A Statistical Analysis of US Cities and Identifying Key Predictors of Air Quality

Authors: Hieu (Calvin) Hoang and Vy Vo

UC Davis

STA 135 (Spring Quarter) - Multivariate Data Analysis

Professor Xiucai Ding

June 08, 2025

Background:

Air pollution is a major topic of discussion and a significant problem in today's society. As civilization continues to grow and develop, air pollution levels have worsened significantly over the years. But just how much worse has it gotten? To explore this, we use the dataset **"Air Pollution in U.S. Cities,"** which contains records from 41 cities across the country. The dataset includes seven variables:

- **SO_2**: Annual mean concentration of sulfur dioxide in micrograms per cubic meter, a key indicator of air pollution.
- **Temp**: Average annual temperature in degrees Fahrenheit.
- **Manu**: Number of manufacturing enterprises employing 20 or more workers, reflecting industrial activity.
- **Popul**: Population size in thousands, based on the 1970 U.S. Census.
- **Wind**: Average annual wind speed in miles per hour, influencing pollutant dispersion.
- **Precip**: Average annual precipitation in inches, affecting pollutant removal from the atmosphere.
- **Predays**: Average number of days with precipitation per year, also impacting air quality.

Using this dataset, our objective is to utilize statistical methods such as PCA, LDA, and QDA to explore correlations and better understand the data. With PCA, we aim to identify patterns and determine the variables most strongly associated with air pollution levels. These selected variables are then evaluated using LDA and QDA to assess their predictive accuracy in classifying pollution levels.

Visualization Analysis:

In figure 1, the U.S. air pollution dataset contains 41 observations. Before conducting data analysis, we examined the distribution of each variable to inform our choice of statistical methods. Figure 1 displays boxplots for the variables in the dataset. A key insight from these plots is the presence of numerous outliers, particularly in variables such as “SO2”, “manu”, and “popul”, which show multiple data points lying beyond the whiskers. This suggests that the dataset contains a substantial number of outliers, and therefore, when selecting a classification method, it is important to consider one that is less sensitive to outliers, such as LDA.

Data Analysis:

PCA Testing

In the cumulative contribution ratio plot, shown in figure 2, a red dashed line highlights 80% on the vertical axis while blue dots connected by lines rise from PC1 through PC6. The 1st dot

appears at roughly 36%, showing that PC1 alone explains just over one-third of the total variance. The 2nd dot climbs to about 63%, indicating that PC1+PC2 together account for nearly two-thirds of all variability. When the 3rd dot reaches around 85%, it crosses the red dashed line for the first time. This visual cue makes it clear that retaining 3 principal components captures most of the meaningful information, with any additional components contributing only a small amount of extra variation.

In the scree plot, shown in figure 3, green dots mark the eigenvalues of each principal component in order, and an orange dashed line is drawn at an eigenvalue of 1. The 1st green dot sits above 2, showing that PC1's eigenvalue is well above average. The 2nd dot appears near 1.5, also above the orange line, and the 3rd lands just above 1. After that, the 4th dot falls below 1, hovering near 0.8, while the 5th and 6th dots drop close to zero. The sharp bend where the 3rd green dot meets the orange line signals an "elbow," indicating that only the first 3 components exceed the usual eigenvalue threshold. In other words, the scree plot confirms that selecting 3 components is the best choice to capture almost all variance without including components that add only negligible information.

As shown in figure 4, the largest values appear on manu (-0.612) and popul (-0.578) for PC1, which means PC1 measures how industrial or urban a city is, with cities that have many factories and large populations lying at one extreme of PC1. Wind (-0.354) also contributes, suggesting that more industrial cities in our sample tend to have slightly lower wind speeds. The fact that these numbers are negative simply reflects how R oriented the eigenvector, and what matters most is their magnitude. Looking at PC2, precip (0.623) and predays (0.708) stand out, indicating that PC2 is driven by wetness, with cities that have heavy rainfall and many rainy days scoring highly. The other variables, temperature, wind, manufacturing, and population, have much smaller influence here, showing that climate factors dominate this component. For PC3, the strongest loadings are on temperature (-0.672) and precipitation (-0.505), with a smaller positive loading on wind (0.297). In simple terms, PC3 separates "hot and rainy" cities, which score negatively, from "cool and windy" ones, which score positively, so it adds a useful climate nuance that PC1 and PC2 do not fully capture. Altogether, PC1–PC3 capture about 71 %, so a fourth component (PC4) is required to hit 80 %, but those three already cover most of the pattern, which means we can reduce six variables to three main dimensions, industrial and urban intensity, wetness, and a temperature versus wind contrast, without losing much information.

