**Hospital Cost Efficiency Analysis using CMS Cost Report Data**

**✅ Project Goal**

To evaluate hospital cost-efficiency by predicting revenue using expense categories, and benchmarking hospital performance using efficiency scores and frontier analysis.

**ᵁ Step 1: Data Selection and Preprocessing**

1. **Obtain CMS Hospital Provider Cost Report Data**
   * Download the most recent year available from CMS website.
2. **Initial Filtering**
   * Keep hospitals with complete data for key variables (expenses, revenue).
   * Optionally filter by type (e.g., General Acute Care) and size (e.g., 100–300 beds).
3. **Remove Outliers (Global)**
   * Use IQR or z-score method on revenue and key expenses.
   * Optionally winsorize top/bottom 1% to keep dataset size.
4. **Standardize Features**
   * Apply scaling (e.g., StandardScaler) to all numeric variables.

**ᵁ Step 2: Clustering (Optional, for Peer Grouping)**

1. **Select Features for Clustering**
   * Beds, Total Expenses, Staff Count, Medicare %, Teaching Status.
2. **Apply Clustering Algorithm**
   * Use KMeans or Hierarchical Clustering.
   * Choose number of clusters using elbow method or silhouette score.
3. **Label Hospitals by Cluster**
   * Use cluster labels for later subgroup analysis or benchmarking.

**ᵁ Step 3: Multiple Linear Regression Model**

1. **Define Inputs (X)**
   * Select expense categories: Labor Costs, Admin, Capital, Supplies, Depreciation, etc.
2. **Define Output (y)**
   * Revenue (e.g., total patient revenue).
3. **Split Data**
   * Use train-test split (e.g., 80/20).
4. **Fit Baseline MLR Model**
   * Use statsmodels or scikit-learn.
   * Check Adjusted R², RMSE, residual plots.
5. **Check Multicollinearity**
   * Compute VIF for each variable. Drop or combine variables with VIF > 10.
6. **Feature Selection**
   * Optionally apply Lasso/Ridge to find most predictive cost drivers.

**ᵁ Step 4: Model Comparison**

1. **Compare Against Alternatives**
   * Ridge, Lasso, Random Forest Regressor.
2. **Use Cross-Validation**
   * Evaluate models using RMSE, R², and interpretability.
3. **Select Best Model**
   * Based on performance and explainability.

**ᵁ Step 5: Benchmarking and Efficiency Analysis**

1. **Choose Benchmark Method**
   * Use Data Envelopment Analysis (DEA) or a custom efficiency ratio.
2. **Compute Efficiency Scores**
   * Efficiency = Revenue / Expected Revenue (from frontier or top-performing hospitals).
3. **Rank Hospitals**
   * Identify top-performing hospitals as benchmarks.
4. **Visualize**
   * Plot efficiency curve (Total Expense vs Revenue).
   * Show histogram of efficiency scores.

**✨ Final Deliverables**

* Clean dataset of selected hospitals
* MLR model with feature importance
* Efficiency scores and benchmark list
* Visualizations: scatter plot, efficiency curve, model comparison

*Optional extensions:*

* Apply to different states or hospital categories
* Add patient outcome metrics for quality-adjusted benchmarking