Analyzing New York City Air Quality (2018-2022)

Omid Emamjomehzadeh*

November 20, 2024

Abstract

This study aims to analyze the Air Quality Index (AQI) values in New York State (NYS) from 2018 to 2022 to determine whether there are statistically significant differences between the years or not.To achieve this, various transformations are applied to stabilize the variance and make the distribution more closely approximate a normal distribution. since the analysis of variance detected a difference between the years. Tukey's method was employed to see which pair of years are different. The results show that AQI of 2018 is different from 2020, 2021, and 2022 also AQI of 2019 is different from 2020. To explain these differences, meteorological variables such as temperature, precipitation, snow depth, and wind speed are considered and used to derive some correlation between AQI these variables. Correlation value show positive correlation between temperature and AQI and negative correlation between wind speed and AQI.

1 Introduction

The Air Quality Index (AQI), established by the United States Environmental Protection Agency (USEPA) in 1977, it has become a critical tool for monitoring air pollution and its health implications. Before the AQI, air quality metrics varied significantly across metropolitan areas, leading to

*Department of Civil and Urban Engineering, NYU Tandon School of Engineering, New York, NY

inconsistent communication and standards. The AQI was designed as a dynamic framework that adjusts to regulatory changes while maintaining its goal of public awareness about air quality [2].

The AQI is calculated based on pollutants such as ozone (O_3), particulate matter (PM_{10} and $PM_{2.5}$), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2) using a formula that translates pollutant concentrations into a unit less value ranging from 0 to 500, corresponding to different levels of health concern. For each pollutant, the concentration (C_p) is compared to break-point values (BP_{Lo} and BP_{Hi}), which define ranges for AQI categories. The index is computed as:

$$I_p = \frac{(I_{\mathrm{Hi}} - I_{\mathrm{Lo}})}{(BP_{\mathrm{Hi}} - BP_{\mathrm{Lo}})}(C_p - BP_{\mathrm{Lo}}) + I_{\mathrm{Lo}}$$

where I_{Lo} and I_{Hi} are the lower and upper AQI values for the break-point range, respectively. The AQI for a region is determined by the pollutant with the highest index value for a given day, referred to as the "critical pollutant." This approach reflects the worst-case scenario for air quality on any given day. The AQI uses a color-coded system to enhance public understanding, with values below 50 signifying "Good" air quality and values above 300 indicating "Hazardous" conditions [1].

Sensitive groups, such as children, the elderly, and individuals with preexisting respiratory or cardiovascular conditions, are specifically highlighted in AQI reports when pollutant levels exceed NAAQS standards. This makes the AQI an effective tool for public health communication, as it translates complex scientific data into actionable information. Despite its simplicity, the AQI has limitations, such as focusing on the highest pollutant index, which may overlook cumulative or synergistic effects of multiple pollutants. Nonetheless, the AQI remains an essential framework for informing the public and guiding policy decisions on air quality [2].

This study aims to use AQI data of New York city from 2018 to 2022 to investigate any significant difference of AQI values between these years, and provide some justification for this differences using the meteorological data. The result of this study will enhance our understanding about AQI in New York city and meteorological forgings correlation with AQI.

The rest of the paper is organized as follows; section 2 (method), explains transforming the data . Then analysis of variance and Tukey's tests used to identify the difference between years. At the end, the correlation analysis was done between AQI and meteorological variables. Section 3 (Results), presents the results of applying the proposed methodology to the dataset. Section 4, provides concluding remarks and during the discussion explains limitation of these approaches which set the stage for future research.

2 Method

At the outset, basic statistics for the AQI values of each year in the dataset are calculated (including sample mean and sample variance). Since many statistical techniques require the assumption of homogeneity of variance, we perform Levene's test to assess whether the variance is significantly different between the years or not. Additionally, distribution normality is a required condition of the methods like ANOVA and Tukey's test which are used in this study, so we apply the Shapiro-Wilk test to check whether the data are normally distributed.

Since environmental data are rarely distributed normally and have homogeneous variance, various transformations are investigated to improve normality and homogeneity of variance. Transformations such as logarithmic, square root, and Box-Cox, Yeo-Johnson, Reciprocal, inverse square root, Arc sin square root and Z=score were applied. These transformations are explained in 4.

