Claudia_Afonso_36273_Rita_Rodrigues_54859

2022-12-01

Assigment A

The data contained in the airpollution.csv was obtained as part of a study on pollution carried out in 41 cities in the USA, in which the following variables were considered:

- so2: Sulphur dioxide content of air in micrograms per cubic meter
- temp: Average annual temperature (F)
- manuf: No. of manufacturing enterprises employing 20 or more workers
- pop: Population size (1970 census) in thousands
- wind: Average wind speed in miles per hour
- precip: Average annual precipitation in inches
- days: Average number of days with precipitation per year

Descriptive analysis

Before the PCA lets perform a descriptive analysis of the data

```
# Run data
airpollution <- read.csv('airpollution.csv', header = TRUE)
head(airpollution)
```

```
##       city so2 temp manuf pop wind precip days
## 1  Phoenix  10 70.3   213 582  6.0   7.05   36
## 2 Little R  13 61.0    91 132  8.2  48.52  100
## 3 San Fran  12 56.7   453 716  8.7  20.66   67
## 4   Denver  17 51.9   454 515  9.0  12.95   86
## 5 Hartford  56 49.1   412 158  9.0  43.37  127
## 6 Wilmingt  36 54.0    80  80  9.0  40.25  114
```

```
# New data only with variables: so2, temp, manuf, pop, wind, precip, days
airpollution_variables <- airpollution[,2:8]
head(airpollution_variables)
```

```
##   so2 temp manuf pop wind precip days
## 1  10 70.3   213 582  6.0   7.05   36
## 2  13 61.0    91 132  8.2  48.52  100
## 3  12 56.7   453 716  8.7  20.66   67
## 4  17 51.9   454 515  9.0  12.95   86
## 5  56 49.1   412 158  9.0  43.37  127
## 6  36 54.0    80  80  9.0  40.25  114
```

```
str(airpollution_variables)
```

```
## 'data.frame':    41 obs. of  7 variables:
##  $ so2   : int  10 13 12 17 56 36 29 14 10 24 ...
##  $ temp  : num  70.3 61 56.7 51.9 49.1 54 57.3 68.4 75.5 61.5 ...
##  $ manuf : int  213 91 453 454 412 80 434 136 207 368 ...
##  $ pop   : int  582 132 716 515 158 80 757 529 335 497 ...
##  $ wind  : num  6 8.2 8.7 9 9 9 9.3 8.8 9 9.1 ...
##  $ precip: num  7.05 48.52 20.66 12.95 43.37 ...
##  $ days  : int  36 100 67 86 127 114 111 116 128 115 ...
```

```
dim(airpollution_variables)
```

```
## [1] 41  7
```

```
# Localization measures
summary(airpollution_variables)
```

```
##       so2              temp           manuf             pop        
##  Min.   :  8.00   Min.   :43.50   Min.   :  35.0   Min.   :  71.0  
##  1st Qu.: 13.00   1st Qu.:50.60   1st Qu.: 181.0   1st Qu.: 299.0  
##  Median : 26.00   Median :54.60   Median : 347.0   Median : 515.0  
##  Mean   : 30.05   Mean   :55.76   Mean   : 463.1   Mean   : 608.6  
##  3rd Qu.: 35.00   3rd Qu.:59.30   3rd Qu.: 462.0   3rd Qu.: 717.0  
##  Max.   :110.00   Max.   :75.50   Max.   :3344.0   Max.   :3369.0  
##       wind            precip           days      
##  Min.   : 6.000   Min.   : 7.05   Min.   : 36.0  
##  1st Qu.: 8.700   1st Qu.:30.96   1st Qu.:103.0  
##  Median : 9.300   Median :38.74   Median :115.0  
##  Mean   : 9.444   Mean   :36.77   Mean   :113.9  
##  3rd Qu.:10.600   3rd Qu.:43.11   3rd Qu.:128.0  
##  Max.   :12.700   Max.   :59.80   Max.   :166.0
```

```
# Dispersion measures
library(dplyr)
```

```
## 
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
## 
##     filter, lag

## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```
airpollution_variables %>% summarise_if(is.numeric, sd)
```

```
##        so2     temp    manuf     pop     wind   precip     days
## 1 23.47227 7.227716 563.4739 579.113 1.428644 11.77155 26.50642
```

Since the measure units are not the same for all variables and the mean and standard deviation are also quite different, the PCA should be done based on the correlation matrix.

Task 1. Perform a Principal Components Analysis of the data set.

```
# Perform PCA

pca_airpollution <- princomp(airpollution_variables, cor = TRUE)
print(summary(pca_airpollution), loadings = TRUE)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.6517021 1.2297702 1.1810897 0.9444529 0.58887916
## Proportion of Variance 0.3897314 0.2160478 0.1992819 0.1274273 0.04953981
## Cumulative Proportion  0.3897314 0.6057792 0.8050611 0.9324884 0.98202821
##                           Comp.6      Comp.7
## Standard deviation     0.3166822 0.159733920
## Proportion of Variance 0.0143268 0.003644989
## Cumulative Proportion  0.9963550 1.000000000
##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## so2     0.490                 0.404  0.730  0.183  0.150
## temp   -0.315         0.677 -0.185  0.162  0.611
## manuf   0.541 -0.226  0.267        -0.164        -0.745
## pop     0.488 -0.282  0.345 -0.113 -0.349         0.649
## wind    0.250        -0.311 -0.862  0.268  0.150
## precip         0.626  0.492 -0.184  0.161 -0.554
## days    0.260  0.678 -0.110  0.110 -0.440  0.505
```

Task 2. Choose the principal components retained and the importance of each one.

```
# Obtain Eigenvalues and Eigenvectors (based on the correlation matrix)

## 1st) Determine the correlation matrix
cor_airpollution <- cor(airpollution_variables)

## 2nd) Obtain eigenvalues and eigenvectors
eigen_airpollution <- eigen(cor_airpollution)
eigen_airpollution
```

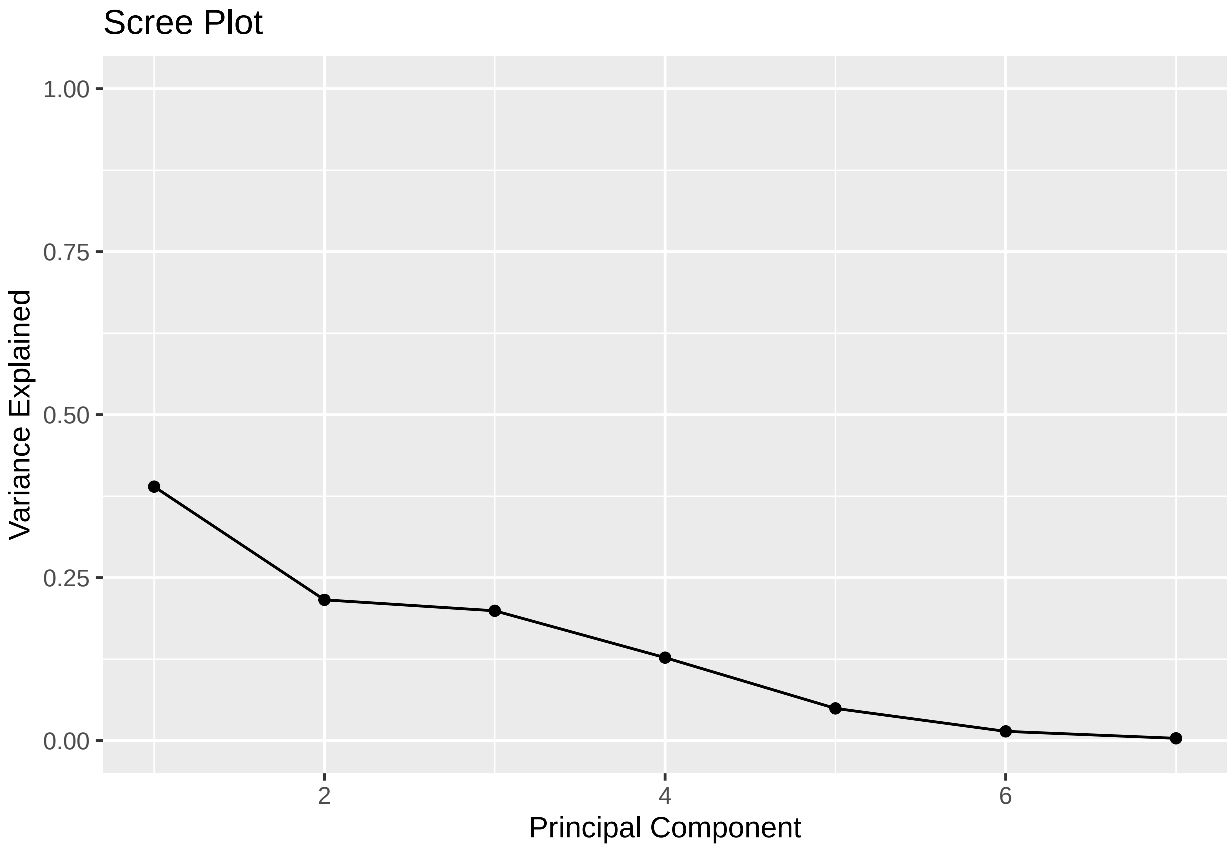
```
## eigen() decomposition
## $values
## [1] 2.72811968 1.51233485 1.39497299 0.89199129 0.34677866 0.10028759 0.02551493
##
## $vectors
##               [,1]        [,2]       [,3]        [,4]       [,5]        [,6]
## [1,]  0.4896988171 -0.08457563 -0.0143502  0.40421007  0.7303942 -0.18334573
## [2,] -0.3153706901  0.08863789 -0.6771362 -0.18522794  0.1624652 -0.61066107
## [3,]  0.5411687028  0.22588109 -0.2671591 -0.02627237 -0.1641011  0.04273352
## [4,]  0.4875881115  0.28200380 -0.3448380 -0.11340377 -0.3491048  0.08786327
## [5,]  0.2498749284 -0.05547149  0.3112655 -0.86190131  0.2682549 -0.15005378
## [6,]  0.0001873122 -0.62587937 -0.4920363 -0.18393719  0.1605988  0.55357384
## [7,]  0.2601790729 -0.67796741  0.1095789  0.10976070 -0.4399698 -0.50494668
##              [,7]
## [1,]  0.149529278
## [2,] -0.023664113
## [3,] -0.745180920
## [4,]  0.649125507
## [5,]  0.015765377
## [6,] -0.010315309
## [7,]  0.008217393
```

According to Kaiser's criterion: we should retain the first three principal components, since these have eigenvalues greater than 1 (2.728, 1.512 and 1.395 > 1).

