**Multiple Notebooks Approach (Recommended):**

**Notebook 1: 01\_data\_loading\_exploration.ipynb**

* Load all 4 essential files
* Initial data inspection and quality checks
* Basic file structure exploration
* Data type optimization
* Save processed files

**Notebook 2: 02\_data\_combination\_preprocessing.ipynb**

* Combine beneficiary files (2008, 2009, 2010)
* Data cleaning and preprocessing
* Handle missing values
* Create master datasets
* Data validation

**Notebook 3: 03\_target\_variable\_creation.ipynb**

* Define 30-day readmissions
* Create readmission target variable
* Validate business logic
* Calculate baseline readmission rates

**Notebook 4: 04\_feature\_engineering.ipynb**

* Create demographic features
* Engineer chronic condition features
* Build prior admission history features
* Feature validation and exploration

**Notebook 5: 05\_model\_development.ipynb**

* Model training and evaluation
* Performance analysis
* Feature importance
* Business impact calculation

**Why Multiple Notebooks:**

**Advantages:**

* ✅ **Better organization** - each notebook has a clear purpose
* ✅ **Faster iteration** - don't need to re-run data loading when working on models
* ✅ **Easier debugging** - isolate issues to specific steps
* ✅ **Cleaner version control** - smaller, focused commits
* ✅ **Collaborative friendly** - team members can work on different notebooks
* ✅ **Professional presentation** - each notebook tells a story

**Workflow Benefits:**

* Run data loading once, use results multiple times
* Experiment with features without reloading data
* Quick model iterations without preprocessing delays

**Alternative: Single Notebook Approach**

If you prefer to start simple, you could begin with:

**hospital\_readmission\_analysis.ipynb**

* All steps in one notebook with clear section headers
* Easier to see the full workflow at once
* Good for initial exploration

**Then split later** when the notebook gets too long (>100 cells)

**My Recommendation:**

**Start with the multiple notebook approach** because:

1. Your project is substantial enough to benefit from organization
2. You'll save time in later phases when iterating on models
3. It's easier to present and document your work
4. More professional and maintainable

**Phase A specifically would use:**

* 01\_data\_loading\_exploration.ipynb - for loading all 4 files
* 02\_data\_combination\_preprocessing.ipynb - for combining beneficiary files

**Notebook Template Structure:**

Each notebook should have:

markdown

# 01 - Data Loading and Exploration

## Objective

Load and explore the 4 essential DE-SynPUF files

## Contents

1. Environment Setup

2. File Loading

3. Data Structure Exploration

4. Data Quality Assessment

5. Save Processed Files

## Key Outputs

- Loaded dataframes

- Data quality report

- Processed files saved to data/processed/