Activity A - PCA and Clustering on Air Pollution Data

Daniel Carvalho nº 64350, Fatma Özel nº 57037, Helton Mendonça nº 56870, Rita Silva nº 56798

2024-11-28

Introduction

In this work we looked at data on pollution in cities in the United States of America, and explored which variables contribute most to understanding it.

Data

Our dataset called airpollution.csv includes the following variables: city, so2, temp, manuf, pop, wind, precip and days.

Descriptive Analysis

Reading the Data

```
airpollution <- read.csv("/Users/fatmabetulozel/1semester/FCD/activities/activity\ 3/A3\ -\ Statistics-
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
```

Types of Variables

```
str(airpollution)
```

```
## 'data.frame':    41 obs. of  8 variables:
##  $ city  : chr  "Phoenix" "Little R" "San Fran" "Denver" ...
##  $ 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 ...
```

Unique Names for Cities

```r
airpollution[,1]
```

```
##  [1] "Phoenix"  "Little R" "San Fran" "Denver"   "Hartford" "Wilmingt"
##  [7] "Washingt" "Jacksonv" "Miami"    "Atlanta"  "Chicago"  "Indianap"
## [13] "Des Moin" "Wichita"  "Louisvil" "New Orle" "Baltimor" "Detroit"
## [19] "Minneapo" "Kansas"   "St Louis" "Omaha"    "Albuquer" "Albany"
## [25] "Buffalo"  "Cincinna" "Clevelan" "Columbus" "Philadel" "Pittsbur"
## [31] "Providen" "Memphis"  "Nashvill" "Dallas"   "Houston"  "Salt Lak"
## [37] "Norfolk"  "Richmond" "Seattle"  "Charlest" "Milwauke"
```

PCA Preparation

We considered the variables temp, manuf, pop, wind, precip and days for the PCA analysis, as shown below.

```r
airpollution_variables <- airpollution[3:8]
rownames(airpollution_variables)<- airpollution[,1]
airpollution_variables
```

```
##          temp manuf  pop wind precip days
## Phoenix  70.3   213  582  6.0   7.05   36
## Little R 61.0    91  132  8.2  48.52  100
## San Fran 56.7   453  716  8.7  20.66   67
## Denver   51.9   454  515  9.0  12.95   86
## Hartford 49.1   412  158  9.0  43.37  127
## Wilmingt 54.0    80   80  9.0  40.25  114
## Washingt 57.3   434  757  9.3  38.89  111
## Jacksonv 68.4   136  529  8.8  54.47  116
## Miami    75.5   207  335  9.0  59.80  128
## Atlanta  61.5   368  497  9.1  48.34  115
## Chicago  50.6  3344 3369 10.4  34.44  122
## Indianap 52.3   361  746  9.7  38.74  121
## Des Moin 49.0   104  201 11.2  30.85  103
## Wichita  56.6   125  277 12.7  30.58   82
## Louisvil 55.6   291  593  8.3  43.11  123
## New Orle 68.3   204  361  8.4  56.77  113
## Baltimor 55.0   625  905  9.6  41.31  111
## Detroit  49.9  1064 1513 10.1  30.96  129
## Minneapo 43.5   699  744 10.6  25.94  137
## Kansas   54.5   381  507 10.0  37.00   99
## St Louis 55.9   775  622  9.5  35.89  105
## Omaha    51.5   181  347 10.9  30.18   98
```

```
## Albuquer 56.8    46  244  8.9   7.77   58
## Albany   47.6    44  116  8.8  33.36  135
## Buffalo  47.1   391  463 12.4  36.11  166
## Cincinna 54.0   462  453  7.1  39.04  132
## Clevelan 49.7  1007  751 10.9  34.99  155
## Columbus 51.5   266  540  8.6  37.01  134
## Philadel 54.6  1692 1950  9.6  39.93  115
## Pittsbur 50.4   347  520  9.4  36.22  147
## Providen 50.0   343  179 10.6  42.75  125
## Memphis  61.6   337  624  9.2  49.10  105
## Nashvill 59.4   275  448  7.9  46.00  119
## Dallas   66.2   641  844 10.9  35.94   78
## Houston  68.9   721 1233 10.8  48.19  103
## Salt Lak 51.0   137  176  8.7  15.17   89
## Norfolk  59.3    96  308 10.6  44.68  116
## Richmond 57.8   197  299  7.6  42.59  115
## Seattle  51.1   379  531  9.4  38.79  164
## Charlest 55.2    35   71  6.5  40.75  148
## Milwauke 45.7   569  717 11.8  29.07  123
```

```r
str(airpollution_variables)
```

```
## 'data.frame':    41 obs. of  6 variables:
##  $ 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 ...
```

```r
dim(airpollution_variables)
```

```
## [1] 41  6
```

Localization Measures

```r
summary(airpollution_variables)
```

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

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

For the PCA we will use the correlation matrix, since the measure units for each variable are different and also taking into account that the standard deviation and mean are different.

