Test_Beneficiarydata-1542969243754.csv					@dropdown	
Feature Name	**Feature Description, what does the feature mean?**	**Possible Values as Examples now**	**dtypes**	**Values after Cleaning**	**NA in Test**	**NA in Train**
Bene ID	Unique ID of the beneficiary	BENE11001	object	BENE11001	No NA's	
DOB	Date of Birth of the beneficiary	1943-01-01	object	Used for create new column "Age"	No NA's	
DOD	Date of Death of the beneficiary	99% is NA, 1% 2009-12-01,	object	Used for create new column "Age"	99,1% NA	
Gender	Gender of the beneficiary	1 = male, 2 = female	int64	1=male, 0=female	No NA's	
Race	Race	1,2,3,4,5	int64	1,2,3,4,5	No NA's	
State	State of the beneficiary	0-51	object	0-51	No NA's	
Country	Country Code for the country where the beneficiary belongs to	230,360,400 etc	int64	230,360,400 etc	No NA's	
NoOfMonths_PartACov	It represents number of months of part-A coverage.	12	int64	Drop, no Value for us	No NA's	
NoOfMonths_PartBCov	It represents number of months of part-B coverage.	12	int64	Drop, no Value for us	No NA's	
Renal Disease Indicator	It represents a code which indicates wether a beneficiary had a long history of kidney disease at the time of buying a plan	Y,0 (Y=Yes=16,51%, 0=No RDI=83,49%)	int64	Y=1, 0=0	No NA's	
ChronicCond_Alzheimer	It represents a code which indicates whether a beneficiary had a chronic condition of a specific disease at the time of buying a plan.	1,2	int64	Y=1, 0=0	No NA's	No NA's
ChronicCond_Heartfailure	It represents a code which indicates whether a beneficiary had a chronic condition of a specific disease at the time of buying a plan.	1,2	int64	Y=1, 0=0	No NA's	No NA's
ChronicCond_KidneyDisease	It represents a code which indicates whether a beneficiary had a chronic condition of a specific disease at the time of buying a plan.	1,2	int64	Y=1, 0=0	No NA's	No NA's
ChronicCond_Cancer	It represents a code which indicates whether a beneficiary had a chronic condition of a specific disease at the time of buying a plan.	1,2	int64	Y=1, 0=0	No NA's	No NA's
ChronicCond_ObstrPulmonary	It represents a code which indicates whether a beneficiary had a chronic condition of a specific disease at the time of buying a plan.	1,2	int64	Y=1, 0=0	No NA's	No NA's
ChronicCond_Depression	It represents a code which indicates whether a beneficiary had a chronic condition of a specific disease at the time of buying a plan.	1,2	int64	Y=1, 0=0	No NA's	No NA's
ChronicCond_Diabetes	It represents a code which indicates whether a beneficiary had a chronic condition of a specific disease at the time of buying a plan.	1,2	int64	Y=1, 0=0	No NA's	No NA's
ChronicCond_IschemicHeart	It represents a code which indicates whether a beneficiary had a chronic condition of a specific disease at the time of buying a plan.	1,2	int64	Y=1, 0=0	No NA's	No NA's
ChronicCond_Osteoporasis	It represents a code which indicates whether a beneficiary had a chronic condition of a specific disease at the time of buying a plan.	1,2	int64	Y=1, 0=0	No NA's	No NA's
ChronicCond_rheumatoidarthritis	It represents a code which indicates whether a beneficiary had a chronic condition of a specific disease at the time of buying a plan.	1,2	int64	Y=1, 0=0	No NA's	No NA's
ChronicCond_stroke	It represents a code which indicates whether a beneficiary had a chronic condition of a specific disease at the time of buying a plan.	1,2	int64	Y=1, 0=0	No NA's	No NA's
IPAnnualReimbursementAmt	It consists of the maximum reimbursement amount allocated to a beneficiary for annual hospitalization on the basis of a insurance plan.	-1000 to 156.000	int64		No NA's	
IPAnnualDeductibleAmt	It consists of the maximum co-payment to be bourne by a beneficiary for annualy if in case gets hospitalized	0 to 38.3000	int64		No NA's	
OPAnnualReimbursementAmt	It consists of the maximum reimbursement amount allocated to a beneficiary for annual non-hospitalization on the basis of a insurance plan.	-60 to 97.5	int64		No NA's	
OPAnnualDeductibleAmt	It consists of the maximum co-payment to be bourne by a beneficiary annualy if in case only visited a hospital without admission.	0 to 13.8k	int64		No NA's	
New Features:						
Bene_Age	Age of the beneficiary, DOD-DOB/365, check if there is a correlation between fraud and age of the beneficiary					
Bene_Alive	Beneficiary is alive or not?:					
Test_Inpatientdata-1542969243754.csv	Dataset of Patients who stayed in a clinic or a hospital					
BeneID	Unique ID of the beneficiary	BENE11014	object		No NA's	
ClaimID	Unique ID of the Claim submitted by the provider	CLM67387	object		No NA's	
ClaimStartDt	start date of the claim	2009-09-09	object		No NA's	
ClaimEndDt	end date of the claim	2009-09-16	object		No NA's	
Provider	Unique ID of the Provider	PRV57070	object		No NA's	
InscClaimAmtReimbursed	Represents the amout re-imbursed for that particular claim	9000	int64		No NA's	
AttendingPhysician	ID of the Physican who attended the patient	PHY317786	object		No NA's	
OperatingPhysician	ID of the Physican who operated the patient	41% NA, PHY427017	object		41,48 NaN!	
