Study on pollution in the USA, Group A

Alexandre Baptista 64506   André Pires 64347   Vram Davtyan 64691

November 2024

Introdution

In this assignment, we analyze the air pollution dataset using Principal Component Analysis (PCA) and K-means clustering. The dataset includes air quality and environmental factors from 41 US cities. It the goal of performing the best models/results for the pollution dataset.

1. Exploratory Data Analysis

The given dataset is already discretized, contaning non-null or blank spaces. It comprises 41 rows and 8 columns. Also the following libraries help visualize the function made, and make functions, there were choosen because are easy and simple to use.

1.1 Load the necessary libraries

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.3

##
## 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
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.3.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.3.3

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

```
library(AMR)
```

```
## Warning: package 'AMR' was built under R version 4.3.3
```

1.2 Load Dataset

```
data <- read.csv("airpollution.csv")
```

1.3 Descriptive analysis

```
head(data)
```

```
##       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
```

```
str(data)
```

```
## '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 ...
```

```
dim(data)
```

```
## [1] 41  8
```

1.4 Localization measures

```
summary(data)
```

```
##      city                so2              temp           manuf       
##  Length:41          Min.   :  8.00   Min.   :43.50   Min.   :  35.0  
##  Class :character   1st Qu.: 13.00   1st Qu.:50.60   1st Qu.: 181.0  
##  Mode  :character   Median : 26.00   Median :54.60   Median : 347.0  
##                     Mean   : 30.05   Mean   :55.76   Mean   : 463.1  
##                     3rd Qu.: 35.00   3rd Qu.:59.30   3rd Qu.: 462.0  
##                     Max.   :110.00   Max.   :75.50   Max.   :3344.0  
##       pop              wind            precip           days      
##  Min.   :  71.0   Min.   : 6.000   Min.   : 7.05   Min.   : 36.0  
##  1st Qu.: 299.0   1st Qu.: 8.700   1st Qu.:30.96   1st Qu.:103.0  
##  Median : 515.0   Median : 9.300   Median :38.74   Median :115.0  
##  Mean   : 608.6   Mean   : 9.444   Mean   :36.77   Mean   :113.9  
##  3rd Qu.: 717.0   3rd Qu.:10.600   3rd Qu.:43.11   3rd Qu.:128.0  
##  Max.   :3369.0   Max.   :12.700   Max.   :59.80   Max.   :166.0  
```

1.5 Dispersion measures

```
data %>% 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
```

2. Calculation of components

Each principal component (PC) has an associated eigenvalue that quantifies the amount of variance explained by that component. The higher the eigenvalue, the more variance that component captures.

```
# 1) Determine the correlation matrix
cor_data <- cor(data[, sapply(data, is.numeric)])
cor_data
```

```
##                so2        temp       manuf         pop        wind      precip
## so2     1.00000000 -0.43360020  0.64476873  0.49377958  0.09469045  0.05429434
## temp   -0.43360020  1.00000000 -0.19004216 -0.06267813 -0.34973963  0.38625342
## manuf   0.64476873 -0.19004216  1.00000000  0.95526935  0.23794683 -0.03241688
## pop     0.49377958 -0.06267813  0.95526935  1.00000000  0.21264375 -0.02611873
## wind    0.09469045 -0.34973963  0.23794683  0.21264375  1.00000000 -0.01299438
## precip  0.05429434  0.38625342 -0.03241688 -0.02611873 -0.01299438  1.00000000
## days    0.36956363 -0.43024212  0.13182930  0.04208319  0.16410559  0.49609671
##               days
## so2     0.36956363
## temp   -0.43024212
## manuf   0.13182930
## pop     0.04208319
## wind    0.16410559
## precip  0.49609671
## days    1.00000000
```

```
# 2) Obtain eigenvalues and eigenvectors
eigen_data <- eigen(cor_data)
eigen_data
```

```
## 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
```

As can be seen we have 2 principal components (PC), because according to Kraiser's criterion: the first eigenvalues >=1 –> Retain the principal components.

3. Perform PCA

The explained varience is demonstrated as the percentage of the total variance explained by each component:

- PC1 explain approximately 39% of the variance.
- PC2 explain approximately 21% of the varience.

