**Comparison of Air Sensor Data Quality Across Utah Counties**

Sensor Air Quality (github.com/huotra/BMI-6016-Group-2-Final-Project)

* Huong Tran (u0578864), [htran735@gmail.com](mailto:htran735@gmail.com)
* Kolby Bray (u01468279), [kolbybray@gmail.com](mailto:kolbybray@gmail.com)
* Harshitha Josyula (u1427994), [u1427994@utah.edu](mailto:u1427994@utah.edu)
* Fnu Kavita (u1418910), [u1418910@umail.utah.edu](mailto:u1418910@umail.utah.edu)

**Background and Motivation:**

Our motivations for choosing this project stem from a combination of personal and professional interests. As residents of Utah, we are acutely aware of the air quality challenges faced by our community. Utah frequently experiences poor air quality due to a combination of factors such as geographic location, weather patterns, industrial activities, and population growth. These issues not only impact our daily lives but also pose significant public health risks, particularly for vulnerable populations such as children, the elderly, and individuals with respiratory conditions. As low-cost sensors become increasingly prevalent in monitoring air quality, there is a pressing need to ensure the reliability and accuracy of the data they produce, particularly in research settings where data integrity is crucial. The background information provided underscores the significance of this issue, highlighting the rapid deployment of low-cost sensors and the associated challenges of data quality and reliability.

**Project Objectives:**

Our main objective for this project is to conduct a comprehensive analysis of data quality pertaining to various Utah counties. We also will provide a visual comparison of air sensor data quality differences across selected counties within the state. Our purpose is to collect, clean, and present air sensor data quality from across the state to show where air sensor data quality is highest and where air sensor data quality is lowest. Our goal is to create visualizations from the data that demonstrate data quality. This project will have many benefits. Principle among them would simply be public knowledge. Once people and organizations are aware of the discrepancies between counties, action can then be taken. Policy making and resource allocation would be affected positively, allowing for more informed decision making. Scientific research would also benefit, with more strategic research in areas with lower data quality.

**Data:**

We will be utilizing air quality data from PurpleAir. PurpleAir operates a network of sensors deployed worldwide, including in Utah. The PurpleAir dataset contains real-time measurements of various air pollutants, including particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, and volatile organic compounds (VOCs), among others. These measurements are reported at regular intervals and are accessible through the PurpleAir website.To access the PurpleAir data in Utah, we will be utilizing the PurpleAir Sensor Map interface, which allows users to select specific geographic areas and timeframes for data retrieval.

Link to PurpleAir data: [PurpleAir Sensor Map](<https://map.purpleair.com/>)

**Data Processing:**

Given the nature of the data from PurpleAir, we expect a large amount of time to be allocated in cleaning the data, which will include checking and accounting for missing data, organizing based on time and location, and determining the appropriate sensors to collect data from to create a meaningful comparison between data quality in different counties. We expect to extract the averages, medians, and standard deviation from the datasets for each county as well as data quantity, comparing the amount of data generated in each county.

Data processing will be implemented by preprocessing to clean and standardize the data across time and the selected counties, quality assurance to identify and address anomalies and missing data, statistical analysis, and finally visualization to present our findings.

**Design:**

For the design of the project, we intend to create visualizations of the data for selected counties and a combined average visualization. Bar graphs showing data quality indicators in each county will be used and a combined graph will be used to visualize data quality across counties. Alternatively, line graphs that show data quality over time would also be effective. We chose these designs because they are the best for showing the differences between counties based on the data quality parameters that we will set.

**Must-Have Features:**

* + Interactive Map Visualization: A core feature of our project will be an interactive map that visualizes air quality data across Utah counties. This map will allow users to select specific counties to view detailed air quality metrics, such as PM2.5 and PM10 levels, ozone, nitrogen dioxide, and VOCs. The map should provide a clear, at-a-glance understanding of air quality across the state, highlighting areas with particularly poor or good air quality.
  + Time-Series Analysis: Incorporate a time-series analysis feature that allows users to track air quality changes over specific periods. This functionality will enable the examination of trends, seasonal variations, and potential impacts of policy changes or environmental events on air quality. It is essential for understanding long-term changes and identifying patterns.
  + Data Quality Indicators: Given the importance of data integrity, the project must include indicators of data quality for each sensor's readings. These indicators could be reliability scores, the age of the sensor, or maintenance records, providing users with confidence in the data they are viewing.
  + Detailed Data Metrics Dashboard: A dashboard that displays detailed metrics of air quality for selected counties or sensors. This dashboard should include average, median, and standard deviation for air pollutants, as well as historical data visualization to track progress or deterioration of air quality over time.

**Optional Features:**

* + Data Export Functionality: Implement a feature that allows users to export processed air quality data in common file formats such as CSV or Excel. This feature would enable users to further analyze the data or integrate it into their own projects or reports.
  + Customizable Data Filters: Add filters to the data visualization interfaces that allow users to customize the display based on specific criteria such as time range, pollutant type, or sensor location. This feature would enhance user experience by providing flexibility in data exploration
  + Email Alerts: Integrate an email notification system that alerts users when air quality in their selected counties exceeds certain thresholds or when there are significant changes in pollution levels. This feature would provide timely updates to users, enabling them to take necessary precautions or actions.
  + Educational Resources: A section dedicated to educating the public about air quality, its health impacts, and ways to improve air quality at a local level. This feature could include tips, policy recommendations, and resources for further reading.

**Project Schedule:**

Week1:

* Make changes as necessary after meeting with professor
* Data Collection & Initial Cleaning:
  + Gather air quality data from PurpleAir.
  + Address missing data and standardize formats.

Week 2:

* Further Data Processing and Basic Visualizations Implementation::
  + Derive averages, medians, and standard deviations.
  + Develop prototype visualizations for map and time-series analysis.

Week 3:

* Advanced Data Analysis and Prototype Refinement:
  + Conduct statistical analysis.
  + Refine visualizations based on insights.
  + Improve design and functionality.

Week4

* Finalize Visualizations:
  + Complete development of interactive map and dashboard.

Week 5:

* Documentation and presentation:
  + Write user guides and project reports and Create presentation slides and practice delivery.