ISyE 3030 Project: Analyzing Impact of Air Pollution on Quality of Life Indicators Across U.S Regions
By: Elias Zarco Gonzalez

. Introduction

In recent times, I have placed growing attention on the environment as it relates to climate change. Many factors influence climate change, the most notable one being air pollution. The objective of this project is to investigate air pollution further and see if it affects the quality of life across various U.S. regions. Specifically, the goal is to see whether areas with poorer air quality have lower indicators of quality of life, including family income, education level, and health insurance coverage. Statistical modeling and analysis is going to be used to establish and quantify potential relationships between air quality and these indicators. To achieve this, publicly available air quality and socioeconomic data sets will be used, utilizing both descriptive statistics as well as inferential methods, and building regression models to establish the strength and nature of such relationships. By analyzing and modeling these relationships, the goal is to draw meaningful conclusions that can support broader discussions about environmental policy and social well-being.

II. Data Collection and Preprocessing

In order to perform a precise statistical analysis of the relationship between overall air quality, and economical and social wellbeing indicators, it was decided to use government provided datasets due to their reliability and geographic coverage. Two publicly available datasets were used—the EPA's 2024 Annual AQI Report and the 2023 1-year American Community Survey—both of which report data at the Core-Based Statistical Area (CBSA) level, ensuring they share the same geographic boundaries for precise comparison. The process of collecting and processing the datasets was the following:
EPA's 2024 Annual AQI Report (https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual):

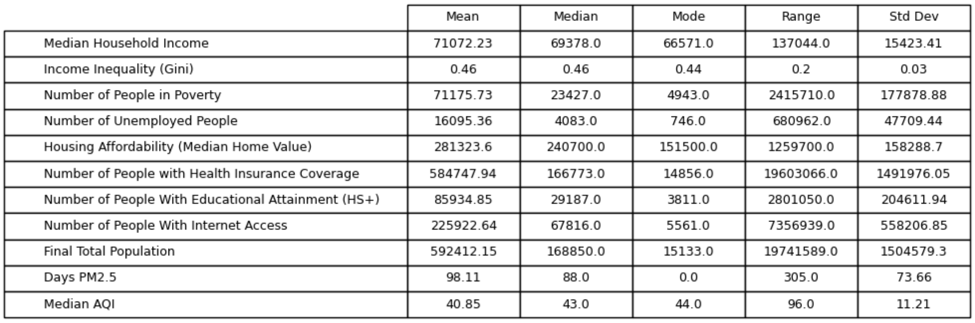
- The AQI Report serves as a benchmark for different air quality indicators like the median and maximum values of AQI, and days with PM(particulate matter) of more than 2.5.
- The dataset can be directly downloaded as a CSV file from the EPA website, almost no preprocessing as most of the data was already clean, only a small amount of NA values were deleted.

American Community Survey 2023 (https://www.census.gov/data/developers/data-sets/acs-1year.html)

- The American Community Survey included several socioeconomic indicators reflecting the overall quality of life of the different CBSAs. Some of the most important data retrieved include Gini's Income inequality Index, Median Household Income, Number of People in Poverty and Number of People with Educational Attainment (High School Diploma +).
- This dataset can be directly retrieved by an API, this made the preprocessing and collection step more complex than usual, however, python modules like Pandas and general strategies learned on the Data Input and Manipulation class proved to be useful.
- Preprocessing includes filtering CBSAs to make sure that the ones retrieved matched with the CBSAs on the AQI file, handling API retrieval errors and renaming all the API-specific variable names using a dictionary. If you want to see more about the data collection and preprocessing step feel free to visit the project's Google Collab: https://colab.research.google.com/drive/1j_hIhv1HEpQj8tHlScZ1g6sOrX7Z60G3.