Together the PC1 and PC2 explain approximately 60% of the total varience, what is a moderate value.

```r
numeric_data <- data[, sapply(data, is.numeric)]
pca_data <- princomp(numeric_data,cor = TRUE)
print(summary(pca_data), 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
```

The importance of variables in each retained principal component is determined by their loading values, which show the correlation between the original variables and the principal components.

- Loading Values: Indicates how much each variable contributes to the PC. A higher absolute value of a loading indicates that the variable has a stronger influence on that PC.
- Sign of Loadings: The sign (positive or negative) shows the direction of the relationship (e.g., positive loading means the variable increases with the component, while negative loading means the variable decreases with the component).

PC1:

- High positive loadings for SO2, temperature, and population, indicate that these variables increase as PC1 increases.
- Negative loadings for precipitation or wind,indicate that these variables decrease as PC1 increases.
- PC1 may be capturing overall urbanization or industrialization.

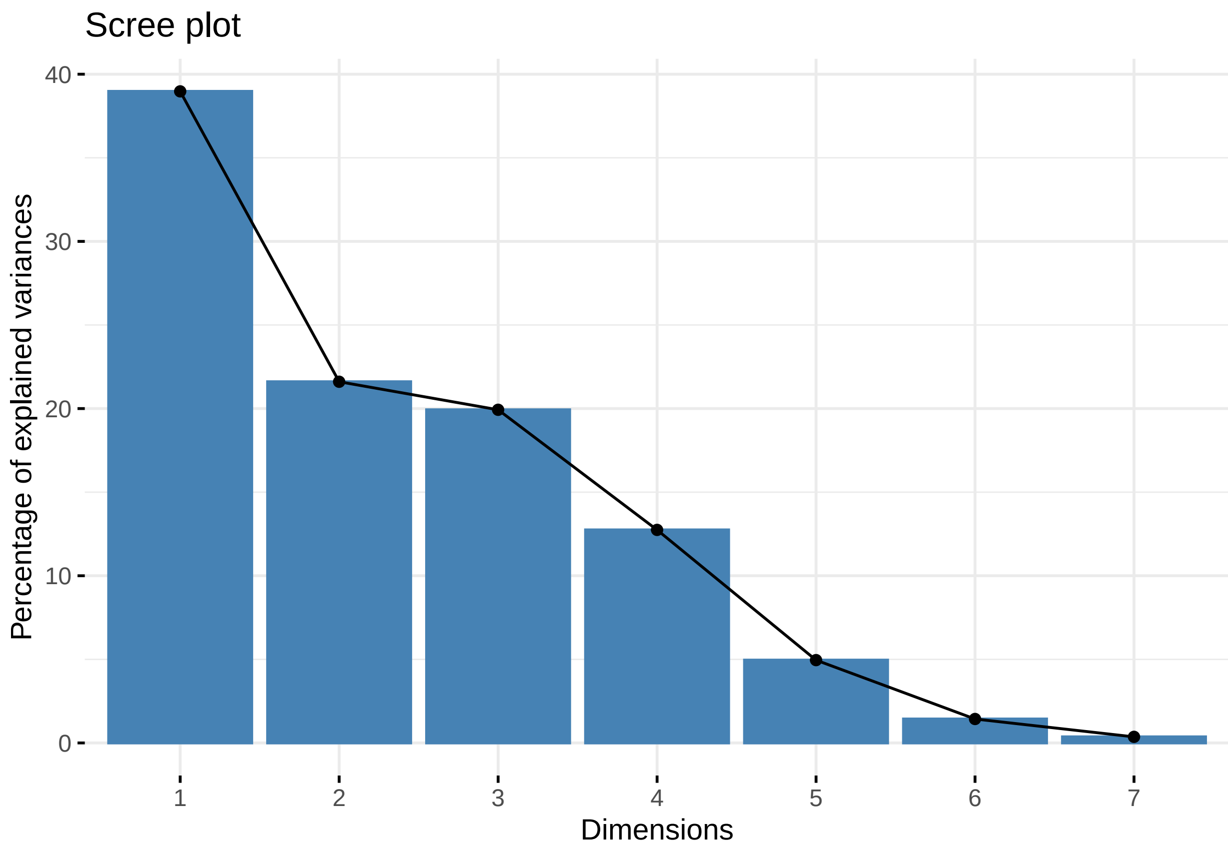
PC2:

- High loadings for precipitation, wind, and temperature.
- PC2 captures more of the environmental factors, like climate, rainfall, and wind conditions, which contrast with industrial factors.

