Data Setup Guide

Purpose: Instructions to obtain and configure the required data files for this analysis.

Overview

This analysis requires **3 data files** in CSV format:

1. **Pesticide usage data** - California agricultural pesticide applications (2000-2022)
2. **Health & demographic data** - COPD hospitalization rates with confounding factors
3. **Population data** - County population numbers for normalization

---

Step 1: Create Your Data Folder

You need a folder named `Datasets` in your project directory.

Option A: Using File Explorer (Windows) or Finder (Mac)

1. Navigate to where you downloaded/cloned this project
2. Right-click inside the project folder
3. Select "New Folder" (Windows) or "New Folder" (Mac)
4. Name it: `Datasets` (capital D, no spaces)

Option B: Using Terminal/Command Line

```bash
# Navigate to your project folder first, then:
mkdir Datasets
```

Success Check: You should now see a folder called `Datasets` inside your project folder.

Step 2: Obtain the Required Data Files

File 1: `historical_data_2000_2022_filtered.csv`

Contents: Pesticide usage records across California counties

Required columns:

- `YEAR` - Year of pesticide application
- `CHEM_CODE` - Chemical identification code

- `TOTAL_LBS_AI` - Total pounds of active ingredient applied
- `TOTAL_ACRES_TREATED` - Total acres treated with pesticides
- `COUNTY_NAME` - California county name

File size: Approximately 1.1 million records (80-120 MB)

File 2: `copd_aqi_poverty_demographics.csv`

Contents: Health outcomes and confounding variables

Required columns:

- `Counties` - County name
- `Year` - Year of observation
- `Median AQI` - Air quality index
- `pct_under_18`, `pct_18_64`, `pct_65_plus` - Age distribution percentages
- `median_age` - Median age of county population
- `pct_AI/AN`, `pct_Asian`, `pct_Black`, `pct_Latino`, `pct_Multi_Race`, `pct_NH/PI`, `pct_White` - Racial/ethnic composition
- `COPD_Hospitalization_Rate` - **TARGET VARIABLE** (what we're predicting)
- `Poverty_AllAges_Percent_Est` - Poverty rate
- `Median_Household_Income_Est` - Median household income

File size: Approximately 1,300 county-year observations (100-200 KB)

File 3: `Population_Census_Numbers_2000_2025.csv`

IMPORTANT: This file is distributed as a ZIP archive (`Population_Census_Numbers_2000_2025.zip`) and **must be extracted before use**.

Contents: County population counts over time

Required columns:

- `County` - County name
- Multiple date columns (format: MM/DD/YY) - Population counts by date

File size: 58 counties × 26 years of data (50-100 KB when extracted)

How to Extract the ZIP File:

Windows:

1. Locate `Population_Census_Numbers_2000_2025.zip`
2. Right-click on the ZIP file
3. Select "Extract All..."
4. Choose your `Datasets` folder as the destination
5. Click "Extract"
6. Verify the CSV file is now in your `Datasets` folder

Mac:

1. Locate `Population_Census_Numbers_2000_2025.zip`
2. Double-click the ZIP file (it extracts automatically)
3. Move the extracted CSV file into your `Datasets` folder
4. Delete the ZIP file if desired

Linux/Command Line:

```bash
unzip Population_Census_Numbers_2000_2025.zip -d Datasets/
```

Step 3: Verify Your Folder Structure

Your project directory should look exactly like this:

```
your-project-folder/
|
|-- XGBoost_pesticide_copd_analysis_Completed.ipynb
|-- requirements.txt
|-- DATA_SETUP.md (this file)
|-- .gitignore
|
`-- Datasets/
    |-- historical_data_2000_2022_filtered.csv
    |-- copd_aqi_poverty_demographics.csv
    `-- Population_Census_Numbers_2000_2025.csv
```

Pre-Flight Checklist:

- □ `Datasets` folder exists in the project directory
- □ All 3 CSV files are in the `Datasets` folder
- □ Population file has been extracted from ZIP (not still as .zip)
- □ File names match exactly (case-sensitive, including underscores and .csv extension)
- □ All files open correctly (verify in Excel or text editor)

Step 4: Update the Notebook File Paths

The notebook currently contains hard-coded file paths that point to the original developer's computer. You must change these to relative paths.

