Formative Assignment: Summarising Multivariate Data and PCA

240609924

1. Introduction

This report analyses the airpollution dataset, which contains measurements from 80 US cities in 1960 across 11 variables, including air pollution concentrations and demographic indicators. The objective of this analysis is to provide numerical and graphical summaries of the data, compute key statistical measures such as total variation and generalised variance, and explore the underlying structure of the dataset using Principal Component Analysis (PCA). By systematically evaluating the data, the report aims to highlight key patterns and relationships among variables, assess the suitability of standardisation for PCA, and identify the most influential principal components that encapsulate the dataset's variability. These insights provide a foundation for interpreting environmental and demographic trends while demonstrating effective application of multivariate techniques.

2. Exploratory Data Analysis (EDA)

In this part, we aim to understand the structure of the airpollution dataset by generating numerical summaries and visualising the relationships among variables. We will calculate key statistics such as means, standard deviations, and ranges for each variable and assess their distributions using histograms and scatterplot matrices.

2.1 Numerical Summaries

Table 1: Summary Statistics for Air Pollution Variables

	Mean	SD	Min	Max
SMIN	47.10000	30.218445	1.0000	155.0000
SMEAN	99.65000	50.427564	26.0000	283.0000
SMAX	219.87500	120.038957	58.0000	940.0000
PMIN	44.50000	18.380645	10.0000	98.0000
PMEAN	116.72500	38.837531	54.0000	247.0000
PMAX	275.53750	159.099041	117.0000	978.0000
PM2	72.85875	154.661688	1.6000	1357.2000
PERWH	87.25750	10.383687	60.0000	99.7000
NONPOOR	81.82875	6.741862	67.8000	93.2000
GE65	85.87500	21.574349	45.0000	171.0000
LPOP	56.55078	3.854450	49.3739	67.9385

Table 1 summarises the key statistics for the air pollution variables, which we calculated using functions like `colMeans()`, `apply()`, and `summary()` to understand the central tendency, variability, and range of each variable across all cities. The table includes variables such as sulphate concentrations (SMIN, SMEAN, SMAX), particulate matter levels (PMIN, PMEAN, PMAX), and demographic indicators (e.g., PERWH, NONPOOR, GE65). From our calculations, we observed that PMAX has the highest variability, with a

standard deviation of 159.09, while LPOP shows the lowest variability, reflecting a more uniform distribution of population logs. These statistics allowed us to identify notable trends, such as significant differences in air pollution levels and demographic distributions, which are critical for our subsequent exploratory and principal component analyses.

Table 2: Air Pollution Data by City (First 10 Rows)

	SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
PROVIL	30	163	349	56	119	223	116.1	97.9	83.9	109	58.5645
JACKSON1	29	70	161	27	74	124	21.3	60.0	69.1	64	52.7195
JOHNST	88	123	245	70	166	452	15.8	98.7	73.3	103	54.4829
JERSEYC	155	229	340	63	147	253	1357.2	93.1	87.3	103	57.8585
HUNTIN	60	70	137	56	122	219	18.1	97.0	73.2	93	54.0617
DESMOIN	31	88	188	61	183	329	44.8	95.9	87.1	97	54.2540
DENVE	2	61	188	54	126	229	25.4	95.8	86.9	82	59.6819
READING	50	94	186	34	120	242	31.9	98.2	86.1	112	54.3999
TOLED	67	86	309	52	104	193	133.2	90.5	86.1	98	56.5985
FRESNO	18	34	198	45	119	304	6.1	92.5	78.5	81	55.6342

Table 2 showcases the air pollution and demographic data for the first 10 cities, which we extracted using the `head()` function to gain a quick overview of the dataset. By including these variables, we explored the variability across cities; for instance, JERSEYC stands out with the highest PM2 value (1357.2), indicating severe air pollution levels compared to other cities. Additionally, we noticed differences in demographic variables, such as PROVIDEN having one of the highest NONPOOR percentages, at 83.9%. This step helps us understand both environmental and social dimensions of the data, providing a foundation for further analysis.

2.2 Graphical Summaries

Figure 1 displays histograms for all variables in the air pollution dataset, created using the `hist()` function to explore their individual distributions. These histograms reveal key patterns in the data. For example, SMIN, SMEAN, and SMAX display right-skewed distributions, indicating that most cities have lower levels, but a few cities experience significantly higher concentrations. Similarly, PM2 shows a heavily right-skewed pattern, reflecting extreme particulate matter values in certain cities. On the other hand, demographic variables like PERWH and NONPOOR have more uniform or slightly left-skewed distributions, with most cities showing high percentages of white and non-poor populations. Interestingly, GE65 (percentage of population aged 65+) and LPOP (log population) demonstrate relatively symmetric distributions.