To complement this result, we also performed a scree-plot (which corresponds to a line plot of the eigenvalues of the principal components).

```
# Calculate total variance explained by each principal component

var_explained_airpollution = pca_airpollution$sdev^2 / sum(pca_airpollution$sdev^2)

# Create scree plot
library(ggplot2)

qplot(c(1:7), var_explained_airpollution) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot") +
  ylim(0,1)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
```



According to the scree plot, we should retain only two principal components. However, according to Kaiser's criterion, we should retain three principal components.

Since the third principal component has a eigenvalue which is significantly higher than 1, the result obtained using Kaiser's criterion takes precedence over the one obtained using the scree plot.

Therefore, three principal components were retained.

With three principal components, we have a total of explained variance of 80.5%, which is reasonably high.

Task 3. Explain the importance of the variables for the explanation of each of the principal components retained.

To interpret the principal components retained, it is necessary to compute the correlations between these and the initial variables.

```
component_matrix <- cor(airpollution_variables,pca_airpollution$scores)
component_matrix
```

```
##              Comp.1      Comp.2      Comp.3      Comp.4      Comp.5
## so2     0.8088365434  0.10400859  0.01694887  0.38175738  0.43011391
## temp   -0.5208984175 -0.10900423  0.79975860 -0.17493906  0.09567234
## manuf   0.8938494595 -0.27778184  0.31553891 -0.02481301 -0.09663572
## pop     0.8053502866 -0.34679989  0.40728458 -0.10710452 -0.20558056
## wind    0.4127189331  0.06821718 -0.36763244 -0.81402520  0.15796972
## precip  0.0003093839  0.76968782  0.58113903 -0.17372001  0.09457328
## days    0.4297383099  0.83374415 -0.12942257  0.10366381 -0.25908903
##             Comp.6       Comp.7
## so2     0.05806232  0.023884898
## temp    0.19338547 -0.003779961
## manuf  -0.01353294 -0.119030670
## pop    -0.02782473  0.103687362
## wind    0.04751936  0.002518265
## precip -0.17530696 -0.001647705
## days    0.15990761  0.001312596
```

The initial variables that are strongly correlated with the 1st principal component are: “so2”, “manuf” and “pop”. Therefore, these are the variables that contribute more to the 1st principal component.

The initial variables that are strongly correlated with the 2nd principal component are: “precip” and “days”. Therefore, these are the variables that contribute more to the 2nd principal component.

The initial variables that are strongly correlated with the 3rd principal component are: “temp” and “precip”. Therefore, these are the variables that contribute more to the 3rd principal component.

```
# Proportion of (standardized) population standard deviation (sd) due to the 1st Principal Component

sqrt(eigen_airpollution$values[1]/7)
```

```
## [1] 0.6242847
```

The 1st principal component explains approximately 62.43% of the (standardized) population standard deviation.

```
# Proportion of (standardized) population standard deviation (sd) due to the 2nd Principal Component

sqrt(eigen_airpollution$values[2]/7)
```

```
## [1] 0.4648095
```

The 2nd principal component explains approximately 46.48% of the (standardized) population standard deviation.

```
# Proportion of (standardized) population standard deviation (sd) due to the 3rd Principal Component

sqrt(eigen_airpollution$values[3]/7)
```

```
## [1] 0.44641
```

The 3rd principal component explains approximately 44.64% of the (standardized) population standard deviation.

Now let's calculate the contribution of the important variables for each principal component retained.

For the 1st principal component, we have:

```
# 1st PC
## so2
a_11_square <-(component_matrix[1,1]/sqrt(eigen_airpollution$values[1]))^2
a_11_square
```

```
## [1] 0.2398049
```

```
# 1st PC
## manuf
a_31_square <-(component_matrix[3,1]/sqrt(eigen_airpollution$values[1]))^2
a_31_square
```

```
## [1] 0.2928636
```

```
# 1st PC
## pop
a_41_square <-(component_matrix[4,1]/sqrt(eigen_airpollution$values[1]))^2
a_41_square
```

```
## [1] 0.2377422
```

For the 2nd principal component, we have:

```
# 2nd PC
## precip
a_62_square <-(component_matrix[6,2]/sqrt(eigen_airpollution$values[2]))^2
a_62_square
```

```
## [1] 0.391725
```

```
# 2nd PC
## days
a_72_square <-(component_matrix[7,2]/sqrt(eigen_airpollution$values[2]))^2
a_72_square
```

```
## [1] 0.4596398
```

For the 3rd principal component, we have:

```
# 3rd PC
## temp
a_23_square <-(component_matrix[2,3]/sqrt(eigen_airpollution$values[3]))^2
a_23_square
```

```
## [1] 0.4585134
```

```
# 3rd PC
## precip
a_63_square <-(component_matrix[6,3]/sqrt(eigen_airpollution$values[3]))^2
a_63_square
```

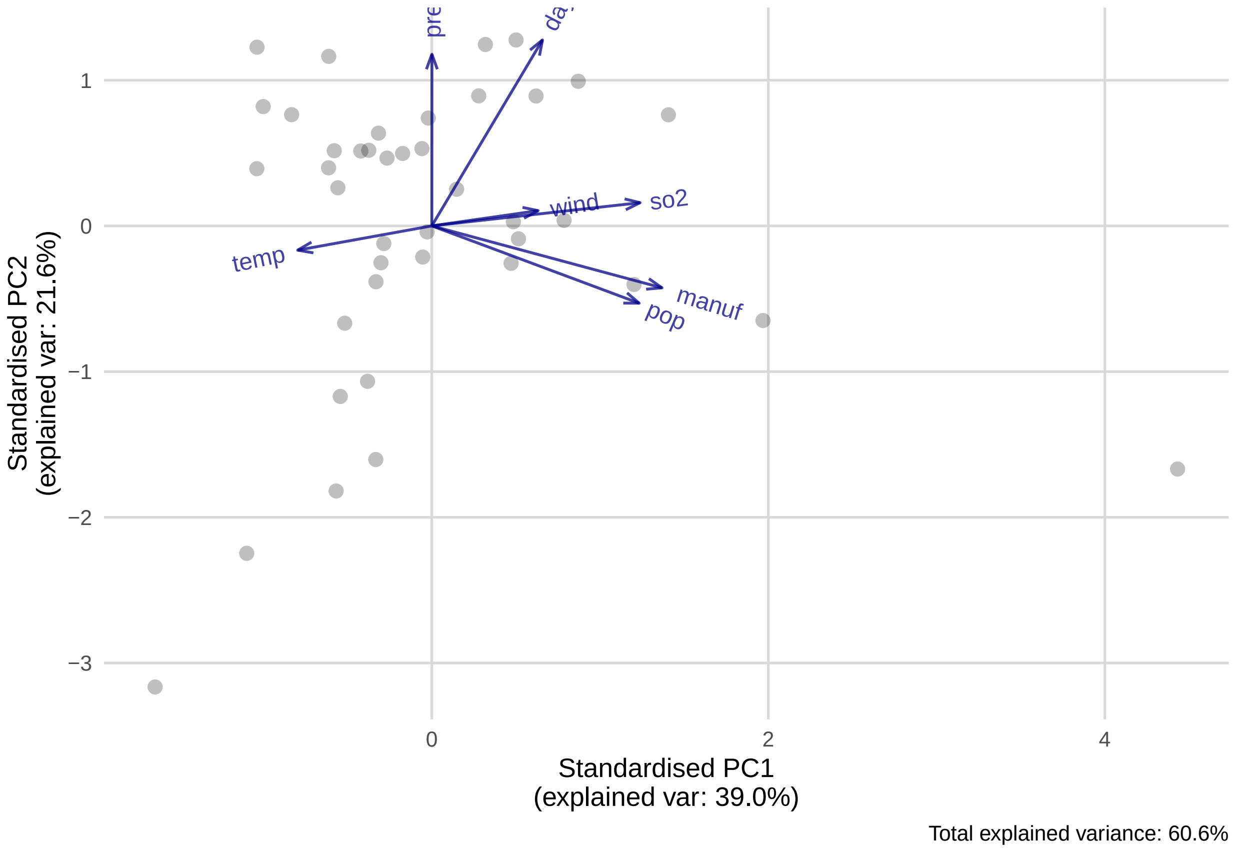
```
## [1] 0.2420997
```

Task 4. Make a graphical representation of the principal components and present relevant results.