Principal Component Analysis

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

## 1st) Determination of the correlation matrix

cor_airpollution <- cor(airpollution_variables)
cor_airpollution
```

```
##               temp       manuf         pop        wind      precip        days
## temp    1.00000000 -0.19004216 -0.06267813 -0.34973963  0.38625342 -0.43024212
## manuf  -0.19004216  1.00000000  0.95526935  0.23794683 -0.03241688  0.13182930
## pop    -0.06267813  0.95526935  1.00000000  0.21264375 -0.02611873  0.04208319
## wind   -0.34973963  0.23794683  0.21264375  1.00000000 -0.01299438  0.16410559
## precip  0.38625342 -0.03241688 -0.02611873 -0.01299438  1.00000000  0.49609671
## days   -0.43024212  0.13182930  0.04208319  0.16410559  0.49609671  1.00000000
```

```
## 2nd) Obtaining Eigenvalues and Eigenvectors

eigen_airpollution <- eigen(cor_airpollution)
eigen_airpollution
```

```
## eigen() decomposition
## $values
## [1] 2.19616264 1.49994343 1.39464912 0.76022689 0.11457065 0.03444727
##
## $vectors
##             [,1]       [,2]        [,3]        [,4]        [,5]        [,6]
## [1,]  0.32964613 -0.1275974  0.67168611 -0.30645728 -0.55805638 -0.13618780
## [2,] -0.61154243 -0.1680577  0.27288633  0.13684076  0.10204211 -0.70297051
## [3,] -0.57782195 -0.2224533  0.35037413  0.07248126 -0.07806551  0.69464131
## [4,] -0.35383877  0.1307915 -0.29725334 -0.86942583 -0.11326688 -0.02452501
## [5,]  0.04080701  0.6228578  0.50456294 -0.17114826  0.56818342  0.06062222
## [6,] -0.23791593  0.7077653 -0.09308852  0.31130693 -0.58000387 -0.02196062
```

According to Kaiser's criteria we need to retain only the principal components which correspond to eigenvalues greater than 1. So, we retain the first three principal components.

Performing 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     1.4819456 1.2247218 1.1809526 0.8719099 0.33848287
## Proportion of Variance 0.3660271 0.2499906 0.2324415 0.1267045 0.01909511
## Cumulative Proportion  0.3660271 0.6160177 0.8484592 0.9751637 0.99425879
##                             Comp.6
## Standard deviation     0.185599752
## Proportion of Variance 0.005741211
## Cumulative Proportion  1.000000000
##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## temp    0.330  0.128  0.672  0.306  0.558  0.136
## manuf  -0.612  0.168  0.273 -0.137 -0.102  0.703
## pop    -0.578  0.222  0.350               -0.695
## wind   -0.354 -0.131 -0.297  0.869  0.113
## precip        -0.623  0.505  0.171 -0.568
## days   -0.238 -0.708        -0.311  0.580
```

