Formative Practical Report: Summarising Multivariate Data and PCA

Jack Young

2022-10-28

1 - Summarising the airpollution data

(a) - Numerical and graphical summaries of the data

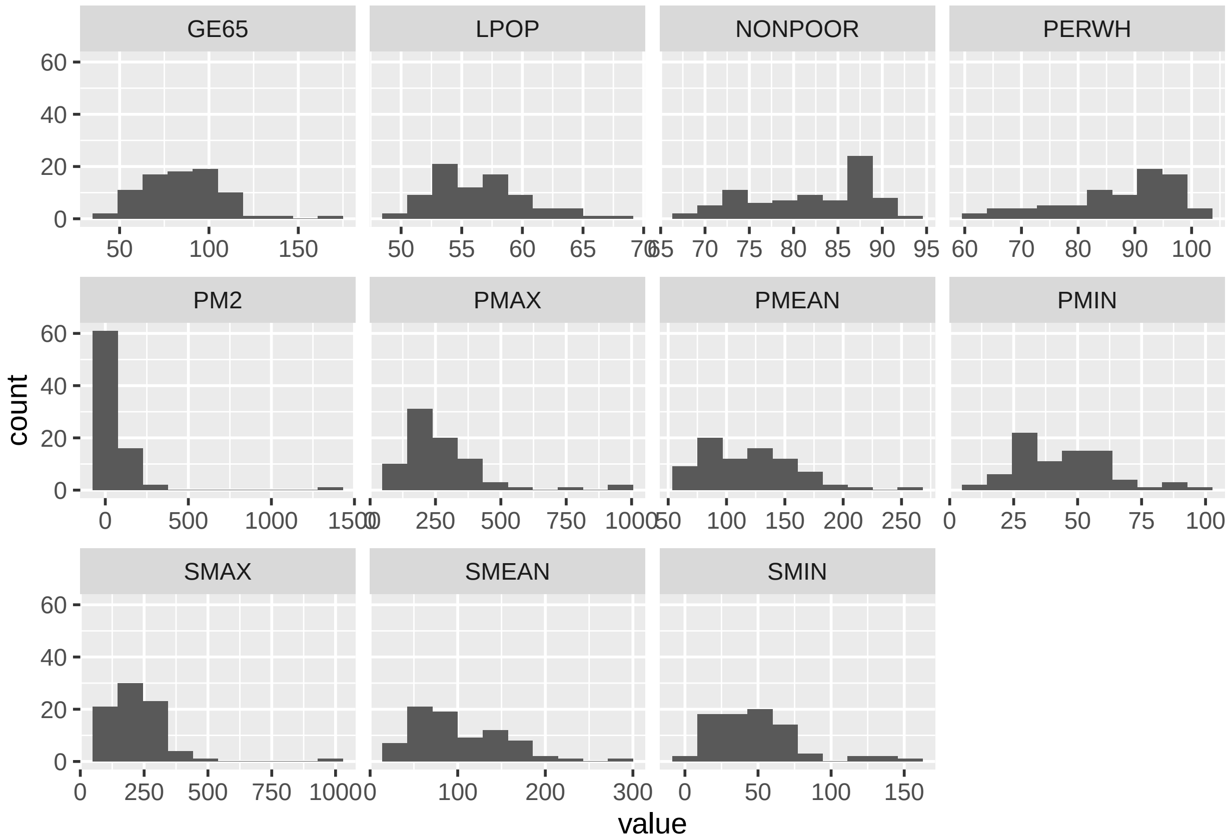
We start by selecting the `airpollution` dataset from our package. We can draw immediate insights using the summary function - as shown below, this shows us all of the variables in our data as well as some summary statistics for each variable:

```
summary(airpollution)
```

```
##       SMIN           SMEAN            SMAX           PMIN      
##  Min.   :  1.00   Min.   : 26.00   Min.   : 58.0   Min.   :10.00  
##  1st Qu.: 25.25   1st Qu.: 64.25   1st Qu.:146.2   1st Qu.:29.75  
##  Median : 45.00   Median : 86.00   Median :201.5   Median :42.50  
##  Mean   : 47.10   Mean   : 99.65   Mean   :219.9   Mean   :44.50  
##  3rd Qu.: 60.00   3rd Qu.:136.00   3rd Qu.:282.0   3rd Qu.:55.25  
##  Max.   :155.00   Max.   :283.00   Max.   :940.0   Max.   :98.00  
##      PMEAN           PMAX            PM2               PERWH      
##  Min.   : 54.0   Min.   :117.0   Min.   :   1.60   Min.   :60.00  
##  1st Qu.: 83.5   1st Qu.:171.0   1st Qu.:  23.88   1st Qu.:81.92  
##  Median :115.0   Median :234.5   Median :  37.95   Median :90.30  
##  Mean   :116.7   Mean   :275.5   Mean   :  72.86   Mean   :87.26  
##  3rd Qu.:142.8   3rd Qu.:327.5   3rd Qu.:  73.30   3rd Qu.:95.28  
##  Max.   :247.0   Max.   :978.0   Max.   :1357.20   Max.   :99.70  
##     NONPOOR          GE65             LPOP      
##  Min.   :67.80   Min.   : 45.00   Min.   :49.37  
##  1st Qu.:76.30   1st Qu.: 72.00   1st Qu.:53.84  
##  Median :83.55   Median : 85.50   Median :56.01  
##  Mean   :81.83   Mean   : 85.88   Mean   :56.55  
##  3rd Qu.:87.20   3rd Qu.: 98.25   3rd Qu.:58.47  
##  Max.   :93.20   Max.   :171.00   Max.   :67.94  
```

