**Healthcare Data Analysis: Exploring Billing Patterns and Identifying Anomalies**

**A person wearing a white coat and mask

Description automatically generated**

**1. Introduction**

This project explores a healthcare dataset to uncover patterns in patient demographics, billing behavior, and potential data anomalies. By applying descriptive statistics and exploratory data analysis (EDA), the goal is to generate insights that support cost optimization, data quality improvements, and informed decision-making in healthcare systems.

**2. Data Source**

The dataset is provided as a CSV file and includes:

* Patient demographics (age, gender, etc.)
* Medical details (conditions, admission type)
* Billing amounts
* Insurance providers
* Hospital and doctor information
* Admission and discharge dates

The data was uploaded using **Google Colab's files. Upload ()** for flexible and dynamic handling of file names.

**3. Data Loading & Preprocessing**

* **Library Imports**: Used pandas, numpy, matplotlib, and seaborn for data handling and visualization.
* **Loading**: Data loaded with pd.read\_csv() using io.BytesIO() to dynamically manage file uploads.
* **Initial Exploration**: Checked data shape, column names, data types, and summary statistics.
* **Null Handling**: Identified and handled missing values.
* **Data Type Conversion**: Converted date columns (Admission Date, Discharge Date) to datetime format for time-based operations.

**4. Exploratory Data Analysis (EDA)**

**➤ Descriptive Statistics**

* Summarized numerical features: Age, Billing Amount, Room Number.
* Explored categorical features: Medical Condition, Admission Type, Insurance Provider.

**➤ Visualizations**

* **Histograms**: Distribution of age, billing amounts.
* **Bar Charts**: Frequencies of admission types, conditions, and providers.
* **Box Plots**: Variance in billing across medical conditions and insurance types.
* **Scatter Plots**: Examined correlation between age and billing.
* **Heatmap**: Correlation matrix for numeric features.

**5. In-depth Analysis**

**Correlation Matrix**

* Used a heatmap to identify relationships among features like age, room number, and billing amount.

**Anomaly Detection**

* Flagged **negative billing amounts** as potential data entry errors.
* Explored patterns in negative billing across providers and conditions.

**Group Analysis**

* Grouped by Medical Condition, Admission Type to evaluate average and median billing.
* Box plots used to visualize group-level billing variability.

**High Cardinality Features**

* Columns like Name, Doctor, and Hospital showed high uniqueness.
* Suggested techniques: label encoding, frequency encoding, or clustering.

**6. Key Findings**

* **Negative Billing**: Indicates potential billing errors; needs validation.
* **High Variability in Billing**: Strong variation across conditions and admission types.
* **Data Quality Issues**: Missing and invalid entries require cleaning.
* **Feature Importance**: Categorical features like Medical Condition and Insurance Provider likely influence billing heavily.

**7. Recommendations & Next Steps**

* **Clean Negative Values**: Investigate or remove rows with negative billing.
* **Feature Engineering**:
  + Encode high-cardinality features.
  + Calculate hospital stay durations.
* **Modeling Potential**:
  + Predict billing amounts using regression.
  + Classify high/low-cost cases using classification models.
* **Time Series or Trend Analysis**: Explore admission volume trends over time.

**8. Conclusion**

* This analysis provided a detailed overview of healthcare billing behaviors and exposed valuable insights regarding costs, demographics, and potential data quality issues. Future work can involve predictive analytics, cost-saving strategy development, and deeper anomaly detection for healthcare systems.