RainInAustralia

Filippo, Antoni, Cristina, Mengxue

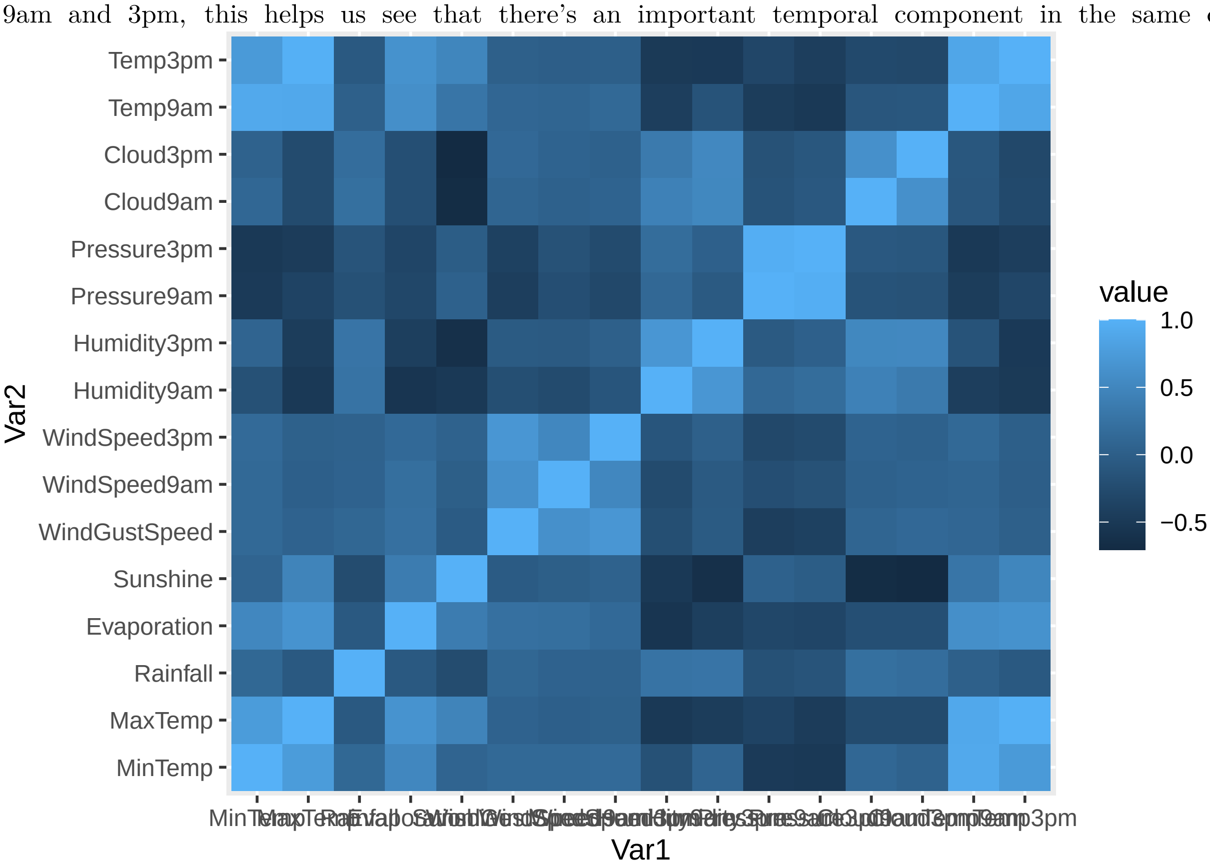
6/11/2021

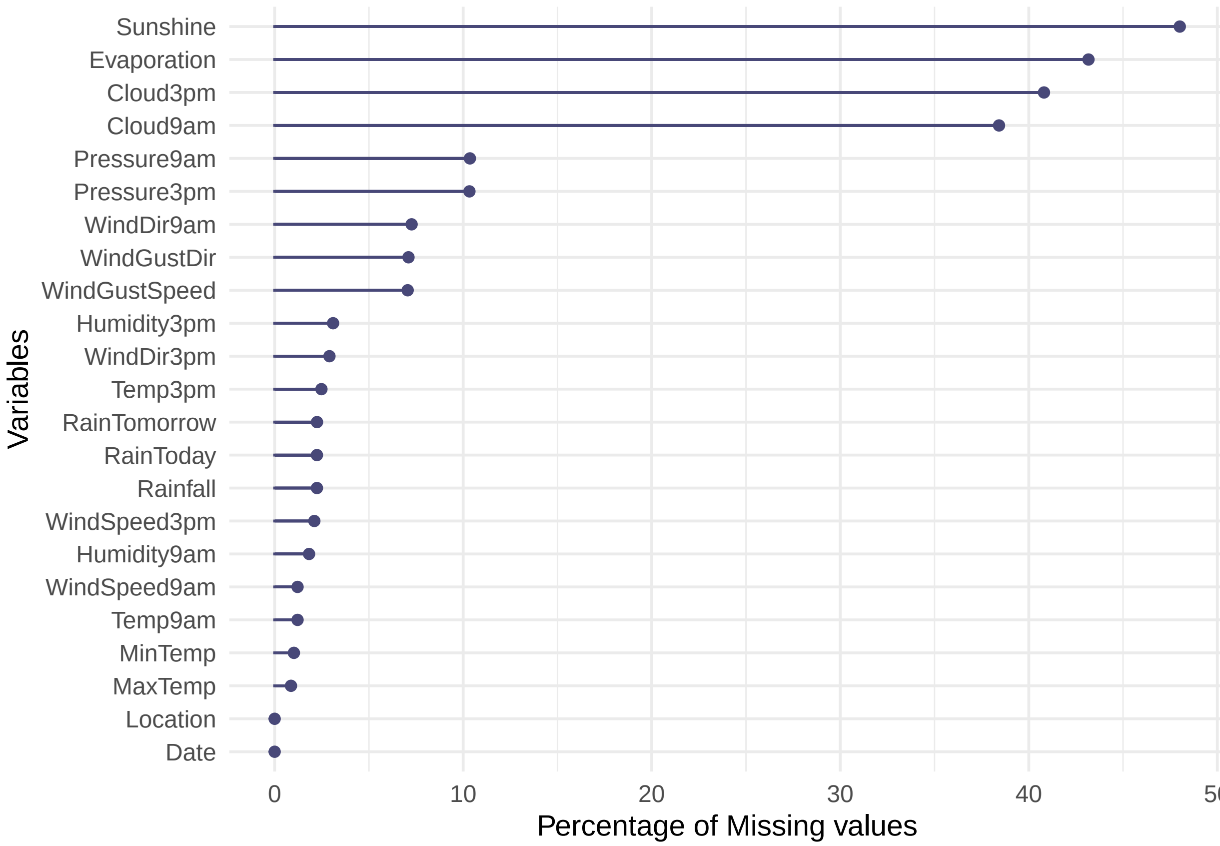
Dataset description

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

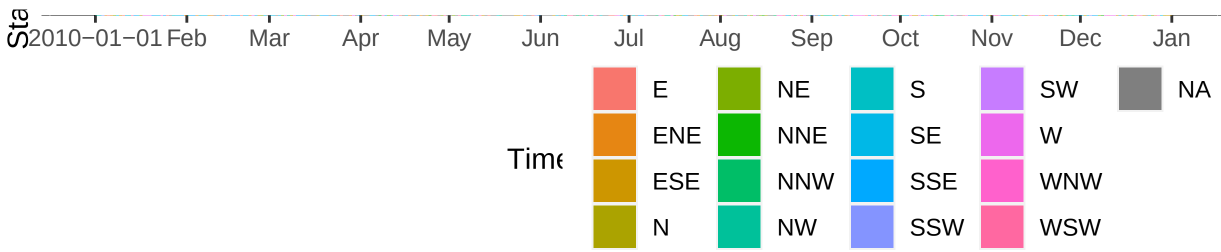
Date- Categorical variable, when the measurements were taken Location - Categorical variable, where the measurements were taken MinTemp - Numerical variable, minimal temperature observed that day MaxTemp - Numerical variable, maximal temperature observed that day Rainfall - Evaporation - Sunshine - WindGustDir - WindGustSpeed - WindDir9am - WindDir3pm - WindSpeed9am - Windspeed3pm - Humidity9am - Humidity3pm - Pressure9am - Pressure3pm - Cloud9am - Cloud3pm - 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

