# RainInAustralia

Filippo, Antoni, Cristina, Mengxue

6/11/2021

## Problem description

The problem is to predict whether it will rain the next day or not based on measurements taken by the australian's government meteorological department. The measurements have been taken daily in various regions of Australia, with over 10 years of measurements. Due to the nature of the instruments used, some measurements are missing heavily in some areas, or are mostly incomplete.

The purpose of our work is to find insights into the weather data, to see if we can ascend some patterns that might help better understand rain prediction.

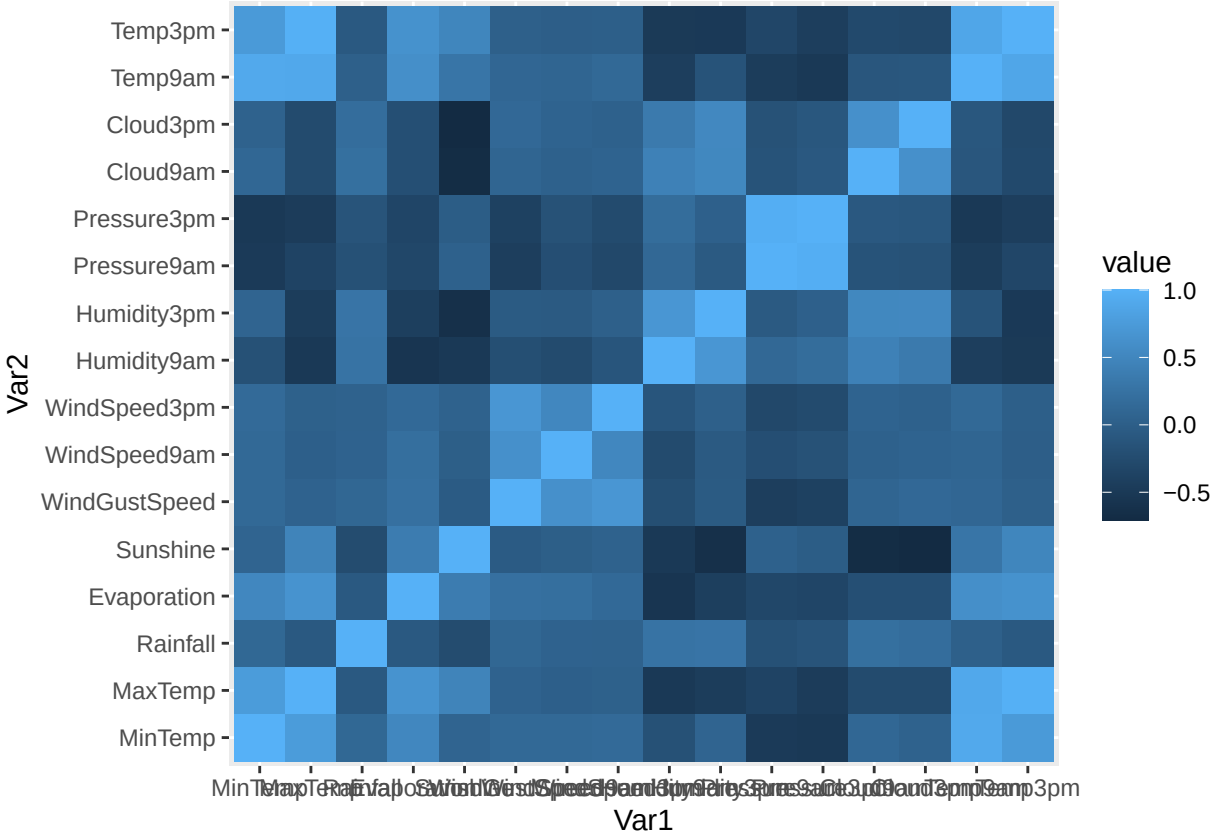
## Dataset description

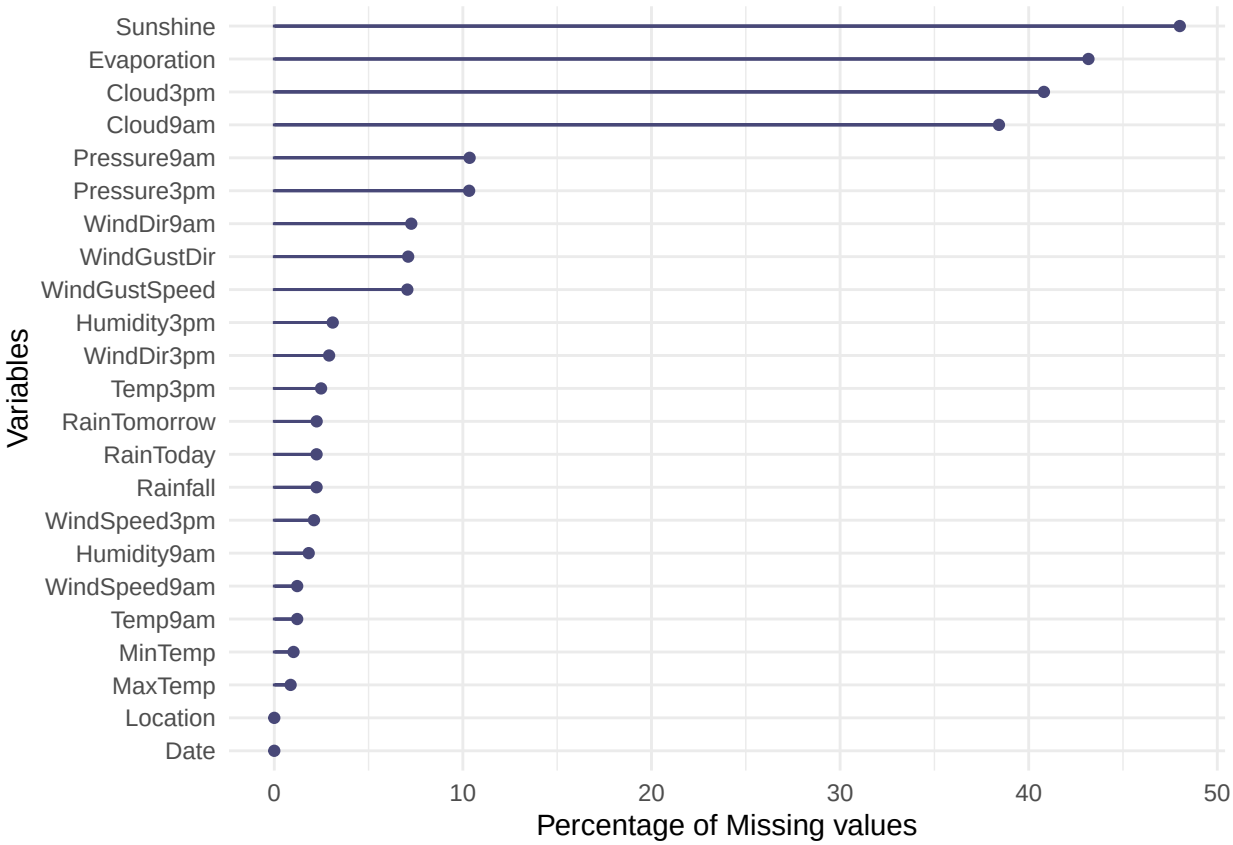
The dataset is compromised of 23 variables, and is a timeseries of australian weather, which the purpose of predicting whether it would rain tomorrow.

- **Date** - Categorical variable, when the measurements were taken
- **Location** - Categorical variable, where the measurements were taken
- **MinTemp** - Numerical variable, minimal temperature observed that day
- **MaxTemp** - Numerical variable, maximal temperature observed that day
- **Rainfall** - Numerical variable, precipitation in the 24hours to 9am
- **Evaporation** - Numerical variable, "Class A" pan evaporation in the 24 hours to 9am
- **Sunshine** - Numerical variablem, bright sunshine hours in the 24 hours to midnight
- **WindGustDir** - Categorical variable, direction of strongest gust in the 24 hours to midnight, 16 compass points
- **WindGustSpeed** - Numerical variable, speed of strongest wind gust in the 24 hours to midnight
- **WindDir9am** - Categorical variable, wind direction averaged over 10 minutes prior to 9 am
- **WindDir3pm** - Categorical variable, wind direction averaged over 10 minutes prior to 3 pm
- **WindSpeed9am** - Numerical variable, wind speed averaged over 10 minutes prior to 9 am
- **Windspeed3pm** - Numerical variable, wind speed averaged over 10 minutes prior to 9 am
- **Humidity9am** - Numerical variable, relative humidity at 9 am
- **Humidity3pm** - Numerical variable, relative humidity at 3 pm
- **Pressure9am** - Numerical variable, atmospheric pressure reduced to mean sea level at 9 am
- **Pressure3pm** - Numerical variable, atmospheric pressure reduced to mean sea level at 3 pm
- **Cloud9am** - Numerical variable, fraction of sky obscured by cloud at 9 am
- **Cloud3pm** - Numerical variable, fraction of sky obscured by cloud at 3 pm
- **Temp9am** - Numerical variable, temperature in Celsius at 9am
- **Temp3pm** - Numerical variable, temperature in Celsius at 3pm
- **RainToday** - Categorical variable, whether it rained today or not
- **RainTomorrow** - Categorical variable, whether tomorrow will rain or not

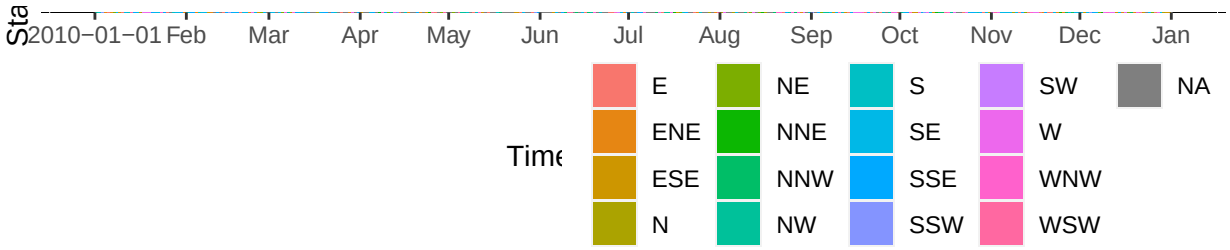
## Preprocessing

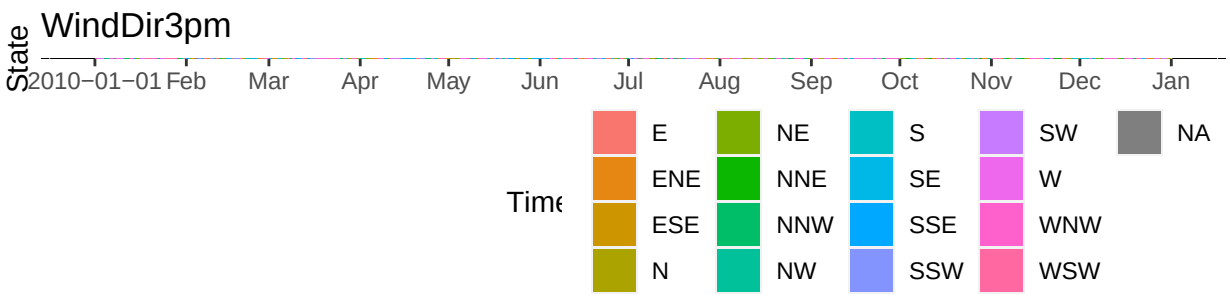
We perform a basic visualization, first of the correlation between variables, which isn't significant with the exception of variables recorded in the same day, that is, those measurements taken at 9am and 3pm, this helps us see that there's an important temporal component in the same day.

