📊 Data Setup Guide

What this is: Instructions to get the required data files so the analysis notebook can run successfully.

🎯 Quick Overview

This analysis needs **3 data files** (CSV format) to work properly:

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 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: Get the Data Files

File 1: `historical_data_2000_2022_filtered.csv`

What it contains: 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: ~1.1 million records (large file - may take time to download)

File 2: **`copd_aqi_poverty_demographics.csv`**

What it contains: 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: ~1,300 county-year observations

File 3: **`Population_Census_Numbers_2000_2025.csv`** ⚠️ **ZIP FILE**

What it contains: 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

⚠️ **IMPORTANT:** This file comes as `Population_Census_Numbers_2000_2025.zip` and **must be extracted** before use.

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. The CSV file should now be 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 (you don't need it anymore)

Linux/Command Line:

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

🗂️ Step 3: Verify Your Setup

Your folder structure should look EXACTLY like this:

```
your-project-folder/
│
├── XGBoost_pesticide_copd_analysis_Completed.ipynb  ⟸ The analysis notebook
├── requirements.txt
├── DATA_SETUP.md (this file)
├── .gitignore
│
└── Datasets/                                  ⟸ Your data folder
    ├── historical_data_2000_2022_filtered.csv      ⟸ Pesticide data
    ├── copd_aqi_poverty_demographics.csv           ⟸ Health data
    └── Population_Census_Numbers_2000_2025.csv     ⟸ Population data (EXTRACTED)
```

✓ 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** (including capitalization, underscores, and .csv extension)
- [ ] All files open correctly (try double-clicking to view in Excel/text editor)

🔧 Step 4: Update the Notebook File Paths

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

What to Do:

1. **Open** the notebook: `XGBoost_pesticide_copd_analysis_Completed.ipynb`
2. **Find Cell 2** (it's near the top, look for code that loads data)
3. **Look for these lines** (they'll have a long file path):

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

4. **Replace with these simple 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 press Ctrl+S / Cmd+S)

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

📊 Data Sources & Attribution

Where This Data Comes From:

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 features)
- **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 & solutions:

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

🔴 Problem: ZIP file won't extract

Solutions:

1. **Windows:** Make sure you have extraction software
   - Windows 10/11 has built-in ZIP support
   - Try right-click → "Extract All" instead of just opening
2. **Mac:** File may be corrupted
   - Try downloading the ZIP file again
   - Double-click should auto-extract
3. **Alternative:** Extract manually
   - Open the ZIP in any program (WinRAR, 7-Zip, Archive Utility)
   - Drag the CSV file to your Datasets folder

🔴 Problem: Notebook crashes or shows errors about columns

Solutions:

1. **Check your CSV files are correct**
   - Open each CSV in Excel or text editor
   - Verify column names match those listed in Step 2
   - Look for any weird characters or extra blank rows
2. **Files might be corrupted**
   - Try downloading them again
   - Check file sizes (should not be 0 KB)
3. **Wrong file format**
   - Make sure files are actual CSV files
   - If they're Excel files (.xlsx), convert to CSV first

🔴 Problem: "ModuleNotFoundError" or "ImportError"

Solution: You need to install Python packages first

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

This installs all necessary software libraries (pandas, xgboost, etc.)

🔴 **Problem: Analysis runs but results look wrong**

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)

💾 File Size Reference

Expected file sizes (approximate):

- `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.

🔒 Privacy & Security

Why aren't data files included in the GitHub repository?

1. **Large file size** - The pesticide dataset is ~100 MB (too large for GitHub)
2. **Data licensing** - Some datasets have redistribution restrictions
3. **Best practice** - Separating code from data makes projects cleaner
4. **Version control** - 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 data files from being uploaded to GitHub if you fork/clone this project.

⏱️ 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

Go through 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
- [ ] Can open notebook without errors

Ready to go? Open the notebook and run all cells! 🚀

💡 Still Need Help?

If you're completely stuck:

1. **Read error messages carefully** - They usually tell you what's wrong
2. **Check spelling and capitalization** - Computers are picky about exact matches
3. **Try the troubleshooting section** - Most issues have solutions above
4. **Ask for help** - Include your error message and what you've tried

Helpful information to include when asking for help:

- Your operating system (Windows 10, macOS 14, Ubuntu 22.04, etc.)
- Python version (run `python --version` in terminal)
- Exact error message (copy and paste the full error)
- What step you're stuck on
- What you've already tried

🎓 Learning Resources

New to this? Here are some helpful resources:

- **What is a CSV file?** https://en.wikipedia.org/wiki/Comma-separated_values
- **Jupyter Notebooks 101:** https://jupyter.org/try
- **Python basics:** https://www.python.org/about/gettingstarted/
- **Understanding file paths:** https://en.wikipedia.org/wiki/Path_(computing)

All set? Head back to the notebook and start analyzing! 🔬 📈