OtherPhysician	ID(?) of the physican who assisted the patient and other physicans	89% NaN, Other 10%	object		89,39 NaN!	
AdmissionDt	The Date on which the patient was admitted in the hospital	2009-09-09	object		No NA's	
ClmAdmitDiagnosisCode	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim	V6109, 78650	object		No NA's	
DeductibleAmtPaid	represents the amount bourne by the patient for that claim. It can be thought of as the co-payment which is some percentage of the total amount to be paid by the patient		float64		NaN 2.05%	
DischargeDT	Represents the date on which the patient was discharged from the hospital		object		No NA's	
DiagnosisGroupCode	Represents a group code for the diagnosis done on the patient. This code, I believe, might be a superset of many other diseases which are being diagnosed in the patient (?)		object		No NA's	
ClmDiagnosisCode_1	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		No NA's	
ClmDiagnosisCode_2	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		No NA's	
ClmDiagnosisCode_3	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		No NA's	
ClmDiagnosisCode_4	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 4,23%	
ClmDiagnosisCode_5	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		Nan 7,53%	
ClmDiagnosisCode_6	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		Nan 12,53%	
ClmDiagnosisCode_7	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 18.18%	
ClmDiagnosisCode_8	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 24.71%	
ClmDiagnosisCode_9	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 33.90%	
ClmDiagnosisCode_10	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 90.71%	
ClmProcedureCode_1	Represents the Code of the medical treatments or proedures performed by the provider or physicians for medication of a patient for a specific claim		float64		NaN 43,12%	
ClmProcedureCode_2	Represents the Code of the medical treatments or proedures performed by the provider or physicians for medication of a patient for a specific claim		float64		NaN 86,87%	
ClmProcedureCode_3	Represents the Code of the medical treatments or proedures performed by the provider or physicians for medication of a patient for a specific claim		float64		NaN 97,67%	
ClmProcedureCode_4	Represents the Code of the medical treatments or proedures performed by the provider or physicians for medication of a patient for a specific claim		float64		NaN 99,70%	
ClmProcedureCode_5	Represents the Code of the medical treatments or proedures performed by the provider or physicians for medication of a patient for a specific claim		float64		NaN 99,98%	
ClmProcedureCode_6	Represents the Code of the medical treatments or proedures performed by the provider or physicians for medication of a patient for a specific claim		float64	Most of them NA's	NaN 100%	
Test_Outpatient-1542969243754.csv	Dataset of Patients wo went to the doctor					
BeneID	Unique ID of the beneficiary		object		No NA's	
ClaimID	Unique ID of the Claim submitted by the provider		object		No NA's	
ClaimStartDt	start date of the claim		object		No NA's	
ClaimEndDt	end date of the claim		object		No NA's	
Provider	Unique ID of the Provider		object		No NA's	
InscClaimAmtReimbursed	Represents the amout re-imbursed for that particular claim		int64		No NA's	
AttendingPhysician	ID of the Physican who attended the patient		object		No NA's	
OperatingPhysician	ID of the Physican who operated the patient		object		NaN 82,83%	
OtherPhysician	ID(?) of the physican who assisted the patient and other physicans		object		NaN 62,16%	
ClmDiagnosisCode_1	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 2,05%	
ClmDiagnosisCode_2	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 37,93%	
ClmDiagnosisCode_3	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 60,85%	
ClmDiagnosisCode_4	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 75,79%	
ClmDiagnosisCode_5	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 85,72%	
ClmDiagnosisCode_6	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 90,62%	
ClmDiagnosisCode_7	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 93,67%	
ClmDiagnosisCode_8	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 95,60%	
ClmDiagnosisCode_9	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 97,17%	