In the bubble heatmap, shown in figure 5, each row corresponds to one principal component and each column to one of the six original variables. Darker red circles show strong positive loadings, and darker blue circles show strong negative loadings. Circle size increases with the absolute value of the loading. In the PC1 row, the largest blue circles appear under manu and popul, confirming that PC1 is driven by manufacturing activity and population size, an industrial

and urban axis. In the PC2 row, large red circles under precip and predays emphasize PC2's role as the wetness and climate axis. In the PC3 row, a large blue circle under temp and a smaller blue circle under precip indicate that PC3 contrasts hot and rainy conditions with windy conditions, which would have a positive loading on wind. This visual makes it easy to see which original variables matter most for each component and whether that influence is positive or negative. Because PC1 through PC3 capture most of the variance, their rows have the most prominent circles, while PC4 through PC6 show smaller circles and therefore contribute only minor additional information.

LDA VS QDA Testing

Based on the loadings from PC1, PC2, and PC3, we selected the two most strongly correlated variables from PC1 and PC2, and the highest contributing variable from PC3. We wanted to ensure that all PCs are accounted for based on their relevance. As a result, the chosen variables for further analysis were "manu," "popul," "precip," "predays," and "temp." To apply LDA and QDA, we needed categorical groupings, so we divided the forty cities into three groups based on their SO_2 levels, a key indicator of air pollution. These groups were labeled "low," "medium," and "high" (see Figure 7 for the distribution of these groups).

Using these groupings, we applied LDA and QDA with the selected variables. While our earlier data exploration revealed many outliers, suggesting that LDA might be more appropriate due to its robustness to outliers and smaller sample sizes, we proceeded to compare both methods. We randomly assigned 80% of the data to training and used the remaining 20% for testing.

The results of 77.78 % for LDA and 55.56 % for QDA showed that LDA outperformed QDA in classification accuracy, indicating that a linear boundary is more suitable for distinguishing between pollution level categories in this dataset. This outcome is consistent with our expectations, as QDA is more sensitive to outliers due to its use of separate covariance matrices for each class. In contrast, LDA is less affected by such data irregularities. Using the selected variables (manu, popul, precip, and predays), LDA achieved approximately 77.78% accuracy in predicting pollution levels. This relatively high accuracy suggests that the chosen variables are strong indicators of SO_2 concentration levels.

For comparison, we also explored an alternative variable selection method by choosing the two strongest variables from PC1, since it had the highest correlation with SO2, and one variable each from PC2 and PC3. This resulted in the selection of "manu," "predays," "popul," and "precip." We applied the same methodology as before, using 80% of the data for training and 20% for testing.

With this variable set, LDA achieved an accuracy of 66.67%, while QDA reached 55.56%. Once again, LDA proved to be the better classifier, which aligns with our earlier observation that

variables like "manu" and "popul" contain many outliers, making QDA less suitable due to its sensitivity to data distribution. Although an accuracy of 66.67% is still relatively high, it is lower than the 77.78% achieved with our original variable selection. This confirms that our initial PCA-based selection provided stronger predictors for SO2 levels.

Conclusion:

Using statistical methods such as PCA, LDA, and QDA, we found that the variables "manu," "popul," "precip," "predays," and "temp" together are strong predictors of SO_2 levels for this datset. The LDA model demonstrated a classification accuracy of 77.78%, indicating a solid performance in predicting pollution levels. While this result is promising, there are several areas for improvement in future work. First, the dataset contains only 41 observations, which is relatively small by modern data standards and limits the generalizability of our findings. A larger dataset would allow for more robust training and testing. Additionally, the data is outdated, as it was collected in 1970. Given the significant changes in population, industrial activity, and environmental policies since then, these variables may no longer hold the same predictive power today. Therefore, our analysis should be viewed as a demonstration specific to this dataset, and further research using current data is necessary to validate its relevance in today's context.