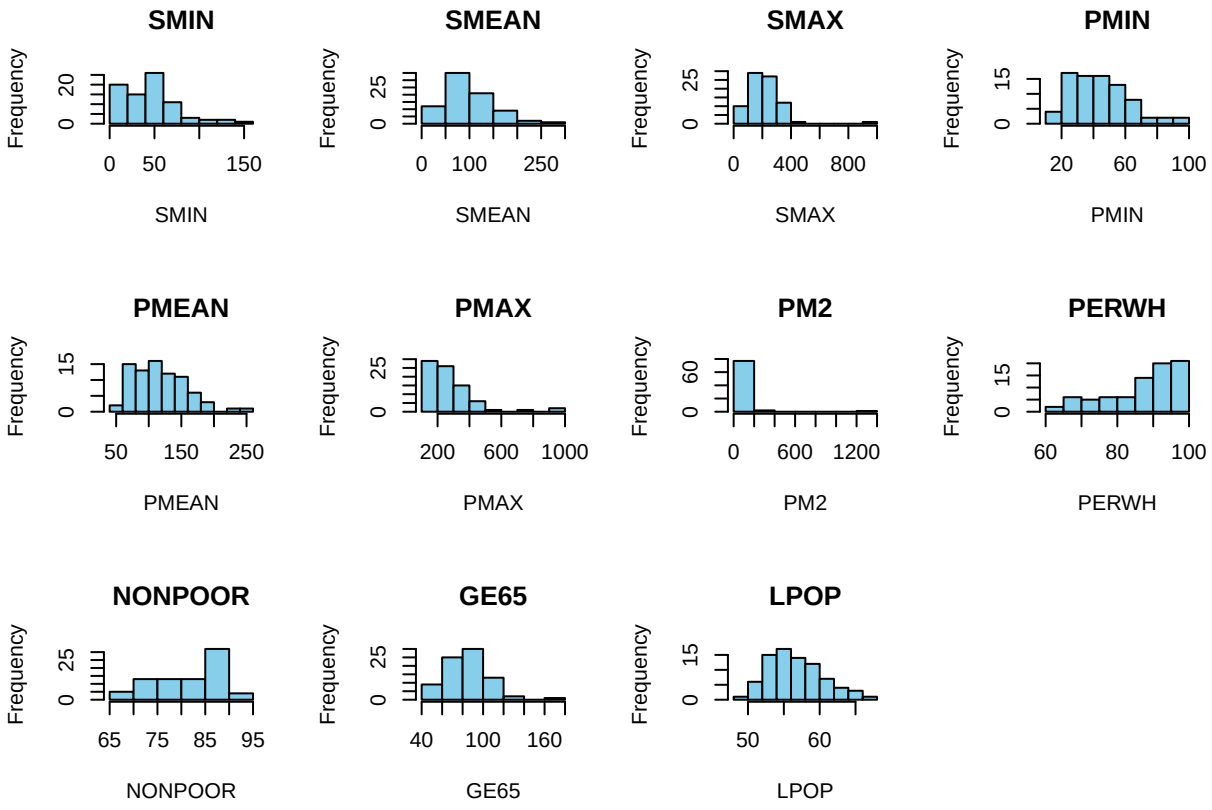

Figure 1: Air Pollution Distribution

Figure 2 displays a scatterplot matrix focusing on sulphate concentrations (SMIN, SMEAN, SMAX) and particulate matter levels (PMIN, PMEAN, PMAX) to investigate relationships between these air pollution variables. The plots reveal strong positive correlations within each category of variables. For example, SMEAN and SMAX, representing average and maximum sulphate concentrations, show a clear linear relationship, suggesting that cities with higher average sulphate levels also experience higher peak values. Similarly, PMEAN and PMAX exhibit a similar pattern for particulate matter. However, the relationships between sulphate concentrations and particulate matter levels are weaker, indicating that while both measure pollution, they may be driven by different sources or environmental factors. This matrix provides a focused view of interactions within and between the key pollution metrics, enhancing our understanding of their distribution and correlation.

