# DATA CLEANING LOG — CareNova Healthcare Dataset

## 1. Reason for Data Cleaning

The dataset consists of three core tables:

- **PatientInfo**
- **DoctorDetails**
- **TreatmentRecords**

Cleaning is required to:

- Ensure consistency before applying constraints (PK/FK).
- Standardize column names and formats.
- Remove duplicate entries.
- Handle missing or inconsistent values.
- Prepare a reliable data model for SQL Server + Power BI analysis.

## 2. Issues Identified

### PatientInfo Table

| Issue ID | Type of Issue | Description | Columns Affected | How It Was Found |
|---|---|---|---|---|
| **P1** | Naming Issue | Column names not standardized | PatientID → Patient_ID | Manual inspection |
| **P2** | Trailing Spaces | Spaces in text fields | Name, Gender, Region, Disease | LTRIM(RTRIM) test |
| **P3** | Duplicate Records | Potential duplicate records | All columns | COUNT(*) OVER PARTITION |
| **P4** | Missing Values | Check for NULLS | All columns | SUM(CASE WHEN… ) |

### DoctorDetails Table

| Issue ID | Type | Description | Affected Columns | Detection Method |
|---|---|---|---|---|
| **D1** | Naming Issue | Non-standard column names | DoctorID, YearsOfExperience | Manual inspection |
| **D2** | Trailing Spaces | In text fields | Name, Specialty, Hospital_Affiliation | TRIM test |
| **D3** | Duplicate Check | No duplicates found | All columns | Window COUNT |
| **D4** | Null Check | Check for missing data | All | SUM(CASE WHEN) |

### TreatmentRecords Table

| Issue ID | Type | Description | Affected Columns | How Found |
|---|---|---|---|---|
| **T1** | Naming Issue | Mixed-case/undocumented naming | Many columns renamed | Schema review |
| **T2** | Datatype Issue | Money → decimal conversion limitation | Treatment_Cost | Error on ALTER |
| **T3** | Trailing Spaces | Strings had spaces | Patient_ID, Doctor_ID, Outcome | TRIM test |
| **T4** | Duplicate Check | No duplicates found | All | Window COUNT |
| **T5** | Null Check | No NULLs | All | NULL scanning query |

## 3. Actions Taken

### PatientInfo

| Issue ID | Fix Applied | Tool Used | Fixed By | Date |
|---|---|---|---|---|
| **P1** | Standardized column names using sp_rename | SQL Server | Sooraj | 25-11-2025 |
| **P2** | Removed leading/trailing spaces | LTRIM(RTRIM()) | SQL | 25-11-2025 |
| **P3** | Duplicate removal logic applied (no duplicates found) | CTE with ROW_NUMBER | SQL | 25-11-2025 |
| **P4** | Verified no NULLs | CASE-based NULL scan | SQL | 25-11-2025 |

### DoctorDetails

| Issue ID | Fix Applied | Tool Used | Fixed By | Date |
|---|---|---|---|---|
| **D1** | Renamed columns to consistent format | sp_rename | Sooraj | 25-11-2025 |
| **D2** | Trimmed whitespace | UPDATE with TRIM | SQL | 25-11-2025 |
| **D3** | Duplicate check executed | Window COUNT | SQL | 25-11-2025 |
| **D4** | Null check performed (0 nulls) | CASE WHEN | SQL | 25-11-2025 |

### TreatmentRecords

| Issue | Fix Applied | Tool | Fixed By | Date |
|---|---|---|---|---|
| **T1** | Renamed columns for clarity | sp_rename | Sooraj | 25-11-2025 |
| **T2** | Converted money → decimal(10,2) | ALTER TABLE | Sooraj | 25-11-2025 |
| **T3** | Trimmed all string fields | TRIM | SQL | 25-11-2025 |
| **T4** | Duplicate check (none found) | COUNT OVER PARTITION | SQL | 25-11-2025 |
| **T5** | Null check (none found) | SUM(CASE WHEN) | SQL | 25-11-2025 |

### 4. Additional Notes / Follow-up Items

- No NULLs found across any tables.
- No duplicates detected, but duplicate removal code implemented for robustness.
- All tables now meet requirements to apply **Primary Key / Foreign Key Constraints**.
- Consistent naming conventions now follow:
  **PascalCase → Snake_Case**

### 5. Final Checks Before Analysis ✓

| Check | Status |
|---|---|
| Duplicate entries removed | ✓ No duplicates found |
| Missing data reviewed | ✓ No NULLs |
| Categorical values standardized | ✓ Gender, Region, Specialty |
| Numeric range validated | ✓ Age, Cost, Satisfaction Score |
| Date formats correct | ✓ YYYY-MM-DD |
| Column naming consistent | ✓ All snake_case |
| Tables ready for constraints | ✓ Yes |

### 6. File Version Tracking

| File Name | Description | Date Saved | Saved By |
|---|---|---|---|
| **PatientInfo.csv** | Original dataset | 24-11-2025 | Sooraj |
| PatientInfo_Cleaned | After renaming + NULL/duplicate checks | 25-11-2025 | Sooraj |
| PatientInfo_Cleaned | Final cleaned tables + ready for SQL constraints | 25-11-2025 | Sooraj |
| **DoctorDetails.csv** | Original dataset | 24-11-2025 | Sooraj |
| DoctorDetails_Cleaned | After renaming + NULL/duplicate checks | 25-11-2025 | Sooraj |
| DoctorDetails_Cleaned | Final cleaned tables + ready for SQL constraints | 25-11-2025 | Sooraj |
| **TreatmentRecords.csv** | Original dataset | 24-11-2025 | Sooraj |
| TreatmentRecords_Cleaned | After renaming + NULL/duplicate checks | 25-11-2025 | Sooraj |
| TreatmentRecords_Cleaned | Final cleaned tables + ready for SQL constraints | 25-11-2025 | Sooraj |

# Summary of Constraints Added to Cleaned Tables

After cleaning the dataset, structural integrity was enforced by adding Primary Keys, Foreign Keys, and CHECK constraints to ensure the data is valid, consistent, and relationally connected.

**1. DoctorDetails_Cleaned**

- **Primary Key**

- PK_DoctorDetails on Doctor_ID
  Ensures each doctor record is unique.

- **Check Constraint**

- Years_Of_Experience must be ≥ 0
  Prevents invalid negative experience values.

**2. PatientInfo_Cleaned**

- **Primary Key**

- PK_PatientInfo on Patient_ID
  Ensures each patient has a unique identifier.

- **Check Constraints**

- Age must be between 0 and 120
  Ensures realistic patient age.

- Gender must be Male, Female, or Other
  Standardizes categorical values and blocks invalid entries.

**3. TreatmentRecords_Cleaned**

- **Primary Key**

- PK_TreatmentRecords on Record_ID

- **Foreign Key Relationships**

- Patient_ID → PatientInfo_Cleaned(Patient_ID)
  Ensures every treatment is linked to a valid patient.

- Doctor_ID → DoctorDetails_Cleaned(Doctor_ID)
  Ensures every treatment is linked to a valid doctor.

These relationships create a proper **ER structure** and maintain referential integrity.

➢ **Check Constraints**

- **Treatment_Duration_Days ≥ 0**
  Blocks negative durations.
- **Treatment_Cost ≥ 0**
  Prevents invalid negative billing.
- **Satisfaction_Score between 1 and 10**
  Ensures valid rating scale.
- **Outcome must be Recovered / Critical / Ongoing**
  Standardizes clinical outcomes.