Group 1: Multivariate analysis of australian climate data

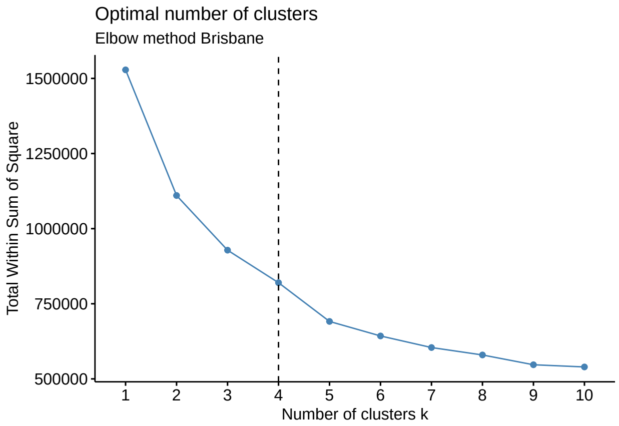
Data input

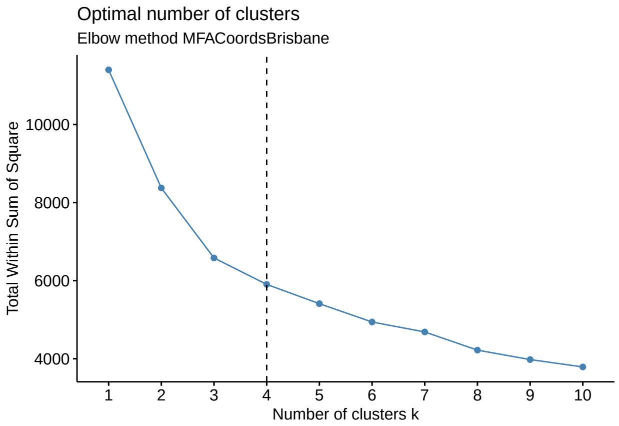
To perform the clustering analysis are used the original datasets (numeric variables) for Brisbane, Perth and Cairns.

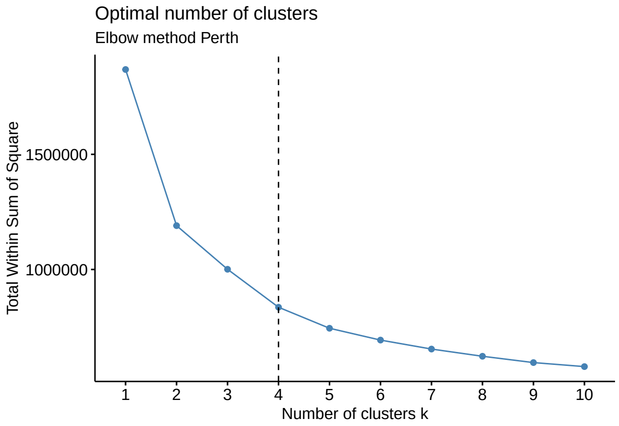
Clustering Analysis

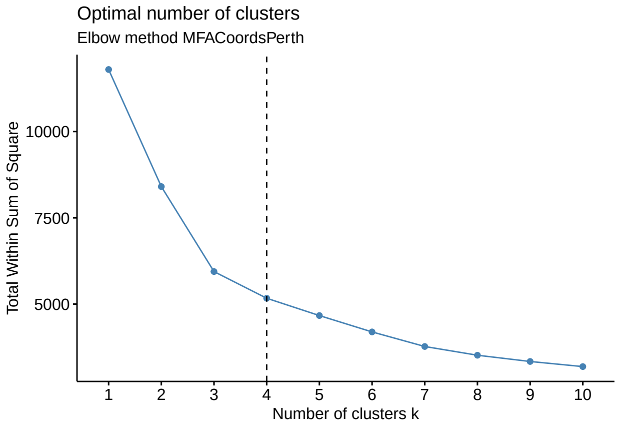
In order to analyze if data presents patterns of association are it is performed a clustering analysis. For this purpose, all incomplete cases remaining are removed and as a first step, the optimal number of clusters are estimated through direct methods: elbow, average silhouette and ASM to choose the most common value of optimal clusters.