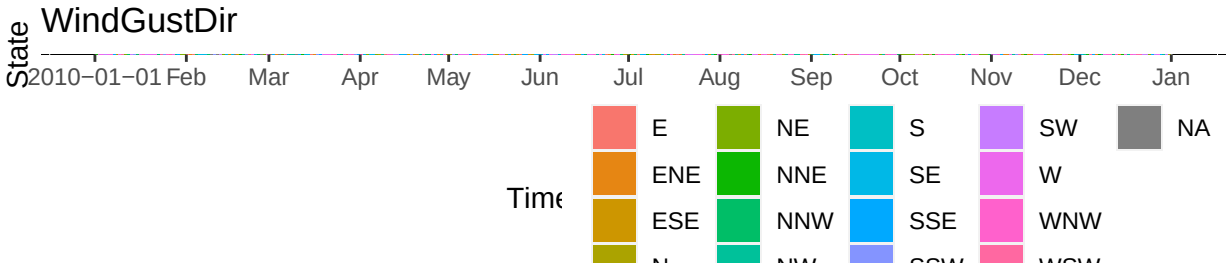





We perform a special method of data imputation, following the timeseries plot of the WinDir9am, WindDir3pm and WindGustDir, we can see that if the last day was a certain category, it will probably be that same category. So we choose this as our method of imputation for categorical NAs.

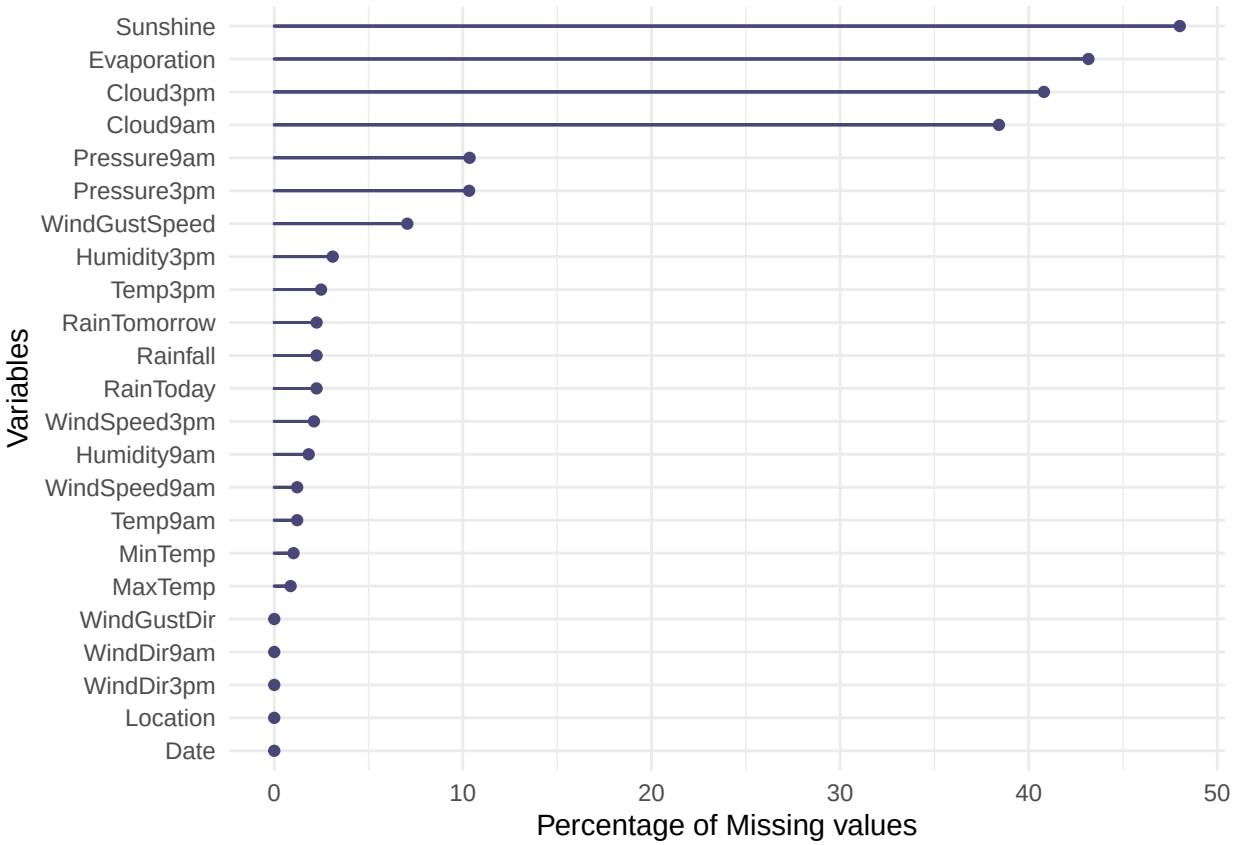


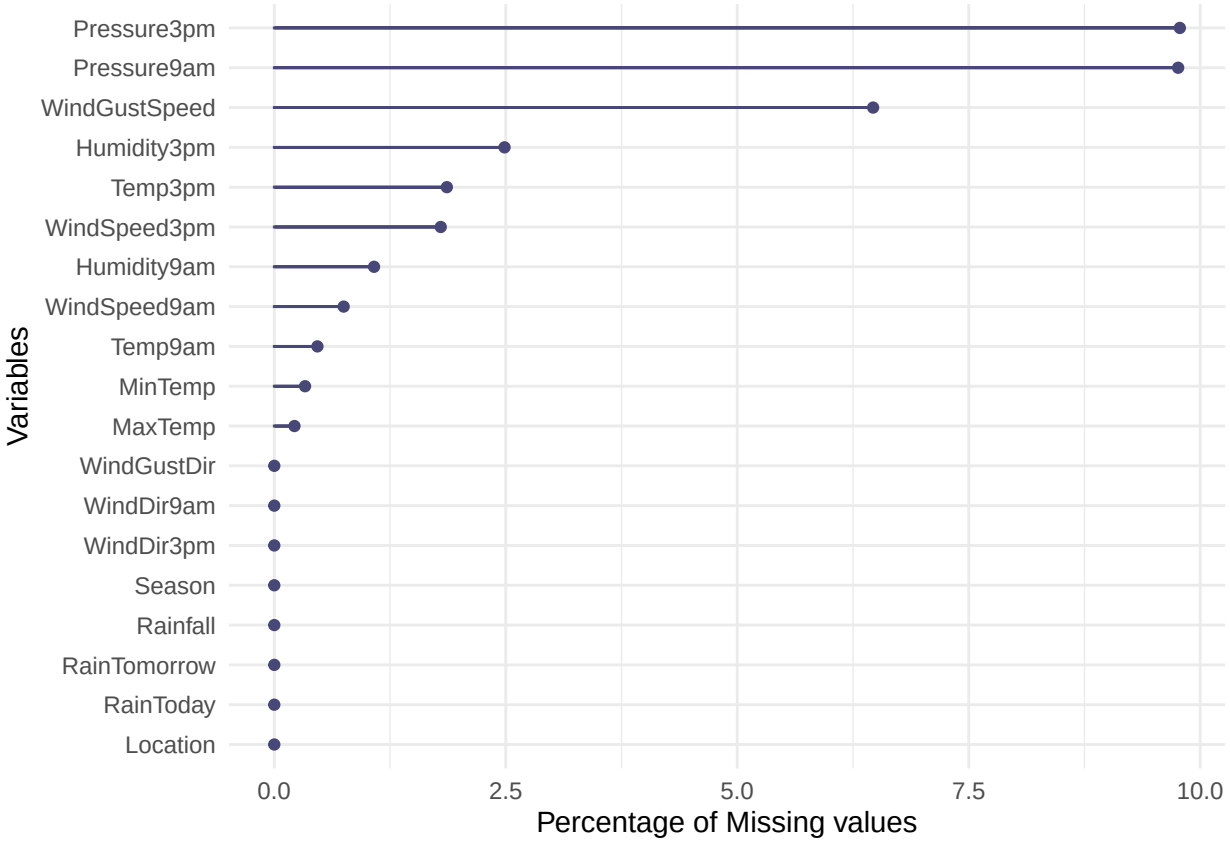

```
## INFO [2021-06-06 09:54:08] Date has been selected as the timestamp column
## INFO [2021-06-06 09:54:08]  has been selected as the numeric column(s)
## INFO [2021-06-06 09:54:08] WindDir9am, WindDir3pm, WindGustDir has been selected as the state column
## INFO [2021-06-06 09:54:08] creating state plot layers
```

We remove the columns with over 30% NAs, as imputation might be too imprecise when over a third of data is missing, and dropping 30% of data might be too excesive. We also remove all NAs, which are 2% from RainToday and RainTomorrow, as RainTomorrow is the variable to predict, and any imputation will change the real space, and RainToday because it is highly rellated to RainTomorrow and might worsen our prediction. To reduce the effect of the temporality of data we transform Date into the new variable Season, which is an approximation of the season to which the date belongs to.




```
##    Location           MinTemp          MaxTemp          Rainfall
##  Length:140787      Min.   :-8.50000   Min.   :-4.80000   Min.   :  0.000000
##  Class :character   1st Qu.: 7.60000   1st Qu.:17.90000   1st Qu.:  0.000000
##  Mode  :character   Median :12.00000   Median :22.60000   Median :  0.000000
##                     Mean   :12.18482   Mean   :23.23512   Mean   :  2.349974
##                     3rd Qu.:16.80000   3rd Qu.:28.30000   3rd Qu.:  0.800000
##                     Max.   :33.90000   Max.   :48.10000   Max.   :371.000000
##                     NA's   :468        NA's   :307
##   Evaporation          Sunshine       WindGustDir        WindGustSpeed
##  Min.   :  0.00000   Min.   : 0.00000   Length:140787      Min.   :  6.00000
##  1st Qu.:  2.60000   1st Qu.: 4.90000   Class :character   1st Qu.: 31.00000
##  Median :  4.80000   Median : 8.50000   Mode  :character   Median : 39.00000
##  Mean   :  5.47252   Mean   : 7.63054                      Mean   : 39.97052
##  3rd Qu.:  7.40000   3rd Qu.:10.70000                      3rd Qu.: 48.00000
##  Max.   :145.00000   Max.   :14.50000                      Max.   :135.00000
##  NA's   :59694       NA's   :66805                         NA's   :9105
##   WindDir9am         WindDir3pm         WindSpeed9am       WindSpeed3pm
##  Length:140787      Length:140787      Min.   :  0.0000   Min.   : 0.00000
##  Class :character   Class :character   1st Qu.:  7.0000   1st Qu.:13.00000
##  Mode  :character   Mode  :character   Median : 13.0000   Median :19.00000
##                                        Mean   : 13.9905   Mean   :18.63114
##                                        3rd Qu.: 19.0000   3rd Qu.:24.00000
##                                        Max.   :130.0000   Max.   :87.00000
##                                        NA's   :1055       NA's   :2531
##   Humidity9am        Humidity3pm        Pressure9am      Pressure3pm
##  Min.   :  0.00000   Min.   :  0.00000   Min.   : 980.500   Min.   : 977.100
##  1st Qu.: 57.00000   1st Qu.: 37.00000   1st Qu.:1013.000   1st Qu.:1010.400
```

