Project Report: Milestone 1 - Data Collection, Preprocessing, and Exploratory Data Analysis

---

1. Project Objective

This project aims to develop a **chatbot** for analyzing and interacting with **U.S. power plant data** and environmental trends. The goal is to compare **U.S. power plant data** with **U.S. pollution levels** and **temperature changes.** Users will be able to query the chatbot for insights and dynamically explore trends within the U.S.

2. Type of Project

- **Conversational Agent (Chatbot)**: Allows users to query the power plant database and related datasets using natural language.

3. Data Used

Primary Dataset: U.S. Power Plant Database

- **Source**: Global Power-Plants (Kaggle)
- **Description**: Contains information on power plants in the United States, including location, capacity, fuel type, and operational status.
- **Dimensions**: 8644 records, 18 variables (Power Plant Database)
- **Key Variables and Descriptions**:
  - `country`: The country where the power plant is located (all entries are 'USA').
  - `country_long`: Full name of the country (United States of America).
  - `name`: Name of the power plant.
  - `capacity_mw`: The maximum electrical output capacity of the power plant in megawatts.
  - `latitude`, `longitude`: Geographic coordinates of the power plant location.
  - `primary_fuel`: The main fuel source used by the power plant (e.g., solar, gas, coal, wind, etc.).
  - `commissioning_year`: The year the power plant began operations.
  - `owner`: The entity that owns the power plant.
  - `year_of_capacity_data`: The year when capacity data was last updated.
  - `generation_gwh_2013-2017`: Annual electricity generation (in GWh) from 2013 to 2017.
  - `cluster`: Cluster ID assigned for power plant grouping.

Additional Datasets

- **U.S. Pollution Dataset** (Collected from AirPure API based on coordinates)
    - **Source**: AirPure API
    - **Description**: Contains air quality index (AQI) data for various locations in the U.S.
    - **Dimensions**: 1506 records, 14 variables (Air Quality Dataset)
    - **Key Variables and Descriptions**:
        - `DateObserved`: The date when the air quality measurement was recorded.
        - `HourObserved`: The hour at which the measurement was taken.
        - `LocalTimeZone`: The local time zone of the reporting station.
        - `ReportingArea`: The city or region where the measurement was recorded.
        - `StateCode`: The U.S. state abbreviation.
        - `Latitude`, `Longitude`: Geographic coordinates where the air quality measurement was taken.
        - `ParameterName`: The pollutant measured (e.g., Ozone, PM2.5, PM10).
        - `AQI`: Air Quality Index value indicating the overall air quality level.
        - `CategoryNumber`: Numeric classification of AQI category.
        - `CategoryName`: Descriptive category of AQI (e.g., Good, Moderate, Unhealthy).
        - `year`, `month`: Extracted fields indicating the year and month of observation.
        - `cluster`: Cluster ID assigned for AQI station grouping.

- **U.S Temperature dataset (Collected from NOAA using Open-meteo API)**
    - **Source:** Open-meteo API
    - **Description:** This dataset contains temperature data for various locations, recorded quarterly.
    - **Dimensions:** 1800 rows and 7 columns
    - **Key Attributes:**
        - `cluster`: Cluster ID assigned to power plants.
        - `latitude`: latitude of cluster.
        - `longitude`: longitude of cluster.
        - `commissioning_year`: The year the power plant began.
        - `quarter`: quarters for 2 years before and after commissioning year.
        - `time`: The time when the temperature was captured.
        - `temperature_2m_mean`: Average temperature of the particular quarter at given latitude and longitude.

4. Tech Stack

- **Programming Languages**: Python
- **Libraries**:
    - **Data Manipulation:** Pandas, NumPy, Scipy
    - **Data Visualization:** Matplotlib, Seaborn, Plotly

5. Exploratory Data Analysis (EDA)

- **Clustering Methodology for AQI Data Collection:** To optimize AQI requests, **DBSCAN clustering** was applied to group power plants based on **geographic proximity** using haversine distance. Key steps:
    - **Data Cleaning:** Removed missing or invalid values for latitude, longitude, and commissioning year.
    - **Filtering Criteria:** Considered only plants commissioned in the **last 13 years**.
    - **Geospatial Clustering:** Applied **DBSCAN with a 100 km proximity threshold** to form clusters.
    - **Cluster Representation:** Selected a representative location per cluster for AQI data retrieval.
- **Descriptive Statistics**: Mean, median, standard deviation.