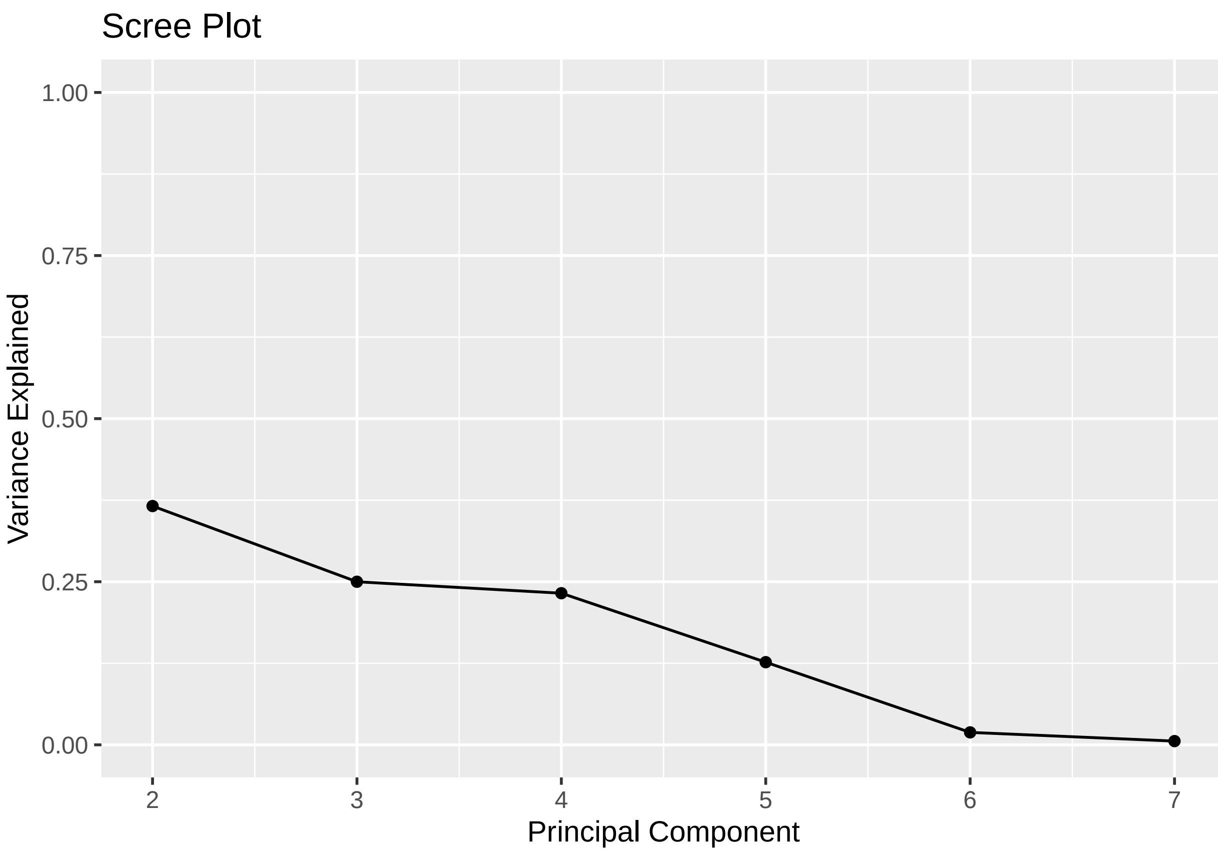
With three principal components we have 85% (0.848) of variance explained.

The first principal component explain 37% (0.366) of the variance. The second principal component explain 25% (0.249) of the variance. The third principal component explain 23% (0.232) of the variance.

```
#Calculating total variance explained by each principal component
var_explained_airpollution = pca_airpollution$sdev^2 / sum(pca_airpollution$sdev^2)

library(ggplot2)

qplot(c(2:7), var_explained_airpollution) +
```

```r
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot") +
  ylim(0,1)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Based in all the methodologies, we should consider first three principal components.

Next we will identify the variables that contribute more for the explanation of each principal component retained.

Contribution of variables for the explanation of each principal component retained

We will use the following formula: $|l_{ij}| \geq \sqrt{\frac{\lambda_j}{p}}$

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

```
##             Comp.1     Comp.2     Comp.3      Comp.4      Comp.5       Comp.6
## temp    0.48851762  0.1562713  0.7932295  0.26720314  0.18889252  0.025276422
## manuf  -0.90627258  0.2058239  0.3222658 -0.11931281 -0.03453951  0.130471152
## pop    -0.85630067  0.2724433  0.4137753 -0.06319713  0.02642384 -0.128925256
## wind   -0.52436980 -0.1601832 -0.3510421  0.75806100  0.03833890  0.004551836
## precip  0.06047377 -0.7628275  0.5958649  0.14922586 -0.19232035 -0.011251469
## days   -0.35257846 -0.8668156 -0.1099331 -0.27143159  0.19632137  0.004075886
```

```r
sqrt(eigen_airpollution$values[1]/6)
```

```
## [1] 0.6050017
```

Variables that must be used in the interpretation of the first principal component: manuf and pop.

```r
sqrt(eigen_airpollution$values[2]/6)
```

```
## [1] 0.4999906
```

Variables that must be used in the interpretation of the second principal component: precip and days.

```r
sqrt(eigen_airpollution$values[3]/6)
```

```
## [1] 0.4821219
```

Variables that must be used in the interpretation of the third principal component: temp and precip.

Importance of the variables for the explanation of each of the principal components retained

We will use the following formula: $a_{ij}^2 = (\frac{l_{ij}}{\sqrt(\lambda_j)})^2$

```r
#1st PC
## manuf
a_21_square <- (component_matrix[2,1]/sqrt(eigen_airpollution$values[1]))^2
a_21_square
```

```
## [1] 0.3739841
```

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

```
## [1] 0.3338782
```

The variable that contributes most to explaining the first principal component is manuf.

```r
#2nd PC
## precip
a_52_square <- (component_matrix[5,2]/sqrt(eigen_airpollution$values[2]))^2
a_52_square
```

```
## [1] 0.3879519
```

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

```
## [1] 0.5009318
```

The variable that contributes most to explaining the second principal component is days.