Three biplots were performed to visualize the results from the PCA. A biplot graphical representation combines the principal component scores and the loadings in a single display.

The loadings correspond to the coefficients of the linear combination of initial variables from which the principal components were derived. Thus, the loadings give information concerning which variables provide the largest contribution to the retained principal components.

```r
# First biplot - PC1 vs PC2
library(AMR)
ggplot_pca(pca_airpollution, c(1,2))
```



The horizontal axis represents the correlation of the variables with the 1st Principal Component (PC). Considering the horizontal axis, it was possible to make the following conclusions:

- 1 variable has no correlation with the 1st PC (precip); therefore, this variable is not explained by the 1st PC;
- 3 variables have a similar high and positive correlation with the 1st PC (so2, manuf, pop); therefore, these variables are well explained by the 1st PC;
- 2 variables have similar medium and positive correlations with the first PC (days and wind); therefore, these variables are explained to a medium extent by the 1st PC;
- 1 variable has a medium and negative correlation with the 1st PC (temp); therefore, this variable is explained to a medium extent by the 1st PC;

The vertical axis represents the correlation of the variables with the 2nd Principal Component (PC). Considering the vertical axis, it was possible to make the following conclusions:

- 1 variable has low and negative correlation with the 2nd PC (temp); therefore, this variable is poorly explained by the 2nd PC;
- 2 variables have medium and negative correlation with the 2nd PC (manuf and pop); therefore, these variables are explained to a medium extent by the 2nd PC;
- 2 variables have low and positive correlation with the 2nd PC (wind and so2); therefore, these variables are poorly explained by the 2nd PC;
- 2 variables have high and positive correlation with the 2nd PC (precip and days); therefore, these variables are well explained by the 2nd PC.

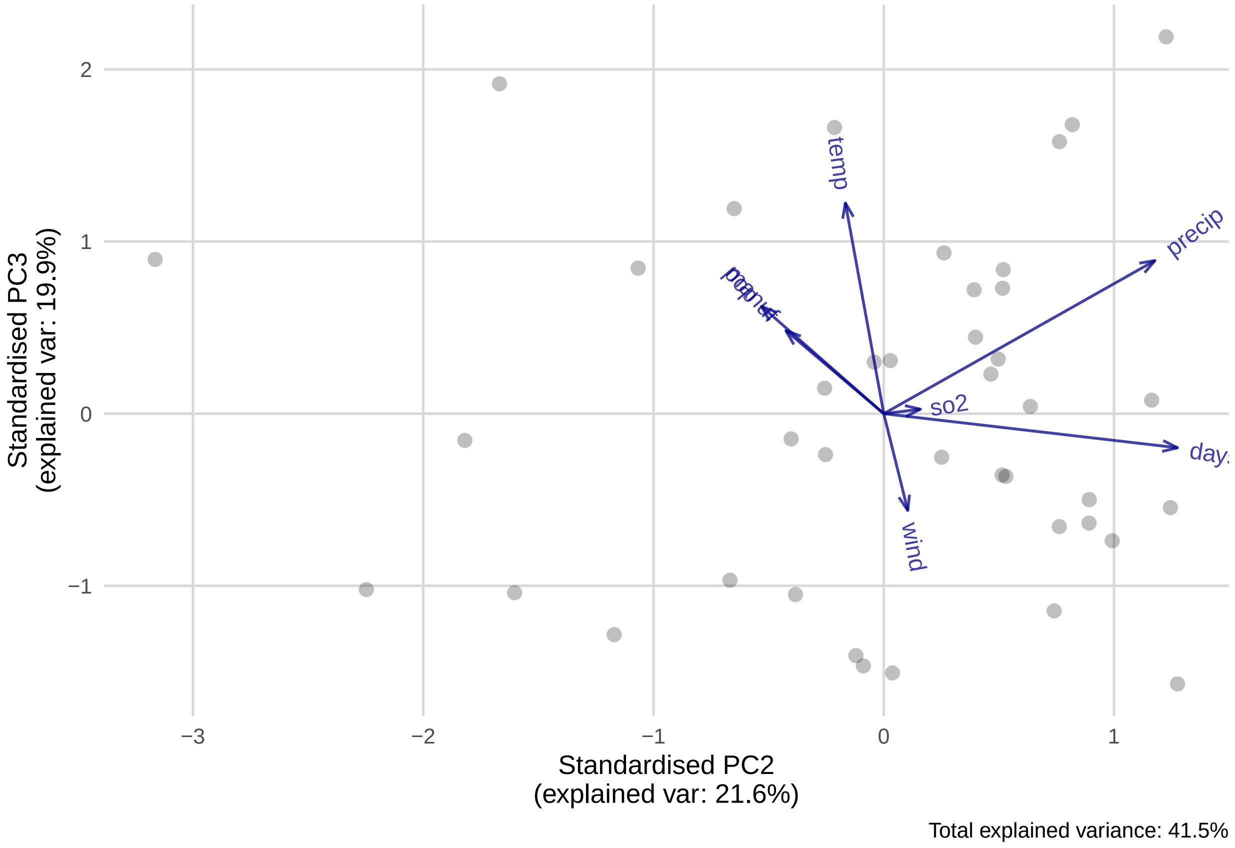
Furthermore, analysing the first biplot it was also possible to conclude that the vector associated with variable "precip" is pratically orthogonal to the vectors associated with variables "wind", "so2" and "temp". Therefore, the variable "precip" is not correlated with variables "wind", "so2" and "temp".

To summarize the data from the first biplot (PC1 vs PC2):

- variable "precip" is not explained by the 1st PC
- variables "temp", "wind" and "so2" are poorly explained by the 2nd PC
- variables "days", "manuf" and "pop" are explained by the 1st and 2nd PC; variables "manuf" and "pop" are explained more by the 1st PC than the 2nd PC, while variable "days" is explained more by the 2nd PC than the 1st PC
- variable "precip" is not correlated with variables "wind", "so2" and "temp"

Since we have 3 principal components retained, it was necessary to perform two additional biplots.

```
# Second biplot - PC2 vs PC3
library(AMR)
ggplot_pca(pca_airpollution, c(2,3))
```



The horizontal axis represents the correlation of the variables with the 2nd Principal Component (PC).

Considering the horizontal axis, it was possible to make the following conclusions:

- 2 variables have high and positive correlation with the 2nd PC (precip and days); therefore, these variables are well explained by the 2nd PC;
- 2 variables have low and positive correlation with the 2nd PC (so2 and wind); therefore, these variables are pooly explained by the 2nd PC;
- 1 variable has low and negative correlation with the 2nd PC (temp); therefore, this variable is poorly explained by the 2nd PC;
- 2 variables have medium and negative correlation with the 2nd PC (manuf and pop); therefore, these variables are explained to a medium extent by the 2nd PC;

The vertical axis represents the correlation of the variables with the 3rd Principal Component (PC). Considering the vertical axis, it was possible to make the following conclusions:

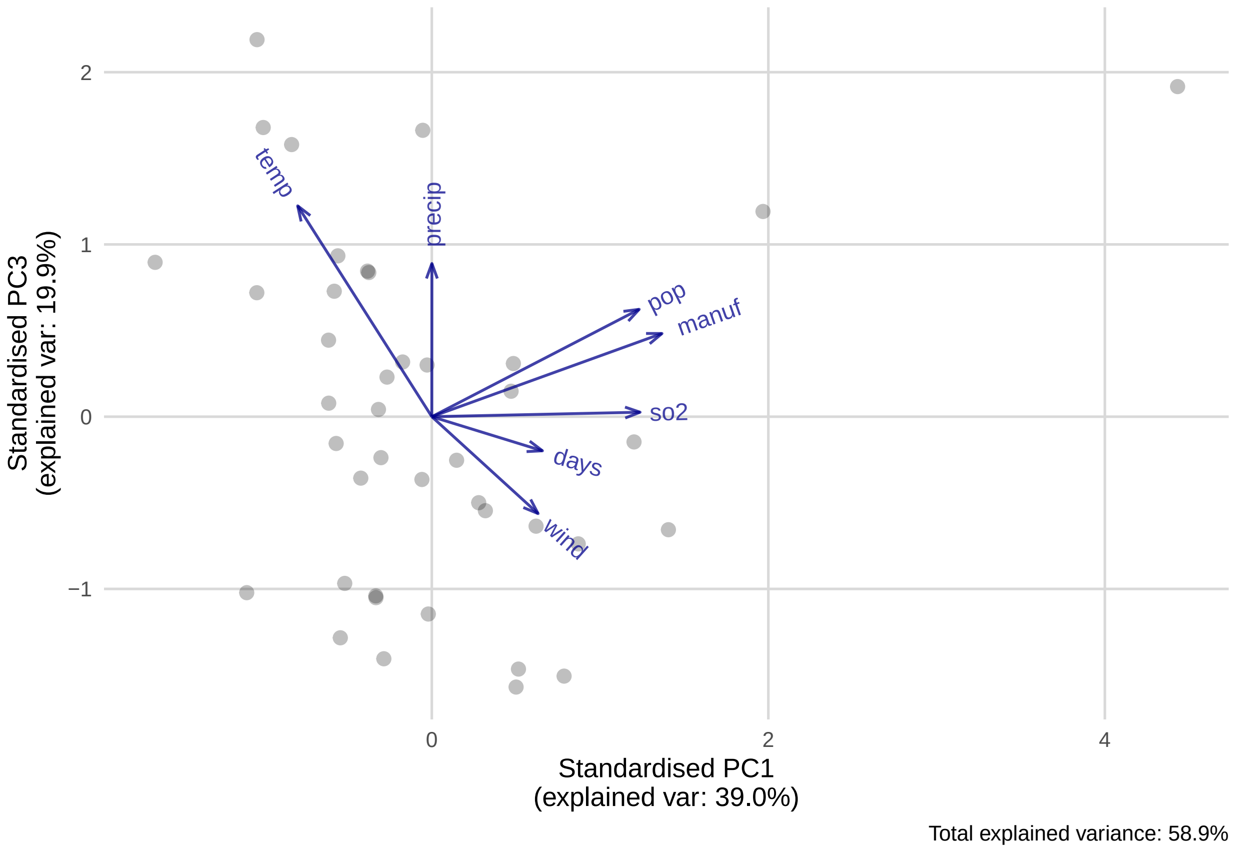
- 1 variable has no correlation with the 3rd PC (so2); this variable is therefore not explained by the 3rd PC;
- 2 variables have high and positive correlation with the 3rd PC (temp and precip); these variables are therefore well explained by the 3rd PC;
- 2 variables have medium and positive correlation with the 3rd PC (manuf and pop); these variables are therefore explained to a medium extent by the 3rd PC;
- 1 variable has negative and low correlation with the 3rd PC (days); this variable is therefore poorly explained by the 3rd PC
- 1 variable has medium and negative correlation with the 3rd PC (wind); this variable is therefore explained to a medium extent by the 3rd PC.

