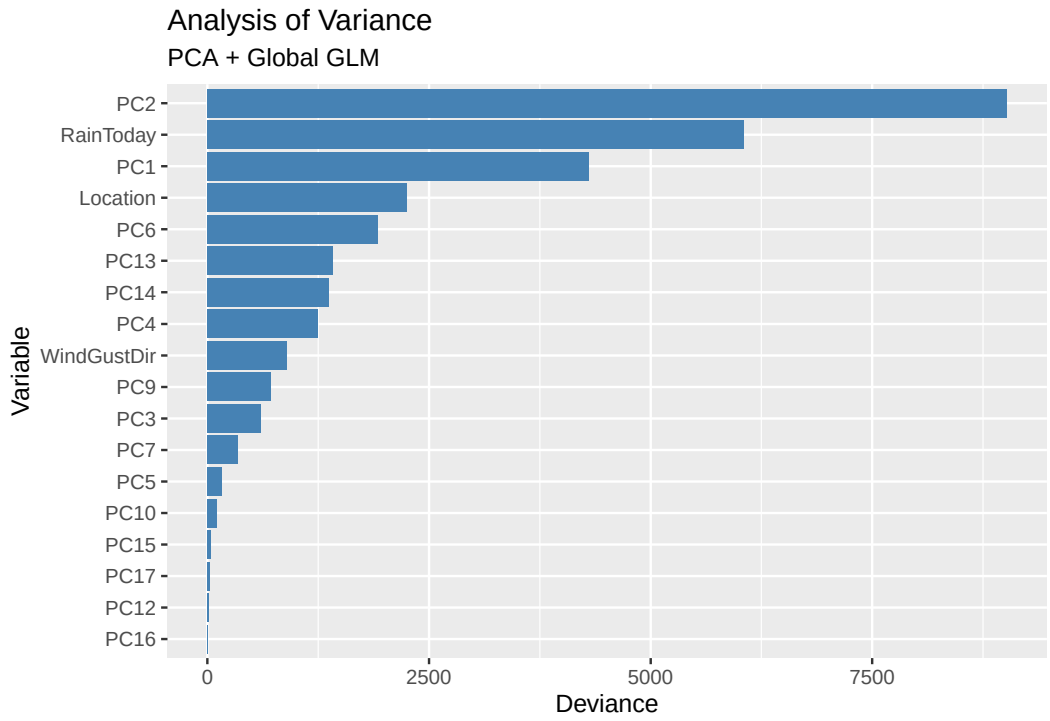
All variables are significant, only certain WindGust directions and a few locations are not, as we can see on the p_val graph (plotting only p_vals greater than 1%):



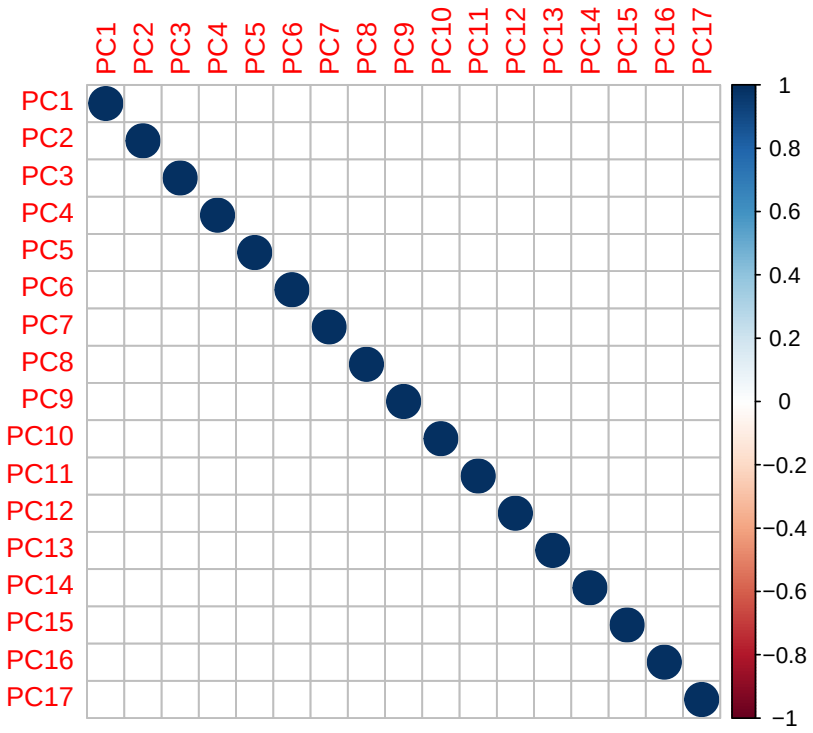
All variables have a VIF around 1, indicating perfect de-correlation. Comparing with the VIF plot of the glm with significant variables, we see that VIFs, which were already good, are even closer to 1 with PCA:



The ANOVA plot shows the most important variables at the top of the graph. PC2, RainToday, PC1, Location are the most significant. There are 3 or 4 PCs of less significance at the bottom of the graph.



