Clustering

With the climate information from Australia, a segmentation analysis was carried out to identify patterns of association between climate variables and establish whether a grouping method could account for the seasons. According to the K-means method, the similarities between the observations of the three cities: Brisbane, Perth and Cairns. In this report, you can find the most outstanding results. The script developed, the graphical and numerical outputs obtained are available in the project repository on [*github*](https://github.com/pgonzalezpr/MVA_Project)*.*

The procedure carried out by city was: identification and evaluation of the input information to the clustering analysis, verification of the ideal cluster number, obtaining the clusters through the K-means algorithm, assessment of the incidence of the variables in the clusters, evaluation of the relationship of the group with the Rainfall variable.

In all three cases (Brisbane, Perth and Cairns), the optimal number of clusters for each was evaluated using the Elbow method (WSS, Silhouette and Gaps), obtaining a result of two or four groups in most cases. This relationship coincides with the number of seasons of the year or dry and wet periods. The most relevant result for the three cities is shown below.

In this case, as expected, the results varied between the raw data and the scaled data; visually a change can be obtained in the three cities, with greater power of segregation in the scaled data. Some of the statistics for each case are presented below:

|  |  |
| --- | --- |
| Brisbane |  |
|  |  |
| Perth |  |
|  |  |
| Cairns |  |
|  |  |

With these results, the calculation of K-means in R was carried out and the corresponding visualization was produced with the *factoextra* package.

|  |  |
| --- | --- |
| Brisbane |  |
|  |  |
| Perth |  |
|  |  |
| Cairns |  |
|  |  |

Although clearly due to the difference in scales, the sums of squares are lower for the cases scaled by city, in proportion, the results are similar.

|  |  |  |
| --- | --- | --- |
| Brisbane |  |  |
|  | *Original data* | *Scaled data* |
| Total Sum of Squares: | 1.528.481 | 30.532 |
| Total within-cluster sum of squares: | 820.083,5 | 17.873,75 |
| Between-cluster sum of squares: | 708.397.3 | 12.658,25 |
|  |  |  |
| Perth |  |  |
| Total Sum of Squares: | 1.868.453 | 30.532 |
| Total within-cluster sum of squares: | 835.939,6 | 16.373,94 |
| Between-cluster sum of squares: | 1.032.514 | 14.158,06 |
|  |  |  |
| Cairns |  |  |
| Total Sum of Squares: | 1.631.837 | 27.931 |
| Total within-cluster sum of squares: | 738.283,2 | 16.063,58 |
| Between-cluster sum of squares: | 893.553,5 | 11.867,42 |

Subsequently, the results were evaluated by variable for each cluster. Two tools were used, the exploration of the boxplots and the condes / catdes methods of the package *factominer* to characterize centers and profiling. Below, the most representative outputs related with the cluster analysis by main important variables in the discrimination of each cluster.

|  |  |  |
| --- | --- | --- |
| *Brisbane* | | |
|  |  |  |
| *Brisbane Scaled* | | |
|  |  |  |
| *Perth* | | |
|  |  |  |
|  |  |  |
| *Perth Scaled* |  |  |
|  |  |  |
| *Cairns* | | |
|  |  |  |
| *Cairns scaled* | | |
|  |  |  |

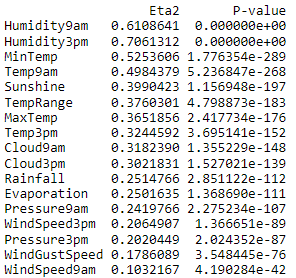
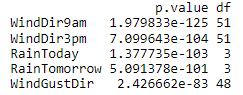
*Analysis for Brisbane*

Using segmentation methods was possible to observe four clusters in Brisbane data. For original data, the cluster were : TempRange, Cloud9am, Cloud3pm and Humidity9am were the variables distributed among the identified cluster with different median. For scaled data, the variables establishing major difference were coincident in: Cloud9am and Cloud3pm.

Comparing the groups using original data and scaled data. The unique pattern of matching clustering was the group 3 of the original data, with the group 3 of the scaled data. As well, comparing the distribution of the variable with the variable season, the majority of the individuals in group 3 of the clustering using the original dataset is related with summer season:

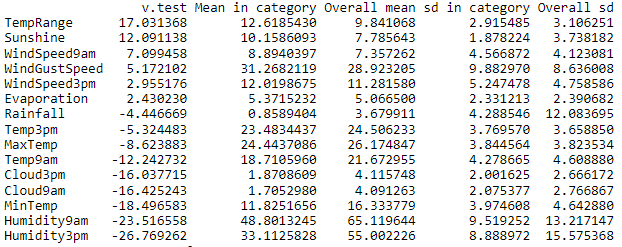
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Scaled data | | | |  |  |  | Scaled data | | | |
|  |  | **G1** | **G2** | **G3** | **G4** |  |  |  | **autumn** | **spring** | **Summer** | **winter** |
| Original | **C1** | 147 | 219 | 551 | 98 |  | Original | **C1** | 259 | 276 | 317 | 163 |
| **C2** | 151 | 2 | 6 | 166 |  | **C2** | 121 | 46 | 88 | 70 |
| **C3** | 0 | 0 | 2 | 41 |  | **C3** | 9 | 6 | 27 | 1 |
| **C4** | 18 | 315 | 81 | 0 |  | **C4** | 71 | 127 | 19 | 197 |

As far as cluster profiling goes, the most statistically correlated variables with the cluster segregation seem to be the ones related with wind direction (categorical) and the ones related with humidity and temperature (numerical). It is worth noting that cluster 1 seems to correspond with less humid than average days, while cluster 3 seems to contain more humid than average days. Average temperature remains fairly stable across clusters.

