**Medical Claims Data Analysis and Feature Engineering**

**1. Introduction to the Data and Variables:**

The data revolves around medical claims, shedding light on potential fraudulent activities within the healthcare sector. Diving deeper, the data is segmented into:

* **Inpatient Data:** Comprehensive insights about claims associated with patients admitted to hospitals.

Key Features: Admission and discharge dates (which helps in determining the length of a patient's stay), Diagnosis codes (indicating the medical reasons behind hospitalizations).

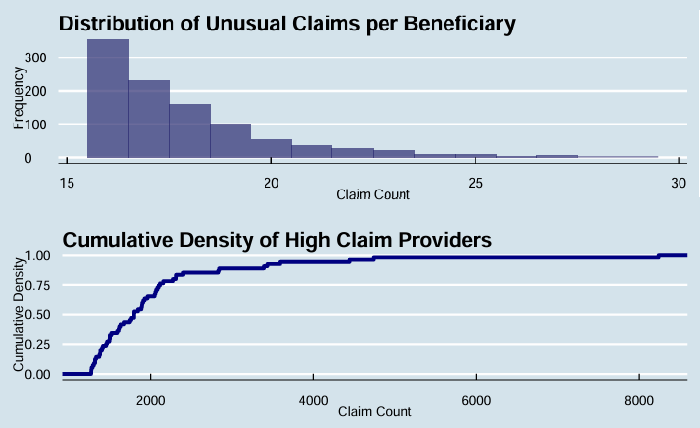
* **Outpatient Data:** This dataset pertains to patients who avail medical care but do not stay admitted in the hospital for long durations.

Key Features: Data indicating the types and frequencies of outpatient services availed.

* **Beneficiary Details Data:** This dataset encapsulates KYC (Know Your Customer) details pertaining to beneficiaries, offering a glimpse into their medical history and affiliations.

Key Features: Health conditions (showcasing the medical history or the current health status of beneficiaries), Regional affiliations (offering a geographical overview which can be pivotal for region-specific analyses).

* **Provider Potential Fraud Data:** This is the cornerstone for our fraud detection analysis. It maps healthcare providers to potential fraudulent activities with a binary distinction – 'Yes' for potential fraud and 'No' for non-fraudulent.



**Distribution of Unusual Claims and Cumulative Density of High Claim Providers**

The dataset comprises 138,556 beneficiaries, 558,211 claims, and 5,410 providers. On average, each beneficiary has around 4.029 claims. Filtering beyond the 99th percentile reveals beneficiaries and providers with notably high claim counts. The Cumulative Density Plot shows a key transition zone, hinting at a rise in unusual claims. Providers with claims exceeding 4,000 require deeper scrutiny due to their deviation from typical patterns.

**2. Merging Datasets and Data Cleaning:**

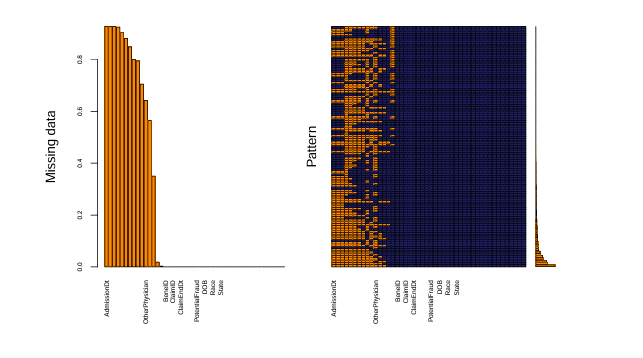
To create a holistic dataset conducive for rigorous analysis, we integrated data from the aforementioned four distinct datasets. This amalgamation resulted in a dataset with 558,211 records spanned across 54 features. Key identifiers like ‘ClaimID’ and ‘BeneID’ were retained for granularity.

Certain columns underwent transformations for consistency:

* ‘Gender’ column was binarized.
* ‘RenalDiseaseIndicator’ was transformed such that 'Y' indicated the presence of renal disease.
* Null values in the ‘DeductibleAmtPaid’ column were imputed with ‘0’.
* In the quest for fraud detection, the ‘PotentialFraud’ column was binarized where 'Yes' indicated potential fraud.

**3. Addressing Missing Data:**

Understanding and navigating through missing data is paramount for any analytical endeavor.



