Project Report - Air Pollution in American Cities

June 8, 2023

Authors:

Tuomas Rickansrud - trickansrud@ucdavis.edu. Contributions: PCA component selection, MLR

Lizzy Stampher - estampher@ucdavis.edu. Contributions: PCA implementation, bivariate box plot implementation

Emilio Barbosa Valdiosera - ebarbosavaldiosera@ucdavis.edu. Contributions: Introduction, exploratory analysis, conclusion

Jianing Zhu - jnzhu@ucdavis.edu. Contributions: City PCA rankings

Instructor: Dr. Xiucai Ding

STA 135 - Multivariate Data Analysis

University of California, Davis

Introduction

Air quality is extremely important to life on Earth. Low air quality has been repeatedly associated with various negative health effects 1 2. As such, the present analysis will focus on the US Air pollution data set and seek to answer which variables might best predict air pollution levels, as well as ranking the performance of the recorded cities. The data set, sourced from Sokal and Rohlf in their book *Biometry*, 2nd Edition (1981), features the following variables:

SO2: Measures of the air pollutant sulphur dioxide, in micrograms per cubic meter.

temp: Average annual temperatures, in Fahrenheit.

manu: The number of manufacturing enterprises with 20 or more workers.

popul: Population size, in thousands, as of 1970.

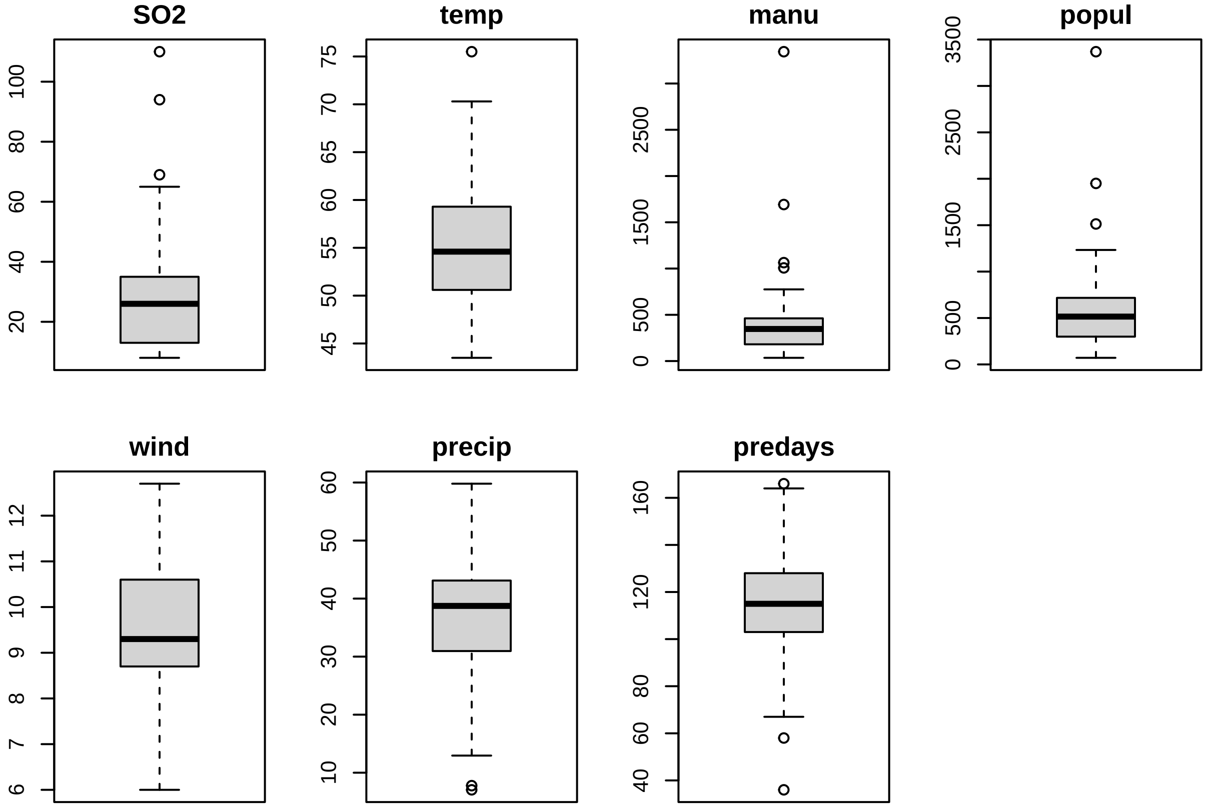
wind: Average annual wind speed, in miles per hour.

precip: Average annual precipitation, in inches.