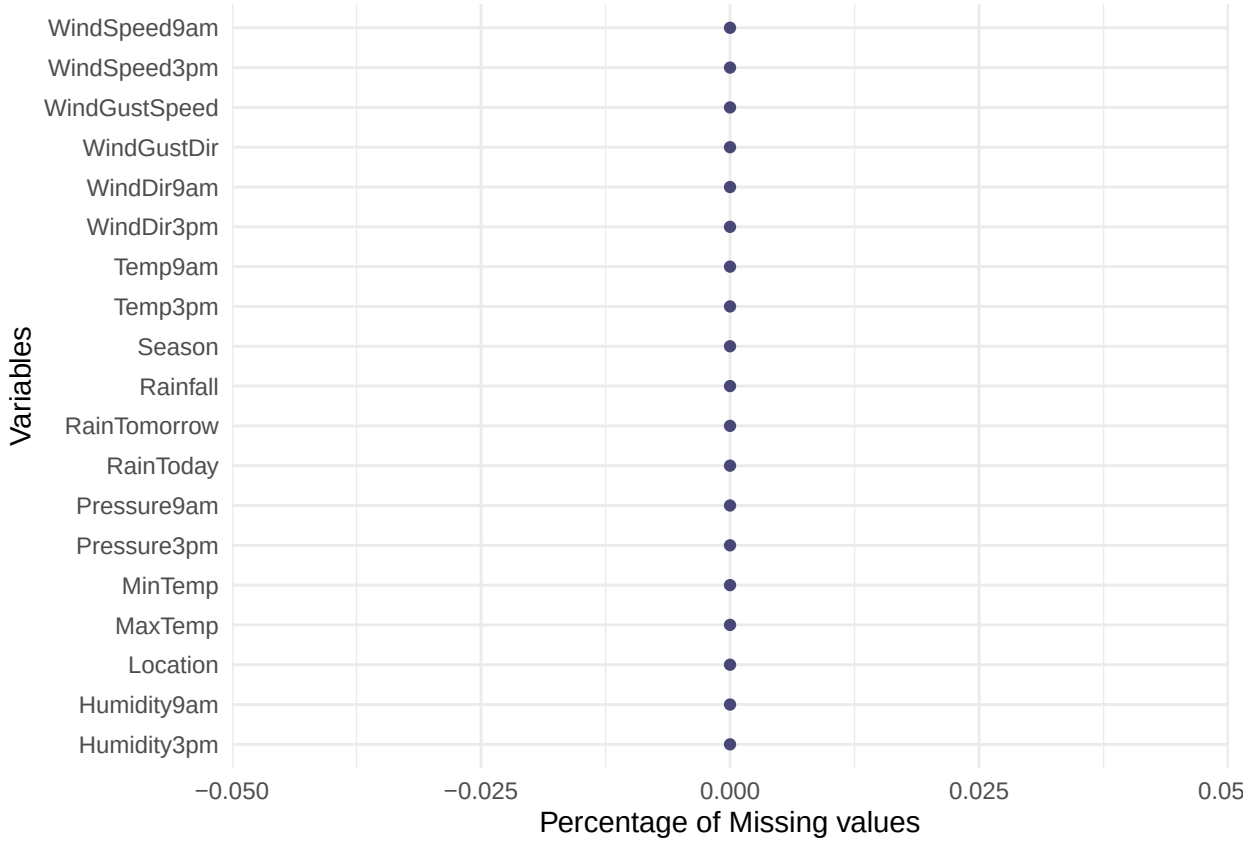
```
##  Median : 70.00000   Median : 52.00000   Median :1017.600   Median :1015.200
##  Mean   : 68.82683   Mean   : 51.44929   Mean   :1017.655   Mean   :1015.258
##  3rd Qu.: 83.00000   3rd Qu.: 66.00000   3rd Qu.:1022.400   3rd Qu.:1020.000
##  Max.   :100.00000   Max.   :100.00000   Max.   :1041.000   Max.   :1039.600
##  NA's   :1517        NA's   :3501        NA's   :13743      NA's   :13769
##     Cloud9am         Cloud3pm          Temp9am           Temp3pm
##  Min.   :0.00000   Min.   :0.00000   Min.   :-7.20000   Min.   :-5.40000
##  1st Qu.:1.00000   1st Qu.:2.00000   1st Qu.:12.30000   1st Qu.:16.60000
##  Median :5.00000   Median :5.00000   Median :16.70000   Median :21.10000
##  Mean   :4.43116   Mean   :4.49925   Mean   :16.98707   Mean   :21.69318
##  3rd Qu.:7.00000   3rd Qu.:7.00000   3rd Qu.:21.60000   3rd Qu.:26.40000
##  Max.   :9.00000   Max.   :9.00000   Max.   :40.20000   Max.   :46.70000
##  NA's   :52625     NA's   :56094     NA's   :656        NA's   :2624
##   RainToday         RainTomorrow         Season
##  Length:140787      Length:140787      winter:33981
##  Class :character   Class :character   spring:37027
##  Mode  :character   Mode  :character   summer:35526
##                                        fall  :34253
##
##
##
```



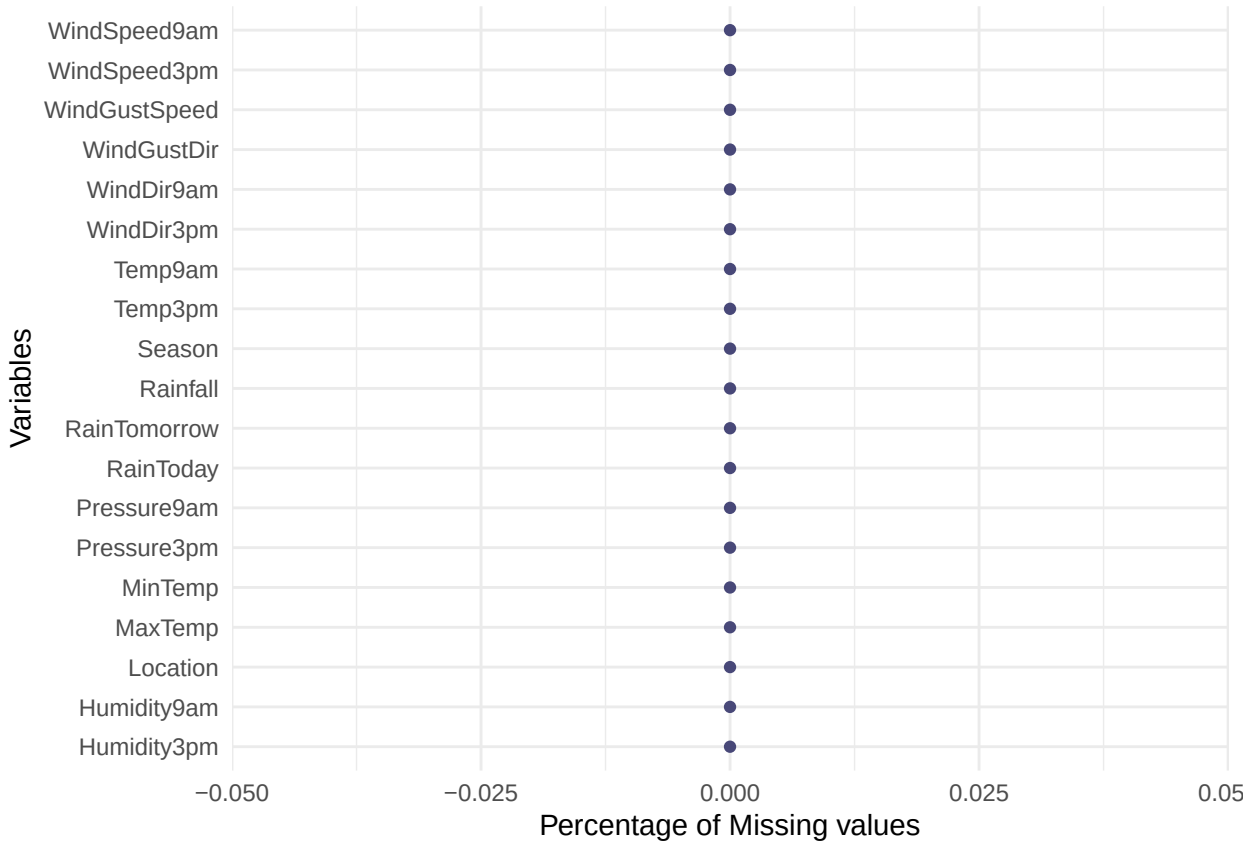

We perform the imputation of the missing continous data, however, to avoid data leakage from train into test, we separate the data into train and test, and build the imputation MICE predictive mean model on the train data, and apply it to both train and test.

```
library(mice)
library(tidyr)
```

```
completeVector=c(1:nrow(australianWeather))

completeVector[trainIndex]=TRUE
completeVector[-trainIndex]=FALSE

cVec=!(!completeVector)

imputed <- mice(australianWeather, m=5,ignore = cVec, maxit = 5, method = 'pmm', seed = 500)

australianWeatherNoNA=complete(imputed,1)

gg_miss_var(australianWeatherNoNA,show_pct = TRUE) + labs(y = "Percentage of Missing values")
```




```
gg_miss_var(australianWeatherNoNA,show_pct = TRUE) + labs(y = "Percentage of Missing values")
```

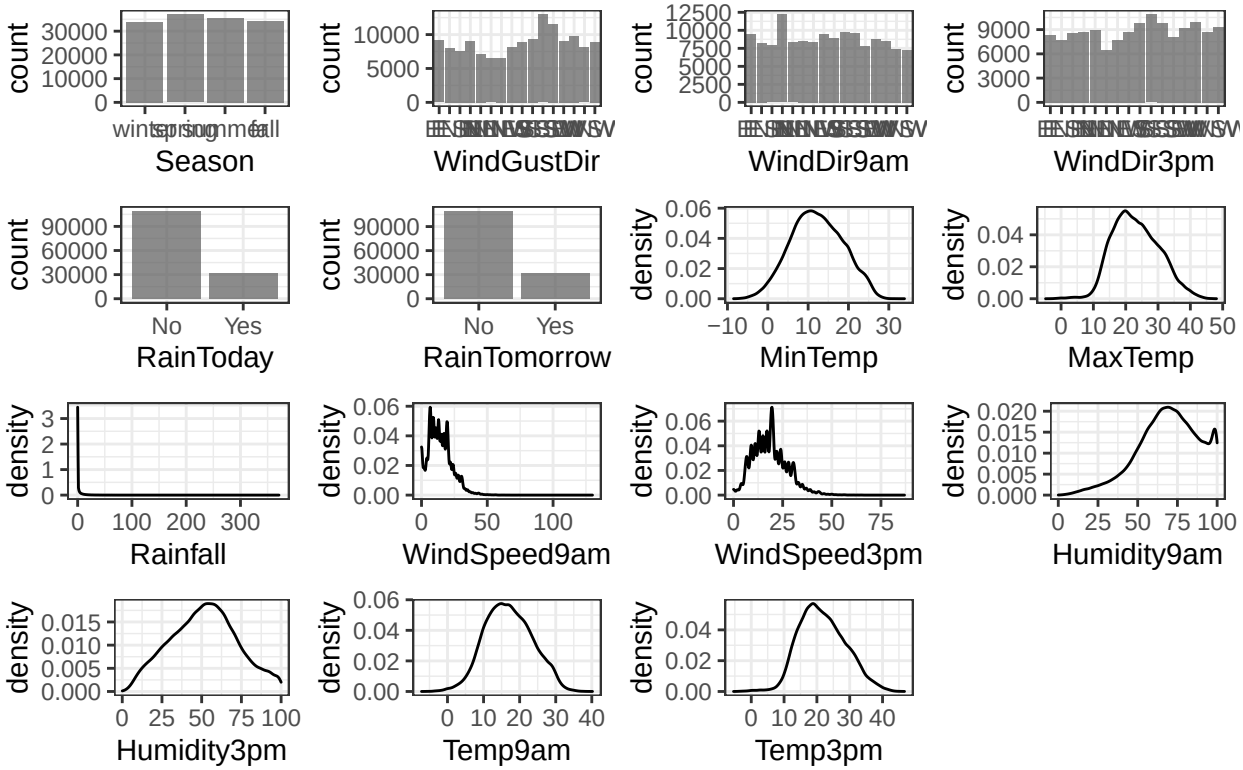



We plot the density distributions of the data, we can observe a gaussian distribution in MinTemp, MaxTemp, Humidity3pm, Temp9am and Temp3pm. A mixture of gaussians can be observed in Humidity9am, and, if we consider each peak in the WindSpeed9am and WindSpeed3pm a gaussian, a extreme version of a mixture of gaussians is present in these variables. All the categorical variables, with the exception of RainTomorrow and RainToday have mostly equal distributions, the only major imbalance being in these two variables.

