Data source and data property

1. **Data Resource:** Weather data is a combined data from data.gov.sg, and the detail information of sub dataset is provided below.

Attribute	Dataset	Citation
wind_speed	Wind Speed across Singapore	National Environment Agency. (2024). Wind Speed across Singapore (2024) [Dataset]. data.gov.sg. Retrieved March 11, 2025 from https://data.gov.sg/datasets/d_7677738484067741bf3b56ab5d69c7e9/view
rainfall	Rainfall across Singapore	National Environment Agency. (2024). Rainfall across Singapore (2024) [Dataset]. data.gov.sg. Retrieved March 11, 2025 from https://data.gov.sg/datasets/d_6580738cdd7db79374ed3152159fbd69/view
relative_humidity	Relative Humidity across Singapore	National Environment Agency. (2024). Relative Humidity across Singapore (2024) [Dataset]. data.gov.sg. Retrieved March 11, 2025 from https://data.gov.sg/datasets/d_2d3b0c4da128a9a59efca806441e1429/view
air_temperature	At Temperature across Singapore	National Environment Agency. (2024). Air Temperature across Singapore (2024) [Dataset]. data.gov.sg. Retrieved March 11, 2025 from https://data.gov.sg/datasets/d_66b77726bbae1b33f218db60ff5861f0/view

2. **Data Crawling**:
   Since data.gov.sg provides a reliable API, we wrote a Python function to scrape the data and compile it into a CSV file. The code is available on GitHub under the repository name ST5188_Weatherdata_Crawl.
3. **Data Processing:**
   - The sub-datasets are all sourced from the Singapore government, ensuring a consistent format. We selected “timestamp”, “station_name”, and “station_id” as key identifiers, which contribute to both data visualization and model building.
   - Rainfall data is collected from multiple testing sites, with observations recorded almost every hour. However, it does not rain at every observation point at all times. Other datasets have limited observation sites due to equipment constraints, leading to a significant number of missing values in the rainfall dataset. To improve data quality, we only collect rainfall data when it is actually raining (i.e., when the recorded value is greater than 0).
4. **Difficulties**:
   Since the weather data is recorded hourly, it results in an enormous dataset—approximately 360,000 rows. Processing the entire dataset can take over 12 hours, and a major challenge is that the crawling script may unexpectedly shut down without any apparent reason. Additionally, running such a large-scale process on a laptop is difficult.