ClmDiagnosisCode_10	Represents the Code of the diagnosis performed by the provider or physicians on the patient for a specific claim		object		NaN 99,79%	
ClmProcedureCode_1	Represents the Code of the medical treatments or proedures performed by the provider or physicians for medication of a patient for a specific claim		float64		NaN 99,97%	
ClmProcedureCode_2	Represents the Code of the medical treatments or proedures performed by the provider or physicians for medication of a patient for a specific claim		float64		NaN 99,99%	
ClmProcedureCode_3	Represents the Code of the medical treatments or proedures performed by the provider or physicians for medication of a patient for a specific claim		float64		NaN 100%	
ClmProcedureCode_4	Represents the Code of the medical treatments or proedures performed by the provider or physicians for medication of a patient for a specific claim		float64		NaN 100%	
ClmProcedureCode_5	Represents the Code of the medical treatments or proedures performed by the provider or physicians for medication of a patient for a specific claim		float64		NaN 100%	
ClmProcedureCode_6	Represents the Code of the medical treatments or proedures performed by the provider or physicians for medication of a patient for a specific claim		float64		NaN 100%	
DeductibleAmtPaid	represents the amount bourne by the patient for that claim. It can be thought of as the co-payment which is some percentage of the total amount to be paid by the patient		int64		No NA's	
ClmAdmitDiagnosisCode			object		NaN 79.49%	
Test_Label-1542969243754.csv						
Provider			object		No NaN	
New Features v.1.0 clean dataframe and merge to one frame:						
Bene_Admitted?	Patient was admitted to the hospital, yes or no, 1 or 0, Idea: if Bene admitted to the hospital, prob. for fraud should be higher	1=Admitted, 0=No				
Claim_Duration	ClaimEndDt-ClaimStartDt					
Admitted_Duration	DischargeDT-AdmissionDT					
New Features v.2.0 Creating Columns for	**Why useful?**					
Change the NAN in Physicians to Zero => Model is able to recognize patterns.	Because there could be the case, Physicians are not "logged in" to the system and it seems like they are not there but in reality they performing some diagnosicts.					
prv_bene_age_sum	The Sum of the Age of the Beneficiaries per Provider: Some Provider will be working with older patients. The avg. Age could be important for fraudulent behaviour					
prv_total_claims	The Number of total Claims per Provider					
prv_total_claims_for_physicians (currently working with the Attending Physician)	The Number of total Claims for the Attending Physicians per Provider					
prv_physician_count	The Number of unique physicians working for each provider differentiated between Attending, Operating and Other					
prv_allphysician_count	The Number of all 3 types of physicians for each Provider					
prv_insc_claim_reimb_amt	Calculate the total insurance reimbursement amount per provider.					
prv_total_bene	Calculate the total number of unique beneficiaries per provider					
prv_total_chronic_bene (now also contains renal disease indicator)	Calculates the total number of beneficiaries per provider for each chronic condition					
prv_diagnosis_count	Count non-null occurrences of the ClaimAdmitDiagnosisCode and ClmDiagnosisCode 1-3 per provider					
prv_most_frequent_claim_codes	The most frequent claim codes per Provider					
prv_most_frequent_physicians	The most frequent appearing Physicians per Provider					
prv_bene_amount_avg	Return a Dask DataFrame with BeneID, AllocatedAmount (as-is), and summed Deductible & Reimbursed amounts					
add_quarter						
avg_claim_cost_indicator_per provider						
avg_claims_per_provider						
prv_avg_claims (we have to run this after the merge)	The Average of Claims per Provider					
prv_avg_claim_cost (after merging)	The Average of Claimcosts per Provider					
prv_median_claim_cost	The Median of Claimcosts per Provider					
prv_	The First Step for Preparing an analysis of the Budgets on Provider Level					
	The First Step for Preparing an analysis of the Budgets on Provider Level					
	To make a difference between the					

Module Name	Tasks	Description
module_1_data_cleaning.ipynb	**Data Loading from:**	**s3://medicare-fraud-data-25-05-2025/raw**
clean_inpatient_function (Test/Train)	Define data types	Defines expected column data types for the inpatient dataset to optimize memory usage and ensure correct parsing.