We can visualize the nil correlation between PCs:



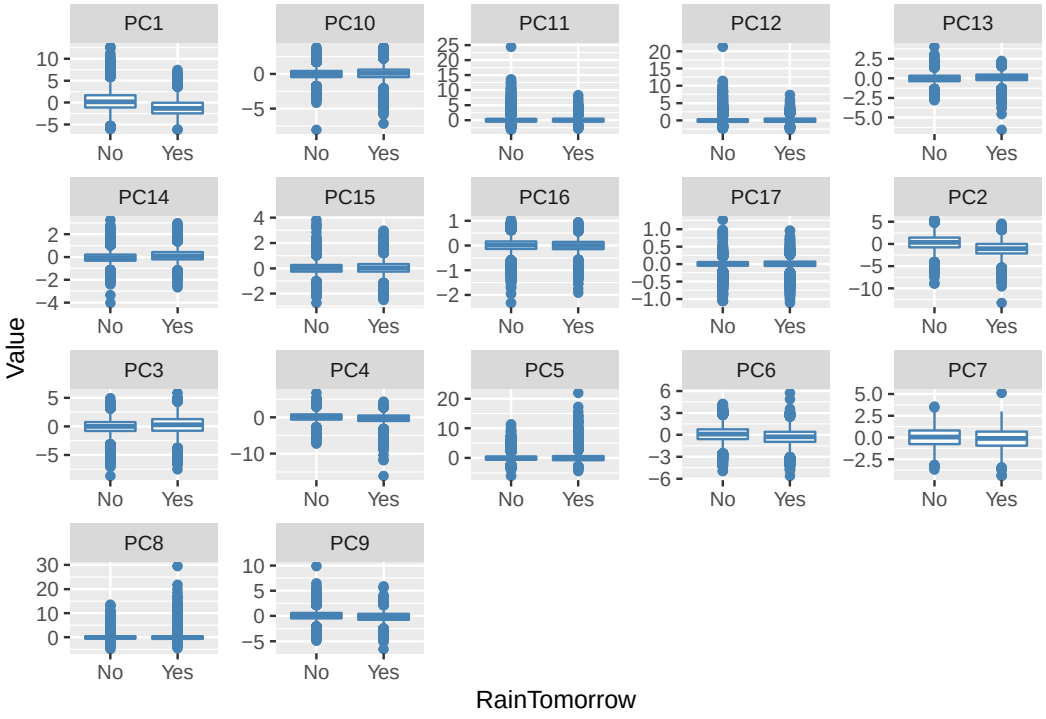
The accuracy of the model is 84.8%, which we can compare to the previously run GLM, even though the number of predictors is different.

The table shows that the global PCA-GLM model has similar results to the global GLM model as expected (differences can be due to the number of parameters retained in the analysis):

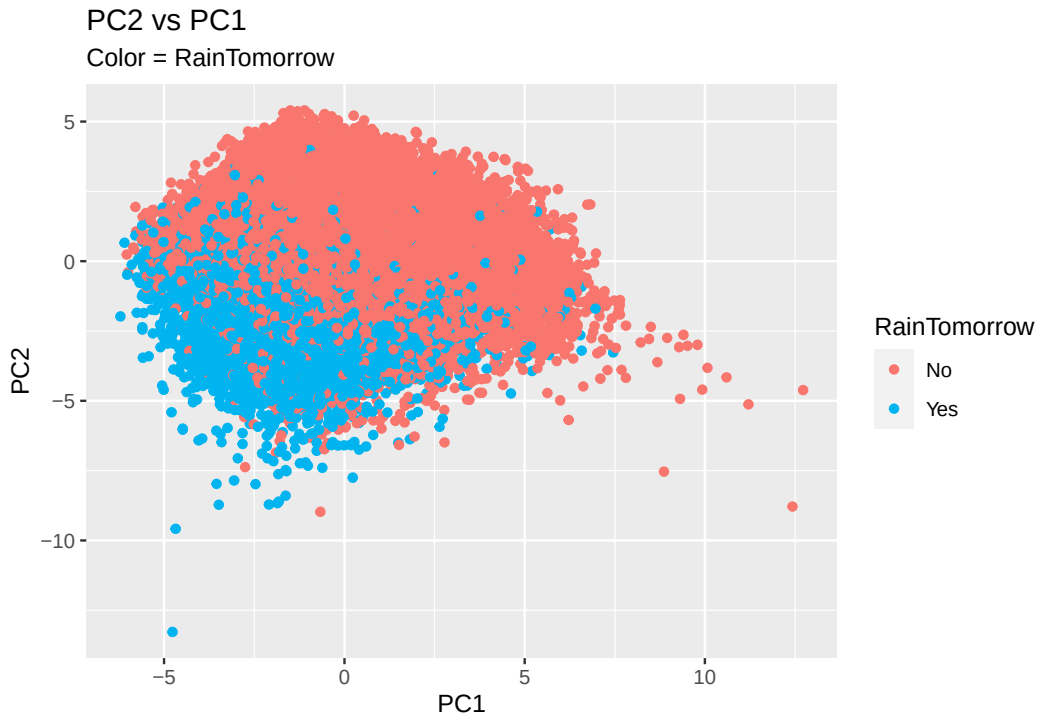
Table 17: Model accuracy on test set

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Global GLM	84.9%	52.3%	94.3%	73.3%
PCA + Global GLM	84.8%	52.1%	94.4%	73.2%