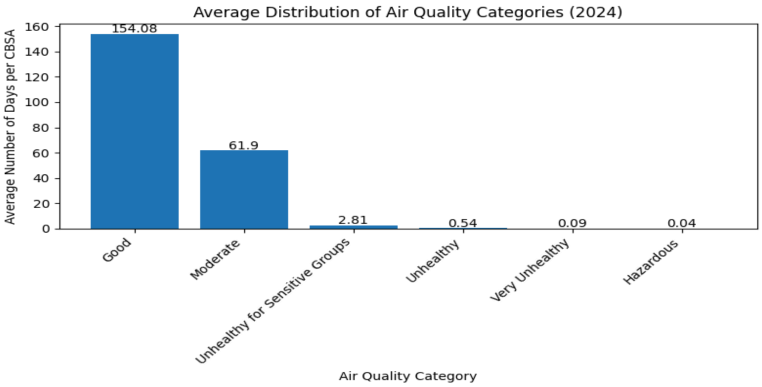
III. Descriptive Statistics

In order to get the summary statistics for the data relevant to the development of this project, the useful **scipy.stats.summary** function was used to get exact statistics for the indicators for the project. Finally, Matplotlib was used to get a simple and concise representation of the descriptive statistics.

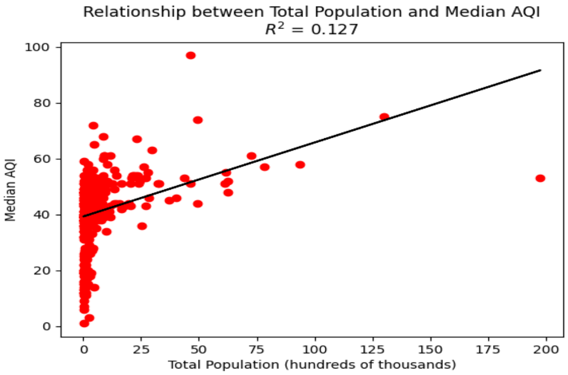


	Mean	Median	Mode	Range	Std Dev
Median Household Income	71072.23	69378.0	66571.0	137044.0	15423.41
Income Inequality (Gini)	0.46	0.46	0.44	0.2	0.03
Number of People in Poverty	71175.73	23427.0	4943.0	2415710.0	177878.88
Number of Unemployed People	16095.36	4083.0	746.0	680962.0	47709.44
Housing Affordability (Median Home Value)	281323.6	240700.0	151500.0	1259700.0	158288.7
Number of People with Health Insurance Coverage	584747.94	166773.0	14856.0	19603066.0	1491976.05
Number of People With Educational Attainment (HS+)	85934.85	29187.0	3811.0	2801050.0	204611.94
Number of People With Internet Access	225922.64	67816.0	5561.0	7356939.0	558206.85
Final Total Population	592412.15	168850.0	15133.0	19741589.0	1504579.3
Days PM2.5	98.11	88.0	0.0	305.0	73.66
Median AQI	40.85	43.0	44.0	96.0	11.21

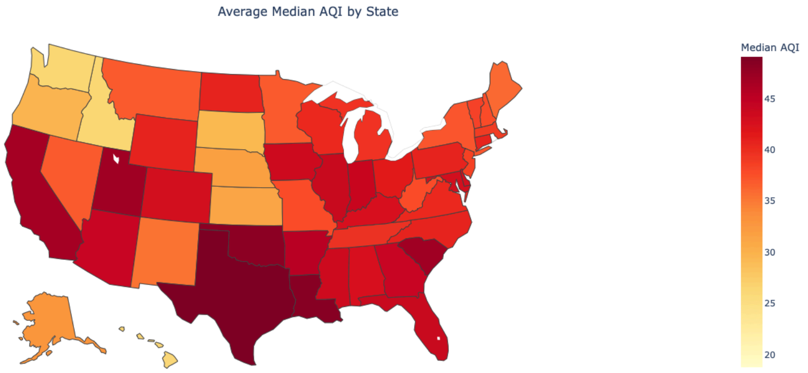
Here is a frequency bin that shows the distribution of the different verbal descriptions of air quality. These descriptions range from Good Days to Hazardous days. This graphical representation will provide an insight on the overall air quality of life across the United States.



