Analyzing New York City Air Quality (2018-2022)

Omid Emamjomehzadeh[*]

November 22, 2024

Abstract

This study seeks to examine the Air Quality Index (AQI) values in New York City (NYC) from 2018 to 2022 to ascertain the presence of statistically significant differences over the years. Various transformations are employed to stabilize variance and close the data distribution to a normal distribution. The analysis of variance revealed a difference between the years. Tukey's technique was utilized to identify the differing pairs of years. The data indicate that the AQI of 2018 differs from that of 2020, 2021, and 2022, and the AQI of 2019 also differs from that of 2020. Meteorological variables, including temperature, precipitation, snow depth, and wind speed, are analyzed to elucidate the correlations between AQI and these variables. The correlation value indicates a positive association between temperature and AQI, and a negative correlation between wind speed and AQI. The correlations do not elucidate the discrepancies in AQI, indicating the presence of confounding factors influencing it.

1 Introduction

The Air Quality Index (AQI), instituted by the United States Environmental Protection Agency (USEPA) in 1977, has emerged as an essential instrument for reporting air pollution to public and its health consequences. Prior

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

to the AQI, air quality indicators exhibited considerable variability within metropolitan regions, resulting in uneven communication [3].

The Air Quality Index (AQI) is determined by assessing pollutants including ozone (O_3), particulate matter (PM_{10} and $PM_{2.5}$), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2). This assessment employs a formula that converts pollutant concentrations into a dimensionless value between 0 and 500, reflecting varying degrees of health risk. Each pollutant's concentration (C_p) is assessed against break-point values (BP_{Lo} and BP_{Hi}), which delineate ranges for AQI categories. The index is calculated using formula 1.

$$I_p = \frac{(I_{\mathrm{Hi}} - I_{\mathrm{Lo}})}{(BP_{\mathrm{Hi}} - BP_{\mathrm{Lo}})}(C_p - BP_{\mathrm{Lo}}) + I_{\mathrm{Lo}} \tag{1}$$

where I_{Lo} and I_{Hi} denote the lowest and upper AQI values for the break-point range, respectively. The Air Quality Index (AQI) for a region is established based on the pollutant exhibiting the highest index value on a specific day, known as the *critical pollutant*. This method represents the most adverse air quality scenario for any certain day. The AQI employs a color-coded scheme to facilitate public comprehension, where values below 50 denote *good* air quality and values beyond 300 imply *hazardous* circumstances [2].

Vulnerable populations, including children, the elderly, and those with previous respiratory or cardiovascular diseases, are notably emphasized in AQI reports when pollutant levels surpass NAAQS requirements. The AQI serves as a great tool for public health communication by converting intricate scientific data into practical information. Although its simplicity, the AQI possesses limitations, including an emphasis on the highest pollutant index, which may neglect the cumulative or synergistic impacts of many contaminants. Nevertheless, the AQI is a crucial foundation for educating the public and directing policy decisions around air quality. [3].

This study seeks to analyze AQI data from New York City from 2018 to 2022 to identify any notable variations in AQI values across these years, while offering explanations for these discrepancies through meteorological data. The findings of this study will improve our comprehension of the association between AQI in New York City and meteorological variables.

The remainder of the study is structured as follows: Section 2 (Method) elucidates the process of data transformation. Subsequently, analysis of variance and Tukey's tests were employed to ascertain the differences across years. Ultimately, a correlation analysis was conducted between the Air

Quality Index (AQI) and meteorological variables. Section 3 (Results) delineates the outcomes of using the recommended methodology on the dataset. Section 4 offers concluding thoughts and discusses the limitations of various approaches, so establishing a foundation 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 analysis of variance and Tukey's test which are used in this study, so we apply the Shapiro-Wilk test to check whether the data is 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 stabilized 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

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 variances are larger than what could be expected by random chance. This is done using the **F-ratio**, which is calculated in formula 2.

$$F = \frac{\text{MSB}}{\text{MSW}} \tag{2}$$

Here, the Mean Square Between Groups (MSB) and Mean Square Within Groups (MSW) are defined in formula 3 :

$$\mathrm{MSB} = \frac{\mathrm{SSB}}{\mathrm{dfB}}, \quad \mathrm{MSW} = \frac{\mathrm{SSW}}{\mathrm{dfW}} \tag{3}$$

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 [4, 5].

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 groups. If ANOVA detects a difference, then Tukey's Test could be used to investigate exactly which two groups are different [5, 4].

2.2 Tukey's test

The goal of Tukey's Test is to identify which pairs of groups 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 [1, 6].