Instructions:

1. Open the notebook: `XGBoost_pesticide_copd_analysis_Completed.ipynb`

2. Navigate to **Cell 2** (near the top, the cell that loads data)

3. Find these lines with long file paths:

```
python

df_pesticides2 = pd.read_csv('/Users/abciii/Library/Mobile Documents/com~apple~CloudDocs/Kil/AI4ALL/XGBoost_sets
df_confounders = pd.read_csv('/Users/abciii/Library/Mobile Documents/com~apple~CloudDocs/Kil/AI4ALL/XGBoost_set
df_population = pd.read_csv('/Users/abciii/Library/Mobile Documents/com~apple~CloudDocs/Kil/AI4ALL/XGBoost_sets/
```

4. Replace them with these relative paths:

```
python

df_pesticides2 = pd.read_csv('Datasets/historical_data_2000_2022_filtered.csv')
df_confounders = pd.read_csv('Datasets/copd_aqi_poverty_demographics.csv')
df_population = pd.read_csv('Datasets/Population_Census_Numbers_2000_2025.csv')
```

5. Save the notebook (File → Save, or Ctrl+S / Cmd+S)

Success Check: The new paths should be much shorter and simply say `Datasets/filename.csv`

Data Sources & Attribution

Original Data Sources

Data Type	Source	Website
Pesticide Usage	California Department of Pesticide Regulation (CDPR)	https://www.cdpr.ca.gov/docs/pur/purmain.htm

Data Type	Source	Website
COPD Hospitalization	California Health and Human Services Open Data Portal	https://data.chhs.ca.gov/
Demographics	U.S. Census Bureau	https://www.census.gov/data.html
Air Quality	EPA Air Quality System (AQS)	https://www.epa.gov/outdoor-air-quality-data

Analysis Coverage

- **Geographic:** 53 California counties
  - *Excluded counties:* Alpine, Lassen, Modoc, Mono, Sierra (insufficient health data)
- **Time Period:** 2000-2022 for raw data; 2005-2022 for analysis (after lag feature creation)
- **Observations:** 943 county-year combinations after data cleaning

Data Processing

- Pesticide data aggregated from individual application records to county-year totals
- Temporal lag features created: 1, 2, 3, 5, 10, 15, 20 years
- Cumulative exposure metrics: rolling windows of 3, 5, 10, 15, 20 years
- All pesticide metrics normalized per 100,000 population

Troubleshooting Common Issues

Problem: "FileNotFoundError: No such file or directory"

Possible causes and solutions:

1. **File names don't match exactly**
   - Verify spelling, capitalization, underscores
   - Ensure file ends with `.csv` (not `.csv.txt` or `.zip`)
2. **Files are in the wrong location**
   - Files must be directly in `Datasets` folder, NOT in a subfolder
   - `Datasets` folder must be in the same directory as the notebook
3. **You didn't update the notebook paths**
   - Return to Step 4 and verify you changed the file paths
   - Save the notebook after making changes

Problem: ZIP file won't extract

Solutions:

1. **Windows:** Ensure extraction software is available
   - Windows 10/11 has built-in ZIP support
   - Right-click → "Extract All" instead of double-clicking
2. **Mac:** File may be corrupted
   - Try downloading the ZIP file again
   - Double-click should auto-extract
3. **Alternative:** Extract manually
   - Use any archive program (WinRAR, 7-Zip, Archive Utility)
   - Drag the CSV file to your `Datasets` folder

Problem: Notebook crashes or shows column errors

Solutions:

1. **Verify CSV files are correct**
   - Open each CSV in Excel or text editor
   - Check that column names match those listed in Step 2
   - Look for unusual characters or extra blank rows
2. **Files might be corrupted**
   - Re-download them
   - Check file sizes (should not be 0 KB)
3. **Wrong file format**
   - Ensure files are actual CSV files
   - If Excel files (.xlsx), convert to CSV first

Problem: "ModuleNotFoundError" or "ImportError"

Solution: Install required Python packages

```bash
pip install -r requirements.txt
```

This installs all necessary libraries (pandas, xgboost, scikit-learn, etc.)

Problem: Analysis runs but results look incorrect

Check:

- Are all 3 data files the correct files? (not test files or wrong datasets)

- Did you extract the population ZIP file?
- Are the files from the correct time periods? (2000-2022)
- Try re-downloading the data files

Expected File Sizes

Approximate file sizes for verification:

- `historical_data_2000_2022_filtered.csv`: 80-120 MB
- `copd_aqi_poverty_demographics.csv`: 100-200 KB
- `Population_Census_Numbers_2000_2025.csv`: 50-100 KB
- `Population_Census_Numbers_2000_2025.zip`: 10-30 KB (compressed)

If your files are dramatically different sizes or 0 KB, they may be corrupted or incomplete.

Privacy & Data Handling

Why aren't data files included in this repository?

1. **Large file size** - The pesticide dataset is approximately 100 MB (exceeds GitHub limits)
2. **Data licensing** - Some datasets have redistribution restrictions
3. **Best practice** - Separating code from data improves version control
4. **Maintainability** - Data files change less frequently than code

Is this data safe to use?

- All data is publicly available from government sources
- Data is aggregated at county level (no individual records)
- No personal information or protected health data
- All sources are official government databases

Note: The `.gitignore` file automatically prevents accidental upload of data files to GitHub.

Expected Runtime

Total time to run notebook: 30-60 seconds on a modern laptop

Time breakdown:

- Data loading: 5-10 seconds
- Feature engineering: 10-15 seconds
- Model training: 10-20 seconds
- Evaluation & visualization: 5-10 seconds

Computer requirements:

- **RAM:** 4 GB minimum, 8 GB recommended
- **Storage:** 500 MB free space
- **Python:** Version 3.8 or higher

Final Checklist Before Running

Complete this checklist before running the notebook:

Data Setup:

□ `Datasets` folder created in project directory
□ All 3 CSV files downloaded
□ Population ZIP file extracted
□ All files in correct location
□ File names match exactly

Software Setup:

□ Python installed (version 3.8+)
□ Jupyter Notebook installed
□ Required packages installed (`pip install -r requirements.txt`)

Notebook Setup:

□ Notebook file paths updated (Step 4)
□ Notebook saved after changes
□ Notebook opens without errors

Getting Help

If you encounter issues:

1. **Read error messages carefully** - They usually indicate what's wrong
2. **Verify spelling and capitalization** - File systems are case-sensitive

3. **Review the troubleshooting section** - Most common issues are addressed

4. **Open a GitHub issue** - Include your error message and system information

Helpful information for bug reports:

- Operating system (Windows 10, macOS 14, Ubuntu 22.04, etc.)

- Python version (run `python --version` in terminal)

- Exact error message (copy and paste)

- What step you're stuck on

- What you've already tried

Learning Resources

New to data science? Here are helpful resources:

- **CSV files:** https://en.wikipedia.org/wiki/Comma-separated_values

- **Jupyter Notebooks:** https://jupyter.org/try

- **Python basics:** https://www.python.org/about/gettingstarted/

- **File paths:** https://en.wikipedia.org/wiki/Path_(computing)

- **XGBoost documentation:** https://xgboost.readthedocs.io/

Ready to analyze? Once your data is configured, open the notebook and run all cells to explore the relationship between pesticide exposure and respiratory health outcomes.