Analyzing the Impact of Air Quality on Life Expectancy Using Data Science Methods

Introduction

This project investigates whether there is a significant relationship between air quality and life expectancy across different countries. The goal is to explore how environmental conditions, particularly air pollution, may affect human health and longevity. I combined datasets on life expectancy and air quality to analyze patterns and correlations using various data science techniques.

The datasets used in this project include:

1. Life Expectancy Data from the World Health Organization (WHO), which covers health and economic data for 193 countries in 2015.

2. Air Quality Data covering 2022 to 2025, providing average Air Quality Index (AQI) values and category classifications (e.g., Good, Moderate, Unhealthy) for 142 countries.

Due to differences in data periods, I used the most recent year for life expectancy (2015) and matched it with average AQI values from 2022–2025, focusing on countries where both data sources overlapped.

Exploratory Data Analysis (EDA)

The EDA phase focused on understanding the general characteristics of the datasets and identifying initial trends. I started by cleaning the data, ensuring consistency in country names, and handling missing values. Then, I merged the datasets based on country.

Key observations from EDA include:

\* A moderate negative correlation between average AQI and life expectancy (-0.3097), suggesting that higher air pollution tends to be associated with shorter lifespans.

\* Countries categorized under 'Low Risk' (AQI ≤ 50) had the highest average life expectancy (76.94 years), while 'High Risk' countries (AQI > 100) had the lowest (70.31 years).

\* Developed countries not only had better air quality but also showed significantly higher life expectancy than developing countries.

These patterns highlighted important disparities and formed the basis for further statistical testing.

Hypothesis Testing

To assess whether the observed differences were statistically significant, I conducted two main hypothesis tests:

1. Correlation Test:

\* Pearson correlation coefficient between AQI and life expectancy: -0.3097

\* p-value: 0.0011

\* Conclusion: The negative correlation is statistically significant, indicating that air quality and life expectancy are inversely related.

2. T-Tests for Group Comparisons:

\* Life expectancy by AQI Risk Groups:

\* Low Risk (AQI ≤ 50): 76.94 years

\* Moderate Risk (AQI 51-100): 73.15 years

\* High Risk (AQI > 100): 70.31 years

\* The difference between Low Risk and High Risk groups is statistically significant.

\* Life expectancy by Development Status:

\* Developed Countries: 80.77 years

\* Developing Countries: 71.16 years

\* The gap of 9.61 years is also statistically significant.

These tests confirmed that air quality is a key factor influencing national life expectancy.

Limitations

Several limitations should be considered when interpreting these results:

\* The datasets are from different time periods, which may affect the accuracy of correlations.

\* Country-level averages may mask regional disparities within nations.

\* Correlation does not imply causation; other socio-economic and healthcare factors likely play a role.

\* The sample includes only \~108 countries with complete data, which may limit generalizability.

Conclusion

This project demonstrates a statistically significant relationship between air quality and life expectancy. Countries with cleaner air tend to have longer lifespans, especially in developed nations. However, due to data limitations and potential confounding variables, further research is needed to better understand causality.

Future work could involve:

\* Using time-aligned datasets to improve accuracy.

\* Incorporating pollutant-specific data (e.g., PM2.5, NO2).

\* Conducting regional or city-level analysis.

\* Applying machine learning models to predict life expectancy based on environmental and socio-economic variables.

This analysis shows how public health and environmental data can be combined to gain insights into global well-being and highlight the importance of pollution control policies.

References

\* WHO Life Expectancy Dataset

\* Air Quality Dataset (2022–2025)