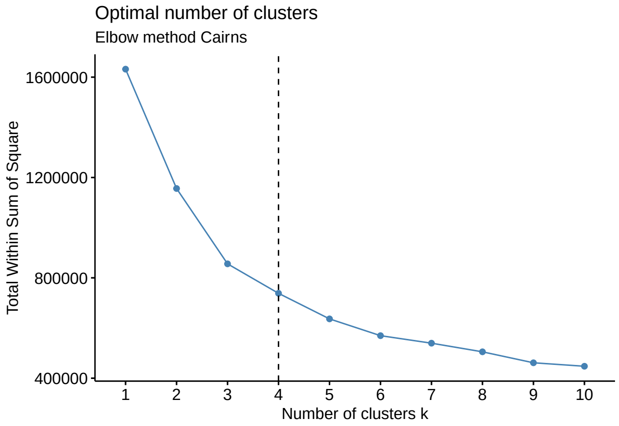
```r
par(mar = c(4,4,.1,.1))
fun01<-function(x){ tmp_df = listall[[x]]
                    tmp_name = names(listall)[x]
                    fviz_nbclust(tmp_df, kmeans, method = "wss") +
                    geom_vline(xintercept = 4, linetype = 2) +
                    labs(subtitle = paste("Elbow method",tmp_name))}
fun02<-function(x){ tmp_df = listall[[x]]
                    tmp_name = names(listall)[x]
                    fviz_nbclust(tmp_df, kmeans, method = "silhouette") +
                    geom_vline(xintercept = 4, linetype = 2) +
                    labs(subtitle = paste("Elbow method",tmp_name))}
fun03<-function(x){ tmp_df = listall[[x]]
                    tmp_name = names(listall)[x]
                    fviz_nbclust(tmp_df, kmeans, method = "gap_stat") +
                    geom_vline(xintercept = 4, linetype = 2) +
                    labs(subtitle = paste("Elbow method",tmp_name))}

wss<-lapply(1:length(listall),fun01)
wss

silhouette<-lapply(1:length(listall),fun02)
silhouette

#Gaps<-lapply(1:length(listall),fun03)
#Gaps
```

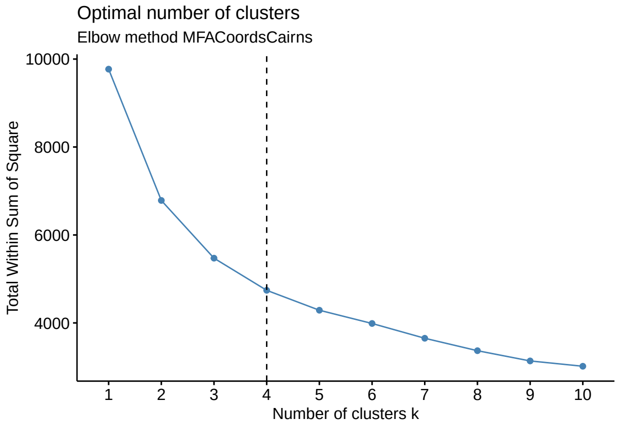

Optimal number of clusters
Elbow method Brisbane
Total Within Sum of Square
1500000
1250000
1000000
750000
500000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

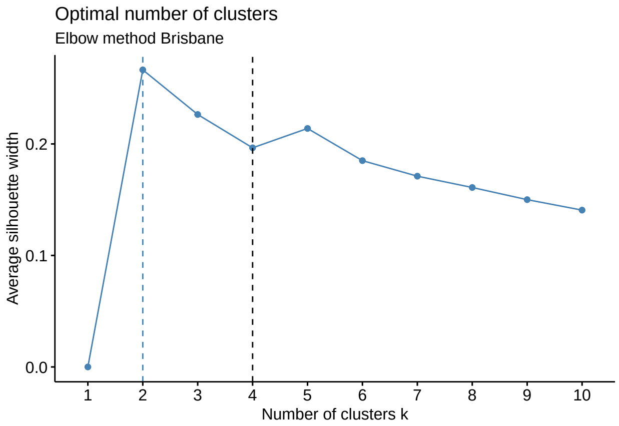

Optimal number of clusters
Elbow method MFACoordsBrisbane
Total Within Sum of Square
10000
8000
6000
4000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