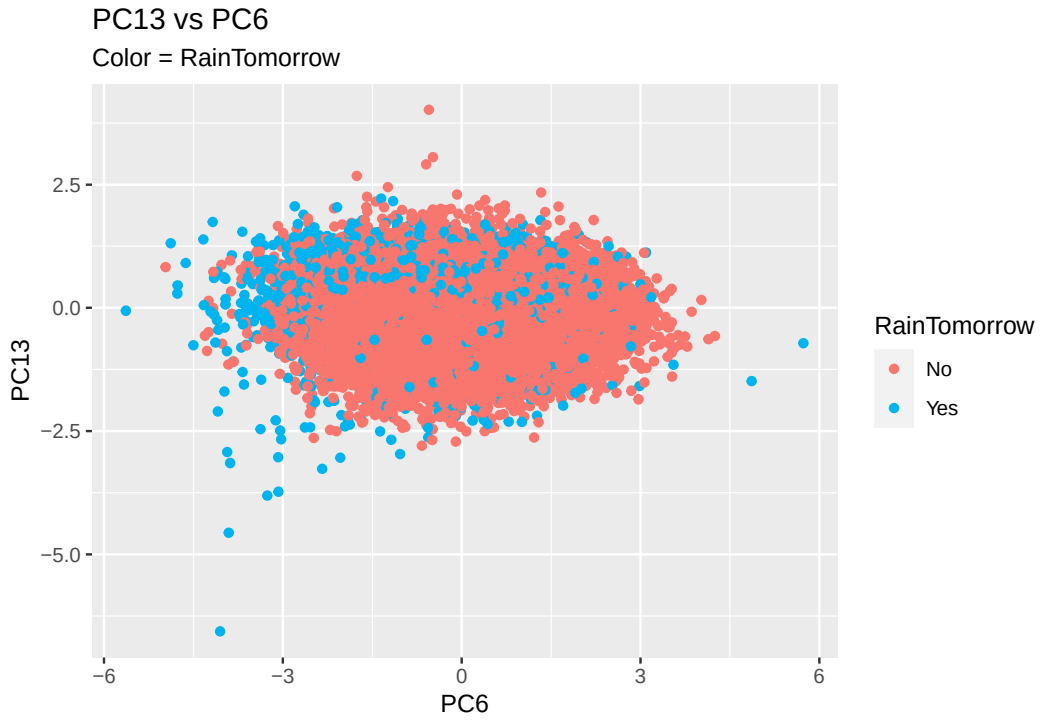
In order to understand better the limitations of the model, we plot the PCs versus RainTomorrow. The plot shows that, apart from PC1 and PC2, most PCs do not have significantly differentiated ranges between RainTomorrow No and Yes:



We can also do scatter plots of the 4 more significant PCs. The first plot shows two areas between Rain and no Rain, and it is perhaps the cleanest such graph observed so far, but still with no definite frontier:



The second plot shows that Rain & No Rain have significant overlaps on a combination of PC6 with PC13:



This confirms that PCA does not allow to split the data more efficiently between Rain and No Rain and highlights the limitations than any model will face in predicting RainTomorrow. However, PCA has the advantage of providing perfectly de-correlated and significant variables, therefore contributing to a more stable overall model.

3.5.2 PCA + Local GLM

We will now use the PCA results and perform a local GLM model. We will retain all PCs in this case, as we do not know which ones are significant at local level.

The GLM parameters will therefore be:

```
## RainTomorrow ~ RainToday + WindGustDir + PC1 + PC2 + PC3 + PC4 +
##     PC5 + PC6 + PC7 + PC8 + PC9 + PC10 + PC11 + PC12 + PC13 +
##     PC14 + PC15 + PC16 + PC17
```

The accuracy is 85.8%.

The below table shows that the performance is similar to the local GLM model, as expected:

Table 18: Model accuracy on test set

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Global GLM	84.9%	52.3%	94.3%	73.3%
PCA + Global GLM	84.8%	52.1%	94.4%	73.2%
Local GLM	85.8%	56.9%	94.2%	75.5%
PCA + Local GLM	85.8%	56.8%	94.2%	75.5%

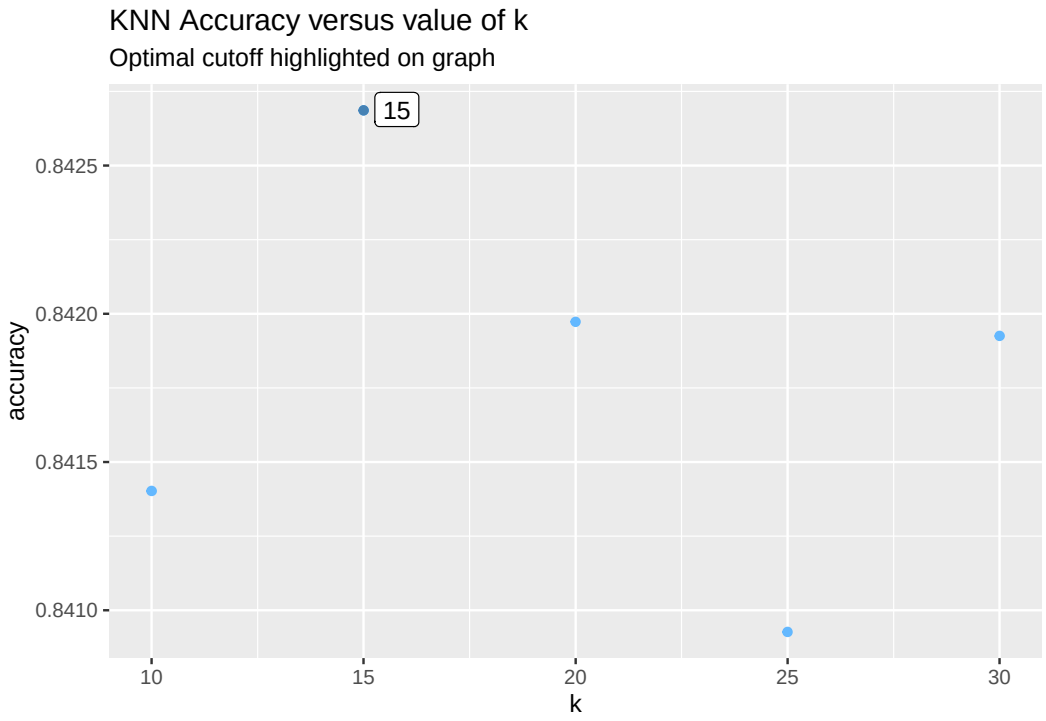
3.6 PCA + KNN model

We now try a k-nearest neighbor model. As KNN requires numeric scaled data, we use the PCA results.