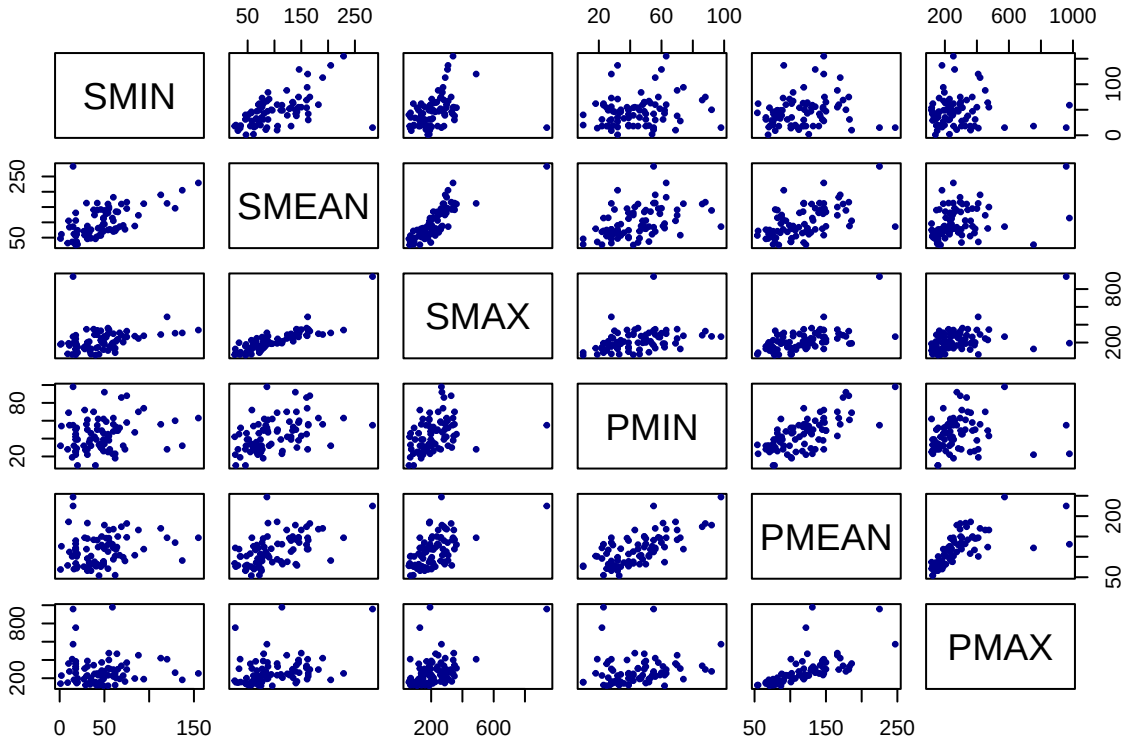

Figure 2: Subset of Air Pollution Variable

Figure 3 presents a correlation heatmap for the air pollution and demographic variables, offering a concise and visually intuitive summary of relationships between variables. The size and colour of the circles indicate the strength and direction of correlations, with darker blue circles representing strong positive correlations, while red circles (not visible here) would represent negative correlations.

From the heatmap, we observe strong positive correlations within the sulphate concentration variables (SMIN, SMEAN, SMAX) and within particulate matter levels (PMIN, PMEAN, PMAX), confirming the findings from the scatterplot matrix. However, it also provides additional clarity about weaker correlations between these two groups of variables. For instance, while PM2 shows moderate correlations with PMAX and PMEAN, its relationship with sulphate variables is weak. Demographic variables like PERWH and NONPOOR exhibit limited correlation with the pollution metrics, except for some mild relationships with LPOP. LPOP itself shows a negative correlation with PERWH, indicating that cities with higher log population tend to have lower percentages of white residents.