Furthermore, analysing the second biplot it was also possible to see that the vector associated with variable “wind” is pratically orthogonal to the vectors associated with variables “precip” and “so2”. Therefore, the variable “wind” is not correlated with variables “precip” and “so2”. In addition, the vector associated with variable “temp” is pratically orthogonal to the vectors associated with variables “precip” and “so2”. Therefore, the variable “temp” is not correlated with variables “precip” and “so2”.

To summarize the data from the second biplot (PC2 vs PC3):

- variable “so2” is not explained by the 3rd PC and is poorly explained by the 2nd PC
- variables “manuf” and “pop” are equally explained by the 2nd and 3rd PCs
- variable “precip” is equally explained by the 2nd and 3rd PCs
- variables “temp” and “wind” are poorly explained by the 2nd PC
- variable “temp” is well explained by the 3rd PC
- variable “wind” is explained to a medium extent by the 3rd PC
- variable “days” is well explained by the 2nd PC, but poorly by the 3rd PC
- variables “wind” and “temp” are not correlated with variables “precip” and “so2”

```r
# Third biplot - PC1 vs PC3
library(AMR)
ggplot_pca(pca_airpollution, c(1,3))
```



The horizontal axis represents the correlation of the variables with the 1st Principal Component (PC). Considering the horizontal axis, it was possible to make the following conclusions:

- 3 variables have high and positive correlation with the 1st PC (so2, pop and manuf); therefore, these variables are well explained by the 1st PC;
- 2 variables have medium and positive correlation with the 1st PC (wind and days); therefore, these variables are explained to a medium extent by the 1st PC;
- 1 variable has no correlation with the 1st PC (precip); therefore, this variable is not explained by the 1st PC;
- 1 variable has medium and negative correlation with the 1st PC (temp); therefore, this variable is explained to a medium extent by the 1st PC;

The vertical axis represents the correlation of the variables with the 3rd Principal Component (PC). Considering the vertical axis, it was possible to make the following conclusions:

- 2 variables have high and positive correlation with the 3rd PC (precip and temp); therefore, these variables are well explained by the 3rd PC;
- 2 variables have medium and positive correlation with the 3rd PC (pop and manuf); therefore, these variables are explained to a medium extent by the 3rd PC;
- 1 variable has no correlation with the 3rd PC (so2); therefore, this variable is not explained by the 3rd PC;
- 2 variables have medium and negative correlation with the 3rd PC (days and wind); therefore, these variables are explained to a medium extent by the 3rd PC;

Furthermore, analysing the third biplot it was also possible to see that the vector associated with variable “precip” is pratically orthogonal to the vector associated with variable “so2”. Therefore, the variables “precip” and “so2” are not correlated with each other.

In addition, the vector associated with variable “temp” is pratically orthogonal to the vectors associated with variables “pop” and “manuf”. Therefore, the variable “temp” is not correlated with variables “pop” and “manuf”.

To summarize the data from the third biplot (PC1 vs PC3):

- variable “precip” is not explained by the 1st PC, but by the 3rd PC
- variable “so2” is not explained by the 3rd PC
- variable “wind” is equally explained by the 1st and 3rd PCs
- variables “manuf” and “pop” are better explained by the 1st PC than the 3rd PC
- variable “temp” is better explained by the 3rd PC than the 1st PC
- variable “days” is better explained by the 1st PC than the 3rd PC
- variable “temp” is not correlated with variables “pop” and “manuf”
- variable “precip” is not correlated with variables “so2”

To summarize the conclusions obtained from all the biplots:

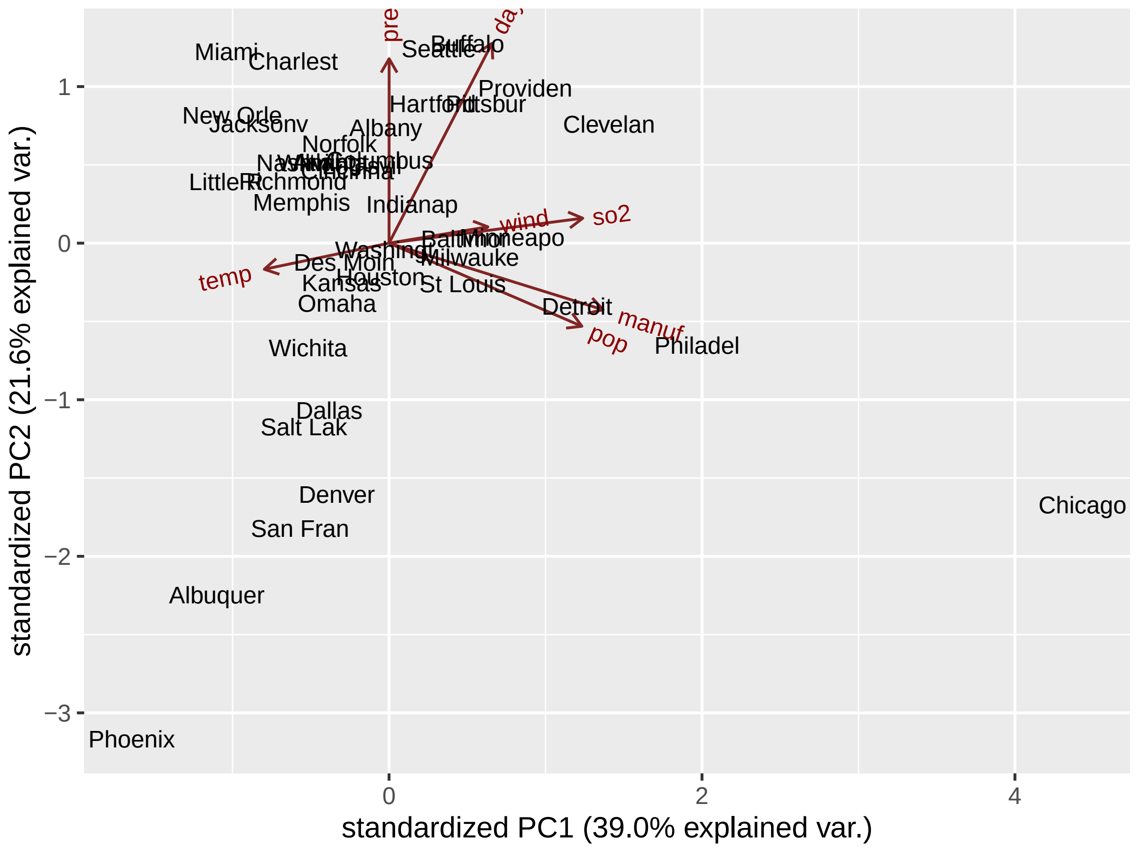
- variable “so2” is well explained by the 1st PC
- variables “manuf” and “pop” are explained by the three PCs, but more by the 1st PC than the 2nd or 3rd PCs
- variable “precip” is explained by the 2nd and 3rd PCs, but slightly more by the 2nd PC than the 3rd PC
- variable “temp” is explained more by the 3rd PC than the 1st or 2nd PCs
- variable “days” is explained more by the 2nd PC than the 1st or 3rd PCs
- variable “wind” is explained mostly by the 1st and 3rd PCs in equal extent

It is also possible to include the labels in the biplots.

```r
#library(devtools)
#install_github("vqv/ggbiplot")
require(ggbiplot)
```

