** Hospital Dataset Analysis**

Overview

Data preprocessing and analysis are crucial steps in the machine learning workflow, significantly influencing model performance and insights gained from the data. Feature engineering enriches the dataset by creating new, informative features, while data cleaning ensures the model is trained on high-quality information. By transforming raw data into a well-prepared dataset, we lay the foundation for developing accurate predictive models that can handle real-world complexities.

In this assignment, you will perform data cleaning, exploratory data analysis (EDA), and visualization on the hospital dataset. Advanced analysis and optional predictive modeling tasks are included for students seeking additional challenges.

Task 1: Data Cleaning and Preprocessing

Proper data cleaning is essential to remove inaccuracies, ensure consistency, and prepare the data for analysis. You are required to:

1. **Datetime Conversion**:
    - Convert the 'Time of Admission' and 'Time of Discharge' columns to datetime format.
2. **Column Name Standardization**:
    - Clean up column names to remove extra spaces and special characters.
3. **Handle Missing or Inconsistent Data**:
    - Identify and address missing, null, or inconsistent values (e.g., negative expenses, admission after discharge).
4. **Data Validation** :
    - Perform basic validation on key columns to ensure no anomalies such as unrealistic expense values or time intervals.

Task 2: Exploratory Data Analysis (EDA)

Gain insights into the dataset by exploring relationships between different features. You should:

1. **Average Stay Duration**:
    - Calculate the average stay duration by finding the difference between 'Time of Admission' and 'Time of Discharge'.
2. **Departmental Analysis**:
    - Group the data by 'Department' and calculate the average 'Medical Expenses' for each department.
3. **Common Discharge Diagnosis**:
    - Identify the most frequent 'Discharge Diagnosis' across all patients.

    Bonus:
    - Visualize distributions of key variables (e.g., using box plots or histograms) to show the spread and outliers in features such as 'Medical Expenses'.
    - Investigate potential seasonal trends in admissions or expenses by grouping the data by month or quarter.

Task 3: Data Aggregation and Grouping

Summarizing and grouping data can provide deeper insights into various segments of the dataset. Perform the following:

1. **Place of Birth Summary**:
    - Summarize total 'Expenses and Outpatient' and 'Medical Expenses' for each 'Place of Birth'.
2. **Departmental Expense Analysis**:
    - Determine the average 'Surgery Expenses' and 'Bed Fees' for each 'Department'.

    Bonus:
    - To analyze medical expenses across these categories, perform multi-level grouping (e.g., by both 'Department' and 'Discharge Diagnosis').
    - Aggregate expenses and patient numbers over time (e.g., monthly trends) to uncover hospital resource usage patterns.

Task 4: Data Visualization

Effective visualizations help in interpreting the data and communicating results. Create the following plots:

1. **Stay Duration Histogram**:
    - Plot a histogram of the 'Days' patients stayed in the hospital.
2. **Departmental Expenses Bar Chart**:
    - Plot a bar chart showing the total 'Medical Expenses' for each department.
3. **Nursing Expenses Over Time**:
    - Generate a line plot showing the trend of 'Nursing Expenses' over time, based on the 'Time of Admission'.

   Bonus:
    - Create more sophisticated plots like heatmaps for visualizing correlations between numerical variables, or stacked bar charts to show expenses broken down by both 'Department' and 'Place of Birth'.
    - Explore using interactive plotting libraries like Plotly to enhance your visualizations.

Task 5: Advanced Analysis (Optional)

For those seeking an additional challenge, try these advanced tasks:

1. **Correlation Analysis**:
    - Perform a correlation analysis between 'Medical Expenses' and other numerical expense-related columns to see if any strong relationships exist.
2. **Pivot Table Analysis**:
    - Use pivot tables to analyze the relationship between 'Department', 'Place of Birth', and 'Medical Expenses'.
3. **Predictive Modeling**:
    - Build a simple regression model to predict 'Medical Expenses' based on other numerical and categorical variables. Experiment with models like linear regression or decision trees.
4. **Clustering**:
    - Perform clustering (e.g., k-means) to group patients based on features like 'Medical Expenses', 'Days Stayed', and 'Discharge Diagnosis' to uncover potential patterns in patient types.

Submission Guidelines

Your submission should include:

- **Preprocessing Code**: All steps taken to clean and preprocess the data.
- **Analysis**: A Jupyter notebook with code and comments for your EDA, aggregations, and visualizations.
- **Visualizations**: Plots with appropriate titles, labels, and legends that clearly convey the data insights.
- Zip the file with your name and mat number