Change	Pre-Cleaning Data	Post-Cleaning Data
C1	Some columns that are supposed to hold numeric data, specifically **Protein1, Protein2, Protein3**, and **Protein4,** contained blank cells or non-numeric values. For example, the values in **Protein1** were **[5.2, , 7.8, invalid, 3.0]**	**All blank or non-numeric values in the mentioned columns were replaced with NULL, ensuring that these columns are clean and ready for further processing. For instance, after the modification, the values in Protein1 became [5.2, NULL, 7.8, NULL, 3.0]**
C2	The columns **Protein1, Protein2, Protein3**, and **Protein4** were still not recognized **as numeric types** due to the presence of blank or non-numeric values. For example, **Protein1** values **were [5.2, NULL, 7.8, NULL, 3.0]**	**The columns were converted to Decimal type, allowing for precise numeric storage and calculations. After the modification, the values in Protein1 appeared as [5.200, NULL, 7.800, NULL, 3.000]**
C3	The columns **(Date of Surgery)** and **(Date of Last Visit)** were not in the proper date format, which could lead to errors in date-related calculations. For example, the values in **(Date of Surgery)** **were [01-01-2020, 15-06-2021, invalid, 03-12-2019]**	**The columns were converted to DATE type, allowing for proper date storage and manipulation. After the modification, the values in (Date of Surgery) appeared as [2020-01-01, 2021-06-15, NULL, 2019-12-03]**
C4	The **Age column** was **not** in an **integer format**, which could cause issues in any numeric operations performed on it. For example, the values were **[25, 30.5, 40, 'unknown']**	**The column was converted to INTEGER type, ensuring that all values are whole numbers. After the modification, the values in Age appeared as [25, 30, 40, NULL]**
C5	The **(Tumour Stage)** column contained non-**numeric values** (like 'I', 'II', 'III') and blanks that needed to be standardized for analysis. For example, values **were ['I', 'II', 'III', '', 'invalid']**	**A new numeric column (Tumor Stage Numeric) was created to represent these stages as integers. The values after modification were [1, 2, 3, NULL, NULL]**
C6	Some records had invalid dates (e.g., 1900-01-01), which needed to be removed to maintain data integrity. For example, **the records were [{(Date of Surgery): '1900-01-01', (Date of Last Visit): '1900-01-01'}, {...}]**	**All records with invalid date values were deleted from the dataset. After the modification, only valid records remained**
C7	The **(Patient Status)** column contained blank values that needed to be standardized. For example, **the values were ['Active', '', 'Inactive', ' ']**	**All blank values in (Patient Status) were replaced with 'Unknown'. After the modification, the values became ['Active', 'Unknown', 'Inactive', 'Unknown']**
C8	**The (year of surgery)** needed to be extracted from the **(Date of Surgery)** column for easier analysis The initial values were dates like **[03-12-2019 ,15-06-2021 ,01-01-2020]**	**A new column (Year of Surgery) was added, extracting the year from (Date of Surgery)** **The values in this new column were [2019 ,2021 ,2020]**
C9	**In the Date of Last Visit column, the date 01/01/1900 appeared, indicating a missing or undefined visit. Additionally, some entries showed the year 2026, which is an invalid future date for a visit** **Example** **Date of Last Visit: 01/01/1900, 2026**	**The entries with the date 01/01/1900 have been removed, as they represent invalid or missing visits. The future date 2026 has also been corrected to reflect a valid past date. Now, the Date of Last Visit only contains accurate dates that match the real visit times of patients** **Example** **Date of Last Visit: NULL, 2023**