Then, the best transformation that can make the variance stabilize and make the distribution normally distributed is selected and applied to data. Then the mean and variance of each year is calculated based on transformed values. After all analysis of variance is used to investigate if there is any significant difference between the years.

2.1 Analysis of Variance (ANOVA)

Analysis of Variance is a statistical method used to test whether there are significant differences between the means of multiple groups. The goal of ANOVA is to determine if the differences between group means are larger than what could be expected by random chance. This is done using the **F-ratio**, which is calculated in formula 1.

$$F = \frac{\text{MSB}}{\text{MSW}} \tag{1}$$

Here, the Mean Square Between Groups (MSB) and Mean Square Within Groups (MSW) are defined in formula 2 :

$$\text{MSB} = \frac{\text{SSB}}{\text{dfB}}, \quad \text{MSW} = \frac{\text{SSW}}{\text{dfW}} \tag{2}$$

Where SSB refers to the Sum of Squares Between Groups, which measures how much the group means differ from the overall mean; SSW stands for the Sum of Squares Within Groups, indicating the variability of data points within each group; dfB represents the Degrees of Freedom Between Groups, calculated as $k - 1$, where k is the number of groups; and dfW denotes the Degrees of Freedom Within Groups, which is calculated as $N - k$, where N is the total number of observations.

The resulting F-statistic is compared to a critical value from the F-distribution table to determine whether the group means are significantly different. If F is large, it indicates that the between-group variability is much greater than the within-group variability, suggesting significant differences between the means. If ANOVA detects a difference, then Tukey's Test could be used to investigate exactly which two groups are different.

2.2 Tukey's Test

The goal of Tukey's Test is to identify which pairs of group means are significantly different from each other. Tukey's Test involves several steps. First, the Mean Difference is calculated by computing the absolute difference between the means of each pair of groups. Next, the Standard Error (SE) is calculated. Then, Tukey's Significant Difference is computed using the formula $HSD = q \cdot SE$, where q is the Studentized range statistic, obtained from Tukey's distribution table for the given degrees of freedom and number of groups, and SE is the Standard Error. Finally, the Mean Differences are compared to the HSD. If $|\bar{x}_i - \bar{x}_j| > HSD$, the difference between groups i and j is considered significant.

3 Results

The data for this analysis consists of the daily AQI values in New York city between 2018 and 2022. The meteorological variables considered include wind speed (m/s), precipitation (inches), snow (inches), and average temperature (F). Figure 1 illustrates the distribution of AQI values and the meteorological variables for the years 2018-2021.

After calculating the basic statistics for AQI each year, the following summary statistics were obtained (Table 3):

Year	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
2018	365	58.93	24.73	25	43	53	67	210
2019	365	56.85	21.48	29	42	51	64	150
2020	366	52.66	18.82	26	40	47	59	140
2021	365	54.81	23.06	22	41	49	60	154
2022	365	54.36	21.64	26	41	49	61	234

Table 1: Summary statistics of AQI values for each year (2018-2022)

Histograms of the AQI data for each year were examined. These histograms clearly show that the data are not normally distributed. Figure 3 displays the histograms for AQI values from 2018 to 2022.

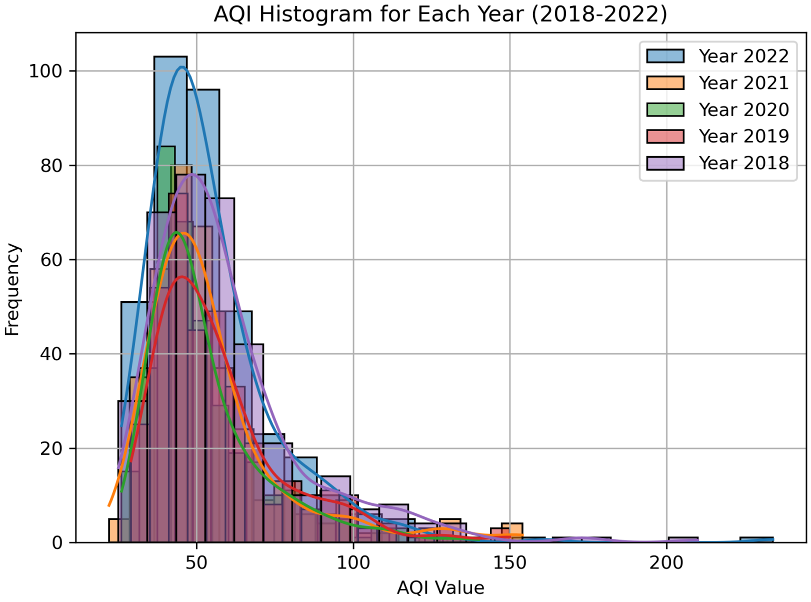

Figure 2: Histograms of AQI values for each year (2018-2022)