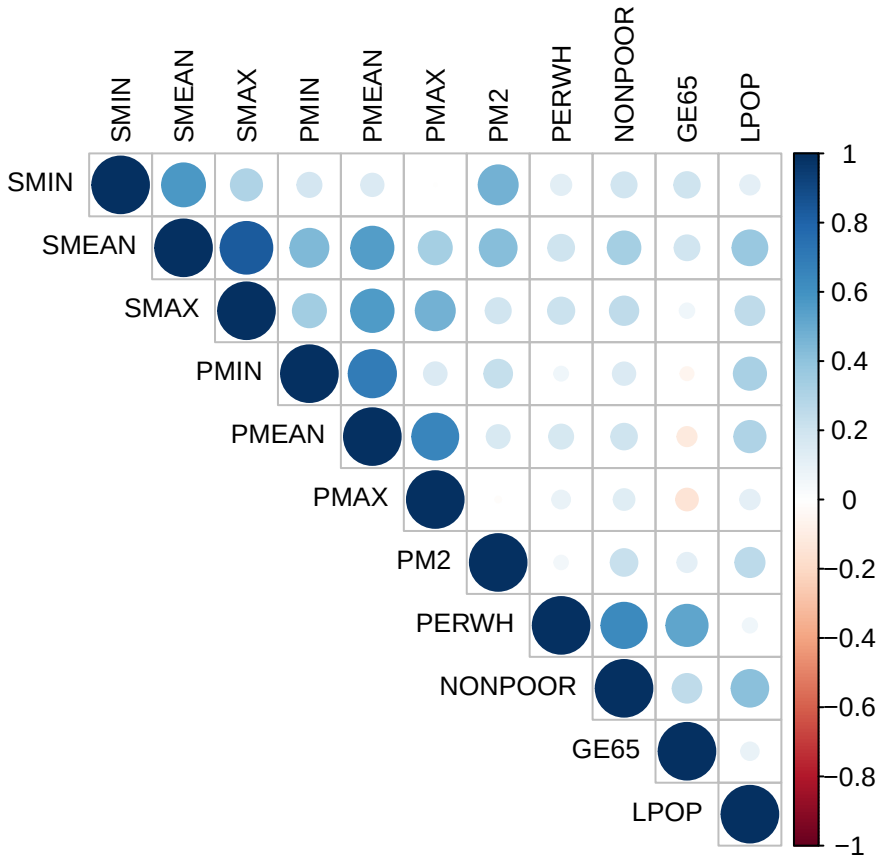

Figure 3: Correlation Heatmap for Air Pollution Variables

3. Statistical Computation and Standardisation

To compute the total variation and generalised variance of the dataset, we first calculated the covariance matrix using the `cov()` function. The total variation, derived as the sum of the diagonal elements of the covariance matrix using `sum(diag(cov_matrix))`, resulted in **69577.97**, reflecting the overall spread of the variables. Next, we computed the generalised variance using the determinant of the covariance matrix with `det(cov_matrix)`, which yielded **8.72131e+29**, indicating substantial multivariate variability. These calculations confirm the high variability present in the dataset, particularly among air pollution variables like PMAX and SMAX, which contribute significantly to the total spread.

Table 3: Total Variation and Generalised Variance

Metric	Value
Total Variation	6.957797e+04
Generalised Variance	8.721310e+29

3.1 Mean Vector

Table 4 presents the mean vector of the variance-scaled data, where each variable was scaled to have unit variance while retaining its original mean. To achieve this, we used the `scale()` function with the `center = FALSE` argument, ensuring that the variables were not mean-centered but were adjusted for variance. This approach allowed us to preserve the original means of the variables, as shown in the table, while ensuring

that all variables contribute equally to subsequent analyses such as PCA. The resulting mean values, such as **0.8379** for SMIN and **0.9915** for LPOP, reflect the original dataset's structure, providing a balance between standardisation and retaining interpretability. This variance-scaling method ensures comparability across variables without losing the context provided by their original averages.

Table 4: Mean Vector of Variance-Scaled Data

	Mean
SMIN	0.8379184
SMEAN	0.8877965
SMAX	0.8734672
PMIN	0.9193029
PMEAN	0.9434948
PMAX	0.8619202
PM2	0.4256757
PERWH	0.9868542
NONPOOR	0.9904164
GE65	0.9641384
LPOP	0.9914587