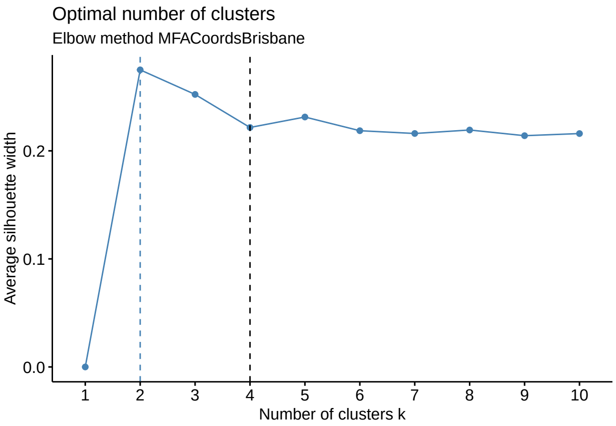

Optimal number of clusters
Elbow method Perth
Total Within Sum of Square
1500000
1000000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

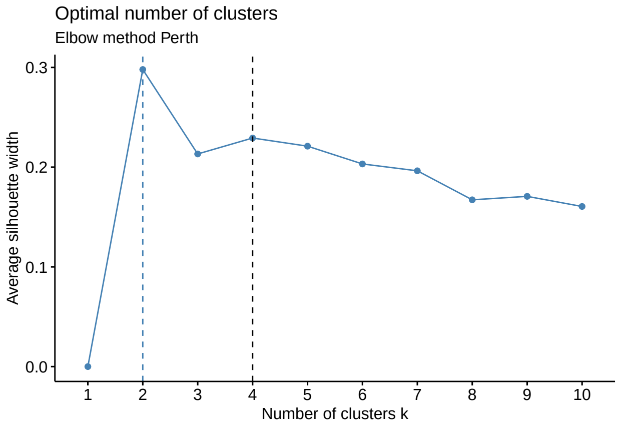

Optimal number of clusters
Elbow method MFACoordsPerth
Total Within Sum of Square
10000
7500
5000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

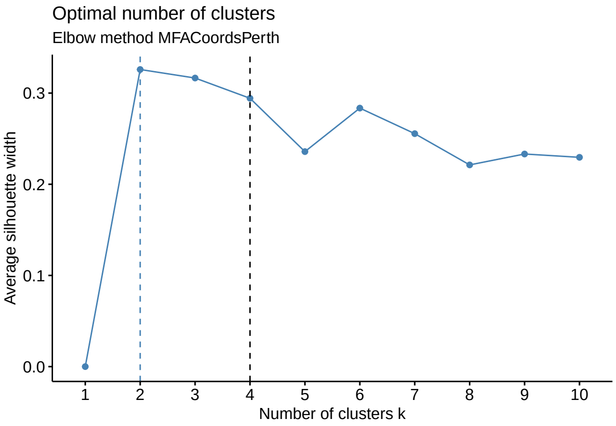

Optimal number of clusters
Elbow method Cairns
Total Within Sum of Square
1600000
1200000
800000
400000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

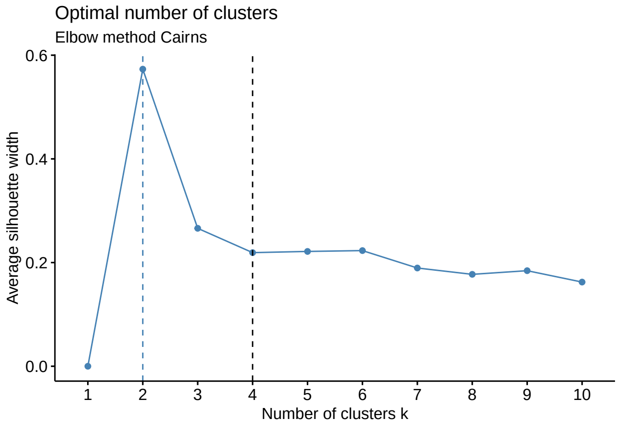

Optimal number of clusters
Elbow method MFACoordsCairns
Total Within Sum of Square
10000
8000
6000
4000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