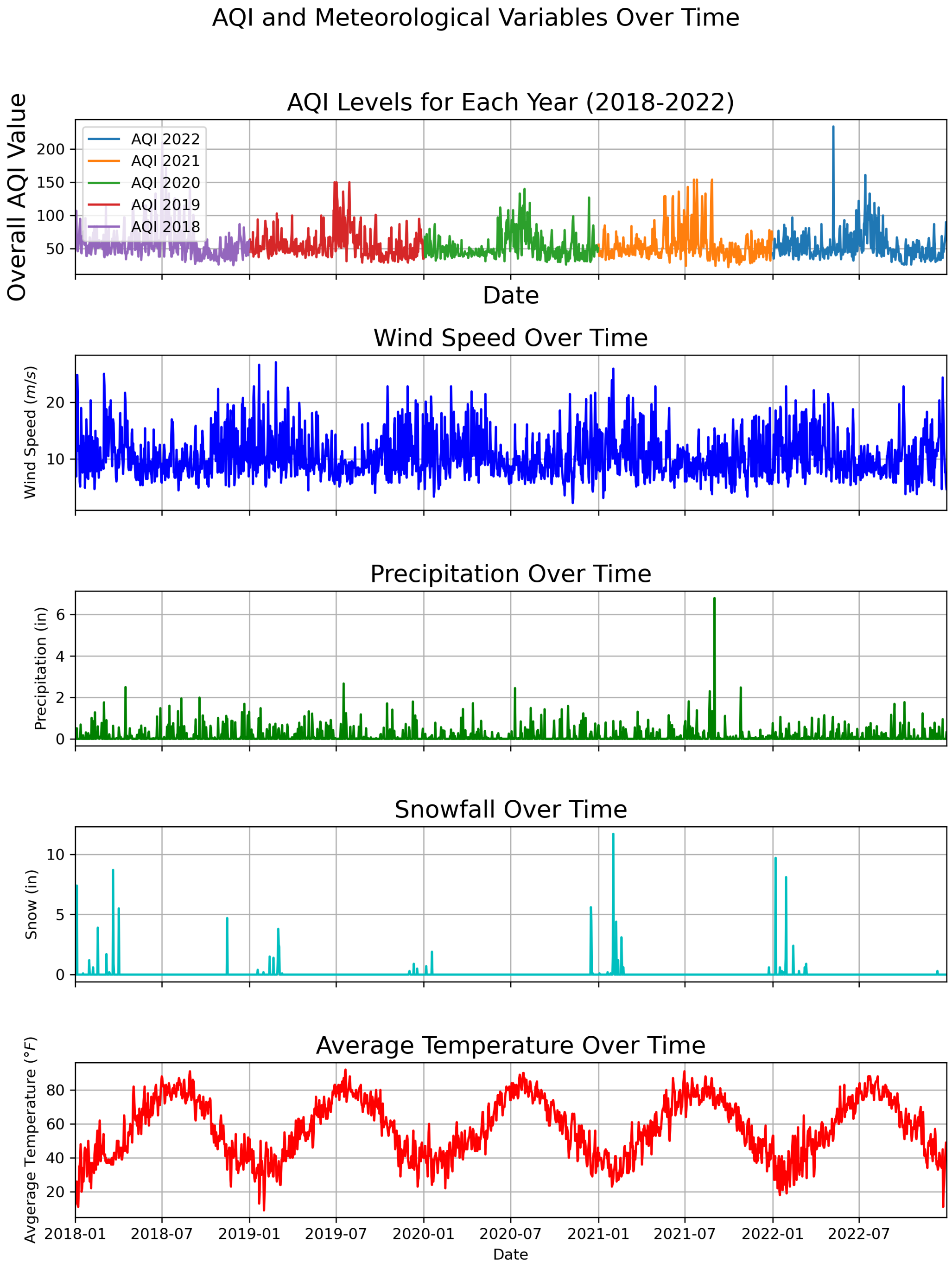

Figure 1: Daily AQI values and meteorological data (2018-2021)

Since many statistical methods, including ANOVA and Tukey's test, require homogeneity of variance, we performed Levene's test to assess variance homogeneity across the years. The results indicate that the variances are indeed homogeneous between the years. However, the data fail the normality assumption according to the Shapiro-Wilk test.