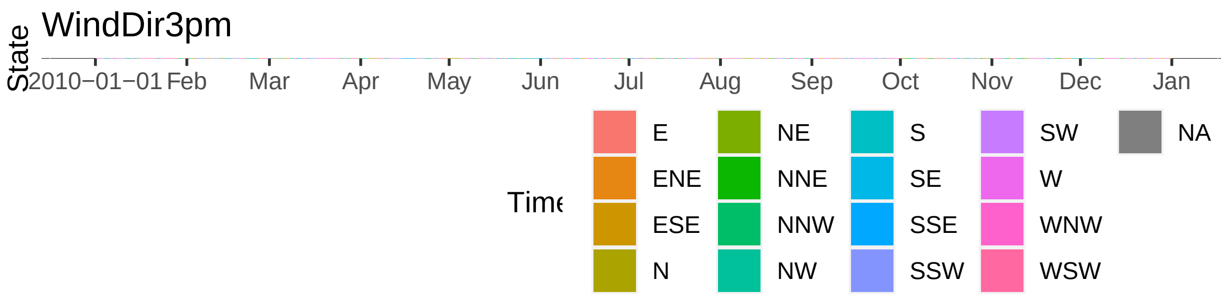
We perform a basic visualization, first of the correlation between variables, which isn't significant with the exception of variables recorded in the same day, that is, those measurements taken at 9am and 3pm, this helps us see that there's an important temporal component in the same day.

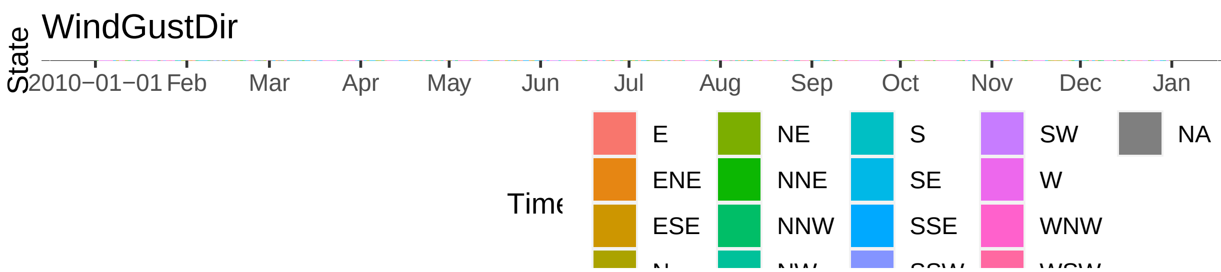




We perform a special method of data imputation, following the timeseries plot of the WinDir9am, WindDir3pm and WindGustDir, we can see that if the last day was a certain category, it will probably be that same category. So we choose this as our method of imputation for categorical NAs.