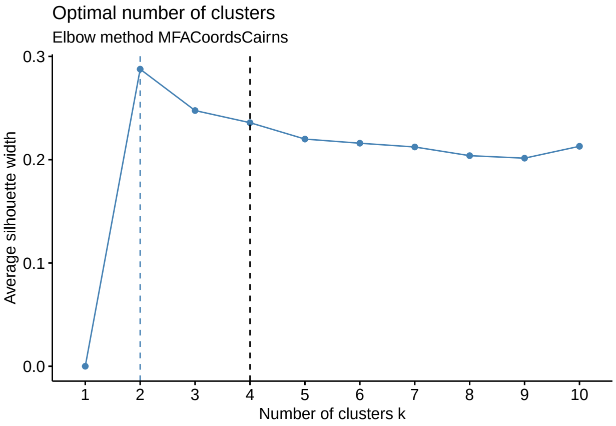












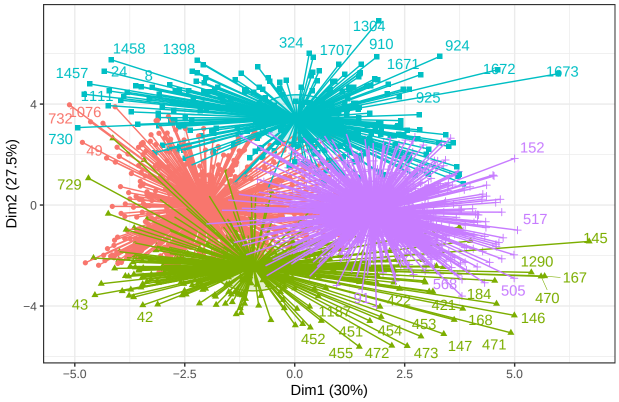
Given the results provided by the methods, it can be concluded the clustering can be performed with 4 cluster for all the dataset, the original numerical variables and the coordinates of the performed MCA.

```
funVizKm<- function(i){ tmp_df = listall[[i]];
                        tmp_kmeans = kmeans(x = listall[[i]], centers = 4)
                        tmp_name = names(listall)[i]
                        fviz_cluster(object = tmp_kmeans, data = listall[[i]],
                                     show.clust.cent = TRUE, ellipse.type = "euclid",
                                     star.plot = TRUE, repel = TRUE) +
                     theme_bw() + theme(legend.position = "none") +
                          labs(title = paste("Results clustering K-means (4 clusters)",
                                             tmp_name))}
```

