Research Article

Beyond the Threshold: How Categorizing Air Quality Metrics Alters Statistical Inference and Model Performance

Audrey Chan

Diamond Bar High School, 21400 Pathfinder Rd, Diamond Bar, CA 91765, United States

ABSTRACT

Air pollution remains a critical environmental and public health issue, with harmful pollutants such as nitrogen oxides, sulfur dioxide, carbon monoxide, and volatile organic compounds posing significant risks to human health and ecosystems. This study investigates key factors influencing air pollution and evaluates how converting continuous pollutant measurements into categorical Air Quality Index (AQI) labels affects statistical inference and model performance. Using a global dataset of air quality measurements, the analysis incorporates a combination of statistical techniques—including correlation analysis and multiple linear regression—to examine pollutant sources and data transformation impacts. The results show that continuous data produced stronger correlations, higher R^2 values, and better prediction accuracy than categorical data. However, categorical models offered clearer interpretability and may still be useful in settings with limited data or for effective public communication. The findings offer valuable insights into the trade-offs between data accessibility and analytical precision, informing policymakers in the development of targeted mitigation strategies and sustainable air quality management practices.

Keywords: Air Quality Metrics; Categorical Data; Statistical Inference; Correlation Analysis; Multiple Linear Model

INTRODUCTION

Air pollution has emerged as one of the most pressing environmental and public health challenges of the 21st century. It affects not only ecological stability but also the quality of life and economic productivity of populations worldwide. According to the World Health Organization, ambient air pollution causes an estimated 4.2 million premature deaths annually due to exposure to fine particulate matter (PM2.5) that penetrates deep into the lungs and cardiovascular system (1). Economically, air pollution imposes substantial costs through healthcare expenditures, workforce productivity loss, and decreased agricultural yields, with the World Bank estimating a global cost of $5 trillion in welfare losses annually (2). Beyond these direct impacts, pollution influences climate systems, contributes to biodiversity loss, and exacerbates existing social inequalities,

Corresponding author: Audrey Chan, E-mail: audrey132510chan@gmail.com.

Accepted August 25, 2025
https://doi.org/10.70251/HYJR2348.3446547l

making air quality monitoring a high-priority issue in environmental governance, public health planning, and urban development.

To manage and mitigate the effects of air pollution, researchers and governments rely on an array of pollutant measurement indices. Common indicators include PM2.5 and PM10 for particulate matter, NO_2 and SO_2 for gaseous pollutants, CO levels, and ozone (O_3) concentrations (3). These measurements are often aggregated into broader indices such as the Air Quality Index (AQI), which provides a standardized summary of pollution levels in categories ranging from "Good" to "Hazardous." Each pollutant has its own threshold values and measurement units, typically reported as micrograms per cubic meter ($\mu g/m^3$) or parts per million (ppm), allowing real-time tracking and historical comparison of air quality levels. The AQI, in particular, has gained traction due to its intuitive structure and utility for public communication, policy development, and regulatory compliance (4).

A wide variety of methodological approaches have been employed to analyze these pollution metrics. Traditional statistical techniques—such as correlation analysis and multiple linear regression—are commonly used to explore associations between pollutants and health outcomes, meteorological conditions, or socioeconomic variables. More recently, advanced machine learning methods, including random forests, support vector machines, and neural networks, have been applied to forecast pollution trends, detect anomalies, and model non-linear relationships between environmental factors (5). Experimental designs have also been utilized in some studies to simulate pollution exposure and examine causality (6). Each of these approaches depends heavily on the availability and granularity of accurate numerical measurements.

Despite these methodological advancements, a persistent challenge in air pollution research lies in data accessibility. High-resolution continuous monitoring systems can be expensive, technically demanding, and limited in spatial coverage—especially in low-resource or rural settings. As a practical workaround, many organizations and researchers convert continuous numerical values into categorical forms, such as AQI bands, to simplify reporting, reduce data noise, and enable low-cost community-level data collection (7). These categorical labels—e.g., "Moderate," "Unhealthy," "Very Unhealthy"—help convey risk to non-technical audiences and are more readily usable in mobile apps, policy dashboards, and emergency communication systems.