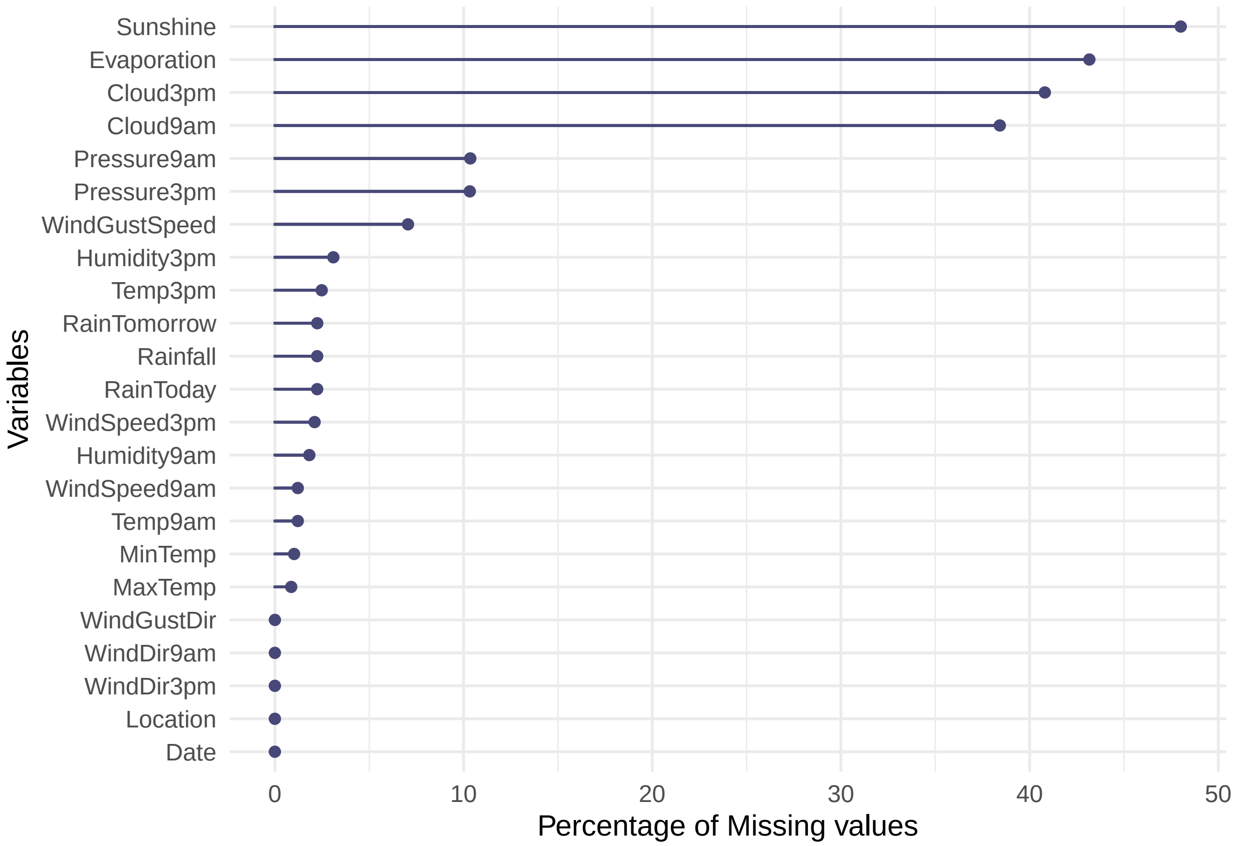






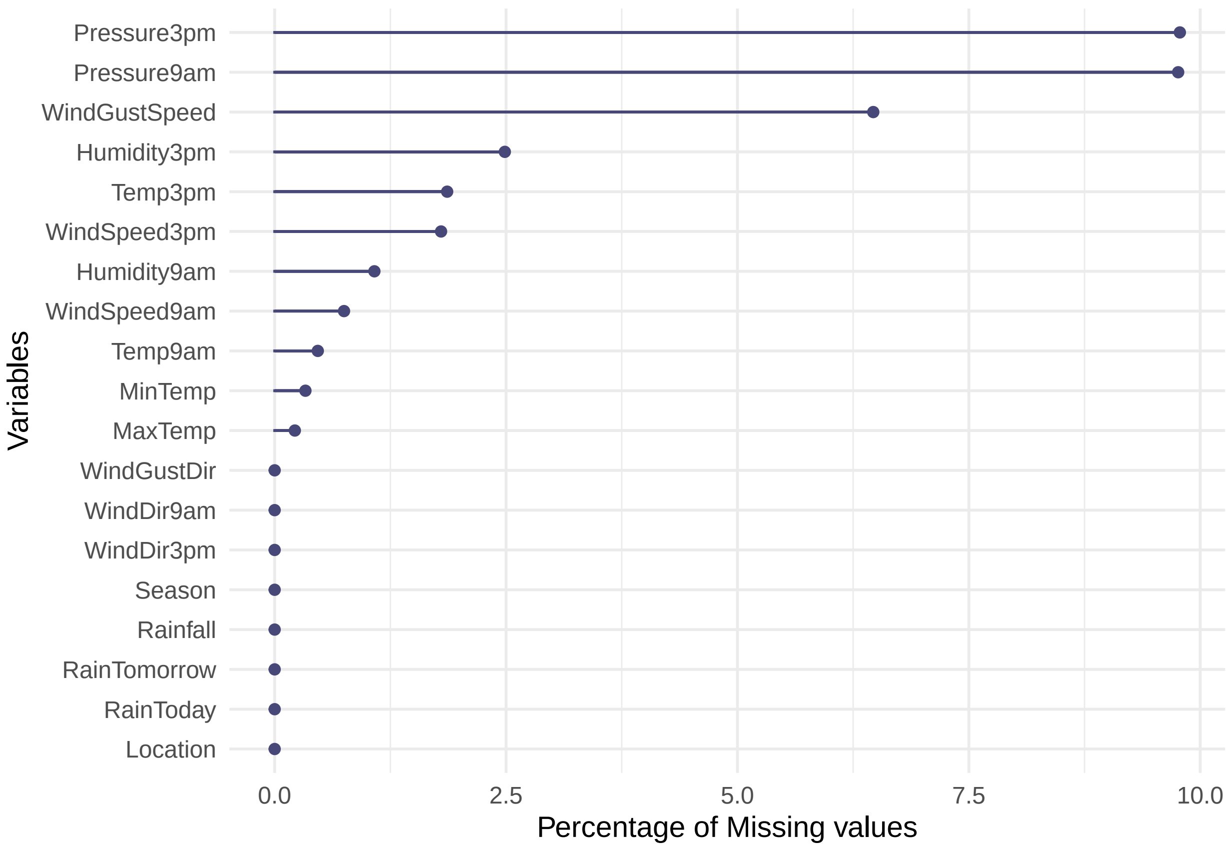

```
## INFO [2021-06-02 16:19:57] Date has been selected as the timestamp column
## INFO [2021-06-02 16:19:57]  has been selected as the numeric column(s)
## INFO [2021-06-02 16:19:57] WindDir9am, WindDir3pm, WindGustDir has been selected as the state column
## INFO [2021-06-02 16:19:57] creating state 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.

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.

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.

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.

To make fiesable in my computer the analysis the dataset have been sampled

Visualization

LDA

first we use numerical variables (except location, wind direction, season) to apply lda.

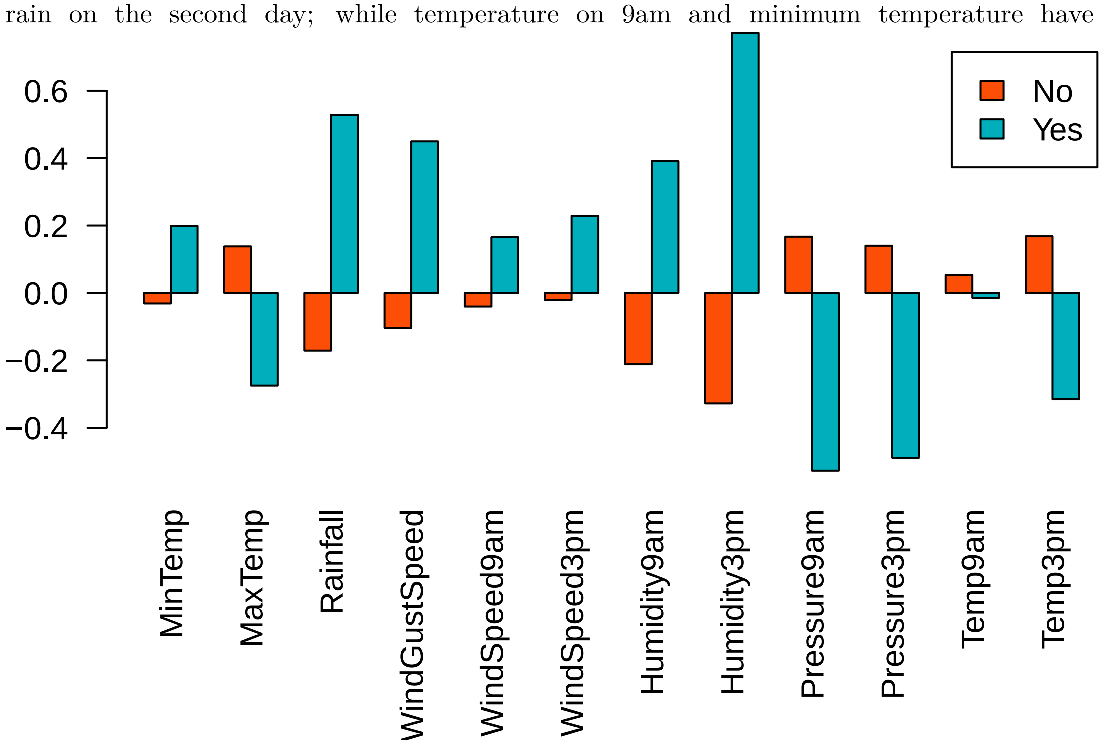
```
## Call:
## lda(RainTomorrow ~ ., data = scaled[, -c(1, 5, 7, 8, 17, 19)])
##
## Prior probabilities of groups:
##    No   Yes
## 0.779 0.221
##
## Group means:
##            MinTemp       MaxTemp      Rainfall WindGustSpeed   WindSpeed9am
## No  -0.03129009203  0.1381194356 -0.1711179629 -0.1037444468 -0.04033780379
## Yes  0.19883140776 -0.2748446964  0.5284213679  0.4497335562  0.16545824105
##        WindSpeed3pm   Humidity9am   Humidity3pm   Pressure9am   Pressure3pm
## No  -0.02108994269 -0.2116395863 -0.3277774578  0.1669145290  0.1402097850
## Yes  0.22904496911  0.3911135215  0.7715055231 -0.5273003287 -0.4889403722
##            Temp9am       Temp3pm
## No   0.05393326701  0.1681473091
## Yes -0.01444686437 -0.3154724759
##
## Coefficients of linear discriminants:
##                         LD1
## MinTemp        0.2497060858
## MaxTemp        0.5175672616
## Rainfall       0.2608917948
## WindGustSpeed  0.5934534923
## WindSpeed9am  -0.1192468627
## WindSpeed3pm  -0.2262787930
## Humidity9am   -0.2078587237
## Humidity3pm    1.0095116531
## Pressure9am    0.8742397009
```

```
## Pressure3pm   -1.2728798720
## Temp9am       -0.4000923765
## Temp3pm       -0.4751598989
```

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 (µk) for response classes for a particular X=x. This indicates means values of different features when they fall to a particular response class.

We see a clear difference between all the variables: they have opposite mean values for class RainTomorrow class. Especially for Humidity3pm, Humidity9am, Rainfall,Pressure9am, 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 second day; while temperature on 9am and minimum temperature have less impact.

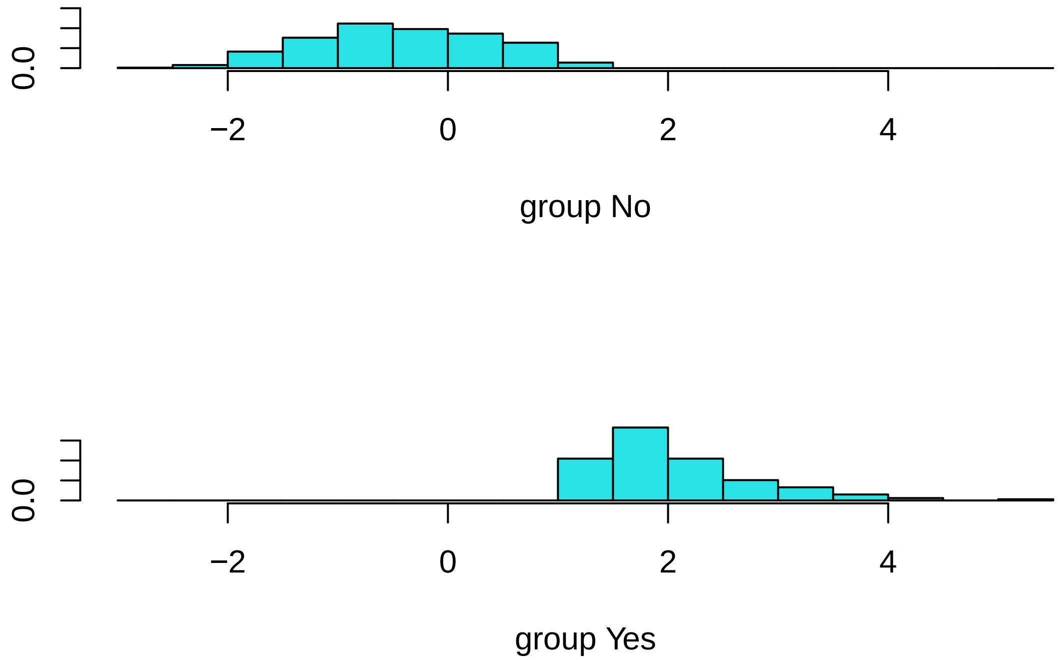


predictions