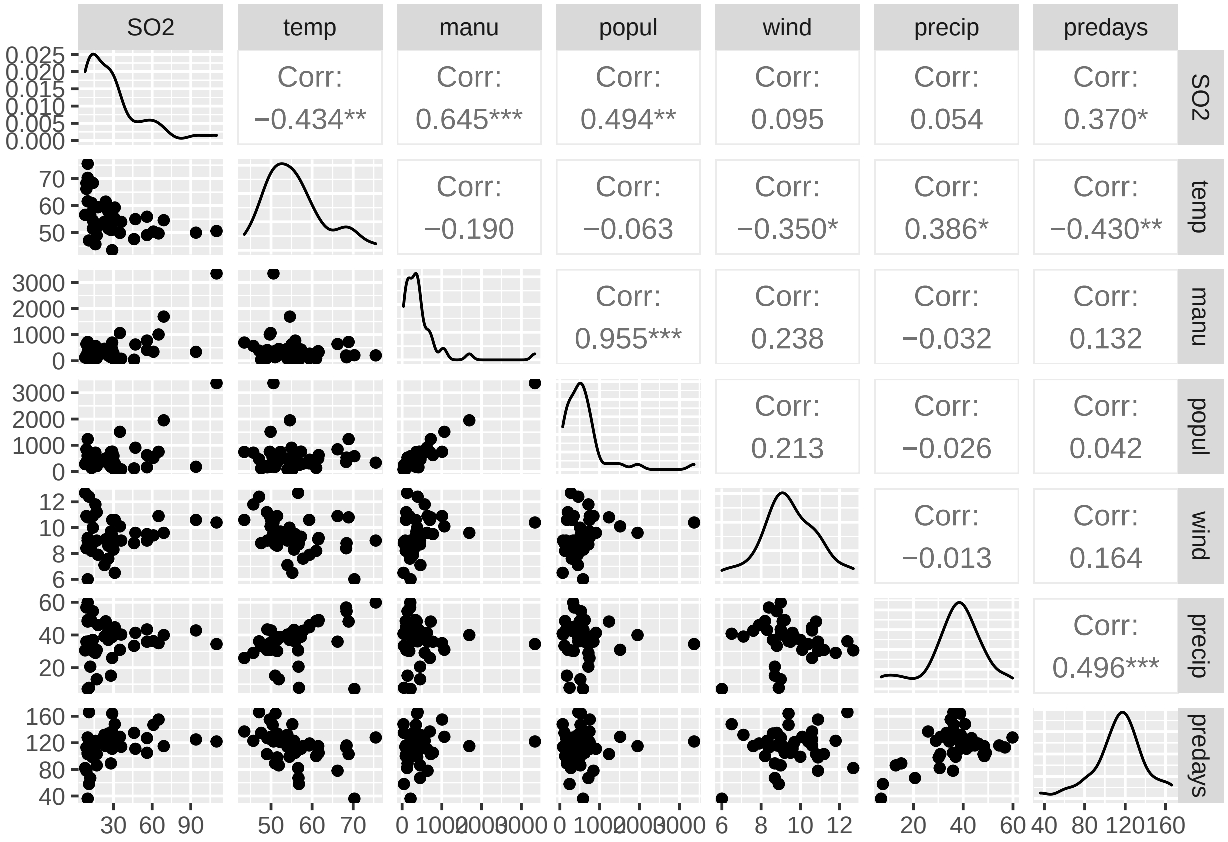
predays: Average number of days with precipitation, per year.

Exploratory analysis

As a first step, box plots for all the variables are plotted to check their distributions and spot potential outliers.

