Matej Popovski

STAT 436

Homework 5: Model Visualization Portfolio Reflection

**PCA of NYC Neighborhoods Based on Air Quality**

**Essential Question**

This project explores: How do air quality profiles differ across NYC neighborhoods, and can dimensionality reduction like PCA reveal latent structure?

Air quality is multivariate—each area’s environment is influenced by pollutants like NO₂, PM2.5, CO, ozone and other health related metrics. Rather than analyzing pollutants individually, PCA compresses these features into principal components that expose underlying pollution patterns across neighborhoods.

**Dataset and Relevance**

The data comes from a publicly available NYC air quality monitoring project, containing ~19,000 observations. After cleaning and reshaping data, each row represented a neighborhood, and each column captured the mean value of a pollutant like NO₂ or PM2.5 (and other 18 columns).

This domain is socially important: air pollution impacts public health, urban planning, and environmental justice. Understanding spatial pollution patterns can inform both citizens and policymakers.

**Design Choices**

I chose PCA because it is interpretable and produces low-dimensional, visual summaries. Before applying PCA, I normalized pollutant values to ensure equal contribution across features. Without normalization, variables measured on larger numeric scales could dominate PCA directions.

For the primary visualization, each point represents a NYC neighborhood plotted by PC1 and PC2 coordinates. I used geom\_text\_repel() to avoid label overlap and highlight neighborhood structure.

The x- and y-axes represent directions of maximum variance, explaining a substantial portion of the data’s structure.

I created a standard scree plot to visualize the proportion of variance explained across principal components. The scree plot showed that the first few components captured most of the variability, supporting the decision to focus on the first two principal components for visualization.

To deepen model interpretation, I also generated a PCA biplot, overlaying pollutant loadings onto the neighborhood scatterplot. This revealed which pollutants (e.g., NO₂, PM2.5) contributed most to variance in certain areas. Although some label overlap remained, the biplot added insight into how pollutant effects align spatially.

The arrows represent pollutant loadings, with the direction showing which neighborhoods are associated with higher values of that pollutant and the length indicating the strength of the influence. The circle provides a reference for the maximum variance captured by the principal components.

Finally, I introduced a clustered PCA biplot. Using k-means clustering (k=3) on the first two principal components, I grouped neighborhoods into clusters and color-coded them. This made latent pollution groupings more visible, helping uncover broader air quality patterns without relying on dozens of individual labels.

**Key Findings**

The PCA scatterplot revealed that:

* Harlem and Bronx neighborhoods, such as East Harlem and Hunts Point–Mott Haven, separated clearly from most other areas, suggesting distinct air quality profiles likely driven by higher concentrations of key pollutants.
* Manhattan neighborhoods, including Union Square and the Upper East Side, clustered tightly together, indicating similar pollution characteristics, particularly elevated levels of NO₂, PM2.5, annual vehicle miles traveled, and outdoor air toxics.
* Staten Island and southern Queens neighborhoods grouped in the upper-right quadrant, reflecting more suburban environments and generally lower overall pollution levels compared to other parts of the city.The scree plot showed that the first two principal components captured enough variance to justify focusing on 2D analysis.
* The directions and magnitudes of the pollutant vectors suggest that traffic-related emissions (e.g., NO₂, PM2.5) and health outcomes (e.g., asthma hospitalizations) form distinct but intersecting patterns of variation, indicating that areas with higher pollution exposure also experience greater health risks.

The PCA biplot highlighted pollutants such as NO₂ and PM2.5 as key contributors to variance, aligning with known relationships between traffic, industrial zones, and air quality.

The clustered biplot further refined this view, grouping neighborhoods into three major profiles—likely corresponding to high-pollution urban centers, moderately polluted transitional zones, and lower-pollution suburban areas.

These patterns validate PCA’s ability to compress complex environmental data into interpretable dimensions, revealing meaningful spatial structures in urban air quality.

**Data Processing Steps**

The original data was reshaped from long format to wide format using pivot\_wider(), turning pollutant names into columns and aggregating by mean.

I normalized features using step\_normalize() and filtered out missing data before running PCA with step\_pca().

Later, I separately applied prcomp() on the final numeric matrix to support biplot construction and clustering analysis.

**Conclusion**

This project demonstrates how PCA can reveal hidden structure in multivariate air quality data. Dimensionality reduction exposed geographic pollution patterns that would be difficult to observe otherwise.

By combining a traditional scatterplot, a biplot with feature loadings, and a clustered biplot based on k-means, I uncovered nuanced patterns of pollution concentration across NYC neighborhoods.

Together, these visualizations offer a clear, data-driven basis for understanding urban environmental variability and can support public health and urban planning initiatives.