```r
#3rd PC
## temp
a_13_square <- (component_matrix[1,3]/sqrt(eigen_airpollution$values[3]))^2
a_13_square
```

```
## [1] 0.4511622
```

```r
#3rd PC
## precip
a_53_square <- (component_matrix[5,3]/sqrt(eigen_airpollution$values[3]))^2
a_53_square
```

```
## [1] 0.2545838
```

The variable that contributes most to explaining the third principal component is temp.

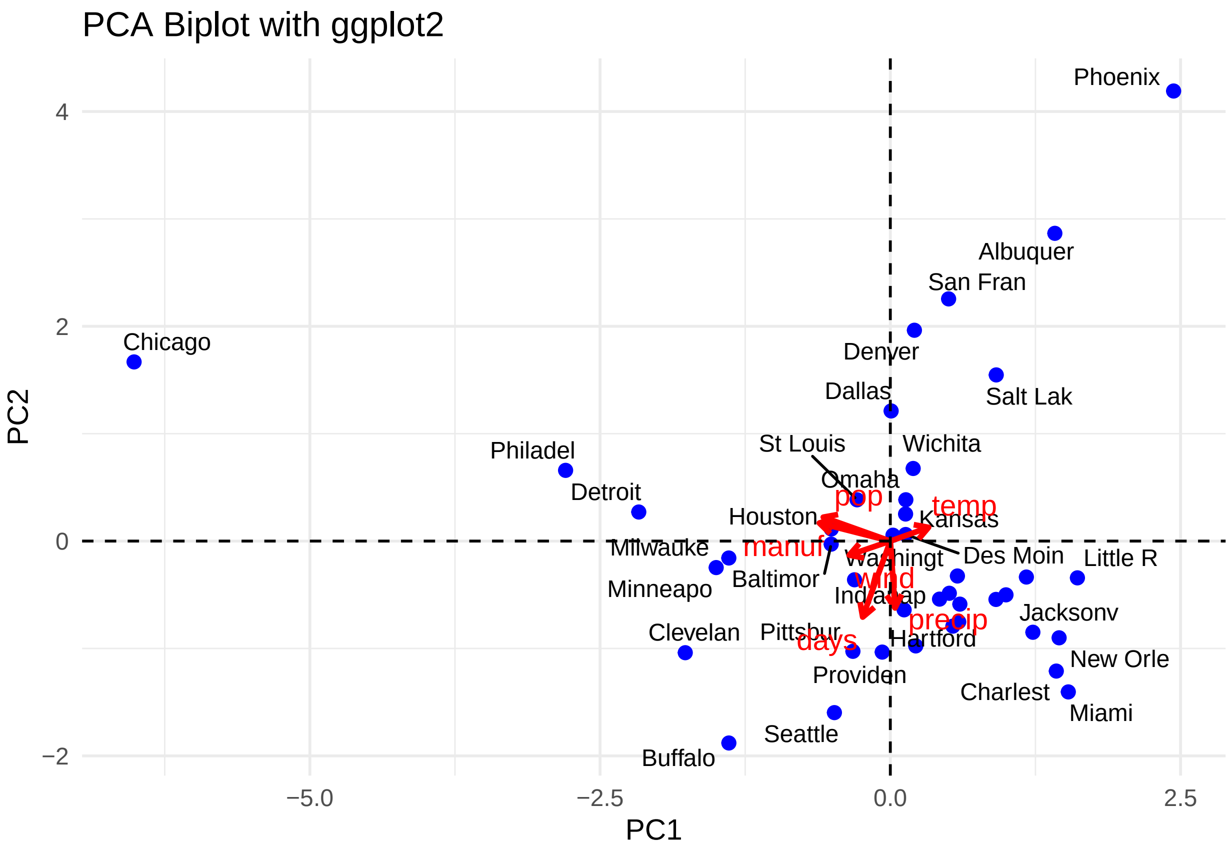
Graphical representation of the principal components

```r
# Extracting scores (PC coordinates for samples) and loadings (contributions of variables)
scores <- as.data.frame(pca_airpollution$scores)  # Principal component scores
loadings <- as.data.frame(pca_airpollution$loadings[, 1:2])  # Loadings for the first two PCs
```

```r
# Renaming the columns for clarity
colnames(scores) <- c("PC1", "PC2")
rownames(scores) <- airpollution$city
loadings$Variables <- rownames(loadings)
colnames(loadings) <- c("PC1", "PC2", "Variables")
```

```r
library(ggplot2)
library(ggrepel)  # For better label placement
library(dplyr)

# Creating a ggplot2 biplot
ggplot() +
  # Plot the points for observations (scores)
  geom_point(data = scores, aes(x = PC1, y = PC2), color = "blue", size = 2) +

  # Add text labels for observations (optional)
  geom_text_repel(data = scores, aes(x = PC1, y = PC2, label = rownames(scores)), size = 3) +