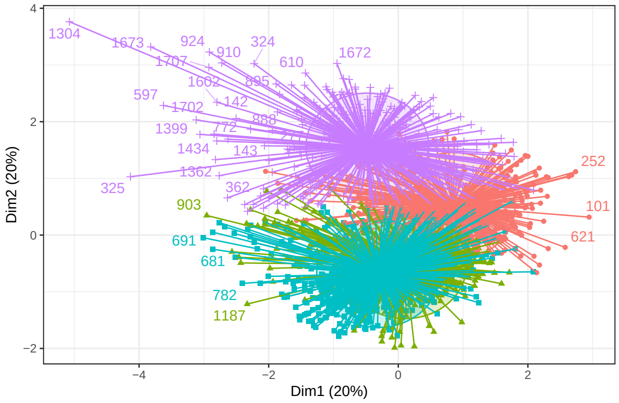
```
VizKmeans<-lapply(1:length(listall),funVizKm)
VizKmeans
```

```
## Warning: ggrepel: 1751 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps

## Warning: ggrepel: 1767 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps

## Warning: ggrepel: 1771 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps

## Warning: ggrepel: 1761 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps

## Warning: ggrepel: 1618 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps

## Warning: ggrepel: 1592 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

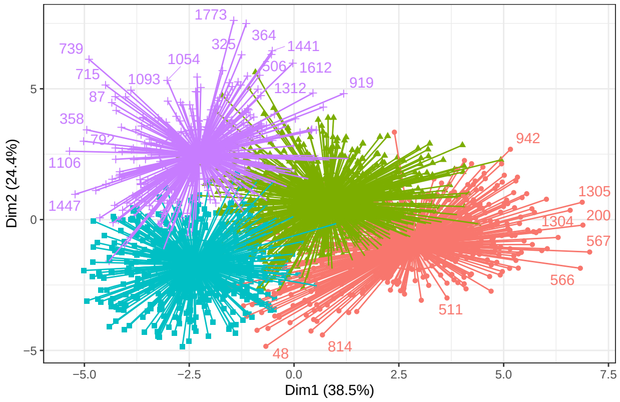
Results clustering K−means (4 clusters) Brisbane



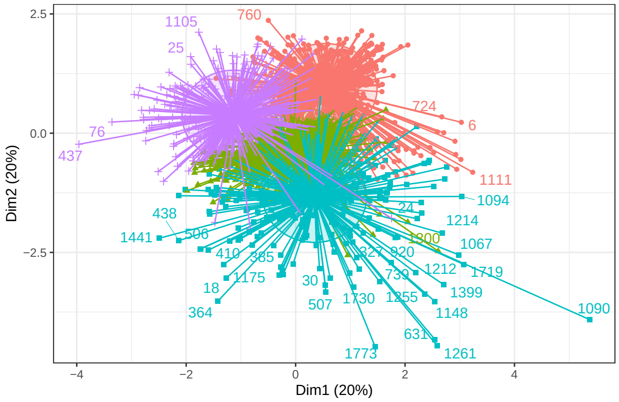
Results clustering K−means (4 clusters) MFACoordsBrisbane

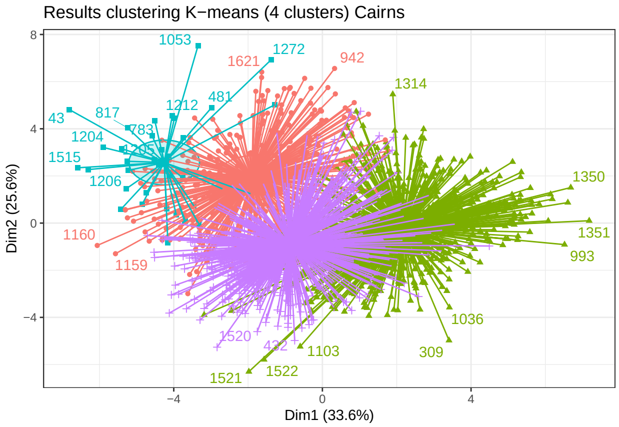


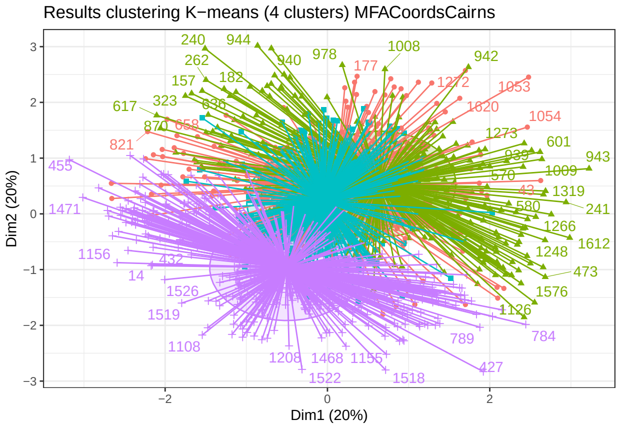
Results clustering K−means (4 clusters) Perth



Results clustering K−means (4 clusters) MFACoordsPerth







```
funKm<- function(i){ tmp_df = listall[[i]];
                        tmp_kmeans = kmeans(x = listall[[i]], centers = 4)
                        listall[[i]]<-add_column(listall[[i]], KmeansCluster =
                                                 tmp_kmeans$cluster)}