Group Discussion

Background: *Vy Vo*
Visualization (Code and Analysis): *Calvin Hoang, Vy Vo*
Data Analysis:

- PCA (Code and Analysis): *Calvin Hoang*
- LDA and QDA (Code and Analysis): *Vy Vo*

Conclusion: *Vy Vo*
Appendix: *Calvin Hoang*

Appendix:

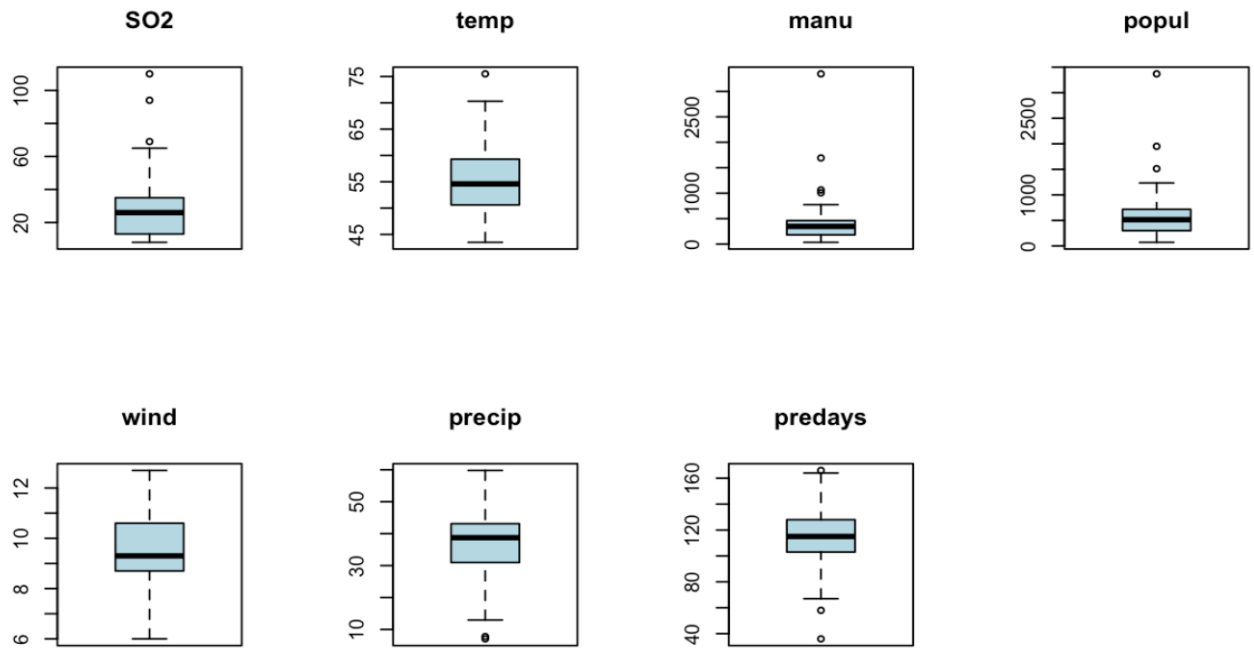

Figure 1: Boxplot Visualization of Air Pollution Dataset Variables

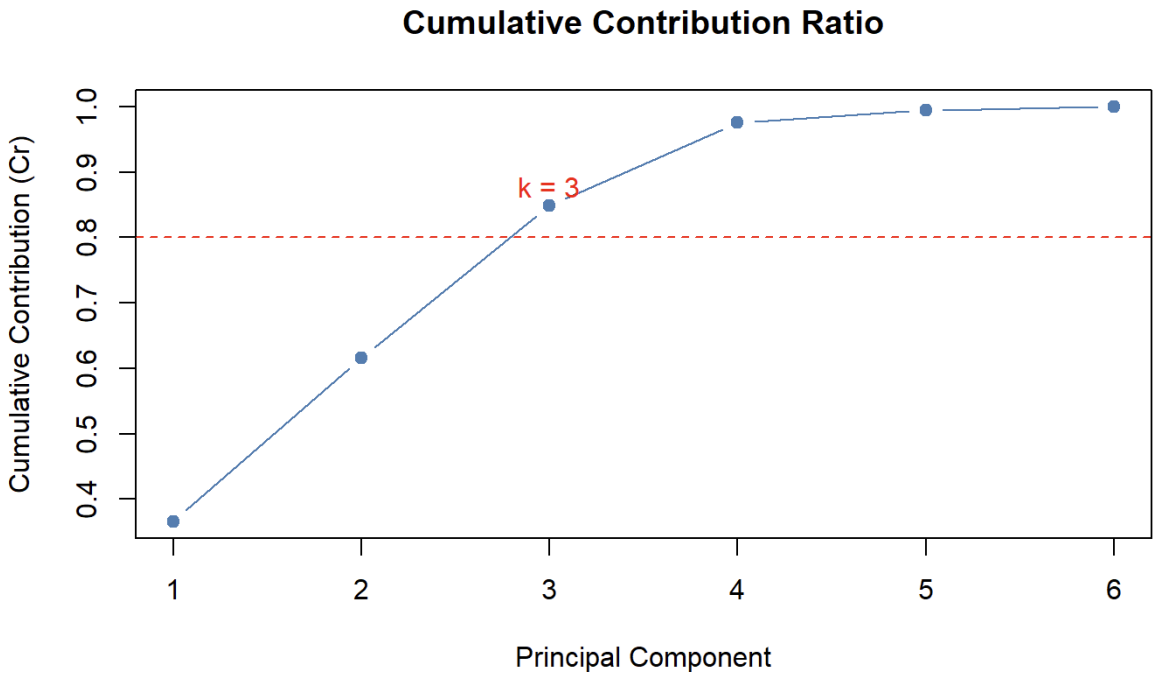