Rainfall does not conform to a Gaussian distribution, and a transformation must be applied specifically for it.

```
##    Location           MinTemp          MaxTemp          Rainfall
##  Length:140787      Min.   :-8.50000   Min.   :-4.80000   Min.   :  0.000000
##  Class :character   1st Qu.: 7.60000   1st Qu.:17.90000   1st Qu.:  0.000000
##  Mode  :character   Median :12.00000   Median :22.60000   Median :  0.000000
##                     Mean   :12.18644   Mean   :23.23593   Mean   :  2.349974
##                     3rd Qu.:16.80000   3rd Qu.:28.20000   3rd Qu.:  0.800000
##                     Max.   :33.90000   Max.   :48.10000   Max.   :371.000000
##  WindGustDir        WindGustSpeed       WindDir9am         WindDir3pm
##  Length:140787      Min.   :  6.00000   Length:140787      Length:140787
##  Class :character   1st Qu.: 30.00000   Class :character   Class :character
##  Mode  :character   Median : 39.00000   Mode  :character   Mode  :character
##                     Mean   : 39.77045
##                     3rd Qu.: 48.00000
##                     Max.   :135.00000
##   WindSpeed9am       WindSpeed3pm     Humidity9am         Humidity3pm
##  Min.   :  0.0000   Min.   : 0.000   Min.   :  0.00000   Min.   :  0.00000
##  1st Qu.:  7.0000   1st Qu.:13.000   1st Qu.: 57.00000   1st Qu.: 37.00000
##  Median : 13.0000   Median :19.000   Median : 70.00000   Median : 52.00000
##  Mean   : 13.9913   Mean   :18.617   Mean   : 68.88555   Mean   : 51.52619
##  3rd Qu.: 19.0000   3rd Qu.:24.000   3rd Qu.: 83.00000   3rd Qu.: 66.00000
```

```
##  Max.   :130.0000   Max.   :87.000   Max.   :100.00000   Max.   :100.00000
##   Pressure9am        Pressure3pm          Temp9am             Temp3pm
##  Min.   : 980.500   Min.   : 977.100   Min.   :-7.20000   Min.   :-5.40000
##  1st Qu.:1013.000   1st Qu.:1010.500   1st Qu.:12.20000   1st Qu.:16.60000
##  Median :1017.700   Median :1015.300   Median :16.70000   Median :21.10000
##  Mean   :1017.688   Mean   :1015.301   Mean   :16.96643   Mean   :21.72553
##  3rd Qu.:1022.400   3rd Qu.:1020.000   3rd Qu.:21.60000   3rd Qu.:26.50000
##  Max.   :1041.000   Max.   :1039.600   Max.   :40.20000   Max.   :46.70000
##   RainToday          RainTomorrow          Season
##  Length:140787      Length:140787      winter:33981
##  Class :character   Class :character   spring:37027
##  Mode  :character   Mode  :character   summer:35526
##                                        fall  :34253
##
##
```



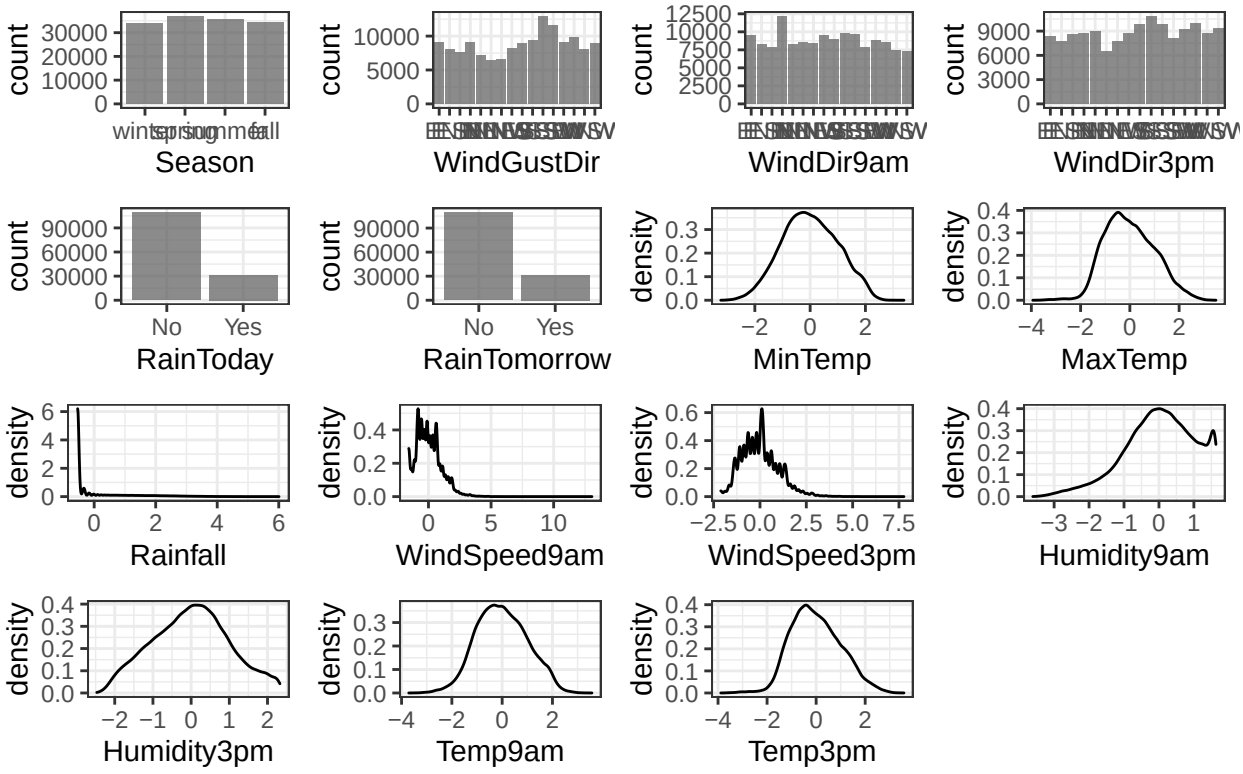

A logarithmic transformation is applied to the rainfall variable, adding a constant value of 1 to deal with zeroes, this is to get Rainfall to a shape closer to a Gaussian, being the variable most far from a Gaussian distribution.

We scale the data to a mean of 0 and variance of 1, so as to be compatible with methods sensible to distance metrics.

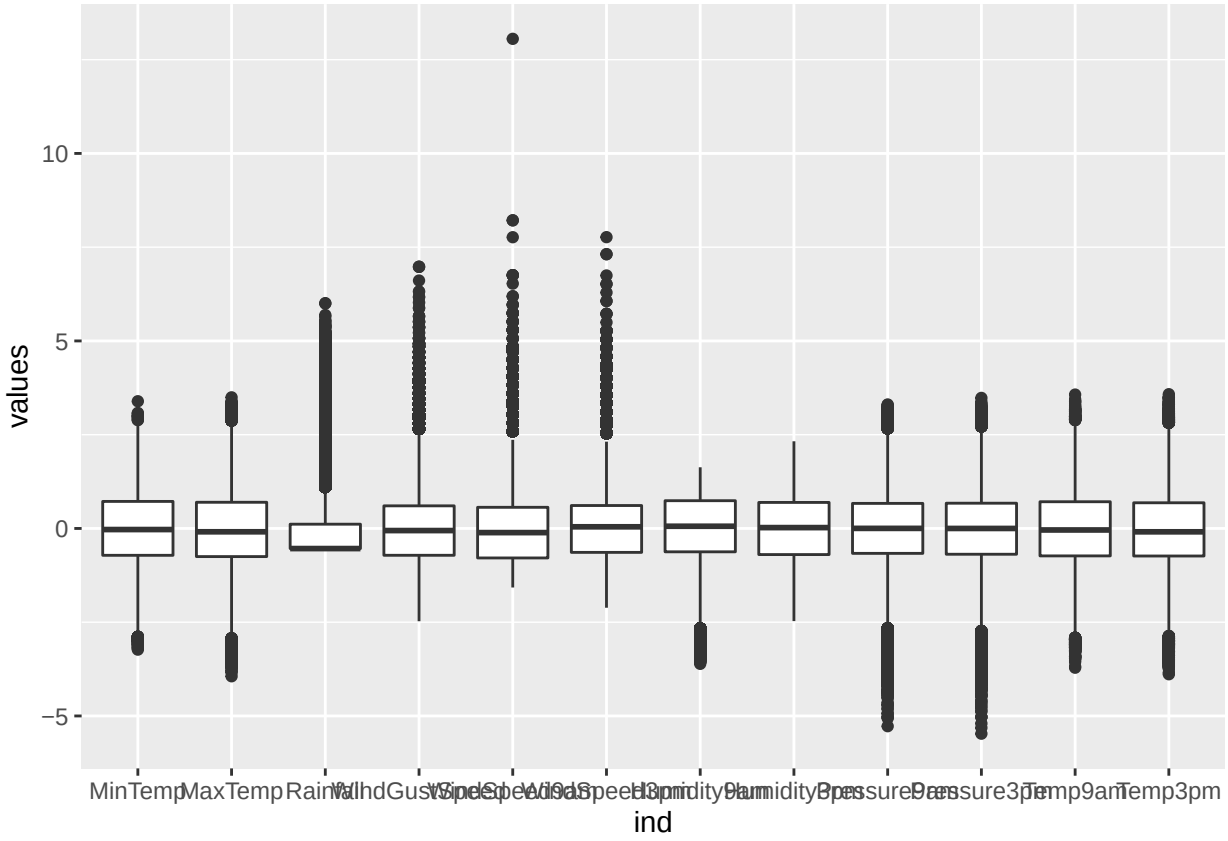
Our new data retains its original shape with the exception of Rainfall, which, even when transformed, is still far away from a Gaussian distribution, but it is however, closer to it.

```
##    Location            MinTemp              MaxTemp
##  Length:140787      Min.   :-3.23146151   Min.   :-3.94157623
##  Class :character   1st Qu.:-0.71645528   1st Qu.:-0.75017942
##  Mode  :character   Median :-0.02912439   Median :-0.08940563
##                     Mean   : 0.00000000   Mean   : 0.00000000
```

```
##                    3rd Qu.: 0.72069114   3rd Qu.: 0.69789931
##                    Max.   : 3.39190894   Max.   : 3.49564366
##     Rainfall          WindGustDir        WindGustSpeed
##  Min.   :-0.5365507   Length:140787      Min.   :-2.47395160
##  1st Qu.:-0.5365507   Class :character   1st Qu.:-0.71576271
##  Median :-0.5365507   Mode  :character   Median :-0.05644187
##  Mean   : 0.0000000                      Mean   : 0.00000000
##  3rd Qu.: 0.1133497                      3rd Qu.: 0.60287897
##  Max.   : 6.0078166                      Max.   : 6.97631373
##   WindDir9am         WindDir3pm         WindSpeed9am
##  Length:140787      Length:140787      Min.   :-1.5745568
##  Class :character   Class :character   1st Qu.:-0.7867888
##  Mode  :character   Mode  :character   Median :-0.1115591
##                                        Mean   : 0.0000000
##                                        3rd Qu.: 0.5636706
##                                        Max.   :13.0554206
##   WindSpeed3pm          Humidity9am           Humidity3pm
##  Min.   :-2.11463131   Min.   :-3.61015072   Min.   :-2.47077295
##  1st Qu.:-0.63801246   1st Qu.:-0.62289738   1st Qu.:-0.69655686
##  Median : 0.04350393   Median : 0.05840602   Median : 0.02271994
##  Mean   : 0.00000000   Mean   : 0.00000000   Mean   : 0.00000000
##  3rd Qu.: 0.61143426   3rd Qu.: 0.73970941   3rd Qu.: 0.69404495
##  Max.   : 7.76735638   Max.   : 1.63064462   Max.   : 2.32440568
##   Pressure9am           Pressure3pm              Temp9am
##  Min.   :-5.272747409   Min.   :-5.472188169   Min.   :-3.71298957
##  1st Qu.:-0.664723067   1st Qu.:-0.687736186   1st Qu.:-0.73232566
##  Median : 0.001668146   Median :-0.000150273   Median :-0.04093455
##  Mean   : 0.000000000   Mean   : 0.000000000   Mean   : 0.00000000
##  3rd Qu.: 0.668059359   3rd Qu.: 0.673110934   3rd Qu.: 0.71191355
##  Max.   : 3.305267136   Max.   : 3.480753415   Max.   : 3.56966349
##     Temp3pm             RainToday         RainTomorrow          Season
##  Min.   :-3.88777316   Length:140787      Length:140787      winter:33981
##  1st Qu.:-0.73461775   Class :character   Class :character   spring:37027
##  Median :-0.08965415   Mode  :character   Mode  :character   summer:35526
##  Mean   : 0.00000000                                         fall  :34253
##  3rd Qu.: 0.68430218
##  Max.   : 3.57947215
```




While there appear to be some outliers, all the outliers in the boxplot almost in its entirety are extremely close together, suggesting highly skewed distributions, not outliers.



Train and test sets are separated for further use in the classification section. The validation method to follow is as follows:

The dataset is divided into train and test, with the train and test data being imputed with the imputation model built from train data. The train data is to be divided into 10 folds for cross-validation, using cross-validation results to do model selection, at which point model validation over the model selected is performed with the test set.