We can also visualize this information using histogram plots for each variable - these give us a little more insight as to how the values are distributed:

```
## Produce histograms of each variable in the airpollution data
ggplot(gather(airpollution), aes(value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~key, scales = 'free_x')
```



As above, we see a variety of distributions present in our data. For example, some seem to roughly follow a normal distribution, whereas others do not, such as the `PM2` variable which appears to approximately follow an exponential distribution.

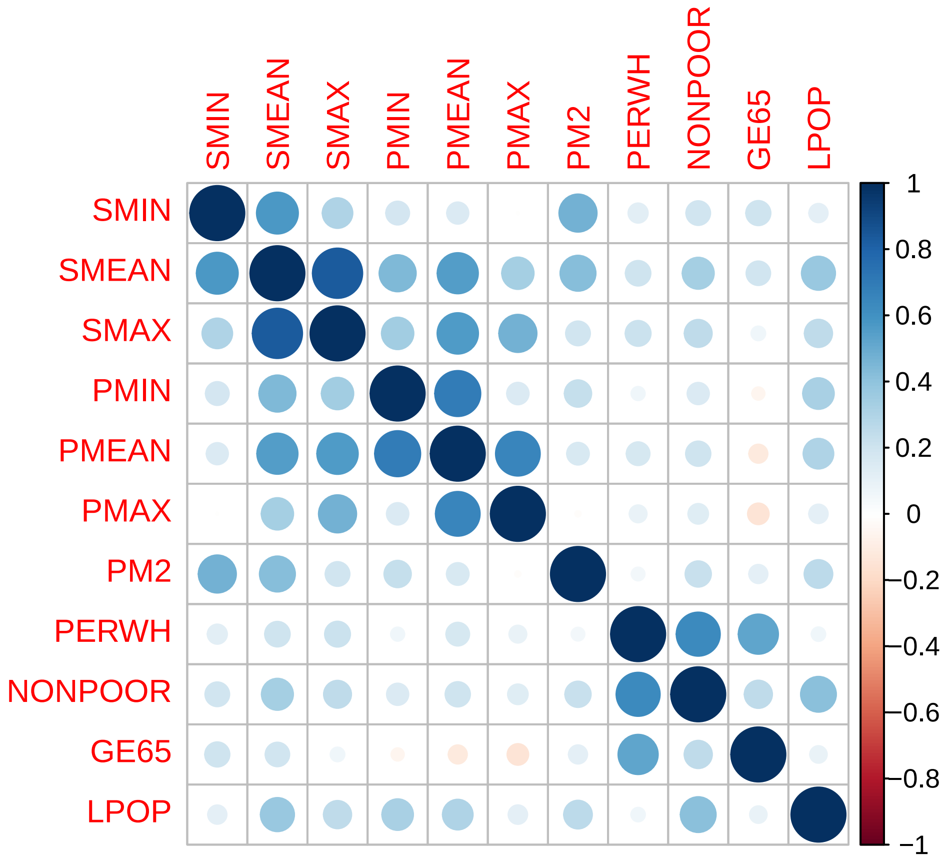
We can check the dimensions of the data like so:

```
dim(airpollution)
```

```
## [1] 80 11
```