	Specify date columns	Identifies which columns should be parsed as dates during CSV reading.
	Read CSV with Dask	Loads the dataset in parallel using Dask, with correct types and missing value handling.
	Drop missing values	Removes rows where Provider or InscClaimAmtReimbursed is null.
	Copy DataFrame	Creates a copy of the DataFrame for safe editing.
	Convert to datetime	Converts date columns to datetime format, coercing invalid values to NaT.
	Calculate ClaimDuration	Creates a column for the duration of the insurance claim in days.
	Calculate HospitalDuration	Creates a column for the hospital stay duration in days.
	Reorder ClaimDuration	Places ClaimDuration immediately after ClaimEndDt in the column order.
	Group hospital date columns	Moves AdmissionDt, DischargeDt, and HospitalDuration after ClaimDuration.
	Apply new column order	Rearranges the DataFrame columns according to the new logical structure.
	Return cleaned DataFrame	Outputs the cleaned and transformed Dask DataFrame.
clean_outpatient_function (Test/Train)	Define data types	Specifies column data types expected in the outpatient dataset for consistency and memory efficiency.
	Specify date columns	Identifies ClaimStartDt and ClaimEndDt to be parsed as datetime fields.
	Read CSV with Dask	Loads the outpatient dataset in parallel using Dask with predefined types and date parsing.
	Drop missing values	Removes rows where Provider or InscClaimAmtReimbursed is null.
	Copy DataFrame	Creates a copy of the DataFrame for safe editing
	Return copied DataFrame	Outputs the copied DataFrame without additional transformations.
clean_bene_function (Test/Train)	Read CSV with Dask	Loads the beneficiary dataset using Dask for scalable processing.
	Copy DataFrame	Creates a working copy of the original DataFrame.
	Normalize gender values	Replaces 2 with 0, making gender binary (0 = female, 1 = male).
	Drop unused coverage columns	Removes NoOfMonths_PartACov and NoOfMonths_PartBCov as they're not needed.
	Transform RenalDiseaseIndicator	Converts "Y" to 1, all other/missing to 0, and casts to integer.
	Normalize chronic condition flags	Replaces value 2 (unknown/missing) with 0 across all chronic condition columns.
	Convert date columns	Parses DOB (Date of Birth) and DOD (Date of Death) as datetime objects.
	Fill missing DOD with survey end	Sets missing DOD values to a default date (2009-12-01) indicating still alive.
	Calculate age	Computes Bene_Age by subtracting DOB from DOD and converting to years.
	Create alive flag	Adds Bene_Alive column: 1 if still alive (DOD = default), otherwise 0.
	Reorder age and alive columns	Inserts Bene_Age and Bene_Alive directly after DOD for better readability.
	Return cleaned DataFrame	Outputs the transformed Dask DataFrame with all applied changes.
clean_label_function	Read CSV with Dask	Loads the label dataset from the specified path using Dask.
	Copy DataFrame	Creates a working copy of the loaded DataFrame for consistency/safety.
	Return copied DataFrame	Returns the unmodified DataFrame copy.
	Data Saving to:	**s3://medicare-fraud-data-25-05-2025/clean/**
module_2_feature_engineering.ipynb	**Data Loading from:**	**s3://medicare-fraud-data-25-05-2025/clean/**
	Provider-Level Aggregations	Numerical columns (e.g., claim duration, deductible amount) with sum, mean, std, max, min. Annual reimbursement and deductible amounts with mean and max. Binary chronic condition flags with sum and mean. Counts of unique values (nunique) for diagnosis codes, physician IDs, gender. Mode (most frequent value) for categorical features like race, state, county.
	Physician-Related Features	Missing values in physician columns are filled with zero. Calculates total claims per provider and per physician type (currently attending physician). Counts unique physicians per provider across different physician roles.