  # Plot the loadings as arrows
  geom_segment(data = loadings, aes(x = 0, y = 0, xend = PC1, yend = PC2),
               arrow = arrow(length = unit(0.2, "cm")), color = "red", size = 1) +

  # Add text labels for variables
  geom_text_repel(data = loadings, aes(x = PC1, y = PC2, label = Variables),
                  color = "red", size = 4) +

  # Add title and labels
  labs(title = "PCA Biplot with ggplot2", x = "PC1", y = "PC2") +

# Add horizontal and vertical axes
  geom_hline(yintercept = 0, linetype = "dashed", color = "black") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "black") +


  # Improving the theme
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## Warning: ggrepel: 10 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



We can see that the city of Chicago is an outlier, as it is quite out of line with the other values. We'll see that this is visible when it comes to clustering. As expected from the previous analysis: manuf and pop are highly correlated and are the variables that best explain the 1st PC. The days and precipitation variables are correlated and are the variables that contribute most to explaining the 2nd PC.

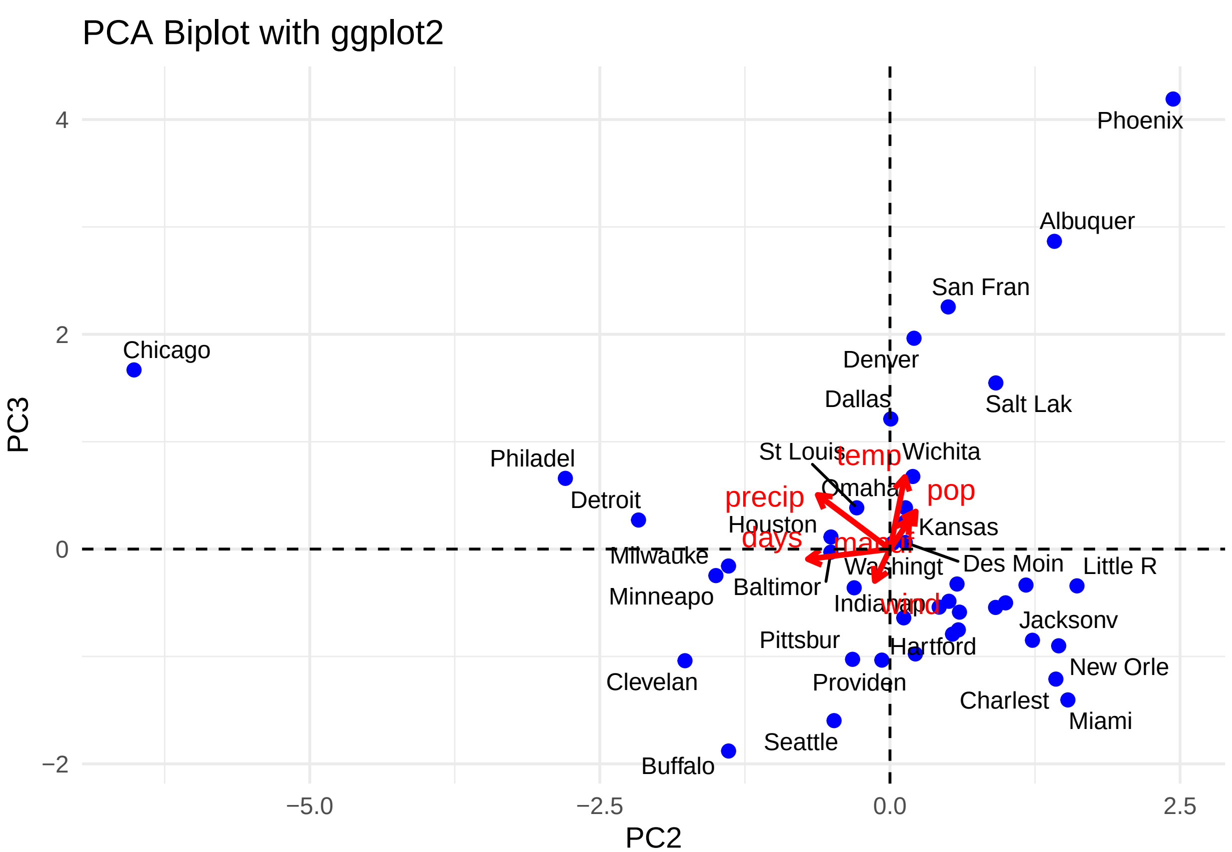
```
# Extracting scores (PC coordinates for samples) and loadings (contributions of variables)
scores <- as.data.frame(pca_airpollution$scores)   # Principal component scores
loadings <- as.data.frame(pca_airpollution$loadings[, 2:3])  # Loadings for the second and third PCs
```

```
# Renaming the columns for clarity
colnames(scores) <- c("PC2", "PC3")