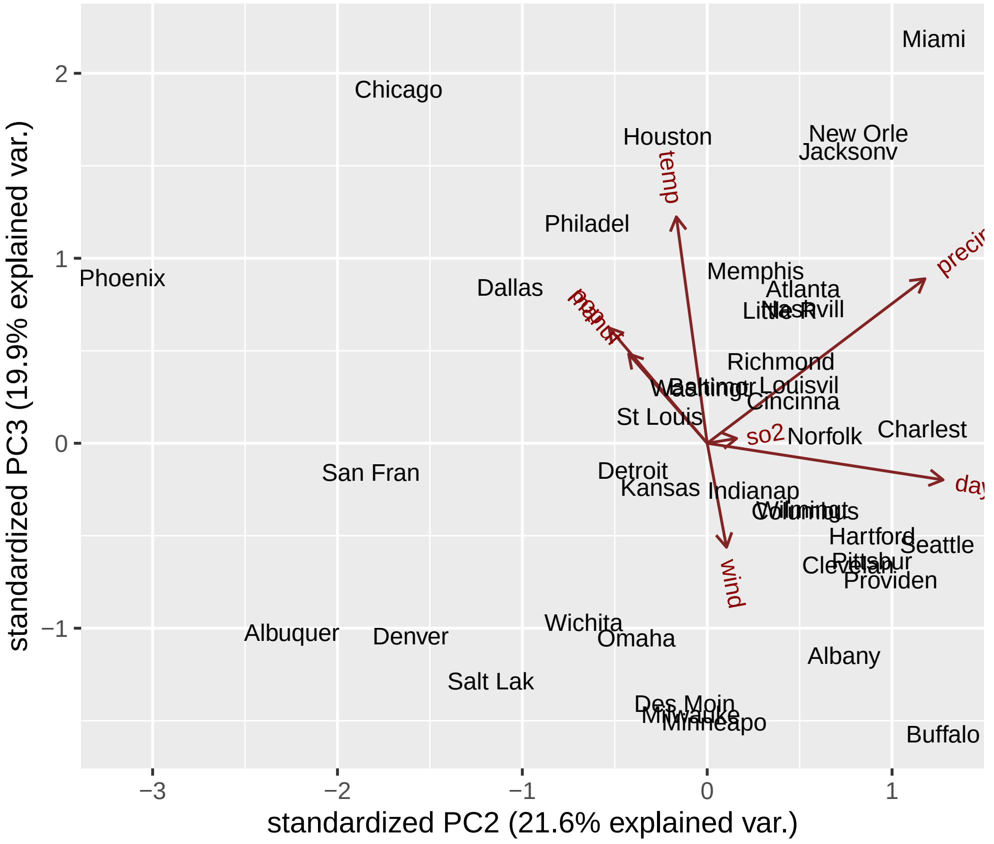
```
## Loading required package: ggbiplot

## Loading required package: plyr

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## Loading required package: scales

## Loading required package: grid
```

```r
# First biplot with labels - PC1 vs PC2
ggbiplot(pca_airpollution, c(1,2), labels = airpollution[,1])
```



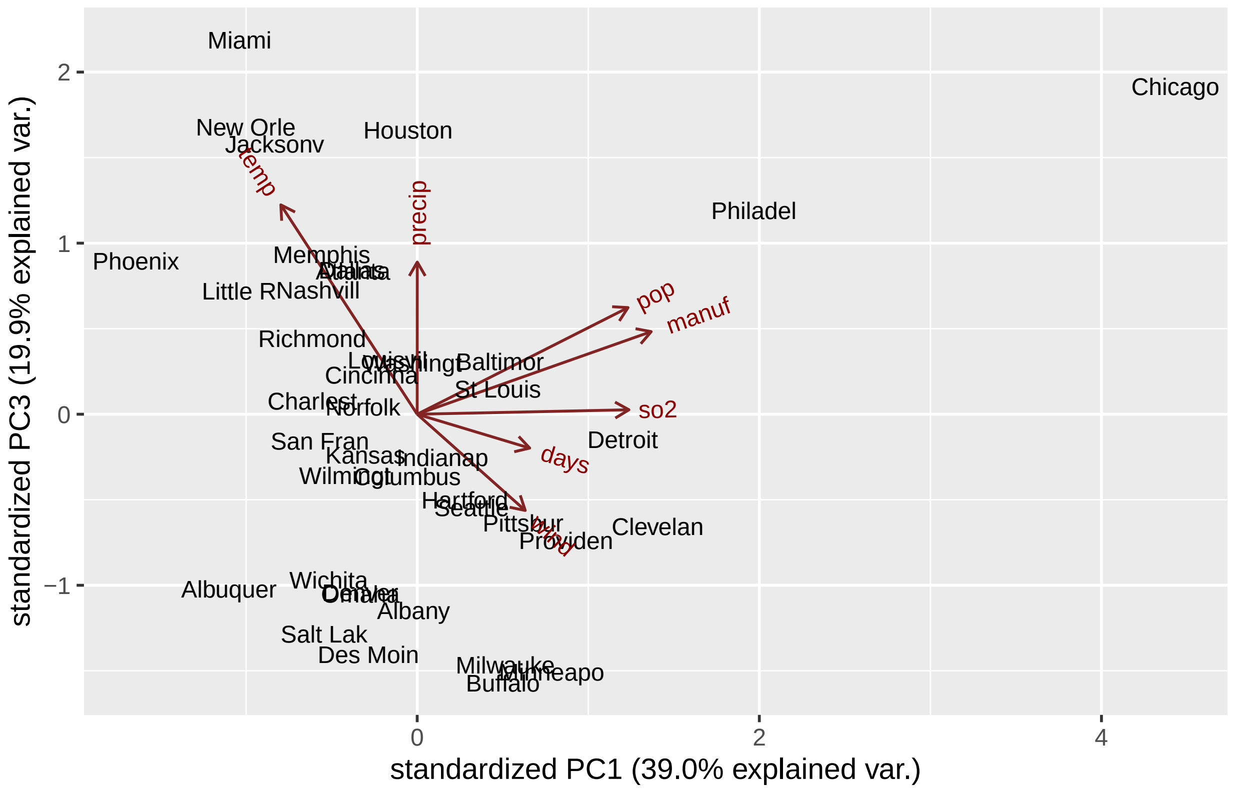
From the 1st biplot with labels, we can see that Chicago presents the highest values for "manuf", "pop" and "so2". We can also see that Phoenix displays the lowest values for "precip" and "days".

```
# Second biplot with labels - PC2 vs PC3
ggbiplot(pca_airpollution, c(2,3), labels = airpollution[,1])
```



From the 2nd biplot with labels, we can see again that Phoenix displays the lowest values for "precip" and "days".

```
# Third biplot with labels - PC1 vs PC3
ggbiplot(pca_airpollution, c(1,3), labels = airpollution[,1])
```



From the 3rd biplot with labels, we can see again that Chicago displays the highest values for "manuf", "pop" and "so2". We can also see that Miami exhibits the highest values for "temp".

Task 5. Perform a k-means clustering.

The Elbow point method, although not very accurate, is widely used to choose the number of clusters (K) in a dataset.

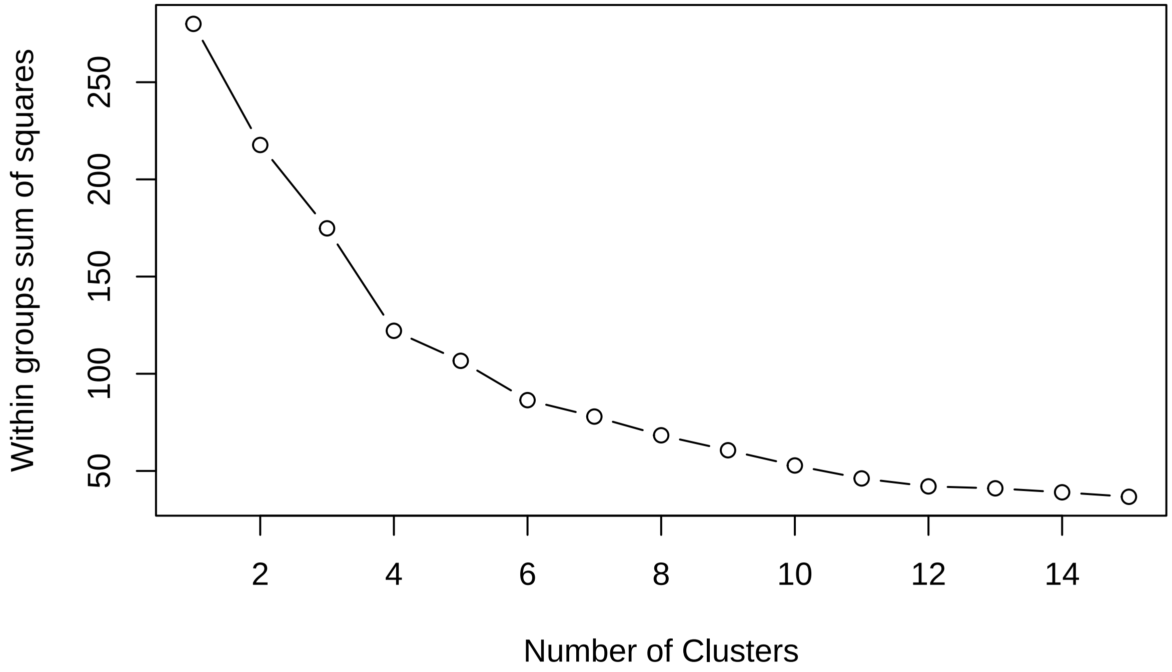
The Elbow point method runs a k-means clustering on the dataset for a predefined range of values for K and then for each value of K computes the "within groups sum of squares", which corresponds to the sum of the squared distance between each point and the centroid in a cluster.