	Beneficiary-Level and Claim-Level Aggregates	Sum of beneficiary ages per provider. Total number of claims per provider. Total unique beneficiaries and chronic condition patients per provider. Counts of non-null diagnosis codes per provider. Most frequent diagnosis and claim codes per provider.
	Financial Features	Computes total insurance reimbursement per provider. For each beneficiary, sums deductible and reimbursed amounts and computes averages per provider. Calculates percentage of allocated amount used (insured claim reimbursed minus deductible relative to allocated amount).
	Temporal Feature	Adds a quarter column derived from claim start date, differentiating quarters for years 2008 and 2009.
	Feature Merging	All engineered features are merged into a consolidated Dask DataFrame keyed by Provider for both train and test datasets.
	Data Saving to:	**s3://medicare-fraud-data-25-05-2025/merged_ready/**
module_3_merge_tables.ipynb	**Data Loading from:**	**s3://medicare-fraud-data-25-05-2025/merged_ready/**
	Data Saving to:	
module_4_preprocessing.ipynb	**Data Loading from:**	
	Data overview	Display of data types (dtypes) and basic information (shape, head, missing values)
	Preprocessor initialisation	Instantiating the MedicarePreprocessor class with df_train
	drop_unused_columns	Remove unused diagnosis/doctor columns
	fill_missing	Fill in missing values: numerical → mean value, categorical → 'MISSING'
	encode_categoricals	Categorise & convert to integer codes
	feature_engineering	New features: avg_cost_per_claim & perc_chronic_alz
	scale_numeric_features	Standardisation of all numerical characteristics except provider & target variable
	get_processed_df	Persistence and return of the processed DataFrame
	Train/Test-Split	Split by target class (0/1), then 80/20 split per class, shuffle, persist
	DaskDMatrix Creation	Conversion to Dask-compatible DMatrix objects for XGBoost
	Modeltraining	Training of an XGBoost model with `scale_pos_weight` for class balancing
	Prediction	Calculation of probabilities and binary predictions (Threshold = 0.5)
	Grid Search	Tuning via hyperparameter combinations (max_depth, learning_rate etc.)
	Final Model	Training with best parameters, prediction with threshold = 0.95, final evaluation
	Data Saving to:	
module_5_model_training.ipynb	**Data Loading from:**	**s3://medicare-fraud-data-25-05-2025/processed_new/train/**
	Load data	Load CSV files from S3 into a Dask DataFrame with predefined dtypes.

	Inspect data types	Print all column names and their data types for verification.	
	Check shape	Print the shape of the training dataset.	
	Preview data	Show the first 3 rows of the dataset.	
	Check missing values	Compute and print all columns with missing values.	
	Initialize Preprocessor	Instantiate the MedicarePreprocessor with the raw Dask DataFrame.	
	Drop unused columns	Remove specific categorical columns not used in training.	
	Fill missing values	Impute missing numeric values with mean and categorical with 'MISSING'.	
	Encode categoricals	Convert categorical columns to category dtype and encode to integers.	
	Feature engineering	Create new features such as ratios like cost per claim or percent with Alzheimer's.	
	Scale numeric features	Standard scale numeric features excluding target and ID columns.	
	Persist processed data	Trigger computation and persist the processed DataFrame in memory.	
	Split by label	Split the dataset into fraud and non-fraud subsets based on target.	
	Random split each class	Perform an 80/20 train/test split for each label subset.	
	Combine splits	Concatenate the 0/1 class splits into full train and test sets.	
	Shuffle datasets	Shuffle both train and test sets based on Provider ID.	
	Separate features/labels	Separate features (X) and labels (y) for both train and test sets.	
	Persist final sets	Persist X_train, y_train, X_test, and y_test in memory.	
	Initialize DaskDMatrix	Create DaskDMatrix objects from X and y for training/testing.	
	Compute scale_pos_weight	Calculate ratio of negative to positive samples for class balancing.	
	Train XGBoost model	Train the model using Dask and XGBoost with given parameters.	
	Get predictions (probabilities)	Predict probability scores on the test set.	
	Threshold probabilities	Convert probabilities to binary class labels using threshold 0.5 or 0.95.	
	Classification report	Print precision, recall, F1-score, and accuracy using sklearn.	
	Confusion matrix	Compute and print confusion matrix to analyze prediction errors.	