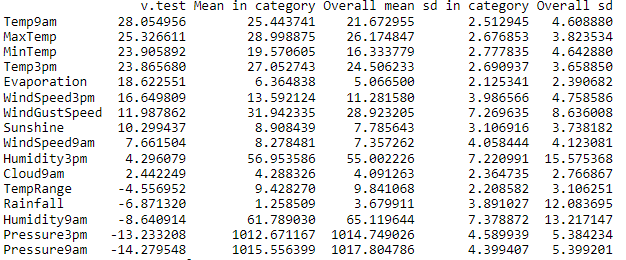


Below the numerical variables summary by clusters in shown

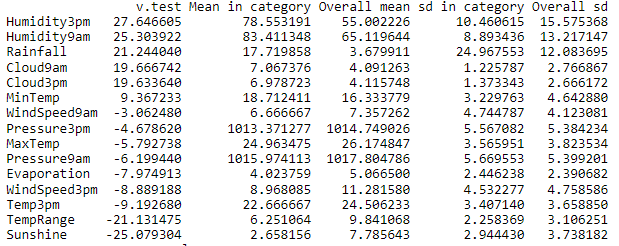
Cluster 1:



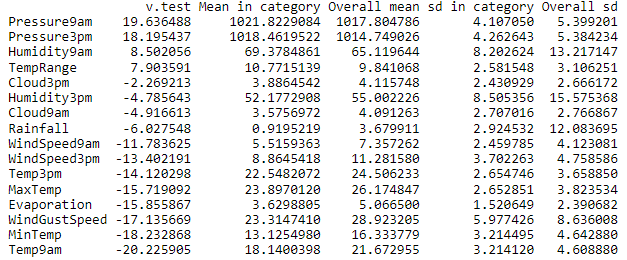
Cluster 2:



Cluster 3:



Cluster 4:



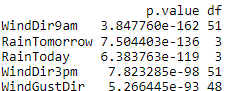
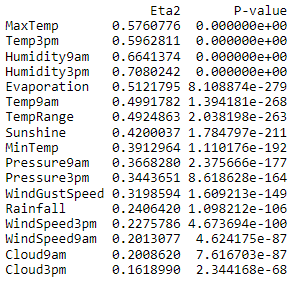
*Analysis for Perth*

For Perth, the clustering analysis using four clusters, the most differentiating of the groups of the original data were: WindSpeed3pm, MaxTemp, TempRange, Temp9am, Humidity9am and Humidity3pm and Cloud9am. For scaled data, the variables establishing major difference were coincident in: MaxTemp, TempRange and were also useful in distinction of clusters MinTemp, Sunshine and Cloud3pm.

Comparing the groups using original data and scaled data there were no pattern of matching clustering. There was no pattern neither for season variable:

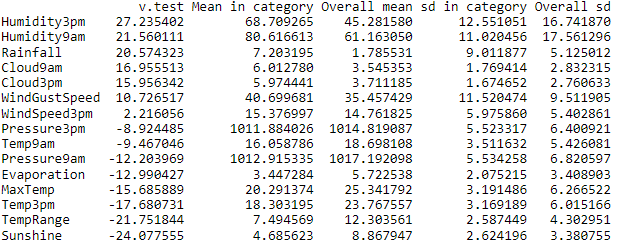
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Scaled data | | | |  |  |  | Scaled data | | | |
|  |  | **G1** | **G2** | **G3** | **G4** |  |  |  | **autumn** | **spring** | **Summer** | **winter** |
| Original | **C1** | 216 | 9 | 204 | 177 |  | Original | **C1** | 149 | 198 | 230 | 29 |
| **C2** | 0 | 54 | 1 | 260 |  | **C2** | 58 | 84 | 22 | 151 |
| **C3** | 327 | 2 | 109 | 2 |  | **C3** | 127 | 97 | 198 | 18 |
| **C4** | 1 | 309 | 109 | 17 |  | **C4** | 126 | 76 | 1 | 233 |

According to the cluster profiling results, variables related with wind direction, boolean rain, temperature and humidity are the ones most correlated with the cluster variable.

