**Data Science**

**By Devi Sri Charan**

[Claim Prediction Model 2](#_Toc1021270703)

[Exploratory Data Analysis on Historical Claims Data 3](#_Toc742763034)

[Preparing Dataset 3](#_Toc1976833366)

[EDA 3](#_Toc897478442)

[Information of the data 3](#_Toc1027415120)

[Statistics of the data 3](#_Toc2025626818)

[Missing Values 4](#_Toc990757989)

[Visual Distribution of Numerical Variables 4](#_Toc1198140445)

[Visual Distribution of Categorical Variables 5](#_Toc924022970)

[Time Series Analysis of Date Columns 6](#_Toc695966883)

[Predictive Model to estimate Claim Amounts 8](#_Toc340308061)

# Claim Prediction Model

## Exploratory Data Analysis on Historical Claims Data

### Preparing Dataset

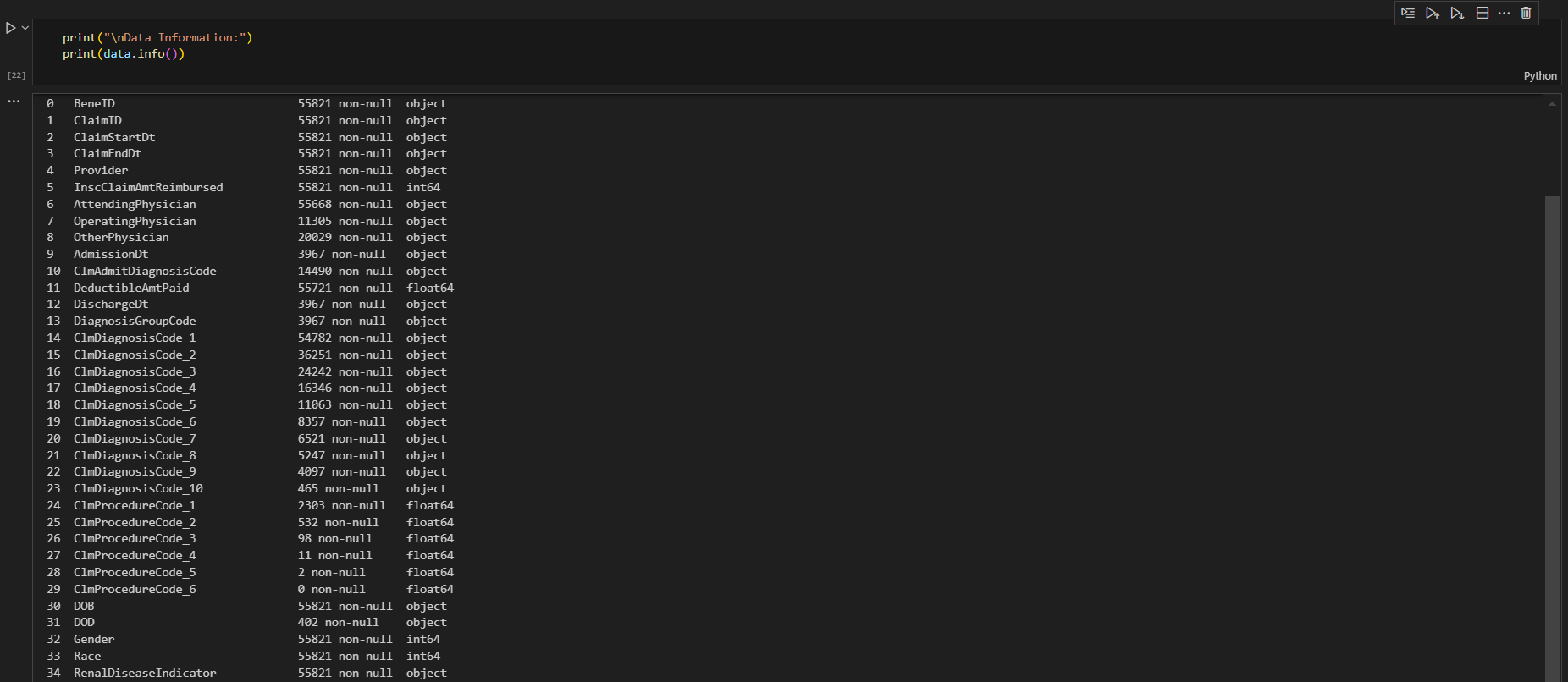
First, I have merged the Inpatient, Outpatient, Beneficiary and Provider Data



