Take home coding exercise (Data Science)

This assignment has three parts:

1. **Synthetic Data Generation (using ChatGPT)**
2. **Exploratory Data Analysis (EDA) & Feature Engineering**
3. **Advanced ML/DL Modeling & Deployment**

1. Synthetic Data Generation

Generate synthetic data to capture hourly pay rates for nurses in the US across the major metros. The key attributes are location (major metros in US), job title (nurse titles vary based on their specialization like ICU Nurses, Oncology etc..), hospital name, hospital type (corporate, not-for-profit, community, veterans etc..), contract start date, contract end date, hourly pay rate

1. The hourly pay rate should show seasonal uptick during flu & Christmas holiday season
2. Pay rates should be higher for large metros where the cost of living is high.
3. Also, desirability of location (good schools, safe neighborhoods, etc..) should affect the pay rates as well.
4. Generate the data for the years 2023 & 2024
5. The contract duration cannot exceed 13 weeks.
6. Generate a total of 250,000 rows

Considerations:

1. Show creativity in your interactions with ChatGPT
2. Use public data sources for cost of living, desirability of a location etc…
3. Ensure your dataset meets the criteria mentioned above
4. Submit your ChatGPT interaction by sharing the ChatGPT URL (no need to copy the content to a word doc)

Part 2: EDA & Feature Engineering

Objective: Analyze the dataset to extract insights and prepare features for modeling.

Evaluation Criteria:

- **Depth of EDA:** Insightfulness of analyses, quality of visualizations, and ability to uncover meaningful patterns.
- **Analysis on Every Feature:** Comprehensive examination of each feature.
- **Feature Engineering:** Detailed explanation of engineering steps taken for handling each feature.
- **Feature Extraction:** Adding additional features based on your understanding.
- **Feature Selection:** Engineering techniques for selecting features for final modeling.

Key consideration:

- How to handle high cardinality features and explain your approach.

Part 3: Advanced ML/DL Modeling & Deployment

Objective: Develop and evaluate sophisticated machine learning and deep learning models to predict hourly pay rates, and deploy these models through an interactive application.

Tasks:

- **Model Implementation:**
  - **Model Selection:** Implement at least two different models, ensuring diversity.
  - **Justification:** Provide rationale for selecting each model based on data characteristics and problem requirements.

    - **Evaluation Metrics:** Choose appropriate metrics and describe their interpretation.
- **Streamlit Application:**
    - **User Inputs:** Allow users to input Job Title, Location (City & State), Hospital Name, Contract Start Date, Contract End Date, Shift Type.
    - **Predictions:** Display predicted Hourly Pay Rate along with confidence intervals or prediction intervals.
    - In your video recording submission, use the Streamlit app to demo the application.

Submission

1. For the first part, submit your link to the ChatGPT session.
    a. This part will contribute about 1/3rd to your grade.
2. For the second & third parts, submit your Google Colab or Python Jupyter notebook along with the data generated in the first part. The following parts should be covered **mandatorily**:
    a. The notebook should contain your EDA with the implementation of the models along with the performance metrics.
    b. Show the performance of your model for both peak & non-peak seasons
    c. Show how you are handling high cardinality fields like hospitals & roles
    d. Show how your model handles a nurse requirement a) from a new hospital that is located in a location currently part of the training b) from a new hospital and also a new location

Notes

1. It is important that you do the work yourself. Submitting other people's work will lead to immediate disqualification
2. During the demo, you will be asked to make changes to demonstrate your understanding of the problem. Hence it is even more important to follow above guideline