* Columns like ‘AdmissionDt’ and ‘DischargeDt’ had substantial missing data (~92.75%). However, these columns are pivotal for computing the ‘length of stay’ feature, a significant metric for fraud detection.
* For the columns ‘ClmDiagnosisCode\_2-10’ that encapsulate a myriad of medical conditions, missing values varied. It's paramount here to judiciously apply domain knowledge to ascertain their inclusion or exclusion from the analysis.
* Columns with missing values above 30% were extricated from the dataset. Those with missing values below this threshold were retained for imputation.

**Procedure for Missing Data Imputation:**

Considering the non-random nature of the missing data, the Multiple Imputation by Chained Equations (MICE) technique was chosen. This method, leveraging other variables within the dataset, is adept at predicting missing values. Noteworthy parameters for this imputation included:

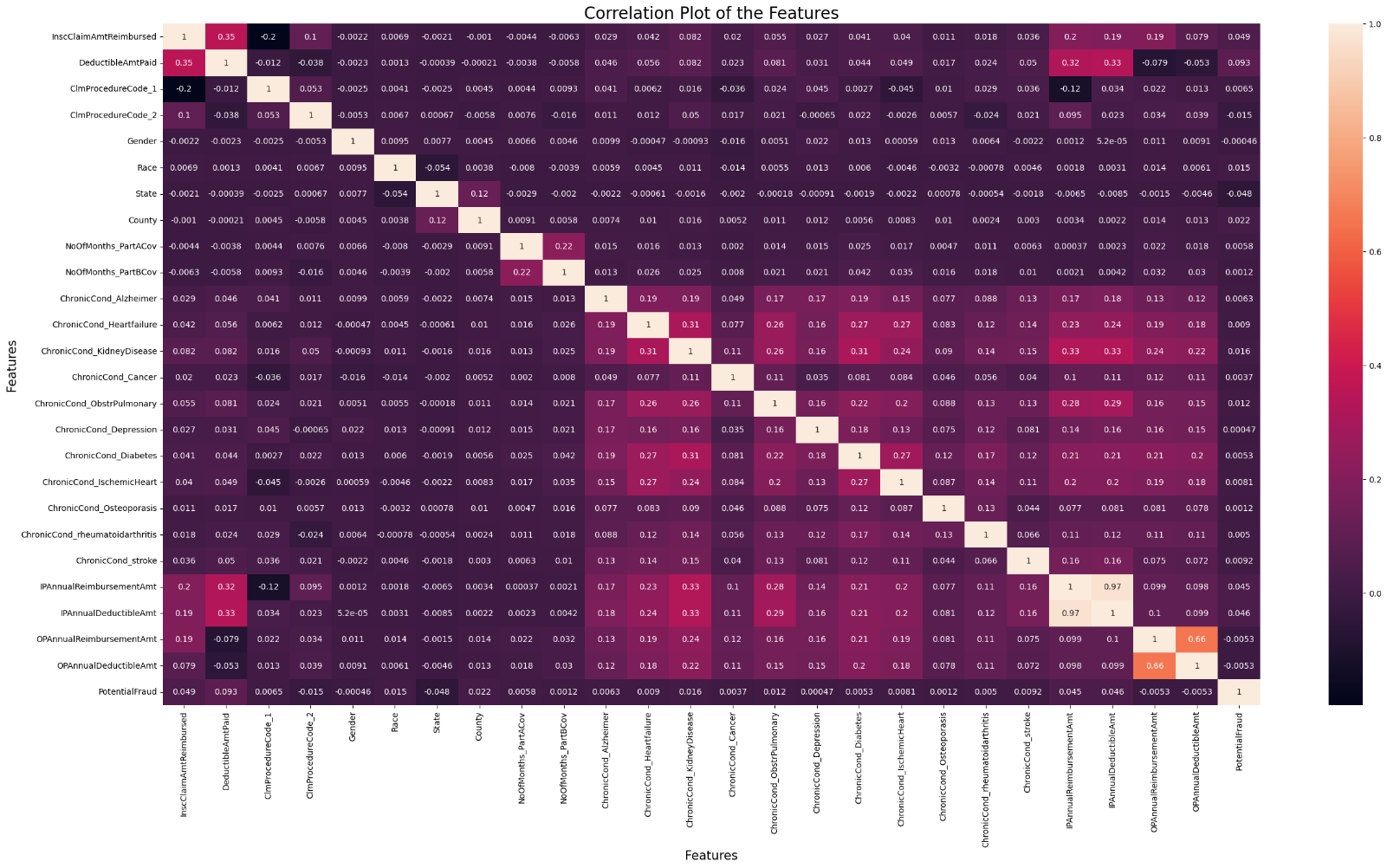
* m=5 for generating five imputed datasets.
* maxit = 50 indicating the number of iterative rounds.
* Adoption of the mean matching method for imputation.