The Elbow point method plots a reduction in the "within groups sum of squares" versus the number of clusters (K). The best value for K, called the Elbow point, is determined visually and corresponds to the number for K after which the slope (rate of reduction) isn't very steep.

Since we had seen before that the measure units are not the same for all variables and the mean and standard deviation are also quite different, it is necessary to standardize the data before performing a k-means clustering.

```r
# Scaling the data
airpollution_s <- scale(airpollution_variables)
```

```r
wssplot <- function(data, nc=15, seed=1234){
wss <- (nrow(data)-1)*sum(apply(data,2,var))
for (i in 2:nc){
set.seed(seed)
wss[i] <- sum(kmeans(data, centers=i)$withinss)}
plot(1:nc, wss, type="b", xlab="Number of Clusters",
ylab="Within groups sum of squares")
wss
}

wssplot(airpollution_s,nc=15,seed = 1234)
```



```
##  [1] 280.00000 217.69973 174.81976 122.09718 106.65860  86.42485  77.96379
##  [8]  68.33546  60.64556  52.80457  46.13515  42.04358  41.05227  39.00970
## [15]  36.70433
```

The plot shows a sharp edge at K = 4, indicating that the optimal number of clusters for our dataset are 4.

However, we also want to compare the k-means clustering to the PCA performed previously. Therefore, we decided to set K = 3.

```
set.seed(1234)
k3 <- kmeans(airpollution_s, centers = 3, nstart = 25)
k3
```

```
## K-means clustering with 3 clusters of sizes 2, 9, 30
##
## Cluster means:
##          so2         temp      manuf        pop       wind      precip
## 1  2.5328276 -0.437678327  3.6468455  3.5414335  0.3892485  0.03533737
## 2 -0.6837336  0.124970604 -0.3576776 -0.3022799  0.1559270 -1.31929404
## 3  0.0362649 -0.008312626 -0.1358198 -0.1454116 -0.0727280  0.39343239
##         days
## 1  0.1734509
## 2 -1.3754402
## 3  0.4010687
##
## Clustering vector:
##  [1] 2 3 2 2 3 3 3 3 3 3 1 3 2 2 3 3 3 3 3 3 3 2 2 3 3 3 3 3 1 3 3 3 3 2 3 2 3 3
## [39] 3 3 3
##
## Within cluster sum of squares by cluster:
## [1]   9.278848  38.079860 117.290353
##  (between_SS / total_SS =  41.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

We have three clusters with sizes 2, 9 and 30.

By assigning the samples to 3 clusters rather than n (number of samples) clusters, we achieved a reduction in the sum of squares of 41.2%, which is low.

```
k3$centers
```

```
##          so2         temp      manuf        pop       wind      precip
## 1  2.5328276 -0.437678327  3.6468455  3.5414335  0.3892485  0.03533737
## 2 -0.6837336  0.124970604 -0.3576776 -0.3022799  0.1559270 -1.31929404
## 3  0.0362649 -0.008312626 -0.1358198 -0.1454116 -0.0727280  0.39343239
##         days
## 1  0.1734509
## 2 -1.3754402
## 3  0.4010687
```

Cluster 1 has a mean value of so2 = 2.533, temp = -0.438, manuf = 3.647, pop = 3.541, wind = 0.389, precip = 0.035 and days = 0.173.

Cluster 2 has a mean value of so2 = -0.684, temp = 0.125, manuf = -0.358, pop = -0.302, wind = 0.160, precip = -1.319 and days = -1.375.

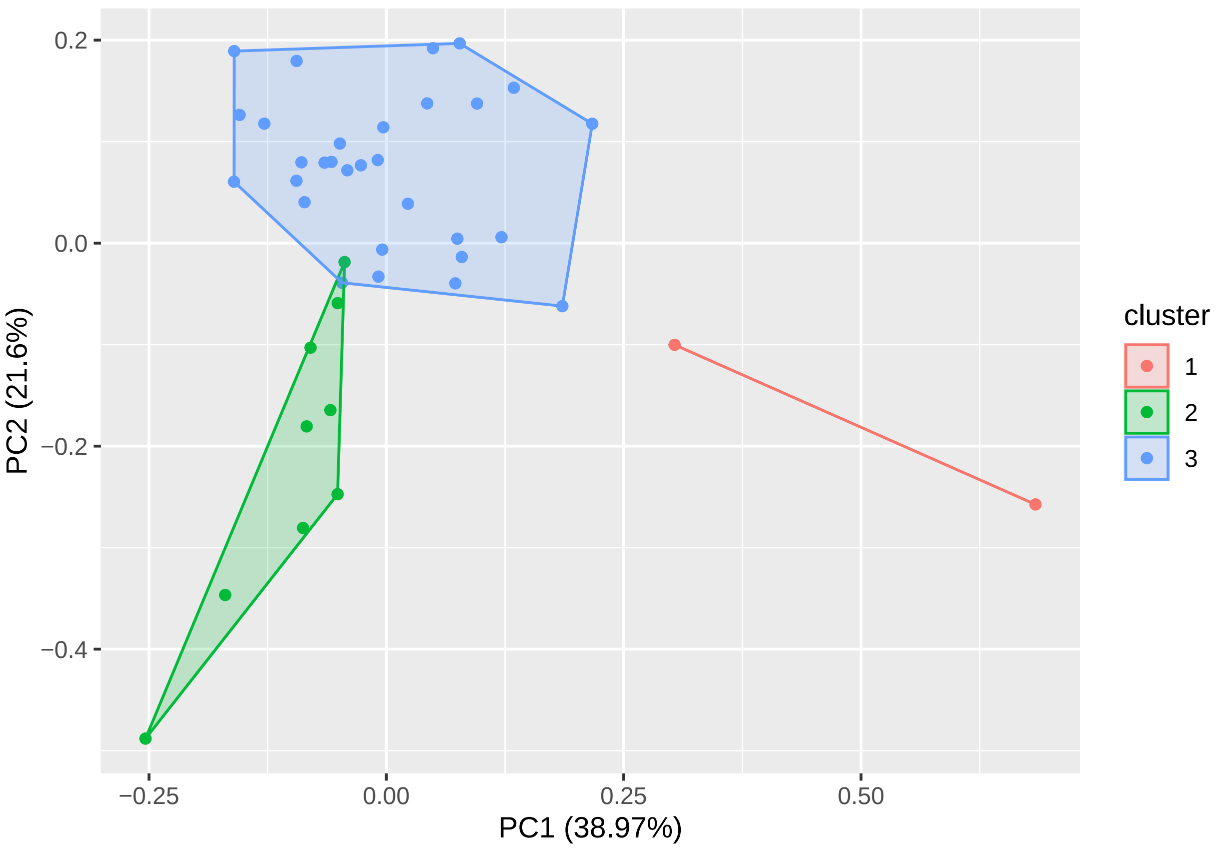
Cluster 3 has a mean value of so2 = 0.036, temp = -0.008, manuf = -0.136, pop = -0.145, wind = -0.073, precip = 0.393 and days = 0.401.

Task 6. Make a graphical representation of the clusterings obtained.

```
library("ggplot2")
library("dplyr")
library("ggfortify")
```

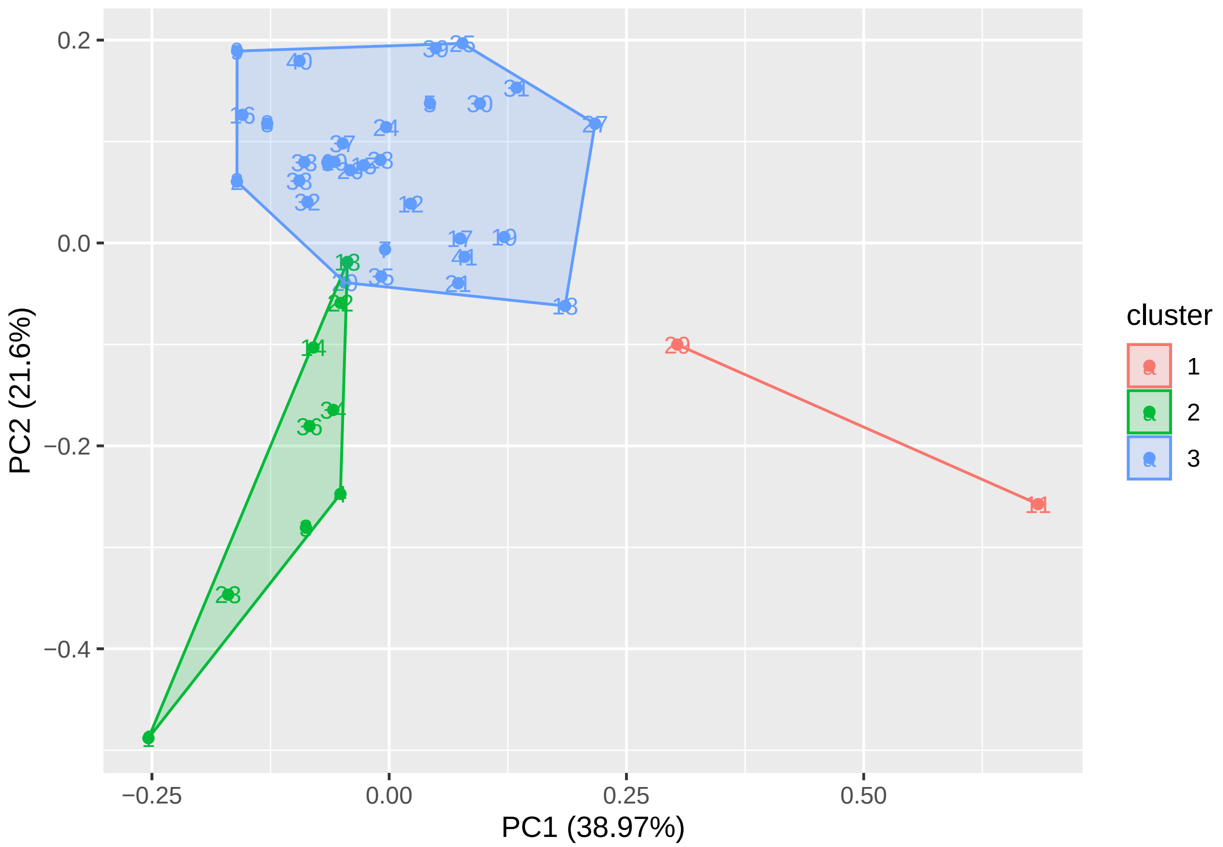
```
##
## Attaching package: 'ggfortify'