4. Scree plot

The scree plot shows the eigenvalues (variance explained) for each PC.

```r
fviz_eig(pca_data)
```



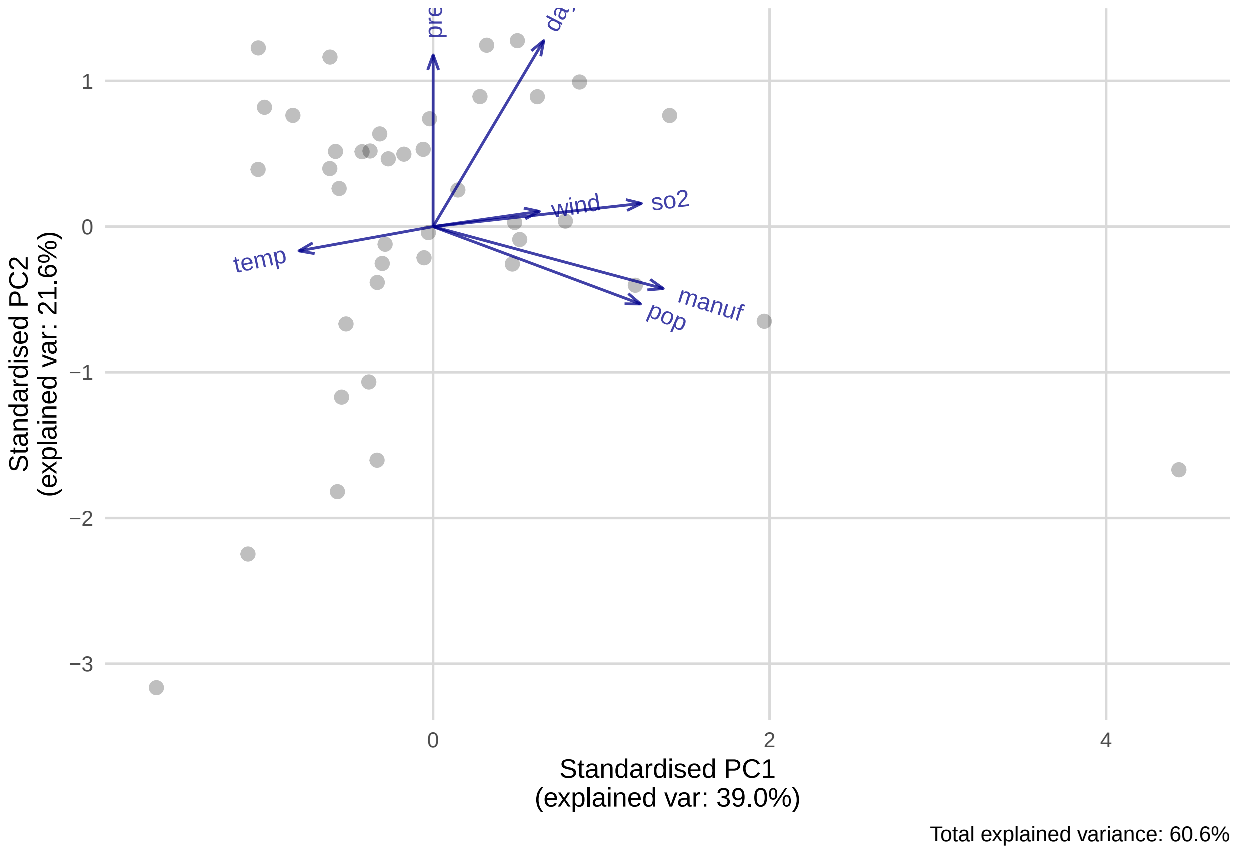
5. Identify the variables that contribute more in relation to the component retained

```r
cor(numeric_data,pca_data$scores)
```

```
##               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
```

```
ggplot_pca(pca_data)
```