```
## [1] 0.854
```

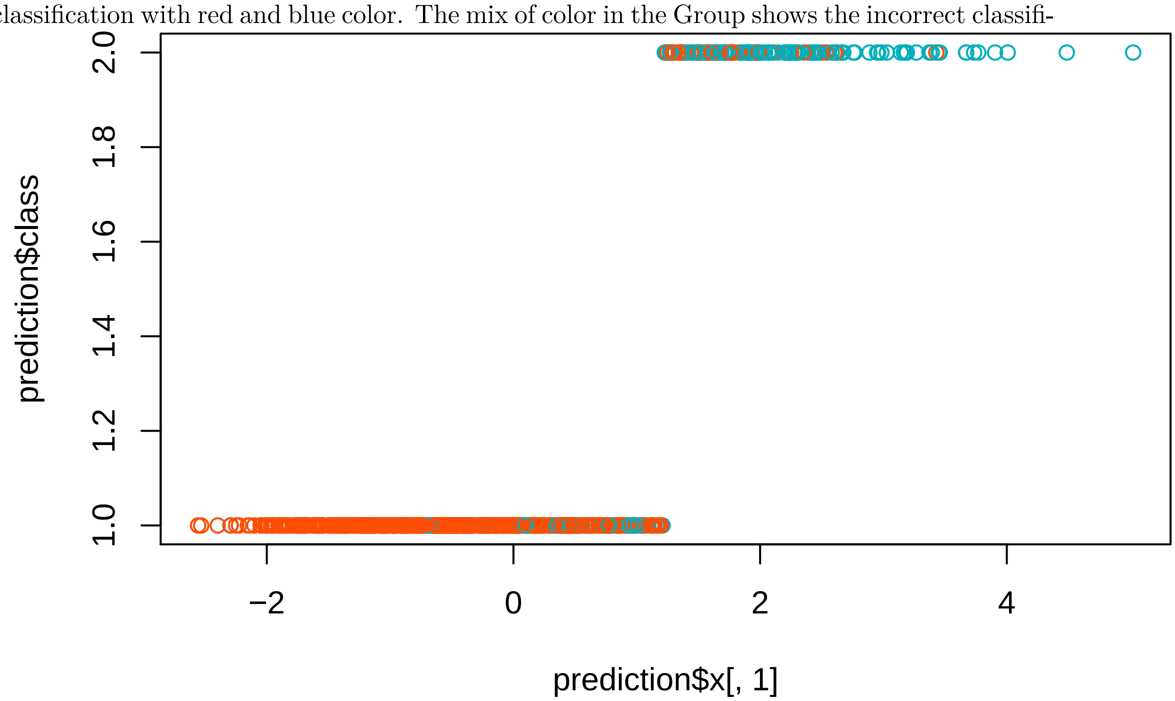
```
##          RainTomorrow
## Predicted  No Yes
##       No  733 100
##       Yes  46 121
```

The below plot 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 groups are the groups for



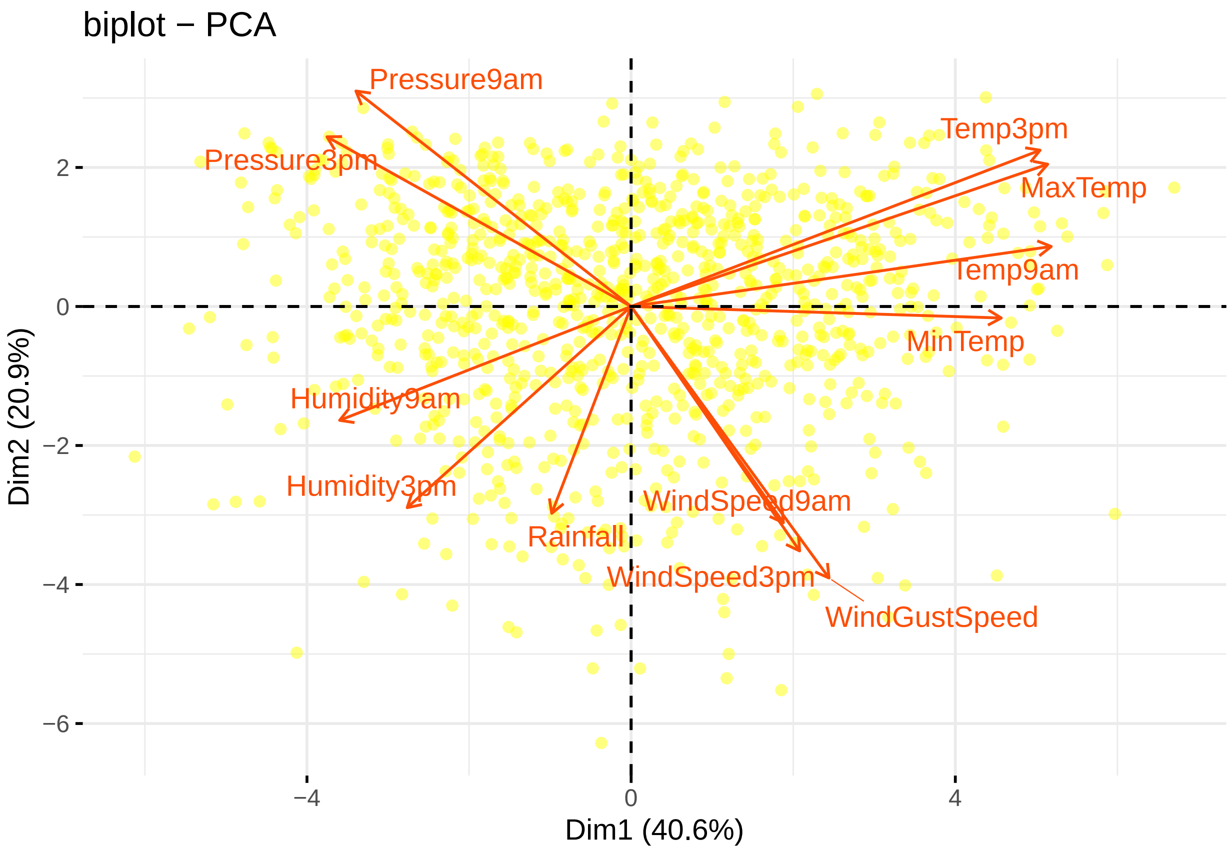
response classes.

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.

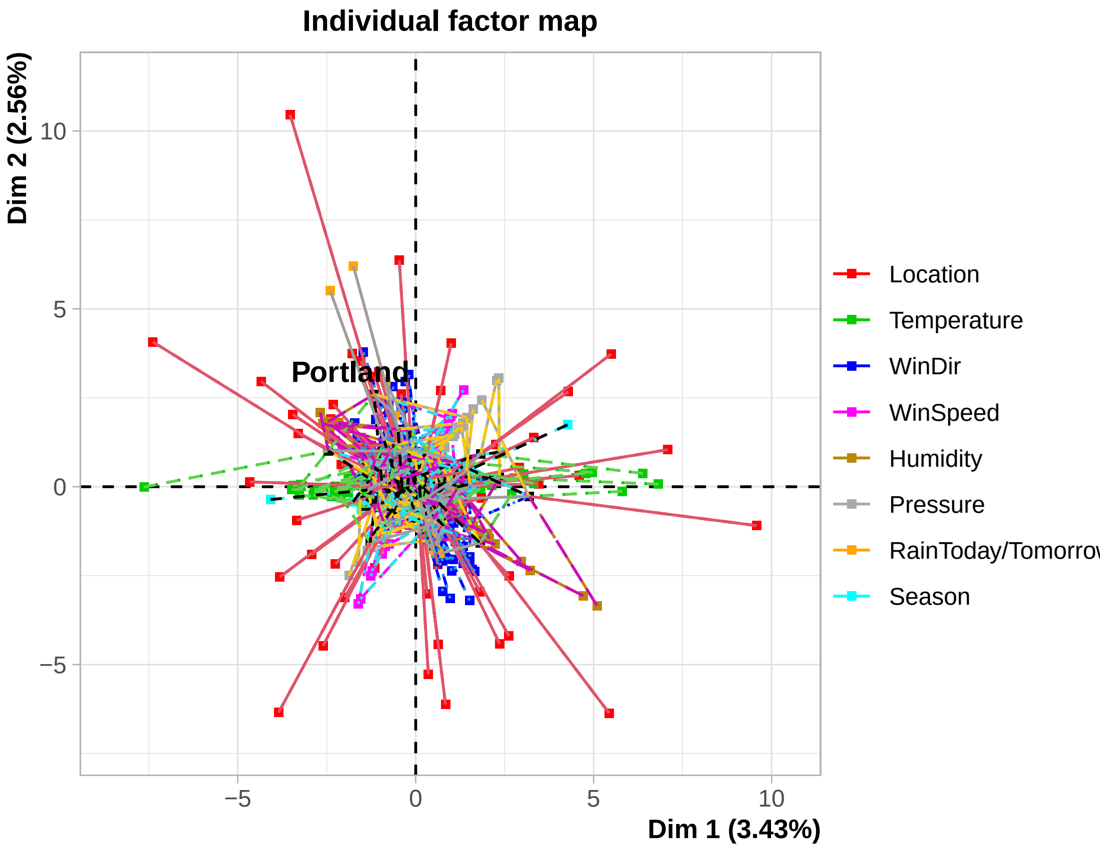


PCA

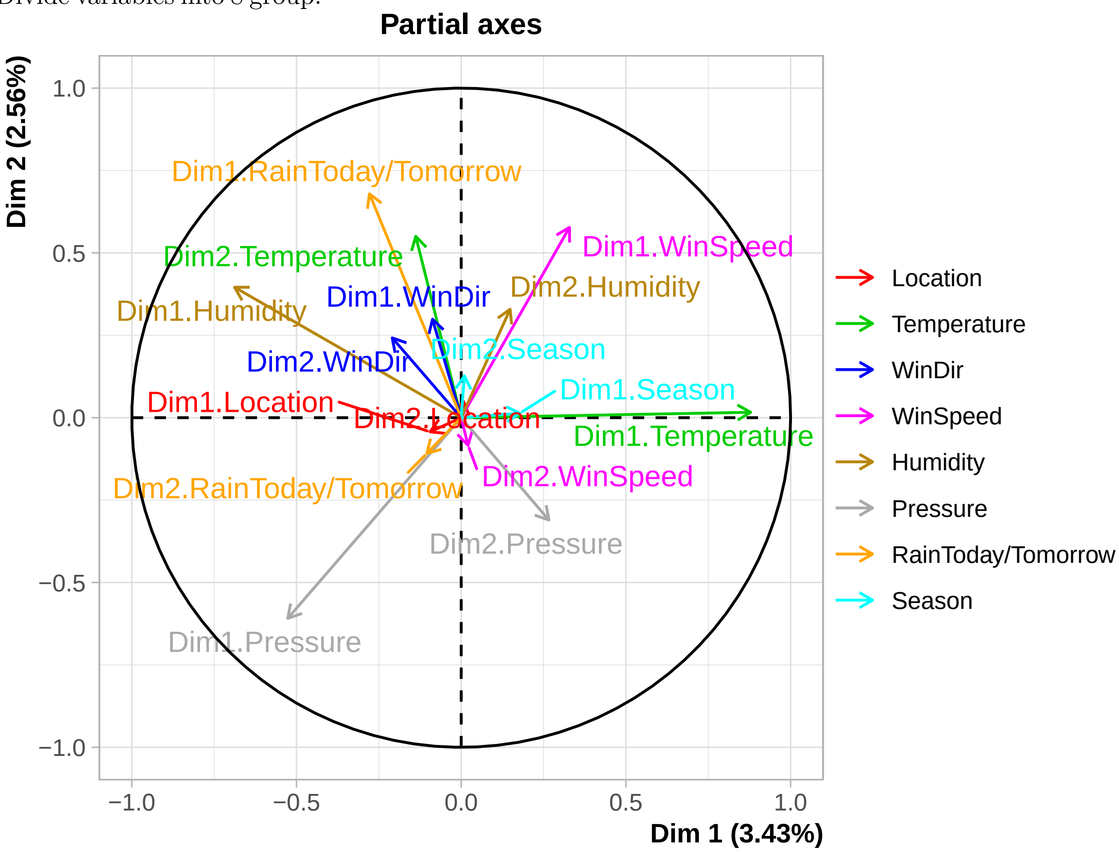
apply pca only on the numerical variables

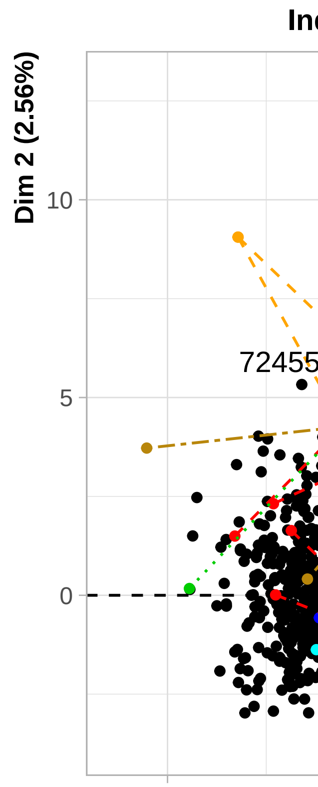

biplot − PCA
Pressure9am
Pressure3pm
Temp3pm
MaxTemp
Temp9am
MinTemp
Humidity9am
Humidity3pm
Rainfall
WindSpeed9am
WindSpeed3pm
WindGustSpeed
Dim2 (20.9%)
Dim1 (40.6%)
2
0
−2
−4
−6
−4
0
4

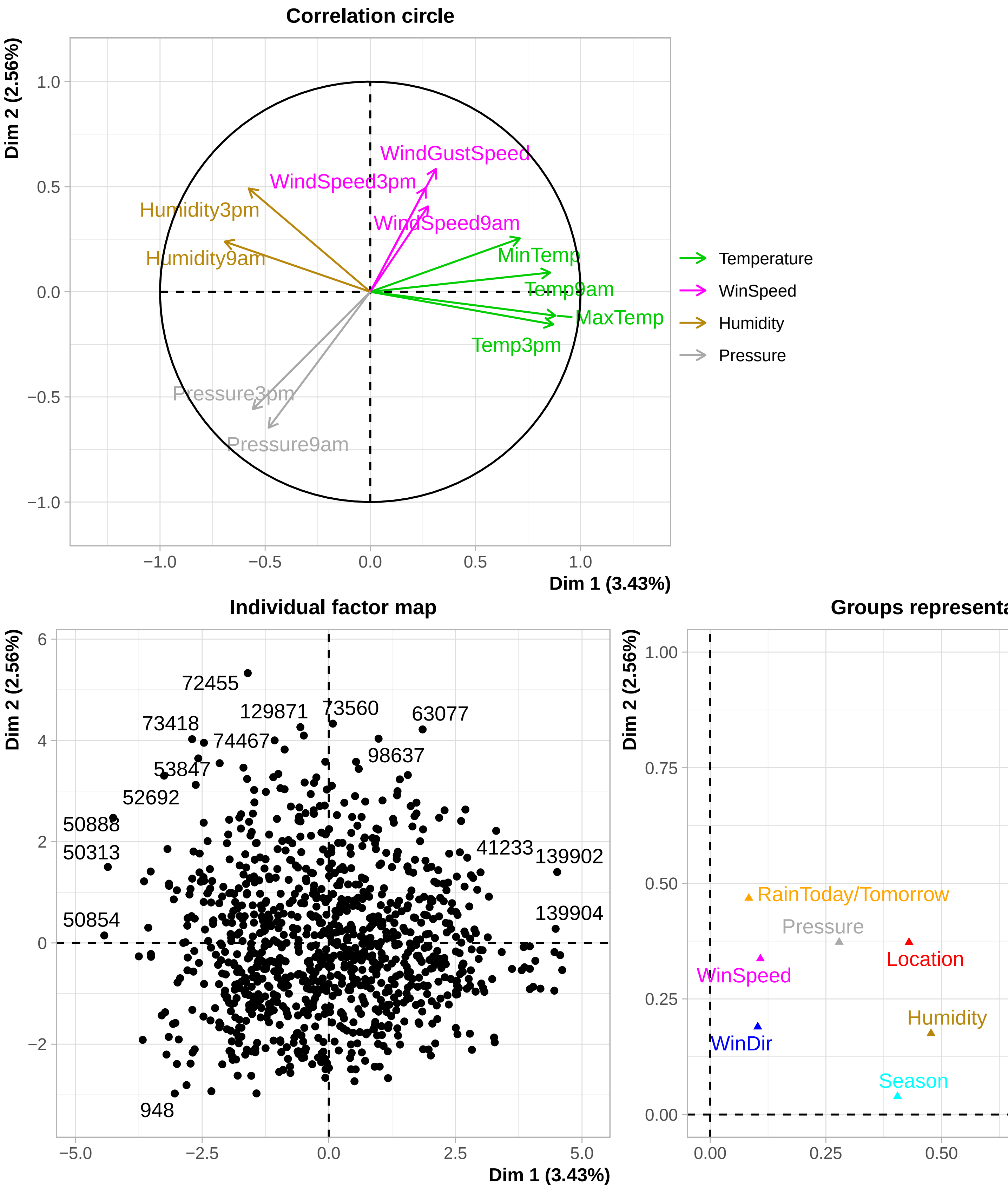
MFA



Divide variables into 8 group.

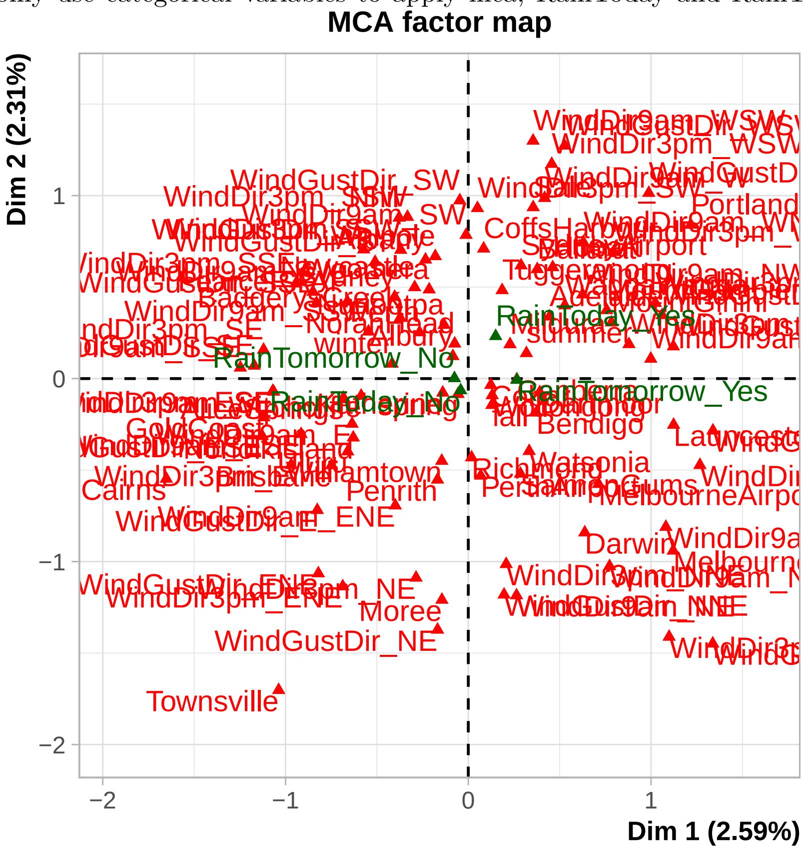


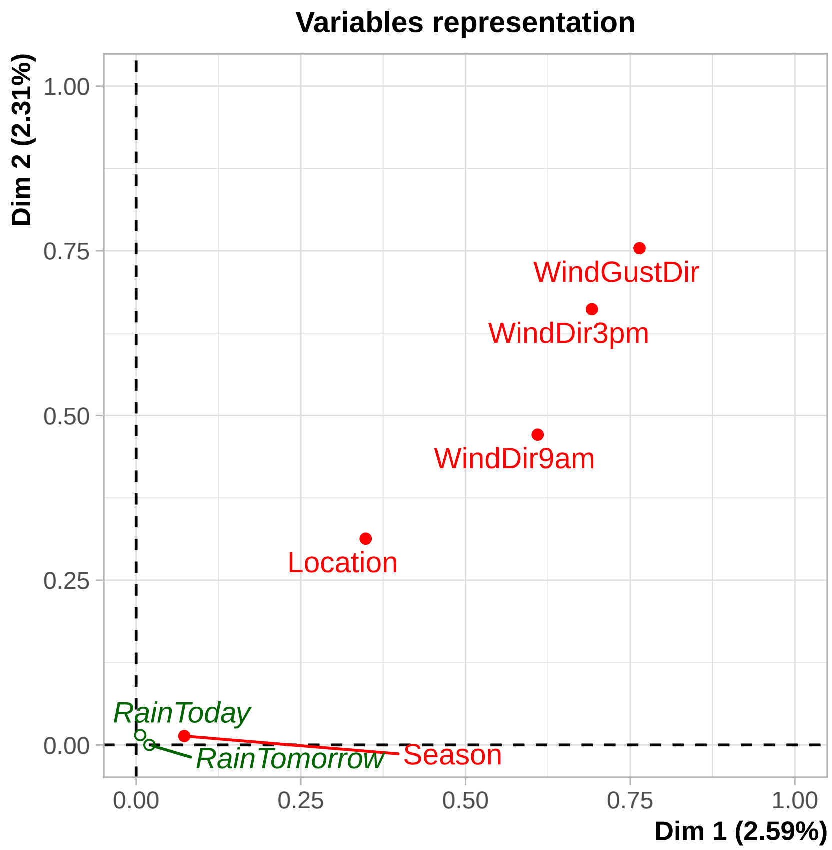



Correlation circle
Dim 2 (2.56%)
Dim 1 (3.43%)
1.0
0.5
0.0
−0.5
−1.0
WindGustSpeed
WindSpeed3pm
WindSpeed9am
Humidity3pm
Humidity9am
MinTemp
Temp9am
MaxTemp