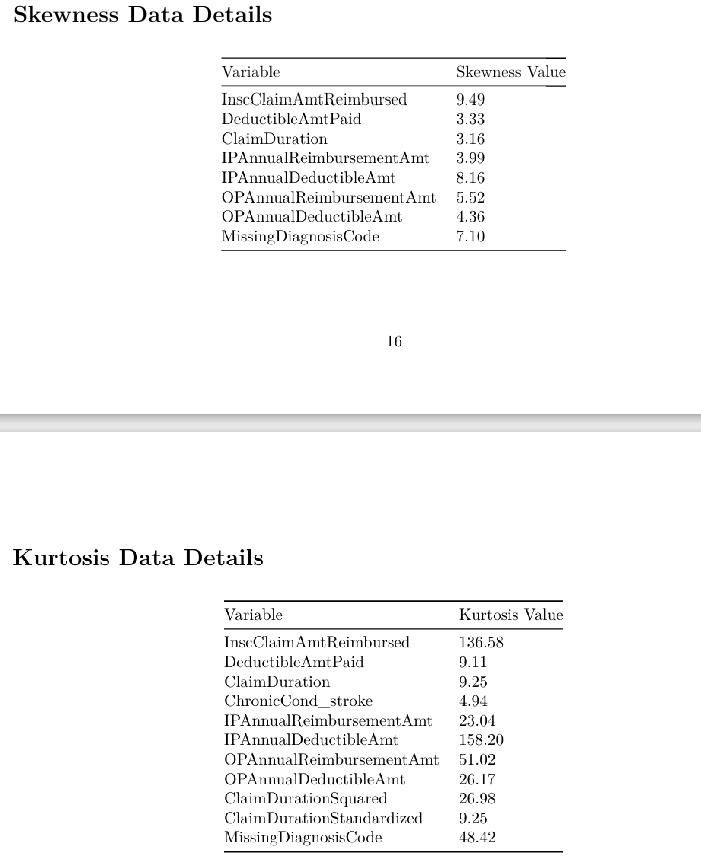
**4. Feature Engineering:**

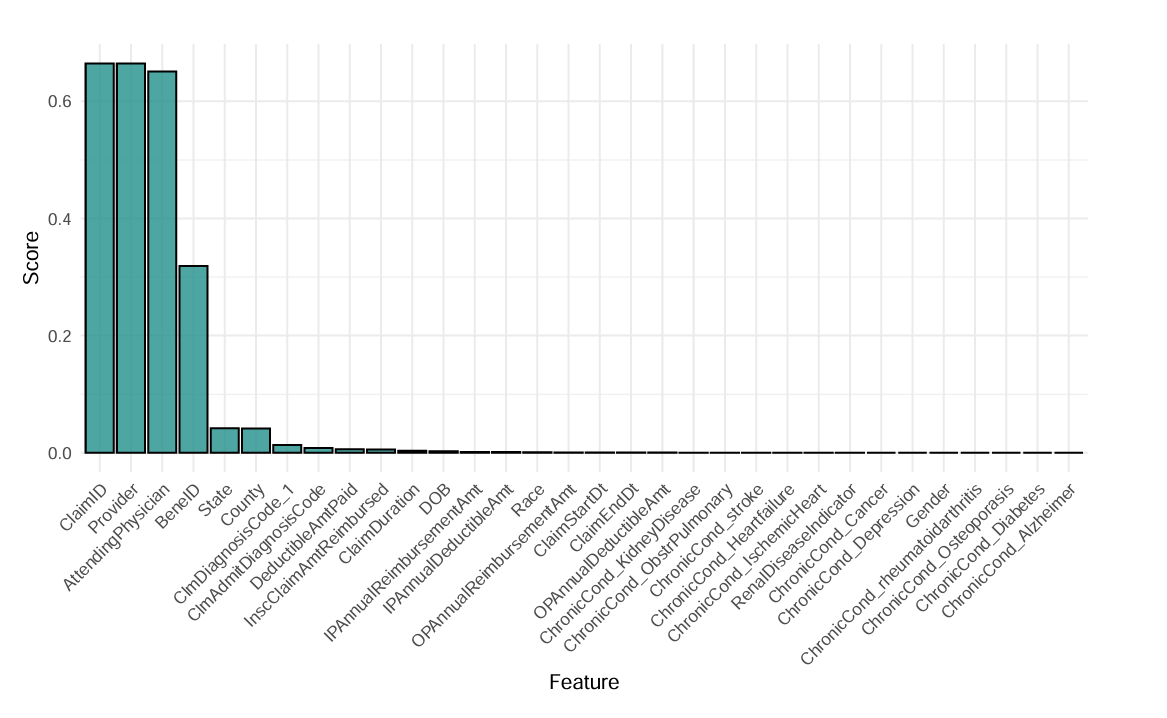
Feature engineering is the backbone of any machine learning project, transforming raw data into insightful features that can significantly improve model performance.