However, while categorical data can enhance interpretability and accessibility, it may compromise the statistical precision of analytical models. Few studies have systematically evaluated how such categorization affects the outcomes of environmental data analysis, especially in modeling-intensive contexts such as regression and classification. This study seeks to fill that gap by quantifying the impact of converting continuous pollution measurements into categorical labels on two types of statistical models: Pearson correlation analysis and multiple linear regression. Specifically, we compare results derived from continuous pollutant variables (such as CO, NO_2, and PM2.5 levels) with those obtained from their corresponding AQI-based categorical forms. By conducting this comparative analysis, the research aims to determine whether categorical transformations meaningfully affect modeling validity, statistical significance, or predictive power.

METHODS AND MATERIALS

In this study, the research aims to investigate the impact of using categorized AQI- related pollutant values on the performance and interpretability of statistical modeling. To address this objective, several methodological steps were undertaken. First, the dataset included pre-categorized pollutant indicators—specifically carbon monoxide (CO), nitrogen monoxide (NO_2), and particulate matter (PM2.5)—each classified into qualitative levels such as Good, Moderate, and Poor according to pollutant-specific thresholds defined by the original data source. These categorizations were not derived within this study but collected directly as part of the dataset. Second, Pearson's correlation analysis was used to examine the relationships between the overall AQI value (used as the continuous response variable) and each pollutant, comparing results for both continuous pollutant measurements and their categorized counterparts. The correlation coefficients and associated p-values were used to evaluate the strength and statistical significance of these relationships. Third, the study compares the modeling performance of multiple linear regression under two conditions: using only continuous pollutant variables, and using only their categorized forms. Through this comparative analysis, the methodology evaluates how the representation of pollutant inputs—continuous versus categorical—affects model accuracy and interpretability in the context of AQI prediction.

Pearson Coefficient Analysis

To analyze the linear relationships between the overall AQI and individual pollutant-specific AQI values—such as NO_2, CO, and O_3 (8)—Pearson's correlation coefficient is employed. This method quantifies the strength and direction of a linear association between two continuous variables. The correlation coefficient (r) ranges from –1 to +1, where values closer to ±1 indicate a stronger linear relationship, and a value near 0 suggests no linear association. For each pairwise comparison, the corresponding p-value is also calculated to assess the statistical significance of the observed correlation. In this study, a significance threshold of 0.05 is used; a p-value below 0.05 indicates that the correlation is statistically significant and unlikely to have occurred by chance. This approach allows for a robust evaluation of whether variations in pollutant-specific AQI values are meaningfully associated with changes in the overall AQI index under normal distribution assumptions. The detailed calculation of r value is shown below (Sedgwick, 2012).

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}} \quad (1)$$

Where x and $\bar{x}$ are individual value and mean of the values for the x-variable, respectively, while y and $\bar{y}$ are the corresponding values for y-variable.

Multiple Linear Regression

To further investigate the influence of pollutant-specific AQI indicators on the overall AQI score, this study applies multiple linear regression analysis. In both models, the dependent variable (Y) is the overall AQI value. The independent variables, however, differ in their format across two separate models. The first model uses the original continuous AQI values for key pollutants, including nitrogen dioxide (NO_2), carbon monoxide (CO), ozone (O_3), and particulate matter (PM2.5). The structure of this continuous model can be expressed as:

$$\text{AQI Overall} = \beta_0 + (\beta_1 \cdot NO_2) + (\beta_2 \cdot \text{CO}) + (\beta_3 \cdot O_3) + \beta_4 \cdot PM2.5 + \varepsilon \quad (2)$$