Given that the data does not meet the normality assumption, we explored several transformations to normalize the data and achieve more homogeneous variances. Transformations such as logarithmic, square root, and Box-Cox, Yeo-Johnson, Reciprocal, inverse square root, Arc sin square root and Z=score were applied to the AQI values to make the data more suitable for statistical analysis. Figure 3 shows transformed AQI histograms.

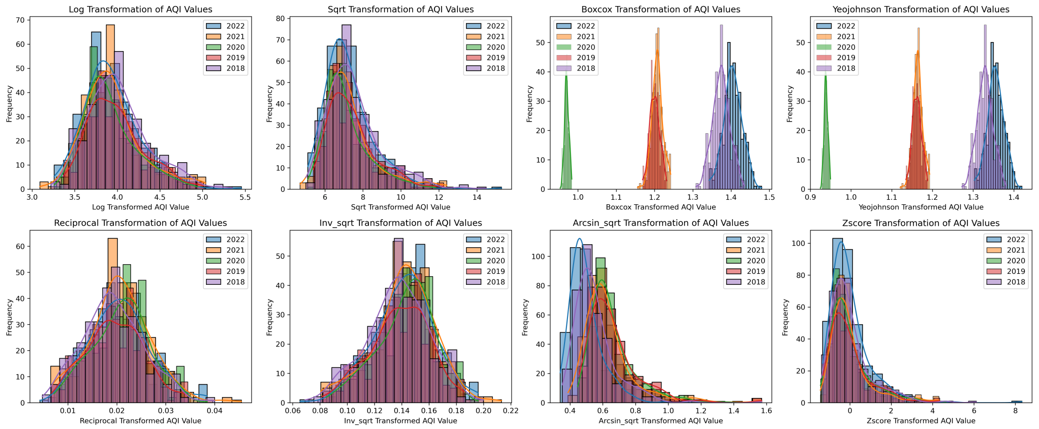

Figure 3: Histograms of AQI values for different transformations

The results showed that none of the transformations can not make the data normal and stablize the variance at the same time. Since ANOVA and Tukey's test are more sensitive to non constant variance and can work with normal like distributions. Hence, the inverse square root transformation applied to AQI. Table 2 shows mean and variance of transformed AQI values.

Table 2: Transformed AQI Statistics (2018-2022)

Statistic	**2018**	**2019**	**2020**	**2021**	**2022**
Transformed Mean AQI	0.13690	0.13839	0.14305	0.14169	0.14165
Transformed Variance AQI	0.00053	0.00047	0.00044	0.00053	0.00050

Transformed values of AQI for each year used as input to ANOVA test. The F-statistic and P-value of ANOVa test was 4.8888, and 0.0006, The small p-value (< 0.05) indicates that there are significant differences in the transformed AQI means across the years. The Tukey HSD test provides pairwise comparisons between the years. the results of Tukey test is reported in the table 3:

Table 3: Tukey's test Results

Group1	**Group2**	**Meandiff**	**P-adj**	**Lower**	**Upper**	**Reject**
2018	2019	0.0015	0.8971	-0.003	0.006	False
2018	2020	0.0062	0.0018	0.0017	0.0106	True
2018	2021	0.0048	0.0305	0.0003	0.0093	True
2018	2022	0.0047	0.0328	0.0002	0.0092	True
2019	2020	0.0047	0.0374	0.0002	0.0092	True
2019	2021	0.0033	0.2640	-0.0012	0.0078	False
2019	2022	0.0033	0.2759	-0.0012	0.0078	False
2020	2021	-0.0014	0.9218	-0.0059	0.0031	False
2020	2022	-0.0014	0.9138	-0.0059	0.0031	False
2021	2022	-0.0000	1.0000	-0.0045	0.0045	False

The **Meandiff** refers to the difference in means between the two groups. The **P-adj** is the adjusted p-value for each pair. The **Reject** indicates whether the null hypothesis (no difference between group means) is rejected (**True**) or not rejected (**False**).