## The following object is masked from 'package:ggbiplot':
##
##     ggbiplot
```

```
ggplot2::autoplot(k3, airpollution_s, frame = TRUE)
```



We can also had the labels to the previous graphical representation.

```
library("ggplot2")
library("dplyr")
library("ggfortify")
ggplot2::autoplot(k3, airpollution_s, frame = TRUE, label = TRUE, label.size = 3)
```

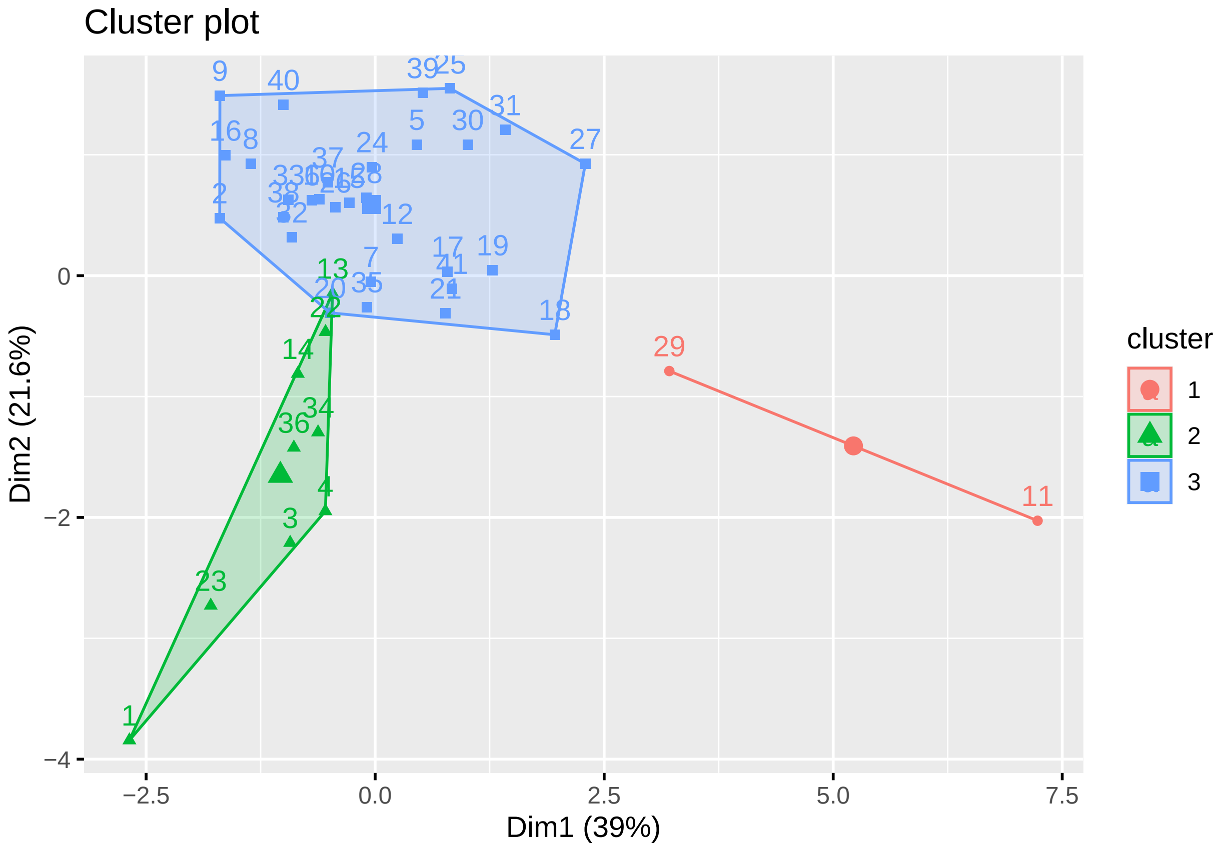


We can also visualize the 3 clusters using fviz_cluster, which displays the clusters using the principal components retained (2 by 2) to define the x and y coordinates of each observation.

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

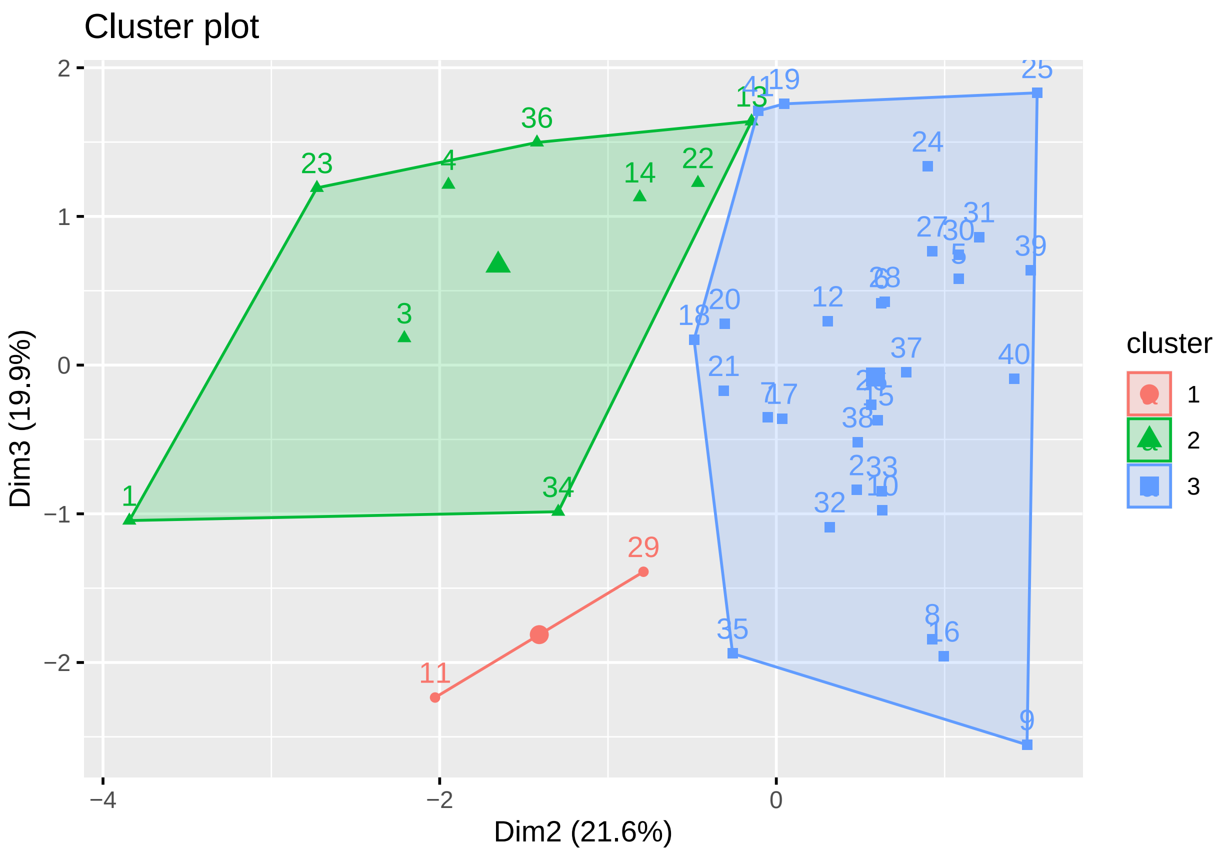
```
# For PC1 and PC2
fviz_cluster(k3, data = airpollution_s, axes = c(1, 2))
```



From the previous graphical representation, we can see that cluster 1 (in red) has a high and positive correlation with the 1st PC. We can also see that cluster 2 (in green) has a medium and negative correlation with the 1st PC.

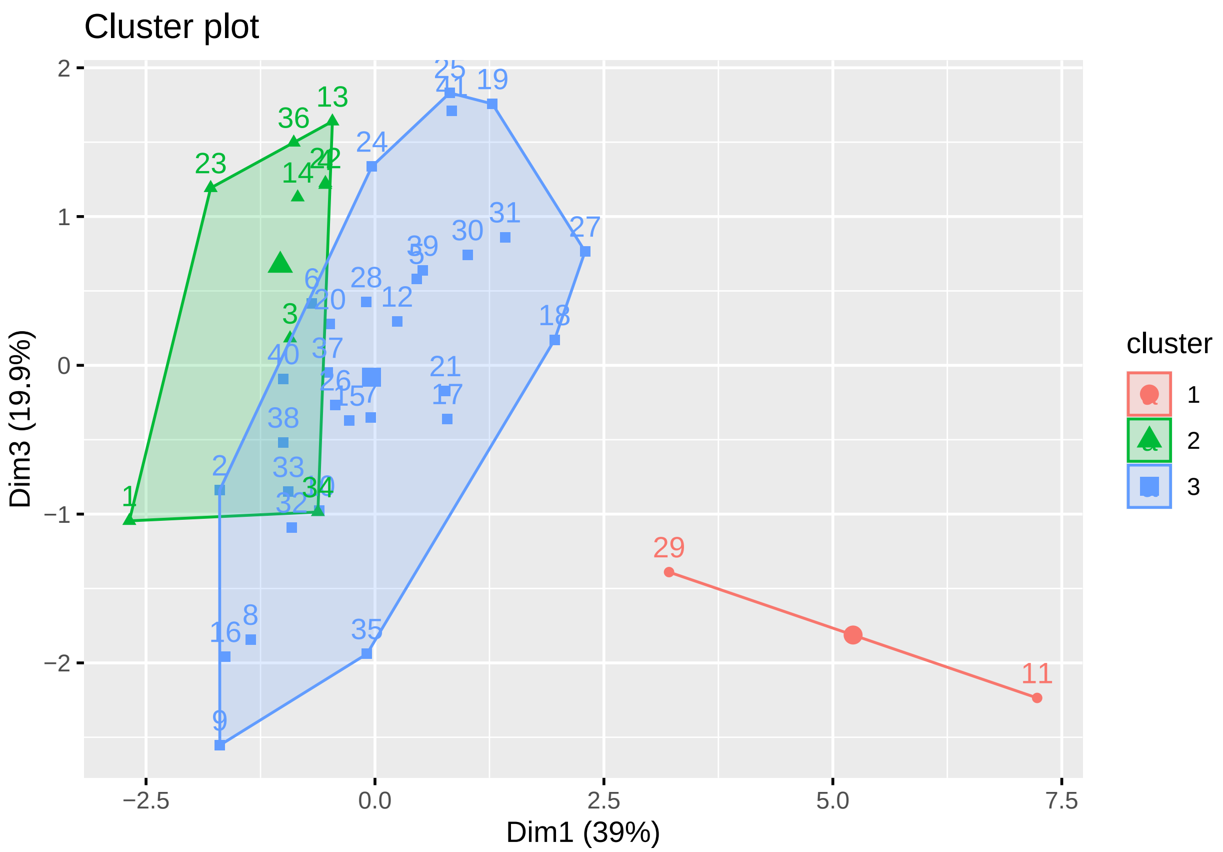
Furthermore, we can see that cluster 2 (in green) has a negative and medium correlation with the 2nd PC, while cluster 3 (in blue) has a medium and positive correlation with the 2nd PC.