In the regression equation, β_0 represents the intercept, or the baseline value of the overall Air Quality Index (AQI) when all pollutant concentrations are zero. The coefficients β_1, β_2, β_3, and β_4 are the weights associated with the respective pollutants NO_2, CO, O_3, and PM2.5, indicating their individual contributions to the overall AQI. ε is the error term, capturing the unexplained variation in AQI not accounted for by the model. The second model replaces the continuous predictors with their corresponding categorical versions, as provided in the dataset. These categories—such as "Good," "Moderate," and "Poor"—are encoded using dummy variables to fit within the linear modeling framework. The goal of this second model is to evaluate whether representing pollutant indicators in categorized form impacts model performance or alters the interpretation of predictor importance. The categorical model can be expressed in general form as:

$$\text{AQI Overall} = \beta_0 + \sum_{j=1}^{k} \beta_j \cdot D_j + \varepsilon \quad (3)$$

where D_j represents the dummy-coded levels of the categorized pollutants, and k is the total number of dummy variables generated across all pollutants. By comparing both models, this study aims to determine whether the categorization of independent variables—though often useful for simplification—leads to loss of predictive accuracy or change in statistical significance relative to using the full continuous data. Model performance is assessed using R-squared values and significance testing for coefficients (9), providing insight into the trade-offs between interpretability and precision in environmental data modeling.

DATA DESCRIPTION

The dataset used in this study contains air quality information from various countries and cities, with the aim of analyzing the relationship between general AQI and individual sub-AQIs. It includes AQI values for specific pollutants—Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Ozone (O_3), and others—alongside the overall Air Quality Index (AQI), which reflects the combined impact of multiple pollutants.

To support meaningful analysis, the raw numerical AQI values were categorized into levels such as "Good," "Moderate," "Unhealthy," and "Very Unhealthy" based on standard threshold values. This categorization enables both correlation analysis and modeling, while also allowing for comparisons between general AQI and its pollutant-specific sub-indices. Thresholds for classification followed U.S. EPA standards (10).

Table 1 summarizes the key variables used in this study, including their types, definitions, and relevant descriptive statistics. The thresholds used to classify AQI levels play a critical role in shaping the outcomes of both the modeling and annotation analyses.

RESULTS

To evaluate the analytical consequences of converting continuous pollution measurements into categorical form, this study conducted a series of comparative analyses. First, Correlation Analysis was used to assess how categorization affects the strength and direction of relationships among key air quality variables. Second, Multiple Linear Regression models were employed to examine the impact of data transformation on model fit, explanatory power, and predictor significance. Together, these approaches provide a foundation for understanding

Table 1. Summary of Key Variables

Variable	Type	Definition	Descriptive Statistic
Country	Categorical	Location where air pollution measurements were taken	94 unique countries/regions (location identifier only)
Overall AQI Value	Numerical	Composite Air Quality Index combining multiple pollutants; higher numbers = worse air quality.	Mean: 72.01 SD: 56.06 Min: 6 Max: 500
Overall AQI Band	Categorical	Overall AQI grouped per U.S. EPA breakpoints ("Good", "Moderate", etc.).	Good: 9,936 Moderate: 9,231 Unhealthy for Sensitive Groups: 1,591 Unhealthy: 2,227 Very Unhealthy: 287 Hazardous: 191
NO_2 AQI Value	Numerical	AQI derived from nitrogen-dioxide concentration.	Mean ≈ 1.07 SD ≈ 0.20 Min = 0 Max = 5
NO_2 AQI Band	Categorical	NO_2 AQI grouped into EPA bands.	Good: 23,448 Moderate: 15
CO AQI Value	Numerical	AQI derived from carbon-monoxide concentration.	Mean = 1.00 SD ≈ 0.06 Min = 0 Max = 5
CO AQI Band	Categorical	CO AQI grouped into EPA bands.	Good: 23,460 Moderate: 2 Unhealthy for Sensitive Groups: 1
Ozone AQI Value	Numerical	AQI derived from ground-level ozone (O_3) concentration.	Mean ≈ 6.5 SD ≈ 7.7 Min = 0 Max = 40
Ozone AQI Band	Categorical	Ozone AQI grouped into EPA bands.	Good: 21,069; Moderate: 1,445 Unhealthy for Sensitive Groups: 491 Unhealthy: 405 Very Unhealthy: 53
PM2.5 AQI Value	Numerical	AQI derived from fine particulate matter (≤ 2.5 μm) concentration.	Mean ≈ 71.3 SD ≈ 55.9 Min = 0 Max = 500
PM2.5 AQI Band	Categorical	PM2.5 AQI grouped into EPA bands.	Good: 10,208 Moderate: 9,075 Unhealthy for Sensitive Groups: 1,624 Unhealthy: 2,129 Very Unhealthy: 255 Hazardous: 172