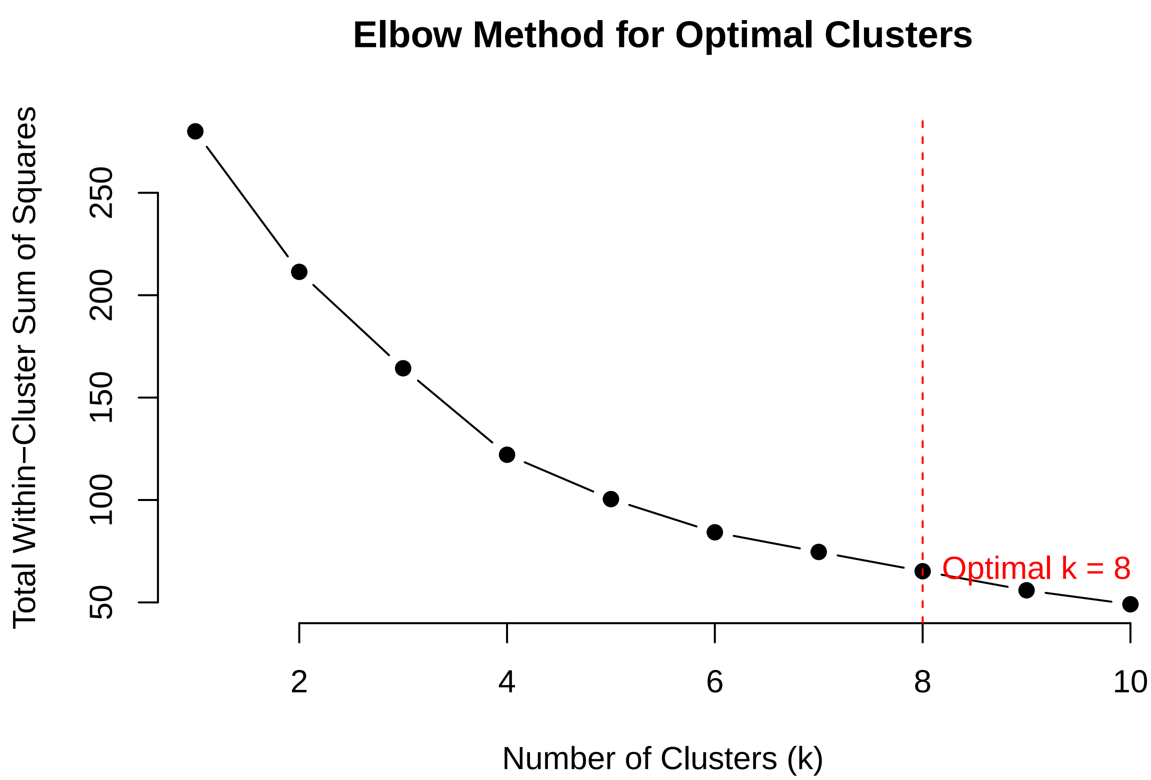


6. Elbow method

The elbow method suggest the optimal number of clusters (k). Looking for the point where the curve flattens and adding a vertical line to be sure, we obtain the optimal K= 8.

```
# Compute WSS (Within-Cluster Sum of Squares) for different k values
data_scaled <- scale(numeric_data)
set.seed(123)
wss <- sapply(1:10, function(k) {
  kmeans(data_scaled, centers = k, nstart = 25)$tot.withinss
})
# Determine the optimal number of clusters (elbow point)
optimal_k <- which.min(diff(diff(wss))) + 1  # Add 1 because we applied two differences
# Plot the Elbow Method
plot(1:10, wss, type = "b", pch = 19, frame = FALSE,
     xlab = "Number of Clusters (k)",
     ylab = "Total Within-Cluster Sum of Squares",
     main = "Elbow Method for Optimal Clusters")
# Add a vertical line at the optimal number of clusters
abline(v = optimal_k, col = "red", lty = 2)
# Add a label to indicate the optimal k
```

```
text(optimal_k, wss[optimal_k], labels = paste("Optimal k =", optimal_k), pos = 4, col = "red")
```