Temp3pm
Pressure3pm
Pressure9am
Temperature
WinSpeed
Humidity
Pressure
Individual factor map
Dim 2 (2.56%)
Dim 1 (3.43%)
−5.0
−2.5
0.0
2.5
5.0
72455
129871
73560
63077
73418
74467
98637
53847
52692
50888
50313
41233
139902
139904
50854
948
Groups representa
Dim 2 (2.56%)
1.00
0.75
0.50
0.25
0.00
RainToday/Tomorrow
Pressure
Location
WinSpeed
Humidity
WinDir
Season

MCA

only use categorical variables to apply mca, RainToday and RainTomorrow as supplementary variables

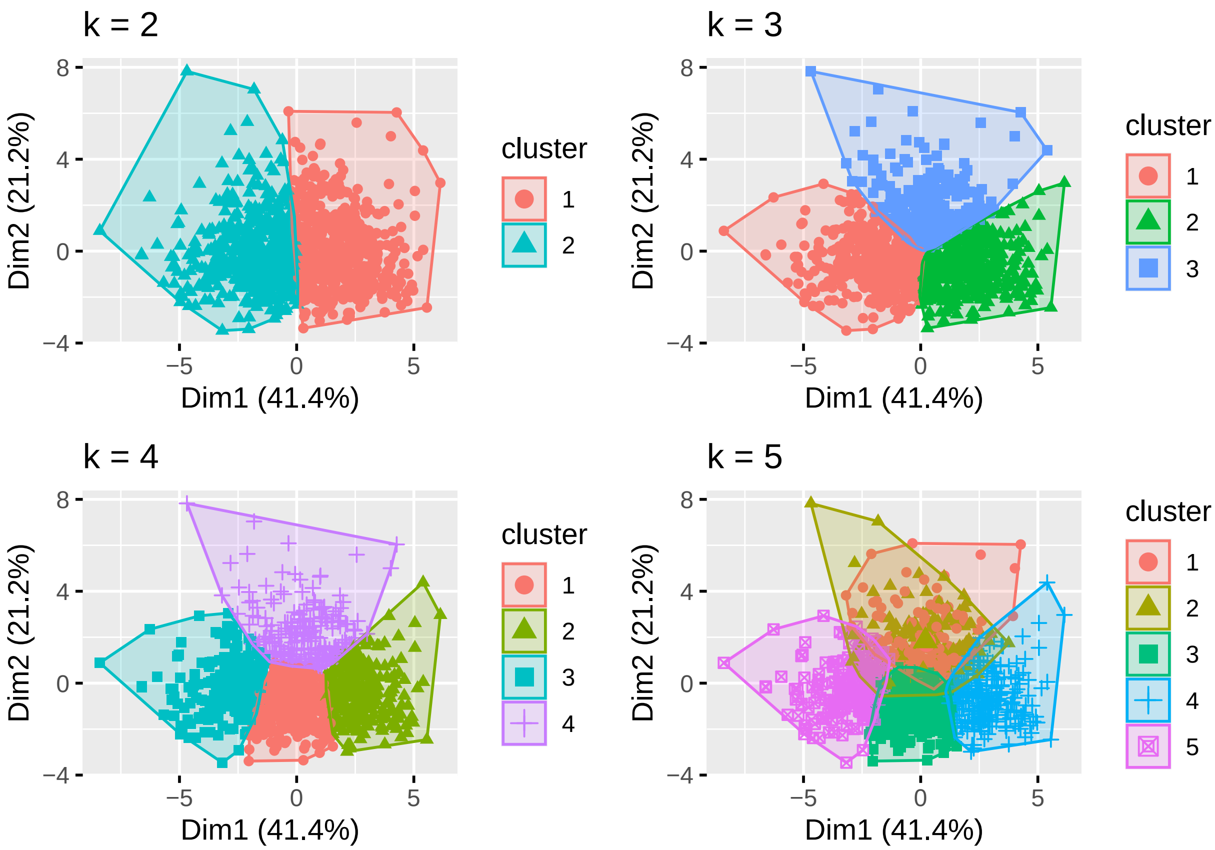




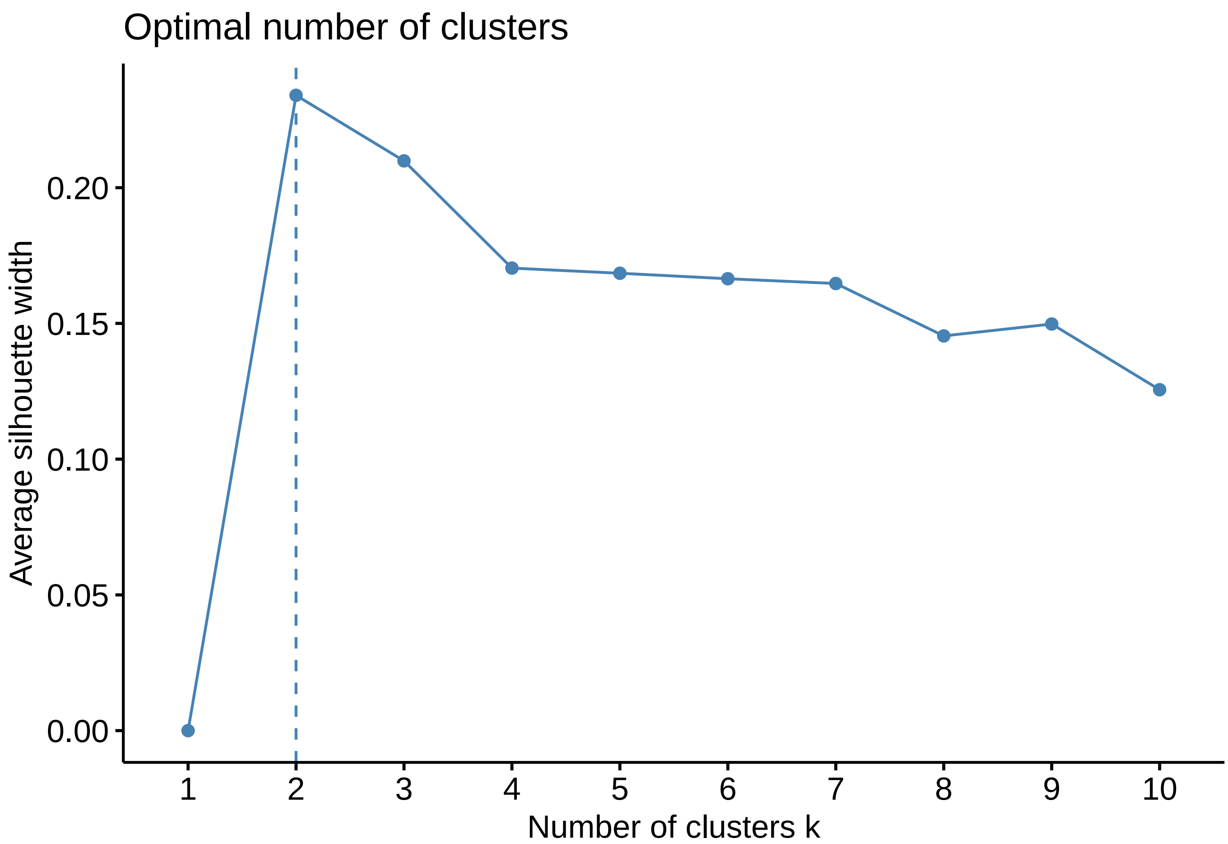
#Clustering

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.

The first approach with clustering method have been with the traditional partition methodology applying K-Means algorithm, since is the computationally less expensive technique. 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.

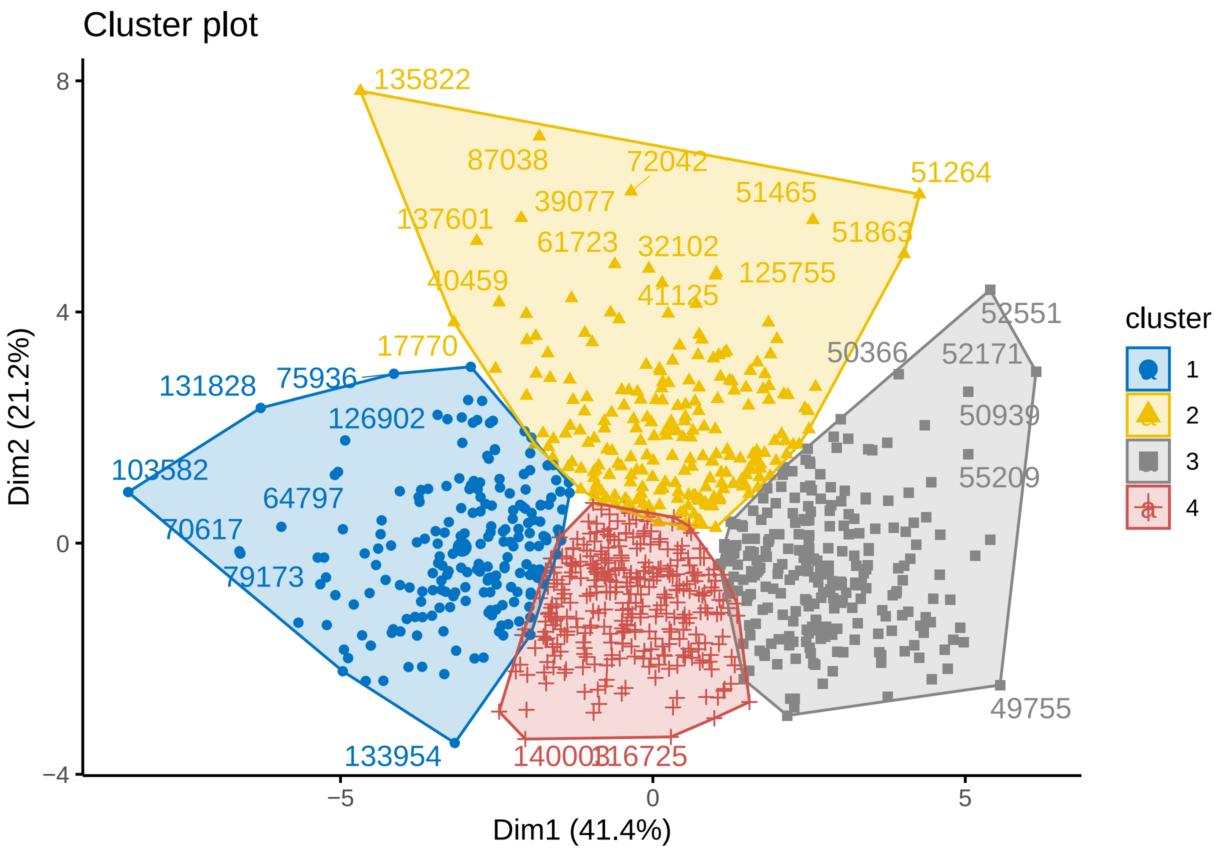


The object of our analysis then will be based on this plot.

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 respectively low and high values.