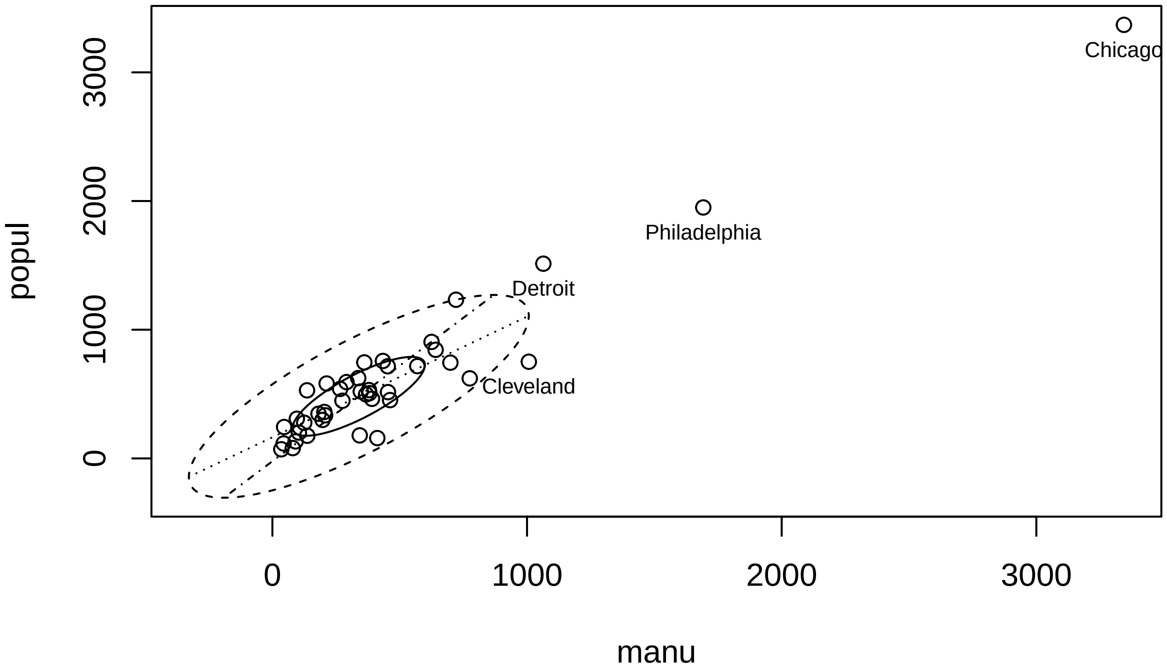


Further, a scatterplot matrix is used to check all combinations of pair-wise scatterplots, as well as their correlation coefficients:



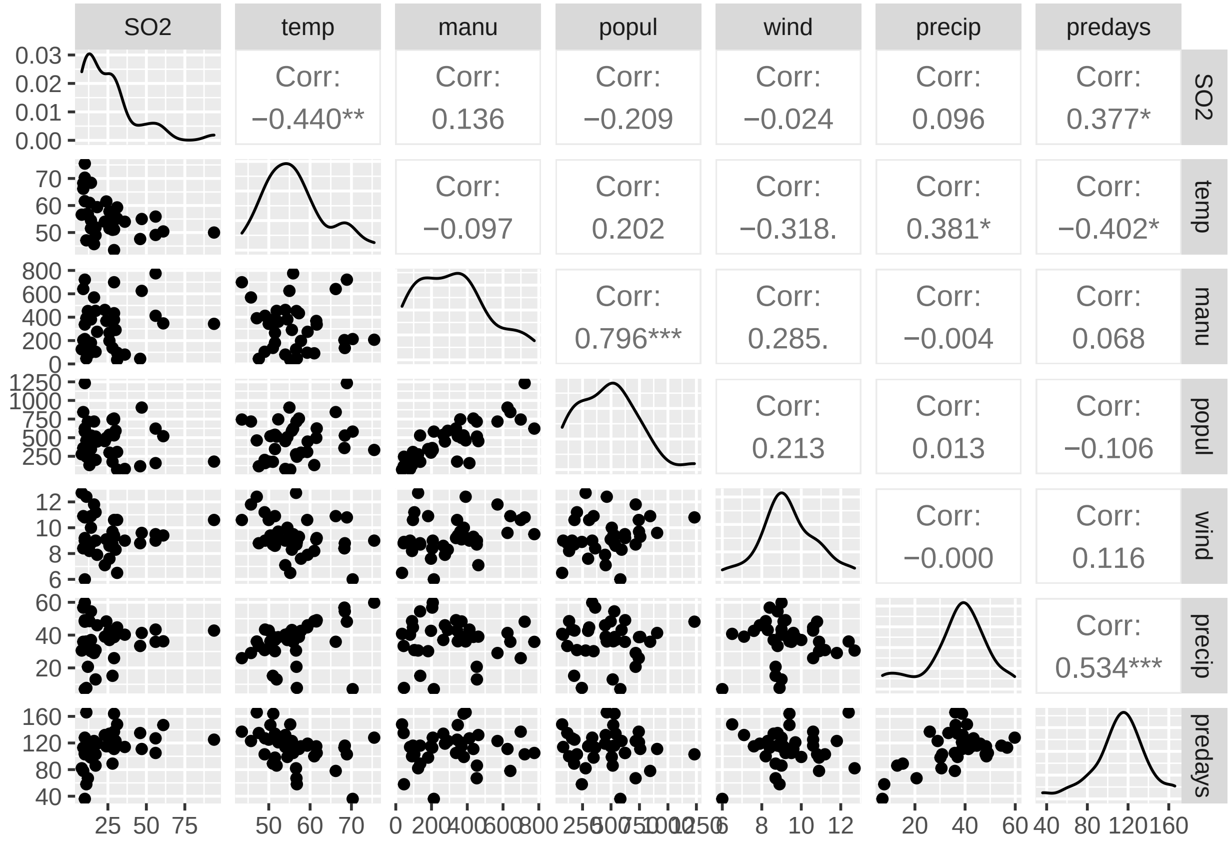
In the box plots, outliers in the manufacturing and population variables immediately stand out, as they appear to be very far from their distributions. This causes their box plots to be drawn very narrowly to fit the full distribution. Extreme values are also apparent across multiple scatterplots. A negative correlation can be seen between temperature and SO2, as well as a positive correlation between the population and manufacturing variables. A positive correlation between the precipitation and days of precipitation variables is also present. These relationships will come up again later when we perform the principal component analysis.

With respect to outliers, the process of identification in the multivariate setting is surprisingly challenging due to the nature of higher dimensional data, and more advanced methods of outlier detection appear to be outside of the scope of this course. However, a somewhat rudimentary method (obtained by Everitt and Hothorn (2011) with further reference to Goldberg and Iglewicz (1992)) involves the use of a bivariate box plot, which is applied to the manufacturing and population variables to further investigate their outliers:



This bivariate box plot of the population and manufacturing variables, with outliers labeled, indicates that the cities which may be useful to exclude from further analysis are Chicago, Philadelphia, Detroit, and Cleveland.