Kmeans<-lapply(1:length(listall),funKm)
names(Kmeans)<-c("Brisbane","MFACoordsBrisbane","Perth","MFACoordsPerth",
                 "Cairns","MFACoordsCairns")

funMetrics<-function(i){ tmp_df = listall[[i]];
                        tmp_kmeans = kmeans(x = listall[[i]], centers = 4)
                        print("ClusterCenter")
                        print(tmp_kmeans$center)
                        print("Total Sum of Squares")
                        print(tmp_kmeans$totss)
                        print("Within-cluster sum of squares")
                        print(tmp_kmeans$withinss)
                        print("Total within-cluster sum of squares")
                        print(tmp_kmeans$tot.withinss)
                        print("Between-cluster sum of squares")
                        print(tmp_kmeans$betweenss)}
lapply(1:length(listall), funMetrics)
```

```
## [1] "ClusterCenter"
##   WindGustSpeed WindSpeed9am WindSpeed3pm  MinTemp  MaxTemp TempRange  Temp9am
## 1      31.26821     8.894040    12.019868 11.82517 24.44371 12.618543 18.71060
## 2      23.31474     5.515936     8.864542 13.12550 23.89701 10.771514 18.14004
## 3      28.78369     6.666667     8.968085 18.71241 24.96348  6.251064 21.62730
## 4      31.94233     8.278481    13.592124 19.57060 28.99887  9.428270 25.44374
##    Temp3pm Evaporation  Sunshine Cloud9am Cloud3pm   Rainfall Humidity9am
## 1 23.48344    5.371523 10.158609 1.705298 1.870861  0.8589404    48.80132
## 2 22.54821    3.629880  7.648207 3.575697 3.886454  0.9195219    69.37849
## 3 22.66667    4.023759  2.658156 7.067376 6.978723 17.7198582    83.41135
## 4 27.05274    6.364838  8.908439 4.288326 4.095640  1.2585091    61.78903
##   Humidity3pm Pressure9am Pressure3pm
## 1    33.11258    1018.128    1014.756
## 2    52.17729    1021.823    1018.462
## 3    78.55319    1015.974    1013.371
## 4    56.95359    1015.556    1012.671
## [1] "Total Sum of Squares"
## [1] 1528481
## [1] "Within-cluster sum of squares"
## [1] 144878.1 153141.1 304244.6 217819.7
## [1] "Total within-cluster sum of squares"
## [1] 820083.5
## [1] "Between-cluster sum of squares"
## [1] 708397.3
## [1] "ClusterCenter"
##        Dim.1      Dim.2       Dim.3      Dim.4       Dim.5
## 1 -0.4679027  1.0628077  1.07076279  0.1362153  0.07018321
## 2  2.3009142 -0.4046405  0.02016907 -0.1334548 -0.10350373
## 3 -1.0466814 -1.1543413  0.03805446 -0.1390103 -0.08686800
## 4 -0.2730182  1.0975399 -0.72408786  0.1954259  0.14755772
## [1] "Total Sum of Squares"
## [1] 11400.06
## [1] "Within-cluster sum of squares"
## [1] 1287.370 2023.041 1464.653 1283.148
## [1] "Total within-cluster sum of squares"
## [1] 6058.213
## [1] "Between-cluster sum of squares"
## [1] 5341.851
## [1] "ClusterCenter"
##   WindGustSpeed WindSpeed9am WindSpeed3pm   MinTemp  MaxTemp TempRange  Temp9am
## 1      38.72158    11.978583     17.74959 15.244646 26.38023 11.135585 20.76145
## 2      26.30435     7.107551     11.03204  7.403204 20.46133 13.058124 13.25080
## 3      40.69968    10.974441     15.37700 12.796805 20.29137  7.494569 16.05879
## 4      36.31591    13.404545     13.90682 15.762727 32.34909 16.586364 23.13932
##    Temp3pm Evaporation  Sunshine Cloud9am