the trade-offs between data simplification and statistical robustness.

Impact of Categorization on Correlation Analysis

To investigate the effect of transforming continuous AQI values into categorical classifications, two correlation matrix plots were constructed—one based on continuous AQI values and the other based on discrete AQI categories (e.g., Good, Moderate, Unhealthy, etc.). In the continuous correlation matrix (Figure 1), several moderate to strong associations emerged. For example, NO_2 AQI values showed a moderately strong positive correlation with CO AQI values ($r = 0.49$, $p < 0.001$), and PM2.5 AQI values also correlated positively with CO ($r = 0.44$) and Ozone ($r = 0.34$). These statistically significant relationships suggest a shared pollution pattern across certain pollutants—particularly those often emitted from common sources such as vehicle exhaust and industrial activity. In contrast, the correlation matrix based on categorical AQI values (Figure 2) displayed a notable reduction in the strength of relationships. For instance, the correlation between NO_2 and CO dropped from $r = 0.49$ to $r = 0.21$, while Ozone and NO_2, which were weakly negatively correlated in the continuous case ($r = -0.18$), showed an even weaker and statistically insignificant association ($r = -0.01$, $p > 0.05$) after conversion. This shift can be attributed to the information loss that occurs during binning, where continuous variability is replaced by threshold-based groupings. While categorization may simplify interpretation for public health communication or regulatory enforcement, it also obscures subtle but meaningful variations in pollutant behavior, reducing the granularity and sensitivity of statistical relationships.

Impact of Categorization on Multiple Linear Regression

The impact of predictor resolution on model performance is evident in the comparison between the high-resolution numerical model (as shown in Table 2) and the low-resolution categorical model (as illustrated in Table 3). The high-resolution model indicates that it

Table 2. Multiple Linear Regression Results based on Numerous Independent Variables

	Coef	Std err	t	P>|t|
CO AQI Value	0.0171	0.040	-5.419	0.667
Ozone AQI Value	0.1575	0.002	0.430	0.000
NO_2 AQI Value	-0.0362	0.013	67.371	0.007
PM2.5 AQI Value	0.9801	0.001	775.385	0.000

Table 3. Multiple Linear Regression Results based on Categorized Independent Variables

	Coef	Std err	t	P>|t|
CO AQI Value	100.8328	9.443	10.678	0.000
Ozone AQI Value	8.5512	0.288	29.707	0.000
NO_2 AQI Value	40.0311	5.990	6.683	0.000
PM2.5 AQI Value	47.5150	0.156	305.543	0.000

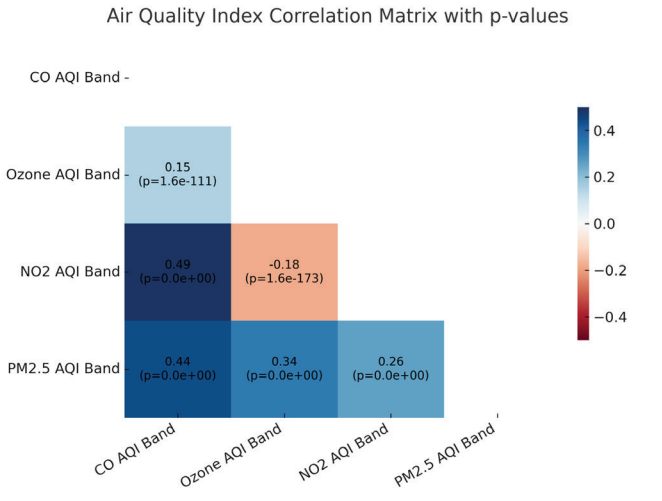

Figure 1. Correlation Matrix Plot Based on Numerous Independent Variables.