In the context of this dataset, this means that we have 80 observations, and 11 variables in total. We could produce a scatterplot matrix for our data, but this may be impractical in this case given the large number of variables present. A good alternative for this is to use a correlation heatmap instead, which colour-codes variable pairs based on their correlation coefficient:

```
## Produce a correlation heatmap based on the data matrix
cor_matrix <- cor(airpollution)
corrplot::corrplot(cor_matrix)
```



For example, we can see here that there is a fairly strong correlation between variables 'PERWH' and 'NONPOOR', however there are very few variables present that show strong intercorrelation.

(b) - Variation in the data

As stated in Section 1.4.4 in the course notes, we have the following two measures of multivariate scatter:

1. **Generalised variance:** $\det(S) = |S|$, the determinant of the sample covariance matrix,
2. **Total variation:** $\mathrm{tr}(S)$, the trace of the sample covariance matrix.

Using R, we can compute the sample covariance matrix S by simply feeding our data matrix to the `var` function like so:

```r
## Compute the sample covariance matrix S
S <- var(airpollution)
S
```

```
##                SMIN      SMEAN       SMAX      PMIN      PMEAN         PMAX
## SMIN     913.154430   874.7443  1097.0127 100.22785  182.48354    -8.256962
## SMEAN    874.744304 2542.9392  5036.0570 415.32911 1084.07975  2716.810759
## SMAX    1097.012658 5036.0570 14409.3513 750.51899 2612.75000  9048.498418
## PMIN     100.227848   415.3291   750.5190 337.84810  496.20253   466.588608
## PMEAN    182.483544 1084.0797  2612.7500 496.20253 1508.35380  4056.858544
## PMAX      -8.256962 2716.8108  9048.4984 466.58861 4056.85854 25312.504905
## PM2     2212.321899 3282.9385  3633.6467 681.14747  982.00370  -246.966155
```

```
## PERWH     39.985316  109.0520   266.8187  11.50000   72.16538   163.838323
## NONPOOR   39.698354  112.7077   202.4796  19.36899   53.31307   143.418528
## GE65     131.417722  208.3861   169.1361 -20.81013  -95.25000  -505.020570
## LPOP      13.543716   73.2376   118.2727  22.85910   45.43667    72.616843
##                 PM2      PERWH   NONPOOR       GE65       LPOP
## SMIN     2212.32190  39.985316  39.69835  131.41772  13.543716
## SMEAN    3282.93854 109.052025 112.70766  208.38608  73.237597
## SMAX     3633.64668 266.818671 202.47959  169.13608 118.272652
## PMIN      681.14747  11.500000  19.36899  -20.81013  22.859099
## PMEAN     982.00370  72.165380  53.31307  -95.25000  45.436669
## PMAX     -246.96616 163.838323 143.41853 -505.02057  72.616843
## PM2     23920.23764  92.016832 230.47854  383.08592 157.772005
## PERWH      92.01683 107.820956  44.59833  118.32120   2.544737
## NONPOOR   230.47854  44.598326  45.45271   37.18592  10.839829
## GE65      383.08592 118.321203  37.18592  465.45253   7.926460
## LPOP      157.77200   2.544737  10.83983    7.92646  14.856788
```