The parameters are:

```
## RainTomorrow ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC9 +
##     PC10 + PC12 + PC13 + PC14 + PC15 + PC16 + PC17
```

We first train the model to select the optimal k, which is 15:



The accuracy of the global KNN model is 84.3%, however we will also try a local model.

The parameters for the local models are:

```
## RainTomorrow ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 +
##     PC9 + PC10 + PC11 + PC12 + PC13 + PC14 + PC15 + PC16 + PC17
```

The accuracy of the local model is 84.2%. The below table shows that KNN has slightly less accuracy than the other models:

Table 19: Model accuracy on test set

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
PCA + Local GLM	85.8%	56.8%	94.2%	75.5%
Local Random Forest	85.5%	55.4%	94.3%	74.9%
Global XG Boost	86.1%	58.1%	94.3%	76.2%
PCA + Local KNN	84.2%	43.5%	96.1%	69.8%

3.7 PCA + QDA

We now try a Quadratic Discriminant Analysis. QDA works with numeric, scaled data. We will use the PCA data and the same parameters as KNN. There is no parameter to be tuned.

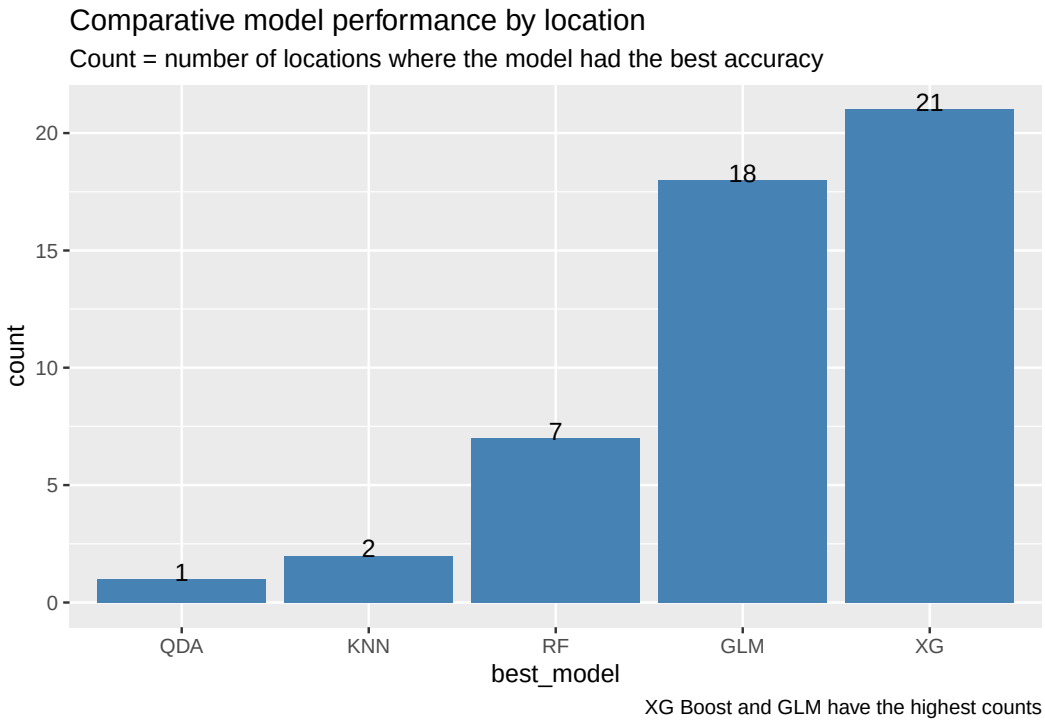
QDA models have lower accuracy than most other models, but, interestingly, they have the best Sensitivity:

Table 20: Model accuracy on test set

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
PCA + Local GLM	85.8%	56.8%	94.2%	75.5%
Local Random Forest	85.5%	55.4%	94.3%	74.9%
Global XG Boost	86.1%	58.1%	94.3%	76.2%
PCA + Local KNN	84.2%	43.5%	96.1%	69.8%
PCA + Local QDA	84.3%	60.4%	91.2%	75.8%

3.8 Model Comparaison

To compare model performance, we count how many times each model has the best accuracy per location. The plot confirms that XG and GLM are the most accurate models, not only globally but per location too. This suggests to build an ensemble to see if performance can be further improved.



3.9 Ensemble

We now build an ensemble ("Ensemble 5") based on the majority vote between the 5 main models:

- Local GLM
- Local Random Forest
- XG Boost
- Local KNN
- Local QDA

We predict Rain Yes if 3 or more of these algorithms predict rain.

We make another ensemble ("Ensemble 3") where we predict RainTomorrow based on the majority vote of 3 models: we select the best two models XG and Local GLM, and we add QDA which has the best Sensitivity (adding RF instead of QDA would improve accuracy slightly but deteriorate Sensitivity).