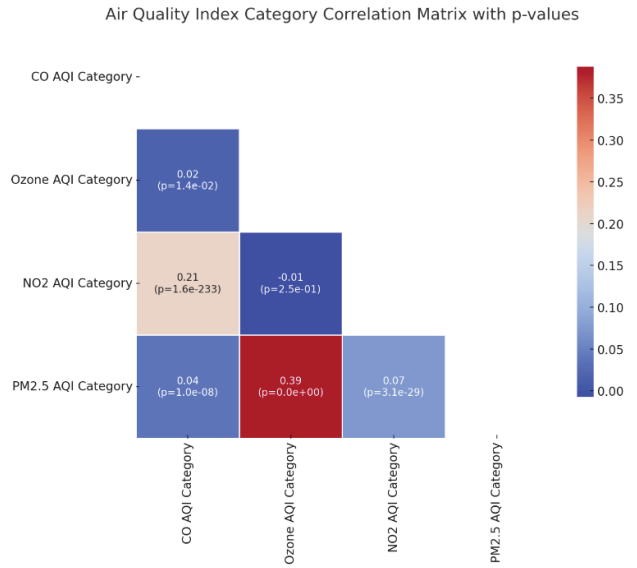

Figure 2. Correlation Matrix Plot Based on Categorized Independent Variables.

explains 97.5% of the variance in AQI Value, whereas the categorical model reflects a loss of predictive power due to discretization. The fine granularity of numerical predictors allows the model to capture more precise relationships between air pollutants and AQI, while categorization leads to information loss. This effect is also observed in the coefficient magnitudes; for example, PM2.5 AQI Value has a coefficient of 0.9801 in the numerical model, whereas PM2.5 AQI Category has a coefficient of 47.5150 in the categorical model. The categorical model, while capturing broader trends, lacks the detailed resolution to represent small-scale variations effectively.

Statistical significance (p-values) also differs between the models. In the numerical model, CO AQI Value has a p-value of 0.667, suggesting it is not a significant predictor, while in the categorical model, CO AQI Category has a p-value of 0.000, implying statistical significance despite reduced granularity. The discretization process reduces noise, making some predictors appear more statistically relevant, but at the cost of model precision. Additionally, AIC and BIC values are lower in the high-resolution model, confirming a better overall fit. Both models exhibit no significant autocorrelation (Durbin-Watson ≈ 2.0), suggesting well-specified error structures. Overall, the high-resolution numerical model provides superior predictive accuracy and finer interpretability, while the categorical model, though statistically significant, introduces coarser approximations that reduce explanatory power.

CONCLUSIONS AND RECOMMENDATIONS

This study demonstrates how the transformation of continuous air quality data into categorical labels significantly impacts the validity and precision of statistical analyses. By examining pollutants such as carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), and particulate matter (PM2.5), the research reveals that while categorization can simplify data interpretation for the public and policymakers, it also introduces information loss that reduces statistical power and model accuracy.

Across Correlation Analysis and Multiple Linear Regression, continuous data consistently yielded stronger relationships, higher R^2 values, and greater predictive precision. In contrast, categorical models, while sometimes enhancing interpretability and showing stronger statistical significance for certain predictors, demonstrated a loss of nuance and granularity.

These findings suggest that while categorized AQI data may be beneficial for public reporting and regulatory enforcement, continuous data should be retained for scientific analysis and forecasting whenever possible.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my AP Research teacher for his continuous guidance and support throughout this project. Special thanks to my professor for statistical advice and feedback on data interpretation. I am also thankful to my peers who reviewed early drafts and to the open data community for making air quality data available on platforms like Kaggle. Lastly, I thank my family for their encouragement and belief in my academic journey.