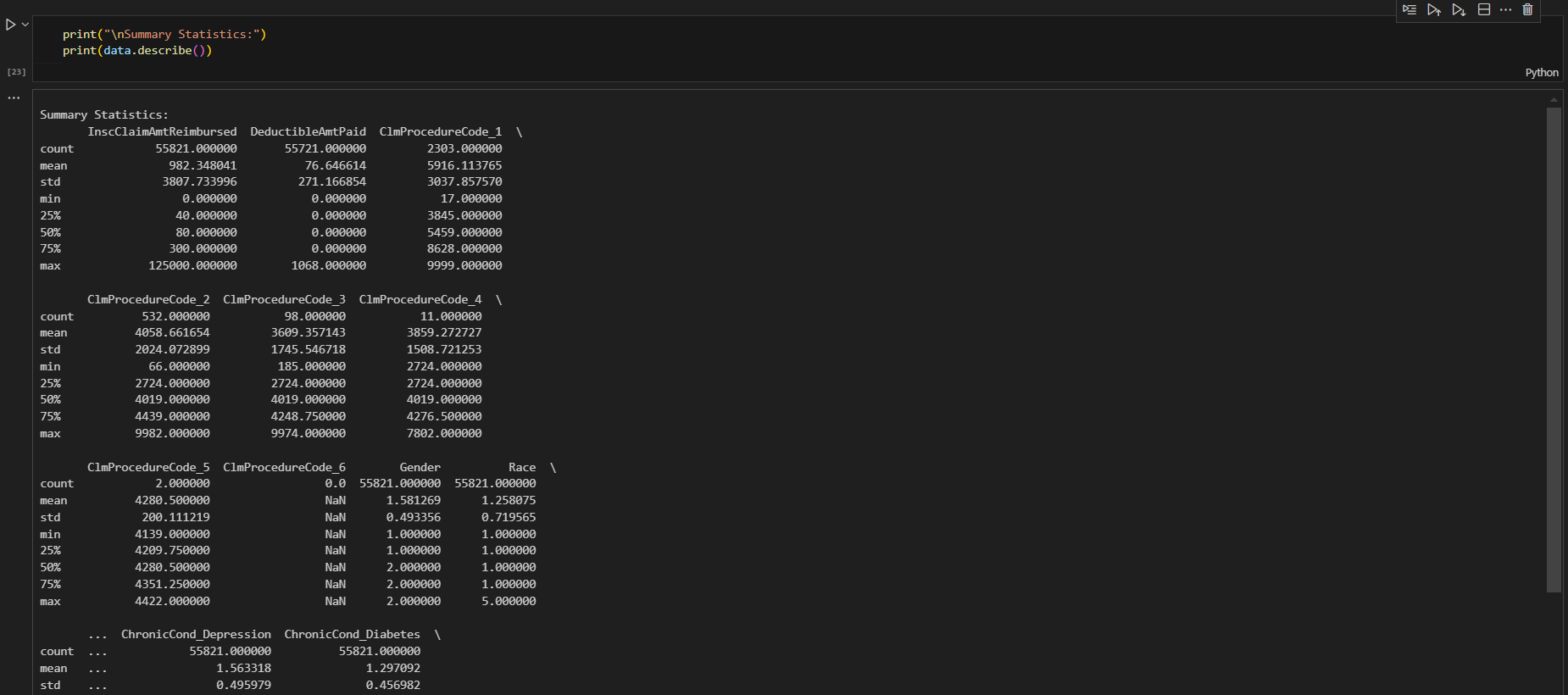
### EDA

#### Information of the data

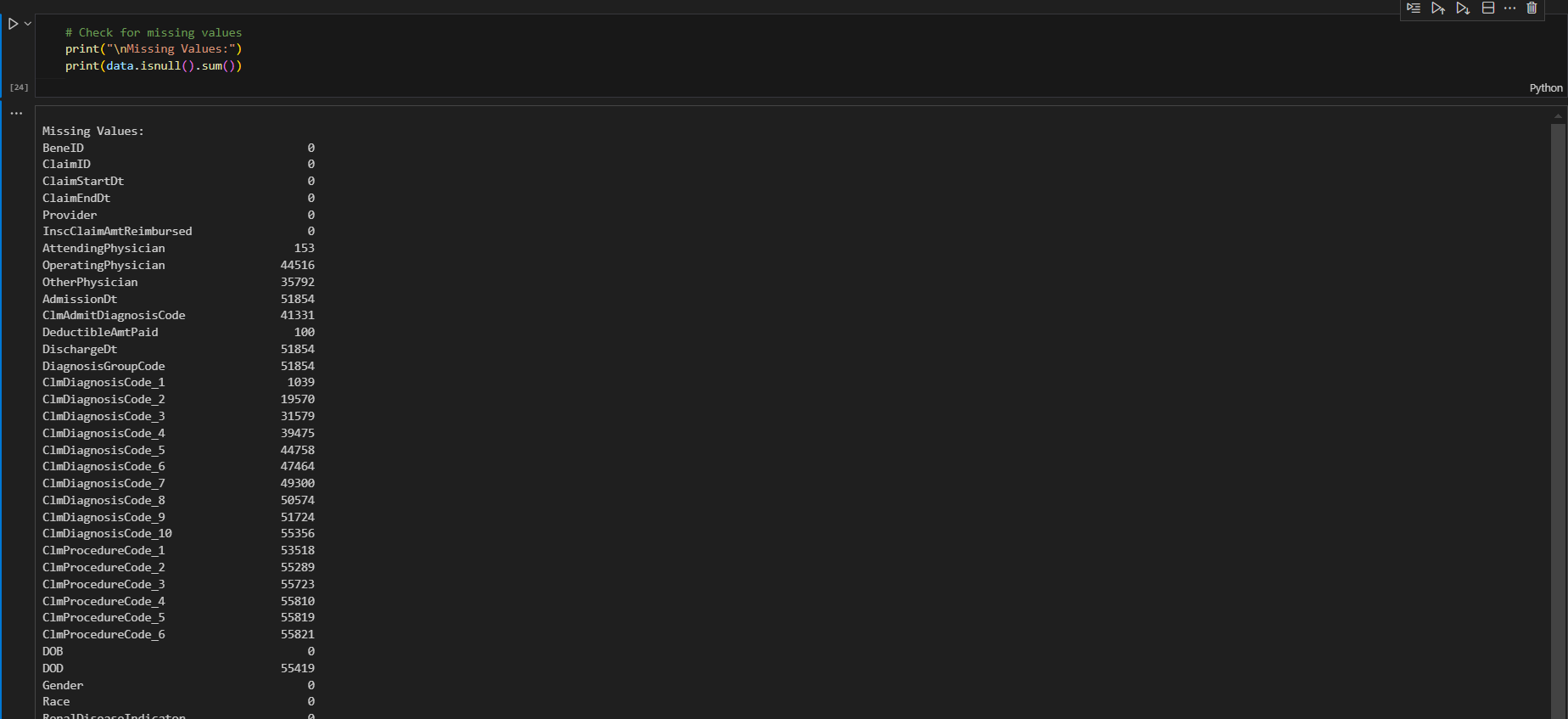