Cloud3pm   Rainfall Humidity9am
## 1 24.48040    6.799671  9.790115 3.680395 3.630972 0.85140033    56.24053
## 2 19.63066    2.689703  8.208924 2.981693 3.320366 0.96430206    73.66133
## 3 18.30319    3.447284  4.685623 6.012780 5.974441 7.20319489    80.61661
## 4 30.78000    8.867273 11.225455 2.163636 2.600000 0.03590909    41.70227
##   Humidity3pm Pressure9am Pressure3pm
## 1    47.47941    1014.771    1012.789
## 2    45.91076    1024.166    1021.405
## 3    68.70927    1012.915    1011.884
## 4    24.95909    1016.649    1013.167
## [1] "Total Sum of Squares"
## [1] 1868453
## [1] "Within-cluster sum of squares"
## [1] 253351.5 171586.3 216609.1 194392.6
## [1] "Total within-cluster sum of squares"
## [1] 835939.6
## [1] "Between-cluster sum of squares"
## [1] 1032514
## [1] "ClusterCenter"
##        Dim.1      Dim.2       Dim.3       Dim.4        Dim.5
## 1 -0.3337137 -1.7020888  0.11261539  0.03317668  0.009607726
## 2  1.6583060  0.5476833  0.02635364 -0.11815647 -0.064714919
## 3 -2.5107697  1.3188696  0.58302477 -0.62322638  0.004254587
## 4 -0.5925200  0.5900040 -0.72246307  0.69861475  0.107150386
## [1] "Total Sum of Squares"
## [1] 11796.84
## [1] "Within-cluster sum of squares"
## [1] 1395.001 1394.982 1294.771 1084.667
## [1] "Total within-cluster sum of squares"
## [1] 5169.419
## [1] "Between-cluster sum of squares"
## [1] 6627.42
## [1] "ClusterCenter"
##   WindGustSpeed WindSpeed9am WindSpeed3pm  MinTemp  MaxTemp TempRange  Temp9am
## 1      39.77215     18.40506     25.26741 19.29620 29.12136  9.825158 24.86060
## 2      41.64865     14.21622     16.37838 23.35676 28.69730  5.340541 25.58649
## 3      32.16850     11.48661     17.36063 22.35134 30.58126  8.229921 27.11118
## 4      40.77059     16.52941     20.61471 22.92088 28.91471  5.993824 25.37265
##    Temp3pm Evaporation Sunshine Cloud9am Cloud3pm    Rainfall Humidity9am
## 1 27.60016    6.545570 9.416139 3.337025 3.148734   0.5332278    60.94462
## 2 26.97297    4.335135 1.172973 7.594595 7.594595 109.8486486    88.02703
## 3 28.99669    5.652283 7.523780 4.725984 4.669291   1.2727559    70.71339
## 4 26.79853    4.998235 3.357647 6.770588 6.691176  15.9423529    83.47059
##   Humidity3pm Pressure9am Pressure3pm
## 1    49.46994    1015.952    1012.753
## 2    82.62162    1007.789    1005.186
## 3    64.85827    1011.756    1008.652
## 4    77.14706    1012.239    1009.499
## [1] "Total Sum of Squares"
## [1] 1631837
## [1] "Within-cluster sum of squares"
## [1] 207087.4 116149.4 176327.9 238718.4
## [1] "Total within-cluster sum of squares"
## [1] 738283.2
## [1] "Between-cluster sum of squares"
## [1] 893553.5
## [1] "ClusterCenter"
##        Dim.1       Dim.2      Dim.3      Dim.4      Dim.5
## 1 -0.1032490  1.25947262  0.2131545  0.3502464  0.2484787
## 2 -1.7693380  0.03153956 -0.2951626 -0.2372490 -0.1526417
## 3  2.3680252  0.43349087 -0.1065151 -0.2885750 -0.2599948
## 4  0.2584844 -1.27848780  0.2075642  0.1684881  0.1411082
## [1] "Total Sum of Squares"
## [1] 9770.433
## [1] "Within-cluster sum of squares"
## [1] 1240.271 1085.424 1030.351 1385.376
## [1] "Total within-cluster sum of squares"
## [1] 4741.422
## [1] "Between-cluster sum of squares"
## [1] 5029.011
```