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 (in), snow (in), and average temperature

($^\circ F$). Figure 1 illustrates the distribution of AQI values and the meteorological variables for the years 2018-2021. After calculating the basic statistics for each year AQI, the following summary statistics were reported in (Table 1).

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

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

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

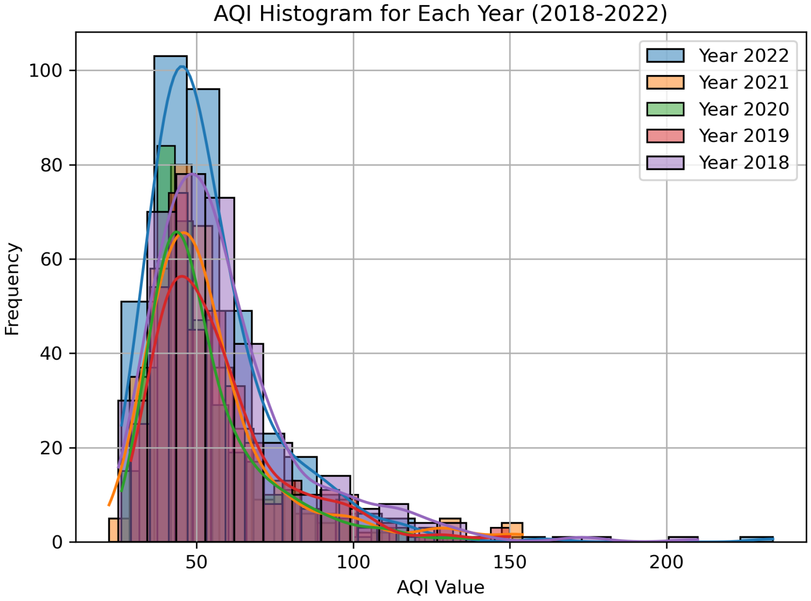

Figure 2: Histograms of AQI values for each year (2018-2022)

Since many statistical methods, including ANOVA and Tukey's test, require homogeneity of variance, we performed Levene's test (explained in de-