The dataset has 54 columns



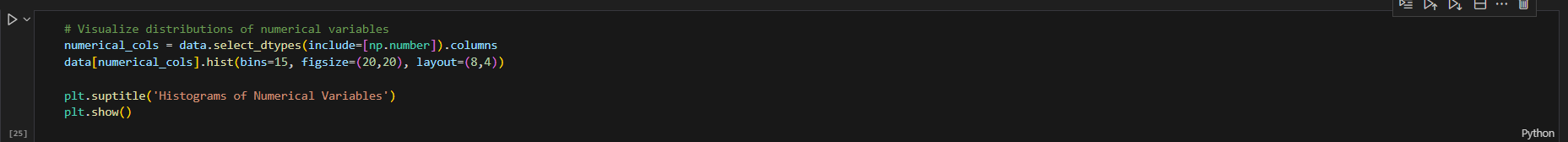
#### Statistics of the data

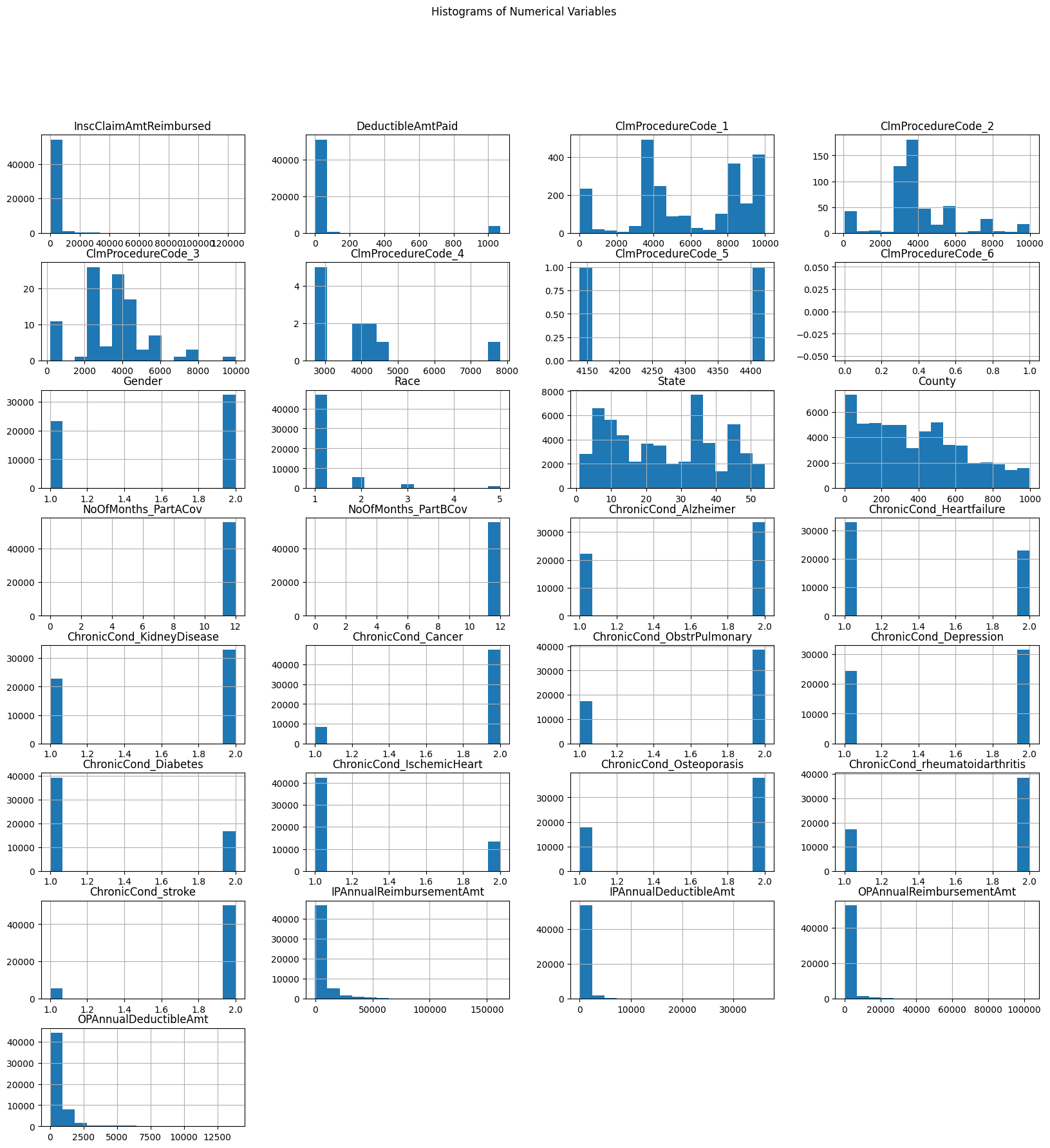


#### Missing Values

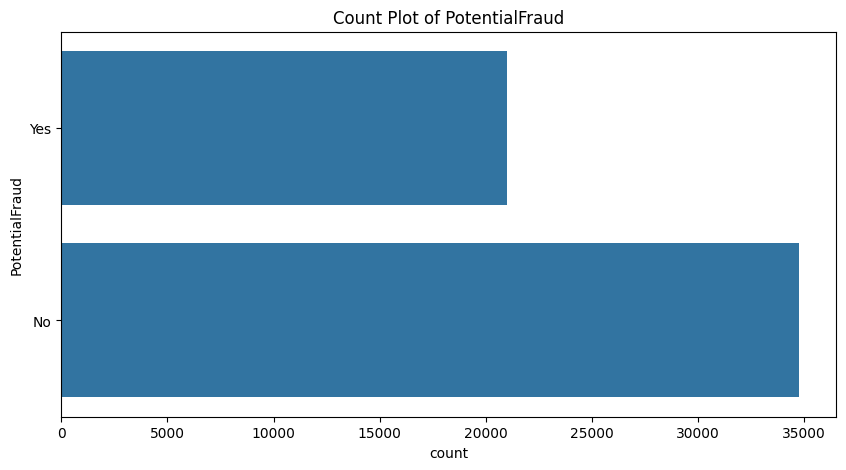