**Deal with computational complexity**

Then, as a checkpoint for the pre-processing phase, the datasets have been saved on memory in order to avoid executing all the tmie this costly phase.

To make feasible in our computers the analysis below the dataset will be sampled all the time keeping `sampling_size` number of records (after reading it). To guarantee reproducibility of the sampling we used a seed `set.seed(1)`. The function `load_df` will be help in this operation (in particular in the **clustering** section).

# Visulization and Interpretation of the latent concepts

In this section, we apply LDA, PCA and MCA to visualize and interpret the latent concepts.

## Linear Discriminant Analysis

As we can observe the numerical variables (except location, wind direction, season) are nearly normally distributed, we use the training data to apply the Linear Discriminant Analysis and the response class is RainTomorrow.

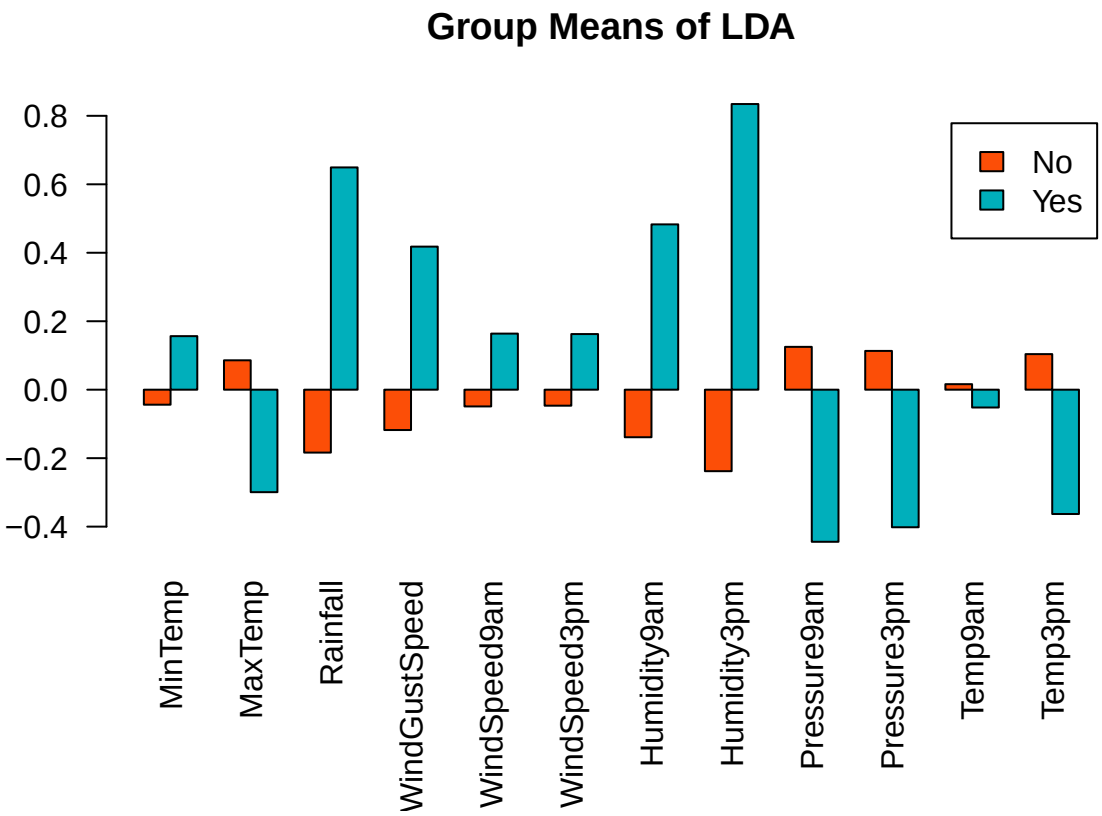
```
## Call:
## lda(RainTomorrow ~ ., data = preprocessedTrain[, -c(1, 5, 7,
##     8, 17, 19)])
##
## Prior probabilities of groups:
##          No         Yes
## 0.778380538 0.221619462
##
## Group means:
##            MinTemp        MaxTemp      Rainfall WindGustSpeed   WindSpeed9am
## No  -0.04383556041  0.08583022013 -0.1837875743 -0.1178372931 -0.04893383371
## Yes  0.15645390443 -0.29937449027  0.6492677534  0.4179249307  0.16380602544
##       WindSpeed3pm   Humidity9am   Humidity3pm   Pressure9am   Pressure3pm
## No  -0.04672073785 -0.1388706018 -0.2380528034  0.1250684274  0.1132302721
## Yes  0.16252334207  0.4828787351  0.8344539076 -0.4443204129 -0.4018589805
##            Temp9am       Temp3pm
## No   0.01610887066  0.1039347420
## Yes -0.05204794413 -0.3631222306
##
## Coefficients of linear discriminants:
##                             LD1
## MinTemp        0.003613389873
## MaxTemp        0.338102046070
## Rainfall       0.343662534783
## WindGustSpeed  0.567172322968
## WindSpeed9am  -0.039287968173
## WindSpeed3pm  -0.268477089568
```

```
## Humidity9am   -0.084122044809
## Humidity3pm    0.974400770234
## Pressure9am    0.758677510733
## Pressure3pm   -1.050292733265
## Temp9am       -0.142056055271
## Temp3pm       -0.252496604535
```

Prior probabilities of groups defines the prior probability of the response classes for an observation. This shows 77.84 % of rain tomorrow and 22.16 % of not rain tomorrow.

Group Means defines the mean value for response classes. This indicates means values of different features when they fall to a particular response class.

To be more specific, from the below diagram we see a clear difference between all the variables: all of them have opposite mean values for response class RainTomorrow. Especially for Humidity3pm, Humidity9am, Rainfall, Pressure9am, WindGustSpeed, their absolute values vary greatly. The more the difference between mean, the easier it will be to classify observation.




We can assume humidity, rainfall, pressure have more impact on the probabilities of rain on the next day; while temperature on 9am and minimum temperature have less impact. The more the humidity on 3pm and 9am, rainfall, speed of strongest wind, and less pressure on 3pm and 9am, the more likely it is to rain the next day.

**predictions**

Next step, we find the model accuracy of 0.8423 for the training data, which is excellent.

```
## [1] 0.8423421824
```

```
##           RainTomorrow
## Predicted    No   Yes
```

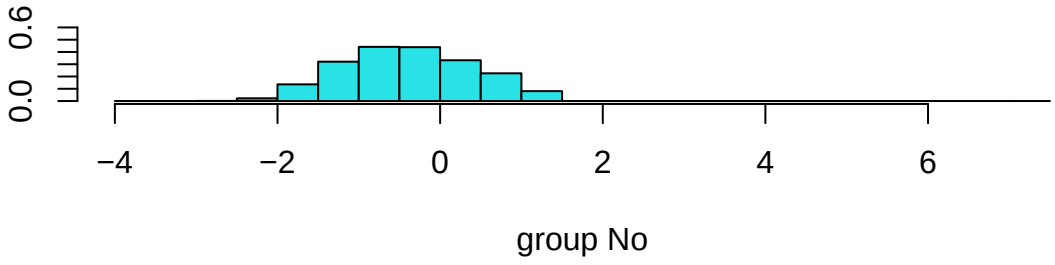
Table 1: Posterior probabilities. RainAustralia data set

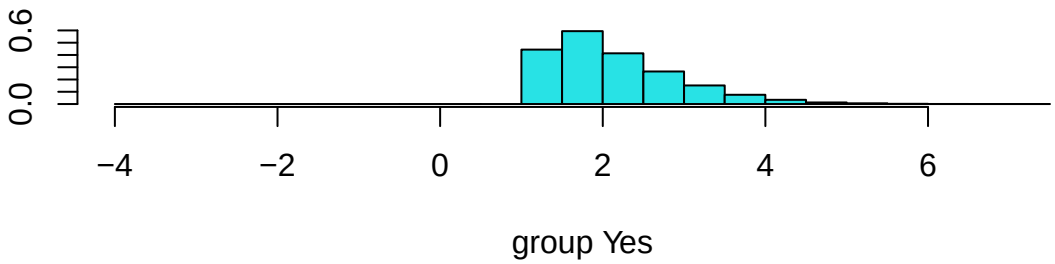
| | No | Yes |
|---|---|---|
| 1 | 0.9829115998 | 0.0170884002 |
| 6 | 0.9286372281 | 0.0713627719 |
| 7 | 0.9818072021 | 0.0181927979 |
| 8 | 0.9853806670 | 0.0146193330 |
| 9 | 0.8294077184 | 0.1705922816 |
| 12 | 0.0680567032 | 0.9319432968 |

```
##       No  82572 12660
##       Yes  5097 12301
```

We check the posterior probabilities of a piece of data, we can find that the classifier basically meets our expectations.

The below stacked histogram shows how the response class has been classified by the LDA classifier. The X-axis shows the value of line defined by the co-efficient of linear discriminant for LDA model. The two Yes/No groups are the groups for response class RainTomorrow.



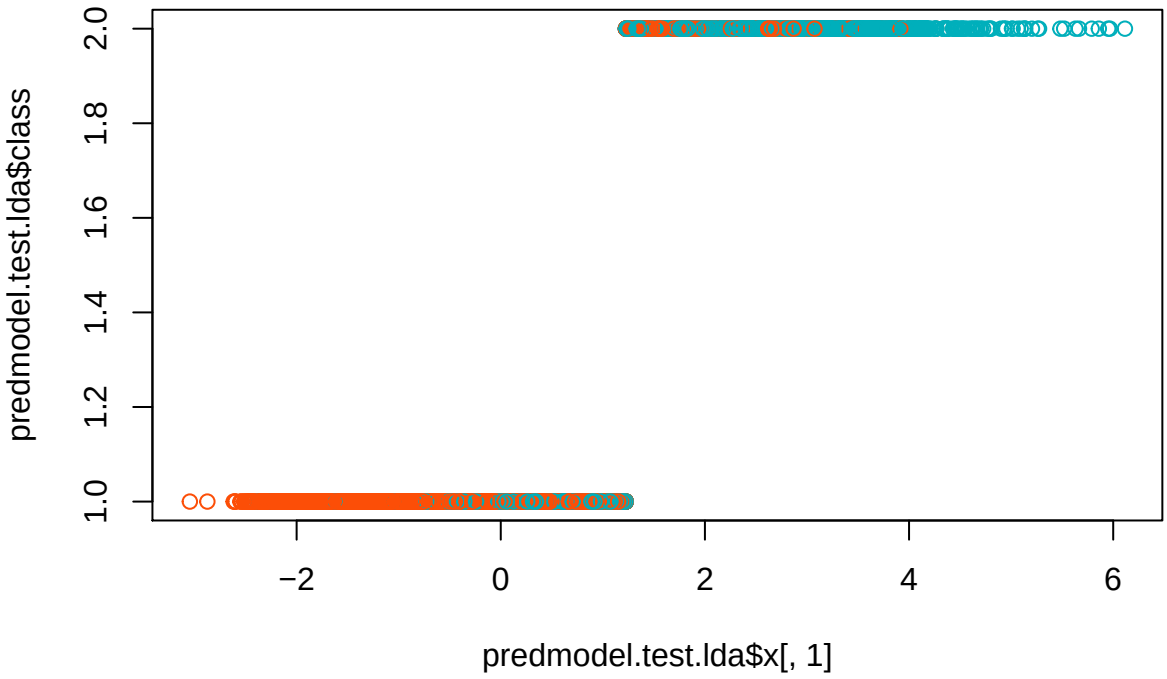




We check the model accuracy of 0.8389 for the test data, which is also excellent.

```
## [1] 0.8389388074
```

