DSA 1080A Group Project SS 2024

Project: Analyzing Global Air Quality Index (AQI) Data

Overview

This project involves analyzing air quality index (AQI) data from various cities around the world. The goal is to identify trends, compare air quality across different regions, and investigate the impact of certain events (like holidays, wildfires, or policy changes) on air quality. The project will use Python's pandas for data manipulation, matplotlib and seaborn for data visualization, and possibly sklearn for some basic predictive modeling.

Week 1: Introduction and Data Collection

Objective: Learn the basics of pandas and how to import data.

Tasks:

Introduction to pandas: Understanding Series and DataFrames.

Collect AQI data from public datasets available online. Websites like OpenAQ provide access to global air quality data.

Basic data inspection and cleaning: Handling missing values, duplicate entries, and data types.

Week 2: Data Exploration and Analysis

Objective: Dive deeper into data manipulation and start exploring the dataset.

Tasks:

Data exploration: Calculating summary statistics and understanding the distribution of data using histograms, box plots, and scatter plots.

Time-series analysis: Analyzing AQI trends over time for selected cities.

Comparing AQI data across different cities or regions to identify areas with the best and worst air quality.

Week 3: Visualization and Interpretation

Objective: Learn data visualization techniques and present findings.

Tasks:

Create compelling visualizations to represent the findings using matplotlib and seaborn. Examples include time series plots, bar charts comparing AQI across cities, and heatmaps of AQI by month.

Interpret the results: Discuss the potential reasons behind air quality trends and differences between cities or regions.

Prepare a final presentation or report summarizing the methodology, findings, and insights.

Optional Extensions:

Predictive Modeling: Use basic machine learning models from sklearn to predict future AQI levels based on historical data.

Impact of Events: Analyze the impact of specific events (e.g., New Year's Eve, major wildfires,

For the project on analyzing the Global Air Quality Index (AQI), a publicly accessible and comprehensive dataset is essential. Here's a suggested dataset and how to obtain it:

OpenAQ Dataset

OpenAQ aggregates and shares open air quality data from around the world. It offers an accessible platform with historical and near-real-time data on various pollutants, which is ideal for your AQI analysis project.

Key Features:

Global Coverage: Data from thousands of stations across the world, covering major cities and regions.

Pollutants Measured: Includes key pollutants such as PM2.5, PM10, CO, NO2, SO2, O3, and more, which are crucial for calculating the AQI.

Open and Free: The dataset is openly available for educational and research purposes.

How to Access the Data:

OpenAQ API: The OpenAQ API allows programmatic access to the data. You can filter the data by date, location, and pollutant. This approach is great for getting the latest data or for specific queries. API documentation is available at https://docs.openaq.org/.

Downloadable Files: For some analyses, working with a static dataset might be preferable. OpenAQ periodically provides data dumps that can be downloaded from their website or GitHub repository. Check https://openaq.org/#/download or their GitHub page for the latest dumps.

Suggested Steps for Using the Data:

Define Your Scope: Decide on the specific pollutants, time frame, and locations you want to analyze. This step will help streamline the data collection process.

Collect Data: Use the OpenAQ API to collect data based on your scope. For beginners, it might be easier to start with a smaller dataset (e.g., one year of data for PM2.5 in several cities).

Data Cleaning: Prepare the data for analysis by cleaning and structuring it as needed. This may involve dealing with missing values, converting data types, and aggregating data points.

Analysis: Conduct your analysis, exploring trends, comparisons, and correlations within the data.

Visualization and Reporting: Create visualizations to illustrate your findings and compile a report or presentation to share your insights, industrial policy changes) on air quality.

Project Report or Presentation

Introduction

Brief overview of the project objectives.

Importance of studying AQI and its impact on health and environment.

Data Collection

Sources of AQI data (e.g., OpenAQ).

Description of the dataset, including the variables and time period covered.

Data Cleaning and Preparation

Summary of data cleaning steps: handling missing values, duplicates, and any data transformation.

Initial data exploration findings: distribution of AQI values, missing data patterns, etc.

Data Analysis

Summary Statistics: Central tendency and variability of AQI across different locations and times.

Comparative Analysis: Comparison of AQI among different cities or regions, highlighting areas with the best and worst air quality.

Trend Analysis: Analysis of AQI trends over time, identifying any seasonal variations or long-term trends.

Data Visualization

Time Series Plots: Showing AQI trends for selected cities over time.

Comparative Bar Charts: Illustrating average AQI levels across different cities or regions.

Heatmaps: Visualizing AQI variations by time (e.g., monthly) for specific locations.

Scatter Plots: (Optional) Exploring relationships between AQI and other relevant variables (e.g., temperature, humidity).

Interpretation and Discussion

Insights on AQI trends and their potential causes (seasonal changes, policy impact, special events).

Regions or times of particular concern and possible reasons behind these patterns.

Limitations of the analysis and suggestions for further research.

Conclusion

Recap of the key findings and their implications for public health and policy.

Recommendations for improving air quality based on the analysis.

2. Code and Notebooks

Well-commented Python notebooks containing all the code used for data collection, cleaning, analysis, and visualization.

Documentation of any challenges faced during the project and how they were addressed.

3. Visual Outputs

A collection of visualizations created during the project, possibly as a separate file or embedded within the project report/presentation.

These should be clear, labeled correctly, and include interpretations.

4. Presentation (if applicable)

A concise, engaging presentation summarizing the project's key points, findings, and visualizations.

Could be in the form of slides, a video, or a live presentation to the class or a wider audience.

5. Reflection

A section reflecting on what the team learned from the project, including both the data science skills and insights about air quality.

Challenges encountered and how they were overcome.