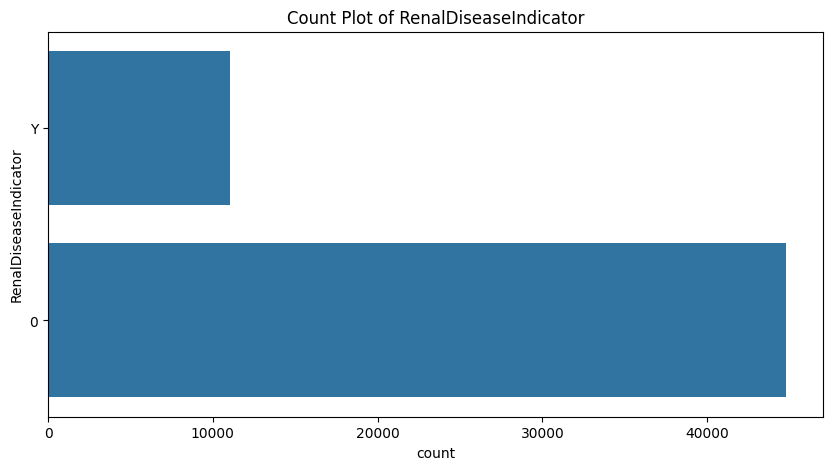
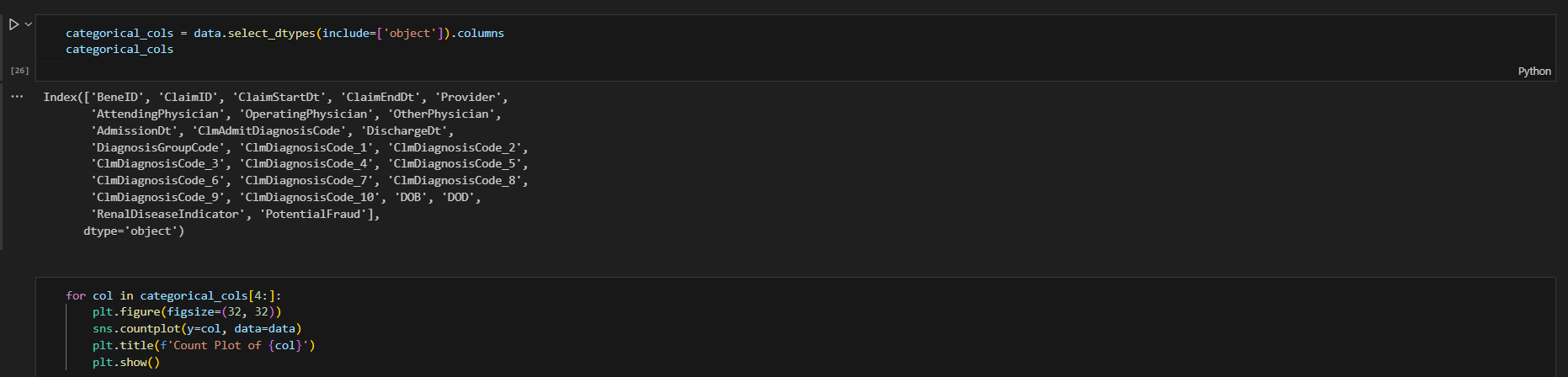