* **Temporal Features & Medical Conditions:** Features like ‘Age at the Time of Claim’ and ‘Claim Processing Time’ were derived. A consolidated feature was also created to encapsulate the severity of chronic conditions, offering an in-depth view of patient health over time.
* **Features Transformation:** Given the skewness and kurtosis in the dataset, transformations, predominantly logarithmic, were used. This not only addressed skewness but also mitigated outlier impact.

**5. Feature Selection:**

The feature selection was data-driven, leveraging the mlr3 package. Based on information gain criteria, features like ‘ClaimID’, ‘Provider’, and ‘Attending Physician’ emerged as paramount predictors. Conversely, features like ‘ChronicCond\_Cancer’, ‘ChronicCond\_Depression’, and ‘Gender’ demonstrated minimal influence on model prediction. Moving forward, our strategy is meticulously crafted, focusing on features with high information gain, ensuring the model's robustness and accuracy.  






|  |  |
| --- | --- |
| Feature Name | Description & Reason for Selection |
| ClmAdmitDiagnosisCode | Binary variable indicating if a claim admission diagnosis code is present. Helpful in identifying the primary reason for medical services. |
| ClmDiagnosisCode\_\* | Binary variables representing different claim diagnosis codes. These codes can provide insights into the types of medical conditions present and may have an association with fraudulent claims. |
| Gender | Binary representation of gender. Gender might play a role in medical service utilization patterns. |
| Race | Categorical variable representing racial categories. Different racial groups might have different medical needs or service utilization patterns. |
| RenalDiseaseIndicator | Indicates the presence or absence of renal disease. Patients with renal diseases might have specific medical service utilization patterns. |
| State, County | Represent geographical locations. Geographical patterns might emerge in fraudulent activities. |
| ChronicCond\_\* | Series of binary variables indicating the presence or absence of various chronic conditions. Chronic conditions might influence the type and frequency of medical services required. |
| PotentialFraud | Binary variable indicating potential fraud. This is the target variable for a fraud detection task. |
| Age | Represents the age of the beneficiaries. Different age groups might utilize medical services differently. |
| WeekendAdmission | Indicates if the admission was during a weekend. Weekends might see different types of claims or emergencies. |
| IsDead | Binary variable indicating if the beneficiary is deceased. Might influence the kind of services rendered. |
| ClaimSettlementDelay\_Cat | Categorical representation of claim settlement delay. Delays in settlements might indicate complications or disputes in the claim. |
| TreatmentDuration\_Cat | Represents the duration of the treatment. Longer durations might be associated with more severe medical conditions. |
| TotalClaimAmount, OPTotalAmount | Amounts related to the claims. High or unusual amounts might be red flags for fraudulent activities. |
| Log\_\* | Log-transformed values of various features to normalize skewed distributions. |
| UniquePhysCount, PhysRoleCount | Numeric values indicating the count of unique physicians and their roles. Multiple roles or frequent changes in physicians might be suspicious. |
| IsSamePhysMultiRole1,2 | Indicates if the same physician plays multiple roles. A single physician playing multiple roles might indicate a lack of specialization or potential fraud. |
| PHY\* | Binary variables indicating presence of specific physician identifications. Specific physicians might be associated with specific types of claims or patterns. |