Figure 2: Cumulative Contribution Ratio Plot

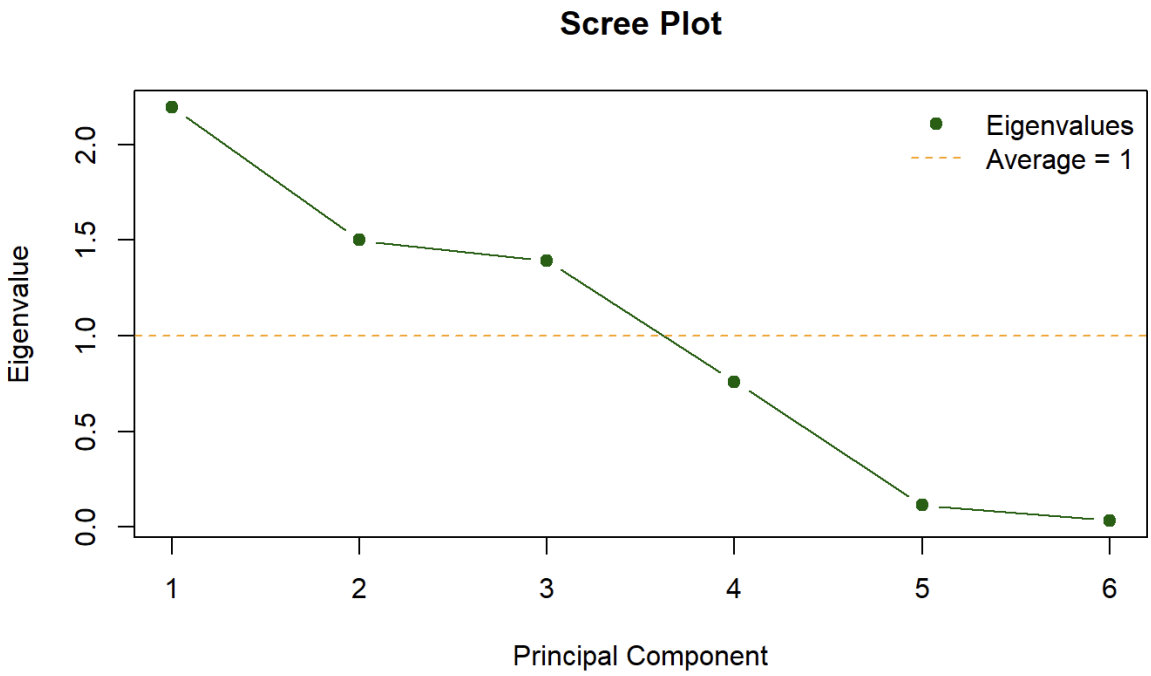

Figure 3: Scree Plot

	PC1	PC2	PC3	PC4	PC5	PC6
temp	0.330	-0.128	-0.672	-0.306	0.558	0.136
wind	-0.354	0.131	0.297	-0.869	0.113	0.025
precip	0.041	0.623	-0.505	-0.171	-0.568	-0.061
predays	-0.238	0.708	0.093	0.311	0.580	0.022
manu	-0.612	-0.168	-0.273	0.137	-0.102	0.703
popul	-0.578	-0.222	-0.350	0.072	0.078	-0.695