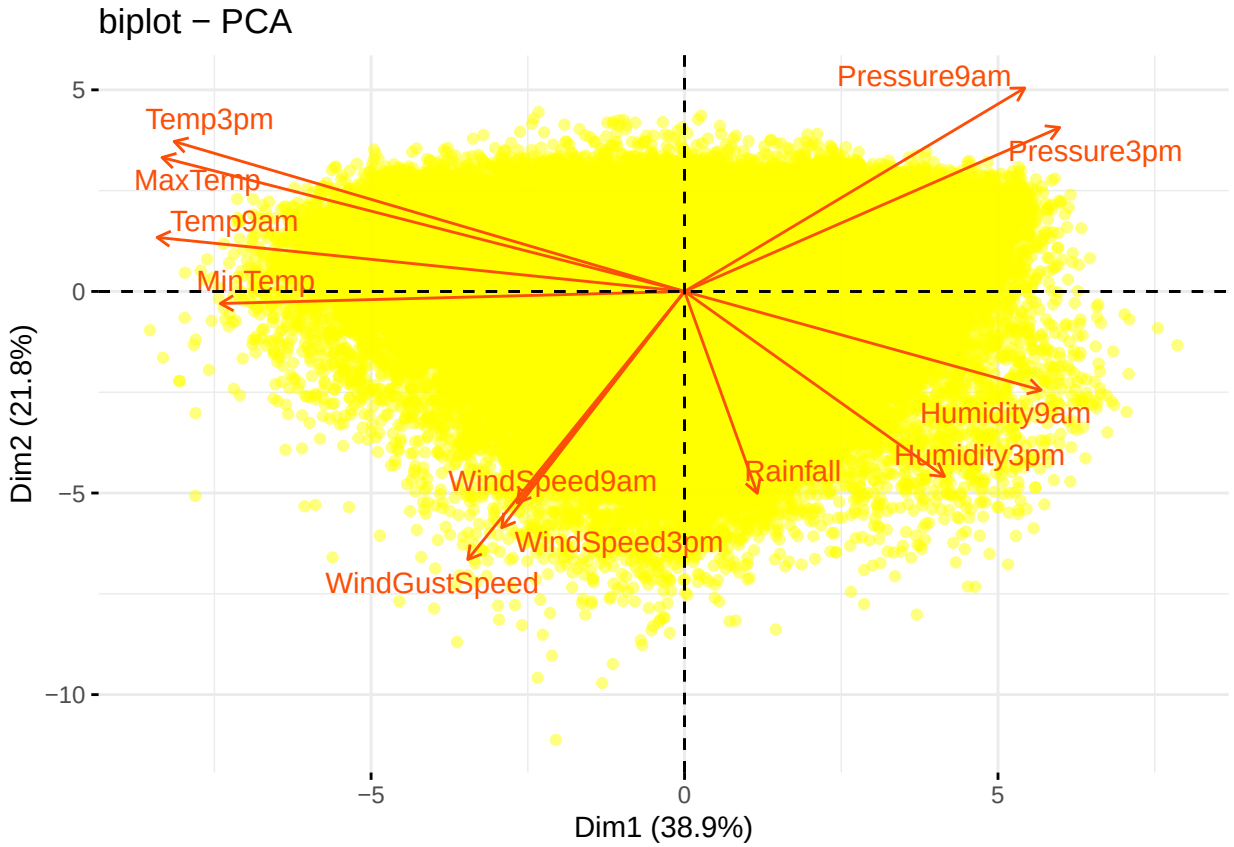
```
##          RainTomorrow
## Predicted    No   Yes
##       No  20609  3227
##       Yes  1308  3013
```

The below figure shows how the data has been classified. The Predicted Group-No and Group-Yes has been colored with actual classification with red and blue color. The mix of color in the Group shows the incorrect classification prediction.




### Principal Components Analysis

We apply PCA only on the numerical variables (already standardized in preprocessing phase).




As we can see, the first two dimensions explain more than 60% variance. The variables are approximately divided into 4 groups: temperature, pressure, wind speed and humidity (where rainfall is included in it).

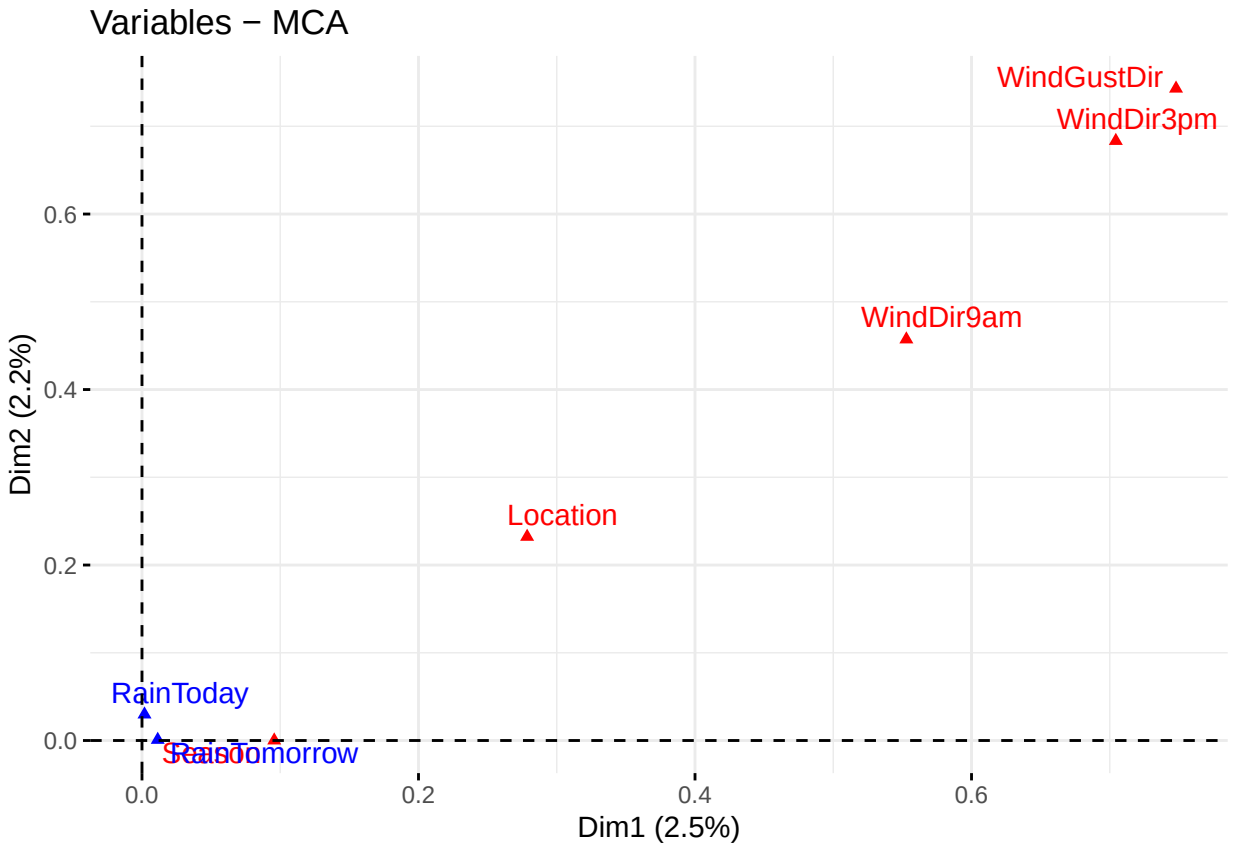
Each of these four groups of variables occupies a quadrant.

In the first dimension, temperature and wind speed have negative projection while pressure and humidity have positive projection. This means that there is a negative correlation between temperature, wind speed and pressure, humidity. The first principal component tells us about whether this observed day is with high pressure, wet, low temperature, low wind weather, or a low pressure, low humid and high temperature, windy day.

In the second dimension, MinTemp and Temp9am have little projection onto it. However, we can still observe that temperature, pressure (positive projection) are negatively correlated with wind speed and humidity (negative projection). Thus, this axis separates wet, windy day from high pressure, high temperature day.

### Multiple Correspondence Analysis

After analyzing the numerical variables, we use the categorical variables to apply MCA, using RainToday and RainTomorrow as supplementary variables.



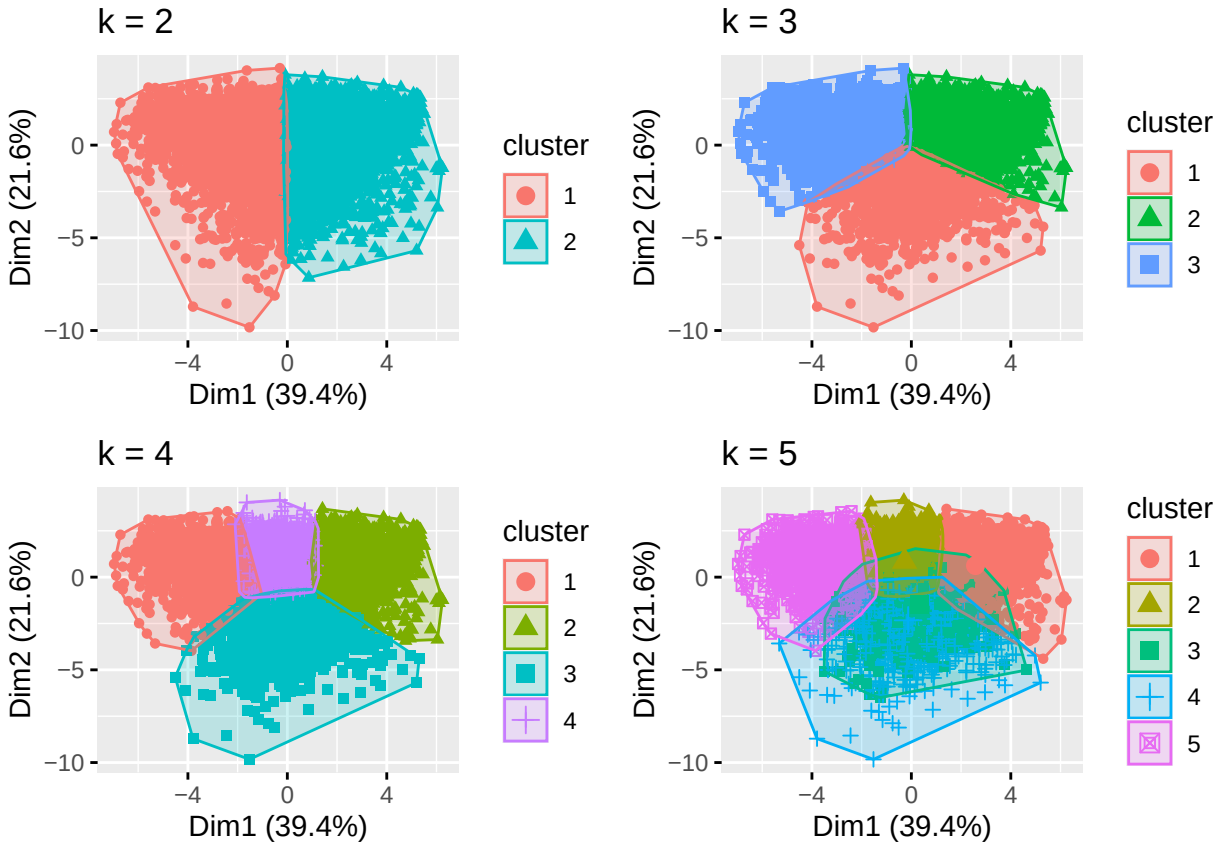

We have very low explanation of variance (only 2.5% and 2.2% of the first and second dimension respectively). The reason may be the values of the categorical variables vary enormously. And we observe that supplementary variable RainTomorrow has slightly correlation with the first dimension, and RainToday has correlation with the second dimension.