A Scatterplot that visualizes the relationship between the population size of the CBSA and the median AQI was constructed. The visualization seeks to answer the question: Does more people imply worse quality of air? On the visualization, the dispersion between the different data points theR^2 value of 0.127 confirms that running a regression model between Total Population and Median AQI would not be effective.



Finally, a Choropleth map that shows the Median AQI distribution across different states in the U.S was constructed. This map shows the different trends of air quality across the different geographical regions.



IV. Statistical Inference

Null Hypothesis (H_0): There is no difference in the mean of median household income between high-pollution and low-pollution areas.
Alt Hypothesis (H_1): There is a difference in the mean of the mean household income between high-pollution and low-pollution areas.

A **two-sample t-test** will be used to compare the mean of median household income between regions with high pollution and regions with low pollution. This test is appropriate because it compares the means of two independent populations with unknown variances.

High- and low-pollution regions will be defined using the top and bottom quartiles of the "Median AQI" variable. Median household incomes from these two groups were compared using a two-sample t-test assuming unequal variances. After running the test in Python, these were the relevant statistical values:

Test Statistic (t_0): -1.955
Critical t-value ($\pm t_{\alpha/2}$): ±1.969
Degrees of Freedom: 270.546
Mean Income (Low Pollution): $68,942
Mean Income (High Pollution): $72,606

Since the absolute value of the test statistic ($|-1.955|$) is less than the critical value (1.969), the null hypothesis can't be rejected. This means that there is not enough statistical evidence to conclude a significant difference in median household income between high- and low-pollution areas.

Even though the result is not statistically significant at the 5% level, the group mean difference observed is still significant and could indicate that other variables (like urbanization or industrialization) could be behind the relationship between pollution and income.

V. Regression Analysis

To determine if a region's air quality has an impact on healthcare access, a simple linear regression analysis was conducted,the median AQI was set as the independent variable and the number of individuals with health insurance coverage in each CBSA as the dependent variable.

The regression model is:

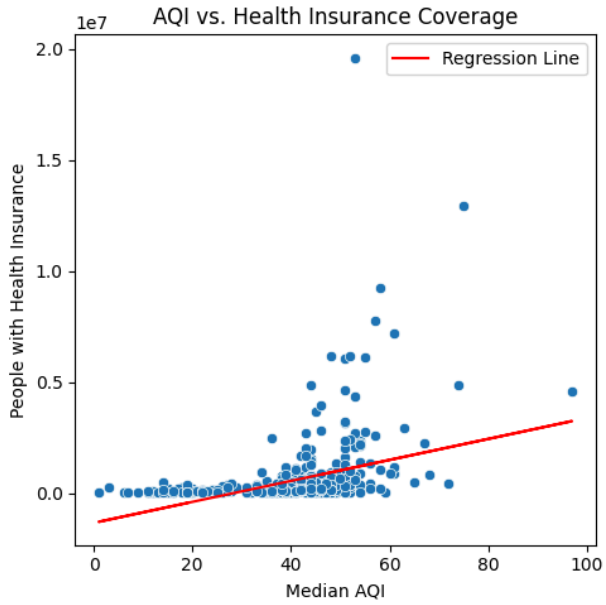
$$\text{HealthInsurance} = \beta_0 + \beta_1 \cdot \text{AQI} + \varepsilon$$

