Rise South City – Team 2 Capstone Documentation

This document summarizes the methodology, data sources, analytical code, and project decisions made by Team 2 for Rise South City's Capstone project. It is organized into the following sections:

- Data — Details on all datasets used and how they were processed.
- Code — Overview of the scripts and notebooks in the code/ directory.

Data

Air Quality Data

We used air quality data from three main sources:

- Clarity Sensor Data (clean_clarity.csv): Provided by Rise South City, this dataset spans from October 30, 2024 to March 31, 2025.
- PurpleAir API Data (clean_api_purpleair.csv): Collected directly by our team, this dataset includes more recent readings from March 30, 2024 to May 3, 2025.
- Historical PurpleAir Data (clean_purpleair.csv): A merged file combining older historical exports with the newer API data, covering a broader period from December 27, 2018 to April 9, 2025.

Final risk scores are saved in:

- tracts_with_combined_aqi.csv and tracts_with_combined_aqi.geojson
  These contain daily PM2.5 medians aggregated by census tract and adjusted with source-specific weights (see Air Quality Risk Index section).

Health Risk Data

- Health Risk Index: Stored in health_risk_index.csv, this file contains normalized (0 to 1) health vulnerability scores at the census tract level.
- Raw Indicator Sources:
  - all_indicators.csv: Consolidates key health-related variables from:
    - CalEnviroScreen 4.0 (via ACSST5Y2023.S2701-Data.csv)
    - San Mateo County's All Together Better Initiative (stored as separate files in smc_indicators/)
  - Health Equity Index: Must be downloaded separately and added to the data/ directory before processing.

Health indicator processing and scoring are conducted in:

- health.ipynb: Computes the final Health Risk Index as a geometric mean of key indicators.
- health_preproc.ipynb: Standardizes and cleans all input metrics before merging.

Figures

Stored in the figures/ folder, these visual assets were used in weekly meetings, final presentations, and the live dashboard. They include sensor distribution maps, PM2.5 daily trend charts and explanatory figure captions for all visualizations

Miscellaneous Data

Stored in the misc/ folder:

- census.geojson: Boundary file for San Mateo County census tracts (used for spatial joins and maps).
- air_traffic.csv: Monthly SFO airport traffic data, used to investigate correlations between flight activity and pollution.

Code

The code/ directory contains all scripts and notebooks, logically grouped by task. In-line comments are provided throughout for clarity and reproducibility.

Preprocessing

Scripts here clean and standardize air quality data from Clarity and PurpleAir:

- clean_clarity.py: Cleans the raw Clarity dataset. Tasks include:
    - Timestamp parsing
    - Renaming columns
    - Removing invalid values
- clean_api_purpleair.ipynb: Processes API-collected PurpleAir data, converting it to daily averages aligned with Clarity's format.
- clean_purpleair.py: Merges and cleans historical and API PurpleAir datasets.
- purpleair_wrapper.py: Wrapper for the PurpleAir API. Allows automated retrieval of sensor readings based on bounding boxes or sensor IDs.
- combine_air_quality_data.py:
    - Filters both data sources to a consistent timeframe (March 30, 2024 – March 31, 2025)
    - Assigns readings to census tracts using spatial joins
    - Aggregates daily PM2.5 by tract

Air Quality Risk Index

- calculate_sensor_weights.py:
  Calculates source weights based on overlapping sensor readings from Clarity and PurpleAir at the same location (Rollingwood Elementary):
    - Clarity: 0.76
    - PurpleAir: 0.24
- combine_air_quality_data.py (continued):
  Applies sensor weights and computes final tract-level AQI scores. Outputs are saved in both CSV and GeoJSON formats.

Health Risk Index

- health.ipynb:
    - Computes the Health Risk Index using indicators from all_indicators.csv
    - Combines respiratory vulnerability and broader structural health inequities
- health_preproc.ipynb:
    - Prepares the dataset by processing raw indicator CSVs
    - Includes guidance on required renaming and standardization steps

Sensor Predictability & Consistency

- predictability.ipynb:
  Computes two scores:
    - Consistency Score: Predicts a sensor's next-day PM2.5 using its previous 7 days
    - Predictability Score: Uses the 5 geographically nearest monitors to predict the current-day PM2.5 reading
    -

Both scores help assess sensor reliability and network redundancy.

Dashboard

- streamlit_app.py:
  Main application file for the Streamlit dashboard. To run locally:

```shell
python -m streamlit run code/streamlit_app.py
```

Key Features:

- Map view of the Composite Risk Score for each tract
- Adjustable slider to set the relative weight w for:
  Risk = w * HRI + (1 - w) * AQI
- Address-based lookup tool and visual overlays of air monitor locations

If you are a future team continuing this project, we recommend starting with the README in the GitHub repository, then walking through the notebooks in the following order:

1. clean_clarity.py and clean_api_purpleair.ipynb

2. combine_air_quality_data.py
3. calculate_sensor_weights.py
4. health_preproc.ipynb → health.ipynb
5. predictability.ipynb
6. streamlit_app.py

For any questions about file dependencies, refer to the Code Book provided in the code/ directory.