After removing the four cities, we check the scatterplot matrix again:



and the scatterplots look much improved, with fewer extreme-looking points. We now consider the data suitably pre-processed for the principal component analysis.

PCA

In order to answer the questions about which variables are most important for predicting air pollution levels and how the cities rank against each other, we use the multivariate analysis method of PCA. PCA is a way of re-expressing the variables in a data set by combining them together into a new set of variables, called principal components. These components are created such that they are mathematically independent from each other (in fact, the principal components are eigenvectors, which are linearly independent given unique eigenvalues), which allows one to select a subset of only the most important combinations of variables, thus reducing the overall complexity and dimensionality of the data set. Because these new variables are generated from the data set's measures of variance, the subset selected represents the variables which "explain" the highest amount of variance in the data.

This data set includes multiple different variables with different scales and units of measurement. Therefore, we prefer to use the correlation matrix in PCA rather than the covariance matrix, as the correlation matrix is the covariance matrix in a standardized form.

We also make another decision for the sake of interpretability: Noting the negative association between the temperature variable and the SO2 variable that can be seen in the scatterplot matrix, we flip the sign of the temperature variable, so that "increasing" values in the negative temperature track with increasing values in the other variables, and thus makes correlations between the principal components easier to see.

The principal components are calculated and output as follows:

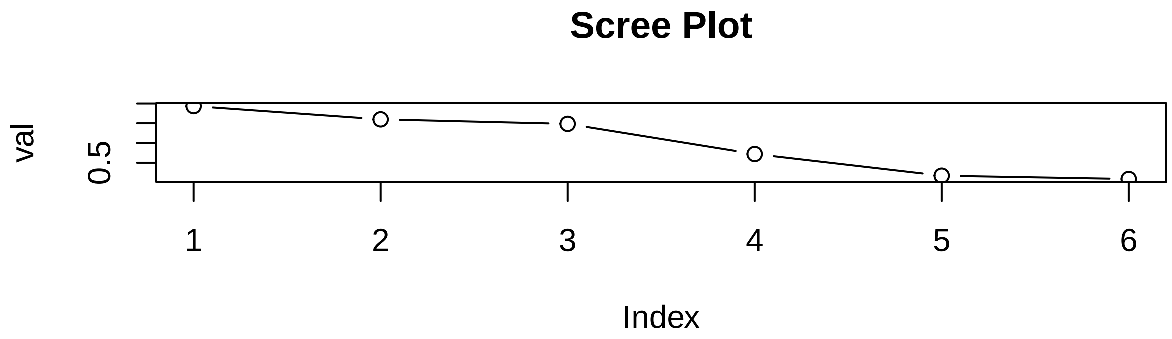
```
##
```

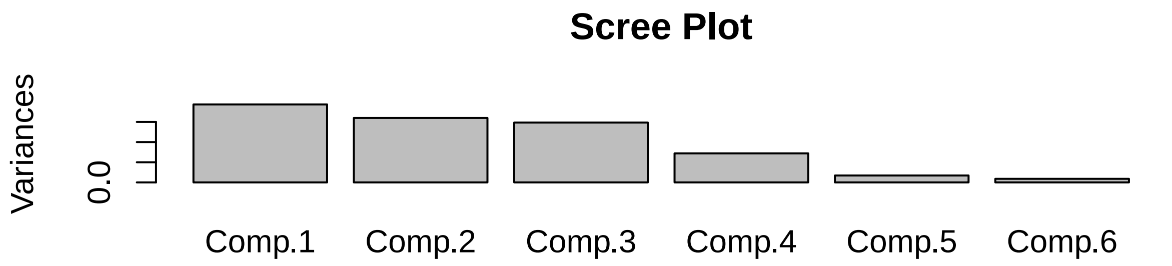
```
## Loadings:
##          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## neg_temp         0.519  0.562  0.276         0.568
## manu      0.666                0.287  0.631 -0.266
## popul     0.627 -0.270 -0.148  0.140 -0.621  0.326
## wind      0.386  0.286  0.223 -0.837        -0.131
## precip           0.287 -0.737 -0.188  0.283  0.509
## predays          0.700 -0.265  0.294 -0.353 -0.474
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.167  0.167  0.167  0.167  0.167  0.167
## Cumulative Var  0.167  0.333  0.500  0.667  0.833  1.000
```