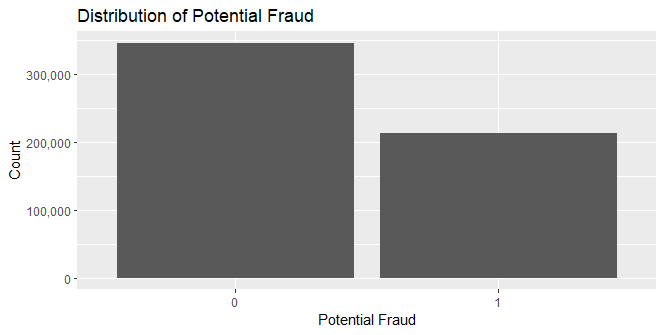
**6.Addressing Feedback:**

**Data Description:** Enhanced explanations of the datasets and their key features.

**Dataset Integration:** Clarified the integration process of the four datasets into one comprehensive dataset.

**Clearer Explanations:** Augmented context and rationale in the data exploration steps for improved understanding.

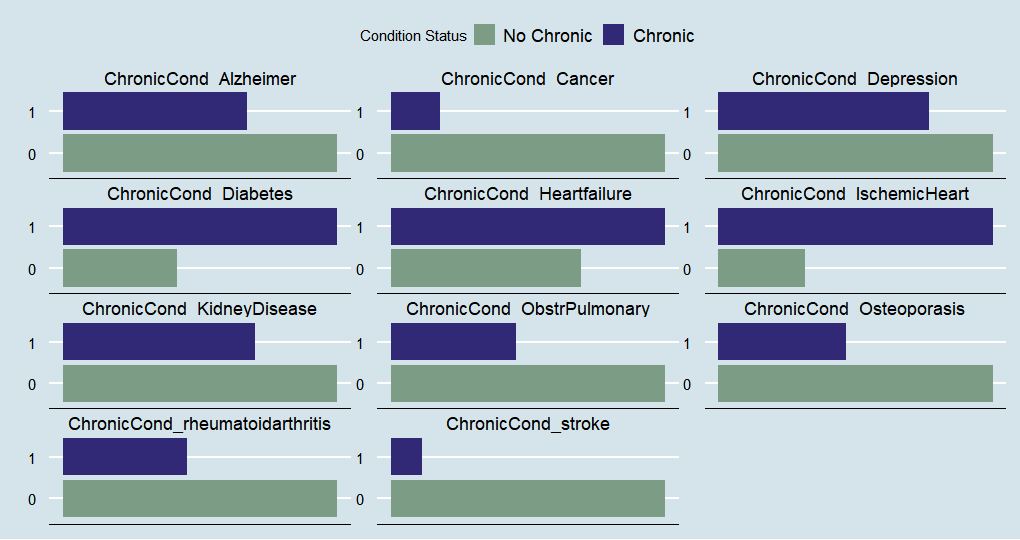
**Plots**

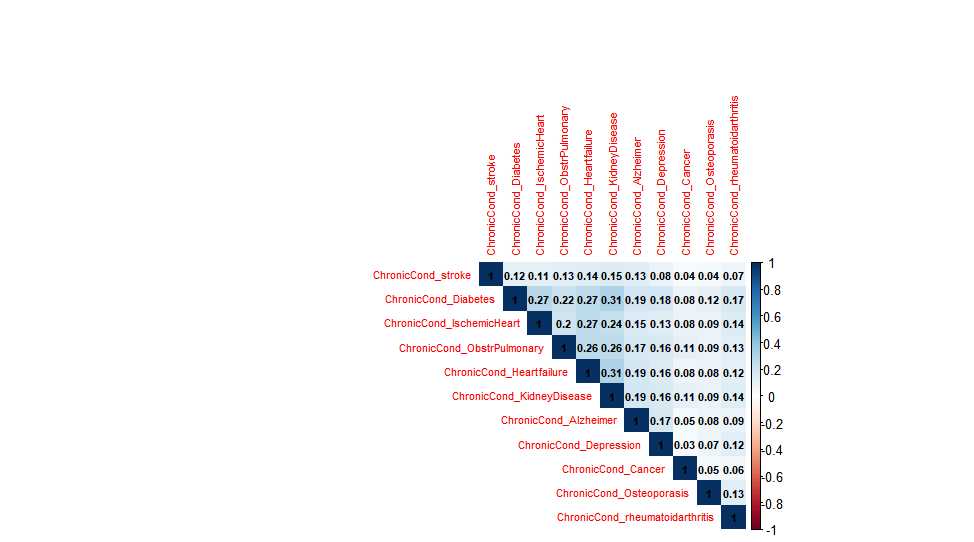


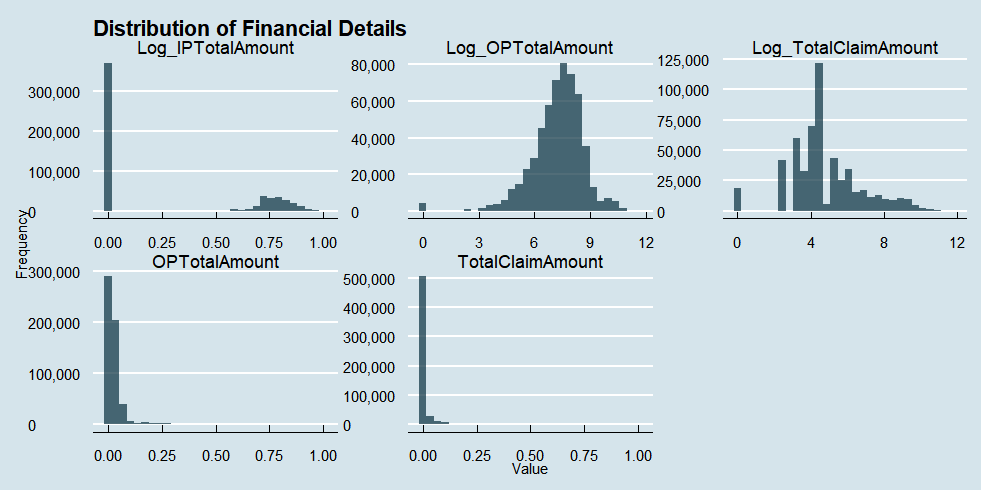
The bar plot showcases the distribution of **potential fraudulent** claims within our dataset. A clear class imbalance is evident, with a vast majority of claims labeled as non-fraudulent (represented by '0'). Such imbalances are pivotal to recognize as they can influence modeling outcomes, potentially skewing predictive accuracy on unseen data. Further investigation could delve into understanding the underlying features that contribute to the minority fraudulent class.



This graph highlights the gender and racial distributions within our dataset. While there's a slight preponderance of one gender over the other, the racial distribution is dominated by one particular category. Delving deeper, one could explore if fraudulent claims are more prevalent within a specific gender or racial group. Such insights can guide targeted fraud detection strategies, ensuring that potential biases are minimized.



The bar chart provides a glimpse into the frequency of various chronic conditions among beneficiaries. 'IschemicHeart' and 'Diabetes' emerge as predominant conditions. However, it would be insightful to investigate if any specific chronic condition is more associated with fraudulent claims. Such findings can be instrumental in healthcare planning, as they can hint at potentially suspicious patterns of claim submissions.

The heatmap portrays the intricate relationships among different chronic conditions. Darker shades, such as the one between 'KidneyDisease' and 'Diabetes', signify stronger positive correlations. These correlations hint that patients diagnosed with one condition might frequently be diagnosed with another. From a fraud detection perspective, unusually high correlations might warrant further investigation, as they could be indicative of repetitive or template-based claim submissions.

The histograms delineate the financial distributions associated with both inpatient and outpatient services. For instance, 'OPAnnualDeductibleAmt' showcases a right-skewed pattern, revealing that most beneficiaries have relatively lower annual deductible amounts for outpatient services. An intriguing line of inquiry would be to assess if there are stark financial distribution differences between fraudulent and non-fraudulent claims. Such differences can act as red flags, indicating potentially suspicious claim patterns.

**Addressing Feedback:**

**Data Description:** Enhanced dataset and variable explanations.

**Dataset Integration:** Detailed the merging process of the four datasets.

**Plots:** All plots are now labeled and referenced in the text.

**Feature Visualization:** Prioritized plots showcasing class distribution differences.

**Correlation Analysis:** Added heatmap and elaborated on its significance.

**Features & Conclusion:** Presented key features in a table and clarified selection criteria.