The significant differences primarily occur when comparing 2018 (the earliest year) to later years like 2020, 2021, and 2022. Also, 2019 and 2020 are different. This suggests that air quality improved (higher values indicate better air quality) in these later years compared to 2018. The lack of significant differences among the years closer together (e.g., 2020 vs. 2021 or 2021 vs.

2022) indicates stability in air quality during those periods. the mean of AQI and the four meteorological variables are reported in table 4.

Table 4: Mean values of AQI and meteorological variables for each year

Year	**AQI**	**Temperature** (F)	**Wind Speed** (m/s)	**Precipitation** (in)	**Snow** (in)
2022	54.4	56.8	10.7	0.1	0.07
2021	54.8	58.1	10.4	0.1	0.07
2020	52.7	58.3	10.5	0.1	0.04
2019	56.8	56.1	10.8	0.1	0.04
2018	58.9	56.7	10.9	0.2	0.10

Figure 3 shows correlation analysis results between meteorological variables and AQI value for every and all years and also once by considering all years together. The results generally show positive correlation between temperature and AQI and negative correlation between windspeed and AQI. percipitation and snow has pretty low correlation with AQI but the correlation is negative.

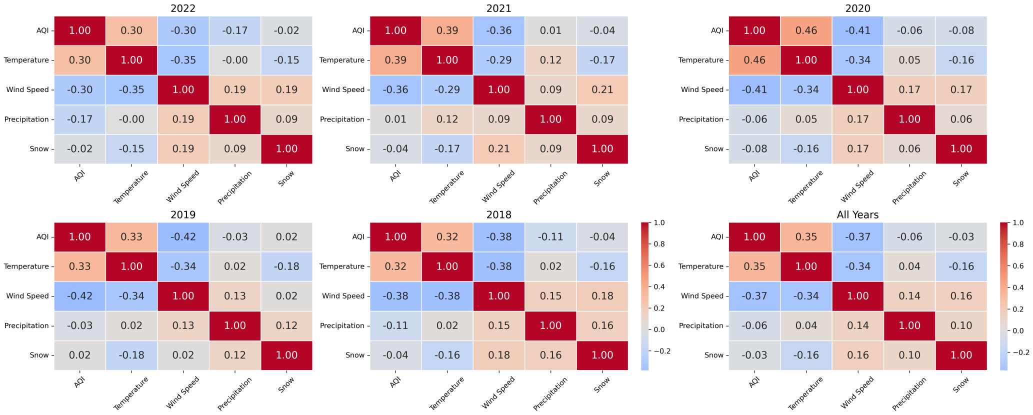

Figure 4: Correlation of meteorologic variables and AQI

4 Conclusion and Discussion

The analysis of AQI data for New York city from 2018 to 2022 revealed that the data do not follow a normal distribution. While variance stabilization was observed, the lack of normality prevents the appropriate application of parametric methods such as ANOVA and Tukey's test. These tests, while more sensitive to unequal variances, require normality for their validity. To address this, number of transformations that simultaneously normalize the data and stabilize variances should be explored, or alternative non-parametric methods should be considered.

Despite these challenges, the analysis identified four pairs of years with statistically significant differences in AQI: 2019-2020, 2018-2020, 2018-2021, and 2018-2022. Additionally, correlation analysis revealed a negative relationship between wind speed and AQI and a positive relationship between temperature and AQI. These findings align with physical principles: higher temperatures can accelerate the chemical reactions that produce pollutants, increasing AQI, while stronger winds can disperse pollutants, thereby reducing AQI levels.

However, inconsistencies arise when considering the mean values of AQI and meteorological variables reported in Table 4. For example, in 2018, the AQI was 4.5 units higher than in 2022, despite 2018 being windier and cooler on average. This inconsistency suggests the presence of confounding variables not accounted for in the current dataset, which may also influence AQI levels.

To resolve these inconsistencies and provide a more comprehensive understanding of AQI variability, further analyses are recommended. These should include the identification and inclusion of additional variables that may affect AQI. Addressing these factors would strengthen the conclusions and provide more reliable insights into the factors driving AQI variations in New York city.