REFERENCES

1. World Health Organization. WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide.2021. https://www.who.int/publications/i/item/9789240034228
2. Owusu PA & Sarkodie SA. Global estimation of mortality, disability-adjusted life years and welfare cost from exposure to ambient air pollution. *Science of the Total Environment.* 2020; 742: 140636. https://doi.org/10.1016/j.scitotenv.2020.140636
3. Karagulian F, Belis CA, Dora CFC, Prüss-Ustün AM, et al. Contributions to cities' ambient particulate matter (PM): A systematic review. *Atmospheric Environment.* 2015; 120: 475–483. https://doi.org/10.1016/j.atmosenv.2015.08.087
4. Wang Y & Chen Y. Understanding air quality index (AQI) and its implications for public health in China. *International Journal of Environmental Research and Public Health.* 2020; 17 (22): 8188. https://doi.org/10.3390/ijerph17228188
5. Zhang Y, Bocquet M, Mallet V, Seigneur C & Baklanov A. Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment.* 2018; 60: 632–655. https://doi.org/10.1016/j.atmosenv.2012.06.031
6. Schikowski T, Altug H & Krämer U. Experimental evidence on the impact of air pollution on human health. *European Respiratory Review.* 2014; 23 (132): 149–158. https://doi.org/10.1183/09059180.00000714
7. Rai P, Fokar M, Sahu S, Singh A & Gaur A. Development of low-cost sensor-based air quality monitoring system for India. *Measurement.* 2020; 152: 107302. https://doi.org/10.1016/j.measurement.2019.107302
8. Sedgwick P. Pearson's correlation coefficient. *BMJ.* 2012; 345: e4483. https://doi.org/10.1136/bmj.e4483
9. Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal.*

2012; 24 (3): 69–71.
10. U.S. Environmental Protection Agency. 2018. *Technical assistance document for reporting of daily air quality – the AQI.* EPA-454/B-18-007.
11. "Global air pollution dataset." (n.d.). Kaggle. https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset
12. Brauer M, Amann M, Burnett RT, Cohen A, et al. Exposure assessment for global burden of disease attributable to outdoor air pollution. *Environmental Science & Technology.* 2012; 46 (2): 652–660. https://doi.org/10.1021/es2025752
13. Chen K, Wang M, Huang C, Kinney PL & Anastasios T. A machine learning approach to estimate PM2.5 concentrations. *Science of the Total Environment.* 2020; 636: 52–60. https://doi.org/10.1016/j.scitotenv.2018.04.251
14. Dominici F, Peng RD, Barr CD & Bell ML. Protecting human health from air pollution. *Epidemiology.* 2006; 21 (2): 187–194. https://doi.org/10.1097/EDE.0b013e3181cc86e8
15. Lelieveld J, Evans JS, Fnais M, Giannadaki D & Pozzer A. The contribution of outdoor air pollution sources to premature mortality. *Nature.* 2015; 525 (7569): 367–371. https://doi.org/10.1038/nature15371
16. Benesty J, Chen J, Huang Y & Cohen I. Pearson correlation coefficient. In *Noise reduction in speech processing* (2009; pp.1–4). Springer. https://doi.org/10.1007/978-3-642-00296-0_5
17. Dancey CP & Reidy J. *Statistics without maths for psychology* (7th ed.). Pearson Education. 2017.
18. Helsel DR & Hirsch RM. *Statistical methods in water resources.* U.S. Geological Survey. 2002. https://pubs.usgs.gov/twri/twri4a3/
19. Guttikunda SK, Goel R & Pant P. Nature of air pollution, emission sources, and management in Indian cities. *Atmospheric Environment: X.* 2019; 3: 100040. https://doi.org/10.1016/j.aeaoa.2019.100040