#### Visual Distribution of Numerical Variables

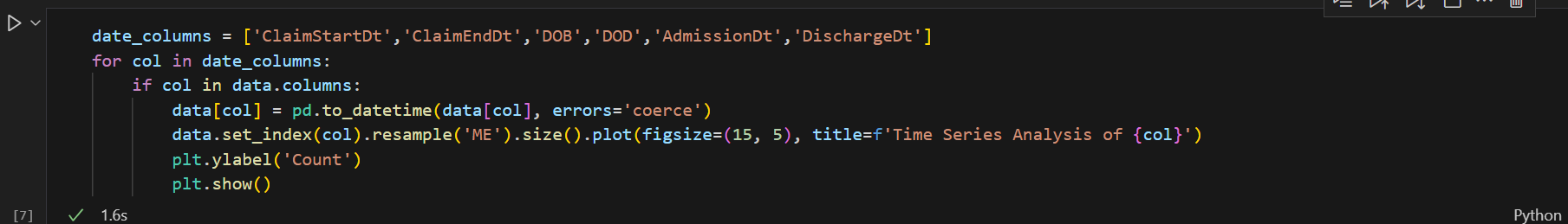


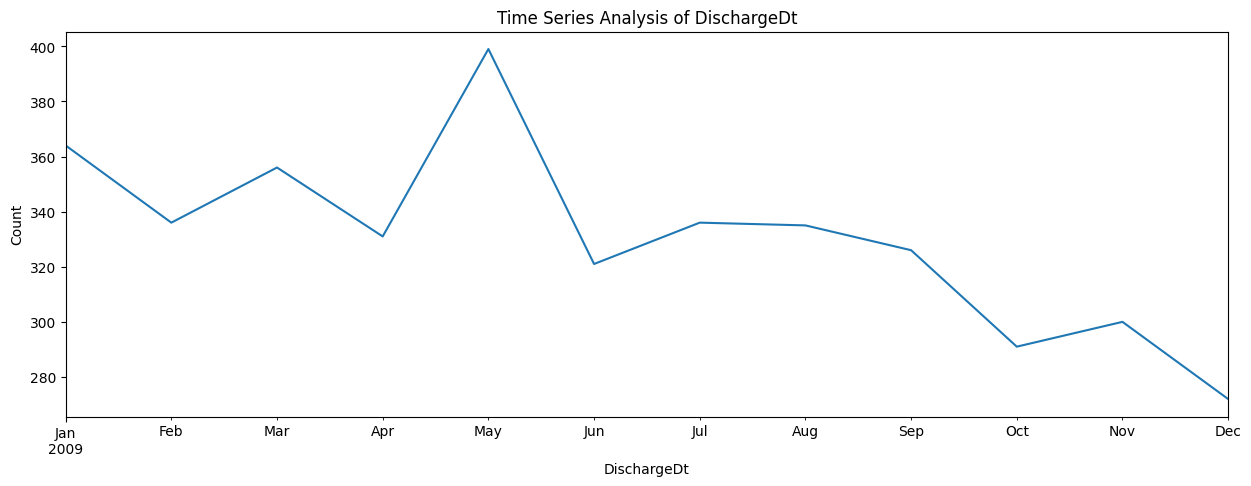
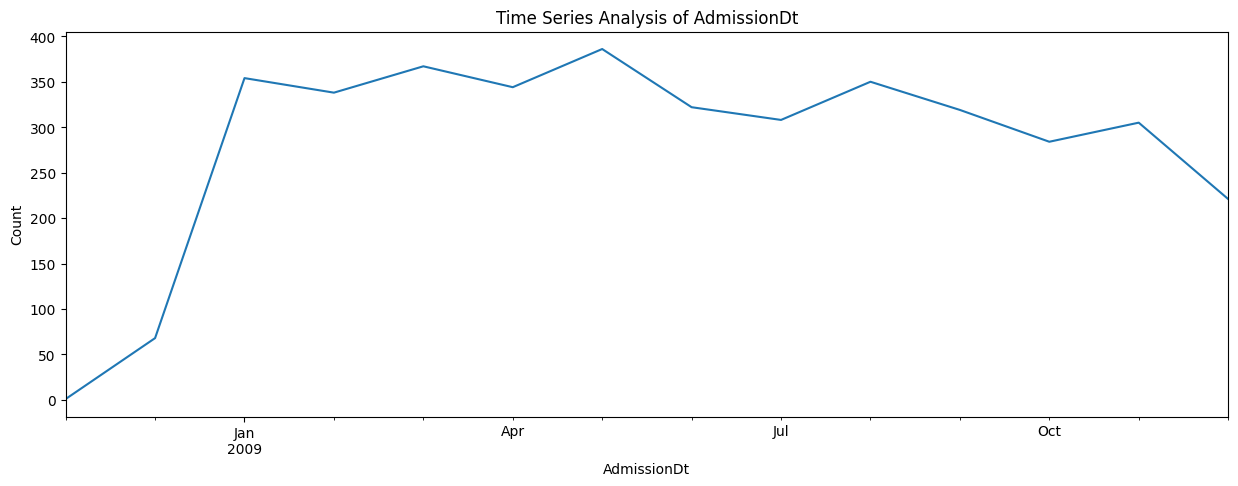
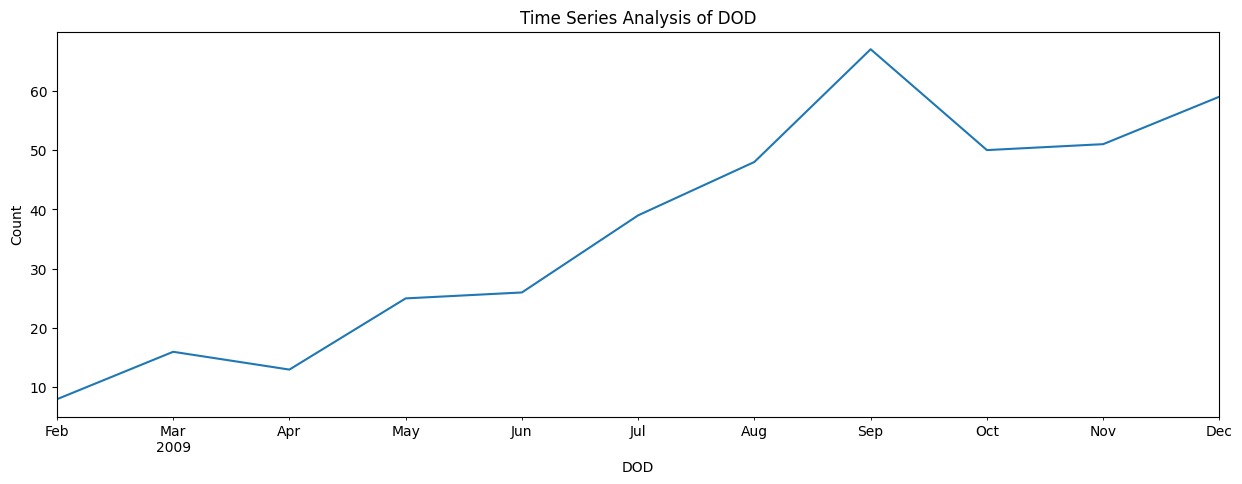
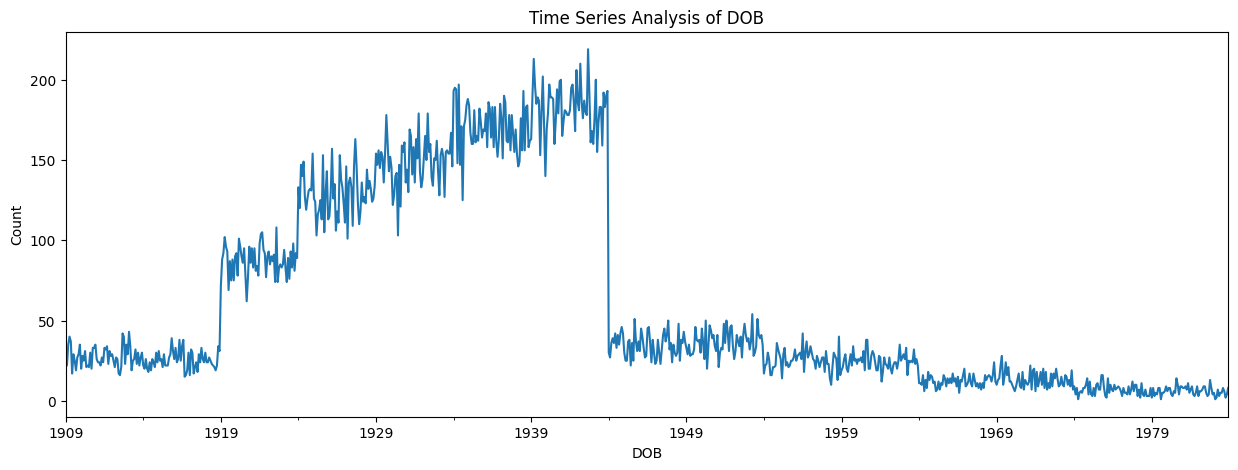
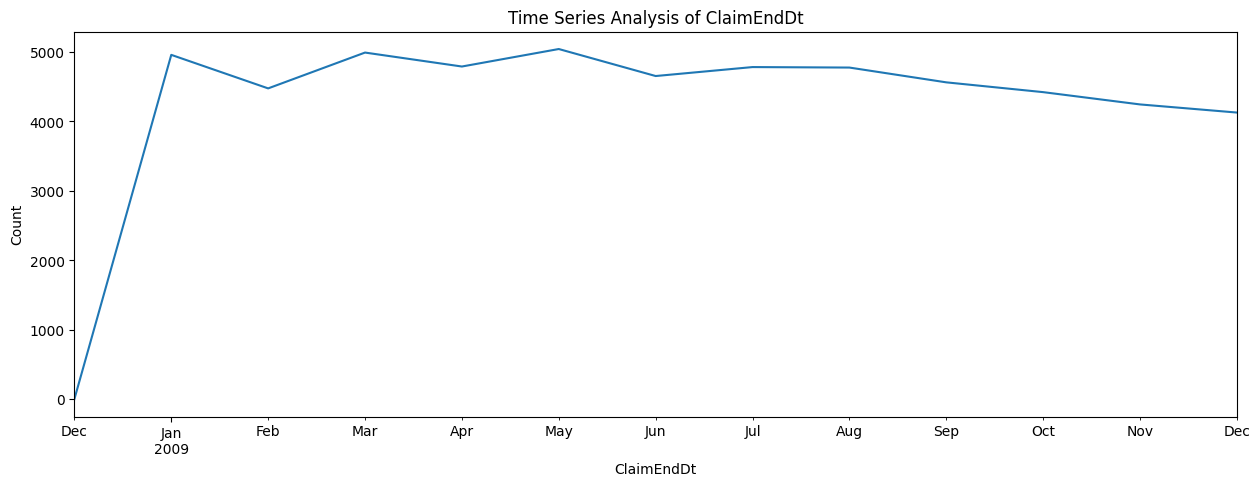
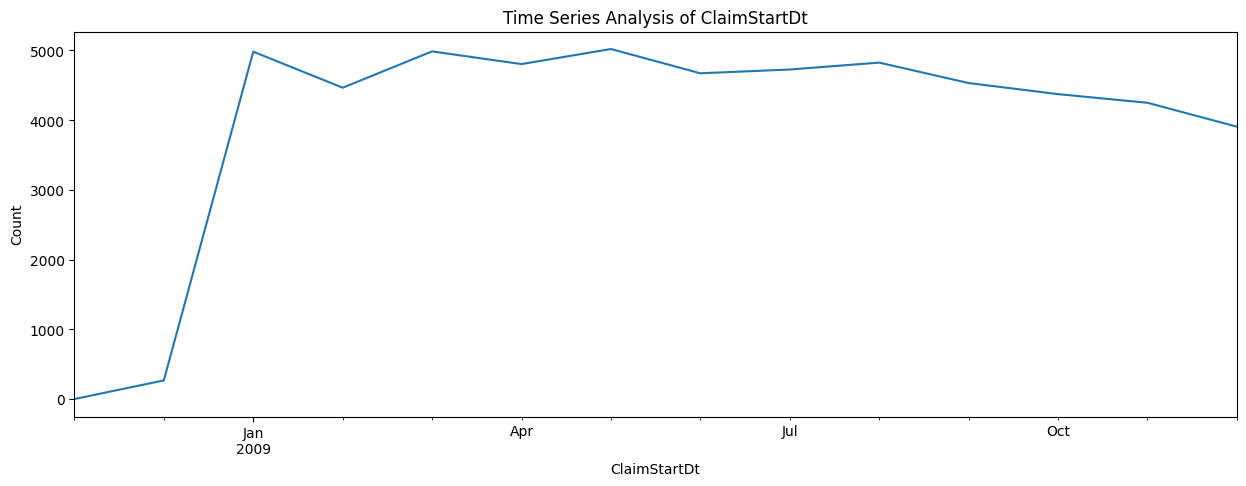


#### Visual Distribution of Categorical Variables



#### Time Series Analysis of Date Columns





## Predictive Model to estimate Claim Amounts

To build a predictive model for estimating the claim amounts (InscClaimAmtReimbursed), we need to identify relevant features from the dataset.

1. **Unnecessary columns**: Some columns, such as BeneID, ClaimID, ClaimStartDt, ClaimEndDt, Provider, and physician IDs (AttendingPhysician, OperatingPhysician, OtherPhysician), are likely unnecessary for predicting claim amounts.
2. **Potentially relevant features**:
   * Demographic information: Gender, Race
   * Health condition indicators: ChronicCond\_... columns
   * Annual reimbursement and deductible amounts: IPAnnualReimbursementAmt, IPAnnualDeductibleAmt, OPAnnualReimbursementAmt, OPAnnualDeductibleAmt
   * Indicator of potential fraud: PotentialFraud

Next, we will:

1. Select the relevant features.
2. Preprocess the data (handle missing values, encode categorical variables, etc.).
3. Split the data into training and testing sets.
4. Build a predictive model using different machine learning algorithms.
5. Evaluate the model's performance.

