Project: Predicting Electrical Bill

✅ Data Loading & Cleaning

- Loaded Electricity Bill Dataset dataset.
- Deleted rows where motor pump column = 0 since these do not contribute to power consumption and would not affect the result, ensuring data relevance.
- Ensured consistent data types for all numerical columns for accurate scaling and modeling.
- Applied `pd.get_dummies` on `city` and `company` columns to convert categorical variables into numerical form, enabling the regression model to utilize these features effectively.

✅ Data Visualization

- Explored trends between monthly hours, tariff rate, motor HP, and electrical bill.
- Plotted scatter plots, heatmaps, and pairplots to check correlations.

✅ Feature Engineering

- Created two versions of the model:
  - **Full model:** Includes *monthly hours* and *tariff rate*.
  - **Reduced model:** Excludes *monthly hours* and *tariff rate* to test dependency on operational features.

✅ Preprocessing

- Used train-test split (80-20) for validation.
- Applied StandardScaler to normalize feature ranges.

✅ Model Building

- Implemented Linear Regression for prediction.
- Evaluated using:
  - **RMSE (Root Mean Squared Error)** for error magnitude.
  - **R^2 Score** for variance explanation.

✅ Results Interpretation

- Compared R^2 and RMSE between *full* and *reduced* models to analyze the impact of excluded features.
- Identified the contribution of *monthly hours* and *tariff rate* on bill predictability.

✅ Streamlit App Deployment

- Built an **interactive Streamlit app**:
  - Option to select:
    - *Full model* (with hours, tariff) or
    - *Reduced model* (without them).
  - Users can **input values** (motor HP, monthly hours, tariff, etc.) to **predict the "Electrical Bill" live**.
  - Displays **predicted bill amount, RMSE, and R^2 of the selected model** for transparency.

✅ Model Evaluation Results:

- **Root Mean Squared Error (RMSE): 0 (My** model's predictions **exactly match the actual electricity bills** in your dataset with **zero error)**
- **R^2 Score: 1 (**Your model **perfectly explains 100% of the variance** in the electricity bill using `Usage_Tariff_Interaction`.**)**
  - `Usage_Tariff_Interaction = Monthly_hours * Tariff_rate`

✅ Feature Importance Evaluation :

By comparing the model performance with and without `monthly hours` and `tariff rate`, I found:

- Models including `monthly hours` and `tariff rate` had significantly higher R^2 scores and lower RMSE, indicating better predictive power.
- Removing these features reduced the model's ability to explain variance in the "Electrical Bill", confirming that:
  - Monthly operational hours (usage patterns) and
  - Tariff rate (cost per unit electricity)

  are critical factors influencing the electrical bill for motor pumps.

Extras:

I have done this using Gradient Boosting, Decision Tree, Elastic Net, and Ridge Regression models on this dataset. The project is available in the miscellaneous folder of my GitHub.

Skills & Concepts Used:

- **Python (pandas, sklearn, matplotlib, seaborn)**
- **Machine Learning (Linear Regression, evaluation metrics)**
- **Feature Engineering & Impact Analysis**
- **Data Normalization (StandardScaler)**
- **Streamlit for interactive ML deployment**