**README for Synthetic Data Generation and EDA Project**

**Project Overview**

This project aims to generate synthetic data for hourly pay rates of nurses in major U.S. metropolitan areas, perform exploratory data analysis (EDA), and implement machine learning/deep learning models to predict these rates. The project also includes a Streamlit application for user interaction.

**Workflow Overview**

**Step 1: Synthetic Data Generation**

* **Goal**: Create a comprehensive dataset of nursing job contracts.
* **Process**:
  1. **Define Locations**: Select major U.S. cities (e.g., Dallas, New York, San Francisco).
  2. **Specify Attributes**:
     + Job Title (e.g., Registered Nurse, Physio Therapist)
     + Location (City & State)
     + Hospital Name (City prefix + Suffix: Corporate, NonProfit, etc.)
     + Contract Start Date
     + Contract End Date
     + Hourly Pay Rate (with seasonal variations)
  3. **Generate Data**: Create 250,000 rows of data for the years 2023 and 2024, ensuring contract durations do not exceed 13 weeks.

**Step 2: Exploratory Data Analysis (EDA)**

* **Goal**: Analyze the generated dataset to extract meaningful insights.
* **Process**:
  1. **Hourly Pay Rate Analysis**:
     + Compare pay rates across different metropolitan areas.
     + Identify seasonal trends (flu season, holidays).
  2. **Desirability Factors**:
     + Correlate pay rates with city desirability (cost of living, schools, crime rates).
  3. **Specialization Comparison**:
     + Analyze pay rates for specialized nursing roles (oncology, cardiology, surgery) versus other job titles.

**Step 3: Machine Learning/Deep Learning Models**

* **Goal**: Predict hourly pay rates using ML/DL techniques.
* **Process**:
  1. **Model Selection**: Choose two appropriate models for prediction.
  2. **High Cardinality Handling**: Address the challenge of high cardinality in hospital names.
  3. **Performance Metrics**: Define metrics for evaluating model accuracy (e.g., RMSE, MAE).

**Step 4: Streamlit Application Development**

* **Goal**: Create an interactive application for user input and predictions.
* **Process**:
  1. **User Input**: Allow users to enter Job Title, Location, Hospital, Contract Start Date, and Contract End Date.
  2. **Prediction Output**: Display the predicted Hourly Rate based on user inputs.

**Step 5: Bonus - Time Series Forecasting**

* **Goal**: Implement a time series forecasting model.
* **Process**:
  1. **Model Selection**: Choose between Prophet, NeuralProphet, or Stacked LSTM for forecasting.

**Getting Started**

1. Clone the repository.
2. Install required packages (see **requirements.txt**).
3. Run the Jupyter notebook or Streamlit app to explore the data and models.