```
## [[1]]
## [1] 708397.3
##
## [[2]]
## [1] 5341.851
##
## [[3]]
## [1] 1032514
```

```
##
## [[4]]
## [1] 6627.42
##
## [[5]]
## [1] 893553.5
##
## [[6]]
## [1] 5029.011
```

```
fun04<-function(x) print(names(x))
lapply(Kmeans, fun04)

fun05<-function(x){x[,ncol(x)]
                   x$KMCluster<-x[,ncol(x)]
                   return(x$KMCluster)}
clusters<-lapply(Kmeans,fun05)

BrisbaneClusters<-as.data.frame(cbind(originaldata[[1]],as.factor(clusters[[1]]),
                                      as.factor(clusters[[2]])))
names(BrisbaneClusters)<-c(names(originaldata[[1]]),"KmeansDF","KmeansMFA")

PerthClusters<-as.data.frame(cbind(originaldata[[2]],as.factor(clusters[[3]]),
                                   as.factor(clusters[[4]])))
names(PerthClusters)<-c(names(originaldata[[3]]),"KmeansDF","KmeansMFA")

CairnsClusters<-as.data.frame(cbind(originaldata[[3]],as.factor(clusters[[5]]),
                                    as.factor(clusters[[6]])))
names(CairnsClusters)<-c(names(originaldata[[3]]),"KmeansDF","KmeansMFA")

DFClusters<-list(BrisbaneClusters,PerthClusters,CairnsClusters)

fun06<-function(x){tmpdf=DFClusters[[x]]
                   levels(tmpdf[,24])<-list(C1="1",C2="2",C3="3",C4="4")
                   levels(tmpdf[,25])<-list(G1="1",G2="2",G3="3",G4="4")
                   return(tmpdf)}
data<-lapply(1:length(DFClusters),fun06)
```

```
funtableKmeans<-function(x){table(x$KmeansDF,x$KmeansMFA)}
funtabseason<-function(x){table(x$KmeansDF,x$Season) }
funtabseason2<-function(x){table(x$KmeansMFA,x$Season) }
funtabseason2<-function(x){table(x$KmeansMFA,x$Season) }

lapply(data, funtableKmeans)
```

```
## [[1]]
##
##       G1  G2  G3  G4
##   C1   0  29 183  90
##   C2  57 411  36 207
##   C3 263  14   1   4
##   C4  71  40 385   6
##
```

```
## [[2]]
##
##       G1  G2  G3  G4
##   C1  12  97 204   0
##   C2  79 192  50 286
##   C3  66  13   0 361
##   C4 382  30  24   1
##
## [[3]]
##
##       G1  G2  G3  G4
##   C1 226  22   0  92
##   C2  54 363 113 105
##   C3   1  92 384 155
##   C4  37   0   0   0
```

```
lapply(data,funtabseason)
```

```
## [[1]]
##
##      autumn spring summer winter
##   C1     48    104     15    135
##   C2    149    204    339     19
##   C3     97     48     96     41
##   C4    166     99      1    236
##
## [[2]]
##
##      autumn spring summer winter
##   C1     57     84     21    151
##   C2    149    198    231     29
##   C3    127     97    198     18
##   C4    127     76      1    233
##
## [[3]]
##
##      dry wet
##   C1 155 185
##   C2 268 367
##   C3 492 140
##   C4   3  34
```

```
lapply(data,funtabseason2)
```

```
## [[1]]
##
##      autumn spring summer winter
##   G1    142     71    113     65
##   G2    114    124    250      6
##   G3    144    134      0    327
##   G4     60    126     88     33
##
## [[2]]
```

```
##
##      autumn spring summer winter
##   G1    163    135      0    241
##   G2    104    114     64     50
##   G3     43     77     18    140
##   G4    150    129    369      0
##
## [[3]]
##
##      dry wet
##   G1  89 229
##   G2  67 410
##   G3 458  39
##   G4 304  48
```

```
funProfile<-function(x){catdes(x, num.var=18, prob = 0.01)
                        catdes(x, num.var=19, prob = 0.01)}


#lapply(temp,funProfile)
```