- **Data Preprocessing:**
    - **Handling Missing Values**: Identifying and imputing missing data.
    - **Outlier Detection**: Using visualization and statistical methods like IQR.
    - **Standardization**: Ensuring consistent units and formats.

- **Visualizations**:
    - Distribution of power plant capacities.
    - Power plants by primary fuel type.
    - Year-wise trends in commissioning power plants.
    - Geographic distribution of power plants.
    - Temperature by cluster location over time.
    - Temperature differences and impact after commissioning.

6. Key Insights from EDA

- **Power Plant Insights**:
    - The dataset contains **8,644** power plants across the U.S.
    - **Solar** is the most common fuel type in terms of the number of plants, followed by gas and hydro.
    - **Gas-based power** plants contribute the most to total capacity **(182,686 MW)**, followed by **wind (80,358 MW), coal (53,892 MW),** and **hydro (50,147 MW)**.
    - The top 5 largest power plant clusters contain the majority of power plants, with the largest cluster having over 1,600 plants.
    - Fuel type distribution across the largest clusters shows that gas, wind, and hydro dominate in terms of capacity.
    - The top power plant owners vary significantly in the number of plants and total capacity:
        - **Cypress Creek Renewables** owns the highest number of plants **(109)** but has a lower total capacity (810 MW).
        - **Pacific Gas & Electric Co**. has the highest total capacity **(3,700 MW)** but owns fewer plants **(81)**.

- **Air Quality Insights**:
    - **AQI Category Distribution**:
        - The majority of recorded AQI values fall into the "Good" category (1,266 records).
        - "Moderate" AQI levels account for 232 records, while only 6 and 2 records were categorized as "Unhealthy for Sensitive Groups" and "Unhealthy," respectively.
        - This suggests that most monitored locations maintain relatively clean air quality.
    - **Pollutant-Specific AQI Trends**:
        - **Ozone (O3)** is the most frequently recorded pollutant (642 records), followed by **PM2.5** (571 records) **and PM10** (293 records).
        - **PM2.5 exhibits the highest AQI variability**, with several extreme values exceeding 175, indicating occasional severe pollution spikes.
        - **PM10 has the lowest AQI levels on average**, but certain regions still show high particulate matter concentrations.
    - **Geospatial Trends**:
        - The highest AQI levels (poor air quality) are concentrated in **urban** and **industrial areas**, while **rural regions** tend to have significantly lower pollution.
        - **Montana (MT), Texas (TX), and Arkansas (AR)** have the highest number of AQI records, suggesting greater monitoring efforts in these states.
    - **AQI Trends Over Time and Clusters**:
        - Certain clusters show significant AQI deterioration, while others exhibit clear improvements.
        - Clusters with the most AQI change show up to **69 points of variation**, indicating areas of concern as well as zones with successful air quality interventions.
    - **Impact of Power Plants on AQI**:
        - **Gas power plants** contribute the **most capacity (3,892 MW)** in **AQI-improving clusters**, suggesting that cleaner energy sources correlate with better air quality.
        - **Hydro, oil, and coal plants** are dominant in **AQI-worsening areas**, reinforcing their association with emissions and air quality decline.
        - **Coal and oil plants** show direct correlations with worsening AQI trends, while **renewables (solar, wind)** are more prevalent in **cleaner-air regions**.

- **Temperature Insights:**
    - There are very few outliers particularly 2 i.e., the temperature is very low at those locations.
    - There are no missing or duplicate values.
    - The mean temperature is around 12 degrees Fahrenheit.

- There is a little bit high temperature in southern parts but over the years the temperature increases around 5 degrees in northern and central parts.
- This gives us the insight that there is impact of power plants on the temperature. The temperature increases over the period.

7. Project Timeline

Milestone	Task	Deadline
Milestone 1	Data Collection, Preprocessing, and EDA	Feb 23, 2025
Milestone 2	Feature Engineering & Data Modeling	Mar 21, 2025
Milestone 3	Tool Development & Final Presentation	Apr 23, 2025

8. GitHub Repository

- Repository Name: cap5771sp25-project
- Link: https://github.com/SainathChettupally/cap5771sp25-project.git
- Contributors: **Nagabhairava, Roop Yaswanth**

9. Next Steps

- Complete data preprocessing and visualize trends.
- Confirm additional datasets for integration.
- Start chatbot prototype development.