rownames(scores) <- airpollution$city
loadings$Variables <- rownames(loadings)
colnames(loadings) <- c("PC2", "PC3", "Variables")
```

```
library(ggplot2)
library(ggrepel)  # For better label placement
library(dplyr)

# Create a ggplot2 biplot
ggplot() +
  # Plot the points for observations (scores)
  geom_point(data = scores, aes(x = PC2, y = PC3), color = "blue", size = 2) +

  # Add text labels for observations (optional)
  geom_text_repel(data = scores, aes(x = PC2, y = PC3, label = rownames(scores)), size = 3) +
```

```r
  # Plot the loadings as arrows
  geom_segment(data = loadings, aes(x = 0, y = 0, xend = PC2, yend = PC3),
               arrow = arrow(length = unit(0.2, "cm")), color = "red", size = 1) +

  # Add text labels for variables
  geom_text_repel(data = loadings, aes(x = PC2, y = PC3, label = Variables),
                  color = "red", size = 4) +

  # Add title and labels
  labs(title = "PCA Biplot with ggplot2", x = "PC2", y = "PC3") +

# Add horizontal and vertical axes
  geom_hline(yintercept = 0, linetype = "dashed", color = "black") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "black") +


  # Improving the theme
  theme_minimal()
```

```
## Warning: ggrepel: 10 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



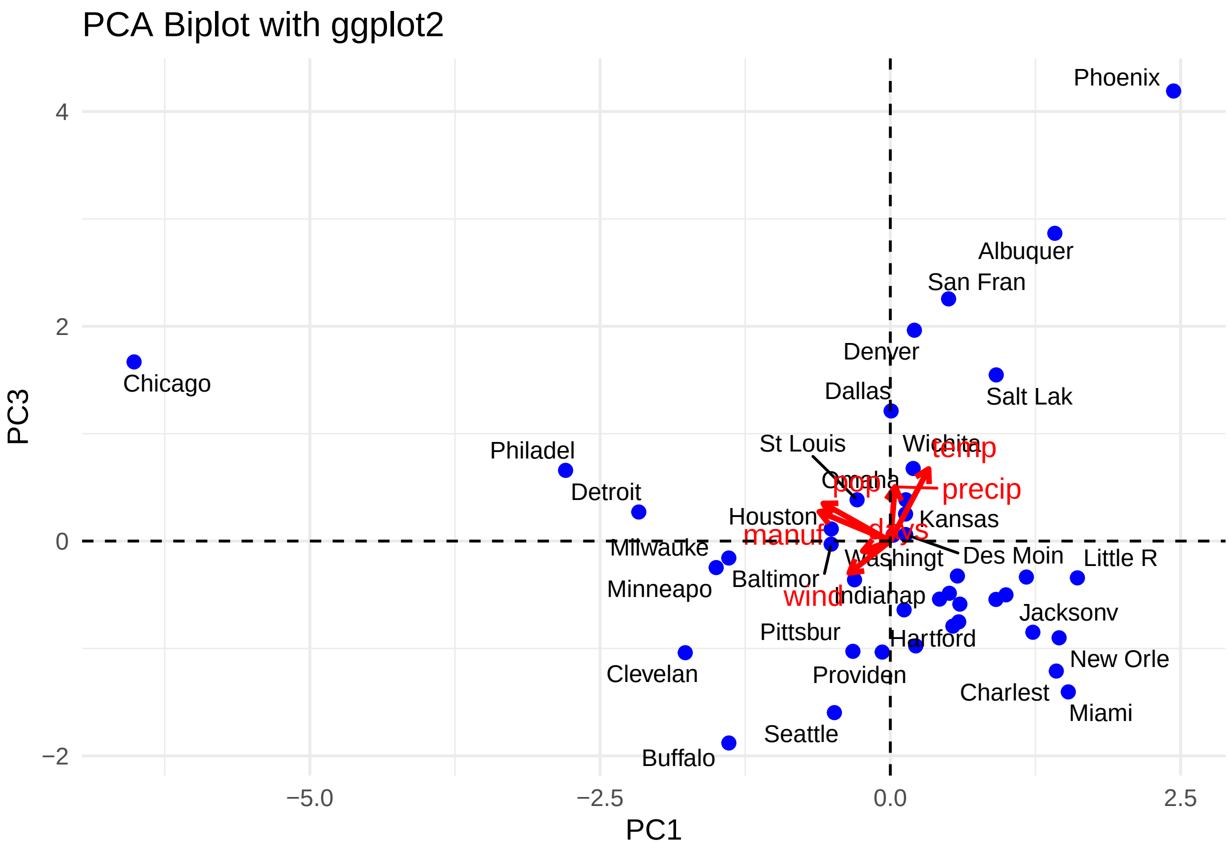
As predicted by the previous analyses, the temp variable is the one that best explains the 3rd PC. Since the precip variable contributes well to explaining both the 2nd PC and the 3rd PC, the angle formed between the arrow and each of the axes is almost the same.