Now we have computed S, we can find the generalized variance by computing the determinant:

```r
genvar <- det(S)
genvar
```

```
## [1] 8.72131e+29
```

We can also compute the total variation by taking the trace of the matrix:

```r
totvar <- sum(apply(airpollution, 2, var))
totvar
```

```
## [1] 69577.97
```

(c) - Standardising the data matrix

An important practice in PCA is the standardizing of data - this means that the features are scaled such that they are distributed around a mean of zero with a standard deviation of one. We may then go ahead and compare covariances for pairs of features in our data, but we shall first check that our assumptions hold by performing standardization on the `airpollution` data.

```r
## Standardize the airpollution data
airpollution_standard <- scale(airpollution)
```

Now we have our standardized data, we can check that the sample mean vector is composed of zeros:

```r
## Round values to 10 decimal places to account for rounding errors
standard_mean_vec <- round(colMeans(airpollution_standard),10)
standard_mean_vec
```

```
##    SMIN   SMEAN    SMAX    PMIN   PMEAN    PMAX     PM2   PERWH NONPOOR    GE65
##       0       0       0       0       0       0       0       0       0       0
##    LPOP
##       0
```

We must also check that the sample covariance matrix is equal to the sample correlation matrix of the original `airpollution` data. We can check these matrices are identical using the `all.equal` function:

```
## Take the covariance matrix of the standardised data
standard_cov_matrix <- cov(airpollution_standard)
## Again accounting for rounding issues, check the new covariance matrix
## is equal to the original correlation matrix
all.equal(round(cor_matrix,10), round(standard_cov_matrix,10))
```

```
## [1] TRUE
```

As we have now verified our assumptions made about the standardised data matrix, we may proceed to perform PCA on the data.

2 - Principal Component Analysis

(a) - Which matrix?

When examining the sample variances for the 11 variables in our data, we find the following:

```
## Take the individual variances for each variable in airpollution
apply(airpollution, 2, var)
```

```
##        SMIN       SMEAN        SMAX        PMIN       PMEAN        PMAX
##   913.15443  2542.93924 14409.35127   337.84810  1508.35380 25312.50491
##         PM2       PERWH     NONPOOR        GE65        LPOP
## 23920.23764   107.82096    45.45271   465.45253    14.85679
```

Notice that, for example, the variance of the `PM2` variable is significantly larger than that of the `LPOP` variable. As PCA is not scale invariant, this could affect our analysis as if several components have a larger mean/variance than others in the data, they will dominate our PCA if based on our covariance matrix S. Therefore, we shall instead choose a PCA based on the spectral decomposition of the sample correlation matrix, which is equivalent to performing the analysis on the standardised data.