```
##         MinTemp       MaxTemp      Rainfall WindGustSpeed  WindSpeed9am
## 1 -0.6508956243 -0.7300318512  0.1316102290 -0.2496811333 -0.2159006832
## 2  0.6948146271  0.8066280912 -0.1583370081  0.2420988455  0.2230822078
##    WindSpeed3pm   Humidity9am   Humidity3pm   Pressure9am   Pressure3pm
## 1 -0.2292815024  0.4375383565  0.3162830386  0.4092452476  0.4748082130
## 2  0.3256221658 -0.5594113195 -0.4315550132 -0.4924192670 -0.5688877065
##         Temp9am      Temp3pm
## 1 -0.7164775017 -0.700733183
## 2  0.8080650099  0.805604279
```

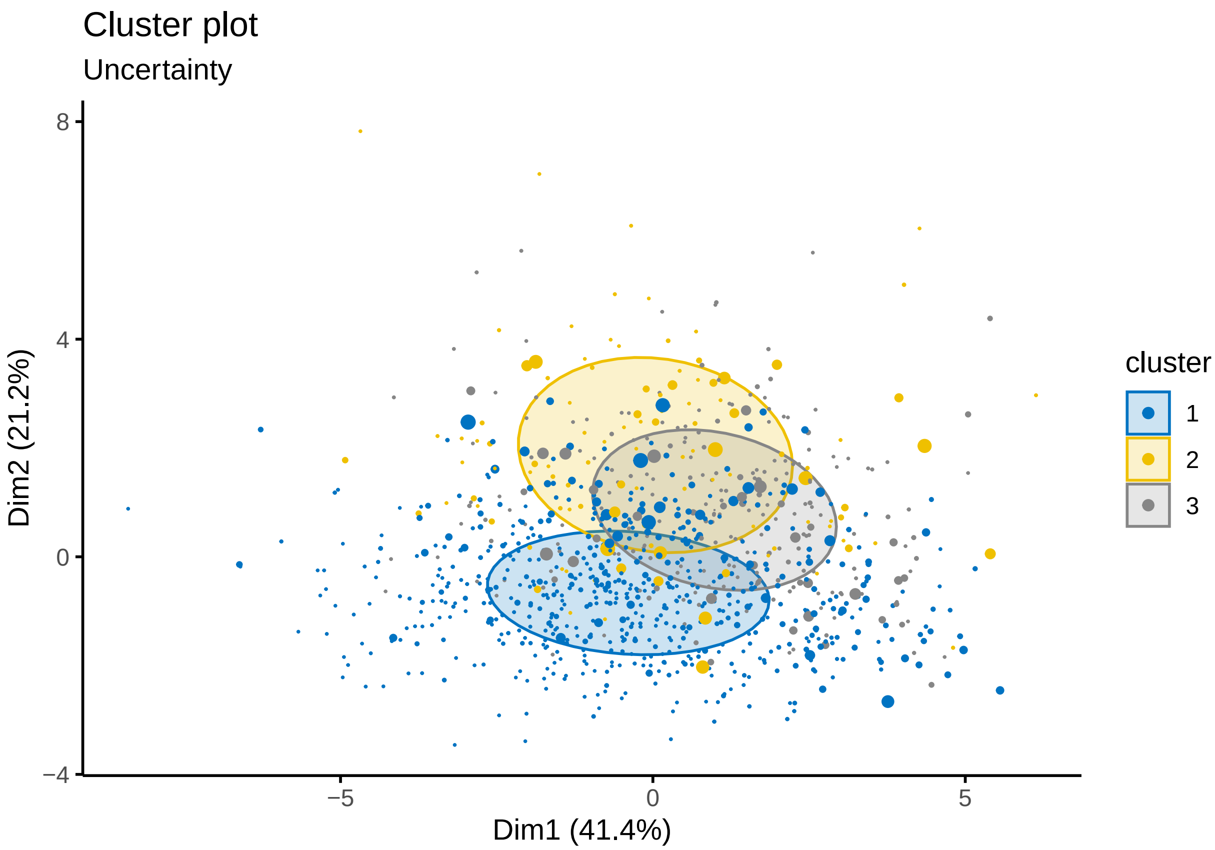
Another trial to identify other kind of clusters shapes have been done applying a mixed approach, using a hierarchical clustering to determine the shape of clusters. The number of clusters will be specified by the parameter `k=4`. This time we will observe the characteristic of four different clusters.



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 3 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 and low pressure and temperatures.

