This notebook performs exploratory data analysis (EDA) on a healthcare fraud detection dataset. Here's a breakdown of the steps:

1. **Load Cleaned Data:** The notebook starts by loading cleaned data for inpatient claims, outpatient claims, beneficiary information, and training labels from parquet files stored in Google Drive. It then prints the shape of each DataFrame to show the number of rows and columns.
2. **Class Distribution & Key Summary Stats:**
   * **Class Distribution (Fraud vs. Legitimate Providers):** This section analyzes the distribution of the target variable (PotentialFraud) to see how many providers are labeled as fraudulent versus legitimate. It prints the counts and percentages and visualizes this distribution using a countplot.
   * **Key Metrics by Fraud Label:** It calculates and compares key metrics for fraudulent and legitimate providers, such as the average number of unique beneficiaries, average inpatient claims per beneficiary, average outpatient claims per beneficiary, and average total claims per beneficiary.
3. **Key Reimbursement Stats by Fraud Label:** This section focuses on reimbursement amounts. It calculates the total reimbursement per provider (summing inpatient and outpatient claims) and compares the average total reimbursement for fraudulent and legitimate providers. A boxplot is used to visualize the distribution of total reimbursements by fraud label, using a log scale to handle skewness and outliers.
4. **Risk Stratification:**
   * **Risk Stratification by Total Reimbursements:** Providers are grouped into risk categories (Low, Medium, High, Very High) based on quartiles of their total reimbursements. The fraud rate is calculated and printed for each risk group. A bar plot visualizes the fraud rate across these risk groups.
   * **Risk Stratification Summary:** This section summarizes the key findings from the risk stratification, highlighting the increasing fraud rate in higher reimbursement groups and the significant fraud rate in the "Very High" risk group.
5. **Outlier Context:**
   * This section examines the top providers in the "Very High Risk" group based on total reimbursements. It prints the provider ID, total reimbursements, and fraud label for these top providers.
   * **Outlier Context Summary:** It provides context for these outliers, suggesting they might be legitimate high-volume institutions or potential fraud rings and recommending domain expert review.
6. **Key Takeaways from EDA:** This section summarizes the major insights gained from the EDA, including the rarity and concentration of fraud, likely predictive features, the challenge of class imbalance, the importance of considering outliers, and the clear separation of features between fraudulent and legitimate providers.
7. **Time-Series Analysis:** This section combines inpatient and outpatient claims to analyze trends over time. It converts claim dates to datetime objects, merges claims with provider fraud labels, and aggregates total reimbursement and claim counts by date and fraud label. Line plots are generated to visualize the total reimbursement and claim count trends for legitimate and fraudulent providers.
8. **Correlation Analysis:** This section calculates and visualizes the correlation matrix for numeric features at the provider level using a heatmap.