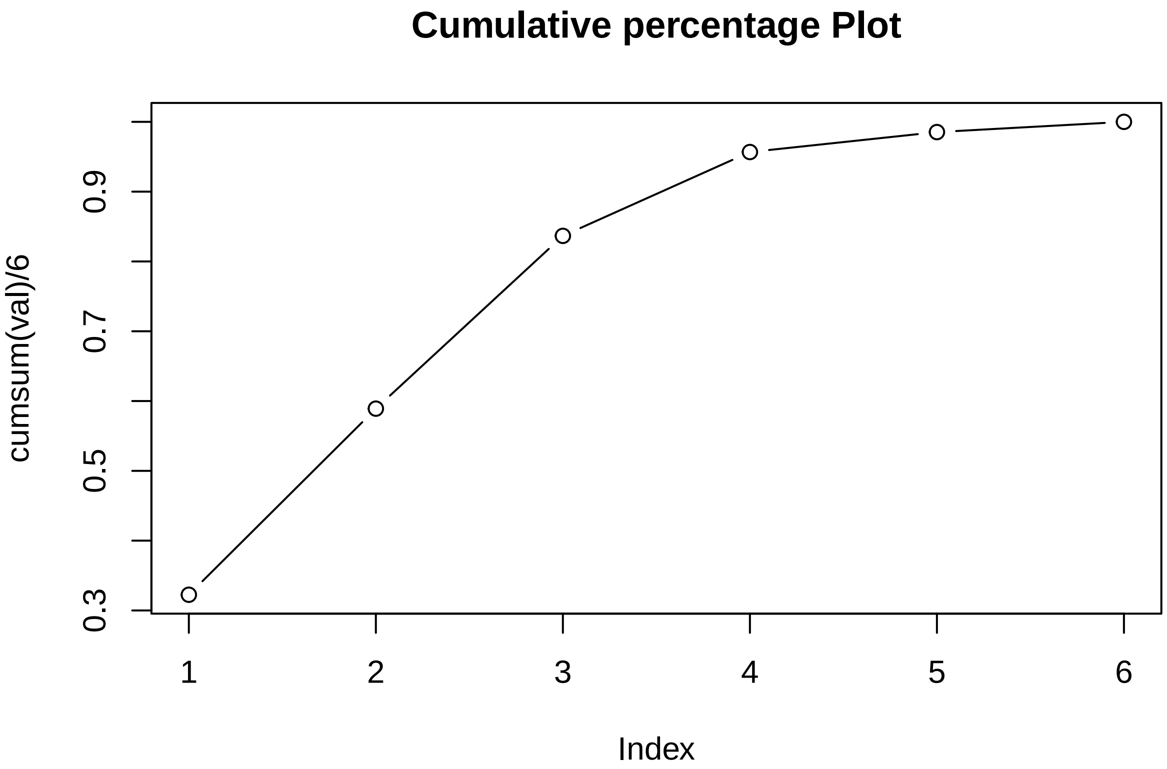
What this output seems to imply is that the first principal component weighs the manufacturing and population variables most highly. In the second component, days of precipitation and (negative) temperature receive the greatest weight. The third component also sees highest weights given to (negative) temperature and annual inches of precipitation. The fourth component gives the highest weight to wind speeds, and the fifth and sixth components appear to somewhat repeat the pattern of components one and two, but with the sign of the inches of precipitation weight flipped.

Selecting components

We now consider how many components to consider. When choosing the number of components to use in a model, it is typically done by selecting the optimal number that explains some set level of variance within the model. For our case we wanted at least 80% of the variance explained through our components which gave us the value three. We plot the scree and cumulative variance plots:

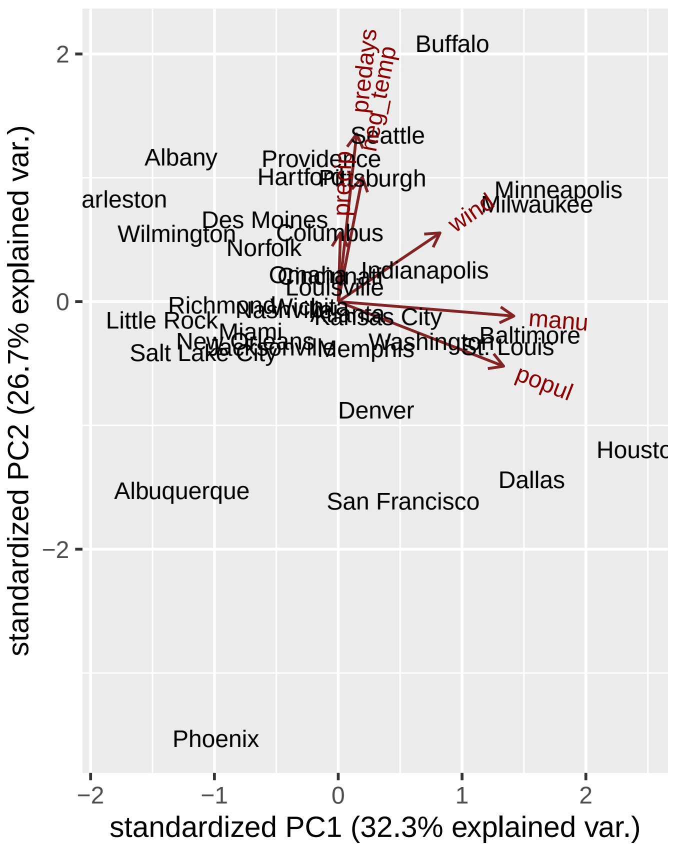

Scree Plot
val
0.5
1
2
3
4
5
6
Index


Scree Plot
Variances
0.0
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6



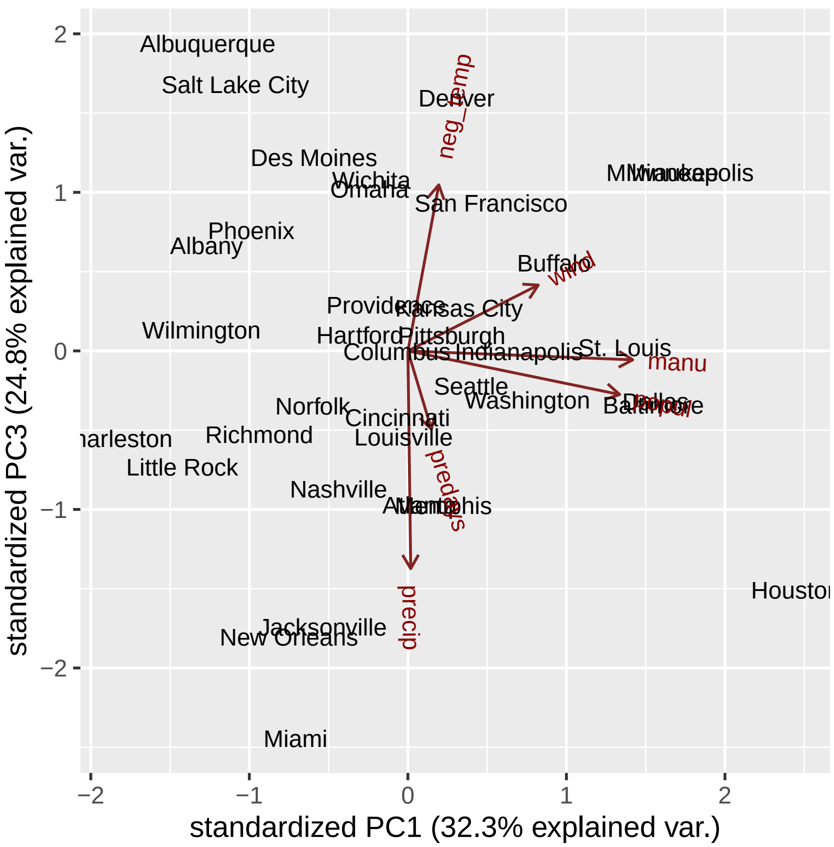
As seen in the Scree plot visualizations -specifically the Cumulative percentage plot- at three components we reach that threshold level at around 83.66% explained variance while the remaining increases in number of components give diminishing returns to the explained variance. This type of selection is typically referred to as using the 'elbow' of the plot because it easily visualizes how the first few components contribute most to the model, allowing the omission of the remaining components without the loss of any important information to the model. This is corroborated when looking at the individual variance percentages shown through the Scree Plot bar chart. Component one is contributing to the overall explained variance by about 32.25%. Component two by 26.66% for a cumulative value of 58.91%. Finally, component three adds an additional 24.75% to the explained variance for a total of 83.66% (Our threshold goal). Furthermore, we can see that component four supplies much less value to the variance at a value of 12.01%, over half of the previous component. This quantitatively displays the elbow technique mentioned earlier.

Now that we know how many components we consider most important, we also generate some biplots of the three selected principal components to look at things more visually:

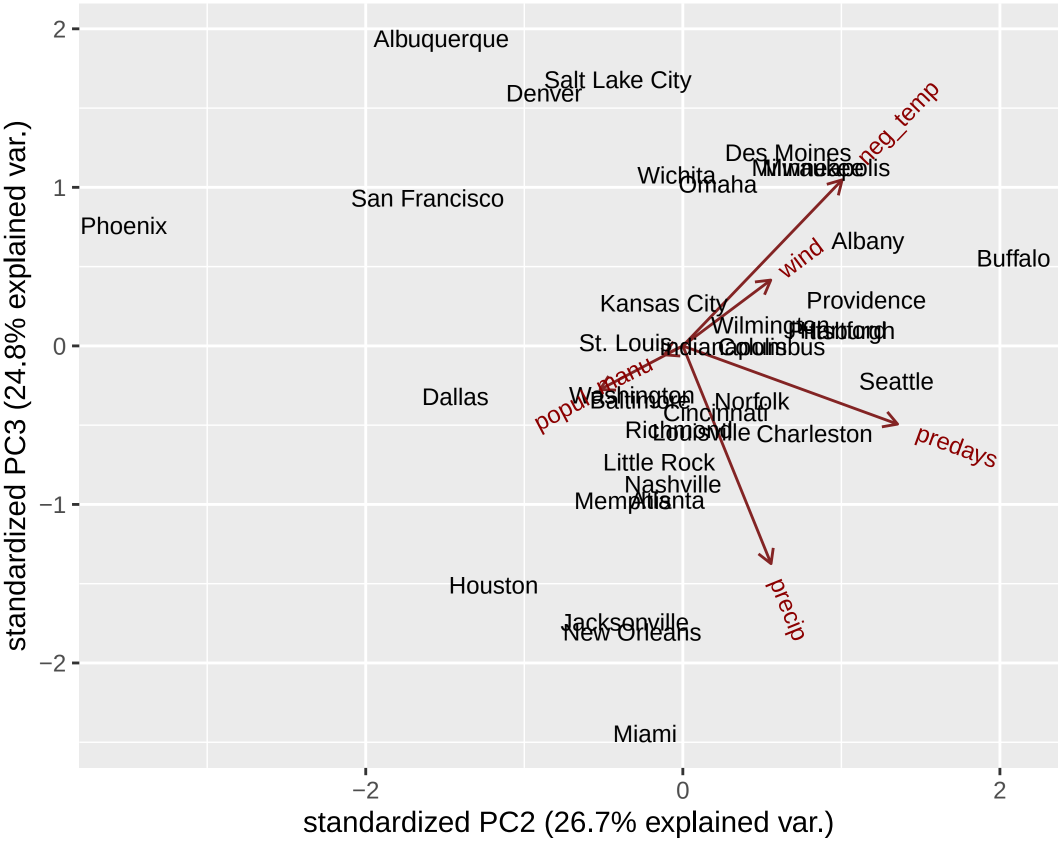


The first biplot of the principal components involves the representation of the variables in the space of best fit, in this case the first two principal components. Here we can see a very interesting picture: The precipitation and (negative) temperature variables are highly correlated with each other, and have the greatest influence on principal component 2. Additionally, the manufacturing and population variables are highly correlated with each other, and have the greatest influence on principal component 1. This is a reflection of what we have seen before in the scatterplot matrices. Interestingly, the wind variable's vector seems to to be traveling right between the other two groups, with perhaps slighly more influence on principal component 1, but implying less of an association with the other directions overall.

Now the other two biplots:



Component 1 against component 3 shows a similar distribution to the first plot, but the precipitation variables have now flipped and have a negative contribution to component 3.



Component 2 against component 3 shows the expected contribution of the variables that can be seen in the previous plots, where manufacturing, population, and the precipitation variables negatively contribute to component 3 and negative temperature and wind have a positive contribution.

Taken together, these biplots imply the following: - manufacturing and population remain highly correlated across all principal component combinations - precipitation variables also maintain a correlation to each other, but to different extents - the contributions of negative temperature to components 2 and 3 can flip in direction - in all cases, variables have a positive contribution to component 1

Since component 1 and component 2 make the highest contributions to the variance in the data, and manufacturing seems to have the highest influence on component 1 and days of precipitation have the highest influence on component 2, we are led to believe that these two variables may hold the highest importance in explaining the variance in the data, and by extension could have important contributions to air pollution. In a future section, we will also use all of the principal components in a multiple linear regression model to see what contributions they make when air pollution is explicitly assumed to be an outcome variable.

Ranking of cities with principal component scores

Cities are ranked in descending order according to their principal component (PC) ranking in the results. The PC scores are calculated as the sum of each city's PC scores. A higher PC score indicates that a city such as Minneapolis and Milwaukee has higher values on the variables that contribute most to the principal components, and have higher levels of the variables associated with air pollution (such as manufacturing, population, etc):

```
##           City PC_Score
## 22 Minneapolis 5.776278
## 21   Milwaukee 4.300699
## 34   St. Louis 2.897588
## 5      Buffalo 2.754176
## 12    Hartford 2.652352
## 4    Baltimore 2.356852
```