```
##         MinTemp       MaxTemp       Rainfall WindGustSpeed  WindSpeed9am
## 1  1.1670099003  1.3188507646 -0.24550873056  0.4978639921  0.3890422658
## 2 -0.1005353883 -0.6058158347  0.77189658866  0.8055273653  0.7137216930
## 3 -0.9957310840 -0.9748379351 -0.06980687292 -0.6216590987 -0.5333489238
## 4  0.1327010344  0.4313491410 -0.38231612681 -0.4435639289 -0.3517579894
##    WindSpeed3pm   Humidity9am   Humidity3pm   Pressure9am    Pressure3pm
## 1  0.6086365708 -0.8740941188 -0.6658488159 -0.7912123743 -0.91693149091
## 2  0.7053517398  0.3626443953  0.4484736763 -0.5864035319 -0.44542230317
## 3 -0.5290437165  0.6216849951  0.4065291537  0.9320913725  0.99134069540
## 4 -0.3519175600 -0.3507035691 -0.3752798884  0.1012723645  0.02021766938
##         Temp9am       Temp3pm
## 1  1.3385931840  1.3140860198
## 2 -0.3530748557 -0.5950152417
## 3 -1.0386737571 -0.9436046064
## 4  0.3091782299  0.4475020009
```

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 five clusters of shape **VEV**.