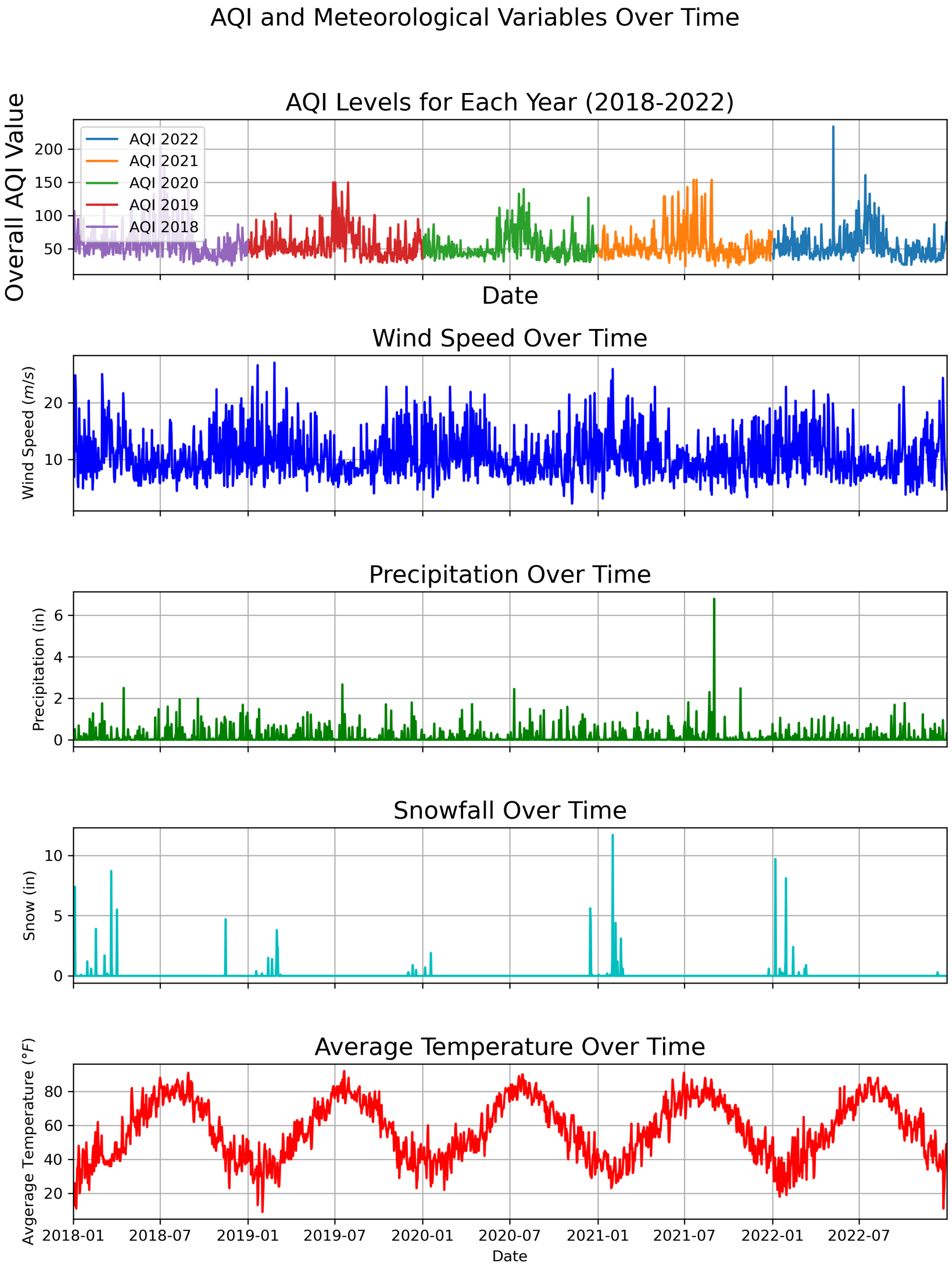

Figure 1: Daily AQI values and meteorological data (2018-2021)

tail in 4) 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 (explained in detail in 4).

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 (transformation formulas are explained in detail in 4) were applied to the AQI values to make the data more suitable for statistical analysis. Figure 3 shows transformed AQI histograms.

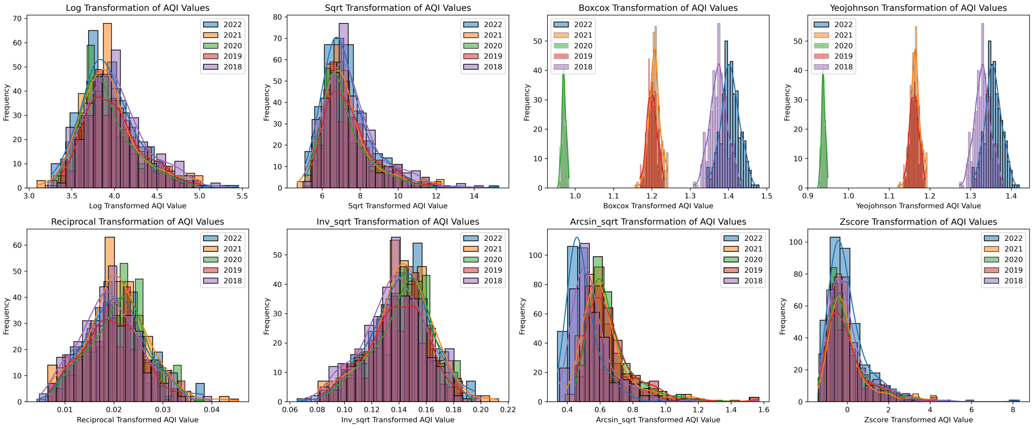

Figure 3: Histograms of AQI values for different transformations

The results show that none of the transformations can make the data normal and stabilize 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 AQI mean	0.13690	0.13839	0.14305	0.14169	0.14165
Transformed AQI variance	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 is significant differences in the transformed AQI means across the years. The Tukey's test provides pairwise comparisons between the years. the results of Tukey's 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 year and once by considering all years together. The results generally show positive correlation between temperature and AQI and negative correlation between wind speed and AQI. Precipitation and snow has pretty low correlation with AQI but the correlation is negative.