(b) - Performing PCA on the standardised data

Now we have decided to perform our analysis on the sample correlation matrix, we can start our analysis.

```
## Perform PCA on the sample correlation matrix
pca_airpol <- prcomp(airpollution, scale = TRUE)
pca_airpol
```

```
## Standard deviations (1, .., p=11):
##  [1] 1.9589090 1.3778958 1.1793324 1.0214751 0.8790417 0.8214578 0.7377982
##  [8] 0.6606760 0.4573085 0.3293824 0.2896070
##
## Rotation (n x k) = (11 x 11):
##              PC1         PC2         PC3         PC4         PC5         PC6
## SMIN   0.2613624  0.19024716  0.48898065  0.30389689 -0.00503646  0.16976617
## SMEAN  0.4503394 -0.01348804  0.17831895  0.22217870 -0.07700225 -0.25640015
```

```
## SMAX    0.3988570 -0.13441659 -0.05261114  0.33299140 -0.16751423 -0.32783475
## PMIN    0.3126485 -0.22716515  0.07514298 -0.35107342  0.67334658  0.02028976
## PMEAN   0.3868269 -0.34029207 -0.19234925 -0.05145100  0.26347970  0.14420155
## PMAX    0.2522820 -0.34479429 -0.37450985  0.25148644 -0.30869420  0.20626628
## PM2     0.2404705  0.14632479  0.51477592 -0.11716178 -0.11430278  0.43974026
## PERWH   0.2073243  0.45946073 -0.43826348  0.08658339  0.18383543  0.25437827
## NONPOOR 0.2764271  0.36544285 -0.27541948 -0.29711596 -0.27193283  0.32090663
## GE65    0.1059282  0.53990856 -0.09313504  0.17056040  0.29745501 -0.43689985
## LPOP    0.2651881  0.04129927  0.04278853 -0.64781366 -0.37227336 -0.42692962
##                 PC7         PC8         PC9        PC10         PC11
## SMIN    -0.23357936  0.65047519 -0.10093751  0.10793791  0.188886532
## SMEAN   -0.16687341 -0.14094009  0.10793985 -0.24130316 -0.725666579
## SMAX    -0.23117364 -0.43584227 -0.07259588  0.21225036  0.529091005
## PMIN    -0.11701411 -0.01835620  0.21795695  0.45080432 -0.056405756
## PMEAN    0.14871000  0.14558294 -0.18129015 -0.68523112  0.242875167
## PMAX     0.44615456  0.27668187  0.19043628  0.37810587 -0.145921101
## PM2      0.48679957 -0.43913254 -0.05078188  0.02049280  0.052870507
## PERWH   -0.07529613 -0.09090255 -0.60313853  0.18709915 -0.187409126
## NONPOOR -0.34131500 -0.01058218  0.54480664 -0.13760757  0.128915709
## GE65     0.50321729  0.10479728  0.30194137 -0.07121415  0.140363292
## LPOP     0.13009585  0.23828579 -0.31924187  0.09954071  0.009544292
```

To begin to draw some insights from our PCA, we can extract components individually, like so:

```
## Compute the variances of each principal component
pca_airpol$sdev^2
```

```
##  [1] 3.83732465 1.89859672 1.39082484 1.04341143 0.77271434 0.67479298
##  [7] 0.54434620 0.43649281 0.20913105 0.10849277 0.08387221
```

```
## Extract the loadings matrix
pca_airpol$rotation
```

```
##               PC1         PC2         PC3         PC4         PC5         PC6
## SMIN    0.2613624  0.19024716  0.48898065  0.30389689 -0.00503646  0.16976617
## SMEAN   0.4503394 -0.01348804  0.17831895  0.22217870 -0.07700225 -0.25640015
## SMAX    0.3988570 -0.13441659 -0.05261114  0.33299140 -0.16751423 -0.32783475
## PMIN    0.3126485 -0.22716515  0.07514298 -0.35107342  0.67334658  0.02028976
## PMEAN   0.3868269 -0.34029207 -0.19234925 -0.05145100  0.26347970  0.14420155
## PMAX    0.2522820 -0.34479429 -0.37450985  0.25148644 -0.30869420  0.20626628
## PM2     0.2404705  0.14632479  0.51477592 -0.11716178 -0.11430278  0.43974026
## PERWH   0.2073243  0.45946073 -0.43826348  0.08658339  0.18383543  0.25437827
## NONPOOR 0.2764271  0.36544285 -0.27541948 -0.29711596 -0.27193283  0.32090663
## GE65    0.1059282  0.53990856 -0.09313504  0.17056040  0.29745501 -0.43689985
## LPOP    0.2651881  0.04129927  0.04278853 -0.64781366 -0.37227336 -0.42692962
##                 PC7         PC8         PC9        PC10         PC11
## SMIN    -0.23357936  0.65047519 -0.10093751  0.10793791  0.188886532
## SMEAN   -0.16687341 -0.14094009  0.10793985 -0.24130316 -0.725666579
## SMAX    -0.23117364 -0.43584227 -0.07259588  0.21225036  0.529091005
## PMIN    -0.11701411 -0.01835620  0.21795695  0.45080432 -0.056405756
## PMEAN    0.14871000  0.14558294 -0.18129015 -0.68523112  0.242875167
## PMAX     0.44615456  0.27668187  0.19043628  0.37810587 -0.145921101
## PM2      0.48679957 -0.43913254 -0.05078188  0.02049280  0.052870507
```

```
## PERWH   -0.07529613 -0.09090255 -0.60313853  0.18709915 -0.187409126
## NONPOOR -0.34131500 -0.01058218  0.54480664 -0.13760757  0.128915709
## GE65     0.50321729  0.10479728  0.30194137 -0.07121415  0.140363292
## LPOP     0.13009585  0.23828579 -0.31924187  0.09954071  0.009544292
```