```
## [1] 3
```

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

Finally let's interpret the output of the clustering. Even this time there is one cluster over representative for the variable rainfall, presenting even higher value than before. As before the features presented by rainy days are almost the same, with the difference that this time the humidity is way higher but than before but the pressure is not that low.

```
##                        [,1]           [,2]            [,3]
## MinTemp        0.04239902300 -0.006341010582 -0.0577299312253
## MaxTemp        0.31796653010 -0.332804452704 -0.5337275830811
## Rainfall      -0.53655048982  0.856034922057  0.8997121262176
## WindGustSpeed -0.15688786499  0.844645239513 -0.0142234681895
## WindSpeed9am  -0.08023932089  0.471916439262 -0.0090454702870
## WindSpeed3pm  -0.05681496572  0.579489078813  0.0563116141571
## Humidity9am   -0.37552526202  0.355070418642  0.5603651060738
## Humidity3pm   -0.39570310708  0.510288700806  0.5471268091921
## Pressure9am    0.05242019853 -0.598805341416 -0.0006673402984
## Pressure3pm   -0.02721738097 -0.575499088274  0.1678063887525
## Temp9am        0.19675844051 -0.127288745935 -0.2965300869339
## Temp3pm        0.33483280468 -0.446471376486 -0.4694766935272
```