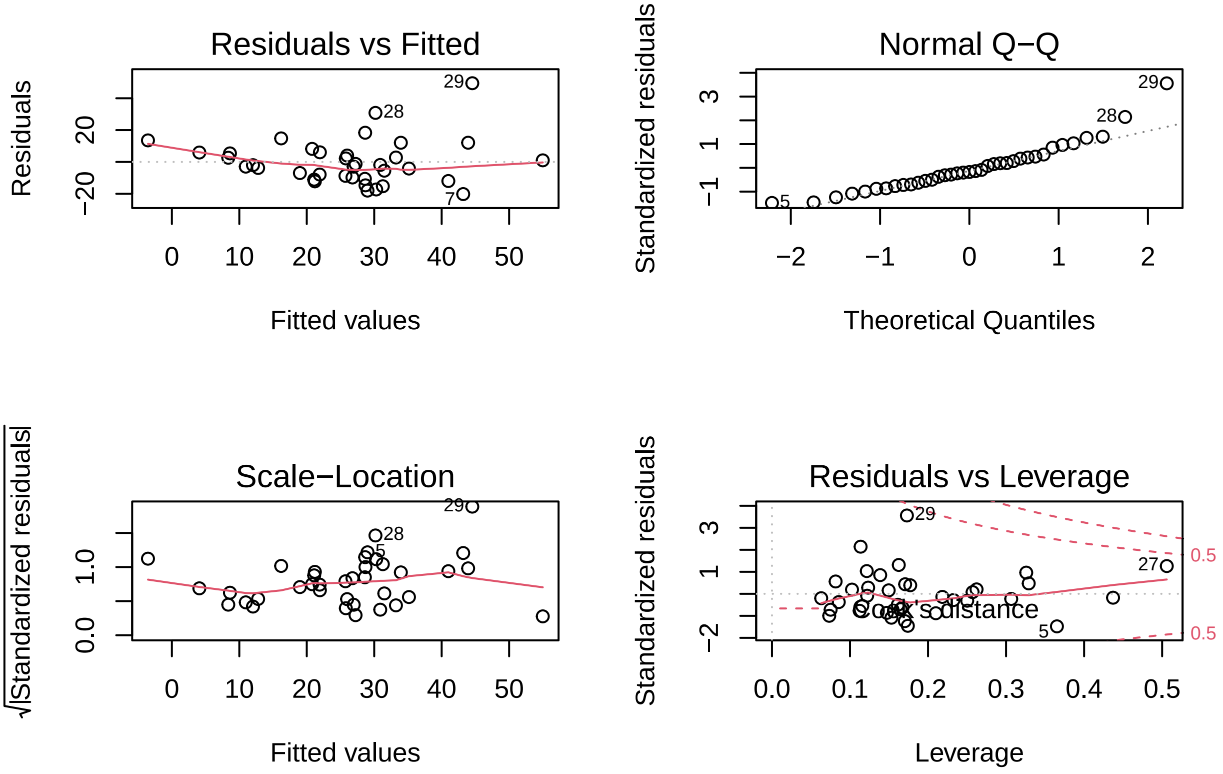
The lower-ranked cities such as Miami and Phoenix have lower scores, indicating they have lower levels of the variables associated with air pollution.

```
##            City  PC_Score
## 20        Miami -5.954391
## 27      Phoenix -4.737998
## 15 Jacksonville -4.344820
## 24  New Orleans -3.869136
## 17  Little Rock -2.587435
## 25      Norfolk -2.290092
```

Multiple linear regression with principal components

The last analysis we will perform is with the use of our principal components in a multiple linear regression model. Everitt and Hothorn (2011) suggest that the use of principal components is advantageous over a direct fitting of the model to the underlying data, because the components are mathematically independent in a way which the raw data does not guarantee (and indeed, a number of the raw variables are correlated with each other). They further suggest that MLR be fitted to all principal components, instead of the subset of selected components, because some of the discarded components may have an unexpected significant contribution to the model. Additionally, the nature of linear regression should assign components with no real contribution very low coefficients.

```
## 
## Call:
## lm(formula = data$SO2 ~ PCA$scores)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -20.195  -9.775  -2.026   5.923  49.457 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       25.7568     2.5132  10.248 2.58e-11 ***
## PCA$scoresComp.1   0.1688     1.8067   0.093   0.9262    
## PCA$scoresComp.2   6.3679     1.9871   3.205   0.0032 ** 
## PCA$scoresComp.3   1.2026     2.0623   0.583   0.5642    
## PCA$scoresComp.4   6.1569     2.9604   2.080   0.0462 *  
## PCA$scoresComp.5  16.1172     6.0692   2.656   0.0126 *  
## PCA$scoresComp.6   3.8860     8.4712   0.459   0.6497    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 15.29 on 30 degrees of freedom
## Multiple R-squared:  0.4254, Adjusted R-squared:  0.3104 
## F-statistic: 3.701 on 6 and 30 DF,  p-value: 0.007157
```



After fitting the model, we see that components two, four, and five are our most significant variables. This is inferred from the low p-values of 0.0032, 0.0462, and 0.0126 respectively. Turning to our model's diagnostic plots we see that the residuals are reasonably randomized and the QQ-Norm plot shows a close fit.

Further looking into these components starting with the most significant: two has highest weights in days of precipitation and negative temperature; four's largest weight is with the feature wind; and five has height weights in manufacturing and population. This model seems to suggest that days of precipitation, followed by negative temperatures, are most important predictors of air pollution, with increasing values tracking with worse pollution values. However, this model has a somewhat weak R-squared value of 0.4254, suggesting that these components may not linearly track particularly well with air pollution overall.

Conclusion

Taken together, the present data and analysis methods yield interesting findings about the relationship between sulphur dioxide air pollution and the given weather, population, and manufacturing variables. PCA analysis indicates that manufacturing enterprises with 20 or more workers and annual average days of precipitation are particularly informative with respect to the data set's variability. A multiple linear regression fit of the principal components also implies that days of precipitation, negative temperature, and manufacturing variables are particularly important. In some sense, this corroborates our PCA findings, but the low R-squared of the linear model makes the conclusions suspect. PCA rankings of city performance revealed that Miami and Phoenix attained the highest ranks in terms of low air pollution factors, and Minneapolis and Milwaukee ranked worst.