Acknowledgments

The manuscript text has been reviewed using AI-powered proofreading tools.

References

[1] Us EPA Office of Air Quality Planning. *Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI).*

[2] S.A. Horn and P.K. Dasgupta. “The Air Quality Index (AQI) in historical and analytical perspective a tutorial review”. In: *Talanta* 267 (2024), p. 125260.

Appendix

This appendix provides details on the data transformations applied to the AQI values for various years, as well as the statistical tests used to assess normality and homogeneity of variances. The transformations are designed to address issues such as skewness, heteroscedasticity, and non-normality in the data.

A1. Data Transformation Formulas

Log Transformation: The log transformation reduces skewness and stabilizes variance. It is applied as follows:

$$AQI_{\text{log}} = \log(AQI_t + |\min(AQI_t)| + 1)$$

Explanation: This ensures that any non-positive AQI values are shifted before taking the logarithm. This transformation helps compress large values and make the distribution more symmetric.

Square Root Transformation: The square root transformation reduces the impact of large values:

$$AQI_{\text{sqrt}} = \sqrt{AQI_t}$$

Explanation: This transformation is useful when data exhibits positive skewness and variance increases with the mean.

Box-Cox Transformation: The Box-Cox transformation stabilizes variance and normalizes data. It is defined as:

$$AQI_{\text{boxcox}} = \frac{AQI_t^{\lambda} - 1}{\lambda}, \quad \lambda \neq 0$$

For $\lambda = 0$, $AQI_{\text{boxcox}} = \log(AQI_t)$

Explanation: The parameter λ is estimated from the data and controls the strength of the transformation.

Yeo-Johnson Transformation: This generalization of the Box-Cox transformation handles both positive and negative values:

$$AQI_{\text{yeojohnson}} = \begin{cases} \frac{(AQI_t+1)^{\lambda}-1}{\lambda}, & \text{if } AQI_t \geq 0, \\ -\left(\frac{(-AQI_t+1)^{2-\lambda}-1}{2-\lambda}\right), & \text{if } AQI_t < 0. \end{cases}$$

Explanation: This transformation is more flexible than Box-Cox and is widely used for normalizing data that includes negative values.

Reciprocal Transformation: The reciprocal transformation compresses large values:

$$AQI_{\text{reciprocal}} = \frac{1}{AQI_t + \epsilon}$$

where $\epsilon = 10^{-5}$ to avoid division by zero.

Explanation: The reciprocal transformation inverts AQI values, compressing large values and amplifying small ones.

Exponential Transformation: The exponential transformation amplifies small values:

$$AQI_{\text{exp}} = \exp(AQI_t)$$

Explanation: This function increases the spread of data and is suitable for heavy-tailed distributions.

Inverse Square Root Transformation: The inverse square root transformation reduces the influence of large values:

$$AQI_{\text{inv\_sqrt}} = \frac{1}{\sqrt{AQI_t + \epsilon}}$$

where $\epsilon = 10^{-5}$ to avoid division by zero.

Explanation: This transformation is similar to the reciprocal transformation but reduces large value influence more gradually.

Arcsine Square Root Transformation: This transformation is typically used for proportions:

$$AQI_{\text{arcsin\_sqrt}} = \arcsin\left(\sqrt{\frac{AQI_t}{\max(AQI_t)}}\right)$$

Explanation: Applied to proportions, this transformation scales the AQI values before applying the arcsine square root.

Z-Score Standardization: Z-score standardization transforms data to have a mean of 0 and a standard deviation of 1:

$$AQI_{\text{zscore}} = \frac{AQI_t - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of AQI values.

Explanation: Z-score standardization normalizes the data, making it suitable for comparison across datasets with different units or scales.

A2. Statistical Tests explanation

Shapiro-Wilk Test for Normality: This test assesses whether the data follows a normal distribution. The null hypothesis is that the data is normally distributed.

$$H_0 : \text{Data is normally distributed}$$

$$H_A : \text{Data is not normally distributed}$$

The Shapiro-Wilk test statistic W is used to evaluate normality.

Levene's Test for Homogeneity of Variances: Levene's test checks the equality of variances across groups (in this case, across years). The null hypothesis is that the variances are equal.

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

$$H_A : \text{At least one variance is different}$$

The Levene's test statistic F and corresponding p-value are used to assess the homogeneity of variances.