The accuracy of "Ensemble 5" is 86.2%.

The accuracy of "Ensemble 3" is 86.3%.

"Ensemble 3" has the best results:

Table 21: Model accuracy on test set

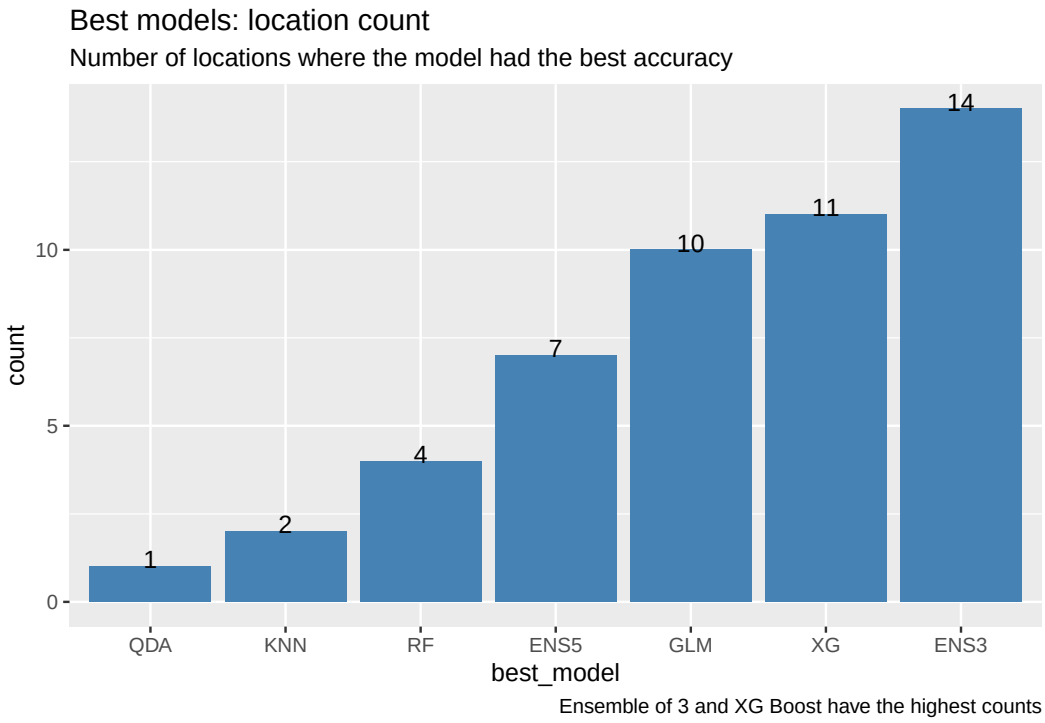
Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
PCA + Local GLM	85.8%	56.8%	94.2%	75.5%
Local Random Forest	85.5%	55.4%	94.3%	74.9%
Global XG Boost	86.1%	58.1%	94.3%	76.2%
PCA + Local KNN	84.2%	43.5%	96.1%	69.8%
PCA + Local QDA	84.3%	60.4%	91.2%	75.8%
Ensemble 5	86.2%	55.4%	95.2%	75.3%
Ensemble 3	86.3%	58.7%	94.4%	76.5%

3.10 General Summary

The table below shows all studied models, ranked by increasing accuracy then balanced accuracy:

Table 22: Models ranked by Accuracy then Balanced Accuracy

	Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
1	Predicting no rain	77.5%	0.0%	100.0%	50.0%
2	Humidity3pm with cutoff	82.3%	35.1%	96.0%	65.5%
3	PCA + Global QDA	82.3%	53.7%	90.6%	72.2%
4	GLM Location & Humidity	83.1%	39.5%	95.8%	67.6%
5	PCA + Local KNN	84.2%	43.5%	96.1%	69.8%
6	PCA + Local QDA	84.3%	60.4%	91.2%	75.8%
7	Global GLM	84.9%	52.3%	94.3%	73.3%
9	Local Random Forest	85.5%	55.4%	94.3%	74.9%
12	Local GLM	85.8%	56.9%	94.2%	75.5%
13	PCA + Local GLM	85.8%	56.8%	94.2%	75.5%
8	Global XG Boost	86.1%	58.1%	94.3%	76.2%
10	Ensemble 5	86.2%	55.4%	95.2%	75.3%
11	Ensemble 3	86.3%	58.7%	94.4%	76.5%



- In view of the table and the plot, we select as best approach the **"Ensemble 3" with three models: GLM, XG Boost and QDA**.
- Whilst it is not strictly necessary, we prefer to run a principle component analysis prior to the GLM & QDA, as this generates perfectly de-correlated predictors.
- There could also be a debate on whether using the ensemble is justified, given that XG Boost has results which are very close (for instance, only lower by 0.6% on Sensitivity). However, the bar plot shows that XG is not always the best at location level. Also, since the computation time of each model is in seconds only, there is no reason to limit ourselves to one model: we can benefit from the ensemble results with little additional cost.

4 Results

We now re-train the selected model on the entire training set ("weather_clean"). We then apply it to the separate and entirely new validation set, in order to assess its performance.

Data Preparation

The validation set contains NA data which need to be filled first. We populate these NAs with the default values already computed on the training set during model development. The validation data is thus not used to define the NA replacement values.

We can compare the number of NA values in the validation set, before NA replacement (57,558) and after NA replacement (0).

We also add the calculated variables (var between 9am and 3pm as well as Year and Month) to the validation set.