```
library(factoextra)
# For PC2 and PC3
fviz_cluster(k3, data = airpollution_s, axes = c(2, 3))
```



From the previous graphical representation, we can see that cluster 1 (in red) has a medium and negative correlation with the 2nd and 3rd PCs. Furthermore, we can see that cluster 2 (in green) has a negative and medium correlation with the 2nd PC.

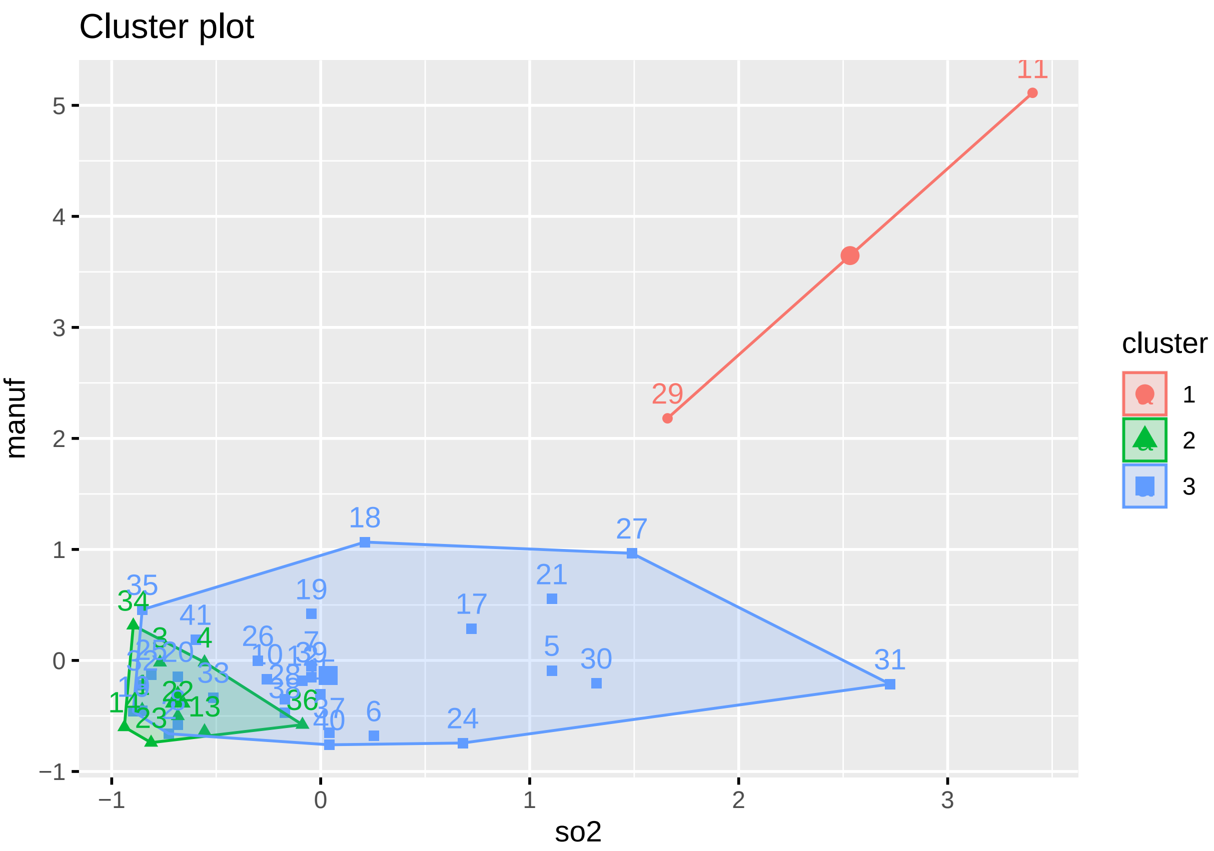
```
library(factoextra)
# For PC1 and PC3
fviz_cluster(k3, data = airpollution_s, axes = c(1, 3))
```



From the previous graphical representation, we can see that cluster 1 (in red) has a high and positive correlation with the 1st PC. Furthermore, we can see that cluster 2 (in green) has a medium and negative correlation with the 1st PC.

Additionally, we can display the clusters using two of the original variables (for instance, “so2” and “manuf”) to define the x and y coordinates of each observation.

```
library(factoextra)
# For variables so2 and manuf
fviz_cluster(k3, data = airpollution_s, choose.vars = c(1, 3))
```



We can conclude from the previous graphical representation that observations corresponding to rows 11 (Chicago) and 29 (Philadelphia) that belong to cluster 1 have the highest values for the variable “manuf”.

Task 7. Write a brief description of each cluster.

The first cluster, which is colored in red in the graphical representations, has a size equal to 2. Cluster 1 has the highest mean values for variables “so2”, “manuf”, “pop” and “wind”. This cluster also has the lowest values for variable “temp”.

The second cluster, which is colored in green in the graphical representations, has a size equal to 9. Cluster 2 has the highest mean value for the variable “temp”. This cluster also has the lowest mean values for variables “so2”, “manuf”, “pop”, “precip” and “days”.

The third cluster, which is colored in blue in the graphical representations, has a size equal to 30. Cluster 3 has the highest mean values for variables “precip” and “days”. It also has the lowest mean values for the variable “wind”.

From the k-means clustering that we performed, we can conclude that observations with higher values for variables “so2”, “manuf” and “pop” and lower values for variable “temp” are grouped together. On the contrary, observations with lower values for variables “so2”, “manuf” and “pop” and higher values for variable “temp” are grouped together.

We can also see that observations with higher values for variables “precip” and “days” and lower values for variable “wind” are grouped together. On the contrary, observations with lower values for variables “precip” and “days” and higher values for variable “wind” are grouped together.

If we compare the results obtained from the PCA with the k-means clustering, we can see that:

- in PCA, variables “so2”, “manuf” and “pop” are well explained by the 1st PC, while in k-means clustering we can see that observations from cluster 1 are also grouped together on the basis of their higher values for variables “so2”, “manuf” and “pop”;

- in PCA, variables “precip” and “days” are more explained by the 2nd PC, while in k-means clustering we can see that observations from cluster 3 are also grouped together on the basis of their higher values for variables “precip” and “days”;
- in PCA, variable “temp” is more explained by the 3rd PC, while in k-means clustering we can see that observations from cluster 2 are also grouped together on the basis of their higher values for variable “temp”.