	Compute metrics	Calculate test AUC and accuracy using scikit-learn.	
	Create param grid	Define grid of XGBoost parameters for tuning.	
	Train with grid search	Evaluate AUC score across all parameter combinations.	
	Select best model	Identify parameter set with highest validation AUC.	
	Predict best model	Predict using best-performing model and print metrics.	
	Data Saving to:		
module_6_frontend.ipynb	**Data Loading from:**		
	UI Setup and Styling	Uses Streamlit to build an interactive web dashboard. Custom CSS styles for input fields, buttons, and page background for a modern look. Page title, layout, and icon are configured.	
	Login System	Simple sidebar login with hardcoded username/password pairs. Only logged-in users can access the dashboard; others see a warning and the app stops.	
	Data Loading	Loads a CSV file with provider-level fraud prediction results and features. Loads a JSON with explanations for each feature (to show human-friendly info).	
	Filtering and Provider Selection	Sidebar slider filters providers by fraud probability (0–100%). Dropdown lets users select one provider from filtered data.	
	Feature Comparison Visualization	Shows a bar chart comparing dollar-based financial features (e.g., allocated budget, reimbursed amount) for the selected provider.	
	Fraud Probability Gauge Animation	Animated gauge showing fraud probability percent for the selected provider. Gauge color switches from green (low risk) to red (high risk).	
	OpenAI GPT-4 Explanation Integration	Uses OpenAI API to generate professional, contextual explanations of the impact of top features on fraud prediction. Cached calls for efficiency.	
	Report Generation and Download	Button to generate a detailed explanation report for the selected provider using the OpenAI explanations. Report displayed in an expandable section in the app. PDF report generated dynamically using ReportLab with formatting and branding. Download button lets users save the explanation report as a PDF.	
	Sidebar Metrics and Info	Shows total providers analyzed and count of high-risk providers. Provides info text explaining the dashboard's purpose.	
module_7_id9cm			

BeneID	Drop
ClaimID	Drop
ClaimStartDt	Drop
ClaimEndDt	Drop
Provider	Keep
InscClaimAmtRe	Drop
AttendingPhysici	Drop
OperatingPhysici	Drop
OtherPhysician	Drop
AdmissionDt	Drop
ClmAdmitDiagno	Drop
DeductibleAmtPa	Drop
DischargeDt	Drop
DiagnosisGroupC	Drop
ClmDiagnosisCo	Drop
ClmDiagnosisCo	Drop
ClmDiagnosisCo	Drop
ClmDiagnosisCo	Drop
ClmDiagnosisCo	Drop
ClmDiagnosisCo	Drop
ClmDiagnosisCo	Drop
ClmDiagnosisCo	Drop
ClmDiagnosisCo	Drop
ClmDiagnosisCo	Drop
ClmDiagnosisCo	Drop
ClmProcedureCo	Drop
ClmProcedureCo	Drop
ClmProcedureCo	Drop
ClmProcedureCo	Drop
ClmProcedureCo	Drop
ClmProcedureCo	Drop
ClaimDuration	Keep
HospitalDuration	Keep
DOB	Drop
DOD	Drop
Gender	Drop
Race	Drop
RenalDiseaseInd	Drop
State	Keep
County	Keep
ChronicCond_Alz	Drop
ChronicCond_He	Drop
ChronicCond_Kid	Drop
ChronicCond_Ca	Drop
ChronicCond_Ob	Drop
ChronicCond_De	Drop
ChronicCond_Dia	Drop

ChronicCond_Isc	Drop
ChronicCond_Os	Drop
ChronicCond_rhe	Drop
ChronicCond_str	Drop
IPAnnualReimbu	Drop
IPAnnualDeducti	Drop
OPAnnualReimb	Drop
OPAnnualDeduc	Drop
Bene_Age	Drop
Bene_Alive	Drop
PotentialFraud	Keep
AllocatedAmount	Drop
Year	Drop
Quarter	Keep

New Features v.1.0 clean dataframe and merge to one frame:	
Bene_Admitted?	Patient was admitted to the hospital, yes or no, 1 or 0, Idea: if Bene admitted to the hospital, prob. for fraud should be higher
Claim_Duration	ClaimEndDt-ClaimStartDt
Admitted_Duratic	DischargeDT-AdmissionDT