By using the loadings matrix, we find that the first principal component is given by:

$$\begin{aligned}\texttt{PC1} = 0.261\texttt{SMIN} + 0.450\texttt{SMEAN} + 0.399\texttt{SMAX} + 0.313\texttt{PMIN} + 0.387\texttt{PMEAN} + 0.252\texttt{PMAX}+ \\ 0.240\texttt{PM2} + 0.207\texttt{PERWH} + 0.276\texttt{NONPOOR} + 0.106\texttt{GE65} + 0.265\texttt{LPOP}\end{aligned}$$

As we can see, the first principal component isn't particularly dominated by any one of our variables here, as all of our coefficients fall between +0.1 and +0.5. We find that the higher coefficients have been generally attributed to sulphate and particulate readings however, so we may interpret our first principal component as a weighted average of pollution rates. Cities with higher readings of pollution will have larger scores for PC1, but more generally cities with high values across the 11 variables will score highly here.

Moving on to the second principal component:

$$\begin{aligned}\texttt{PC2} = 0.190\texttt{SMIN} - 0.013\texttt{SMEAN} - 0.134\texttt{SMAX} - 0.227\texttt{PMIN} - 0.340\texttt{PMEAN} - 0.345\texttt{PMAX}+ \\ 0.146\texttt{PM2} + 0.459\texttt{PERWH} + 0.365\texttt{NONPOOR} + 0.540\texttt{GE65} + 0.041\texttt{LPOP}\end{aligned}$$

The second principal component differs from the first in that it contains both positive and negative coefficients for the variables. Generally speaking, the demographic factors have positive coefficients, with the `GE65` variable the largest in absolute value of these. On the other hand, the pollution-related variables generally have been attributed with positive coefficients - especially the particulate readings. Therefore, we could interpret that cities with lower pollution rates and more white, less deprived and older populations will have a high PC score for PC2, and vice versa. This principal component allows us to contrast high pollution rates with our numerical demographic factors.

(c) - How many Principal Components?