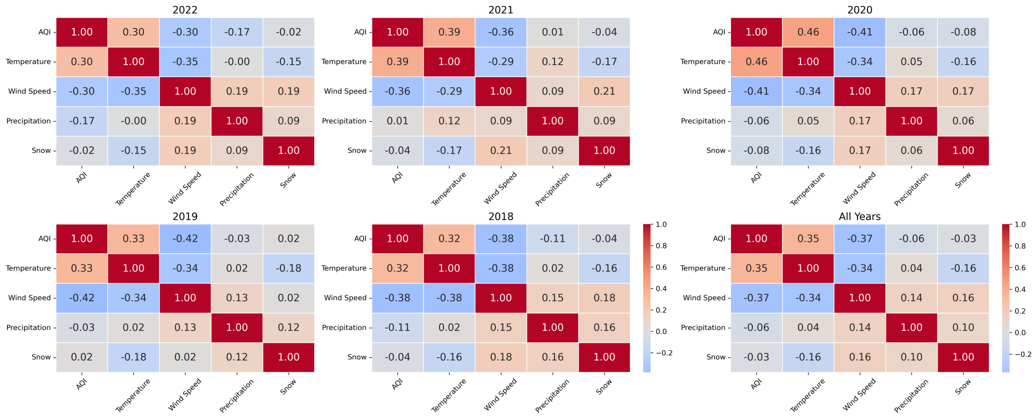

Figure 4: Correlation of meteorologic variables and AQI

4 Conclusion and Discussion

The analysis of AQI data for New York city from 2018 to 2022 revealed that the data do not follow normal distribution. While stabilized variance 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 in future studies.

Despite these challenges in transformation of the data to normal distribution and stabilize variance, 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.

References

[1] Herve Abdi and Lynne Williams. *Tukey's Honestly Significant Difference (HSD) Test.* `https://personal.utdallas.edu/~herve/abdi-HSD2010-pretty.pdf`. 2021.

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

[3] 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.

[4] R. Lowry. *Concepts and Applications of Inferential Statistics, Chapter 14*. http://vassarstats.net/textbook/. 2014.

[5] G.H. McDonald. *Handbook of Biological Statistics, One-way ANOVA*. http://www.biostathandbook.com/onewayanova.html.

[6] NIST/SEMATECH. *e-Handbook of Statistical Methods, "7.4.7.1. Tukey's Method."*. https://www.itl.nist.gov/div898/handbook/prc/section4/prc471.htm. Accessed: 2020-11-28. Nov. 2020.

[7] NIST/SEMATECH. *e-Handbook of Statistical Methods, Section 2.1.3*. https://www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm. DOI: 10.18434/M32189.

[8] NIST/SEMATECH. *e-Handbook of Statistical Methods, Section 3.5a*. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm.

Acknowledgments

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

Data availability

Data, code, and report (latex) are available in the NYC_AQI GitHub repository: https://github.com/omidemam/NYC_AQI.git.

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, heterogeneity of variance, and non-normality in the data.

A1. Data transformation formulas

Log Transformation: The log transformation reduces skewness and stabilizes variance. It is applied using formula 4.

$$AQI_{\log} = \log(AQI_t + |\min(AQI_t)| + 1) \tag{4}$$

where AQI_t represents the AQI at time t. The term $\min(AQI_t)$ refers to the minimum AQI value observed over the entire time series. The transformed value, $AQI_{\log}$, is useful in normalizing the distribution and stabilizing the variance of the AQI data.

Square root transformation: The square root transformation reduces the impact of large values, is caculated using formula 9.

$$AQI_{\text{sqrt}} = \sqrt{AQI_t} \tag{5}$$

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 computed using the formula 6.

$$AQI_{\text{boxcox}} = \begin{cases} \frac{AQI_t^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(AQI_t), & \lambda = 0 \end{cases} \tag{6}$$

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 and is computed using formula 7.

$$AQI_{\text{yeojohnson}} = \begin{cases} \frac{(AQI_t+1)^{\lambda}-1}{\lambda}, & \text{if } AQI_t \geq 0, \\ -\left(\frac{(-AQI_t+1)^{2-\lambda}-1}{2-\lambda}\right), & \text{if } AQI_t < 0. \end{cases} \tag{7}$$

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, and is computed using formula 8.

$$AQI_{\text{reciprocal}} = \frac{1}{AQI_t + \epsilon} \tag{8}$$

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

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

Inverse square root transformation: The inverse square root transformation reduces the influence of large values and is calculated using formula 9.

$$AQI_{\text{inv\_sqrt}} = \frac{1}{\sqrt{AQI_t + \epsilon}} \tag{9}$$

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

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 and is calculated using formula 10.

$$AQI_{\text{arcsin\_sqrt}} = \arcsin\left(\sqrt{\frac{AQI_t}{\max(AQI_t)}}\right) \tag{10}$$

This transformation scales the AQI values before applying the arcsine square root.

Z-Score standardization: Z-score standardization normalizes the data, making it suitable for comparison across datasets with different units or scales. It transforms data to have a mean of 0 and a standard deviation of 1, and is computed using formula 11.

$$AQI_{\text{zscore}} = \frac{AQI_t - \mu}{\sigma} \tag{11}$$

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

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[7].

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 [8].