## Clustering

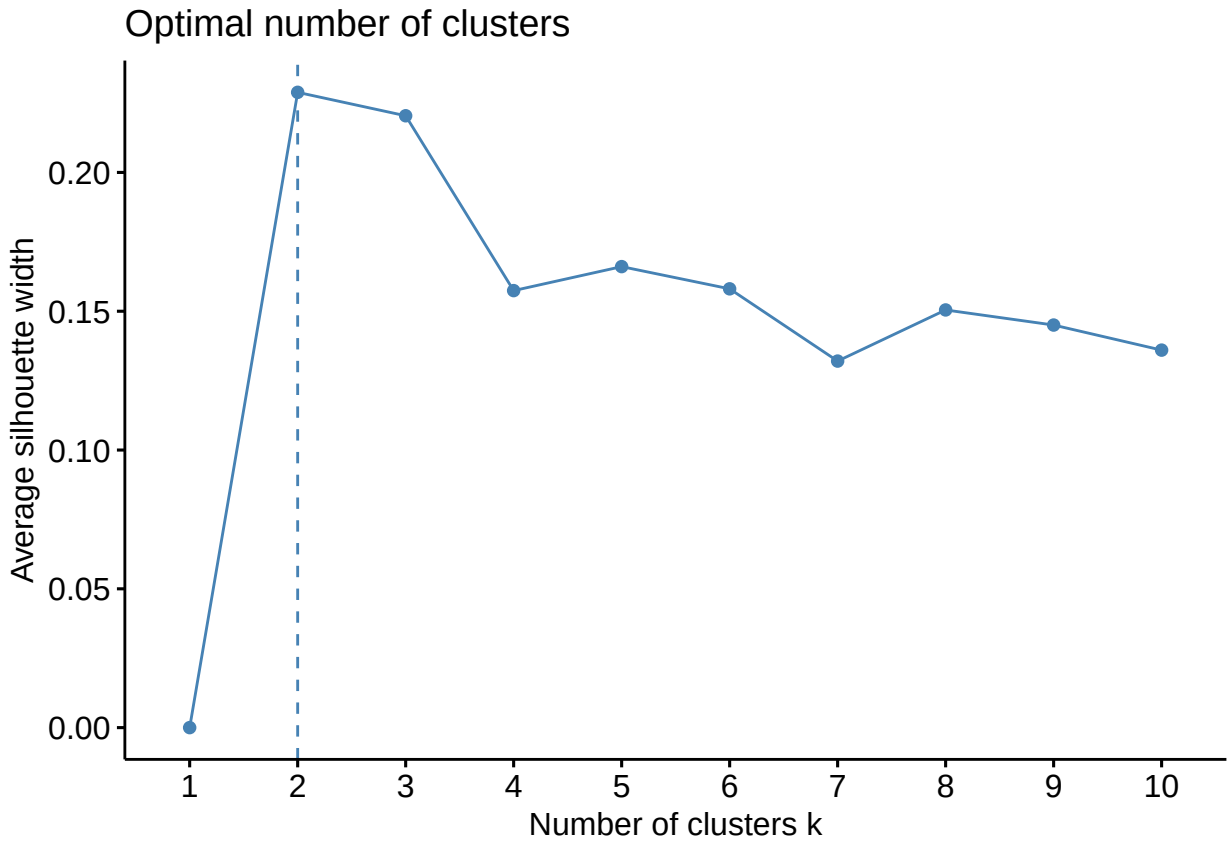
In the following chunk of code a tiny data pre processing will be applied to the dataset in order to prepare it to execute few clustering algorithms on top of it. To apply the clustering algorithms below the input dataset must be composed by **numeric variables**, therefore not numeric data will be discarded. The analysis will be performed considering just climatic descriptors.

### Partitioning method

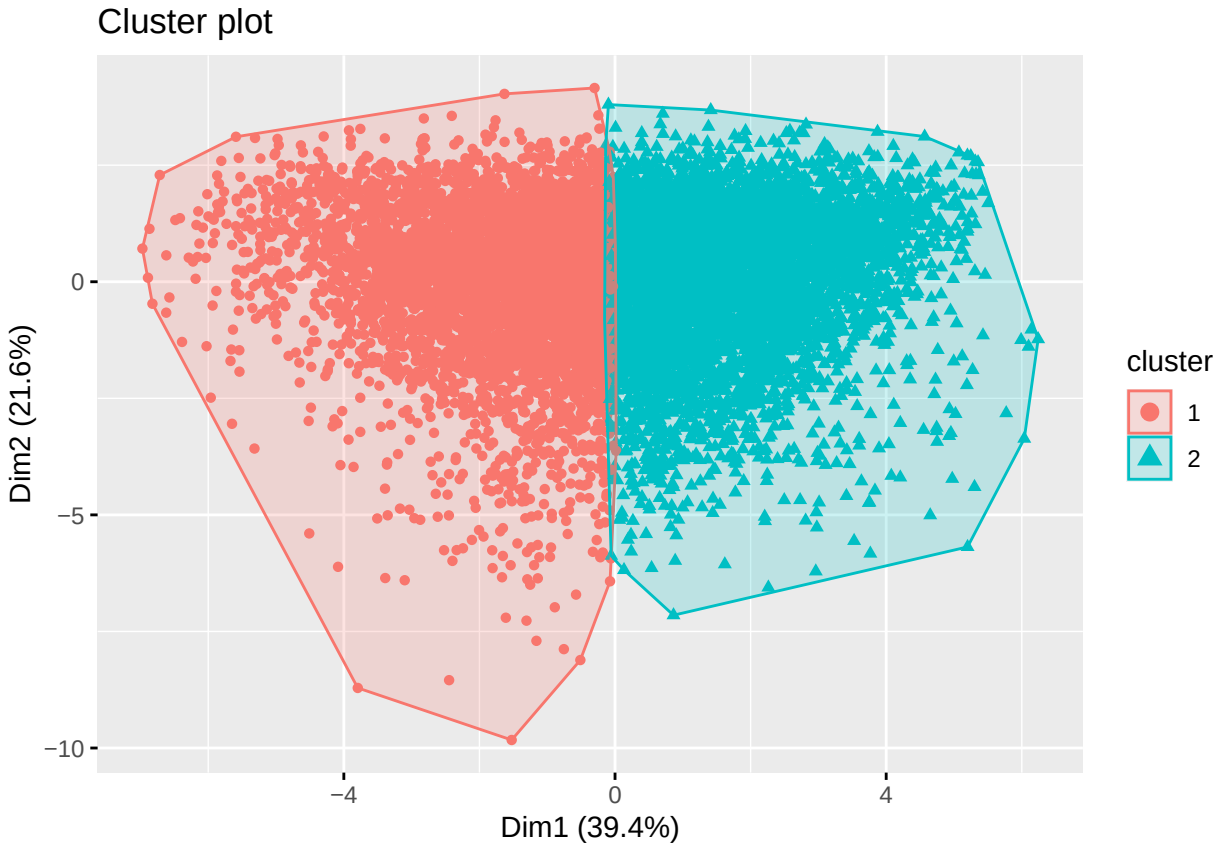
The first approach with clustering method have been with the traditional partition methodology applying K-Means algorithm, since is the computationally less expensive technique studied. The algorithm have been executed, looking for 2, 3, 4 and 5 clusters (`centers = x`) in order to look for some likely shapes of the clusters. It is plain that datas have the hape of a cloud, therefore it is not going to be possible distinguish clean clusters.




To determine the optimal number of clusters we adopted the **silhouette** method, with the respective code `method = "silhouette"`. The output suggest an optimal number of clusters equal to two.



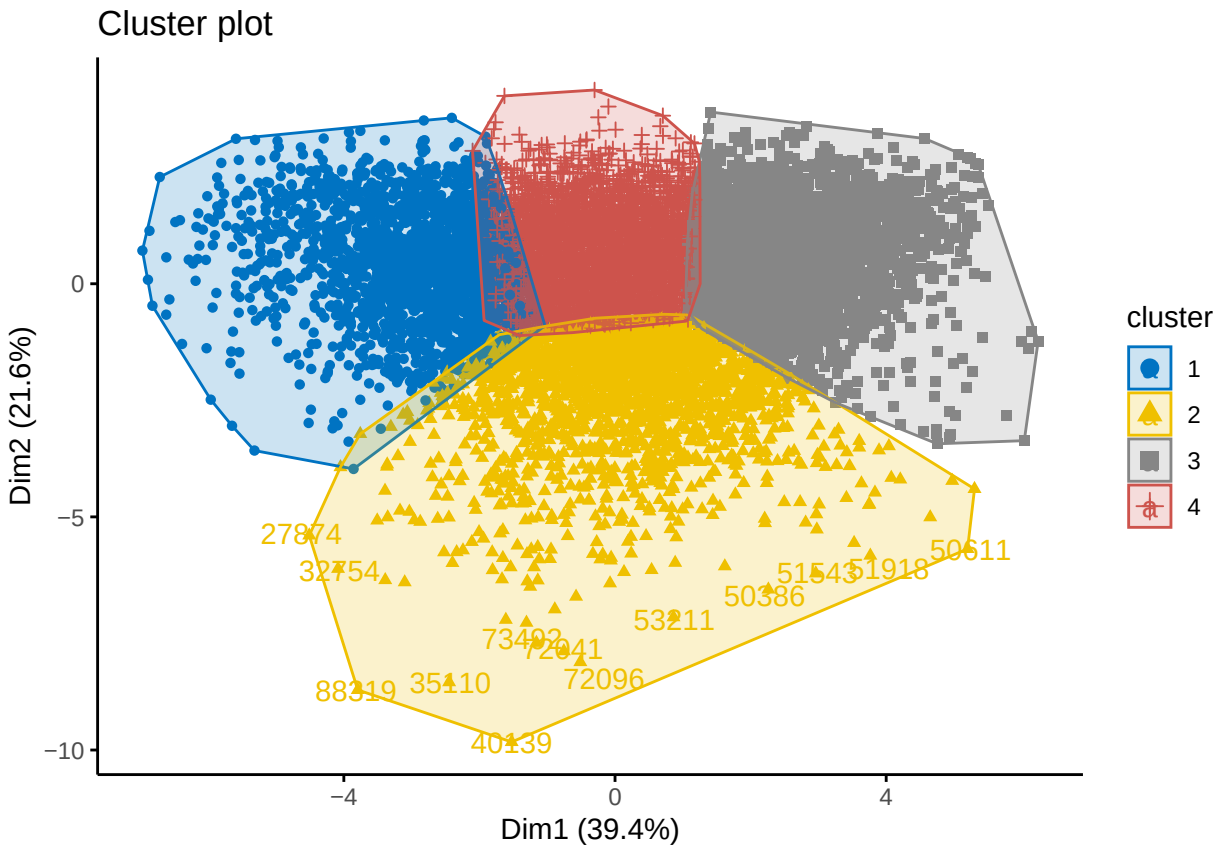

The object of our analysis then will be based on this plot.




As the silhouette method suggested will be studied the clustering with k equals to 2. For the interpretation of the obtained results, showing the centers `k2$centers` will help to associate each cluster to particular feature. It is clear that the first cluster (1) is more representative for the high **temperature** sampling while the second cluster (2) is more representative for the low temperatures. High temperature cluster and low temperature cluster differ also in term of **humidity** and **pressure**, presenting lower values.

```
##         MinTemp       MaxTemp      Rainfall WindGustSpeed  WindSpeed9am
## 1  0.6802116553  0.7475873448 -0.1197340529  0.2929090744  0.2266163697
## 2 -0.6501723808 -0.7038839606  0.1087038555 -0.3080095658 -0.2368670556
##    WindSpeed3pm   Humidity9am   Humidity3pm   Pressure9am   Pressure3pm
## 1  0.2958781166 -0.4742667678 -0.3288269687 -0.4932828429 -0.5416438026
## 2 -0.2939882830  0.4420399216  0.2978180954  0.4976378378  0.5464055603
##         Temp9am       Temp3pm
## 1  0.7656852292  0.7274544433
## 2 -0.7126177982 -0.6814944811
```

Another trial to identify other kind of clusters shapes have been done applying a mixed approach, using a hierarchical clustering to determine the shape of clusters. The number of clusters will be specified by the parameter `k=4`. This time we will observe the characteristic of four different clusters.