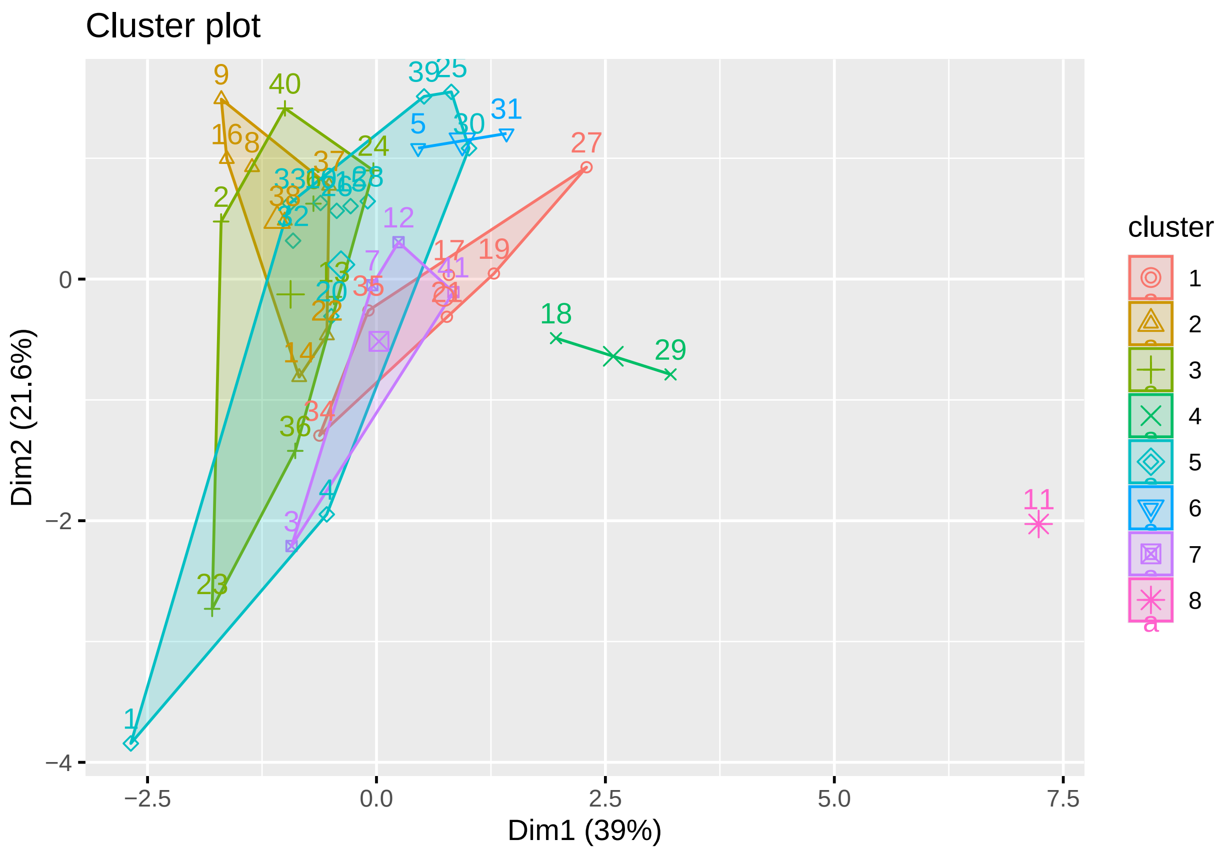
7. k-means

```
kmeans <- kmeans(numeric_data, 8)
kmeans
```

```
## K-means clustering with 8 clusters of sizes 6, 7, 7, 2, 12, 2, 4, 1
##
## Cluster means:
##      so2     temp      manuf       pop      wind   precip     days
## 1  36.00 56.53333  744.66667  849.8333 10.383333 37.04333 114.8333
## 2  16.00 62.48571  163.71429  350.8571  9.714286 45.58143 109.7143
## 3  26.00 53.51429   76.71429  145.7143  8.757143 30.95286 106.7143
## 4  52.00 52.25000 1378.00000 1731.5000  9.850000 35.44500 122.0000
## 5  22.75 55.74167  347.00000  522.7500  8.866667 35.89333 118.8333
## 6  75.00 49.55000  377.50000  168.5000  9.800000 43.06000 126.0000
## 7  21.25 53.00000  454.25000  734.0000  9.875000 31.84000 105.5000
## 8 110.00 50.60000 3344.00000 3369.0000 10.400000 34.44000 122.0000
##
## Clustering vector:
##  [1] 5 3 7 5 6 3 7 2 2 5 8 7 3 2 5 2 1 4 1 5 1 2 3 3 5 5 1 5 4 5 6 5 5 1 1 3 2 2
## [39] 5 3 7
##
## Within cluster sum of squares by cluster:
```

```
## [1] 327491.342  57002.318  40686.705 293403.900 114757.075   3326.877  26159.885
## [8]      0.000
##  (between_SS / total_SS =  96.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
fviz_cluster(kmeans, numeric_data)
```



Observing the plot, each point in the plot represents an individual observation, colored according to its assigned cluster. The shapes of the points represent different clusters, and the shaded areas indicate the convex hulls that enclose the observations for each cluster.