Figure 4: PCA Loadings Table

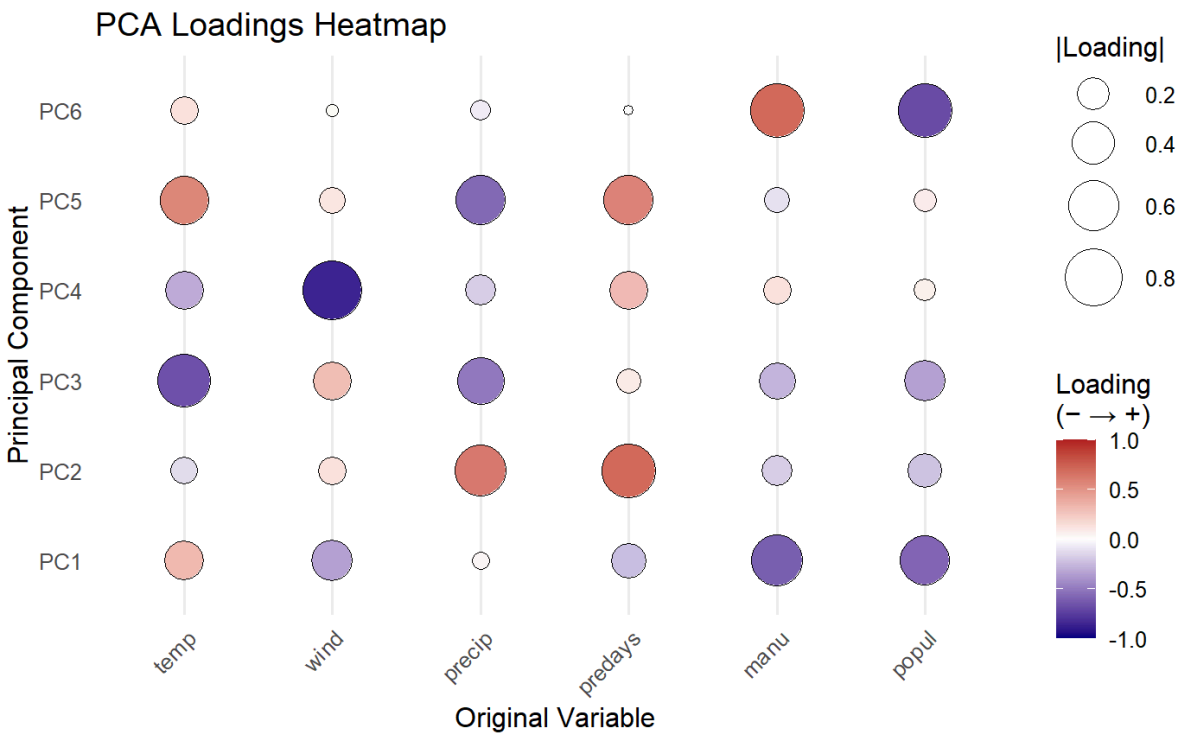

Figure 5: Bubble‑style heatmap of PCA loadings. Each circle's color indicates the sign and magnitude of the loading (blue = negative, red = positive), and its size reflects the absolute value |loading|. Rows correspond to principal components PC1–PC6; columns correspond to original variables (temp, wind, precip, predays, manu, popul).