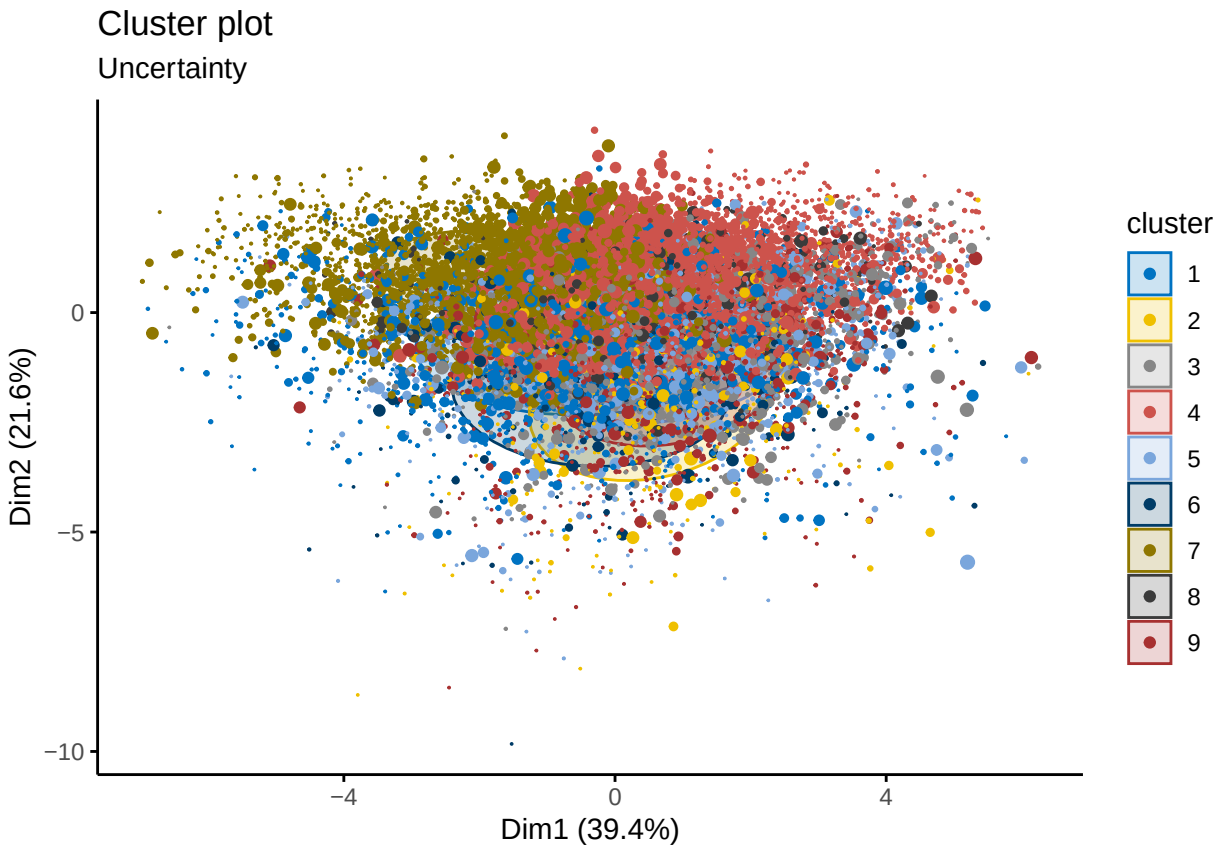



Adopting a higher number of cluster is easier to notice a higher variation in term of clusters specialization. The most important cluster in this analysis is clearly the number (2) since it is represented by a high value of the `Rainfall` attribute and therefore it is representing the rainy days, that are very important for our analysis, since the goal of the following prediction phase will be focused on classify correctly the variable `Raintomorrow`. According with this cluster, rainy days are characterized by high wind values, low pressure and temperatures and high humidity.

```
##          MinTemp       MaxTemp       Rainfall WindGustSpeed  WindSpeed9am
## 1  1.13214469075  1.3572778823 -0.25253880808  0.2524517185  0.1664201068
## 2 -0.05342667522 -0.5937801789  0.93114131530  1.0605498863  0.8626590828
## 3 -0.94665282984 -0.9014480978 -0.06194323488 -0.6936974024 -0.5509865022
## 4  0.07091398596  0.2214141667 -0.34084897070 -0.2622680903 -0.1946454890
##    WindSpeed3pm   Humidity9am   Humidity3pm   Pressure9am   Pressure3pm
## 1  0.2164179499 -0.7960227954 -0.6272703960 -0.7360802876 -0.8515996741
## 2  0.8972333623  0.3706700722  0.6383509436 -0.7197830497 -0.5386249443
## 3 -0.6107832089  0.6535193932  0.3528005675  0.8996042972  0.9209142881
## 4 -0.1852179020 -0.2561717076 -0.2838592503  0.2050477106  0.1605795972
##         Temp9am       Temp3pm
## 1  1.3196986513  1.3346694076
## 2 -0.3048875098 -0.6425014240
## 3 -1.0004426703 -0.8601902729
## 4  0.1713493220  0.2359141485
```

## Model Based Clustering

Since the biggest part of the dataset shows a gaussian distribution, a Gaussian finite mixture model fitted by EM algorithm should achieve good results in terms of clustering.



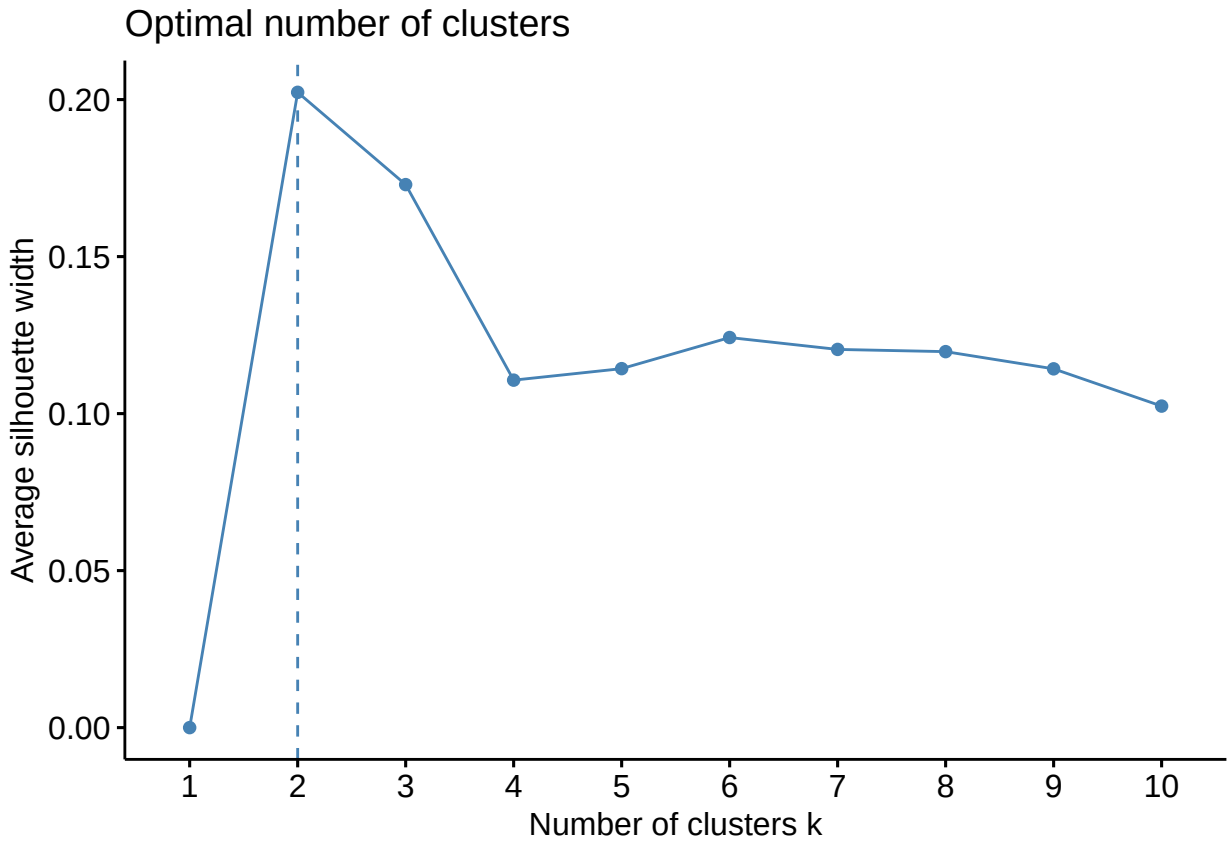

Gaussian mixture produced as output 9 clusters of shape **VEV**. A such high number of cluster (9) suggest that, as we hypotized before, the shape of the data is a cloud point and that's why MBC is actually failing in findig clusters.

```
## [1] 9
```

```
## [1] "VEV"
```

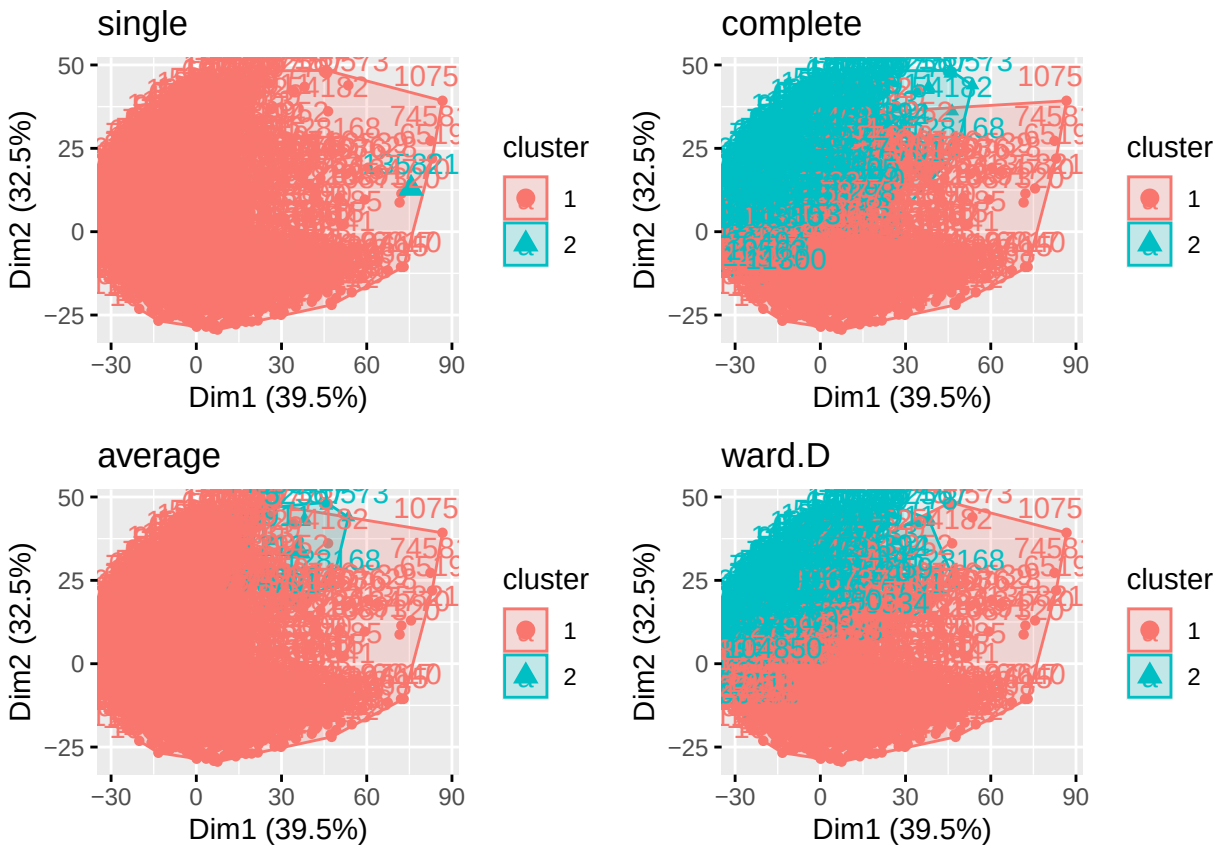
### Hierarchical Clustering

Since the dataset doesn't shows explicit cluster so far, we decided to exploit the cloud shape of the dataset applying the hierarchical clustering. The metric chosen to compute distances is the `euclidean`. As before the `silhouette` method helped us to cut the tree to have the optimal number of clusters.




Then the hierarchical clustering algorithm have been executed considering the number of clusters equal to two (`k = 2`) allying all the known metrics to link clusters.

The plotted graphs actually don't show kind of new information we didn't observed in the previous analysis: As for the single and average method we can observe a clear connection (clustering) between the points on the extreme left, while the ward and complete linking method suggest a more clear separation between the top and the bottom. The analysis by mean of the ward method actually reminds the group observed with kmeans algorithm.


single
complete
average
ward.D
Dim1 (39.5%)
Dim2 (32.5%)
cluster
1
2