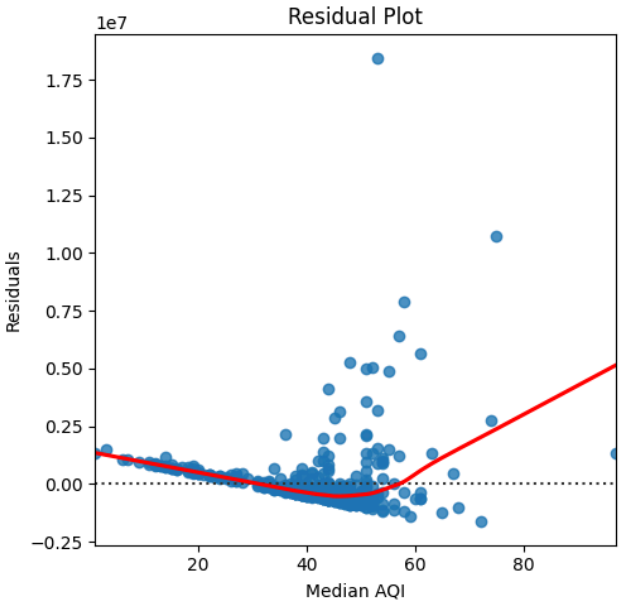
After fitting the model using the dataset, the following estimates were obtained:
- **Intercept (β_0):** -1,346,769.50
- **AQI Coefficient (β_1):** 47,279.69

These results show that with every one-unit increase in AQI, approximately 47,280 individuals have additional health insurance coverage in a CBSA. The model's R2 value was 0.126, indicating air quality accounts for 12.6% of variation in health insurance coverage across regions.

Residual analysis revealed no violations of linearity assumptions, and the residuals were roughly normally distributed. Even though the relationship is counterintuitive, it may be driven by the fact that elevated levels of AQI are most commonly found in larger, urban communities, and locales that are also likely to have greater access to medical care and higher insurance coverage rates.

Overall, this regression reaffirms the assumption that air quality is statistically linked to certain quality-of-life indicators, although this specific correlation is brought about by related factors of urbanization.





VI. Advanced Statistical Method

The R-squared value was very low in the original Linear Regression, so a **Multiple Linear Regression** was performed with the median household income as the response variable. A group of predictors were adjusted based on their potential effect on income, public health and access prominence: Median AQI, Number of Unemployed People, Number of People With Educational Attainment (HS+), Number of People with Health Insurance Coverage, Number of People With Internet Access, Final Total Population.

The Multiple Linear Regression Model:

Median Household Income =
$\beta_0 + \beta_1 \cdot$ Median AQI + $\beta_2 \cdot$ Number of Unemployed People +
$\beta_3 \cdot$ Number of People With Educational Attainment (HS+) +
$\beta_4 \cdot$ Number of People with Health Insurance Coverage +
$\beta_5 \cdot$ Number of People With Internet Access + $\beta_6 \cdot$ Final Total Population

Model Summary:

R-squared: 0.234
Adjusted R-squared: 0.225
F-statistic: 24.296
Model p-value: < 0.001

These values indicate that the model explains approximately 23.4% of the variability in median household income, and the model as a whole is statistically significant.

Coefficients and significance:

Variable	Coefficient	P-value
Intercept (const)	70323.486	0
Median AQI	-47.234	0.443
Number of Unemployed People	-0.422	0
Educational Attainment (HS+)	-0.135	0
Health Insurance Coverage	-0.095	0.243
Internet Access	0.021	0.378
Final Total Population	0.121	0.132

VII. Conclusion

A statistical analysis was conducted to examine the relationship between quality of life and air pollution across the United States. An initial hypothesis test was conducted for comparing median incomes of households in low- and high-pollution areas; however, the results did not yield statistical significance. Differences in group means noted though suggested that some variables could influence the correlation between income levels and pollution.

A simple linear regression provided a low R^2 of 0.126, which would indicate that air quality in isolation has minimal explanatory ability for variation in measures of quality of life. To allow for the effect of multiple factors to be considered, a multiple linear regression model was specified. This had an R^2 of 0.234, implying that socioeconomic outcomes are likely influenced by a collection of intercorrelated variables and not due to pollution itself.

The relatively low statistical result is possibly a consequence of multicollinearity among predictors and the deficiencies of analyzing aggregate data, which often will obscure larger individual-level trends. Despite these challenges, the analysis provided useful insight into the complex interplay among environmental and socioeconomic factors and gained hands-on experience with advanced statistical methods.