PCA

We retrain PCA on the full weather training set and apply the result to the validation set.

Table 23: PCA Variance

	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	2.064	0.251	0.251
PC2	1.711	0.172	0.423
PC3	1.292	0.098	0.521
PC4	1.120	0.074	0.595
PC5	1.023	0.062	0.656
PC6	1.010	0.060	0.716
PC7	0.998	0.059	0.775
PC8	0.926	0.050	0.825
PC9	0.856	0.043	0.869
PC10	0.739	0.032	0.901
PC11	0.712	0.030	0.930
PC12	0.651	0.025	0.955
PC13	0.542	0.017	0.973
PC14	0.462	0.013	0.985
PC15	0.424	0.011	0.996
PC16	0.247	0.004	0.999
PC17	0.103	0.001	1.000

4.1 Global Results

We apply the local GLM, local QDA and XG Boost models to the validation set.

The accuracy on the validation set is 86.2%. The results table is as follows:

Table 24: Model accuracy on Validation set

Model	Accuracy	Sensitivity	Specificity	Balanced_Acc
Validation set	86.2%	58.2%	94.4%	76.3%

We can also view the predictive values:

- The accuracy of predicting Rain is 75.1% (precision or positive predicting value)
- The accuracy of predicting No-Rain is 88.6%

Table 25: Predictive values

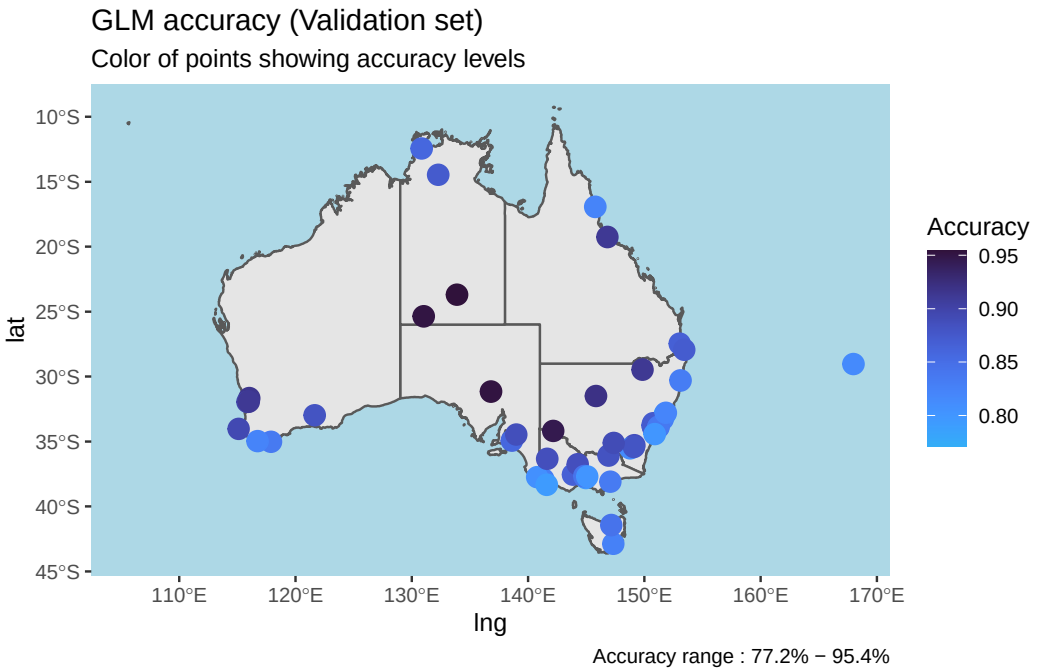
	Rain_Accuracy	No_Rain_Accuracy
Validation set	75.1%	88.6%

The confusionMatrix is as follows:

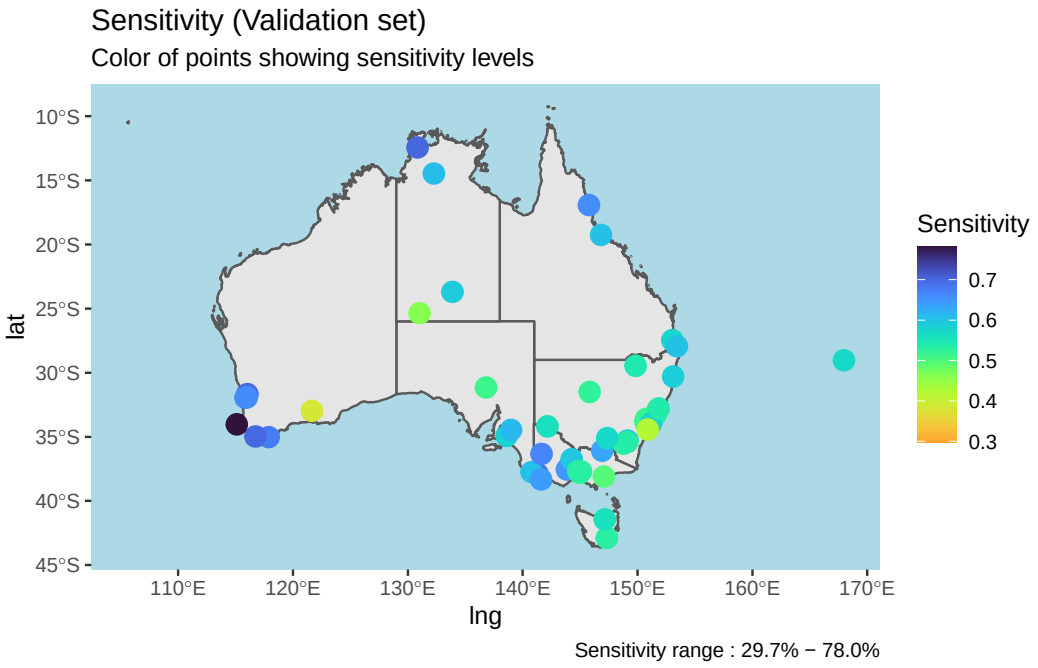
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    No   Yes
##        No  19228  2475
##        Yes  1140  3442
##
##                Accuracy : 0.8625
##                  95% CI : (0.8582, 0.8666)
##     No Information Rate : 0.7749
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5715
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.5817
##             Specificity : 0.9440
##          Pos Pred Value : 0.7512
##          Neg Pred Value : 0.8860
##              Prevalence : 0.2251
##          Detection Rate : 0.1309
##    Detection Prevalence : 0.1743
##       Balanced Accuracy : 0.7629
##
##        'Positive' Class : Yes
##
```

4.2 Local results

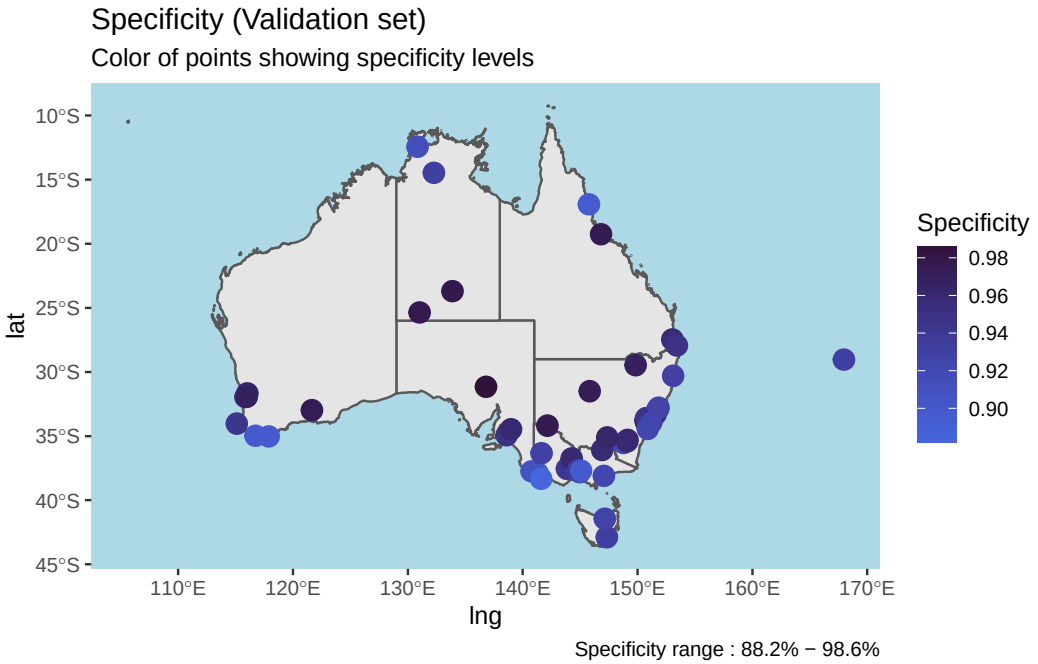
We plot the accuracy of each location on the map. Although the overall accuracy is quite good in many places, the accuracy remains lower in coastal locations. As stated earlier, these are areas where we expect the weather to change frequently, due to the oceanic influence.



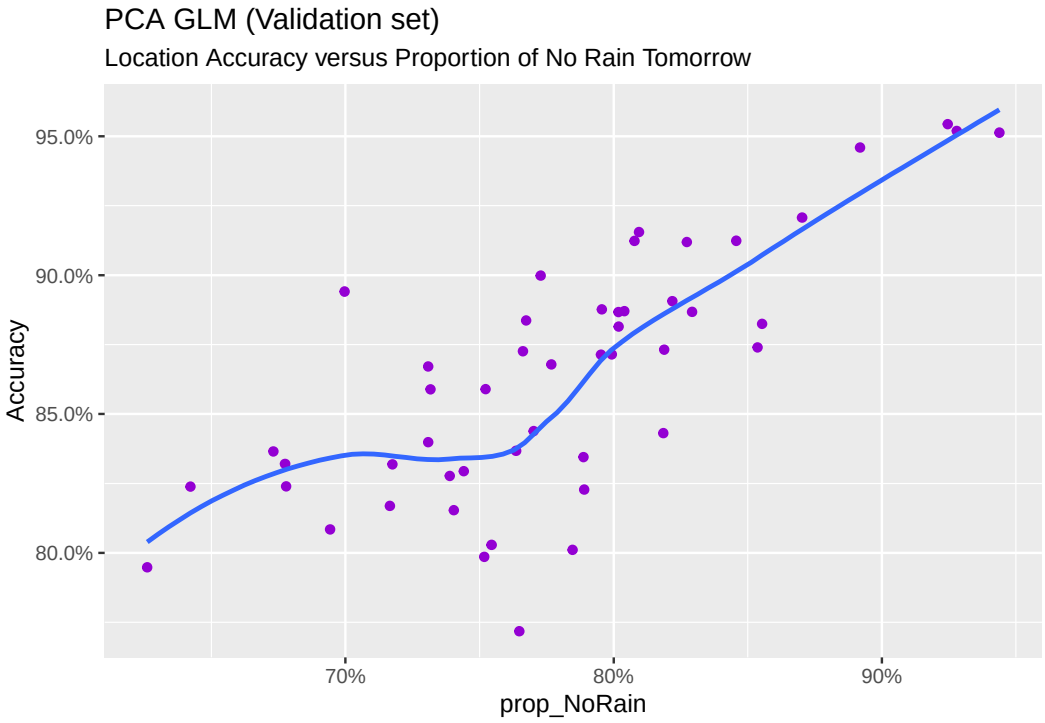
It is interesting to also plot the Sensitivity and Specificity by location. We start with Sensitivity. The map shows the sensitivity is very weak (<50%) in a few locations only. Sensitivity is best in Western and Northern locations.



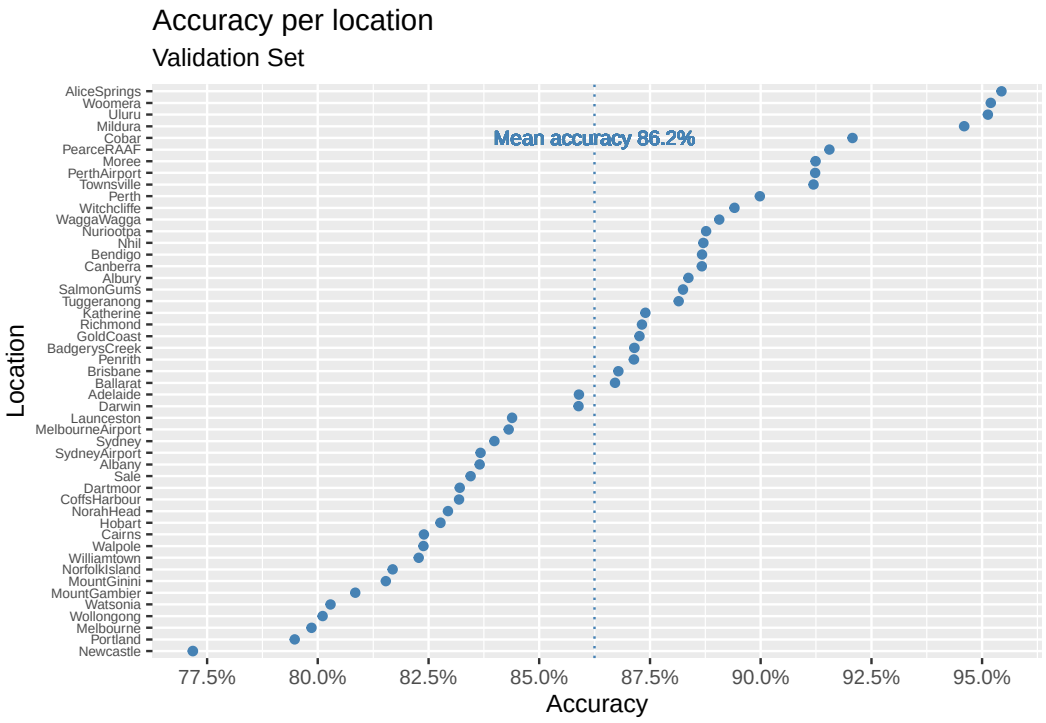
We now plot Specificity. The map shows strong specificity across virtually all locations



We also plot the accuracy for each location versus the No-Rain percentage in the location. The plot shows some link between No-Rain prevalence and accuracy, albeit not a very strong one.



Finally, we plot the locations by increasing accuracy: Three locations have an accuracy below 80% (the minimum being Newcastle around 77.5%), all others have greater accuracy than 80%, of which 10 have accuracies 90% or above.