```r
# Extracting scores (PC coordinates for samples) and loadings (contributions of variables)
scores <- as.data.frame(pca_airpollution$scores)  # Principal component scores
loadings <- as.data.frame(pca_airpollution$loadings[, c(1, 3)])  # Loadings for the first and third PCs
```

```r
# Renaming the columns for clarity
colnames(scores) <- c("PC1", "PC3")
rownames(scores) <- airpollution$city
loadings$Variables <- rownames(loadings)
colnames(loadings) <- c("PC1", "PC3", "Variables")
```

```r
library(ggplot2)
library(ggrepel)  # For better label placement
library(dplyr)

# Create a ggplot2 biplot
ggplot() +
  # Plot the points for observations (scores)
  geom_point(data = scores, aes(x = PC1, y = PC3), color = "blue", size = 2) +

  # Add text labels for observations (optional)
  geom_text_repel(data = scores, aes(x = PC1, y = PC3, label = rownames(scores)), size = 3) +

  # Plot the loadings as arrows
  geom_segment(data = loadings, aes(x = 0, y = 0, xend = PC1, yend = PC3),
               arrow = arrow(length = unit(0.2, "cm")), color = "red", size = 1) +

  # Add text labels for variables
  geom_text_repel(data = loadings, aes(x = PC1, y = PC3, label = Variables),
                  color = "red", size = 4) +

  # Add title and labels
  labs(title = "PCA Biplot with ggplot2", x = "PC1", y = "PC3") +

# Add horizontal and vertical axes
  geom_hline(yintercept = 0, linetype = "dashed", color = "black") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "black") +


  # Improving the theme
  theme_minimal()
```

```
## Warning: ggrepel: 10 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Clustering using K-means Algorithm