3.2 Covariance Matrix

Table 5: Covariance Matrix of Air Pollution Variables

	SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
SMIN	913.15	874.74	1097.01	100.23	182.48	-8.26	2212.32	39.99	39.70	131.42	13.54
SMEAN	874.74	2542.94	5036.06	415.33	1084.08	2716.81	3282.94	109.05	112.71	208.39	73.24
SMAX	1097.01	5036.06	14409.35	750.52	2612.75	9048.50	3633.65	266.82	202.48	169.14	118.27
PMIN	100.23	415.33	750.52	337.85	496.20	466.59	681.15	11.50	19.37	-20.81	22.86
PMEAN	182.48	1084.08	2612.75	496.20	1508.35	4056.86	982.00	72.17	53.31	-95.25	45.44
PMAX	-8.26	2716.81	9048.50	466.59	4056.86	25312.50	-246.97	163.84	143.42	-505.02	72.62
PM2	2212.32	3282.94	3633.65	681.15	982.00	-246.97	23920.24	92.02	230.48	383.09	157.77
PERWH	39.99	109.05	266.82	11.50	72.17	163.84	92.02	107.82	44.60	118.32	2.54
NONPO	39.70	112.71	202.48	19.37	53.31	143.42	230.48	44.60	45.45	37.19	10.84
GE65	131.42	208.39	169.14	-20.81	-95.25	-505.02	383.09	118.32	37.19	465.45	7.93
LPOP	13.54	73.24	118.27	22.86	45.44	72.62	157.77	2.54	10.84	7.93	14.86

Table 5 summarises the covariance matrix for the air pollution dataset, calculated using the `cov()` function to measure how variables co-vary. The results were rounded to two decimal places using the `round()` function for better readability. Strong positive covariances, such as 5036.06 between SMEAN (mean sulphate concentration) and SMAX (maximum sulphate concentration), indicate a significant linear relationship. Conversely, near-zero covariances, like -8.26 between SMIN (minimum sulphate concentration) and PMAX (maximum particulate matter level), suggest minimal association. These findings provide valuable insights into how variables vary together, laying the groundwork for further multivariate analysis, such as Principal Component Analysis (PCA).

3.3 Correlation Matrix

Table 6: Correlation Matrix of Air Pollution Variables

	SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
SMIN	1.00	0.57	0.30	0.18	0.16	0.00	0.47	0.13	0.19	0.20	0.12
SMEAN	0.57	1.00	0.83	0.45	0.55	0.34	0.42	0.21	0.33	0.19	0.38
SMAX	0.30	0.83	1.00	0.34	0.56	0.47	0.20	0.21	0.25	0.07	0.26
PMIN	0.18	0.45	0.34	1.00	0.70	0.16	0.24	0.06	0.16	-0.05	0.32
PMEAN	0.16	0.55	0.56	0.70	1.00	0.66	0.16	0.18	0.20	-0.11	0.30
PMAX	0.00	0.34	0.47	0.16	0.66	1.00	-0.01	0.10	0.13	-0.15	0.12
PM2	0.47	0.42	0.20	0.24	0.16	-0.01	1.00	0.06	0.22	0.11	0.26
PERWH	0.13	0.21	0.21	0.06	0.18	0.10	0.06	1.00	0.64	0.53	0.06
NONPO	0.19	0.33	0.25	0.16	0.20	0.13	0.22	0.64	1.00	0.26	0.42
GE65	0.20	0.19	0.07	-0.05	-0.11	-0.15	0.11	0.53	0.26	1.00	0.10
LPOP	0.12	0.38	0.26	0.32	0.30	0.12	0.26	0.06	0.42	0.10	1.00

Table 6 displays the correlation matrix, computed with the `cor()` function to quantify the strength and direction of linear relationships between variables. The matrix was also rounded to two decimal places for clarity. Key findings include strong positive correlations, such as 0.83 between SMEAN and SMAX, and 0.86 between PMEAN and PMAX, reflecting consistent patterns within related variables. Weak or negligible correlations, such as 0.00 between SMIN and PMAX, highlight independent variable behavior. This matrix offers a comprehensive view of variable interdependencies, essential for identifying patterns and relationships in PCA.

To verify the accuracy of the standardisation process, we compare the covariance matrix of the standardised data to the correlation matrix of the original data. Standardisation transforms the data to have a mean of 0 and a standard deviation of 1, effectively removing the influence of units or scales. As a result, the covariance matrix of the standardised data should match the correlation matrix of the original data. To ensure this, we use the `all.equal()` function, which performs an element-wise comparison of the two matrices and returns TRUE if they are identical. This step serves as a validation checkpoint to confirm that the data has been correctly prepared for subsequent analyses, such as PCA.

The output of the code confirms that the covariance matrix of the standardised data is indeed equal to the correlation matrix of the original data, as indicated by the result **TRUE**. This verification demonstrates that the standardisation process was performed correctly, and the data is now in a consistent and comparable format. This alignment ensures the integrity of subsequent multivariate analyses, particularly PCA, which often relies on a correlation or covariance matrix to identify patterns and reduce dimensionality effectively.

