# **📄 Data Ingestion & Initial Profiling Report**

## **1. Data Ingestion**

**Actions:**

* Mounted Google Drive and created a clean project structure:  
  + /data/raw/, /data/processed/, /models/, /reports/, /notebooks/, /src/
* Placed all raw data files in /data/raw/

**Raw Files Loaded:**

* Train.csv, Test.csv, Train\_Beneficiarydata.csv, Test\_Beneficiarydata.csv
* Train\_Inpatientdata.csv, Test\_Inpatientdata.csv
* Train\_Outpatientdata.csv, Test\_Outpatientdata.csv

## **2. Table Structure & Schema**

| **File** | **Rows** | **Columns** | **Key Columns** | **Notes** |
| --- | --- | --- | --- | --- |
| Train.csv | 5,410 | 2 | Provider (target), PotentialFraud | No missing values |
| Test.csv | 1,353 | 1 | Provider | No missing values |
| Train\_Beneficiarydata.csv | 138,556 | 25 | BeneID | DOD mostly missing |
| Train\_Inpatientdata.csv | 40,474 | 30 | BeneID, Provider, ClaimID | See missing summary |
| Train\_Outpatientdata.csv | 517,737 | 27 | BeneID, Provider, ClaimID | See missing summary |

## **3. Table Relationships**

* **Provider**: Primary key in Train.csv/Test.csv; foreign key in claims files.
* **BeneID**: Primary in beneficiary file, foreign in claims files.
* **ClaimID**: Unique per claim (not used for provider-level features).

## **4. Initial Data Profiling**

### **Row & Unique Counts**

* Providers in Train.csv: 5,410  
  + Providers in Inpatient claims: 2,092
  + Providers in Outpatient claims: 5,012
* Unique beneficiaries:  
  + Inpatient: 31,289
  + Outpatient: 133,980
  + Beneficiary master: 138,556

## **5. Missing Value Analysis**

### **Inpatient Claims:**

* **100% missing:** ClmProcedureCode\_6
* **>99% missing:** ClmProcedureCode\_5, ClmProcedureCode\_4
* **~88% missing:** OtherPhysician
* **>90% missing:** ClmDiagnosisCode\_10
* **~40% missing:** ClmProcedureCode\_1, OperatingPhysician
* Most diagnosis/procedure codes: missing increases with code number.

### **Outpatient Claims:**

* **100% missing:** ClmProcedureCode\_5, ClmProcedureCode\_6
* **~99% missing:** All other procedure codes and ClmDiagnosisCode\_10
* **>95% missing:** Higher-order diagnosis codes (8, 9, 10)
* **~80%+ missing:** OperatingPhysician, ClmAdmitDiagnosisCode
* **~62% missing:** OtherPhysician
* Core columns always present.

#### **Interpretation:**

* Diagnosis/procedure columns are designed this way (few claims have >3 codes).
* Physician columns often not applicable.
* No missing in identifiers, dates, or reimbursement amounts.

## **6. Visualizations**

* Created heatmaps to visualize missing value patterns in claims files.
* Observed blocks of missingness for advanced diagnosis/procedure/physician columns.