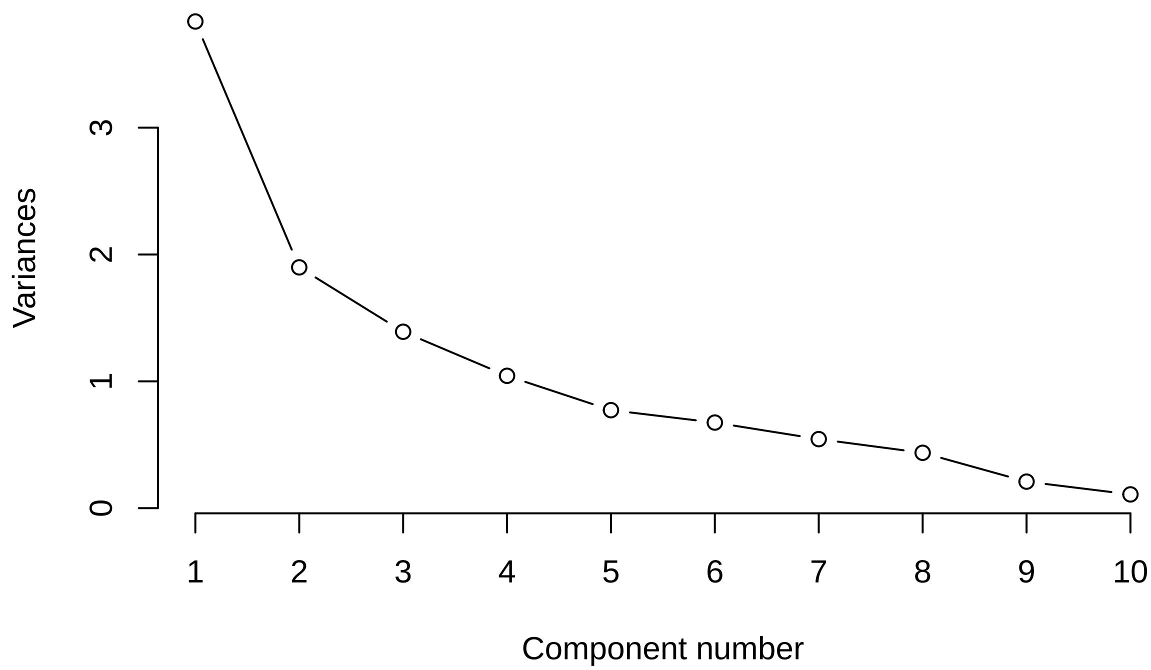
Now we have our principal components, we can decide how many to use. To do this, we use Result 2.1 - that is, we can take the sum of the variances of the principal components to be equal to the total variation in the original data. Therefore, we may use the variance of one principal component divided by the sum over all principal components to be the proportion of variation accounted for by our one principal component. R, using the `summary` function calculates the proportion of variance and cumulative proportion, as displayed below:

```
summary(pca_airpol)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.9589 1.3779 1.1793 1.02148 0.87904 0.82146 0.73780
## Proportion of Variance 0.3488 0.1726 0.1264 0.09486 0.07025 0.06134 0.04949
## Cumulative Proportion  0.3488 0.5214 0.6479 0.74274 0.81299 0.87433 0.92382
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.66068 0.45731 0.32938 0.28961
## Proportion of Variance 0.03968 0.01901 0.00986 0.00762
## Cumulative Proportion  0.96350 0.98251 0.99238 1.00000
```

We can also use a scree plot to help visualize this:

```r
plot(pca_airpol, type="lines", main="")
title(xlab="Component number")
```