4. Principal Component Analysis (PCA)

4.1 Standardisation vs Raw Data for PCA

PCA should be based on the **standardised data** due to the differing units and scales of the variables in the air pollution dataset. The dataset contains variables such as sulphate concentrations (e.g., SMIN, SMEAN, SMAX) measured in micrograms per cubic meter, particulate matter levels (e.g., PMIN, PMEAN, PMAX), and demographic indicators (e.g., PERWH, NONPOOR, GE65) represented as percentages or population measures. These variables are inherently on different scales, where some may exhibit larger ranges or variances compared to others. For example, PM2 (particulate matter levels) might have values spanning thousands, while demographic variables like LPOP (log population) are relatively small.

Standardisation addresses this disparity by transforming all variables to have a mean of 0 and a standard deviation of 1, effectively removing the influence of scale and ensuring each variable contributes equally to the PCA. This is crucial because PCA identifies patterns based on variances; variables with larger variances or units would dominate the principal components if left unstandardised, leading to results biased toward those variables. By standardising, PCA highlights the true relationships between variables, independent of

their scales or units, allowing for a more balanced and meaningful dimensionality reduction. In summary, standardising the data ensures fairness and comparability among variables, enabling PCA to extract components that truly represent the underlying structure of the dataset rather than being skewed by variables with larger numerical ranges or variances.

4.2 Performing PCA and Interpreting the Principal Components

Table 7: Loadings of the First Two Principal Components

	PC1	PC2
SMIN	0.26	0.19
SMEAN	0.45	-0.01
SMAX	0.40	-0.13
PMIN	0.31	-0.23
PMEAN	0.39	-0.34
PMAX	0.25	-0.34
PM2	0.24	0.15
PERWH	0.21	0.46
NONPOOR	0.28	0.37
GE65	0.11	0.54
LPOP	0.27	0.04

Table 7 presents the loadings of the first two principal components (PC1 and PC2) derived from the Principal Component Analysis (PCA) on the air pollution dataset. These loadings represent the contribution of each variable to the respective principal component. **PC1** is primarily associated with sulphate and particulate matter variables (e.g., SMEAN, SMAX, and PMEAN) with relatively high positive loadings (e.g., SMEAN at 0.45 and SMAX at 0.40), indicating that it captures overall pollution levels. In contrast, **PC2** highlights socio-demographic factors, as evidenced by high positive loadings for PERWH (0.46), NONPOOR (0.37), and GE65 (0.54), suggesting it reflects demographic patterns such as population characteristics. Negative loadings for PMIN (-0.23) and PMEAN (-0.34) in PC2 suggest an inverse relationship with certain pollution metrics. These interpretations provide insights into the primary dimensions of variability in the dataset: PC1 summarises pollution intensity, while PC2 captures socio-demographic variability across cities.

4.3 Determining the Number of Principal Components to Retain

Table 8: Explained and Cumulative Variance by Principal Components

	Explained_Variance	Cumulative_Variance
PC1	0.35	0.35
PC2	0.17	0.52
PC3	0.13	0.65
PC4	0.09	0.74
PC5	0.07	0.81
PC6	0.06	0.87
PC7	0.05	0.92
PC8	0.04	0.96
PC9	0.02	0.98
PC10	0.01	0.99
PC11	0.01	1.00

Table 8 summarises the explained variance and cumulative variance for each principal component (PC) derived from the PCA. **PC1** explains the largest proportion of variance (35%), followed by **PC2** (17%), together accounting for 52% of the total variance. By including **PC3** (13%), the cumulative variance increases to 65%, indicating that the first three components capture the majority of the variability in the dataset. Components beyond PC3 contribute less than 10% individually, with diminishing returns on the variance explained. Based on the "elbow rule" and the goal of retaining at least 70-80% of cumulative variance, it is reasonable to recommend using the first three principal components for this analysis. This ensures that most of the dataset's variability is captured while avoiding overfitting by including components that contribute minimally to the variance.