We chose 3 clusters based on the number of principal components retained.

```
#standardize data
airpollution_scaled <- scale(airpollution_variables)
```

```
head(airpollution_scaled)
```

```
##                temp       manuf         pop       wind     precip         days
## Phoenix   2.0112281 -0.44384938 -0.04594916 -2.4106088 -2.5246484 -2.939002785
## Little R  0.7245145 -0.66036338 -0.82299955 -0.8706873  0.9982522 -0.524493296
## San Fran  0.1295825 -0.01792019  0.18543918 -0.5207052 -1.3684710 -1.769474751
## Denver   -0.5345277 -0.01614549 -0.16164333 -0.3107159 -2.0234400 -1.052667247
## Hartford -0.9219254 -0.09068309 -0.77810330 -0.3107159  0.5607567  0.494127895
## Wilmingt -0.2439795 -0.67988513 -0.91279204 -0.3107159  0.2957109  0.003680655
```

```
set.seed(90)
kmean <- kmeans(airpollution_scaled, centers = 3)
#kmean <- kmeans(airpollution_scaled, centers = 3, nstart = 25)
kmean
```

```
## K-means clustering with 3 clusters of sizes 16, 23, 2
##
## Cluster means:
##         temp      manuf         pop       wind      precip        days
## 1  0.9346432 -0.2996546 -0.1804220 -0.5863268  0.61024043 -0.15194203
```

```
## 2 -0.6121276 -0.1086616 -0.1824397  0.3740318 -0.42758790  0.09061612
## 3 -0.4376783  3.6468455  3.5414335  0.3892485  0.03533737  0.17345085
##
## Clustering vector:
##  Phoenix Little R San Fran   Denver Hartford Wilmingt Washingt Jacksonv
##        1        1        2        2        2        2        1        1
##    Miami  Atlanta  Chicago Indianap Des Moin  Wichita Louisvil New Orle
##        1        1        3        2        2        2        1        1
## Baltimor  Detroit Minneapo   Kansas St Louis    Omaha Albuquer   Albany
##        2        2        2        2        2        2        2        2
##  Buffalo Cincinna Clevelan Columbus Philadel Pittsbur Providen  Memphis
##        2        1        2        2        3        2        2        1
## Nashvill   Dallas  Houston Salt Lak  Norfolk Richmond  Seattle Charlest
##        1        1        1        2        1        1        2        1
## Milwauke
##        2
##
## Within cluster sum of squares by cluster:
## [1] 60.259298 74.526644  7.753294
##  (between_SS / total_SS =  40.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

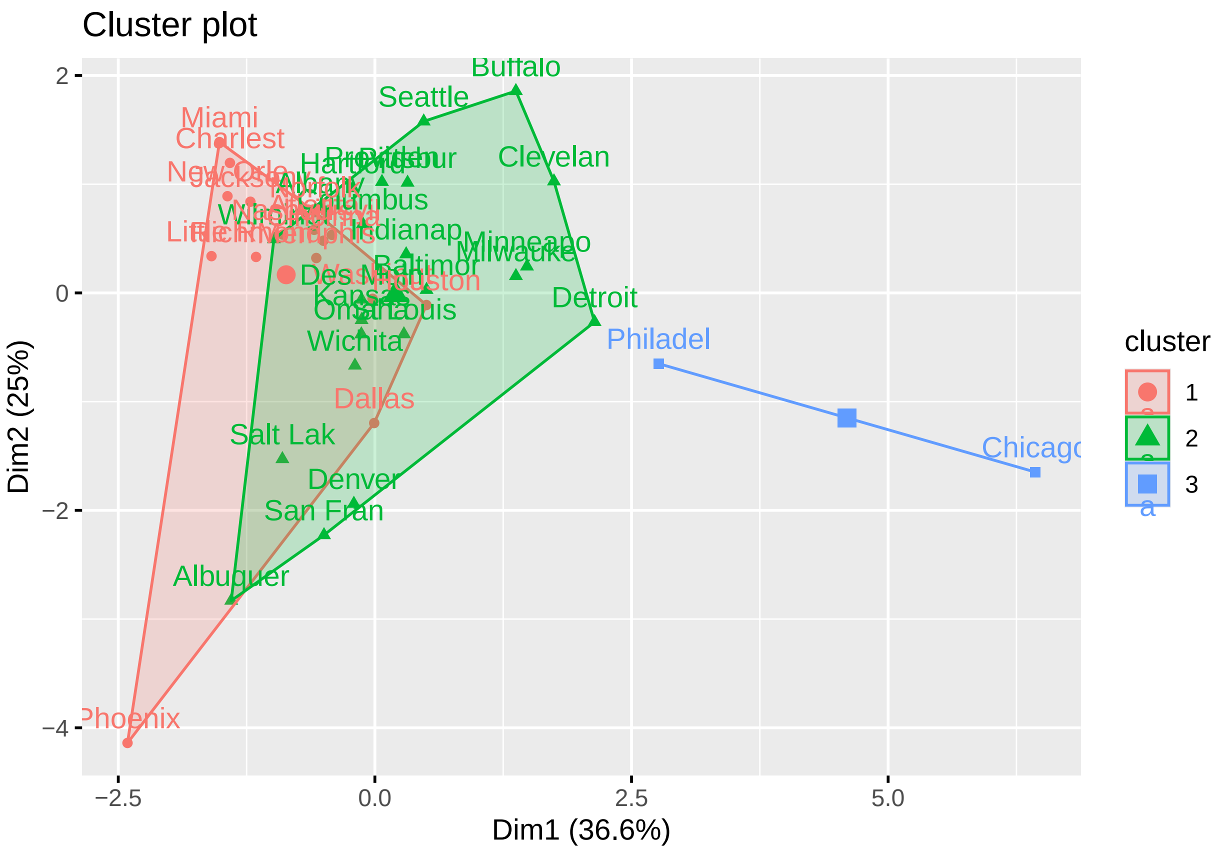
```r
kmean$cluster
```

```
##  Phoenix Little R San Fran   Denver Hartford Wilmingt Washingt Jacksonv
##        1        1        2        2        2        2        1        1
##    Miami  Atlanta  Chicago Indianap Des Moin  Wichita Louisvil New Orle
##        1        1        3        2        2        2        1        1
## Baltimor  Detroit Minneapo   Kansas St Louis    Omaha Albuquer   Albany
##        2        2        2        2        2        2        2        2
##  Buffalo Cincinna Clevelan Columbus Philadel Pittsbur Providen  Memphis
##        2        1        2        2        3        2        2        1
## Nashvill   Dallas  Houston Salt Lak  Norfolk Richmond  Seattle Charlest
##        1        1        1        2        1        1        2        1
## Milwauke
##        2
```

```r
library(factoextra)
```

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

```r
fviz_cluster(kmean, data = airpollution_variables)
```



We interpret this graph from left to right, with the three clusters formed. The first cluster is made up of cities whose most significant pollution indicators are in the temperature and precipitation variables (e.g. Miami and Charlest). The second cluster has to do with the cities that are the most polluting according to manufacturing, population and days indicators (e.g. Detroit and Clevelan). The overlap of the first two clusters shows us that there are a lot of cities that share indicators, which makes sense given the problem of pollution. The third cluster only has two cities, that are Chicago and Philadel. We can possibly interpret Chicago as an outlier because of the very high values that it has compared to other cities.