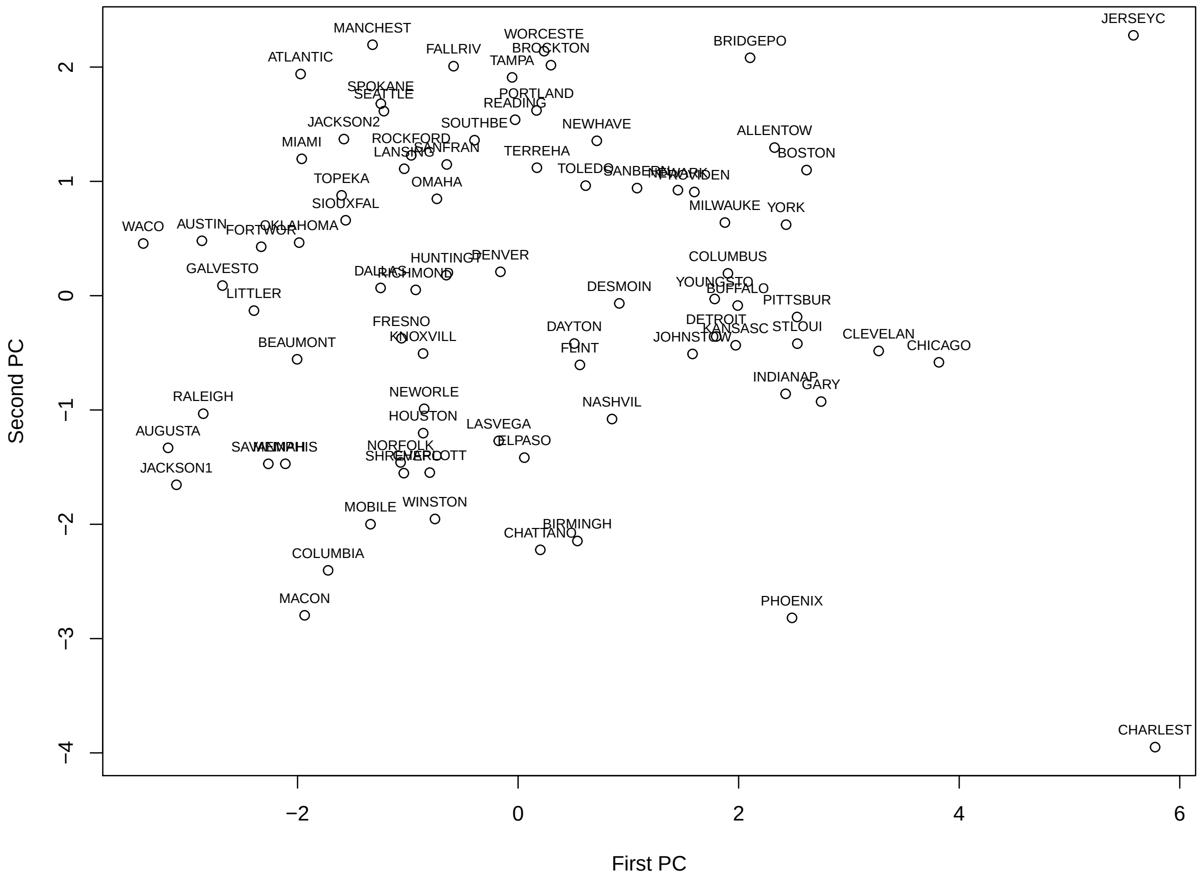


Usually when carrying out an analysis such as this, we would look for a 'kink' in our scree plot - i.e. where the gradient of our plot flattens out. Therefore, with an appropriate threshold in mind, we notice that the first 4 PCs explain around 74% of the total variation, and the remaining 6 components provide little in comparison. Therefore with the goal of dimension reduction in mind, we could probably disregard the final 6 for the purpose of our analysis.

(d) - Plotting the first two components

We can plot the first principal component scores against each other, labelling the points by the city they represent:

```r
# Plot the first two principal components against each other
plot(pca_airpol$x[,1], pca_airpol$x[,2], xlab="First PC", ylab="Second PC")
# Add labels representing the cities
text(pca_airpol$x[,1], pca_airpol$x[,2], labels=rownames(airpollution_standard), cex=0.7, pos=3)
```



From this plot, we can begin to draw some insight from our data based on the characteristics we managed to infer about each of the first two principal components earlier. For example, *CHARLESTON* stands out instantly, as we notice that it has the highest score in the data for our first PC, yet the lowest score for PC 2. Applying the interpretation we formulated in part (b), this would suggest that this city has high rates of pollution present in the air, and that its population is less white, more deprived and younger than most cities in our data. Now consider *JERSEYC* on the top right of the graph - this scores highly on both the first and second PC axes. With a high PC 1 score, we infer that this city generally had high scores across the 11 variables. However, with a high score for PC 2, this interpretation may change. High PC 2 values indicate more white, less deprived, and older populations, and as 5 out of 6 of our pollution variables are negatively weighted in this PC, this would also suggest low pollution levels in this city. Therefore combining these two interpretations, it is suggested that this city in particular will have high values across the board on our demographic factors outlined. However, as we have only taken the first two principal components, which we found to represent just 52% of the total variation in our data, we may have to take such interpretations with a pinch of salt.