4.4 Visualising the Principal Components

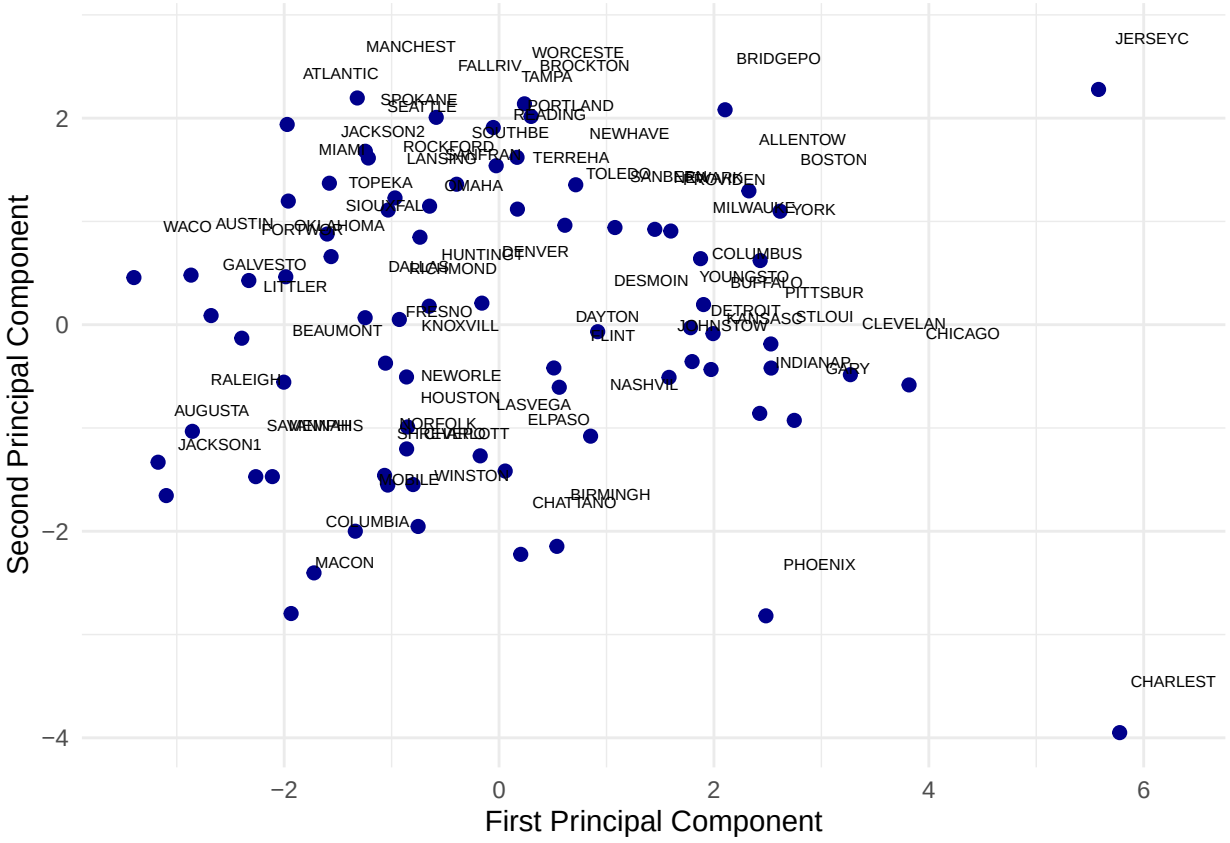

Figure 4: Subset of Air Pollution Variable

Figure 4 illustrates the scatter plot of the first two principal components (PC1 and PC2) for the air pollution dataset, with each point representing a city and labelled accordingly. The x-axis corresponds to **PC1**, which primarily captures overall pollution levels, while the y-axis corresponds to **PC2**, associated with socio-demographic factors. Cities like **JERSEYC** and **CHARLEST** stand out as outliers, with **JERSEYC** having the highest PC1 value, indicating extreme pollution levels, and **CHARLEST** showing a distinctive demographic profile. Most cities cluster near the origin, suggesting similar moderate values for both components.

5. Conclusions

The analysis revealed significant variability in air pollution levels and demographic characteristics across 80 U.S. cities. Sulphate concentrations and particulate matter levels showed substantial differences, with certain cities like JERSEYC exhibiting extreme pollution levels. Socio-demographic factors, such as the percentage of white and non-poor populations, also varied, contributing to distinct profiles for different cities.

Principal Component Analysis (PCA) highlighted two main dimensions of variability: PC1 captured overall pollution intensity, while PC2 reflected socio-demographic characteristics. The first three principal components explained 65% of the dataset's variance, effectively summarising key patterns. The scatter plot of PC1 and PC2 revealed clusters of cities with similar profiles and identified notable outliers, such as JERSEYC and CHARLEST. These findings provide a deeper understanding of the dataset, enabling targeted strategies to address environmental and demographic disparities.