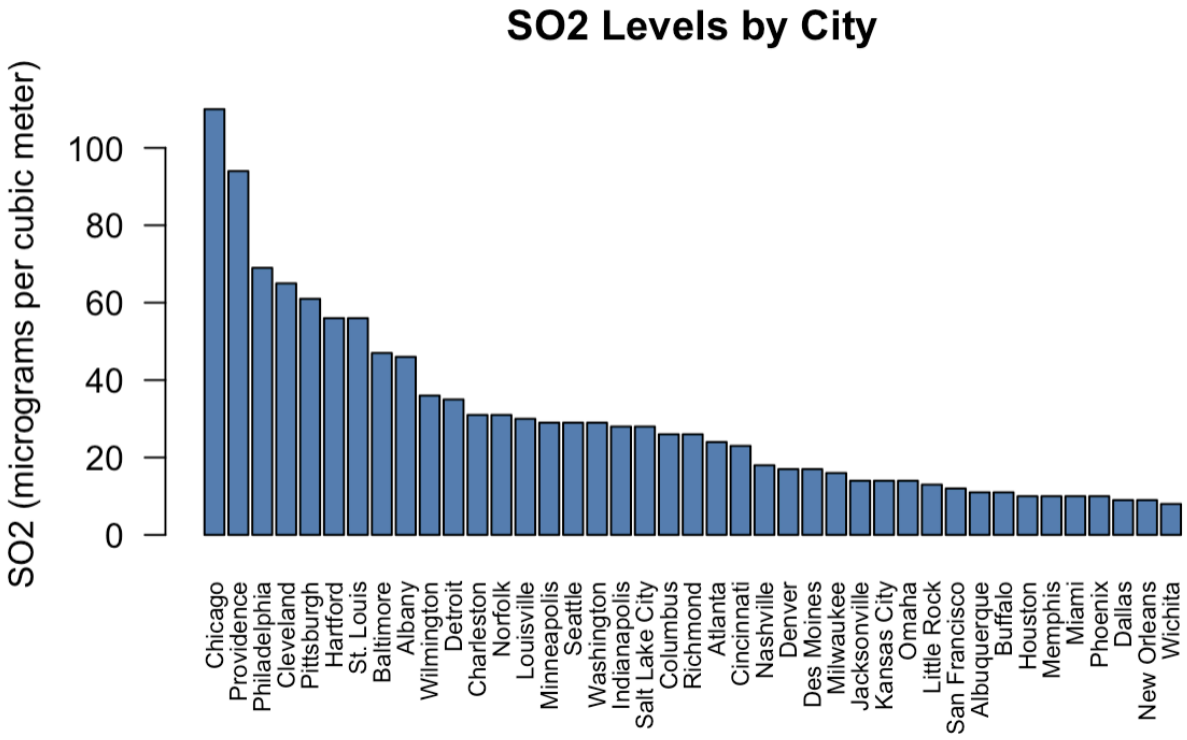

Figure 6: Histogram distribution of SO2 Levels by City

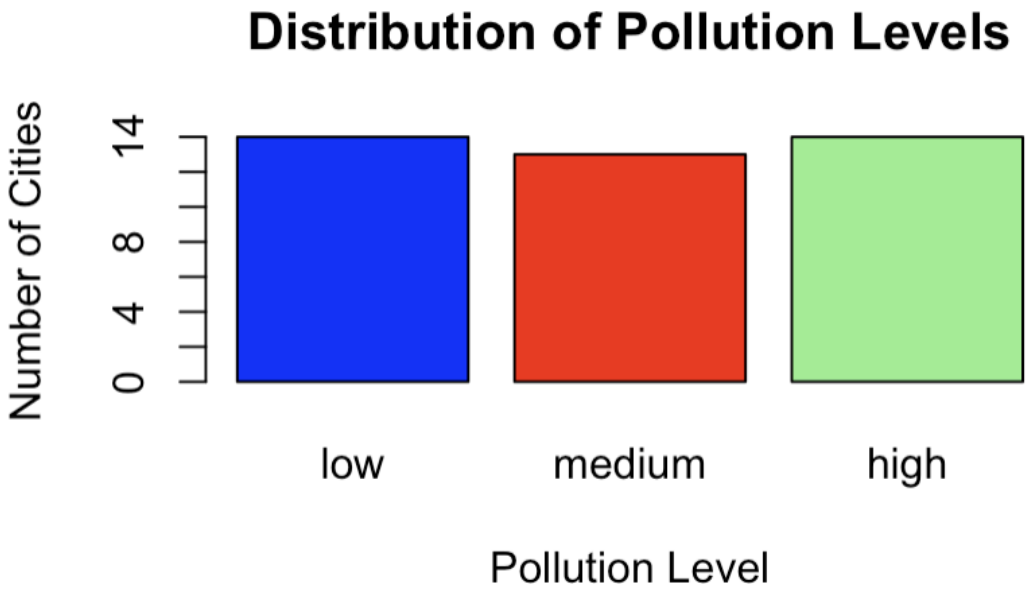

Figure 7: Histogram Distribution of Pollution Levels