7.1 Cluster Sizes:

- Cluster 1: 6 observation.
- Cluster 2: 7 observation.
- Cluster 3: 7 observation.
- Cluster 4: 2 observation.
- Cluster 5: 12 observation.
- Cluster 6: 2 observation.
- Cluster 7: 4 observation.
- Cluster 8: 1 observation.

7.2 Cluster Means:

This provides a summary of the average values for each variable within each cluster.

7.3 Plot Interpretation:

- The plot uses the 2 principal components to reduce dimensionality.
- We can see that cluster 8 (pink) is a distint outlier in terms of pollution (SO2), with very high values compared to the other clusters.

7.4 Cluster characteristics in the plot:

- Cluster 1 (green) and Cluster 5 (blue) are near the center of the plot, with moderate pollution and moderate temperature, population, and manufacturing levels.
- Cluster 4 (yellow) stands out with high pollution (SO2), representing industrialized regions with large populations.
- Cluster 8 (pink) is an outlier with very high SO2, manufacturing, and population, showing extreme values compared to the other clusters.

8 Clusters description

8.1 Cluster 1 (Red):

- Is the PC1, which indicates a unique positioning compared to other clusters. It's separated from the other clusters in terms of Dim1 and Dim2 that suggest disting patterns in the original data.
- Has high concentration of SO2, manufacturing, and population, with moderated precipitation and wind values. It appears to be an outlier in terms of Dim1.

8.2 Cluster 2 (Brown):

- It is in the middle of the plot, showing a mix of data points but generally concentrated around Dim1 and Dim2 axes. It has a moderate to high population and moderate levels of SO2, temp, and precipitation.
- This cluster represents areas with moderate air pollution and moderate temperatures, but with a diverse range of industrial activity.

8.3 Cluster 3 (Ligh Green):

- It is located at the positive side of Dim1 and relative lower in Dim2. It tends to have low to moderate values across most features but is grouped tightly, indicating consistent values across this cluster.
- It has lower levels of SO2, temperature, and manufacturing, and are marked by higher precipitation and wind.This could be a rural or less industrialized region with a high focus on environmental factors.

8.4 Cluster 4 (Dark Green):

- It is located in the lower right quadrant of the plot, with high Dim1 and low Dim2 values.The points are scattered but still tightly grouped.
- It has higher values for temperature and manufacturing, along with moderate SO2 levels and population density. This could represent urban areas with moderate industrial activity.

8.5 Cluster 5 (Cyan):

- It is located towards the upper left of the plot, with data ponts scattered around both axes.
- It has higher SO2, population, and manufacturing values, suggesting that these could be densely populated urban areas with moderate industrialization and air pollution levels.

8.6 Cluster 6 (Blue):

- It is scattered in top right corner of the plot. With relatively high SO2, temperature, and precipitation values.
- It most likely represents hotter and more polluted areas, possibly liked with industrial activities producing higher SO2 emissions.

8.7 Cluster 7 (Purple):

- It is spread across the lower-left quadrant of the plot, showing distints patterns in Dim1 and Dim2.
- It has areas with lower SO2, temperature, and manufacturing, with moderate precipitation. These might be rural or less industrialized regions.

8.8 Cluster 8 (Pink)

- It it located towards the far right of the plot in the Dim1 positive area.
- It has the highest SO2, temperature, and manufacturing values. It's likely an extremely industrialized area with high levels of pollution.

8.9 Summary of the Clusters:

- Cluster 1: High SO2, manufacturing, population, but distinct from other clusters.
- Cluster 2: Moderate values across most features, indicating a mix of pollution and environmental factors.
- Cluster 3: Low pollution and manufacturing, with high environmental factors like precipitation.
- Cluster 4: Urban area with moderate manufacturing and SO2 levels.
- Cluster 5: Densely populated urban areas with high industrial activity.
- Cluster 6: Hotter and polluted areas with high precipitation.
- Cluster 7: Rural or less industrialized with low SO2, temperature, and manufacturing.
- Cluster 8: Highly industrialized area with high SO2 and temperature levels.