Accuracy per location
Validation Set
Mean accuracy 86.2%
Location
AliceSprings
Woomera
Uluru
Mildura
Cobar
PearceRAAF
Moree
PerthAirport
Townsville
Perth
Witchcliffe
WaggaWagga
Nuriootpa
Nhil
Bendigo
Canberra
Albury
SalmonGums
Tuggeranong
Katherine
Richmond
GoldCoast
BadgerysCreek
Penrith
Brisbane
Ballarat
Adelaide
Darwin
Launceston
MelbourneAirport
Sydney
SydneyAirport
Albany
Sale
Dartmoor
CoffsHarbour
NorahHead
Hobart
Cairns
Walpole
Williamtown
NorfolkIsland
MountGinini
MountGambier
Watsonia
Wollongong
Melbourne
Portland
Newcastle
77.5%
80.0%
82.5%
85.0%
87.5%
90.0%
92.5%
95.0%
Accuracy

5 Conclusion

The model is successful in predicting Rain Tomorrow with reasonable accuracy. We note, however, two areas that would certainly deserve more research:

- improving results where the model performs less well, mainly coastal locations in southern Australia
- improving the sensitivity of the model, ie the capacity to correctly predict Rain Yes

The following developments could be explored:

- Trying other algorithms: we have seen that Rain Tomorrow remains a highly random variable versus the various predictors in the data set, and there will be a limit to what algorithms can do. However some improvements can certainly be sought.
- Performing more work on the NA replacement techniques. Whilst we have seen than NA replacement had little impact on the general results, this could still be an area of improvement
- Adding data to the data set: having only two daily measurements might no be enough especially for coastal locations. Adding some measurements could probably improve results.

We close this report with thanks to Joe Young for posting this very interesting challenge.

END OF REPORT.