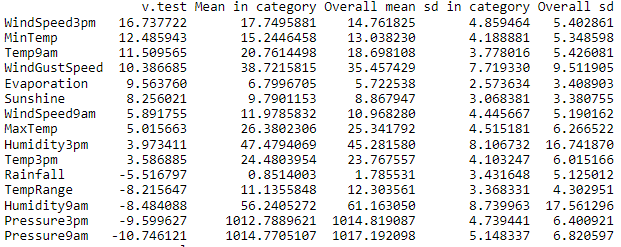
 

The summary for numerical variables according to cluster classification is as follows. Cluster 1 contains more humid than average days, while cluster 1 less humid than average days.

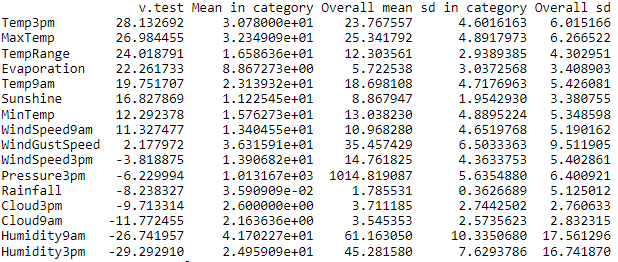
Cluster 1:



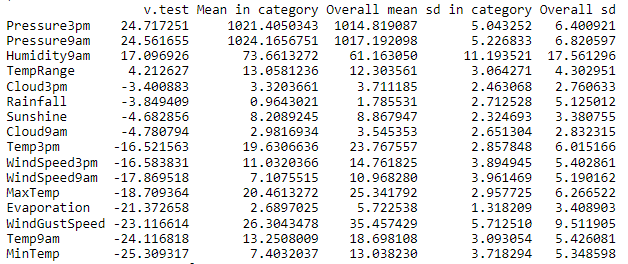
Cluster 2:



Cluster 3:



Cluster 4:



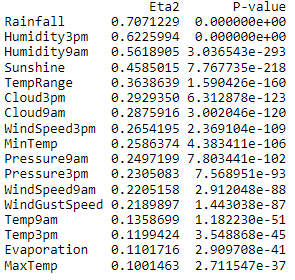
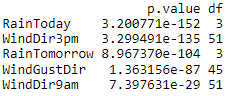
*Analysis for Cairns*

For Cairns, the clustering analysis in the original data were differentiate by variables: TempRange, Sunshine, Cloud9am, Cloud3pm and humidity9am. For scaled data, the main variables was Cloud9am.

In this case, comparing the groups using original data and scaled data there was a coincidence of the group 3 identified using the original data and cluster 1 using the scaled data. Additionally, in this case, the group 3 has association with wet season.

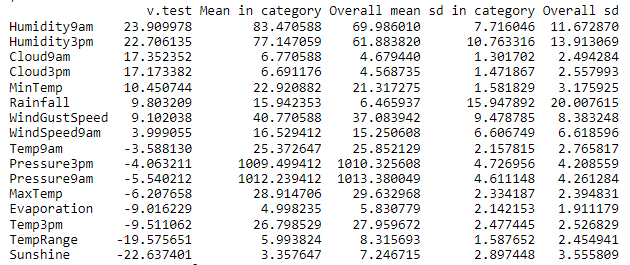
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Scaled data | | | |  |  |  | Scaled data | |
|  |  | **G1** | **G2** | **G3** | **G4** |  |  |  | **dry** | **wet** |
| Original | **C1** | 36 | 98 | 405 | 96 |  | Original | **C1** | 268 | 367 |
| **C2** | 0 | 317 | 110 | 205 |  | **C2** | 492 | 140 |
| **C3** | 36 | 0 | 1 | 0 |  | **C3** | 3 | 34 |
| **C4** | 199 | 0 | 61 | 80 |  | **C4** | 155 | 185 |

In this case, the variables associated with rain, wind direction, rainfall and humidity are the most statistically correlated with the cluster variable.

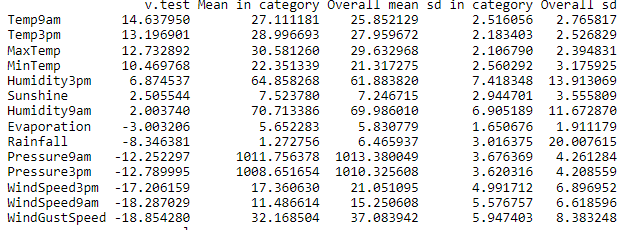


Numerical variables are described by cluster as follows. Cluster 1 corresponds to more humid than average days, clusters 2 and 3 are less rainy than average days and cluster 4 clearly groups those days with a fairly higher than average amount of rainfall.

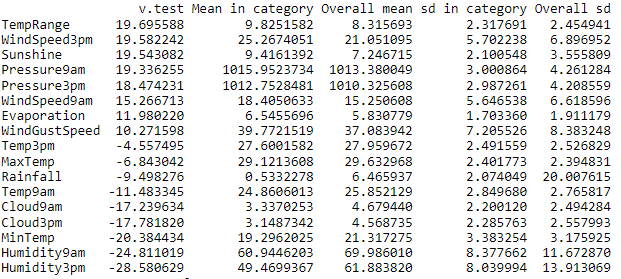
Cluster 1:



Cluster 2:



Cluster 3:



Cluster 4:

